THESIS SUBMITTED IN PARTIAL FULFILMENT OF THE REQUIREMENTS FOR THE DEGREE OF DOCTOR OF PHILOSOPHY

Statistical HLA type imputation from large and heterogeneous datasets

Supervisor: Professor Gil McVean

Trinity 2012

Alexander Dilthey

Statistical Genetics

St Peter's College, Oxford

Department of Statistics, University of Oxford

To my parents,

for always encouraging my curiosity.

Abstract

Statistical HLA type imputation from large and heterogeneous datasets

Alexander Dilthey
St Peter's College
Submitted in partial fulfilment of the requirements for the degree Doctor of Philosophy.
Trinity 2012
Supervisor: Professor Gil McVean.

An individual's Human Leukocyte Antigen (HLA) type is an essential immunogenetic parameter, influencing susceptibility to a variety of autoimmune and infectious diseases, to certain types of cancer and the likelihood of adverse drug reactions.

I present and evaluate two models for the accurate statistical determination of HLA types for single-population and multi-population studies, based on SNP genotypes. Importantly, SNP genotypes are already available for many studies, so that the application of the statistical methods presented here does not incur any extra cost besides computing time.

HLA*IMP:01 is based on a parallelized and modified version of LDMhc [Leslie et al., 2008], enabling the processing of large reference panels and improving call rates. In a homogeneous single-population imputation scenario on a mainly British dataset, it achieves accuracies (posterior predictive values) and call rates $\geq$88% at all classical HLA loci (*HLA-A*, *HLA-B*, *HLA-C*, *HLA-DQA1*, *HLA-DQB1*, *HLA-DRB1*) at 4-digit HLA type resolution.

HLA*IMP:02 is specifically designed to deal with multi-population heterogeneous reference panels and based on a new algorithm to construct haplotype graph models that takes into account haplotype estimate uncertainty, allows for missing data and enables the inclusion of prior knowledge on linkage disequilibrium. It works as well as HLA*IMP:01 on homogeneous panels and substantially outperforms it in more heterogeneous scenarios. In a cross-European validation experiment, even without setting a call threshold, HLA*IMP:02 achieves an average accuracy of 96% at 4-digit resolution ($\geq$91% for all loci, which is achieved at *HLA-DRB1*). HLA*IMP:02 can accurately predict structural variation (*DRB* paralogs), can (to an extent) detect errors in the reference panel and is highly tolerant of missing data. I demonstrate that a good match between imputation and reference panels in terms of principal components and reference panel size are essential determinants of high imputation accuracy under HLA*IMP:02.

Acknowledgements

I would like to thank my supervisor, Professor Gil McVean, for his supervision, guidance and encouragement. Thank you Gil, working with you has been a great experience, and I would always do it again!

Without Dr "SB" Loukas Moutsianas, my closest collaborator, this thesis would look quite differently – and without you, importantly, getting here would have been much less fun.

Zam and Denise, thank you for being really good (sometimes very dramatic, but never boring) company in my endeavours.

Volker, Marko, Marcus, Shobha, Elsbeth, Laura – without you, Oxford wouldn't have been the same.

Franzi, thank you. In general and for introducing me to the secret of cookies.

Max, Markus, Chris ("Gagi") – the Wine & Philosophy group in Düsseldorf. Jan and Kevin – Wine & Philosophy in other places.

Last, but certainly not least, I would like to thank my parents for their love and support, and my aunt Daniela.

Contents

List of Figures

Chapter 1

Introduction

It has long been known that an individual's HLA type is an essential immunogenetic parameter: it describes the primary structure (and therefore the biochemical binding affinities) of the highly polymorphic classical HLA proteins, which present endo- and exogenous antigens to immune cells, and defines transplant immunocompatibility. HLA types are determined by the allelic state of the classical HLA genes in the Major Histocompatibility Complex (MHC) on the short arm of chromosome 6 [Janeway, 2001], and it is well established that HLA types and genetic variation in the surrounding MHC are associated with many biomedically relevant phenotypes.

This is true for many autoimmune and infectious diseases [Blackwell et al., 2009; Hill, 2001], susceptibility to certain types of cancer [Brennan and Burrows, 2008; Moutsianas et al., 2011] and the likelihood of adverse drug reactions [Chung et al., 2007]. In many autoimmune diseases, the HLA contributes a major fraction of genetic disease risk, often as much as 30 to 40% [Shiina et al., 2004]. Despite many attempts to characterize the role of genetic variation in the MHC influencing disease risk, most of the described associations, in particular those in complex autoimmune diseases, have remained elusive.

A statistical geneticist's role in solving that puzzle is quite confined: to identify phenotypically associated genetic variants in the MHC or combinations of variants, in order to offer guidance to functional studies of the underlying biology. Even this, however, is not as simple as one may imagine:

- First, the phenotypes under study are complex, multifactorial and therefore inherently hard to study. For example, there is evidence for secondary HLA-based genetic effects and HLA-based interactions in some autoimmune diseases [Evans et al., 2011; Strange et al., 2010], and many studies report (often contradictory) evidence for gene-environment interactions. Large study sample sizes will typically be required to have the statistical power to characterize these effects, and sometimes cross-country sampling can be helpful in studying assumed gene-environment interactions.

- Second, the MHC region is highly geographically stratified. Any multi-country study of HLA variation which fails to control for population stratification is prone to reporting misleading results [Altshuler et al., 2008; Price et al., 2006] (and unfortunately this applies to many studies of the past).

- Third, the MHC contains a number of highly polymorphic loci which exhibit unusual and strong patterns of linkage disequilibrium, in particular between the classical HLA genes. Studies should thus try to genotype as many of these polymorphic loci as possible, otherwise the likelihood of a reported association being driven by LD with an untyped variant is quite high: although this is a general concern in genetic association studies, the MHC's specific LD structure makes it quite possible that this untyped variant may be located in an entirely different gene, potentially confusing any functional interpretations [Bodmer and Bonilla, 2008].

In summary, to detect and disentangle LD, interactions and secondary effects in the MHC, sampling many samples from possibly multiple countries and exhaustively genotyping them in the MHC would be the most promising strategy. Unfortunately, however, there is no genotyping technology which meets these requirements for the classical HLA genes, which are the MHC's *a priori* most interesting genes and most consistently found to be associated with relevant phenotypes. Although knowing the HLA types of samples would be hugely beneficial in many studies, getting there for sufficient cohort sizes is not easy.

Why is this? Classical HLA typing technologies, which can reliably determine HLA types by targeted sequencing or hybridization, are typically too cost- and labour-intensive to be employed in studies with thousands of participants. SNP genotyping, on the other

hand, is more cost-effective and the current genotyping technology of choice for genome-wide association studies – many allelic variants at the classical HLA genes are, however, not adequately captured by standard SNP genotyping interpretation methods like tagging [de Bakker et al., 2006].

In this thesis, I improve, develop and validate methods for the statistical imputation of HLA types, based on SNP genotypes. These methods are particularly useful in situations where SNP genotyping for study samples is already available and small individual error probabilities can be tolerated: that is, they are particularly useful in genome-wide disease association studies, and have already proved their utility in studying the genetic architecture of autoimmune diseases.

How does HLA type imputation work? Genetic information is transmitted in chromosomal units (haplotypes), which can be broken up by recombination. There are well-established statistical models for capturing haplotype structure, which can be used to assess the degree of relatedness between two chromosomes or chunks of chromosomes. Therefore, if it is possible to use a reference panel of SNP- and HLA-genotyped individuals to learn about the joint haplotype structure of SNPs and HLA types (alleles), it should be possible to leverage this information into an imputation panel consisting of SNP-only genotyped individuals.

The most important requirements for an HLA type imputation technology are clearly accuracy and applicability (including computational tractability). As obvious from the previous discussion, both criteria should ideally extend to multi-country studies. How can we determine whether a technology is well-behaved with respect to these requirements? To formalize this discussion, suppose that a reference panel X should be used to impute HLA types into an imputation panel Y. Now, imagine that there is a particular individual $y \in Y$. We will only be able to impute y reliably if the information in X can be used to this end, i.e. if X contains information on what the two haplotypes of y could look like. From population genetics theory, we know that the chance of this being the case is maximized if we have many samples in X with a short time to the most recent common ancestor of them and y. We can achieve this by expanding X, in a way that makes X structurally

similar to Y (i.e. proportionally sampling new individuals from the populations present in Y – this is the case whether Y is a multi-population sample or not; the diversity of HLA alleles present in single populations is large enough to demand for large reference panels). This will obviously increase the size and complexity of X.

This allows for the definition of a criterion for well-behaved HLA type imputation methods: any well-behaved method should be able to *use* the increased size and complexity of larger reference panels to increase imputation accuracy, given that imputation and reference panels are well-matched. Note that this criterion applies to the technical (mainly computational) as well as the modeling (dealing with more complex haplotype structure) challenges in HLA type imputation.

To illustrate this point, I will define three exemplary scenarios which provide a guide to what this thesis aims to achieve (and which later chapters will make reference to):

- Scenario 1: X and Y are from essentially the same population, for example imputing from a mainly British reference panel into British samples. Extending X is more challenging computationally than in terms of dealing with population structure within X, as all samples come from the same population.

- Scenario 2: X and Y are from the same ethnicity, for example imputing from a cross-European reference panel into a cross-European imputation cohort. X needs to be larger than in the first example and has, in order to accommodate for the increased diversity in Y, to be more complex as well. Both computational and modeling challenges need to be overcome. We say that X is of *medium heterogeneity*.

- Scenario 3: X and Y are structurally matched, but both contain samples from multiple ethnicities. The modeling challenges can be expected to outweigh the computational challenges. We say that X is of *high heterogeneity*.

All three scenarios are relevant to disease association studies; to an extent, Scenario 2 is the most important one. Scenario 3 can probably be solved to an extent based on Scenario 2 (by splitting X and Y according to ethnicity estimates, and by applying the methods from Scenario 2 separately to each group), but is still relevant for practical reasons and for

individuals whose ancestry cannot be unambiguously assessed. However, it seems quite unlikely that it would be possible to assemble population-specific panels for each cohort present in a study (in many current studies, this would require 10 or more country-specific panels).

Developing methods which effectively deal with the challenges posed by these three scenarios is the major aim of this thesis, to make HLA type imputation as useful as possible in a variety of biomedical study contexts.

In *Chapter 2*, I will provide a comprehensive review of the HLA's immunological role, the structure of the MHC, imputation models and algorithms.

In *Chapter 3*, I describe HLA*IMP:01, which is based on extensions to an existing algorithm (LDMhc, Leslie et al. [2008]), increasing call rates and enabling the processing of large datasets by parallelization. HLA*IMP:01 is well suited to deal with Scenario 1.

In *Chapter 4*, I describe HLA*IMP:02, a novel HLA type imputation method, which is based on haplotype graph models. HLA*IMP:02 avoids averaging of allelic backgrounds and allows for detection of errors in the reference panels; it is well suited for Scenario 2.

In *Chapter 5*, I compare the methods' performance on heterogeneous reference panels (Scenarios 2 and 3) and show that HLA*IMP:02 deals much better with heterogeneity; in particular, it fulfills the "more data, higher accuracy" condition formulated for well-behaved models.

In *Chapter 6*, I describe how imputation-derived genotype uncertainty can be integrated into the standard statistical Bayesian framework for disease association. I also carry out a small simulation study, comparing power to detect disease-associated alleles under HLA*IMP:01 and HLA*IMP:02, and present some results on a real dataset.

It is worth pointing out that the methods presented here have already been applied in a variety of studies, some of which I participated in.

HLA*IMP:01 was used in three recent Wellcome Trust Case Control Consortium 2 studies, of psoriasis [Strange et al., 2010], ankylosing spondylitis (AS, Evans et al. [2011]) and multiple sclerosis (MS, Sawcer et al. [2011]). I provide a sketch of the results on MS in

Section 6.3.

HLA*IMP:01 has also been used in studies of cancer [Hosking et al., 2011; Moutsianas et al., 2011] and on the connection between genetic risk factors of HIV and psoriasis [Chen et al., 2012]. Other publications employing HLA*IMP:01, with no involvement from myself or close collaborators, start appearing at the time of writing (Davies et al. [2012], for example). HLA*IMP:02 will be used in the analyses of the Immunochip consortium, and, despite the web service implementing it still officially being in "beta" status, many external research groups have started using it.

Chapter 2

Background

2.1 The classical HLA genes and the MHC region

In many autoimmune and infectious diseases, genetic variation in the Major Histocompatibility Complex (MHC) region of the human genome on the short arm of chromosome 6 is the strongest predictor of disease risk or outcome. The region is also associated with the likelihood of adverse drug reactions and susceptibility to certain types of cancer. Often, these associations map to the classical Human Leukocyte Antigen (HLA) genes, which play an essential role in activating and mediating adaptive immune responses. Perhaps it should therefore not come as a surprise that the MHC exhibits remarkable genomic features (most notably hypervariability and long-range linkage disequilibrium), which are often interpreted as the result of selection (in the context of this thesis, "selection" always refers to "natural selection").

This section aims at providing the necessary background knowledge to understand the HLA's crucial role in immunity, and how variation in the MHC was shaped by evolutionary forces and is involved in disease risk.

2.1.1 The classical HLA genes

There are six classical HLA proteins: A, B, C, DQ, DR, DP. They present antigens, i.e. immunologically recognizable peptide fragments, to immune cells.[1]

We distinguish between two classes of antigen-presenting classical HLA proteins:

- HLA class I proteins (A, B, C) are integrated in the cell membrane of all nucleated cells and present endogenous antigens to white blood cells (leukocytes), i.e. antigens originating from the inside of the cell. Typically, these antigens are characteristic of the cell's inner state, as a molecular cleavage and transportation machinery loads the class I proteins with fragments of arbitrarily sampled peptides from the cytoplasm. For example, if a virus has infected a cell and is using it for reproduction, the HLA class I molecules will start presenting some characteristic fragments of the virus' peptides: the infection has become visible to the outside. The host immune system is now enabled to launch countermeasures, e.g. kill the infected cell.

- HLA class II proteins are located in the membrane of certain white blood cells and present exogenous antigens, i.e. antigens originating from outside the cell, to other leukocytes. Again, a cleavage machinery is involved, which degrades the exogenous molecules to smaller fragments before presenting them. If, for example, a macrophage is digesting a recently uptaken (phagocytosed) bacterium, surface proteins of this bacterium can appear on the HLA class II proteins. That – if it happens often enough to indicate an ongoing infection – triggers a complicated cascade of immune reactions that are mounted specifically against the infecting type of bacteria, e.g. by the proliferation of cells that produce specific antibodies.

- Of note, the proteins sometimes referred to as "HLA class III" proteins have no antigen-presenting role and are therefore not discussed here

For now this should suffice to convince the reader of the outstanding importance of the

[1] Although it was for a considerable time assumed that the proteins listed there (the so-called "classical alleles') represent all antigen-presenting molecules that are encoded in the MHC region, this view does no longer hold [Kumanovics et al., 2003]. However, the classical alleles are still the best-characterized and best-understood antigen-presenting proteins; and they seem to be a major driver of the many phenotype associations found in the MHC.

HLA proteins. A detailed functional description in the context of immunity is given later. The remainder of this section is devoted to describing the basic *genetics* of the classical HLA proteins and set out HLA terminology.

All HLA molecules are heterodimers, i.e. they are composed of two subunits. In the case of the class I molecules, the α-subunit (or -*chain*) is variable (i.e. polymorphic in the human population), and the β-subunit (β2-microglobulin) is not (in the sense that the same chain is used in all individuals and for the A, B and C molecules). The class II molecules consist of a variable α- and a variable β-chain. Each chain peptide is encoded as a separate gene. All *variable* genes involved reside in the MHC (see Section 2.1.3 for details).

Interestingly, expression of the variable genes is codominant, i.e. the two copies each human carries on its two homologous chromosomes 6 are both active (for this paragraph and the remainder of this section, Janeway [2001] is the principal reference).

After the two principal subunits have been synthesized, they combine, together with other parts such as the *invariable chain*, and appear on the cell surface. The cells of most individuals present six types of distinct HLA class I molecules: 3 loci x 2 α-chain alleles x 1 invariable β2-microglobulin. The number of possible distinct class II molecules is higher and depends on the variability of the two chains in the human population. There are, for example, 46 known DQα alleles and 160 known DQβ alleles (see Table 2.1, according to IMGT/HLA, Robinson et al. [2003]), so that one individual could well carry two distinct alleles for both chains. As they are under codominance, one cell may exhibit four different isoforms of the DQ protein, plus variants for DP and DR.

Sometimes confusingly, when the term "HLA" is used, it may refer to a variety of different concepts:

- the final HLA proteins in their assembled form
- the pre-final chain peptides
- or the region in the human genome which encodes the structure of the variable chains (the term MHC also refers to the general mammalian homologue; if applied to humans, it is synonymous to this usage of the term "HLA")

"HLA types" (or synonymously: "HLA codes") are used to distinguish between variants of the α- and β-subunits or their coding sequences. They are specified according to a given syntax which includes the locus they refer to; a star is used to differentiate locus from allele code, and double colons within the allele code are used to separate different levels of resolution. The *digit level* or *resolution* of an HLA type can be inferred from the number of double colons and specifies the biological features that are captured by the code:

- 2-digit HLA types describe the general serological/antigenic features of the subunits.
- 4-digit HLA types describe the sequence (primary structure) of amino acids that the subunits consist of. Therefore, they are a complete specification of the subunits' molecular features.
- 6-digit HLA types refer to the structure-coding DNA sequences of the subunits.
- 8-digit HLA types refer to the complete DNA sequences at the subunits' loci in the HLA region, including polymorphisms in the noncoding regions

In this thesis, most codes are specified at 4-digit level; a typical example for notation is HLA-A*02:01.

Every HLA code at a lower resolution refers to a set of higher-resolution HLA types: an n-digit HLA code determines the $n-2$-digit HLA code. Usually, a 4-digit type's 2-digit classification is determined by the first two digits of the 4-digit code, e.g. HLA-A*02:01 becomes HLA-A*02.[2]

In HLA nomenclature, α and β are used for protein names, whereas their Latin equivalents ("A" and "B") refer to the corresponding genes and HLA codes.

Table 2.1 gives an overview of nomenclature and the relationships between HLA code identifiers and synthesized peptides.

[2]However, this rule does not always apply; there are, for example, 4-digit HLA types whose serological properties are not well characterized, so that their "true" 2-digit code may not be equivalent to the first two digits of the 4-digit-level code. The $n-2$ digit type is then of course still determined by the n digit type, but not in a straightforward way. See Holdsworth et al. [2009] for details.

HLA gene (and code identifier)	Corresponding protein	Number of known alleles
HLA-A	HLA-A protein, α-chain	1729
HLA-B	HLA-B protein, α-chain	2329
HLA-C	HLA-C protein, α-chain	1291
HLA-DPA1	HLA-DP protein, α-chain	33
HLA-DPB1	HLA-DP protein, β-chain	150
HLA-DQA1	HLA-DQ protein, α-chain	46
HLA-DQB1	HLA-DQ protein, β-chain	160
HLA-DRA	HLA-DR protein, α-chain	7
HLA-DRB1	HLA-DR protein, β-chain	1150

Table 2.1: Classical HLA type identifiers and how they relate to the HLA proteins – source for the number of alleles: http://www.ebi.ac.uk/imgt/hla/stats.html, as of 22/11/2011

2.1.2 Determining HLA types

There are a variety of techniques to determine HLA types at different levels of resolution [Dunn, 2011]. Modern methods rely on the polymerase chain reaction (PCR) to survey sequence variation at the classical HLA loci to varying degrees of accuracy (older serological methods are based on examining the biomolecular properties of the HLA proteins themselves; they are less accurate and not commonly employed anymore [Middleton, 1999]):

- SSP (Sequence Specific Primers) is based on primers that bind to predetermined allele sequences or groups of alleles; it usually allows for reliable discrimination of alleles at 2-digit resolution.
- SSOP (Sequence Specific Oligonucleotide Primers) is based on probes that are specifically designed to hybridize with certain alleles or groups of alleles; it usually enables reliable allelic discrimination at 2-digit resolution.
- So-called "intermediate resolution" typing assays, usually based on SSP, enable reliable allelic resolution at 2-digit level for all major allelic groups and also deliver 4-digit information for some alleles.
- SBT (Sequence Based Typing) is based on specific primers to amplify the code-defining regions of the HLA genes (usually the exons encoding the antigen-binding sites of the HLA proteins). It is considered to be the most accurate typing technology and can often fully discriminate between alleles at 4-digit levels (in some cases, it still results in ambiguous calls; some 4-digit alleles are identical at the sequence positions

usually targeted by SBT, but in most cases these alleles exhibit very similar binding affinities).

Accuracy differences between these methods beyond the recognized systematic ambiguities are influenced by the employed lab and sample preparation protocols; to my knowledge, there are no generally accepted statistics on systematic re-test accuracy differences between these methods.

2.1.3 The extended human MHC

The definition of sensible boundaries for what should be denoted "MHC" in the human genome was subject to some scientific controversy. Undisputedly, the term "MHC" refers to a region on the short arm of chromosome 6, encoding the classical antigen-presenting HLA proteins A, B, C, DQ, DR, DP and a variety of other genes, many of them also being related to the immune system. Horton et al. [2004] present a detailed gene map of the region. They also specify the term "extended MHC" (xMHC) to describe a region of immunity-related genes, spanning from *SLC17A2* to *DAXX*, which is used in the literature and adopted here. To give a specific definition for all further analyses, the xMHC region shall be defined as spanning from position 25,921,129 to position 33,535,328 in the human genome build 36 (based on Horton et al. [2004]). The xMHC can be divided into the following sub-regions [Horton et al., 2004], ordered from centromere to telomere:

- an extended class I region, $\sim$ 3.6 Mb
- the classical class I region (containing the genes for the classical class I proteins), $\sim$ 1.9 Mb
- the classical class III region, $\sim$ 0.7 Mb
- the classical class II region (containing the genes for the classical class II proteins), $\sim$ 0.9 Mb
- the extended class II region, $\sim$ 0.2 Mb

, spanning 7.6 Mb in total. Of the 421 distinct loci defined by Horton et al. [2004], 282 seem to be transcriptionally active genes, and 139 appear to be pseudogenes. 28% of the functional genes are assumed to be related to immune function. Remarkably, the class III region seems to be the region in the human genome with the highest gene density [Horton et al., 2004].

Besides the loci coding for the classical HLA proteins, a variety of other genes is located in the MHC, for example

- genes for the non-classical class I molecules HLA-E, -F, -G, -HFE, which are not completely understood. Some of them are polymorphic, and some of them play a role in antigen presentation, but usually not in all cell types and not as abundantly as the classical class I molecules [van den Elsen et al., 2004].
- the class I-like genes *MICA* and *MICB*
- *TAP1* and *TAP2*, genes related to antigen transport and class I biosynthesis
- genes for the non-classical class II molecules DM and DO/DN, involved in loading the classical class II molecules
- tRNA-coding sequences
- olfactory receptor molecule sequences
- *TNF* (Tumor Necrosis Factor) genes

(Flajnik and Kasahara [2001], Kumanovics et al. [2003] and Horton et al. [2004])

Figure 2.1 summarizes the organization of the human MHC.

On a genomic scale, probably the two most interesting features of the MHC are the high degree of polymorphism and the long-range haplotype structure. Many genes in the MHC harbor polymorphisms. This applies to the classical HLA genes (*HLA-B*, for example, is hypervariable with more than 2000 known alleles) and other genes in the region [Horton et al., 2008]. The study of Horton et al. [2008], based on eight completely sequenced MHC haplotypes, has significantly contributed to our knowledge on polymorphisms and

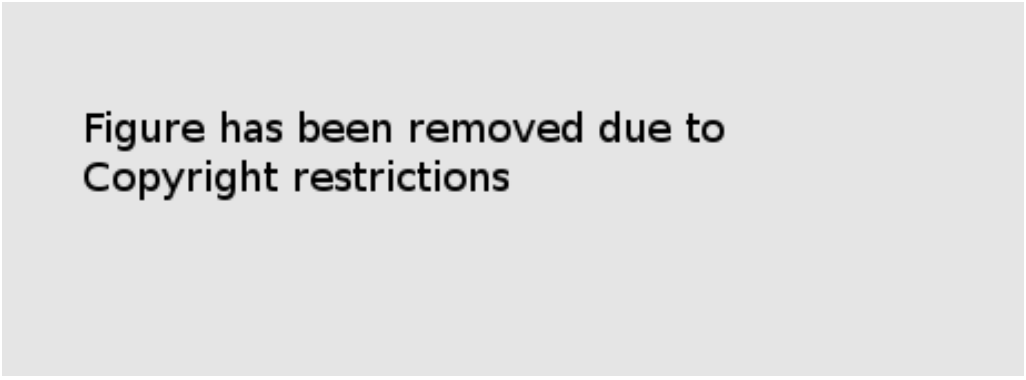

Figure 2.1: Illustration of the genomic structure of the human MHC, displaying the positions of the classical HLA and TAP genes. Figure reproduced from Janeway [2001].

MHC haplotype structure, and on structural variation (SV) in the region in particular. Interestingly and maybe counterintuitively, the MHC exhibits strong long-range linkage disequilibrium on a population scale, sometimes over a couple of megabases, apparent from single nucleotide polymorphisms (SNPs) as well as alleles at the classical loci [de Bakker et al., 2006]. Some classical HLA class II alleles, for example, are in LD with SNPs and classical alleles in the class I region, and certain haplotypic combinations of *HLA-DRB* duplications (active and pseudogenic) are nearly always transmitted together. This is consistent with the MHC being a region of low recombination rates [de Bakker et al., 2006]. Taken together, these features suggest that the MHC is a region of extraordinary allelic diversity and haplotypic stability at the same time. This is often related to the region's role in immunity and the selective pressures probably arising from that (see below for some specific arguments and information on global MHC population divergence).

For what follows, it is important to appreciate that these genomic features make it sometimes difficult to statistically or directly genotype variants in the region, in particular with more recently developed genotyping methods. "Next-generation" sequencing generates relatively short reads, which are potentially difficult to interpret in a region with pronounced haplotypic differences and abundant structural variation. SNP genotyping sometimes suffers from quality problems for the same reasons, and substantial amounts of genetic variation in the MHC, in particular the classical HLA genes, are not captured by SNP-based standard methods such as tagging [de Bakker et al., 2006]. Even many of the wetlab-based methods that were specifically developed to genotype the allelic state of the

classical HLA genes draw on statistical models to distinguish between some high-resolution allele codes, for example allele-specific hybridization. In summary, Sanger sequencing remains the gold standard for surveying variation in the MHC, for the classical HLA alleles (with targeted primers) as well as on an MHC-wide scale [Horton et al., 2008].

2.1.4 A functional view: The HLA's role in immunity

At the beginning of this chapter, I have already alluded to the important role the classical HLA proteins play in establishing adaptive immunity. Here, this role is explored in more detail, at the example of a schematic immune reaction. The description has, due to the enormous complexity of the immune system, to remain partial, and is based on Janeway [2001].

In immunology, it is helpful to distinguish between *innate immunity* and *adaptive immunity*. The components of the immune system that are responsible for innate immunity are steadily active; adaptive immunity is in contrast activated by a specific pathogenic threat and takes 4 - 5 days to become effective. During this period, innate immunity has to control and restrict the spread of an infection; its components are also involved in activating the adaptive immune response. Most known physiological interactions that involve the HLA proteins happen within the scope of adaptive immunity, though it has been found that at least the class I molecules also play a role in innate immunity, for example in interaction with the killer-cell immunoglobulin-like receptors (KIRs).

To cope with the complexity of immune reactions, it is helpful to regard an exemplary case. What happens if a typical pathogen enters the human body? Two important components of innate immunity, *macrophages* and *neutrophils*, are steadily present in blood and tissue. If they are able to recognize the pathogen as a threat, they will attack it by phagocytosis (therefore, these cells belong to the class of *phagocytes*). Macrophages and neutrophils possess receptors that allow them to detect typical pathogen surface characteristics, e.g. bacterial lipopolysaccharide (LPS). Another component of innate immunity, the *complement system*, is aiding detection: blood plasma proteins that bind to typical pathogen characteristics, like mannose-containing carbohydrates, are recognized by phagocytes in

bound form. Phagocytes that are attacking pathogens start emitting signaling molecules that attract other phagocytes and cause inflammation.

Most pathogens are neutralized by this first line of immune defense. However, some pathogens enter the body in a high concentration, overwhelming initial immune reactions, or employ strategies to escape detection. This is where adaptive immunity comes into play. An adaptive immune reaction usually begins with peripheral dendritic cells (DCs) attracted by signals of inflammation: they steadily take up particles that they, via surface receptors, recognize as pathogenic, as well as arbitrary particles from their environment (e.g. host proteins, camouflaged pathogens), degrade them, and start presenting them on their MHC class II proteins. Thus, at some point, the DC will start presenting fragments that are characteristic of the pathogen. Dendrites also possess intracellular receptors to examine the cleaved fragments for pathogen characteristics: unmethylated CpG dinucleotide motifs typical of bacterial DNA, for example, activate TLR-9 from the Toll-like receptor family. Activation of any of the receptors sensitive to pathogen characteristics will induce a switch to a "mature" DC phenotype: this includes activation of particular cell surface molecules which will later indicate to other immune cells that the DC sensed pathogen characteristics, an upregulation of the density of MHC surface molecules, and migration to lymph nodes, where the DC will meet other agents of adaptive immunity. Note that the biochemical binding affinities of the HLA class II proteins of the dendrites determine which parts of a pathogen are preferentially presented. Interestingly, dendrites are also able to present exogenous antigens on their class I molecules (*cross-presentation*), thus representing an exception to the rule that class I molecules are exclusively used for endogenous peptides (see Figure 2.2).

As a next step, CD4+ helper T cells are activated. Helper T cells themselves do not combat pathogens, but they recruit and coordinate other immune cells (killer T cells, macrophages and B cells) that are specifically effective against the pathogen that triggered adaptive immunity. T cells are generated in the thymus and possess a so-called T Cell Receptor (TCR), which can bind to specific antigen+MHC molecule complexes. The genes coding for the structure of the TCR are continuously altered while new T cells are being produced ("*V(D)J recombination*", mediated by the *RAG1* and *RAG2* recombinases), so that there

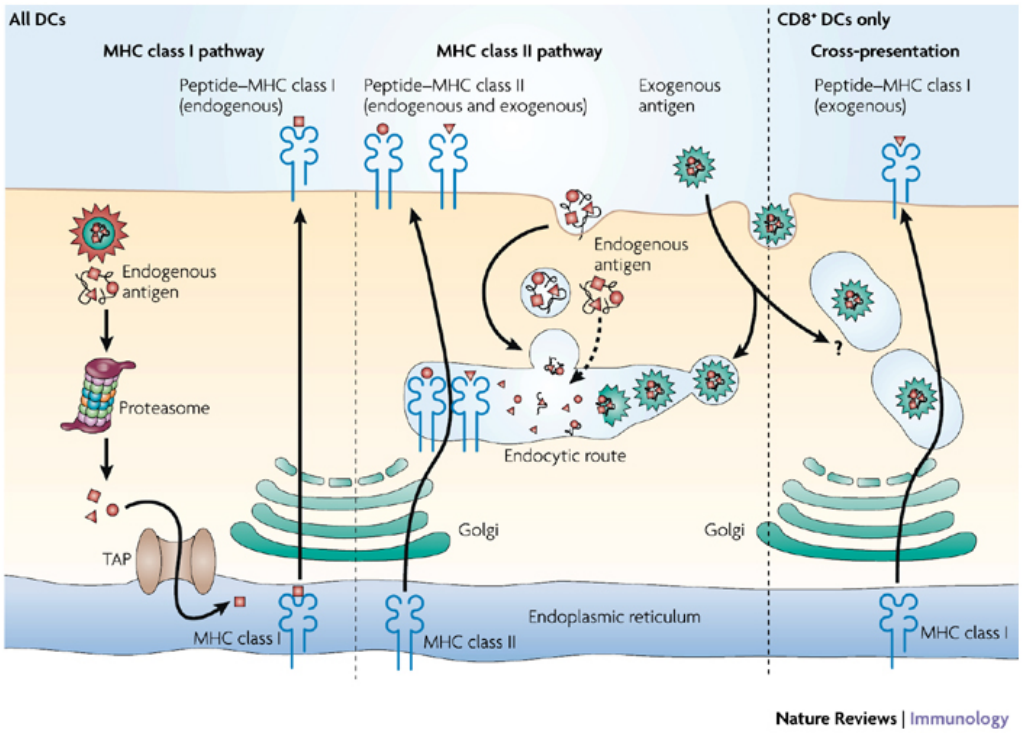

Figure 2.2: Dendritic cells (DCs) activate other immune cells by presenting antigens to them. This figures illustrates the different presentational pathways in DCs. Endogenous antigens are cleaved by proteasomes and TAP transports the fragments to the endoplasmatic reticulum (ER), where they are loaded onto MHC class I molecules and presented to CD8+ T cells. Exogenous antigens reside in endocytic vesicles, where they are cleaved. MHC class II proteins are transported from the ER to these endocytic vesicles; to ensure transport stability, they are loaded with the so-called "invariant chain" (not pictured), which is degraded upon arrival of the MHC class II protein in the endocytic vesicles. The freed MHC class II proteins can then bind to the cleaved antigen fragments and migrate to the cell membrane. This pathway is not completely isolated against the entry of endogenous antigens, so that MHC class II proteins can also present endogenous antigens. In certain types of DCs, there is a separate, not completely understood pathway for loading exogenous antigens on MHC class I proteins ("cross-presentation"). Only the cross-presentation pathway is exclusive to DCs. Figure reprinted from Villadangos and Schnorrer [2007]

exists a wide variety of T cells with different TCR structures. In fact, the range of antigen-specific TCRs that can be produced via V(D)J recombination is so large – $> 10^8$ – that there will usually exist T cells to bind characteristic antigens of every possible pathogen. A selection procedure takes place before the T cells leave the thymus and assures that newly generated TCR variants confirm to two criteria: during the first stage, known as *positive selection*, only those cells survive that have TCRs with a sufficient binding affinity to the host's HLA molecules. During the second stage, known as *negative selection*, all

cells that bind to MHC complexes loaded with normal host proteins are eliminated. That second step helps preventing autoimmune reactions, although not all proteins that exist in an organism are expressed in the thymus – and are therefore not presented in cleaved form on the cells that the immature T cells pass.

A helper T cell that can dock with its TCR to the loaded MHC class II proteins of a dendrite has a TCR structure that is able to recognize antigens that are specific of the putative pathogen just being processed by the dendritic cell (this usually happens in the lymph nodes). To make sure that the helper T cell is not going to become active against the signature of a host protein, the helper T cell also needs to bind to the dendrite's surface proteins that indicate that the dendrite recognized a pathogenic signature. If this does not happen, the T cell deactivates itself – a mechanism known as *peripheral tolerance*, preventing autoimmune reactions. If, on the other hand, the second signal – known as *verification* – is present, the naive helper T cell starts proliferating – known as *clonal expansion* – and differentiating. After a few days, its successors have become *armed helper T cells*.

Armed helper T cells search for the signature of their TCR, presented by MHC class II molecules – now without the need for a second verificatory signal. They usually find it on the surface of macrophages and B cells.

B cells produce antibodies (*immunoglobulins*) and are similar to T cells in that they possess a unique B Cell Receptor (BCR). The corresponding genes also undergo somatic recombination and hypermutation. Contrasting T cells, however, B cells are able to bind to their specific antigenic signature wherever they encounter it, without needing mediation of an MHC molecule – e.g. to a virus in the intercellular space. Whenever that happens, the B cell encloses the object bound to the BCR, reduces it to antigenic fragments and starts presenting them on its MHC class II proteins. An armed T helper cell with a TCR that is compatible with the antigens being presented by the B cell (not necessarily the same antigen that was bound by the BCR!) can now activate the B cell – the latter will start to produce antibodies specific against the signature of the B cell's BCR and undergo clonal expansion. In rare cases, B cells may recognize a pathogenic signature and start

producing antibodies without the help of a T cell just on encountering a BCR-compatible antigen. Antibodies, once bound to their target, can react with the complement molecules described earlier and thus trigger phagocytosis.

If an armed helper T cell encounters a macrophage that is presenting compatible antigens, it emits signaling molecules that make the macrophage produce more superoxide anions and oxygen and nitrogen radicals, increasing the macrophage's effectivity against pathogens.

It is possible to distinguish between two extreme types of armed T helper cells, Th1 and Th2. Whereas it was long assumed that this distinction is a strict one, it is now clear that there exist intermediate forms. Th1 cells preferentially activate macrophages, Th2 cells rather bind to B cells. This implies that the balance of Th1 / Th2 cells is an important determinant of the specific nature of an immune response, specifically to what extent it is based on the activity of B cells or macrophages. The factors that influence Th1 / Th2 differentiation are not fully understood, but a link to the density of MHC molecules on the (usually dendritic) cell that activates the naive T cell seems likely.

In addition to activating/stimulating the cells that they encounter, armed T cells (and immune cells in general) also emit a range of cyto- and chemokines with various effects, e.g. that stimulate the expression of the MHC molecules and foster inflammation. Despite their importance, signaling molecules are beyond the scope of this thesis.

So far, only MHC class II molecules were involved. How can the immune system react to viruses which infect host cells and are thus not directly accessible to B cells and macrophages? CD8+ armed killer T cells examine the antigens presented by MHC class I molecules – from the inside of the host's cells – and destroy the cell, if their TCR is compatible with the antigens presented. Killer T cells are, in terms of their origin in the thymus, very similar to helper T cells, i.e. in that new structural variants of the TCR are continuously generated and that they undergo two stages of positive and negative selection. Like helper T cells, killer T cells start their lives outside the thymus as naive, i.e. ineffective, cells. To become effective, i.e. to initiate the cascade of clonal expansion and differentiation, naive killer T cells need a second confirmatory signal apart from just

recognizing a compatible antigen+MHC complex – usually, this is either delivered by a dendritic cell (i.e., the DC is either itself infected by the virus or cross-presents the viral fragments, and its internal pathogen sensing receptors have been activated) or by a helper T cell that is bound to the same potential target cell. However, once the single naive cell has proliferated into an armada of armed killer T cells, the second confirmatory signal is no longer necessary to effectively eliminate a body cell presenting the right antigens.

Another subtype of T cells, regulatory T cells (T_{REG}), plays an important role in down-regulating immune responses and preventing autoimmunity. This has been established by T_{REG} cell depletion in mice, leading to fatal autoimmune reactions [Sakaguchi et al., 2007]. Is is also known that certain mutations in *FOXP3*, a central and necessary gene for conversion to T_{REG} cell status, lead to severe autoimmunity in humans [Gambineri et al., 2003]. In general, T_{REG} cells and their interactions with other cell types are not as well understood as other types of T cells [Sakaguchi et al., 2010].

The MHC class I proteins also play a role in non-adaptive immune responses, for example in interaction with natural killer T (NKT) and natural killer (NK) cells. NK cells kill cells which display abnormal cell states. Their activity is partly controlled by the KIR proteins, some of which are inhibitory and some of which are activatory [Vilches and Parham, 2002]. For example, the KIR2DL1 receptor binds to HLA-C proteins, which protects cells with normal MHC class I expression from being killed. On the other hand, cells in which class I expression is downregulated, for example by a cancerous process or viral infection, are not being protected. Activatory KIRs are not well understood. KIR2DS1, for example, binds to particular HLA-C alleles and has been linked to autoimmune disease risk, but it is not clear why an NK-activatory KIR recognizes HLA-C molecules. Interestingly, the genetic region that codes for the KIR proteins (chromosome 19) is also complex and exhibits strong haplotypic patterns, similar to the MHC [Uhrberg, 2005]. Most individuals carry haplotypes with a substantial number of KIR genes, but the majority of NK cells express only 1 - 3 (different) KIR proteins [Vilches and Parham, 2002]. The underlying process that determines the distribution of cells expressing different combinations of KIRs is assumed to be stochastic, but influenced by an individual's HLA type. This leads to an increase of the proportion of NK cells which express inhibitory KIRs compatible with

the individual's HLA type [Schoenberg et al., 2011]. NKT cells also kill cells and can be characterized as somewhere in between adaptive and innate immunity. They share characteristics of both T and NK cells. They express TCRs, which recognize predefined pathogenic or cancerous patterns via MHC class I presentation. In contrast to normal T cells, however, NKT TCRs are invariant, i.e. do not undergo somatic recombination and mutation [Pellicci et al., 2011].

As a general remark, it has to be stressed that the whole MHC-based antigen recognition system is based on binding compatibility – which doesn't necessarily imply perfect lock-and-key molecular interactions. In some cases, similar general structural features or just a similar subunit of a peptide may be sufficient to trigger binding of the TCR/BCR. This phenomenon is known as *cross-reactivity* and is often assumed to be one of the mechanisms underlying autoimmune disease (see below).

2.1.5 The HLA and disease

As early as in the 1960s, studies suggested an association between certain HLA types and susceptibility to cancer [Amiel, 1967; Bodmer and Bonilla, 2008]. In 1973, Bodmer [1973] first pointed out the importance of LD when studying markers in the HLA. The list of HLA-associated conditions has been expanding ever since. This includes autoimmune and infectious diseases, certain types of cancer and adverse drug reactions. For example, for common autoimmune diseases like rheumatoid arthritis [Deighton et al., 1989; McMichael et al., 1977; Stastny, 1978] and type 1 diabetes [Mein et al., 1998; Nerup et al., 1974; Singal and Blajchman, 1973], the HLA type determines more than 30% of the genetic risk. The HLA region is strongly associated with multiple sclerosis (MS) [Compston et al., 1976; Oksenberg et al., 2008; Terasaki et al., 1976]. It influences HIV disease progression and mortality [Leslie et al., 2004] and response to abacavir, an important HIV drug [Hetherington et al., 2002]. Some types as cancer, for example multiple hodgkin lymphoma [Moutsianas et al., 2011], exhibit HLA type associations.

A full review of HLA-based conditions would be beyond the scope of this section. Blackwell et al. [2009] provide a comprehensive list of associations between HLA types and infectious

diseases, including HIV, hepatitis B /C, leprosy, malaria and tuberculosis. Gebe et al. [2002] describe possible HLA-mediated autoimmune disease mechanisms on a molecular level, and Shiina et al. [2004] present an extensive summary of putative associations. Chung et al. [2007] describe some cases of HLA-associated drug hypersensitivity.

It seems likely that many of the observed associations are causally related to the allelic state of the HLA alleles: it is well known that different alleles exhibit different binding and therefore presentation affinities. For example, an allele that preferentially presents auto-antigens may contribute to increasing autoimmune disease susceptibility. However, as is generally the case is association studies, it is important to appreciate that association does not necessarily imply a causal link. The MHC harbors more than 60 immunity-related genes, and it could be that an observed association with a classical allele is actually due to LD with another causal variant. It is also well known that there exist functionally relevant polymorphisms in non-coding regions of the HLA: for example, allelic differences in the promoters of the class II genes show clear association with class II transcription levels [Glimcher and Kara, 1992], and some authors argue that these polymorphisms may influence disease risk by affecting the balance between Th1 and Th2 cells [Mueller-Hilke and Mitchison, 2006].

However, even if there is a causal effect of the classical HLA alleles, it is not always clear how this translates into disease risk on a functional level. For some viral infections like HIV, the association between HLA type and outcome is well characterized: the protective HLA allele B*57 enables an effective response of cytotoxic T cells against the virus and is associated with long-term HIV control [Leslie et al., 2004]. HLA-associated susceptibility to cancer is often related to the viral infections which are assumed to play a role in certain types of cancer [Amiel, 1967; Hosking et al., 2011; Moutsianas et al., 2011]. Adverse drug reactions are thought to occur because some HLA types can present the active drug components, or complexes of the drug and self antigens, enabling a T cell based response against the drug. For autoimmune diseases, the picture is substantially less clear. Many theories are presented in the literature, and it is worth reproducing some of them here:

- Disease-associated HLA alleles preferentially bind to self antigens, for example HLA-

DRB1:15*01 to Myelin Basic Protein (MBP) in MS [Sospedra and Martin, 2005]

- Disease-associated HLA alleles do not bind well to particular self antigens (auto-antigens); negative selection in the thymus against the T cells which bind to these antigens is therefore less effective [Sospedra and Martin, 2005]

- In combination with the last two hypotheses, sharing or not sharing of a particular sequence of peptides (the "shared epitope") may determine whether an HLA allele is disease-predisposing or not [Sospedra and Martin, 2005]. The shared epitope may be present in HLA alleles from different 2-digit groups.

- Involvement of pathogens: peptides of the Epstein-Barr virus (EBV) are structurally similar to myelin basic protein (MBP) when presented on HLA-DR proteins, and it is likely that MBP is one of the targets of the autoimmune reaction that defines MS. T cells with cross-reactive TCRs, binding to HLA+EBV and HLA+MBP complexes, may be activated upon EBV proliferation and become autoreactive [Sospedra and Martin, 2005]. Sometimes, this situation is referred to as "molecular mimicry": it is assumed that pathogens which resemble auto-antigens when presented on HLA proteins have a selective advantage, as most T cells with compatible TCRs would be deleted during negative thymic selection.

- Cross-HLA cross-reactivity: it has been shown that there exist TCRs which can bind to EBV peptides presented on a HLA-DRB5 chain and to MBP presented on a HLA-DRB1 chain [Lang et al., 2002]. Further complicating that picture, there is also evidence for epistatic interactions between HLA-DRB1 and HLA-DRB5 proteins in the context of MBP-driven immune responses [Gregersen et al., 2006].

- Interactions: recent studies have found evidence for a risk-modifying interaction effect between *HLA-B*/*HLA-C* and *ERAP1* alleles in ankylosing spondylitis and psoriasis, respectively [Evans et al., 2011; Strange et al., 2010]. One of the functions of *ERAP1* is to "trim" peptides before they are presented on HLA class I proteins. These studies therefore suggest that the risk effect mediated by certain HLA alleles also depends on peptide pre-processing.

In summary, the pathogenesis of many autoimmune diseases is far from understood. Elucidation of the HLA-based effects is certainly not trivial, partly due to the complexity (e.g. interactions) of the underlying biological processes. Further functional and association studies will be required, and these studies will benefit from reliable and cost-effective ascertainment of HLA types. As outlined in the previous section, classical HLA typing methods, based on sequencing, are still the gold standard – but prohibitively expensive for many large-scale studies. This thesis describes methods for the statistical imputation of HLA types, based on SNP genotypes, which are expected to be of particular utility in studies of autoimmune diseases.

To complete this section on HLA-related diseases, another important type of condition has to be mentioned: transplant rejection after transplantation, in particular graft-versus-host disease (GVHD) after bone marrow or stem cell transplantation [Szabolcs et al., 2010]. Careful HLA type matching of donor and recipient is an essential step in preventing GVHD. However, the methods presented have probably not achieved the accuracy required to be useful in donor and recipient HLA type matching. Therefore, all issues of transplantation are not covered here.

2.1.6 Population genetics of the human MHC

At least two questions follow immediately from the MHC's haplotype structure, combining high levels of polymorphism combined with long-range LD:

- Can selection account for the MHC's structure?
- From a purely empirical point of view, are the patterns of LD and diversity globally uniform, or are there population-specific differences? Any population-specific differences could influence disease susceptibility on a population-level and be informative of ancestry.

Selection

Without doubt, selection has strongly influenced the large- and fine-scale structure of the HLA. Firstly, a basic understanding of the functional role of the HLA molecules within the immune system intuitively justifies to assume strong selection. How could, for example, trying to escape immune detection by adapting to the HLA molecules' binding affinities not be an element of particular viral strategies? Secondly, the existence of HLA-associated diseases (with early onset, thus influencing reproductive success – e.g. HIV, see Leslie et al. [2004] and Section 2.1.5) is direct evidence that the intuitive argument just presented has some truth to it. In fact, viral strategies against HLA-based detection include *suppression of HLA expression* [Alcami and Koszinowski, 2000]. Thirdly, in a series of studies, links between survival probability, disease progression or pathogen load and particular HLA types and diversity were clearly established (Lohm et al. [2002], Schad et al. [2005], Froeschke and Sommer [2005], Prugnolle et al. [2005], Hill [2001]). Finally, from the standpoint of population genetics, the patterns observed in the MHC make a strong influence of selection seem likely [Hughes, 2002]:

- in terms of allele frequencies, there are strong departures from the expectations under neutrality [Hedrick and Thomson, 1983] – which could well be explained by selection. Bergstrom et al. [1998] computed phylogenetic trees for *HLA-DRB1* alleles and found out that the majority of alleles is comparably young, whereas a minority of alleles has been maintained for a long time, predating at least the separation of *Homo* and *Pan* (*trans-species polymorphisms*). This pattern suggests that some ancient alleles were beneficial enough over time not to be eliminated by random genetic drift [Hughes, 2002].

- a majority of sequence substitutions in the coding exons of the HLA alleles is non-synonymous [Hughes and Nei, 1988, 1989].

- there are fewer polymorphisms in the intron sequences of the HLA alleles than synonymous substitutions in the exon sequences, suggesting that balancing selection on the exons, genetic drift and recombination act towards a homogenization of the

introns [Cereb et al., 1997; Hughes, 2002].

More recent studies, focused on whole-genome detection of signals of positive selection, have also found strong evidence for selection in the HLA region [Albrechtsen et al., 2010; Andres et al., 2009; Pickrell et al., 2009].

Hedrick [2002] briefly reviews several mathematical models of selection that were invoked to explain the patterns seen at the HLA, including heterozygote advantage, frequency-dependent selection and variable selection in time/space, pointing out that the effects captured by these models may actually combine.

In terms of the region's long-range LD structure, the observed LD between alleles at different classical loci would be consistent with epistatic selection on combinations of alleles; both gene conversion and recombination could play a role in reducing LD in the regions in between the genes.

Accepting selection as a major force shaping the HLA helps to explain the region's fine-scale structure, including the extensive polymorphism. It remains to be seen how selection can explain the larger structures of vertebrate MHCs, in particular the remarkable fact that the MHC has universally stayed intact as a cluster in nearly all jawed vertebrates [Litman et al., 2010].

Global HLA diversity

Probably because of the HLA's relevance in clinical settings, in particular in transplantation medicine, many studies have examined HLA diversity patterns – allele frequencies and haplotypes – in worldwide populations [Cao et al., 2001, 2004; Piancatelli et al., 2004; Prugnolle et al., 2005; Qutob et al., 2011; Torimiro et al., 2006; Trachtenberg et al., 2007]. Most studies have focused on HLA class I.

One of the most recent and comprehensive studies, carried out by Qutob et al. [2011], finds that substantial proportions of global HLA type diversity can be accounted for by distance from Africa, pathogen load, and, interestingly, statistical interactions with frequencies of particular KIR alleles. The component which relates to distance from Africa is consistent

with general models of human demography, which assume an Out-of-Africa bottleneck event. This of course implies that the greatest genetic diversity is to be expected in sub-Saharan Africa, which is certainly true for HLA alleles [Cao et al., 2001]. The finding that the second component, pathogen load, significantly contributes to explaining variance in HLA diversity is consistent with models which assume that MHC diversity is driven by an evolutionary arms race with pathogens. Finally, the fact that HLA and KIR frequencies are related is an interesting result, and the study tries to relate certain KIR subgroups to the impact of bacterial and viral load.

Another useful resource in comparing global haplotype diversity, including class II loci, has been produced by the National Marrow Donor Program (NMDP) of the United States: http://www.haplostats.org provides an online interface to an algorithm (based on Expectation Maximization) which estimates global haplotype frequencies, using worldwide transplant donor databases. The results can be graphically displayed on world maps, with the color of a country indicating a specified haplotype's frequency (see Figure 2.3).

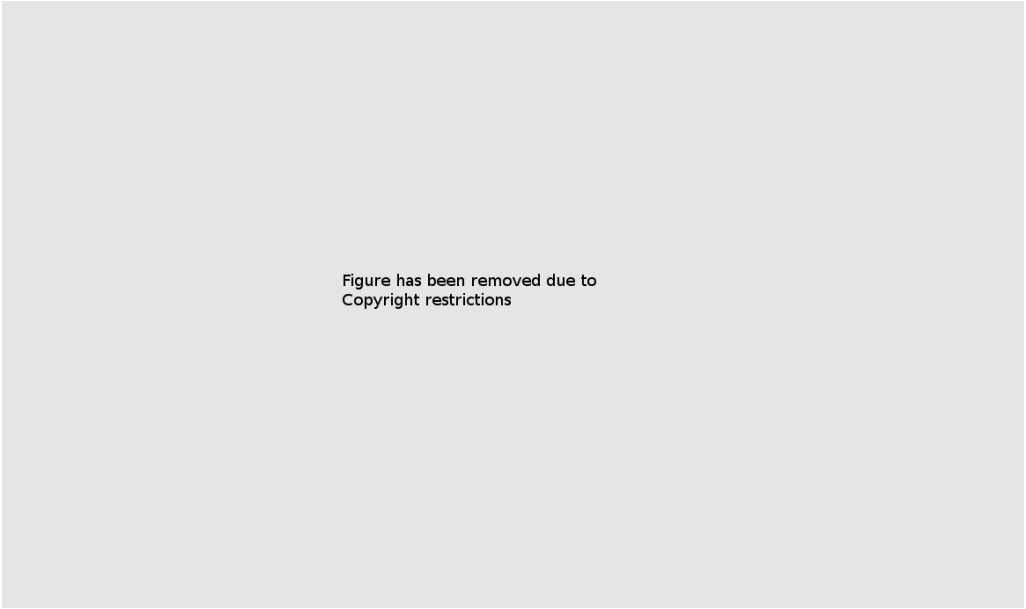

Figure 2.3: An example map from the NMDP's HaploStats (http://www.haplostats.org/home.do) application, providing estimates for global HLA type frequencies. Figure reproduced from the HaploStats manual.

Many of the studies cited above provide very valuable data to guide the search for potential transplant donors and contribute to the global picture of HLA diversity, but as this thesis

is not concerned with particular alleles or haplotypes, an in-depth review of the findings is not presented here for all but one study.

Cao et al. [2004] have surveyed HLA class I variation in five sub-Saharan African populations (Kenyan Nando, Kenyan Luo, Malians, Ugandans, Zambians), genotyping between 45 and 265 individuals from each population (852 in total). As HLA type variation is greatest in Africa, their results reflect an upper limit on expected HLA type variation in other populations, and give some insights into the pre-bottleneck pattern of human HLA type variation. Many alleles observed in the study are Africa-specific and do not appear in other human populations. The five groups vary in the observed patterns of diversity, globally and on a per-locus basis. 20 *HLA-A* alleles are exclusive to one or two groups, and 34 *HLA-B* alleles. For *HLA-B*, no single allele is found in a "large" fraction of the total population (*predominant* in the terminology of that paper, but not clearly defined), apparently consistent with global *HLA-B* diversity patterns. The same is true of *HLA-A* in both Kenyan populations, whereas there are predominant *HLA-A* alleles in the other populations. The fraction of the population which carry one of the 15 most frequent $A/B/C$ haplotypes varies between 0.5 (Zambians) and 0.26 (Ugandans), and only 5 haplotypes are shared between two or more populations. LD between B and C is strongest, with most B alleles being nearly exclusively transmitted with a single C allele. According to the analysis of the authors, many alleles predate the split of the language groups, and overall LD is, despite the high number of haplotypes, comparable to other global populations.

In general, it is worth noting that genetic variation in the HLA is far from completely characterized. In 2010, the NMDP alone has identified 72 new HLA alleles[Lazaro et al., 2011].

2.1.7 Evolutionary history of the MHC and the development of adaptive immunity

Human population genetics provide a framework to think about the more recent evolutionary history of the MHC, but many of its features are only understood in the more general context of the evolution of vertebrate adaptive immunity.

Adaptive immunity arose during the early stages of vertebrate evolution, most probably around 500 million years ago in a species of jawed fish [Flajnik and Kasahara, 2001; Litman et al., 2010], and is a universal feature of all extant jawed vertebrates. Many authors argue that acquisition of the recombination-activating genes 1 and 2 (*RAG1* and *RAG2*) was a necessary condition for developing a system of adaptive immunity, as *RAG1* and *RAG2* universally mediate the V(D)J process which generates TCR and BCR variation. Interestingly, some jawless fish have undergone convergent evolution and developed a different system of adaptive immunity, independent of RAG1/2, but able to generate diversity through somatic rearrangement [Flajnik and Kasahara, 2010; Litman et al., 2010]. Some authors argue that the development of adaptive immunity (broadening the spectrum of possible responses against colonization by parasites) has conferred a selective advantage, but there is no obvious reason to believe that vertebrates would benefit any more from adaptive immunity than complex invertebrates and no systematic evidence that invertebrates succumb to infection more often than vertebrates [Hedrick, 2004].

There have been various attempts to characterize the differences between different species' adaptive immune systems and MHC regions (e.g. Litman et al. [2010], Trowsdale [1995] and Kelley et al. [2005]), resulting in the identification of a distinctive set of four components, present in all jawed vertebrates:

- MHC class I/II-like molecules (except for the Atlantic Cod, which lacks MHC class I despite being classified as jawed vertebrate [Star et al., 2011])
- T/B cell receptor (TCR/BCR) equivalents, modified by *RAG1*/*RAG2*
- genes coding for immunoglobulins
- two lymphoblastoid cell lines (B- and T-like cells)

Does this imply that the genes coding for these functions are orthologous? Apparently not [Nei and Rooney, 2005]. The sequences found in different MHCs show hardly any evidence of common ancestry – and, even more surprisingly, the genes fulfilling the essential immune functions are different – for example, the classical HLA proteins A, B, C are only found in hominid species, but not in New World monkeys. Nei and Rooney [2005] and Kumanovics

et al. [2003] interpret this as evidence for a "Birth & Death" model of evolution in the MHC region: a series of expansions and contractions of the MHC region, marked by duplication and deactivation events. The duplicates can mutate functionally and, for example, become coding for new antigen-presenting molecules, whereas the deactivated genes end up as pseudogenes. This model convincingly explains why there is a substantial amount of pseudogenes in the HLA region; it is also in accordance with the assumption of strong selection, because it allows the duplicated genes to rapidly adapt to new immunological challenges – as the original, duplicated genes are still in place. Finally, it accounts for the fact that the MHC regions are functionally homologous among all jawed vertebrates without being orthologous on the sequence-level. Although the Birth & Death model is very appealing on a large scale, it is not clear how much other effects – e.g. gene conversion – have contributed to the evolution of the MHC. Nei and Rooney [2005] argue in strong favour of a dominating influence of mutation and positive selection, whereas other authors (e.g. Andersson and Mikko [1995], Zangenberg et al. [1995] and Hogstrand and Bohme [1999]) emphasize the importance of gene conversion and recombination. Indeed, it is an interesting (and open) question why certain HLA alleles exhibit stretches of unusual similarity [Nei and Rooney, 2005].

Is the MHC cluster structure a random artifact, or does the clustering confine any kind of selective advantage? Remarkably, nearly all jawed vertebrates – with the exception of some fish species – share the chromosomal connection between the MHC's different components. What is more, according to Kumanovics et al. [2003], there seems to be an affinity of functionally equivalent genes to stick to their relative position – e.g. in the mouse – human comparison, where functional homologs, yet no orthologs, occupy the same positions between other shared MHC genes present in both species. There is evidence for the existence of a *proto-MHC* region, mainly stemming from the observation that there are at least three paralogous regions to the MHC in the human genome, on chromosomes 1, 9 and 19. This pattern is in agreement with the "2R" hypothesis, according to which there were 2 genome duplications early in the vertebrate lineage – in this case, involving the hypothetical proto-MHC. Alternatively, one may assume that the cluster structure of immunity-related genes improves fitness, e.g. by facilitating co-regulation. In general, it is

worth noting that the fact that functional homologs of the classical HLA genes are present in all jawed vertebrates' MHC regions justifies the notion that the classical HLA genes are the most important and indeed defining genes of the human MHC.

In conclusion, the "Birth & Death" model offers a good explanation for the large-scale structures of vertebrate MHCs. It is, in generating gene duplicates that are under relaxed functional constraints, in accordance with the general assumption that the MHC is coined by the need for rapid adaptation. Still, the plasticity of MHC complexes among vertebrates remains impressive.

2.2 Genotype imputation

Defined generally, genotype imputation uses information from a reference panel, typed at high marker density, to infer genotype probabilities for untyped markers in individuals from an imputation panel, typed at lower marker density.

Why is this a sensible idea, in particular for genome-wide association studies (GWAS)?

It is well-known that the power to detect an association (i.e. the probability to detect an association if there is one) in a GWAS depends on sample size as well as on the maximum correlation between any typed variant and the assumed causal variant. Under certain conditions (one marker, equal numbers of cases and controls, multiplicative risk model),

$$\mathrm{E}(\chi^2) \propto N \times \gamma^2 \times p(1-p) \times r^2 \,,$$

where χ^2 is the chi-squared test statistic of no association, N the number of cases and controls, γ the effect size, p the risk allele variant frequency and r the correlation between any typed variant and the causal variant [Chapman et al., 2003; Spencer et al., 2009]. Although this simple risk model is almost certainly an oversimplification for the majority of diseases, the formula provides useful intuition on the importance of the correlation between typed and examined markers in GWAS. In the context of genotype imputation in GWAS, the correlation between causal variant and imputed variant is therefore a crucial determinant of power (in Section 6.2, I will give a practical example of how improving the correlation between causal marker and imputed marker can lead to more accurate results in GWAS).

"Maximum correlation" is identical to the maximum linkage disequilibrium (LD) between any typed marker and the assumed causal variant (and local LD typically also determines the size of the region around the associated marker that may plausibly harbor the causal variant). The underlying principle of imputation is to leverage the effect of LD across multiple markers: if the genotype of an untyped marker can be statistically inferred ("imputed") from the genotypes of surrounding typed markers by taking into account their joint LD structure, and if this imputed marker correlates more strongly with the causal

variant than any typed marker, the power to detect the association increases. This process critically depends on an accurate model of regional LD structure, which in most current imputation algorithms takes the form of an explicit representation of underlying haplotypes. The information on the haplotypic relationship between typed and untyped markers is typically extracted from a separate reference panel, in which all markers are present. Often, experimentally determined or deterministically phased haplotypes are available for neither panel, so that determining the chromosomal phase of markers – *phasing* – becomes an essential part of imputation.

Imputation technology can, to give some specific examples, be used to impute the genotypes of recently discovered variants into older datasets, to optimize the ratio between typed variants and typed individuals to obtain optimal power [Spencer et al., 2009], or to determine the genotype of complex variants like HLA types based on simpler markers like SNPs.

A variety of different algorithms and models have been developed for genotype imputation. Useful high-level questions for categorizing them include

- What model is used to represent haplotype structure?
- How are the parameters or the structure of the model inferred from the data? In particular, is the model built for the reference panel and then applied to the imputation dataset, or is a combined model built for both datasets?
- How are missing genotypes in the imputation dataset inferred?
- How well does the model scale, and how computationally efficient is it?
- How accurate and complete are the model's imputations?
- Is the model optimized for imputing particular kinds of markers, for example SNPs or HLA alleles, or is it generic?

The first of these questions – how to model haplotype structure – is arguably the most fundamental one. As it turns out, the coalescent, a well-parameterized genealogical model

from population genetics, is too complex to be computationally tractable in haplotype inference and imputation scenarios. This inspired the development of the so-called Li & Stephens approximation (L&S), which uses a Hidden Markov Model to model identity by descent (IBD) relationships between haplotypes [Li and Stephens, 2003]. Then, inspired from computer linguistics, there is another class of models: haplotype graph models, which model haplotypes in a way which is independent of any biologically meaningful parameterizations. Both haplotype graph models and the L&S approximation provide a suitable basis for high-accuracy imputation algorithms [Marchini and Howie, 2010], and both will play an important role in this thesis.

The next sections are structured as follows: first, I will describe the class of leveled Hidden Markov Models (HMM), as they are important in both the L&S approximation and haplotype graph models. Then, I will describe the two main haplotype representation models used in imputation frameworks, i.e. the L&S approximation and haplotype graphs. Finally, I will discuss two popular implementations of these models (BEAGLE and IMPUTE) and how they combine information from imputation and reference panels.

2.2.1 Haplotype representation models

This section will discuss three important classes of haplotype structure models, one of which (the coalescent) will turn out to be computationally intractable for the questions we are interested in here. As an important preparation, I will describe the structure and properties of leveled Hidden Markov Models.

In the context of this thesis, haplotypes are represented as strings, i.e. ordered sequences of symbols. The symbols usually refer to the four nucleotides (A, C, G, T) or specific complex alleles, such as HLA alleles, and possibly another symbol which indicates missingness. Each position in a haplotype is associated with a genetic position, and when I refer to a sample of haplotypes, this denotes a set of same-length haplotype strings, with homology between all i-th positions.

Leveled Hidden Markov Models

Hidden Markov Models have successfully been employed in a variety of areas, for example natural language processing and physics [Rabiner, 1989]. They, and in particular the subclass of Leveled Hidden Markov Models (Browning and Browning [2009], LHMMs), have many important applications in genetics (they are also a subclass of inhomogeneous HMMs). As it will turn out later, both haplotype graphs and the L&S approximation belong to the class of LHMMs.

General Hidden Markov Models and useful algorithms To set up basic notation, I give a formal definition of Hidden Markov Models, following Rabiner [1989] – see Figure 2.4 for a graphical illustration of the properties of HMMs and the notation employed here. An HMM M is a system of N states, denoted S_1 to S_N, and the system is in any one of the N states at any time during the observation period. We denote the state of the system at time t as q_t, but q_t is not directly observable. The system evolves in regular units of discrete time from $t = 1$ to $t = T$: after one unit of time has elapsed, the system can change its state. For each state i and each state j, there is a probability distribution $A = a_{ij}$ which specifies the probability of changing state from i to j. Note that this implies that the HMM follows a Markov process: $\mathrm{P}(q_{t+1}|q_t, .., q_1) = \mathrm{P}(q_{t+1}|q_t)$. There is also a probability distribution π_i which defines the probability that i is the first state: $\mathrm{P}(q_1 = i) = \pi_i$. At time t, the HMM will emit an observable symbol O_t from a discrete set of allowed emission symbols, the *model alphabet*. O_t is selected according to a state-dependent probability distribution $b_j(k)$, i.e. $\mathrm{P}(O_t = k|q_t = j) = b_j(k)$.

It is easy to see that the following algorithm will generate strings (sequences of symbols) of length T from M:

1. Set $t := 1$.
2. Select q_1 according to π_i.
3. Select O_t according to b_{q_t}.
4. If $t \neq T$, select q_{t+1} according to a_q, set $t := t+1$, go to step 3; otherwise, terminate.

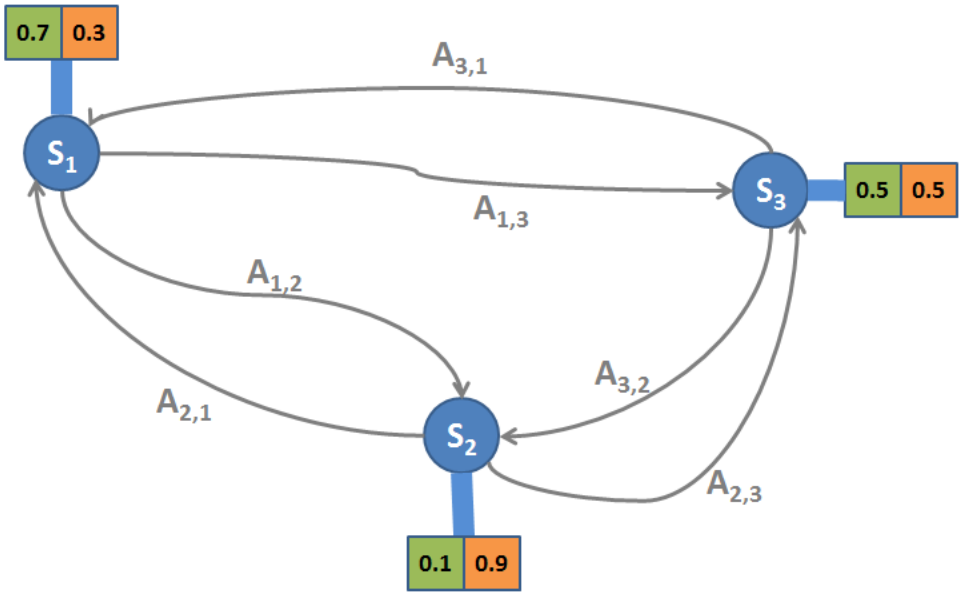

Figure 2.4: This figure illustrates a general HMM: each state (blue circles) has defined transition probabilities to all other states (specified by the matrix A, grey lines) and emission probabilities for all symbols of the model alphabet (here represented by two coloured squares and the emission probabilities printed therein).

Observing the sequence $O_1..O_T$ from M does not allow for direct inference of the states $q_1..q_T$ the model visited while generating the sequence (this is the property that the word "hidden" refers to). However, there are a couple of well-known and computationally efficient algorithms that are useful in drawing inferences from an observed output sequence [Rabiner, 1989].

The *forward algorithm* efficiently computes $\mathrm{P}(q_t = j|O_1..O_h, M)$, i.e. the probability of the model being in state j conditional on the first h observations. Define $\alpha_h(j) := \mathrm{P}(O_1..O_t, q_h = j|M)$, for $1 \leq j \leq T$. Initialize with $\alpha_1(j) := \pi_j \times b_j(O_1)$. For $t = 2..h$, continue with $\alpha_t(j) := \sum_{i=1}^{N}(\alpha_{t-1}(i) \times a_{ij}) \times b_j(O_t)$. It is clear that $\mathrm{P}(q_t = j|O_1..O_h, M) = \frac{\alpha_t(j)}{\sum_{i=1}^{N}\alpha_t(i)}$. The general forward algorithm as presented here is in the complexity class $O(T \times N^2)$.

The forward algorithm allows us to compute the likelihood of an observed sequence under M: $\mathrm{P}(O_1..O_T|M) = \sum_{i=1}^{N}\alpha_T(i)$.

The forward algorithm also allows us to sample paths from M, conditional on $O_1..O_T$. We use the word *path* to denote a possible sequence through the model's states, i.e.

an ordered sequence $p = (p_1, .., p_T) \in \{x \in \mathbb{N}^T | \forall i \in (1..T) : 1 \leq x_i \leq N\}$. Define $\mathrm{P}(p|O_1..O_T, M) := P(q_1 = p_1, .., q_T = p_T|O_1..O_T, M)$. Computing $\mathrm{P}(p|O_1..O_T, M)$ for a particular p is trivial. Sampling from that distribution, however, follows not immediately – the number of possible paths is usually too large to exhaustively calculate all probabilities. The forward algorithm provides a solution. Initialize with selecting p_t according to $\mathrm{P}(p_t) := P(p_T = p_t|O_1..O_T, M)/\sum_{i=1}^{N}(\mathrm{P}(q_T = i|O_1..O_T, M))$. Now, for $t = (T-1), .., 1$, draw p_t from $P(p_t) := \frac{\mathrm{P}(q_t=p_t|O_1..O_t,M)\times a_{p_t p_{t+1}}}{\sum_{i=1}^{N}\mathrm{P}(q_t=i|O_1..O_t,M)\times a_{i p_{t+1}}}$.

The *backward algorithm* approaches the same problem from behind. Define $\beta_h(j) := \mathrm{P}(O_{h+1}..O_T|q_h = j, M)$, for $1 \leq j \leq T$. Initialize with $\beta_T(j) := 1$. For $t = (T-1)..1$, continue with $\beta_t(j) := \sum_{i=1}^{N}(\beta_{t+1}(i) \times b_i(O_{t+1}) \times a_{ji})$. The backward algorithm is in the complexity class $O(T \times N^2)$.

The forward and backward algorithms enable us to compute the likelihood of an observed sequence in another way. For arbitrary $1 \leq k \leq T$, $\mathrm{P}(O_1..O_T|M) = \sum_{i=1}^{N}(\alpha_k(i) \times \beta_k(i))$. Also, we can now compute the probability of being in a particular state k at time t: $P(q_t = k|O_1..O_T, M) = \frac{\alpha_k(i)\times\beta_k(i)}{P(O_1..O_T|M)}$.

The *Viterbi algorithm* computes the Maximum Likelihood (ML) path estimate, conditional on having observed $O_1..O_T$. Its structure is very similar to the forward algorithm. Define $\gamma_h(j) := max_{(p_1,..,p_{h-1})}\mathrm{P}(q_1 = p_1, .., q_{h-1} = q_{h-1}, q_h = j, O_1..O_h|M)$, for $1 \leq j \leq T$. $\gamma_h(j)$ returns the the maximum likelihood of all paths of length h ending in state j. We also define δ to keep track of where the best path comes from. Initialize with $\gamma_1(j) := \pi_j \times b_j(O_1)$ and $\delta_1(j) := 0$. For $t = 2..h$, continue with $\gamma_t(j) := \max_{i=1..N}(\gamma_{t-1}(i)\times a_{ij})\times b_j(O_t)$ and $\delta_t(j) := \arg\max_{i=1..N}(\gamma_{t-1}(i) \times a_{ij}) \times b_j(O_t)$. Terminate with $\gamma\star := \max_{i=1..N}\gamma_T(i)$ and $\delta\star := \arg\max_{i=1..N}\gamma_T(i)$. $\gamma\star$ is the likelihood of the ML path. Backtracking from $\delta\star$ will reconstruct the states traversed by the ML path.

Missing data in the observed sequence Missing data in the observed sequence can be dealt with in a straightforward way: extend the model alphabet by a "missing data" character and assign equal missing data emission probabilities to all states in the model. We say that this way of dealing with missing data is *agnostic*, because all states are assigned equal probabilities to emit missing data. The Viterbi algorithm (or sampling

techniques) can now be used to estimate which states generated the missing data, and to infer a distribution over possible non-missing symbols at each missing position.

Leveled Hidden Markov Models and Graphs In LHMMs, states and transitions between states follow a linear ordering. An HMM M belongs to the class of LHMMs if the function $l(s)$, as defined below, is well-defined for all states s of M[3]. For all states $s \in 1..N$,

- if $\pi_s \neq 0$, define $l(s) := 1$ (all initial states have level 1).
- for all states $s' \in 1..N$, define $l(s) := l(s_2) + 1$ if $a_{s';s} \neq 0$ (if there is a transition between state s and state s', the level of s' must be equal to the level of s plus 1).
- if $a_{ss_2} = 0$ for all states $s_2 \in 1..N$, define $l(s) = T$ (all terminal states with no further transitions are at the same level T).

In LHMMs, no state is visited more than once, and each state can be reached. We refer to $l(s)$ as the *level* of state s. Importantly, the observed symbol O_i at the i-th position of the observed emission sequence is guaranteed to have been emitted by a state s of level $l(s) = i$.

Sometimes, it is convenient to recast notation for LHMMs in a way that makes explicit reference to their level structure. We can refer to the i-th state at the l-th level as $S_{l,i}$, and we denote the number of states at level l with N_l. Each $S_{l,i}$ has an associated probability distribution over the model alphabet: $b_{l,i}(k)$. We define π_i as the probability to select $S_{1,i}$ as a first state, and change the transition probability notation to $a_{l,i;l+1,j}$, denoting the probability to jump from state $S_{l,i}$ to state $S_{l+1,j}$.

Forward and backward algorithm can exploit the state structure of LHMMs. Each iteration of the algorithms is explicitly tied to a level l, and instead of summing over the full set of N model states in computing α and β, it is possible to constrain summation to the states

[3]The last condition implies that we relax the requirement that each model state has to have a proper probability distribution over possible transitions – in an LHMM, all states s with $l(s) = T$ have probability 0 for all possible transitions. The majority of HMMs currently in use in population genetics belong to the class of LHMMs, so that the reader may well already be familiar with their properties. I explicitly introduce them to stress the implications of their restricted transition matrices, and for notational convenience.

at level $l-1$ (forward) or $l+1$ (backward). The complexity of the two algorithms (for computing all α and β values) for LHMMs is in the class $O(T \times (max_{i \in 1..N} N_i)^2)$.

Sometimes, it is also convenient to define a *level-specific model alphabet*, containing all symbols that can be emitted by states of a particular level. Output emission probability calculations, in particular, can sometimes be simplified by considering the level-specific model alphabet.

See Figure 2.5 for an illustration of leveled HMMs and the related notation.

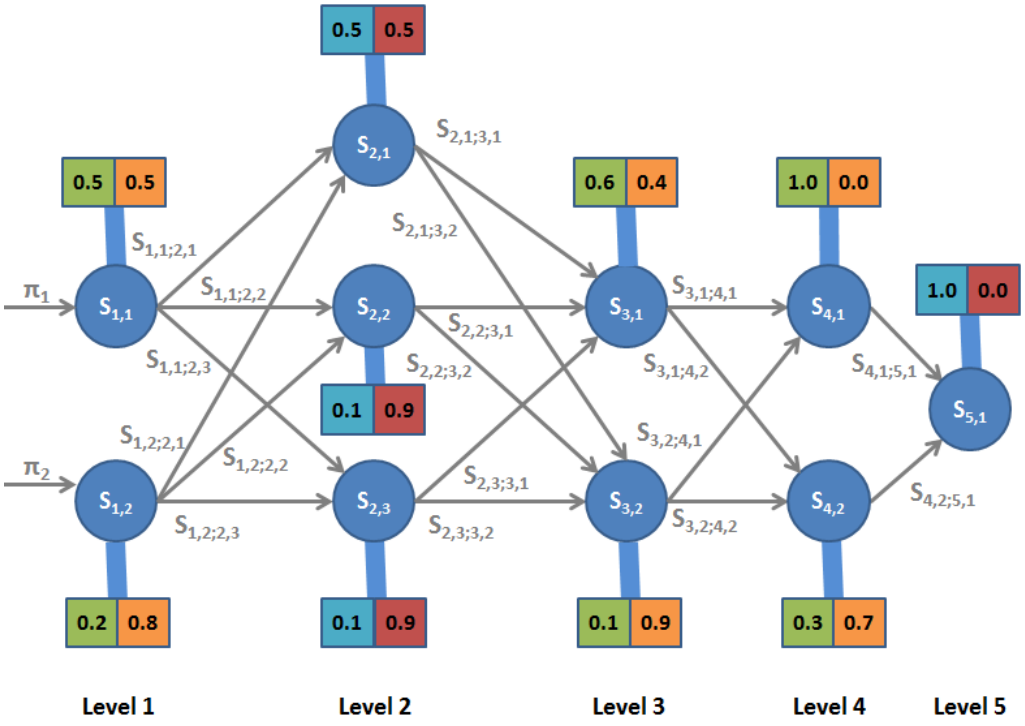

Figure 2.5: States (blue circles) in leveled HMMs follow a linear ordering: each state has an assigned level l (see bottom line), and there is always at least one state transition to a state of level $l+1$ (and no state transitions to other levels). State transition probabilities (grey lines) are specified by the matrix S, and initial probabilities by the vector π. Each state has assigned emission probabilities for all symbols of the level-specific model alphabet (here represented by two coloured squares and the emission probabilities printed therein). Different levels can have different model alphabets (hence different colours for emission symbols at states 2 and 5 in this example). Note that the number of specified state transitions grows much more slowly with the total number of states than in a general HMM.

For later purposes, I similarly constrain the class of directed graphs to the class of leveled directed graphs (LDGs). A directed graph G is the ordered tuple (V, E). V is the set of

the graph's vertices, and E is a set of ordered tuples of the form $(V \times V)$. If $(v_1, v_2) \in E$, we say that there is an edge *from* v_1 *to* v_2.

A directed graph G belongs to the class of leveled directed graphs if there exists a set $S \subseteq V$ and a $T \in \mathbb{N}^{>1}$ for which the function $l(v)$, as defined below, is well-defined for all vertices $v \in V$. For all vertices $v \in V$,

- if $v \in S$, define $l(v) := 0$.
- for all vertices $v_2 \in V$, define $l(v) := l(v_2) + 1$ if $(v_2, v) \in E$.
- if $(v, v_2) \notin E$ for all vertices $v_2 \in V$, define $l(v) = T$.

LDGs are acyclic and connected.

By applying the following algorithm, the topology of an LHMM with $T > 1$ can be mapped onto an LDG. Set $V := \{\}$ and $E := \{\}$. For each HMM state $s \in \{1, .., N\}$, create a vertex s. For all ordered tuples of states $(s_1, s_2) : s_1 \in \{1..N\} \wedge s_2 \in \{1..N\}$, create an edge (s_1, s_2) if $a_{s_1 s_2} \neq 0$. Finally, set $S := \{s \in 1..N | \pi_s \neq 0\}$.

Haplotype HMMs and a diploid generalization For the purpose of this thesis, it has already been defined that haplotypes are to be represented as strings. This immediately implies that HMMs can be used to model haplotypes, and we denote them as *(haploid) haplotype HMMs* in this case. Often, however, we want to make inference based on genotype data. We represent a genotype stretch G of length T as an ordered sequence of T unordered tuples, generated from two underlying haplotypes H_1 and H_2 of length T. The i-th genotype tuple is of the form (e_1, e_2), where e_1 is the the i-th symbol of haplotype H_1, $H_{1,i}$, and e_2 is the the i-th symbol of haplotype H_2, $H_{2,i}$ – or the other way around. Reconstructing H_1 and H_2 from G is known as the problem of *phasing*.

Generating genotype stretches from a haplotype HMM is trivial: generate two underlying haplotypes and generate the genotype stretch for each position from 1 to T according to the definition given above, randomly assigning the underlying haplotype for e_1 (e_2 then follows, of course).

The algorithms presented above to solve the problems of total likelihood, marginal state

probabilities and best path do not apply directly to genotype stretch data. However, by defining a new *diploid haplotype HMM*, which essentially consists of two connected haploid haplotype HMMs, we can apply the algorithms presented above to genotype stretch data.

Define a new HMM with N^2 states of the form $(S_i, S_j), i, j : i \in 1..N, j \in 1..N$. Informally, the state (S_i, S_j) refers to the first haploid haplotype HMM being in state S_i, and the second one being in state S_j. We define the state transition probability from (S_i, S_j) to (S_m, S_n) as $a_{im} \times a_{jn}$, and the probability of (S_i, S_j) being the initial state as $\pi_i \times \pi_j$. The diploid emission probability distribution for (S_i, S_j) over all unordered tuples (k_1, k_2) of the model alphabet is defined as $b_{(S_i,S_j)}(k_1, k_2) := b_i(k_1) \times b_j(k_2)$ if $k_1 = k_2$, and $b_{(S_i,S_j)}(k_1, k_2) := b_i(k_1) \times b_j(k_2) + b_i(k_2) \times b_j(k_1)$ otherwise.

It is easy to see that connected leveled haploid haplotype HMMs result in leveled diploid haplotype HMMs, and that the structure of leveled diploid haplotype HMMs is more tractable than that of general diploid haplotype HMMs. In either case, the forward, backward and Viterbi algorithms are applicable to the diploid haplotype HMM, with each state in the diploid HMM relating to two haploid states. By applying the Viterbi algorithm, it is, for example, possible to determine the most likely pair of paths through the two haploid models, conditional on some observed genotype stretch data.

The Coalescent

The coalescent (see, for example, Kingman [2000] for a historical sketch) is one of the most popular models in population genetics: it models the genealogical relationships between a sample of chromosomes belonging to a single species, backwards in time, until a common ancestor for all genetic material has been reached. More specifically, the coalescent models haplotype structure in an observed sample by modeling the identity-by-descent relationships between these haplotypes, explicitly accounting for the effects of *mutation* and *recombination* (as far as the coalescent with recombination is concerned).

This thesis offers a short treatment of the coalescent, because some of the models later presented can be thought of as informal approximations. Also, it is important to appreciate why one of the field's favourite models can in fact not be used for imputation purposes.

The coalescent emerges as an ancestral limit process from a range of population genetic models, but it is probably most easily understood in the context of a Wright-Fisher model (compare Wakeley [2009]). It makes the assumption of a constant-sized population of N individuals which is reproducing in a discrete-generation manner and which is not subject to selection. In each generation, every individual selects a predecessor from the previous generation. The probability that two specific members from the same generation have different ancestors in the r-th previous generation is therefore $(1 - \frac{1}{N})^r$. This implies that the effective number of ancestors is shrinking from generation to generation – until all ancestral lines have "coalesced" into a single individual, the most recent common ancestor (MRCA).

The coalescent is defined on a sample of n individuals from a population of size N, with $n \ll N$. n individuals imply $n - 1$ coalescent events (each coalescing into one of the remaining effective ancestral lines) until having reached the MRCA. Each event n_i has a coalescent time T_i associated with it. As N goes to infinity, the T_i are distributed independently and exponentially, with

$$f_{T_i} = \binom{i}{2} e^{-\binom{i}{2} T_i},$$

supposing that time is measured appropriately, e.g. in units of N generations for the Wright-Fisher model. $\mathrm{E}[T_i]$ is $\frac{2}{i(i-1)}$, so that the expected time between two coalescent events increases as the number of effective ancestral lines is shrinking.

The mathematical derivation of the coalescent makes the following assumptions:

- No selection and no subdivided population – i.e. exchangeability between the n individuals
- A constant-sized population (fixed N over time)

The order in which the lineages coalesce directly implies a bifurcating genealogical tree – actually, a sample of size n implies $\frac{n!(n-1)!}{2^{(n-1)}}$ possible tree branching structures, the branch lengths corresponding to the T_i random variables. As exchangeability was assumed, all branching structures are equally likely.

The model structure which has been specified so far explicitly accounts for time, but not for the probability of observing a mutation, which we know increases over time. Conditional on a particular tree t (which specifies via its branch lengths how long it takes for two samples to coalesce), we can superimpose a mutation process acting along the branches. For example, we can define the probability to observe x mutations along a branch to follow a Poisson process, with rate proportional to the length of the branch and a mutation rate parameter θ.

Having observed some data, i.e. a sample of n chromosomes genotyped at some positions, we can now identify each of the chromosomes with one sample in the coalescent process (i.e. attach the chromosome to the *leaves* of any possible underlying tree) and use likelihood techniques to perform inference on the parameters of the process. For example, the likelihood of a particular tree *topology* t, averaging over the space of possible branch lengths b_l, is

$$\mathrm{L}_\theta(t) = \mathrm{P}(data|t;\theta) = \int \mathrm{P}(data|t, b_l;\theta)\,\mathrm{dP}(b_l)\,.$$

We could also use likelihood techniques to impute untyped markers: simply treat missing genotypes as parameters to the likelihood and use standard ML techniques to find values. However, as should be clear by now, any full likelihood approach has to take into account that each tree is a priori equally likely, and that the mutation process is defined conditional on a particular underlying tree.

Unfortunately, the enormous size of the space of possible tree topologies renders this approach computationally unfeasible. In fact, the issue is further complicated if a more realistic extension to the original coalescent, the coalescent with recombination, is used. This extension is based on the so-called *Ancestral Recombination Graph* (ARG, see Griffiths and Marjoram [1996]) and is more intricate than the conventional genealogical tree, marginally generating one tree for each locus (e.g., each base in a sequence of DNA) under consideration (see McVean and Cardin [2005]). The parameter vector specifying recombination probabilities between loci is usually denominated ρ.

Note that it is not clear how well the standard coalescent with recombination would perform in the HLA region, even if it was computationally tractable. There is strong evidence that the haplotype structure of the HLA has been shaped by selection and bottleneck effects (see Section 2.1.6), violating the coalescent's basic model assumptions.

The Li & Stephens Approximation

The Li & Stephens approximation [Li and Stephens, 2003] was initially developed to model and estimate recombination. However, it is now clear that it can be applied in a variety of other scenarios, including phasing [Stephens and Scheet, 2005] and imputation [Leslie et al., 2008; Marchini et al., 2007].

Fundamentally, the L&S approximation is a leveled HMM which generates new haplotypes by combining chunks from a set of observed haplotypes; often, it is said that the new haplotypes resemble an "imperfect mosaic" of the observed haplotypes [Leslie et al., 2008].

Li & Stephens were interested in finding an approximation to the expression

$$\mathrm{P}_{\text{coalescent}}(c_1, c_2, ..., c_n; \theta, \rho)\,,$$

where c_1, c_2, ..., c_n are haplotypes sampled from a population, all typed at the same T loci. θ is a global parameter that relates to the probability of mutations and ρ is a $(T-1)$-dimensional genetic map, relating to the probability of a recombination event occurring between each of the k loci (for example from Myers et al. [2005]).

They noted that $\mathrm{P}_{\text{coalescent}}(c_1, c_2, ..., c_n; \theta, \rho)$ is equal to

$$\mathrm{P}_{\text{coalescent}}(c_1; \theta, \rho) \times \mathrm{P}_{\text{coalescent}}(c_2|c_1; \theta, \rho) \times ... \times \mathrm{P}_{\text{coalescent}}(c_n|c_1, c_2, ..., c_{n-1}; \theta, \rho)\,.$$

In the Li & Stephens approximation, these conditional probabilities are expressed as emission probabilities of a leveled Hidden Markov Model; as this does introduce a dependency on the order of the chromosomes, it is recommended to repeat the calculations for different permutations and average over the results.

In the following, let C_N denote a set of N already sampled haplotypes, and c another haplotype, typed at the same T loci. Let $H_{a,b}$ denote the genotype of position b in haplotype $a \in C_N$ and c_b the value of the b-th position of c. In this notation, $\mathrm{P}_{\text{coalescent}}(c|C_N, \theta, \rho)$ is the quantity that shall be approximated. Li & Stephens use C_N, θ and ρ to construct an LHMM that can emit chromosomal haplotypes. The emission probability of that HMM, $P_{\text{L\&S}}(c|C_N, \theta, \rho)$ is then used as an approximation for $\mathrm{P}_{\text{coalescent}}(c|C_N, \theta, \rho)$.

The Li & Stephens LHMM (see Figure 2.6 for a graphical illustration) has a 2-dimensional state space: T levels with N states each. The initial probabilities at $l = 1$ are uniform, i.e. $\pi_i := \frac{1}{N}$. The state transition matrix is such that, being in state (l, i), only jumps to $l + 1 \times \{1..N\}$ are allowed. It is easy to see that this can be interpreted as the HMM "moving" along the markers in chromosomal order; and at each particular locus, the second state dimension refers to any of the N given haplotypes. The transition probabilities are calculated according to the vector of recombination probabilities, ρ.

In the notation of the introductory section on leveled HMMs (see Section 2.2.1), we define the following transition probabilities.

For $l \in 1..(T-1)$:

$$a_{l,i;l+1,j} = \begin{cases} \exp(-4N_e\rho_l/N) + [1 - \exp(-4N_e\rho_l/N)](\frac{1}{N}) & \text{i = j} \\ [1 - \exp(-4N_e\rho_l/N)](\frac{1}{N}) & \text{i} \neq \text{j} \end{cases}$$

where N_e is the *effective population size*. All undefined state transition probabilities are set to 0.

We restrict the model alphabet here to two symbols – major and minor allele, 0 and 1, respectively. This development is, in terms of the models presented in this thesis, sufficient. An extension to more alleles per locus is straightforward.

At state (l, i), the symbol e is emitted with probability

$$b_{l,i}(e) := \begin{cases} k/(k+\theta) + (1/2) \times \theta/(k+\theta) & e = H_{i,l} \\ (1/2) \times \theta/(k+\theta) & e \neq H_{i,l} \end{cases}$$

The mutation parameter, θ, is defined as follows:

$$\theta := \left(\sum_{m=1}^{N-1} \frac{1}{m} \right)^{-1}.$$

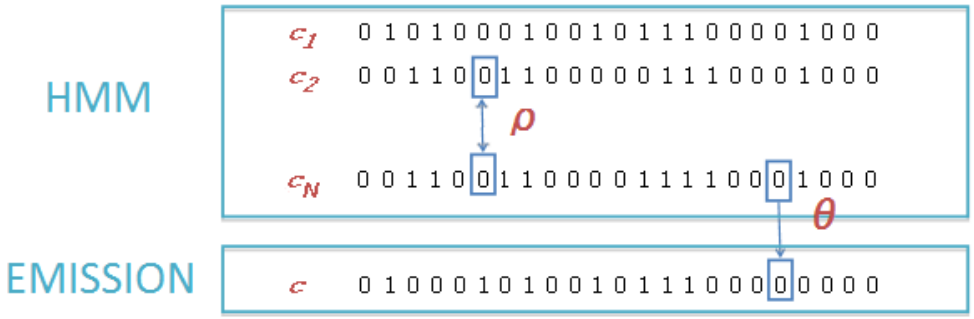

Figure 2.6: The Li & Stephens LHMM. Recombination, i.e. jumps between the chromosomal states, occurs according to the genetic map ρ. Emissions allow for mutations, controlled by θ.

These choices deserve a bit of interpretation. ρ is a genetic map, and ρ_l specifies the expected number of recombination events in a single meiosis between the two adjacent loci l and $l+1$. Multiplying that number by the effective population size and dividing by N, the number of sampled chromosomes, generates a rate for seeing recombination events – and taking the exponential consequently gives us the probability not so see any recombination event, if the number of these events is following a Poisson distribution. Regarding θ, $\left(\sum_{m=1}^{N-1} \frac{1}{m}\right)$ times the underlying mutation rate is the expected number of mutation events at a single site in a coalescent tree relating N samples, so that the chosen form of θ relates to an a priori expected number of 1 mutation event per site. However, these are not formal arguments, and Li and Stephens themselves stress that the relation of these choices to the ARG is unclear.

The L&S approximation has the advantage of being computationally quite tractable in the haploid case: this is partly due to the general decision to use an LHMM framework, but also

related to the fact that the state transition probabilities are equal for all states except one (the "current chromosome"), simplifying the forward and backward algorithms. In terms of algorithmic complexity, the optimized approximation is within the $O((N+N)*T)$ class for forward and backward algorithms, so that it is possible to handle substantial amounts of data. Calculating the likelihood of c and of paths through the model conditional on a set of given haplotypes and recombination and mutation parameters follows from the algorithm described above (Section 2.2.1).

Every approximation raises the question how good it is. Stephens and Scheet [2005] present some evidence that it captures important features of haplotype structure. What are, beyond empirical evidence, the principal reasons for believing that the L&S approximation works? Or, in other words, how can we qualitatively describe its behaviour?

Clearly, a chromosome generated by the L&S HMM resembles the chromosomes from C_N to the extent that it looks like a mosaic; haplotype breaks are related to known population-level recombination probabilities. Conversely, a chromosome that looks like a mosaic of C_N will be attributed a high likelihood. Why does this behaviour make sense? Li and Stephens list five criteria that should by fulfilled by coalescent approximations:

- Newly sampled chromosomes should have a higher probability of resembling haplotypes that are already frequent in the sample than resembling rare haplotypes
- The bigger the present sample, the smaller the probability to observe new haplotypes
- The chance of observing new haplotypes increases with θ
- Smaller differences to existing haplotypes are more probable than large differences
- Newly sampled chromosomes can be put together from existing haplotypes. This should not count as a large difference. Depending on the distribution of recombination probabilities, block-like structures of non-broken haplotype segments are possible or likely. This feature reflects the fact that identity-by-descent relationships between chromosomes under the coalescent change with recombination events.

Indeed, the model proposed by Li and Stephens seems to be in good accordance with these

criteria. Essentially, they reflect the notion that, as more chromosomes are sampled, our knowledge about haplotypes that are present in the population becomes more representative, and that we can make qualitative statements about the influences of recombination and mutation. It is possible to extend the Li & Stephens model by accommodating for other known sources of a population's haplotypic composition, e.g. gene conversion [Gay et al., 2007] or potentially even selection. Unfortunately, every such extension nearly inevitably increases the algorithmic complexity. What is more, the L&S approximation in the presented form has shown to be highly valuable in applications, be it accurate haplotype phasing [Marchini et al., 2006] or imputation [Leslie et al., 2008].

Haplotype graph models

Consider a leveled directed graph (see Section 2.2.1) $G = (V, E)$ with $S = v_0, v_0 \in V$ and a given $T > 1$.

We informally recapitulate what this means: G has one starting vertex v_0 with level $l(v_0) = 0$, and $l(v)$ is well-defined for every other vertex v. The graph is connected, i.e. for every vertex v in V, there is a path from v_0 to v. Whenever a directed edge connects v_i and v_j (i.e. going from v_i to v_j), we have $l(v_j) = l(v_i) + 1$. All vertices with no outgoing edges are at level T.

We now describe the extension of LDGs to *haplotype graph models* (HGMs). Each edge $e \in E$ carries an *attached symbol* or *emission symbol*. We use the notation $b(e)$ to denote this symbol for each edge. There are no two edges with the same emission symbol originating from the same vertex. We say that there is a *level-specific model alphabet*, a set A_l which consists of all symbols emitted by all edges emanating from vertices of level l and a *model alphabet*, which is the union of all level-specific model alphabets. Each vertex v at level $l(v) \neq T$ has a specified number of edges emanating from this vertex. Now, for each v with $l(v) \neq T$, and for the set of edges $E_v := \{(v_i, v_j) \in E \,|\, v_i = v, v_j \in V\}$ emanating from v, there is a probability distribution $\mathrm{P}(e|v), e \in E_v$, specifying the probability of *following* e, conditional on being at v (also denoted as the *edge probability distribution*). Picture 2.7 illustrates the properties of haplotype graph models.

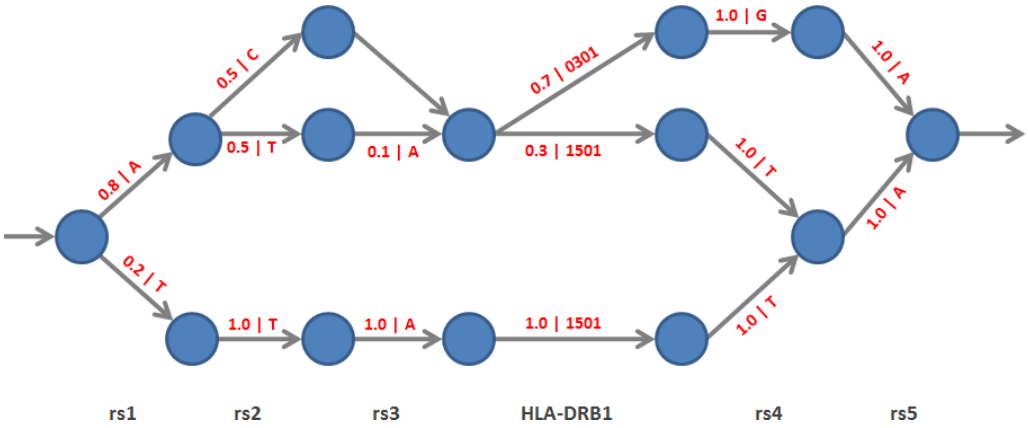

Figure 2.7: Illustration of the features of haplotype graph models. Haplotype graphs are a subclass of connected directed graphs. Their most important properties are illustrated here: 1) They are leveled, i.e. each vertex v has an associated positive number l, and all edges emanating from v at level l lead to a vertex at level $l+1$. A couple of vertices at level T are final vertices with no outgoing edges, and there is a path from every vertex in the graph to one of the final vertices. 2) Edges carry "emission symbols" which are emitted when an edge is traversed (in the figure: the symbols after the "|" character adjacent to the edges), and there are no two edges emanating from the same vertex which carry the same symbol. 3) Each vertex has an edge probability distribution over its attached edges (in the figure: the numbers in front of the "|" character adjacent to the edges), according to which an edge is selected conditional in being at that vertex.

By applying the following algorithm, haplotype graphs probabilistically generate strings of the same lengths T (haplotypes).

1. Set $v := v_0$.
2. If $l(v) = T$, terminate.
3. Otherwise, select an edge $e = (v_i, v_j)$ according to $\mathrm{P}(e|v)$.
4. Emit the symbol associated with e.
5. Set $v := v_j$, go to step 2.

As it turns out, haplotype graph models can be represented as LHMMs – and in analogy to HMM terminology, we speak of a *path* through the graph, which is defined by the edges and vertices traversed by the model while generating a haplotype. We now give a formal algorithmic definition of how to construct an LHMM H from a haplotype graph model G. We refer to H as the HMM *induced* by G. We use H and G indices in cases of ambiguous

notation. Also note that the level indices for haplotype graphs start at 0 and the level indices for HMMs at 1.

1. Generate an arbitrary ordering over the edges in G, by assigning a unique positive number to each $e \in E$.

2. Iterate from $l = 0$ to $l = (T - 1)$:

   (a) Define $V_l := \{v \in V \mid l(v) = l\}$.

   (b) Define the ordered set $E_l := \{(v_i, v_j) \in E \mid v_i \in V_l\}$ (E_l is the set of all edges emanating from nodes at level l), and order E_l according to the arbitrary ordering we have defined. $E_{l,i}$ denotes the i-th element of E_l.

   (c) We define H to have $N_{(l+1)} := |E_l|$ states at level $l+1$ (i.e. every edge becomes a state), and $S_{l,i}$ shall refer to the state in H which relates to the i-th element of E_l.

   (d) Emission probabilities: $\forall i \in \{1..N_l\} : b_{H;l+1,i}(k) := 1$ if $b_G(E_{l,i}) = k$, and 0 otherwise for all members of the model alphabet.

3. Initial probabilities: $\forall i \in \{1..N_1\} : \pi_i := \mathrm{P}(E_{0,i}|v_0)$.

4. Again, iterate from $l = 0$ to $l = (T - 2)$ – we now assign transition probabilities:

   (a) Define $P := \{(i, j) : i \in \{1..N_{l+1}\},\ j \in \{1..N_{l+2}\}\}$.

   (b) For each $p = (i, j) \in P$, we note that by definition $p = (i, j)$ can be mapped to two edges $e_x = E_{l,i} = (v_k, v_l)$ and $e_y = E_{l+1,j} = (v_m, v_n)$, respectively.

   (c) In this notation, if $v_l \neq v_m$, define $a_{l,i;l+1,j} := 0$.

   (d) Otherwise, $a_{l,i;l+1,j} := \mathrm{P}(e_y|v_m)$.

Inspired by computational linguistics [Ron et al., 1998], haplotype graph models were introduced to the field of genetics by Browning [2006], and to the problem of haplotype inference by Browning and Browning [2007]. The connection between haplotype graph models and formal population genetics is not well explored. The model, as specified here, does not directly account for recombination or mutation. A mutation-like effect can be

introduced to the induced HMM by spreading some probability mass in $b_{H;l,i}(k)$ over symbols other than the one carried by the underlying edge. Recombination-like effects could be modeled by modifying the state transition probabilities in the induced HMM, but it is not clear which effects (besides complicating the HMM's structure) this would have. In the present form of the model, LD is reflected in the branching patterns of paths through the models: once two different paths merge, i.e. traverse the same node at level l, the information from the left-hand side of l becomes irrelevant to predicting the remaining parts of the haplotype string, to the right-hand side of l. Conditional on having reached the same node at level l, the probability distributions over the remaining *suffix* parts of the haplotypes are identical. Parts of a haplotype graph in which many edges end up in the same nodes are therefore probably areas of high historical recombination rates.

Conveniently, all standard algorithms described above to deal with the problems of total likelihood, path likelihood, missing data and best path can be directly applied to the LHMMs induced by haplotype graph models.

Building haplotype graphs

I have described how to generate haplotypes from haplotype graph models, and how haplotype graphs relate to HMMs. Clearly, what is missing is a description of how to *construct* a haplotype graph model, based on some observed haplotypes. Browning [2006] has presented an algorithm which serves that purpose. I will describe and discuss this algorithm in the following. Note that the algorithm can be understood both in terms of inference (i.e. we assume that a haplotype graph model has generated the haplotype sample, and we're now trying to make inference on the parameters that led to the generation of the sample) or learning (i.e. we want to construct a haplotype graph model which captures the structure of some observed data).

Suppose that we have a set H of haplotypes (i.e. strings) of length T. The parameters that we are interested in determine the topology of the haplotype graph and the edge transition probabilities.

H induces a level-specific model alphabet A, which can be extended if necessary (note that

A_0 consists of all symbols present at the first position of the haplotypes in H). Consider, for a given number of levels T (corresponding to the lengths of the haplotypes in the set H), the most general possible haplotype graph topology (V, E), i.e. every vertex at each level has edges for all symbols of the level-specific model alphabet. Note that there are no edge probabilities attached yet. Each $h \in H$ corresponds to a unique path through the graph topology, and we say that h is *attached* to all vertices that the path passes through. We can reduce the general topology to a new topology, the *reduced sample topology for* H, by removing vertices that do not have any $h \in H$ attached to them, along with their associated edges.

For each vertex $v \in V$ we define a function count(v, x). If x is the empty string $''$, count(v, x) returns the total number of haplotypes in H attached to v. If x is a string of length ≥ 1, count(v, x) returns the number of haplotypes attached to v that continue with the string x (from position $l(v)$ onwards) – x can be a partial or complete suffix, i.e. of length $1..[T - l(v)]$. Conditional on being at vertex v, the probability $\mathrm{P}(x|v)$ of continuing with suffix x is defined as

$$\mathrm{P}(x|v) := \mathrm{count}(v, x)/\mathrm{count}(v, '') \,.$$

We use this definition to define the edge probability distribution for all edges emanating from a given node. Define $\mathrm{P}(e|v)$ as the probability to follow edge e conditional on being at vertex v, and let s denote the symbol that is attached to e. Then we set

$$\mathrm{P}(e|v) := \mathrm{count}(v, s)/\mathrm{count}(v, '') \,,$$

which results in a fully specified haplotype graph. For reasons to be discussed in the following section, we refer to this object as *proto-graph*.

We observe i) that the proto-graph exhibits a considerable topological complexity, if built from a reasonably sized set H, possibly leading to computational difficulties in later stages and ii) that the topology of the proto-graph is the most general one which is consistent with the observed data. If we assume that H was actually sampled from a haplotype graph, and

if we want to recover the original graph's underlying structure, we have to take into account the possibility that the original graph's structure may have been simpler, i.e. that one node in the original graph corresponds to more than one node at the same level in the protograph. Introducing a criterion of similarity that is based on comparing nodes' conditional suffix distributions addresses both points. Informally speaking, nodes with very similar suffix output distributions can be merged into one node to reduce computational demands without substantially changing the model's haplotype frequencies. Also, if two nodes were actually identical or not distinguishable in the original haplotype graph model, we would expect their suffix output distributions to be similar (see below, and Ron et al. [1998] for a formal treatment).

We formalize the notion of similar suffix distributions following Ron et al. [1998] and Browning and Browning [2007] by defining the function similar(v_1, v_2) as the maximum difference between the two conditional suffix probability distributions of v_1 and v_2:

$$\text{similar}(v_1, v_2) := \max_{x \in S_{v_1,v_2}} \left| \mathrm{P}(v_1, x) - \mathrm{P}(v_2, x) \right|,$$

where S_{v_1,v_2} is the set of possible suffixes originating from v_1 or v_2 (partial or complete, determined by the haplotypes attached to the two nodes). We apply similar to all pairs of nodes at all levels to identify pairs of nodes that can be merged, following the following algorithm. For all levels $l = 0...[T-1]$, compute the similarity measure similar(v_j, v_k) for all pairs of nodes (v_j, v_k) at level l. If similar(v_j, v_k) is below a certain threshold ϵ, merge v_j and v_k.

To merge v_j and v_k,

1. redirect all incoming edges of v_k to v_j, and attach all haplotypes that were attached to v_k to v_j.

2. attach all outgoing edges of v_k to v_j, and delete v_k.

3. note that step 2 may result in a structure violating the haplotype graph assumptions, as it may lead two edges (v_j, v_n) , (v_j, v_m) with the same attached symbol. In this case, merge v_n and v_m as described (i.e. recursively from step 1, if necessary), and

delete one of the two resulting (v_j, v_n) edges.

4. finally, update $P(e|v)$ for all modified nodes. [Ron et al., 1998] define an efficient algorithm for collapsing nodes in the reduced sample topology (which we do not reproduce here).

ϵ is defined dependent on how H is populated – see below.

Intuitively, it seems clear that the algorithm presented here captures some structural features of the data. After all, only nodes with sufficiently similar suffix distributions are merged – this captures the idea of a "breakdown of LD" for certain paths through the graph, and will make sure that the haplotype emission probability distribution of the constructed model is similar to the observed data. However, these points are rather *ad hoc*. Reassuringly, there are some more formally motivated arguments to believe that the presented algorithm will yield good results. Ron et al. [1998] have examined the learning behaviour of a very closely related algorithm. The main differences between the algorithm considered in Ron et al. [1998] and the one presented here are

- Ron et al. [1998] assume that the observed data has actually been generated by a haplotype graph (acyclic probabilistic finite automaton, APFA, in their notation – a class of string-generating automata well-known from computer science), and their research question is to what extent the parameterization of the generating APFA can be re-constructed from the data. That is, Ron et al. [1998] deal with a well-defined inference problem.

- the Ron et al. [1998] algorithm uses the same *similar* function, but employs a particular way of determining the threshold for merging.

- in Ron et al. [1998], nodes with small numbers of associated haplotypes are merged into one node at each level.

Conditional on some assumptions, most importantly sample size and sufficient "distinguishability" between the automaton's states, Ron et al. [1998] then prove that the Kullback-Liebler (KL) divergence between the generating and the inferred automaton is

small with high probability. “Distinguishability” refers to sufficiently different conditional suffix output distributions between nodes. Put in the context of genetics, this can be interpreted as rare haplotypes being difficult to represent accurately in haplotype graph models.

In summary, it seems plausible that the slightly modified version of this algorithm, presented by Browning [2006] and here, has a good learning behaviour. Simulations and empirical data support this conclusion [Browning and Browning, 2007].

Finally, it should be noted that the problem of haplotype graph inference is in principle tractable by standard statistical techniques, such as reversible jump Markov Chain Monte Carlo (rjMCMC) and ML. However, depending on the observed sample, the space of possible graph topologies can become larger than the space of possible genealogies in the coalescent without recombination, rendering exhaustive treatment nearly impossible. Which sampling methods could be used to appropriately explore the space of possible haplotype graph topologies is an open research question.

Summary

We have explored two tractable models of haplotype structure, the L&S approximation and haplotype graph models. Both fit in the class of leveled Hidden Markov Models, which enables the use of optimized versions of standard HMM algorithms for inference problems like total probability and path probabilities. By assuming agnostic missing data probabilities (i.e. assigning equal probabilities to emit the “missing data” symbol, see Section 2.2.1) , the models deal with missing data in a straightforward way.

Both models, however, require an existing sample of haplotypes, which may not always exist. How imputation frameworks deal with this problem and use haplotype representation models to infer missing genotypes remains to be seen.

2.2.2 Imputation frameworks

Besides choosing a basic model for haplotype structure representation, designing an imputation framework requires addressing a couple of statistical and technical questions. For example, in what way shall the framework combine information from various sources and deal with missing data? From a theoretical perspective, inference on missing data should be based on a joint probability distribution of all observed and all missing data [Howie et al., 2009], but large amounts of missing data may render exploring this distribution more difficult than expected. Could it therefore be beneficial not to treat all data at the same time and adopt a model of step-wise conditional inference? Also, how is the imputation framework going to address the fact the haplotype information is often not present? What computational challenges arise, and how to deal with them?

In the following, I will review IMPUTE and BEAGLE, two popular and accurate general-purpose imputation frameworks [Marchini and Howie, 2010]. I will highlight the different ways of addressing some of the questions mentioned and then go on to discuss some HLA-specific aspects.

IMPUTE

There are two versions of IMPUTE – IMPUTE version 1 essentially requires a phased reference panel [Marchini et al., 2007], for example deterministically phased HapMap samples [Consortium et al., 2007]. IMPUTE version 2 (IMPUTE2) extends the original model and enables the combination of phased (haploid) and unphased (diploid) reference panels [Howie et al., 2009]. IMPUTE2 is probably the most accurate SNP genotype imputation program [Marchini and Howie, 2010].

Based on the concepts introduced in the previous sections, it is straightforward to describe the IMPUTE2 algorithm. Denote the haplotypes in the haplotype reference panel as H_{HR}, the haplotypes in the diploid reference panel as H_{DR} and the haplotypes in the diploid imputation panel as H_{DI} (note that the last two sets need to be populated by the algorithm). For terminological clarity, we also define the respective genotype panels,

G_{HR}, G_{DR} and G_{DI}, and we use the index i to denote the i-th individual in these sets, e.g. $H_{DR,i}$.

Define three sets of SNPs: type A SNPs are typed in all panels, type B SNPs are typed in the diploid reference panel and in the haploid reference panel, type C SNPs are only typed in the haploid reference panels. We use A, B and C as indices if we want to constrain the SNPs in any of our genotype or haplotype panels to a particular set, for example as in G^A_{DR}.

The space of possible individual haplotypes in the diploid panels is now going to be explored using Markov Chain Monte Carlo (MCMC). Howie et al. [2009] use a diploid version of the L&S approximation to specify haplotype pair probabilities, and a haploid version of the L&S approximation to impute missing data into haplotypes (by treating missing SNP values as missing data, and applying path sampling algorithms). A fine-scale recombination rate estimate is used in the L&S approximation, for example from Myers et al. [2005]. We need to define a final notational convention: the symbol "-" is going to be used to denote exclusion of individual data from a set, so that $H^A_{DR;(-i)}$ denotes the set of haplotypes of the SNPs in A in the diploid reference panel, with all data coming from individual i excluded.

To set up MCMC, start with a "guess" of haplotype pairs for all diploid individuals, i.e. a random draw from a uniform distribution of haplotype pairs compatible with an individual genotype. Now, start iterating:

1. For all individuals i in the DR group, select a new haplotype pair for SNPs in A and B, $h^{A,B}$ according to $\mathrm{P}_{L\&S}(h^{A,B}|G^{A,B}_{DR,i}, H^{A,B}_{DR;(-i)}, H^{A,B}_{HR})$.

2. From step 1, we have haplotypes $H^{A,B}_{DR}$ for all individuals in set DR at SNPs A and B. SNPs C are missing in the haplotypes. For all individuals i in DR, impute the missing C SNPs into $H^{A,B}_{DR}$, according to $\mathrm{P}_{L\&S}(h^{A,B,C}|H^{A,B}_{DR,i}, H^{A,B,C}_{HR})$ (in two independent haploid steps). Haplotypes in DR are now complete.

3. Phase SNPs in set A in the DI panel. For each individual i in the diploid imputation panel, sample a pair of haplotypes for SNPs in A, h^A, according to

$\mathrm{P}_{L\&S}(h^A|G^A_{DI,i}, H^A_{HR}, H^A_{DR}, H^A_{DI;(-i)})$.

4. SNPs in set B are missing in the DI panel. For each individual i in DI, we have haplotypes H^A_{DI}. Now impute (in two independent haploid steps) set B in each haplotype for each individual i in DI, according to $\mathrm{P}_{L\&S}(h^{A,B}|H^A_{DI}, H^{A,B}_{HR}, H^{A,B}_{DR})$.

5. SNPs in set C are missing in the DI panel. For each individual i in DI, we have haplotypes H^A_{DI}. Now impute (in two independent steps) set C in each haplotype for each individual i in DI, according to $\mathrm{P}_{L\&S}(h^{A,C}|H^A_{DI}, H^{A,C}_{HR})$.

IMPUTE uses burn-in iterations, comprising only steps 1 and 3, to generate plausible haplotypes before imputing any genotypes. When the algorithm has finished, results from steps 2, 4 and 5 can be used to generate genotype distributions for missing loci in the imputation panel.

The algorithm exemplifies how interrelated the steps of phasing and imputation are. The phasing step is very similar to that of PHASE [Stephens and Scheet, 2005]. PHASE, however, operates on a set of pre-computed set of "plausible" haplotype pairs for each individual and assigns these pairs to individuals in a Gibbs sampling step, each iteration requiring the full probability distribution over all plausible pairs. IMPUTE2, in contrast, allows for a new path through the other individuals' haplotypes in each iteration. To limit the state space of the induced diploid HMM, not all available individuals are included in the HMM, only those which meet a similarity criterion based on Hamming distance.

Also, note how IMPUTE2 models the flow of information: only data at typed SNPs is used to infer haplotype phase, and once phase has been inferred, haploid imputation is carried out. By designing the algorithm in this way, Howie et al. [2009] avoid having to take into account missing data at untyped loci while determining haplotype phase. The authors view this as an essential advantage over the BEAGLE algorithm [Howie et al., 2009], which is presented in the next section.

BEAGLE

The BEAGLE algorithm represents the most popular implementation of haplotype graph models in population genetics. Whereas the haplotype graph model used by BEAGLE was introduced by Browning [2006], explicit application to haplotype phase inference and sporadic missing data imputation came only with Browning and Browning [2007]. Later implementations were optimized for the imputation of systematically missing data in imputation panels, informed by reference panel data [Browning and Browning, 2009].

The BEAGLE algorithm is an iterative procedure, aimed at improving the fit of a haplotype graph model to some observed data D. D may consist of haplotypic data D_H and genotypic data D_G, and no individual may be represented in both sets. The algorithm is most easily understood by focusing on genotypic data. Generate a set H by randomly "guessing" haplotypes compatible with D_G for each individual (i.e. randomly draw a haplotype pair which is compatible with each individual's observed genotype). Now, iterate the following steps a predefined number of times:

1. Build a haplotype graph model G for the haplotypes H, as described in Section 2.2.1.
2. As described, G induces haploid and diploid HMMs. Use the diploid HMM to sample R new haplotypes (paths through the model) for each individual.
3. Empty H and re-populate with the samples from step 2.

As a threshold for merging two nodes v_1 and v_2, Browning and Browning [2007] propose the following variance-based value:

$$\epsilon := d \times \sqrt{R} \times (\text{count}(v_1,'')^{-1} + \text{count}(v_2,'')^{-1})^{1/2}\,,$$

where d is a scale parameter (Browning and Browning [2007] suggest that 0.7 is a sensible choice) and R is the number of haplotype samples from each individual. R needs to be part of the definition of ϵ to take into account the non-independence of multiple samples from the same individuals.

If haplotypic data D_H is available from the beginning, it is either possible to put the unmodified data into H, or to use the haplotype graph's induced haploid HMM to sample new paths through the model for each element of D_H.

The standard BEAGLE algorithm does not explicitly account for missing data. When populating H for the first time, for each missing genotype value, an allele which is compatible with the respective level-specific model alphabet is randomly selected according to allele frequencies. The induced diploid HMM uses agnostic emission probabilities for missing data, so that later iterations of the algorithm (conditional on missing and non-missing genotype data) are expected to improve genotype (haplotype) estimates at missing positions. In general, distributions over possible genotypes can be derived from integrating over possible underlying haplotype pairs.

Comparing the BEAGLE algorithm to IMPUTE2, information from all panels and loci is equally combined to fit a haplotype graph model. This may, as observed by Howie et al. [2009], lead to the problem that estimates for missing data may negatively affect the accuracy of haplotype estimates, and thus decrease imputation accuracy. In later implementations of BEAGLE [Browning and Browning, 2009], a weighting scheme is introduced to make the algorithm more robust against this effect: effectively, during the first iterations of the algorithm, the number of samples taken from an individual is inversely related to the proportion of missing data in the individual's genotypes. This will have the effect of increasing the importance of the reference panel during the first iterations, allowing for more robust inference of missing data during the later stages of the algorithm.

Interestingly, the original BEAGLE induced HMM does not allow for deviations from the underlying edge genotype values: the emission probability is set to 0 for all other members of the level-specific model alphabet. This restriction, however, is relaxed in other works [Browning and Yu, 2009].

The accuracy of BEAGLE in phasing as well as in SNP genotype imputation increases with the size of the input dataset [Browning and Browning, 2011; Marchini and Howie, 2010]. It is noteworthy that BEAGLE is extremely fast: analysis of thousands of markers on hundreds of individuals is often performed in a couple of minutes.

Li&Stephens and BEAGLE in the HLA

The standard L&S approximation, the model IMPUTE2 is based on, is not well-suited to deal with the haplotype structure of the MHC region. Imputation under the L&S model assigns more weight to markers close to the locus which is going to be imputed – a behaviour which is consistent with standard population genetics: LD is expected to decline with genetic distance, rendering the allelic state of distant markers less informative than the allelic state of nearby markers. Paradoxically, however, many markers which are substantially correlated with the classical HLA alleles are in considerable distance from the classical HLA loci [de Bakker et al., 2006]. More formally, the long-range linkage relationship $\mathrm{P}(S_{l,i}|S_{l-x,i}) = 1\ , x > 0$ cannot be modeled within the standard L&S framework, and therefore the L&S approximation cannot be expected to perform well in the MHC region. Experiments have confirmed this expectation [Moutsianas, 2011, p. 194]: an standard Li&Stephens algorithm produced much worse (between 3% and 31%) results in a one-population leave-one-out cross validation setting than LDMhc (Leslie et al. [2008], see next chapter).

The unmodified BEAGLE algorithm achieves high accuracies when predicting classical HLA alleles (see later chapters for some comparative experiments). In contrast to the L&S approximation, which spreads transition probabilities in case of a recombination event uniformly over all haplotypes in the model, the cluster structure (allowing for $\mathrm{P}(S_{l,i}|S_{l-x,i}) = 1$) of haplotype graphs may be more appropriate when dealing with MHC haplotypes.

However, it is not clear whether the BEAGLE implementation in its current form is optimal for performing inference in the HLA.

- The standard BEAGLE HMM does not allow for deviations from underlying edge symbols; this may lead to problems when encountering SNP genotyping quality problems, to which the MHC is, because of abundant structural variation, particularly prone.

- The haplotype graph model construction algorithm does not account for the fact

that there is uncertainty in in the set H of haplotypes, for example induced by genotyping error.

- The haplotype graph model construction algorithm does not allow for specifying particular loci of interest that imputation performance should be optimized for.

- Finally, it is not clear whether the way BEAGLE deals with missing data is optimal. Howie et al. [2009] raise the point that focusing on the haplotype structure of the actually typed markers first may be beneficial. Integrating missing data into the model building process may be worth investigating.

2.2.3 Imputation in the HLA

Because of the classical HLA proteins' importance in the immune system, and because of these genes being associated with many important diseases, surveying variation in the HLA in a cost-effective manner is an important goal.

There is no reason to believe that it should be impossible to reliably impute classical HLA alleles (however, as I have discussed in the preceding section, there are reasons and some evidence which suggest that standard models do not achieve optimal performance in the HLA region).

In the following chapters, I will describe two HLA-specific imputation methods, how they represent HLA alleles and how they deal with the particularities of the MHC. I also describe how to use their imputations to inform disease association studies in a logistic regression framework. One of the methods, HLA*IMP, is based on the Li & Stephens approximation. HLA*IMP:02, the other method, employs haplotype graph models, naturally building on the material presented in this introduction.

I will throughout focus on 4-digit HLA types, as they capture the primary structure of the HLA proteins and not just their broader serological features.

Chapter 3

HLA*IMP: an integrated HLA type imputation framework

In this chapter, I will present the model and implementation of HLA*IMP, an integrated HLA type imputation framework [Dilthey et al., 2011]. HLA*IMP was designed with two main goals in mind: high accuracy and good availability. The first goal specifically requires an implementation which is able to deal with large reference panels. The second aim refers to practical availability for the scientific community - HLA*IMP was supposed to be implemented in a way which allows biologists and biomedical researchers to impute HLA types, without much statistical expertise.

HLA*IMP is based on the LDMhc algorithm, which was presented by Leslie et al. [2008]. To meet the requirements outlined above, LDMhc was modified and extended in a couple of ways:

- A new likelihood-based SNP selection function was developed, which increases call rates in many scenarios.
- The model building algorithm was parallelized, which allows for the efficient use of large (a couple of thousand haplotypes) reference panels. In order to make efficient use of computational resources, an automated stopping rule for the algorithm was devised.

- A large reference panel, consisting of 5024 densely typed chromosomes, was assembled from various sources.
- A front-end / back-end web application was developed, which allows external users to make use of the HLA*IMP imputation services in a user-friendly way.

In the following, I will describe the original LDMhc algorithm and the modifications made for HLA*IMP. I will also present two validation experiments to demonstrate that HLA*IMP meets the requirement of high accuracy.

3.1 LDMhc

In contrast to BEAGLE and IMPUTE, which are mainly aimed at SNP imputation, LDMhc was specifically developed to impute the allelic state of the classical HLA genes [Leslie et al., 2008].

The LDMhc algorithm is based on the observation that the distribution of markers in the human MHC which are informative of IBD at the classical loci follows an unusual pattern. Some HLA alleles, for example, can be tagged by certain SNPs, but it is not always the case that SNPs which are highly correlated with a classical HLA allele are in the direct proximity of the corresponding genetic locus [de Bakker et al., 2006]. The approach of Leslie et al. [2008] therefore tries to construct a representation of allele-specific haplotype backgrounds and use this information to impute missing data. More specifically, for each locus L, LDMhc tries to isolate a set of SNPs S_L which, taken together, can be used to distinguish between the haplotypic backgrounds of as many alleles as possible.

Assume that there is a phased haplotype reference panel H containing H_N haplotypes, typed at T loci, and we use $H_{i,l}$ to denote the genotype of the l-th typed locus at haplotype i. There is no missing data in H. Furthermore assume that there is an additional haplotype c, with data on the same T loci, and we use c_l to denote the genotype of the l-th typed locus on c. Again, there is no missing data in c. All T positions are SNPs.

The reference panel is typed at another genetic marker L. L can be multi-allelic, for

example a classical HLA locus. We use L_i to denote the genotype of the i-th haplotype in H at L. Let A_L denote the set of alleles present at T, i.e. $A_L := \{L_1..L_{H_N}\}$. For all $a \in A_L$, we define $C_{H;L;a}$ to be the set of haplotypes in H which carry allele a at locus L, i.e. $C_{H;L;a} := \{H_i \in H | L_i = a\}$.

c^S, H^S and $C^S_{H;L;a}$ denote the respective haplotype / sets of haplotypes, constrained to the set of SNPs specified by S. S can be thought of as an index of SNPs which are going to be considered in a particular step of the algorithm.

LDMhc now defines a (posterior) probability distribution over A_L for the untyped locus c_L, corresponding to haplotype c, for a particular set of considered SNPs S:

$$\mathrm{P}(c_L = a | c, S, H, L) := \frac{\Pr(a) \times \mathrm{P}_{L\&S}(c^S | C^S_{H;L;a})}{\sum_{x \in A_L} \Pr(a) \times \mathrm{P}_{L\&S}(c^S | C^S_{H;L;x})},$$

where $\Pr(a)$ is a prior distribution over A_L, which could for example refer to population allele frequencies. $\mathrm{P}_{L\&S}$ needs to be specified with suitable recombination and mutation parameters ρ and θ for the SNPs in S.

The decision on what SNPs S should include (*SNP selection*) is based on an optimality measure $O(S)$. Leslie et al. [2008] propose to measure optimality in a leave-one-out cross-validation setting based on the reference panel, by counting for how many haplotypes the maximum likelihood allele prediction under S is consistent with the haplotype's classically determined genotype. We define

$$o(i, S, H, L) := 1_{L_i}(\arg\max_{a \in A_L} \mathrm{P}(a | H_i, S, H_{(-i)}, L_{(-i)})),$$

where $H_{(-i)}$ and $L_{(-i)}$ are the sets H and L without the i-th element, and $1_{L_i}(x)$ is the indicator function which is 1 if $L_i = x$ and 0 otherwise. $o(i, S, H, L)$ is only well-defined if $C_{H_{(-1)};L_{(-1)};a}$ is not empty, i.e. if there are at least two chromosomes carrying allele L_i.

We define the full optimality function $O(S)_{H,L}$ as

$$O(S)_{H,L} := \sum_{a \in \{x \in A_L \,|\, |C_{H;L;x}| > 1\}} \left[\sum_{i \in C_{H;L;x}} o(i, S, H, L) \right].$$

The larger $O(S)_{H,L}$, the better the set S (Leslie et al. [2008] actually define an equivalent optimality function that is supposed to be *minimized* – in the context of this thesis, framing the problem in terms of a maximization is probably more intuitive. There is also an extension of the optimality function in Leslie et al. [2008] to optimize for particular call thresholds, but the authors note that this function is not used for any of the experiments presented in their paper).

To select a subset S from the available SNPs T, carry out the following forward-backward algorithm:

- Set $S = \{\}$.
- If $|S| > 41$, terminate.
- Compute $s = \arg\max_{x \in (\{1..T\} \setminus S)} O(S \cup x)_{H,L}$.
- Set $S = (S \cup s)$.
- Compute $s_2 = \arg\max_{\{x \in S\}} O(S \setminus x)_{H,L}$.
- If $s_2 = s$, go to step 2.
- Otherwise, set $S = (S \setminus s_2)$, go to step 2.

The algorithm as described here terminates after having selected 42 SNPs (as proposed by Leslie et al. [2008] without motivation, assumedly for practical reasons). Leslie et al. [2008] propose to track the development of S and of $O(S)_{H,L}$, and to retrospectively select the best S with the smallest number of elements.

Leslie et al. [2008] also present some validation data: leave-one-out cross-validation performance of the algorithm is in the 90% range ($\geq$88%) for all six classical HLA loci (*HLA-A*, *-B*, *-C*, *-DQA1*, *-DQB1*, *-DRB1*) at 4-digit resolution without setting a threshold. The HapMap CEU and YRI cohorts were used as reference panels [Consortium et al., 2007].

However, cross-validation is prone to overfitting. A second validation experiment, based on using the CEU cohort as training data and data from the 1958 Birth Cohort as validation, yields accuracies between 72 and 91% for four classical loci. Setting a threshold substantially increases accuracy, but sometimes at the expense of markedly decreased call rates (for example, *HLA-DRB1* with Affymetrix SNPs: accuracy increases from 72% to 83% when applying a call threshold of 0.9, but the corresponding call rate goes down to 51%).

Before proceeding to HLA*IMP, is is worthwhile discussing some of the characteristics of LDMhc:

- LDMhc is computationally intensive - it scales to the order of $O(2H_N^2 \times T)$. From the algorithm as presented in Leslie et al. [2008], it is not clear what an efficient parallelization could look like.

- $O(S)_{H,L}$ only takes into account the number of correctly predicted chromosomes - is it possible to develop a more finely granulated metric?

- LDMhc is presented in a Bayesian framework, the posterior probabilities over HLA alleles being based on $\Pr(a) \times P_{L\&S}(c^S|C^S_{H;L;a})$. To some extent unorthodoxically, $P_{L\&S}(c^S|C^S_{H;L;a})$ does not take the full data into account – the chromosomal haplotype groups, as defined by the HLA allele they carry, are treated as independent. If one relates the L&S approximation back to the population genetics concepts it is supposed to capture, this effectively means that LDMhc assumes that there is no recombination between these chromosomal haplotype groups (the structure of a haplotype carrying allele a only depends on the other haplotypes carrying a).

- The allele-specific grouping approach of LDMhc assigns equal weight to each selected SNP; this circumvents some of the L&S approximation's general problems with modeling long-range LD (see Section 2.2.2).

- The good performance of the algorithm, even on comparably small reference panels, can probably be explained by SNP selection: SNP selection eliminates SNPs which do not aid imputation or which otherwise conflict with the model's assumptions.

SNP selection, however, may have adverse effects if the reference panel is contaminated with classical HLA typing errors, as it will push the model towards predicting the corresponding chromosomes as consistent with their classical types.

3.2 Core model and implementation of HLA*IMP

As outlined above, HLA*IMP modifies the model and implementation of LDMhc to increase call rates and to allow for an efficient parallelization. Instead of giving a full description of the HLA*IMP model and reproducing the parts which are identical with LDMhc, I will highlight the differences between LDMhc and HLA*IMP.

SNP selection

LDMhc tries to optimize for the number of correctly imputed chromosomes. This is an important goal; however, it is arguably also important to optimize the confidence one has in correct calls, as reflected by the posterior probability distributions. By computing $\mathrm{P}(\text{truth})$ for each haplotype and maximizing the sum over all individuals, both requirements are met (that is, LDMhc counts an optimality value of 1 if the maximum likelihood prediction matches the classically determined type. HLA*IMP, in contrast, takes into account the probability of the "correct" allele under the current model). More formally, in HLA*IMP, $o(i, S, H, L)$ is re-defined as

$$o(i, S, H, L) := \mathrm{P}(L_i | H_i, S, H_{(-i)}, L_{(-i)}) \,.$$

It should be noted that this function – and the original function used by LDMhc – makes the assumption that L_i is actually a correct reflection of the individual's genotype. That is, $o(i, S, H, L)$ does not take into account that a typing error may have occurred. To the knowledge of the author of this thesis, there are no general evaluations of classical HLA typing accuracy; in re-test experiments, accuracy strongly depends on the employed protocols, which may change over the course of a couple of years (Johannes Fischer,

personal communication). Typing uncertainty is therefore difficult to quantify based on prior knowledge. In theory, cross-validation techniques could be used to estimate whether a classical typing is correct, but this would furthermore increase the computational demands of the method. I will return to the question of classical typing uncertainty in Chapter 5.

Parallelization

As a first naive approach, one may parallelize by distributing the number of SNPs that are to be evaluated during the forward selection phase over computing nodes, without significantly altering the algorithm. In fact, the only major necessary change would be to add a function that stores the local best SNPs (from the set that was assigned to each computing node) and determines, once all nodes have finished evaluating SNPs, the globally best SNP.

This approach is not very efficient. By exploiting the structure of the problem at hand, better implementations can be found.

Suppose already having selected a set S of SNPs, ordered by their genetic position. A SNP s shall be added to the set, to compute the L&S HMM emission probabilities for all groups of training chromosomes. Keep in mind that the L&S approximation is a leveled HMM (Section 2.2.1). Calculate the position of s in the ordered set S, by determining the index i of the level after which s has to be inserted, if ordered by genetic position (using 0 if s is in front of all SNPs in S). Now, computing the emission probabilities for an arbitrary chromosome based on the group $S \cup s$, using the forward algorithm, one performs, up to the i-th SNP, exactly the same calculations that one had performed in case of evaluating performance for S, without s. The same principle applies to the backward algorithm. What is more, by combining the results from forward and backward algorithm at an arbitrary level (here: SNP), one can calculate the full emission probability (see Section 2.2.1 and Rabiner [1989]). Consequently,

- calculate the forward- and backward tables (α_{a,H_i} and β_{a,H_i}) corresponding to $\mathrm{P}(a|H_i, S, H_{(-i)}, L_{(-i)})$ for S, all elements in $a \in A_L$, and all $H_i \in H$.

- note that the structure of the underlying leveled HMM for α_{a,H_i} and β_{a,H_i} changes if $L_i = a$ – in this case, H_i needs to be removed from the corresponding LHMM.

- order the set of available (i.e. $\notin S$) SNPs by their genetic position.

- iterate over the ordered set of possible SNPs. Perform the usual "leave one out" algorithm for each possible SNP s, but when it comes to calculating $\mathrm{P}(a|H_i, S, H_{(-i)}, L_{(-i)})$ for allele group a, use the pre-calculated α_{a,H_i} and β_{a,H_i} tables instead of re-populating the full forward and backward tables. This is done by re-calculating column s in α_{a,H_i} and β_{a,H_i}, conditional on $H_{i,s}$, and combining that data as described in Section 2.2.1 to obtain the total emission probability.

- Informally, the column introduced by adding s in the α and β tables is the only one that has to be re-calculated (see Figure 3.1). Its values depend on the genotypes of s in the set of reference chromosomes minus i and position s on H_i, but also on the recombination probabilities along the levels. Processing SNPs in their linear order speeds up the process of adding up the recombination probabilities of skipped SNPs along the chromosome (skipped, because they are not currently members of S).

All these steps can be applied in a completely non-parallel setting and make calculations significantly faster. The effect is that the practical computational effort for adding one SNP remains nearly constant, independent of the size of the set of already selected SNPs. The algorithm is still in the same class of complexity, but faster by a factor of 20 - 30 (strongly depending on the particular use case). What is more, an efficient parallelization is immediately at hand: parallelize by distributing SNPs in their genetic order over computing nodes, but make sure that each node has access to α and β. Now compute the optimality measure for all SNPs as described, and select the globally best SNP from the per-node local best SNPs.

In implementing the described parallelization, a couple of technical details deserve some attention. The most important question is how to achieve sharing of α and β among nodes. This can be implemented by a variety of means, e.g. by employing Ethernet-based messaging protocols. In such cases, communicational overhead becomes potentially import-

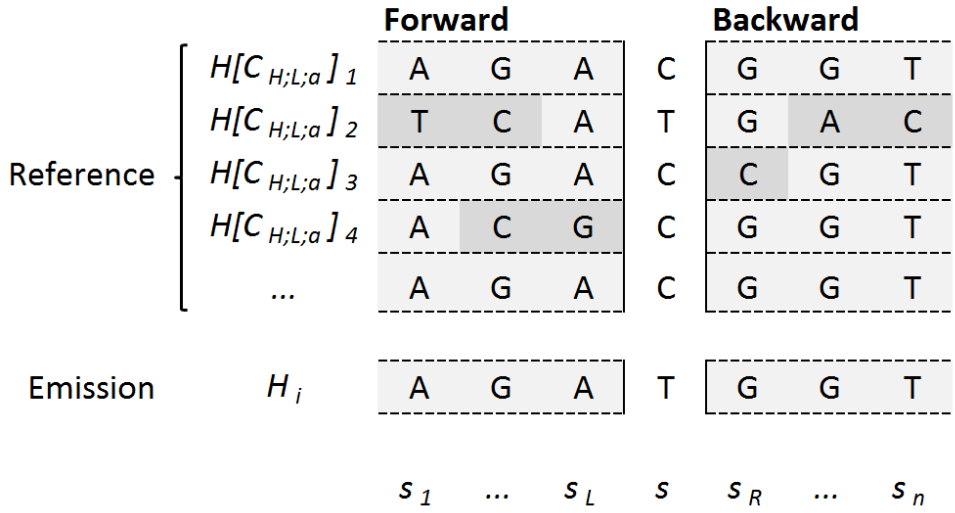

Figure 3.1: Visualization of the L&S Hidden Markov Model states for a group of reference chromosomes carrying the a allele, here denoted as $H[C_{H;L;a}]_1$, $H[C_{H;L;a}]_2$, ... Usually, the computation of an emission probability for a given chromosome H_i would involve filling the corresponding forward-table from states s_1 to s_n and summing over the entries in s_n. However, the emission probability can also be calculated at any point s in the HMM, by combining the forward- and backward-tables up to s. Both tables for each chromosome in are computed in advance (grey cells in the figure, polymorphisms highlighted in dark grey). The specific transition and emission probabilities for any given SNP s (middle column) are then added in parallel, which can be performed without changing the pre-computed table values. Figure adapted from Dilthey et al. [2011].

ant. For HLA*IMP, an openMP-based implementation for shared-memory machines was chosen; shared-memory machines range from multi-core workstations up to medium-sized supercomputing devices, such as the ORAC Itanium2 cluster at the Oxford Supercomputing Center (OSC). The shared memory architecture conveys the additional benefit that the whole machine's memory, 1 TB for ORAC, is available to all CPUs. This is important as caching the α and β tables can require substantial amounts of memory, for example up to 25 GB for one HLA locus of the standard reference panel which was assembled for HLA*IMP (growing quadratically with the number of chromosomes). All code was written in a way that allows for compilation by the highly optimizing Intel ICC compilers. The core programs, written in C and C++, are embedded in a Perl-based framework of supporting programs, that facilitate use and establish communication between the core programs.

Automated stopping

For reasons of cost-efficiency, computational resources on supercomputing facilities should be used as effectively as possible. LDMhc requires selection of a full set of x SNPs, and then a retrospective review to decide at which point the set of selected SNPs was optimal in terms of performance as well as set size. For HLA*IMP, this protocol was replaced with an automated procedure. HLA*IMP monitors how $O(S)_{H,L}$ changes as the number of SNPs in the model increases (i.e., at each time point used for monitoring purposes, forward selection has added a SNP and backward selection has not eliminated another SNP). If the increase in $O(S)_{H,L}$ is not above a certain threshold for two subsequent steps, SNP selection aborts. Initial experiments have suggested that 0.5 is a good choice for that threshold.

Validation methodology

In order to be able to validate the imputations generated by HLA*IMP, I define three accuracy metrics:

- Haplotype validation. I compare an imputation I with a validation result L. I carries a posterior probability. If a call threshold has been defined and the posterior probability of I is below the threshold, or if L is not defined up to 4-digit (2-digit for 2-digit validation) accuracy, the haplotype is ignored. Otherwise, it is counted as correct if and only if $I = L$.

  Haplotype validation is employed for all experiments presented in this chapter (with one exception; see below).

- Genotype validation. I compare two unordered sets with two elements each for each individual, one set (I) representing the imputation results and the other (L) containing the lab-derived types. I only consider individuals who carry two HLA alleles typed at 4-digit resolution at the locus under validation or one allele at 4-digit resolution and one missing allele. For 2-digit validation, I consider the same individuals, but I set the last 2 digits of each HLA allele to '00' (this will lead to an

underestimation of accuracy in some cases, as there are some serologically defined 2-digit allele groups that map to more than one pair of leading two digits). A posterior probability call threshold may or may not be applied before validation.

If there is no missing data in L, there are three possible cases:

- 0 imputations left after thresholding: I count 0 correctly imputed alleles out of 0.
- 1 imputation (I_1) left after thresholding: I count 1 correctly imputed alleles out of 1 if $I_1 \in L$, otherwise 0 out of 1.
- 2 imputations left after thresholding: I count 0 correct imputations out of 2 if $(I_1 \notin L) \wedge (I_2 \notin L)$, 1 out of 2 if $(I_1 \in L) \veebar (I_2 \in L)$, 2 out of 2 otherwise.

If $L = \{\text{missing}, A\}$ (i.e. only one allele has been typed), there are also three possible cases:

- 0 imputations left after thresholding: I count 0 correctly imputed alleles out of 0.
- 1 imputation (I_1) left after thresholding: I count 1 correctly imputed alleles out of 1 if $I_1 = A$, otherwise 0 out of 1.
- 2 imputations left after thresholding: I count 1 correct imputations out of 1 if $I_1 = A$ or $I_2 = A$ or both.

Genotype validation is applied throughout the next chapters, to assess and compare the performance of HLA*IMP and HLA*IMP:02 (see next chapter).

- "Genotype overvalidation". This is the validation methodology from Leslie et al. [2008] and it is only employed here once to enable a direct comparison (Table 3.1) with data presented in their paper. Genotype overvalidation validates on the genotype level, but in a slightly nonintuitive way. Suppose there are two imputed alleles A and B for one individual and that there is a set R of classically typed reference alleles for the same individual. The approach of Leslie et al. [2008] now assesses A and B independently, and counts each one of them as correct if $A \in R$ ($B \in R$, respectively). That is, for example, if the two imputed alleles are 0201, and the set R

is $\{0201, 0101\}$, *both* imputed alleles are counted as correct. Genotype overvalidation will overestimate accuracy.

According to these metrics, every imputation can be classified as either correct, not correct or ignored (for the presence of missing data or a call threshold). We can aggregate these measures at the level of a single locus (per-locus validation) or at the level of single alleles (per-allele validation).

At the per-locus level, accuracy is measured as concordance. That is, the number of correct imputations is divided by the number of correct imputations plus the number of incorrect imputations. This is equivalent to the definition of PPV (positive predictive value).

At the per-allele level (for allele A, say), there are various meaningful metrics. PPV measures the proportion of A imputations that are actually A. Sensitivity measures the the proportion of true A alleles that are imputed as A. Finally, and most importantly in the context of GWAS (see the discussion on page 32), there is r^2, the squared correlation coefficient between the imputed and true allele number. r^2 directly relates to the power of detecting an effect mediated by A. In the context of evaluating the expected power in GWAS, r^2 should be measured at the genotype validation level.

A small validation experiment

To assess the impact of the described changes (SNP selection and stopping) on imputation accuracy, a small validation experiment, comparing the original implementation of LDMhc with HLA*IMP, was carried out. Data, validation methodology ("genotype overvalidation") and results for LDMhc were replicated from Leslie et al. [2008]. The experiment presented here uses CEU HapMap data [Consortium et al., 2007] as reference panel and 1958 birth cohort data as imputation panel [Consortium, 2007]. The Affymetrix SNP set scenario is, based on Leslie et al. [2008], slightly more challenging, and therefore used for comparison.

Table 3.1 summarizes the results. The new algorithm based on optimizing posterior probabilities typically outperforms the old SNP selection algorithm, in particular when

a threshold is applied. This effect is largely driven by increases in call rates rather than increased accuracies (conditional on a set call rate). For example, call rate increases from 25% to 75% for *HLA-DQB1* at a call threshold of 0.9. At T = 0.9, the total number of correctly predicted alleles increases by 44% over all loci. At lower thresholds this number is typically (though not consistently) increased.

Threshold	Locus	# Validated	2008			2010		
			Accuracy (overvalidation)	Call Rate	# Correct	Accuracy (overvalidation)	Call Rate	# Correct
T = 0.00	*HLA-A*	876	0.89	1.00	780	0.91	1.00	794
	HLA-B	1630	0.82	1.00	1337	0.78	1.00	1276
	HLA-DRB1	834	0.72	1.00	600	0.68	1.00	566
	HLA-DQB1	1088	0.77	1.00	838	0.83	1.00	901
T = 0.50	*HLA-A*	876	0.91	0.93	741	0.91	0.99	786
	HLA-B	1630	0.85	0.88	1219	0.80	0.97	1269
	HLA-DRB1	834	0.76	0.88	558	0.70	0.94	551
	HLA-DQB1	1088	0.80	0.88	766	0.84	0.98	895
T = 0.90	*HLA-A*	876	0.97	0.58	493	0.94	0.87	720
	HLA-B	1630	0.93	0.66	1000	0.90	0.79	1159
	HLA-DRB1	834	0.83	0.51	353	0.80	0.67	445
	HLA-DQB1	1088	0.93	0.29	293	0.93	0.75	756

Table 3.1: Accuracy ("genotype overvalidation"), call rate and the number of correctly imputed genotypes for the Affymetrix data from Leslie et al. [2008], comparing the 2008 implementation of LDMhc with HLA*IMP at different class threshold (T). "# Validated" refers to the number of validated alleleles (pre-thresholding).

3.3 Assembling a large reference panel: the "Golden Set" (GS)

A large and carefully quality-controlled reference panel is a necessary condition for accurate imputations on a population level. The author of this thesis assembled, in very close collaboration with Loukas Moutsianas, the reference panel which is used in HLA*IMP from a variety of data sources. Referring to the carefully designed steps of quality control that were applied to it, this dataset is subsequently referred to as the "Golden Set" (GS). It was created by combining HLA- and SNP-genotyped samples from three cohorts:

1. The 1958 Birth Cohort (http://www.b58cgene.sgul.ac.uk/), typed on the Illumina 1.2M (2589 samples) and the Affymetrix Genome-Wide Human SNP Array 6.0 (2711 samples).

2. The HapMap CEU samples (60 samples) [Consortium et al., 2007]

3. The CEPH CEU+ additional samples (32 samples) [de Bakker et al., 2006]

3.3.1 Step 1: Cohort-specific protocols

1958 BC. We removed all SNPs not present in HapMap r27 and those outside the extended MHC region (xMHC). The region considered was that defined by Horton et al. [2004] with 50Kb flanks at the two ends. We removed all samples and SNPs which were not typed on both the Affymetrix and Illumina chip in order to be able to compare Affymetrix and Illumina typing data for each genotype in the remaining set (overlap 2462 samples).

We applied an EM-based procedure to align SNP strandedness to HapMap (cohort 2) separately for the Illumina and Affymetrix data. First, non-complementary (i.e not A/T or G/C alleles) SNPs' strandedness was adjusted to HapMap. Then, for each complementary SNP, we selected the two nearest non-complementary neighbors (with the complementary SNP in the middle) and calculated the likelihood of the observed SNP triplet genotypes under two hypotheses and conditional on the observed HapMap haplotypes. H0: the complementary SNP's strand is not inverted compared to HapMap. H1: the complementary SNP's strand is inverted compared to HapMap. We selected the hypothesis with the higher likelihood, set the SNP strand accordingly and used an EM algorithm to estimate haplotype frequencies in the target dataset. Finally, we compared the likelihood of the observed SNP triplet genotypes under the EM-estimated haplotype frequencies and under the HapMap-observed haplotypes and flagged SNPs for removal from both datasets if the two likelihoods showed gross deviations. In some datasets, a substantial proportion of SNPs has to be inverted (for example, 2642 out of 5345 SNPs in the xMHC region in a recent Immunochip-based datasets).

To merge data from the Affymetrix and Illumina datasets, we applied the following procedure: call each SNP genotype based on a call threshold of T = 0.9 on the posterior probability. Non-called SNP genotypes were marked as "missing data". Compare each called SNP genotype between the two datasets and mark them as "missing data" if they do not agree. Finally, calculate the percentage of missing data for all SNPs and all samples and remove each SNP and each individual with missing data >5% from both sets. For the final step of merging, average the posterior genotype probabilities for each genotype and call, employing a threshold of T = 0.9. All genotypes below this threshold are marked as

"missing data". The resulting consensus dataset (2420 samples, 7733 SNPs) was phased using IMPUTE v2 [Howie et al., 2009], including phased HapMap samples as haplotype templates. Missing SNP data was imputed as part of this process.

HapMap. SNPs outside the xMHC region were removed. Missing data thresholds on the SNP genotype and individual level as described for cohort 1 were applied to the HapMap datasets. No data had to be removed. Finally, we removed all SNPs which were not present in the merged version of cohort 1.

CEPH CEU+. SNPs outside the xMHC region were removed. Missing data thresholds on the SNP genotype and individual level as described for cohort 1 were applied. No data had to be removed. IMPUTE v2 was used for phasing, including deterministically phased HapMap samples as reference. Missing SNP genotypes for SNPs present in the merged version of cohort 1 were imputed as part of the phasing process.

In total. The described quality control and data preparation measures led to a set of 5024 high-quality SNP haplotypes in 7733 SNPs in the xMHC region.

3.3.2 Combining classically typed HLA data and SNP haplotypes

In order to obtain full SNP and HLA haplotypes (i.e. HLA alleles phased into their surrounding SNP contexts), we had to merge classically typed HLA genotypes and the SNP haplotypes created in step 1. We applied the following procedure: for each cohort, remove individuals without HLA typing. Use PHASE [Stephens and Scheet, 2005] to phase HLA alleles into the SNP haplotypes inferred in step 1, employing standard settings for multiallelic loci. To do so, treat the SNP haplotypes from step 1 as invariable, so that only the phase of the HLA alleles, determined by the surrounding SNP context, is determined.

3.3.3 Step 3: Final merging procedure & Summary

Finally, we merged SNP and HLA data from all cohorts to obtain the Golden Set, comprising 2474 (*HLA-A*), 3090 (*HLA-B*), 2022 (*HLA-C*), 175 (*HLA-DQA1*), 2629 (*HLA-DQB1*), 2665 (*HLA-DRB1*) HLA- and SNP-genotyped chromosomes. Table 3.2 gives a summary

of the GS on the level of individuals, and provides information on the allelic diversity within the GS.

Number of individuals with at least one allele typed at 4-digit resolution (2-digit for DRB3, DRB4, DRB5)

	SNPs	HLA-A	HLA-B	HLA-C	HLA-DQA1	HLA-DQB1	HLA-DRB1	HLA-DPB1	HLA-DRB3	HLA-DRB4	HLA-DRB5
GS	7733	1556	1570	1153	87	1585	1517	0	0	0	0
GSK_EU	7568	308	1060	349	279	446	897	74	282	282	282
GS&GSK_EU	6056	1864	2630	1502	366	2031	2414	74	282	282	282
GS&GSK_EU 2/3	6056	1253	1758	1017	250	1359	1592	50	187	187	187
GS&GSK_EU 1/3	6056	611	872	485	116	672	822	24	95	95	95
GS&GSK_ALL	7632	2028	3063	1675	521	2223	2809	112	353	353	353
GS&GSK_ALL 2/3	7632	1356	2055	1129	354	1495	1853	77	242	242	242
GS&GSK_ALL 1/3	7632	672	1008	546	167	728	956	35	111	111	111

Number of 4-digit HLA alleles (2-digit for DRB3, DRB4, DRB5)

	HLA-A	HLA-B	HLA-C	HLA-DQA1	HLA-DQB1	HLA-DRB1	HLA-DPB1	HLA-DRB3	HLA-DRB4	HLA-DRB5
GS	27	46	21	7	18	34	0	0	0	0
GSK_EU	34	60	28	14	17	43	18	5	3	4
GS&GSK_EU	38	65	30	14	19	47	18	5	3	4
GS&GSK_EU 2/3	33	60	27	13	19	42	15	5	3	3
GS&GSK_EU 1/3	28	52	24	11	17	37	13	5	3	4
GS&GSK_ALL	58	110	39	15	21	62	21	5	3	4
GS&GSK_ALL 2/3	48	99	34	14	20	56	17	5	3	4
GS&GSK_ALL 1/3	45	85	30	13	20	49	17	5	3	3

Table 3.2: Dataset characteristics summary. The upper part of this table shows the number of individuals that are available for building the HLA*IMP:02 graphs in a locus-specific manner. For HLA*IMP, the number of available haplotypes is approximately double the individual number. The bottom part of the table shows allelic diversity for all reference and validation datasets used in our study. Note that the allelic diversity in the GSK and in the GS&GSK 2/3 datasets is bigger than in the GS.

3.4 HLA*IMP server model

One of the explicitly stated aims of HLA*IMP was to enable non-statisticians, for example biomedical researchers, to use HLA type imputation methodology in their own studies. To this end, HLA*IMP was implemented as a user-friendly front-end / back-end web application framework.

The front-end is designed to assist end users in preparing their data – it has built-in modules for quality control, SNP strand alignment and haplotype phasing (quality control and SNP strand alignment essentially replicating the steps described above for creating the Golden Set). Users are guided through these steps in a wizard-like sequential manner (see Figure 3.2). Output files from some popular genotype callers, including PLINK [Purcell et al., 2007], Birdsuite [Korn et al., 2008] and CHIAMO [Marchini et al., 2007], can be read in directly, as well as a simple generic format.

The back-end part, implemented as an online web-service, carries out the computationally intensive parts of the imputation process. It can automatically process the files generated by the front-end and notifies the end-user via email of completed processes. As the result of the parallelized SNP selection depends on the initial set of available SNPs, SNPs for some popular Affymetrix and Illumina SNP genotyping platforms were pre-selected. HLA*IMP is free for academic use and available from http://oxfordhla.well.ox.ac.uk. It currently (November 2011) has more than 100 registered users and is used by many international research groups.

3.4.1 Back-end security model

Some of the data submitted to HLA*IMP may be sensitive. To be in agreement with UK data protection laws (the submitted data will be stored on the server for some time), the server only accepts sufficiently anonymized datasets. Still, unauthorized access to imputation datasets may have negative consequences, for the reputation of HLA*IMP itself as well as for the research group which has submitted the underlying SNP genotype data. Suppose, for example, that someone gets access to the imputations of a research group

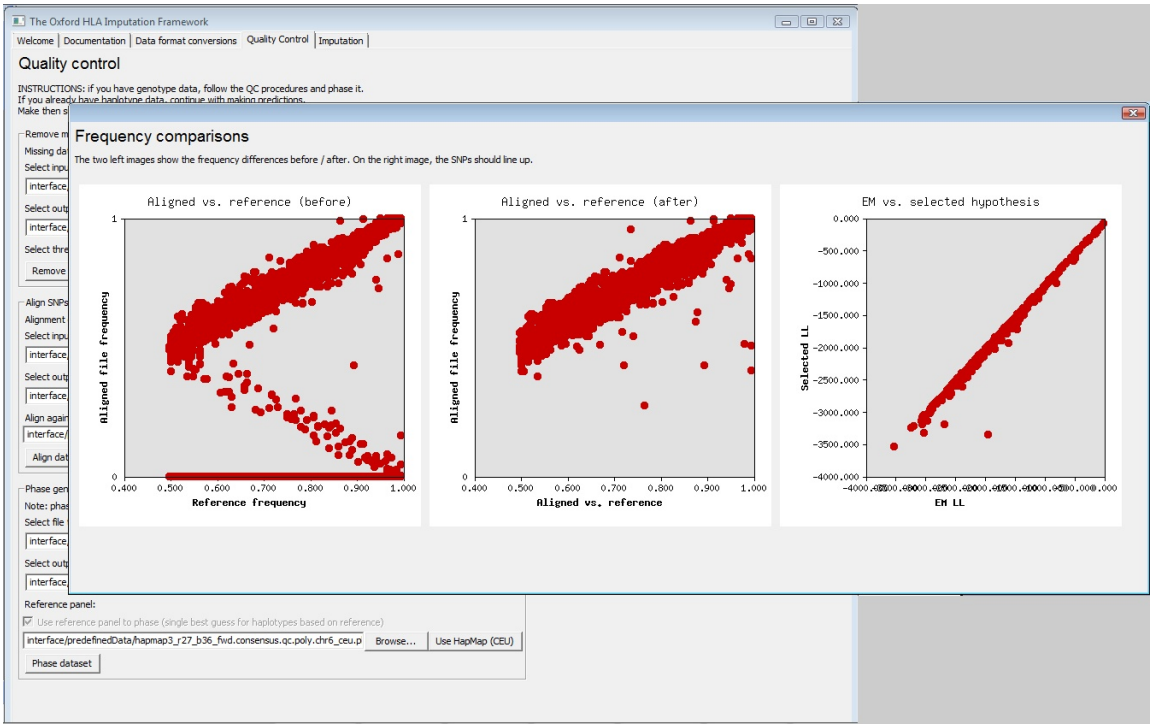

Figure 3.2: The front-end of HLA*IMP [Dilthey et al., 2011] controls for missing data, aligns complementary SNPs and phases haplotypes in a largely automated manner. In this screenshot: graphical output from the alignment procedure, comparing SNP allele frequencies in the user dataset to HapMap allele frequencies, before (left) and after (middle) alignment. Complementary SNPs are aligned using an EM-based procedure. A straight line of data points (right) indicates that there are no gross deviations between EM-estimated and HapMap frequencies.

working on Multiple Sclerosis (MS), which point to previously unknown MHC associations, or manages to manipulate user datasets.

In order to reduce the probability of such events, a sophisticated authentication and security system manages access to data stored on the HLA*IMP server (see Figure 3.3).

The core concept is the separation of the system in two realms: the first being the "sandbox", which runs the standard web server and processes incoming user data. Once the sandbox receives user files, they are securely stored to a file system area with a restrictive security configuration, the most important feature being that the sandbox has only write access to that area. For ease of implementation, the sandbox runs a couple of web application frameworks. Importantly, even a total breach of security in the sandbox area would not enable access to existing user files. This concept is, on database level, also applied to the system's main database which manages workflows and server load.

The second realm, the "back-end", has full access to the protected areas, but does not directly interact with any outside processes. It carries out imputation and safely stores the results in the protected file system area. Once imputations are done, the back-end programs directly send a "security credential" to the user, which is necessary to access the imputations. Importantly, the communication channel between back-end and user does not involve the sandbox, and could be secured by encryption if desired.

Finally, there is third component, the "download.pl", sitting in between the other two realms. download.pl is executed in its own context, with read access to the secure areas and the main user database in the sandbox realm. download.pl uses no framework components, to minimize the risk of an externally introduced security problem. If a user is able to authenticate himself against the main user database (in the sandbox context) and with the security credential received from the back-end, he is granted access to his imputations.

3.5 Validation

To evaluate the accuracy of HLA*IMP, two validation experiments are carried out. In the first experiment, 2/3 of the GS are used as reference panel to impute the remaining

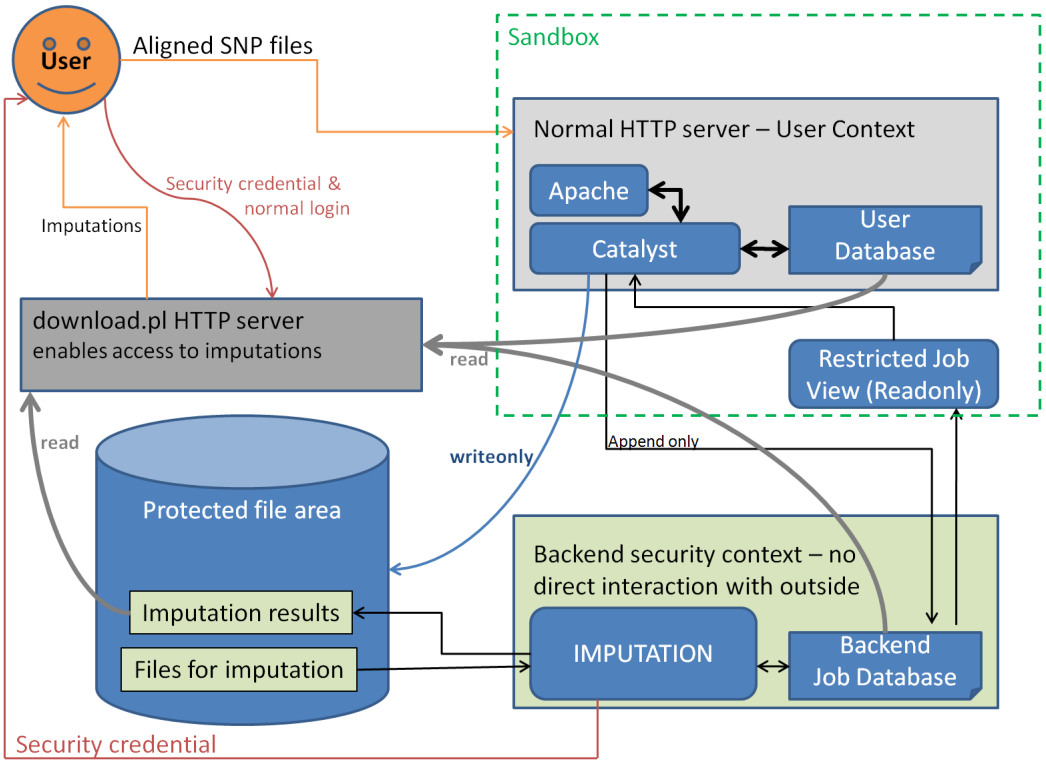

Figure 3.3: The security concept of the HLA*IMP [Dilthey et al., 2011] web server is based on a separation of access rights between different components. Importantly, the "sandbox" context may be compromised, without giving access to existing imputation files. The "back-end" security context carries out the imputations and sends a security credential to the user. Only if the user is able to authenticate himself against the main user database in the sandbox context and using the security credential, he is granted access to his imputation files.

1/3. In the second experiment, the whole GS is used to impute HLA types into another, independent dataset, denoted "GSK_EU" (and described below). Arguably, the second experiment is more characteristic of many situations in which HLA*IMP is supposed to be useful: it utilizes the whole set of available training data to impute HLA types into a cross-European validation panel – this is very comparable to the way that HLA type imputation was employed in recent studies [Sawcer et al., 2011]. In-depth analyses of the determinants of accuracy therefore focus on the second experiment. Accuracy (concordance / PPV, see page 3.2) is assessed by comparing imputed HLA types to classically typed HLA types at the haplotype level. Both experiments are based on data for the six classical loci *HLA-A*, *-B*, *-C*, *-DQA1*, *-DQB1* and *-DRB1*.

3.5.1 2/3 - 1/3 cross validation

In this experiment, 2/3 (random split) of the haplotypes in the GS are used to impute the HLA types of the remaining 1/3 (the data actually comes from a cross-validation experiment used to assess some of the accuracies presented in Sawcer et al. [2011], i.e. the SNP set was restricted to the SNPs available in this study). Table 3.3 summarizes the results for call thresholds of T = 0.00 and T = 0.70. In the non-thresholded scenario, accuracy at 4-digit resolution is ≥95% for all loci but *HLA-DRB1*, where it is at 90%. Setting a threshold improves accuracies in the 1 - 2% range, at the expense of a drop in call rate of 1 - 5%. Measured at 2-digit resolution, accuracies are ≥96% even without a call threshold. *HLA-DRB1*, apparently the most problematic gene at 4-digit resolution, is at 98%.

3.5.2 GSK validation

In this experiment, the full GS is used to impute HLA types into another dataset, which has been provided by GlaxoSmithKline and is therefore denoted as "GSK" (SNP genotyping is based on the Illumina 1M platform, HLA genotypes were derived by sequence based typing).[1].In this chapter, I restrict the dataset to samples of European ancestry (the last

[1]Again, the dataset preparation process took place in very close collaboration with Loukas Moutsianas

Threshold	Locus	# Validated	Call Rate	Accuracy	Call Rate (2-digit)	Accuracy (2-digit)
T= 0.00	*HLA-A*	816	1.00	0.95	1.00	0.96
	HLA-B	1009	1.00	0.95	1.00	0.97
	HLA-C	635	1.00	0.96	1.00	0.97
	HLA-DQA1	51	1.00	0.98	1.00	0.98
	HLA-DQB1	867	1.00	0.97	1.00	0.99
	HLA-DRB1	858	1.00	0.90	1.00	0.98
T = 0.70	*HLA-A*	816	0.98	0.97	0.98	0.98
	HLA-B	1009	0.98	0.96	0.98	0.98
	HLA-C	635	0.97	0.97	0.98	0.97
	HLA-DQA1	51	0.98	0.98	1.00	0.98
	HLA-DQB1	867	0.99	0.98	1.00	0.99
	HLA-DRB1	858	0.95	0.92	0.99	0.98

Table 3.3: Non-thresholded and thresholded cross-validation results for HLA*IMP (see Section 3.5.1): 2/3 of the GS are used to impute the remaining 1/3. Accuracy (PPV) is measured at 4-digit resolution, unless otherwise specified in the column headers. "# Validated" refers to the number of validated alleleles/haplotypes (pre-thresholding).

chapter will address HLA type imputation and differences in ethnicity) and denote it "GSK_EU", to ensure a good match between reference (GS) and imputation panel. After filtering for individuals of (self-declared) White ethnicity (European origin) and applying the same quality control measures as described for the GS (see Section 3.3), the dataset GSK_EU consists of 1084 individuals x 6056 SNPs, typed at a differing number of HLA loci (this includes *DRB* orthologs and *HLA-DPB1*, but the analysis presented in this chapter focuses on the six classical loci described above). Table 3.2 provides a summary of the GSK_EU dataset and of the contained allelic diversity. Notably, for each classical locus, GSK_EU contains more alleles than GS alone. Table 3.4 details the countries of origin of the samples, underlining the point that the GSK_EU dataset is considerably more diverse than the GS dataset.

Non-thresholded and thresholded results are presented in Table 3.5. In the non-thresholded scenario, accuracies at 4-digit resolution are substantially lower than in the 2/3 - 1/3 cross validation experiment: *HLA-A* and *HLA-DRB1* are at 80 and 81%, respectively. For the other loci, PPV ranges from 89% to 96%. Remarkably, the corresponding numbers measured at 2-digit resolution are much higher, most of them in the upper 90% range. Setting a call threshold markedly improves accuracies for *HLA-A* (91%) and *HLA-DRB1* (87%), and slightly improves accuracies for the other loci (improvement between 1 and 2%). This, at least at *HLA-A* and *HLA-DRB1*, comes at the expense of a drop in call rates to 86% and 85%, respectively. At 2-digit resolution, for all loci but *HLA-A*, call

GSK_EU		GSK_ALL		GSK_ALL	
Country	# Samples	Country	# Samples	Ethnicity	# Samples
United States	452	United States	531	White	1075
Poland	80	Peru	88	Hispanic	187
UK	73	Poland	80	Asian	153
Australia	60	UK	74	Black	36
Russian Federation	56	Australia	66		
France	47	Russian Federation	56		
Belgium	42	France	51		
Germany	36	Belgium	47		
Italy	31	Pakistan	41		
Czech Republic	30	Germany	37		
Canada	29	Italy	35		
Austria	22	Czech Republic	32		
Hungary	14	Canada	30		
Spain	14	Spain	29		
Greece	13	India	28		
Latvia	12	Austria	22		
Ukraine	9	Singapore	16		
New Zealand	7	Argentina	14		
Ireland	6	Chile	14		
Lithuania	6	Hong Kong	14		
Slovakia	6	Hungary	14		
South Africa	6	Mexico	14		
Romania	4	China	13		
Finland	3	Greece	13		
Sweden	3	Latvia	12		
Switzerland	3	Republic Of Korea	9		
Argentina	1	Ukraine	9		
Bulgaria	1	Lithuania	7		
Croatia	1	New Zealand	7		
Estonia	1	South Africa	7		
Portugal	1	Ireland	6		
Turkey	1	Slovakia	6		
		Malaysia	5		
		Japan	4		
		Romania	4		
		Switzerland	4		
		Finland	3		
		Sweden	3		
		Bulgaria	1		
		Croatia	1		
		Estonia	1		
		Portugal	1		
		Tunisia	1		
		Turkey	1		

Table 3.4: Country and ethnicity of the samples in the GSK_EU and GSK_ALL datasets.

rates and accuracies are ≥97%. At *HLA-A*, accuracy is at 95%, at a call rate of 91%.

Threshold	Locus	# Validated	Call Rate	Accuracy	Call Rate (2-digit)	Accuracy (2-digit)
T= 0.00	*HLA-A*	595	1.00	0.80	1.00	0.91
	HLA-B	2061	1.00	0.89	1.00	0.97
	HLA-C	647	1.00	0.96	1.00	0.99
	HLA-DQA1´	610	1.00	0.90	1.00	0.98
	HLA-DQB1	988	1.00	0.94	1.00	0.99
	HLA-DRB1	1762	1.00	0.81	1.00	0.97
T = 0.70	*HLA-A*	595	0.86	0.91	0.91	0.95
	HLA-B	2061	0.97	0.91	0.99	0.97
	HLA-C	647	0.99	0.97	0.99	0.99
	HLA-DQA1	610	0.97	0.92	0.99	0.99
	HLA-DQB1	988	0.99	0.94	1.00	0.99
	HLA-DRB1	1762	0.85	0.87	0.99	0.98

Table 3.5: Non-thresholded and thresholded GSK validation results for HLA*IMP (see Section 3.5.2): the complete GS is used to impute GSK_EU samples. Accuracy (PPV) is measured at 4-digit resolution, unless otherwise specified in the column headers. "# Validated" refers to the number of validated alleleles/haplotypes (pre-thresholding).

Examining the validation results on a per-locus basis is instructive. Figures 3.4 and 3.5 show the results for *HLA-B*, the most diverse gene in the HLA region, and *HLA-DRB1*, a gene with comparably low 4-digit imputation accuracy in both validation experiments.

Focusing on *HLA-B* first, it is clear that the majority of alleles is predicted reliably (green bars). This is true for the three alleles with the highest coverage in the reference panel. Then there is a variety of alleles which are never imputed correctly, because they do not appear in the training data (bright blue bars). Finally, we can observe that there are some well-covered alleles which exhibit substantial rates of error – 2705, for example. Many of these errors arise because of an incorrect classification at the 4-digit level, which disappears at the 2-digit level (data not shown).

The picture for *HLA-DRB1* is similar in some aspects and different in others. First, the problem that errors arise from alleles which are not present in the reference panel also affects *HLA-DRB1*. Note that we can reliably predict the top 3 best-covered alleles, but the general picture is quite different from *HLA-B*: many alleles either cannot be called or imputed reliably, even if they are well covered in the reference panel. Again, many of the errors at 4-digit resolution disappear at 2-digit resolution (data not shown).

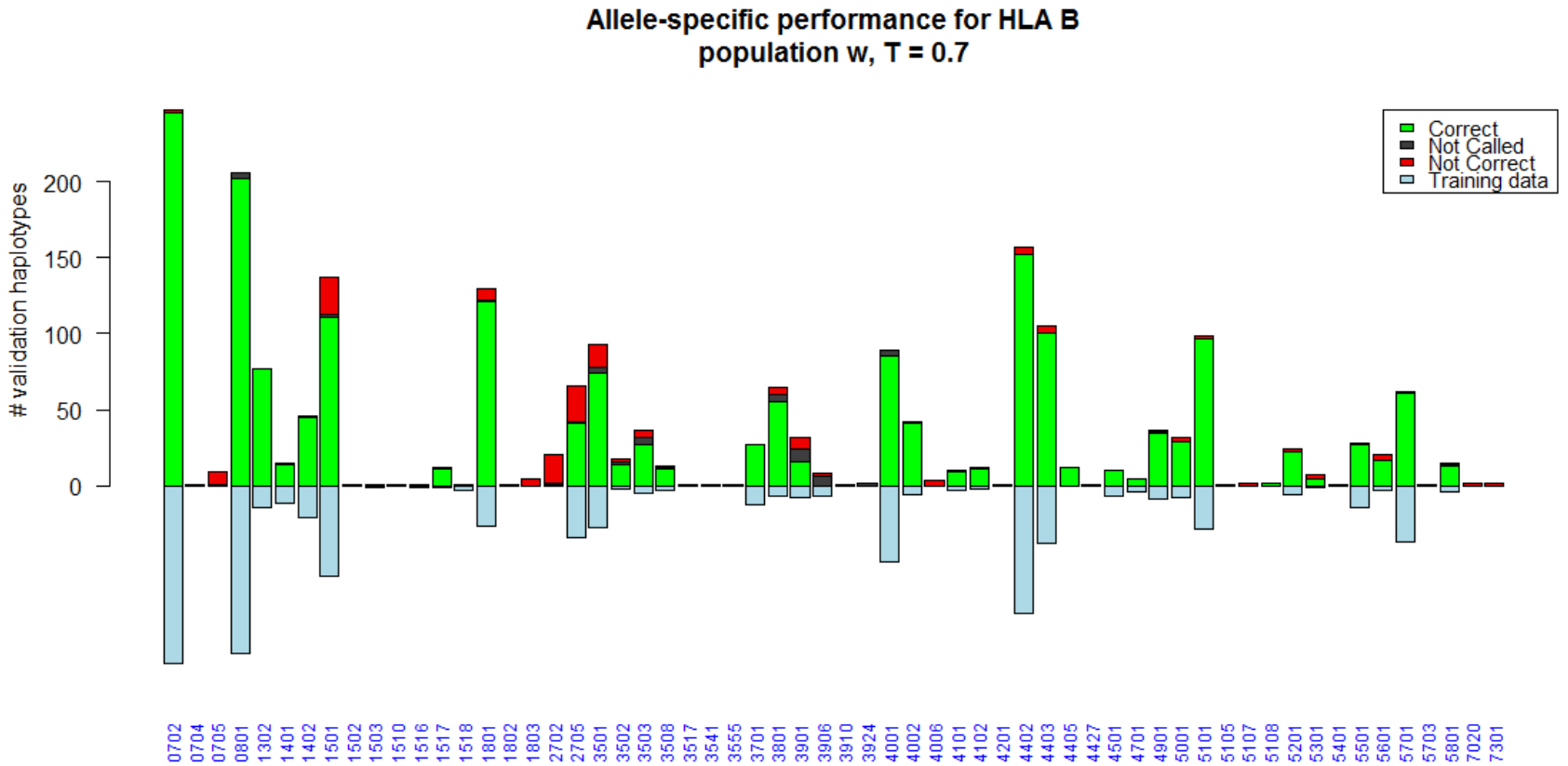

Figure 3.4: Per-allele analysis of HLA*IMP imputation accuracy for *HLA-B* in the GSK validation experiment at a call threshold of T = 0.7 (see Section 3.5.2). The x-axis represents the different HLA alleles in the validation panel. The downward blue bars indicate how often each allele appears in the reference panel (the GS dataset). Imputation success is indicated by the upward stack plots: green indicates correct imputations (per-haplotype "best guess" ML calls), red incorrect imputations, and black haplotypes which were not called. Note that there is a connection between how well an allele is represented in the reference panel and how well it is imputed (see text).

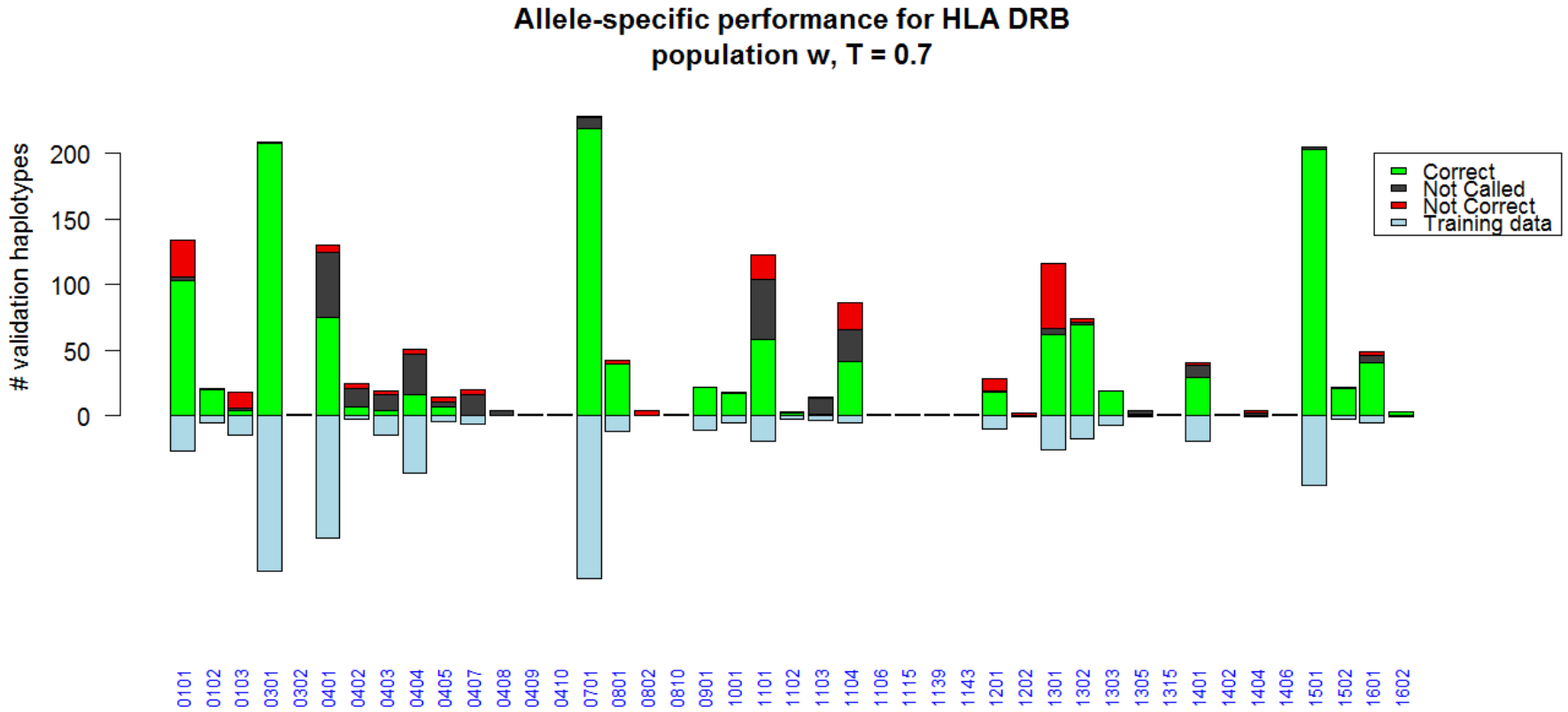

Figure 3.5: Per-allele analysis of HLA*IMP imputation accuracy for *HLA-DRB1* in the GSK validation experiment at a call threshold of T = 0.7 (see Section 3.5.2). The x-axis represents the different HLA alleles in the validation panel. The downward blue bars indicate how often each allele appears in the reference panel (the GS dataset). Imputation success is indicated by the upward stack plots: green indicates correct imputations (per-haplotype "best guess" ML calls), red incorrect imputations, and black haplotypes which were not called. Note that there is a connection between how well an allele is represented in the reference panel and how well it is imputed (see text).

3.5.3 Discussion

What conclusions can we draw from these results?

Firstly, note that the two validation experiments represent two extremes that most studies with individuals of European descent probably lie somewhere in between. The first experiment's imputation panel, the 1/3 GS panel, is very closely matched to the reference panel – as close as it can get, actually. The results from the first experiment therefore represent an upper boundary on accuracy and are probably a good proxy for disease-association studies with mainly British participants (in terms of the scenarios discussed in the introduction: Scenario 1. Strictly speaking, British and Central European samples are not from the *same* population, but PCA analyses presented below will reveal that there is not much population structure separating them – see Figure 5.1). The second experiment's imputation panel, the GSK_EU dataset, is much more diverse, with participants from Southern and Eastern as well as Central Europe. The results obtained here are probably closer to the lower accuracy boundary for high-quality imputation panels of general European descent and the present GS (in terms of the scenarios discussed in the introduction: in between Scenarios 1 and 2). The make-up of the GSK_EU dataset resembles panels used in many recent cross-European disease association studies [Sawcer et al., 2011], and the results obtained here are probably a good proxy for what accuracies to expect in pronouncedly cross-European studies.

Taken together, these results imply that employing a call threshold of $T = 0.70$ is generally a good idea – in closely matched panels, the loss in call rate is negligible, and in more diverse panels, the call threshold effectively identifies haplotypes which cannot be imputed reliably. Scenario 1 can now be addressed with a satisfactory degree of accuracy: we can expect to obtain accuracies and call rates at least as good as the ones presented for the GSK experiment, i.e. $\geq$90% for most loci.

Secondly, we can identify two important factors influencing imputation accuracy:

- most importantly, distance between imputation panel and reference panel, as obvious from the differences between the two experiments. We will return to this topic in

the last chapter.

- amount of available reference data: many allelic errors occur because the allele is not present in the reference panel, and the best-represented alleles are generally imputed reliably

As apparent from the observed differences between *HLA-B* and *HLA-DRB1*, there also seem to be locus-specific accuracy differences between different HLA genes, possibly independent of the amount of available training data. It is not clear why this is. Possible reasons include the complexity of haplotype structure around the loci (which may be different for class I and class II genes) and the quality of classical HLA typing data (the presence of *DRB* orthologs may, for example, make it more difficult to obtain a correct classical *HLA-DRB1* genotype).

Finally, how should we interpret the difficulties HLA*IMP sometimes experiences in differentiating between 4-digit alleles of the same 2-digit type group? The fact that these errors disappear at 2-digit resolution implies that the algorithm at least correctly identifies a signal of identity by descent, probably reflecting the evolutionary relatedness of alleles in the same 2-digit group. It also not clear that every single allele in the reference panel is assigned correctly; in fact, many classical HLA typing methods (for example allele-specific hybridization) are also prone to mistake alleles in the same 2-digit group for one another. It is quite possible that such systematic errors contribute to some of the errors observed here.

3.6 Conclusion

HLA*IMP, the integrated HLA type imputation frameworks, achieves accuracies and call rates $\geq$90% for most classical loci considered here when applied to individuals of European descent (haplotype validation). It employs a slightly modified version of the LDMhc model and is implemented in a highly efficient parallelized way, which enables the use of large reference panels. All of these factors contribute to the accuracies described here. By examining the results of two validation experiments, two important factors contributing

to imputation accuracy could be identified (see page 91): distance between reference and imputation panel and the amount of allele-specific training data.

By offering a user-friendly front-end program and a back-end imputation web-service, HLA type imputation is now available for biomedical researchers without expert knowledge in bioinformatics or statistics. As of November 2011, HLA*IMP has already been employed in several of high-profile disease association studies [Evans et al., 2011; Sawcer et al., 2011; Strange et al., 2010].

Note that HLA*IMP requires phased input data; maybe more importantly, it also requires the selected SNPs to be present in the imputation panel. Imputing missing SNP genotypes prior to imputation is a workaround. However, given that SNP selection is carried out in order to identify SNPs which can differentiate between different haplotypic backgrounds, it is not clear whether SNP genotype imputation is appropriate here.

In addition, the problem of population structure, as observed in the GSK_EU panel (in the form of a not perfectly matched imputation dataset), deserves further attention. In Chapter 5, I am going to examine how the model of HLA*IMP behaves when GS and GSK_EU are combined and then randomly split into well-matched, but diverse, reference and imputation panels. I will also address the question of whether it is possible to detect HLA genotyping errors in the reference panel (Chapter 4).

Chapter 4

HLA*IMP:02

In this chapter, I present the model and implementation of HLA*IMP:02, a novel HLA type imputation method. The development of HLA*IMP:02 was motivated by the insight that haplotype graph models should be able to capture the MHC's long-ranging haplotype features, and that they should be more tolerant of errors and heterogeneity in the reference panel. However, existing haplotype graph construction algorithms are not necessarily ideal for the purpose of HLA type imputation, as they do not allow for uncertainty in the current guess of haplotypes or incorporation of prior knowledge on the rich LD structure of the MHC. HLA*IMP:02 addresses these points and also allows for missing data in the set of estimated haplotypes; the graph-building algorithm I propose operates on a set of distributions over haplotypes instead of a set of deterministic haplotypes.

I will also present validation results for HLA*IMP:02; they suggest that HLA*IMP:02 very slightly outperforms HLA*IMP:01 (for clarity, HLA*IMP is denoted as HLA*IMP:01 in the following) in terms of accuracy when evaluated on a non-heterogeneous reference panel. I also demonstrate that HLA*IMP:02 can be used to impute structural variation (*DRB* orthologs). In the following Chapter 5, I evaluate the model's performance on heterogeneous reference panels and show that it indeed outperforms HLA*IMP:01 on heterogeneous reference panels, that it is highly tolerant of missing data and that it has some limited ability to detect errors in the reference panel.

4.1 The model

4.1.1 Overview

Before describing any algorithmic details, I will give a high-level overview of the HLA*IMP:02 model.

As discussed in the introduction, genotype imputation is based on using a reference panel with high marker density and a statistical model of population haplotype structure to impute missing markers in imputation panels with lower marker density. This basic approach can be implemented in a variety of ways [Marchini and Howie, 2010]. In HLA*IMP:02, the following structure is adopted:

1. A reference panel with SNP- and HLA-genotyped individuals is used to construct a statistical model of population haplotype structure, the haplotypes comprising HLA alleles and their surrounding SNP contexts.

2. Any data in the imputation panel (individuals' SNP genotypes, usually with no HLA typing available) is treated as if it had been generated from the model, and standard statistical techniques for dealing with missing data are applied to infer the genotype distribution at the untyped HLA loci.

The haplotype structure model I employ here belongs to the class of haplotype graph models (see Section 2.2.1). I informally recapitulate the essential characteristics of haplotype graph models: Haplotype graph models exhibit an acyclic structure of probabilistically connected states. Each state refers to a genetic position (e.g. SNP or HLA locus) and emits a symbol (nucleotide or HLA allele). By jumping between the states, more precisely: by following probabilistically defined *paths* through the model's states according to the chromosomal ordering of their associated genetic positions, it is possible to simulate haplotypes from the model.

It is well known that haplotype graph models can be thought of as leveled Hidden Markov Models, which leads to convenient computational properties and enables the application of

standard statistical techniques such as the forward-backward algorithm (see Section 2.2.1). Specifically, conditional on assuming that observed (haploid or diploid) genotype data has been generated by a haplotype graph model, for every possible single path (observed haploid data) or pair of paths (observed diploid data) through the model, the likelihood of this path / pair of paths having generated the observed data can be computed. This property immediately enables HLA genotype inference for individuals in the inference dataset: for each individual, we treat the (untyped) HLA alleles as missing data, compute the probability of each pair of paths through the population haplotype graph model, conditional on the individual's SNP genotypes, and marginalize over the HLA allele information carried by each pair of paths.

To deal with the particularities of the MHC region, I introduce a new probabilistic haplotype graph construction algorithm. Specifically, compared to other approaches [Browning and Yu, 2009; Browning and Browning, 2007], I introduce two additional parameters: one modeling uncertainty in the set of haplotype estimates (enabling borrowing of information between similar haplotypes, thus potentially capturing genotyping error and local phase ambiguities), and the other one allowing for "localization" of the graph, i.e. tailoring the graph to effectively capture linkage with classical HLA alleles. The algorithm is also extended in a way that allows it to deal with missing data. Note that HLA*IMP:02, in contrast to BEAGLE, separates inference from reference panel-based model construction (one graph can be applied to many inference panels). To the extent that HLA*IMP:02 does not allow the data in the imputation panel to influence haplotype choice in the reference panel, it is more similar in spirit to IMPUTE2 [Howie et al., 2009] than BEAGLE [Browning and Browning, 2007].

The model construction algorithm operates in an *iterative* manner: it tries to gradually improve the fit of the haplotype graph model to observed population data. The algorithm is similar to that of BEAGLE [Browning and Browning, 2007] and is the most complex component of HLA*IMP:02 (the following description assumes that there is no prior haplotype information for the individuals in the reference panel – if this information is available, it can be integrated. See page 104 for details).

1. The process is based on a reference panel of SNP- and HLA-genotyped individuals. For each individual present in the reference panel, HLA*IMP:02 randomly generates possible haplotypes, consistent with the individual's genotypes.

2. In the first iteration, these haplotypes are used to build a haplotype graph model (according to the algorithm presented below). Informally, this means trying to learn the edge structure (topology) and transition probabilities of the graph from the haplotypes, involving a series of comparisons and a complex merging procedure.

3. As noted above, haplotype graph models allow for inference of likely paths through the model, conditional on observed genotypes. This property is used to obtain a new set of haplotypes for the individuals in the reference panel, this time conditional on the observed genotypes and the haplotype graph of the first iteration. The new set of haplotypes is used to build the haplotype graph for the next iteration.

4. This procedure is repeated a specified number of times, and the haplotype graph of the last iteration is stored as the final graph for HLA genotype inference.

Note that, although HLA*IMP:02 will produce estimates of SNP haplotypes, correct SNP phasing is not the aim here: validation will exclusively focus on whether the allelic state of the HLA genes can be imputed accurately.

4.1.2 Probabilistic haplotype graph construction

Haplotype estimate uncertainty and missing data

Conceptually, the algorithm presented here for building haplotype graphs is a probabilistic adaptation of the works of Ron et al. [1998] and Browning and Browning [2007]. For convenience, wherever possible, notation follows Section 2.2.1.

Suppose that we want to build a haplotype graph, based on a set H of haplotypes of length T. So far, we have always assumed that we know the elements in H with certainty. In the context of genetics, it is easy to see why this assumption may be violated. For example, a SNP genotyping error may lead to a haplotype being present in H which does not really

exist in the population. Some SNP genotypes may also be missing, and we would like to have a model which can either directly deal with missing data or at least allow for incorporating some of the imputation-associated uncertainty, if missing data gets imputed at some stage. So far, however, we have always assumed that the elements in H contain no missing data. I will first describe an error / uncertainty model and then how to deal with missing data in H.

As in the basic case, suppose that we have the most general possible haplotype graph topology (for a given level-specific model alphabet and haplotype length T). To introduce uncertainty in the graph-building process, we now assume that an error process may have modified the set of observed haplotypes. We assume that this error process acts independently on each character position of the haplotypes in H and that, if an error occurs, a new observed value is drawn from a uniform distribution over possible alternative alleles at the affected position (this could, if desired, be easily generalized to more complex error models). If we observe string s_1 of length T, the likelihood that string s_2 is the true underlying string is

$$\prod_{l=1..T} [I_{s_1(l)==s_2(l)} \times (1 - m_B) + (1 - I_{s_1(l)==s_2(l)}) \times \frac{m_B}{|A_{l-1}| - 1}] ,$$

where we define $I_{s_1(i)==s_2(i)}$ to be 1 if the i-th symbol of s_1 is equal to the i-th symbol of s_2 and 0 otherwise. A_l is the level-specific model alphabet for position l (in haplotype graph indexing). For simplicity, although m may capture other effects than error, I refer to m_B as the *graph building error* probability. m_B is conceptually (though not formally) related to mutation parameters of population genetics models, e.g. the Li&Stephens approximation of the coalescent [Li and Stephens, 2003].

To allow for missing data in the set H, we define a new symbol E_M which indicates missingness. E_M may not appear in any position's model alphabet (and does not become a member of these sets). We redefine H to contain strings of symbols of length $L > 0$, where at any given position the symbols come from the union of the position's model alphabet and E_M. E_M is *not* allowed as an edge emission symbol. To integrate the concept of missingness with the error model, we define equal probabilities for each model alphabet

symbol to be observed as missing data (and we say that this approach is "agnostic" – an alternative would be to use observed site-specific marginal genotype frequencies).

I now describe the novel graph building algorithm to take into account missing data and uncertainty according to the error model we have defined. Conceptually, the algorithm is based on attaching haplotypes to the nodes of the most general graph topology, but in a probabilistic way.

For each vertex, we introduce a list of probability-weighted potentially attached haplotypes. More formally, for each vertex $v \in V$, $\mathrm{P}_H(v, h)$ shall denote the probability of h being attached to v. At each level of the graph, the sum of attachment probabilities has to be 1 for each haplotype. All haplotypes are attached to the start vertex v_0 with probability 1 by defining $\mathrm{P}_H(v_0, h) := 1$ for all $h \in H$; they are then distributed along the graph according to our error model. That is, if haplotype h is attached to v_1 at level $l(v_1)$ with probability a, and if the next observed haplotype symbol (at index $l(v_1) + 1$ within the corresponding string) is $s \neq E_M$, we have the following attachment probabilities for the children v_2 of v_1 at level $l(v_1) + 1$: if the edge (v_1, v_2) carries the attached symbol s, the attachment probability of h at v_2 is $a \times (1 - m_B)$, i.e. we define $\mathrm{P}_H(v_2, h) := \mathrm{P}_H(v_1, h) \times (1 - m_B)$. Otherwise, the attachment probability is $a \times \frac{m_B}{A_{l(v_1)}-1}$, and we define $\mathrm{P}_H(v_2, h) := \mathrm{P}_H(v_1, h) \times \frac{m_B}{A_{l(v_1)}-1}$. If $s = E_M$, we attach h in an agnostic manner, i.e. we define $\mathrm{P}_H(v_2, h) := \mathrm{P}_H(v_1, h) \times \frac{1}{|A_{l(v_1)}|}$.

Returning to the most general possible haplotype graph topology for a given haplotype length T, we see that each node in this graph now carries (possibly small) haplotype attachment probabilities (see Figure 4.1). The motivation for merging nodes as discussed in Section 2.2.1 – increasing computational efficiency and approximating a possible generating haplotype graph while not changing haplotype frequencies much – holds for the uncertainty-aware case described here. In order to compare nodes, we consider the conditional suffix distributions.

For notational convenience, let $H(v)$ denote the set of haplotypes attached to v with attachment probability $\mathrm{P}_H(v, h) > 0$. I now redefine the count functions, the subscript

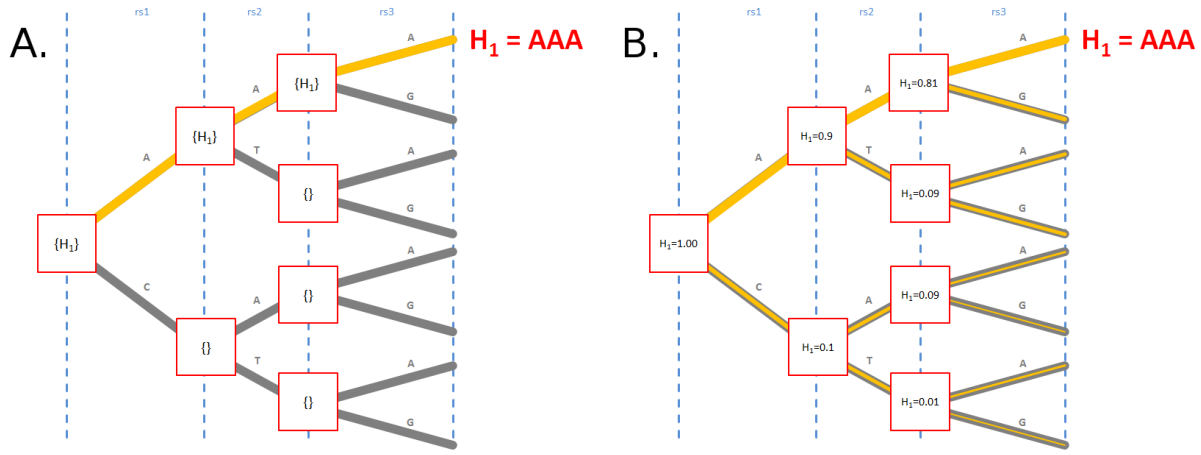

Figure 4.1: A: a non-probabilistic haplotype graph construction algorithm. Each haplotype in the set H follows one defined path (orange) through the graph's possible topology (orange and gray branches), here depicted for $H_1 = \text{AAA}$. Each node (red squares) carries a list of attached haplotypes. B: the probabilistic haplotype graph construction algorithm presented in this chapter. Each haplotype in the set H induces a probability distribution over possible paths through the graph, here pictured as orange lines. The width of the lines indicates how probable a path is according to the path probability distribution (not drawn to scale). At each node, the path follows the edge carrying the haplotype's next symbol with probability $1 - m_B$, and the remaining probability mass is split over the remaining available edges. Each node carries a list of attached haplotypes with the respective attachment probability. The figure is based on a path distribution for "AAA", with the graph-building error probability m_B set to 0.1.

'U' emphasizing that we allow for uncertainty:

$$\text{count}_U(v,'') := \sum_{h \in H(v)} \text{P}_H(v,h)$$

$$\text{count}_U(v,x) := \sum_{h \in H(v)} (\text{P}_H(v,h) \times \prod_{i=[l(v)+1]..[l(v)+L(x)]} [I_{h(i)==x(i)} \times (1-m) + (1 - I_{h(i)==x(i)}) \times \frac{m_B}{A_{(i-1)} - 1}]),$$

where L(x) denotes the length of suffix x. For notational convenience, I have assumed a fixed m_B here for all loci, but it is easy to see that this is not necessary. It is easy to see that the two count functions represent expected values of the number of attached haplotypes (continuing with a particular suffix).

Conditional on being at vertex v, we define the probability $\mathrm{P}(x|v)$ of continuing with suffix x as

$$\mathrm{P}_U(x|v) := \mathrm{count}_U(v,x)/\mathrm{count}_U(v,'')\,.$$

The similar function compares two nodes' conditional suffix distributions

$$\mathrm{similar}_U(v_1,v_2) := \max_{x \in S_{v_1,v_2}} |\mathrm{P}_U(v_1,x) - \mathrm{P}_U(v_2,x)|\,,$$

(and I present a modified version of this function in the next subsection).

Two nodes are merged if the similar function returns a value below a certain threshold:

$$\mathrm{similar}_U(v_1,v_2) < \epsilon\,.$$

We apply similar to all pairs of nodes at all levels to identify pairs of nodes that can be merged, following the following algorithm. For all levels $l = 0...[T-1]$, compute the similarity measure $\mathrm{similar}_U(v_j,v_k)$ for all pairs of nodes (v_j,v_k) at level l.

If $\mathrm{similar}_U(v_j,v_k) < \epsilon$,

1. create a new node v_n at the same level as v_k and v_j.
2. redirect all incoming edges of v_k and v_j to v_n, and for all $h \in \{H(v_k) \cup H(v_j)\}$, set $\mathrm{P}_H(v_n,h) := \mathrm{P}_H(v_k,h) + \mathrm{P}_H(v_j,h)$.
3. attach all outgoing edges of v_k to v_j to v_n, and delete v_k and v_j.
4. note that step 2 will result in a structure violating the haplotype graph assumptions, as it will result in two edges (v_n,v_x) , (v_n,v_y) with the same attached symbol. Merge v_x and v_y as described for all such cases (i.e. recursively from step 1, if necessary), and delete one of the two resulting (v_n,v_x) edges.
5. finally, update $\mathrm{P}_U(e|v)$ for all modified nodes.

See Figure 4.2 for an illustration of the merging process.

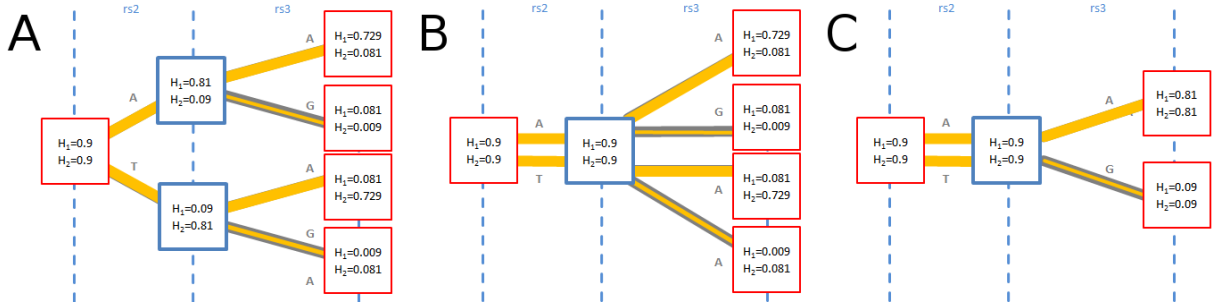

Figure 4.2: The essential steps of merging nodes in the probabilistic framework described in this chapter. A: two haplotypes (AAA and ATA) have been attached to the topology shown in Figure 4.1 (the graph's first level is not shown) with $m_B = 0.1$. The conditional suffix distributions of two nodes (pictured as blue squares) are identical and the nodes will be merged. B: all outgoing edges from the two nodes have been attached to one newly created joint node (blue square). The resulting structure is no haplotype graph, because two edges emanating from the new node carry the same symbols as two other edges emanating from the same node. C: The nodes that the conflicting edges lead two are recursively merged, resulting in a haplotype graph structure.

Referring to the fact that haplotypes are probabilistically instead of deterministically attached here, I refer to the algorithm as a probabilistic haplotype graph building algorithm.

Localization

Localization aims at improving imputation performance for specified loci of interest. Consider the following example to see why localization can be sensible. Suppose that two haplotypes from H are attached to v_1: '00A' and '11A' (the allele identifiers are arbitrary). Suppose further that v_2 has also two attached haplotypes, '00B' and '11B'. If we compare the conditional suffix distributions, we find no difference for suffixes of length up to 2 ('00' and '11' are attached to both nodes, at equal frequencies). For suffixes of length 3 (i.e. '00A', '00B', '11A', '11B') and a small m, we find that the maximum difference is just below 0.5 (because none of the 3-character suffixes present in one node is present in the other one). Depending on our choice of ϵ, we may decide to merge the two nodes. The problem here is that the nodes actually exhibit quite different patterns of LD to the third position – the maximum conditional probability difference is almost 1. The localization element extends the function similar to take into account such situations for a set S_L of

predefined loci and could prevent merging the two nodes.

I go on to introduce the localization parameter S_L, which is a (possibly empty) set of level indices. Define the indicator function $1_{\mathrm{L=l}}(h, a)$ to be 1 if haplotype h carries allele a at position l, and 0 otherwise. We define the probability of observing a particular allele a at level l, conditional on having reached vertex v in the haplotype graph, as

$$\mathrm{P}_{\mathrm{LOCALIZE}}(v, a, l) := \frac{\sum_{h \in H(v)} \mathrm{P}_H(v,h) \times (I_{\mathrm{L=l+1}}(h,a) \times (1-m) + (1 - I_{\mathrm{L=l+1}}(h,a)) \times \frac{m}{|A_l|-1})}{\mathrm{count}_U(v,'')}$$

(note that the +1 in $L = l+1$ comes from the fact that the first level of a haplotype graph has index 0, but the first position of a string has position 1).

We note that this conditional probability integrates over the uncertainty in the intermediate SNP genotypes (see Figure 4.3) and redefine the similar function to include all loci specified in S_L:

$$\mathrm{similar}_U(v_1, v_2) := \max(m_1, m_2)$$

$$m_1 := \max_{x \in S} |\mathrm{P}_U(v_1, x) - \mathrm{P}_U(v_2, x)|$$

$$m_2 := \max_{l \in S_L} \left[\max_{a \in A(\mathrm{l})} |\mathrm{P}_{\mathrm{LOCALIZE}}(v_1, a, l) - \mathrm{P}_{\mathrm{LOCALIZE}}(v_2, a, l)| \right]$$

The localization element has the effect that two nodes with differing linkage patterns to a level specified in S_L will not be merged, irregardless of the similarity of the SNP haplotypes leading to the elements in S_L.

4.1.3 Computational efficiency

The algorithm I have described to build localized haplotype graphs from an uncertain set of haplotypes requires substantial computational resources: to calculate the conditional suffix distributions for each vertex, it is necessary to sum over all attached haplotypes

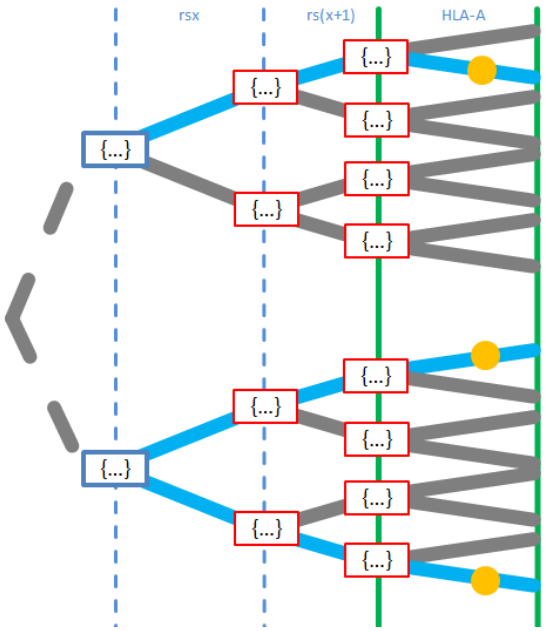

Figure 4.3: Localization at the example of an HLA locus. When comparing the conditional HLA allele probabilities for two nodes (blue squares) for a particular *HLA-A* allele (marked with an orange circle in the graph), the probabilities of all paths leading to this allele are added up (separately for each node). Note that the two blue paths for the lower node would count as two distinct suffixes without localization.

with attachment probability > 0. Haplotype attachment distributions of single haplotypes are typically (depending on the uncertainty model and the number of alleles at the involved loci) very skewed: a few vertices at any level usually account for the majority of available probability. Therefore, a threshold F is introduced: if $\mathrm{P}_H(v,h) < F$, $\mathrm{P}_H(v,h)$ is set to 0 and the removed probability is spread mass proportionally over all nodes with $\mathrm{P}_H(v,h) \geq F$. This has the effect of removing paths (potential haplotypes) with low attached probabilities from the graph. Also, when computing $\mathrm{similar}_U(v_1,v_2)$ for two nodes, only such $x \in S$ that are present in at least one of the haplotypes attached to v_1 or v_2 are evaluated – suffixes x merely induced by error processes on both vertices will carry smaller probabilities than the original strings and therefore lead to a smaller absolute difference in probability.

Iterative Refinement

As outlined in the introduction, the iterative algorithm for constructing haploid haplotype graphs from diploid genotype data presented in Browning and Browning [2007] is adopted here: the aim is to iteratively refine the set H. Initially, the set H is filled with R random haplotypes for each individual which are compatible with the observed genotypes (in contrast to BEAGLE, missing genotype data is preserved as missing data in the initial set of haplotypes). The following algorithm is then iterated a predefined number of times (usually 12):

- build a haplotype graph G for H (based on the algorithm presented in the previous subsections).
- construct the diploid and haploid HMMs induced by G (as described in Section 2.2.1).
- re-populate H with R haplotype samples for each individual from the diploid HMM, conditional on the observed genotypes. Specifically, this means that the diploid HMM is fitted to each individual in turn and samples from the path pair probability distribution are taken. By concatenating the symbols associated with the edges each path traverses, we obtain new haplotypes.
- If haplotype phase is available for some individuals (from deterministically phased HapMap samples, say), the haploid haplotype HMM, induced by G, can be used to obtain new haplotype samples for these individuals in the previous step.

The emission probabilities for the haploid HMM (step 2) are parameterized in accordance with the haplotype uncertainty error model (and the diploid emission probabilities follow as described in Section 2.2.1). Specifically, conditional on not emitting E_M, the state relating to edge e in the haplotype graph will emit the emission symbol of e with probability $1-m_S$, and uniformly choose from one of the other members of the level-specific model alphabet with probability m_S. All states have equal emission probabilities for E_M. m_S is reduced

for HLA loci, in order to preserve rare HLA alleles in the set H as the graph is iteratively refined. We refer to m_S as the *sampling error* probability.

Following Browning and Browning [2007] I define the threshold for merging nodes as

$$\epsilon := d \times \sqrt{R} \times (\text{count}_U(v_1,'')^{-1} + \text{count}_U(v_2,'')^{-1})^{1/2}\,,$$

where d is a scale parameter (usually 0.8 here, determined by initial experiments) and R is the number of haplotype samples from each individual. Note that more refined ways to determine node similarity were considered: the count functions could, for example, return probability distributions, and we could then define $\text{P}_U(x|v)$ in a hierarchical way and let similar compare two hierarchies of distributions (and adapt ϵ accordingly). However, this route is not taken in the remainder of this thesis: the model as currently specified yields satisfactory results, and the proposed extensions would probably add to the (already substantial) computational demands of the model.

4.1.4 Integrating HLA information

HLA loci are treated as multi-allelic SNPs, i.e. the observed HLA types are part of the haplotype strings H and appear as edges in the haplotype graph. Their chromosomal position is taken to be the mean of the beginning and end of the corresponding gene, which leads to an unambiguous positioning with respect to surrounding SNPs. SNP- and HLA-genotyped individuals can be used as input for the "iterative refinement" algorithm, without prior phasing. Only individuals with at least one genotyped 4-digit HLA allele are used for constructing the haplotype graph model, and 4- and 2-digit alleles are treated in the same way, i.e. as unrelated, separate entities (initial experiments, in which 2-digit alleles were modeled as unions of 4-digit alleles, were not successful – data not shown).

To optimize imputation performance at the classical HLA loci, the graph building algorithm is localized for all classical HLA loci except *HLA-B* and *HLA-DRB1* (see below).

Note that the results from step 3 of the model building algorithm can be used to quantitatively assess whether a lab-based HLA typing result in the reference dataset is consistent

with the graph or not; the posterior probabilities follow from summing over the haplotype samples. To minimize the impact of mis-typed HLA alleles in the reference panel, after a specific number of graph-building and sampling iterations (usually 8), the number of sampled haplotypes for a specific individual is weighted by the internally estimated probability that the individual's lab-based HLA type is consistent with the graph.

4.1.5 HLA type inference

To carry out HLA type inference, a locus-specific haplotype graph is constructed from the specified reference panel (or loaded, if the same reference panel was used before and the graph has been saved), based on 300-SNP-windows each side of the locus (for example, ranging from 29807232 to 30202819 for *HLA-A* in the GSK validation experiment). The choice of 600 SNPs is motivated by initial experiments that demonstrated good performance at substantial computational gains; it is, however, possible that using more SNPs would lead to slightly increased accuracies (this has not been investigated here and is a possible direction for future research). The resulting diploid HMM is then applied to the individuals in the imputation dataset. HLA type inference can be carried out by sampling haplotypes from the diploid HMM, conditional on the observed genotypes for each individual in the inference dataset and the haplotype graph. This leads to posterior distributions over possible pairs of HLA types that can be processed in an uncertainty-aware way or thresholded. To call alleles, first the most likely single allele for each individual is determined, and then the most likely second allele, conditional on the individual's first allele. The marginal probability to observe the first allele (i.e. summed over all samples from the haplotype pair distribution) is used as quality score for the first allele and the joint probability for the first and the second allele is used as quality score for the second allele.

4.1.6 Discussion

I have presented a probabilistic generalization of existing haplotype graph construction algorithms [Browning and Browning, 2007] that introduces two additional parameters and

can deal with missing data. m_B is a haplotype estimate uncertainty parameter and S_L is the set of localization loci that can be used to adapt graph construction to complex patterns of LD (graph building error m_B and graph sampling error m_S are usually both set to m. Note that Browning and Yu [2009] allow for graph sampling error, but not for graph building error). By setting m to 0, S_L to the empty set, by ignoring missing data, and by combining the inference panel with the reference panel, one obtains the BEAGLE model.

I briefly discuss some properties of the generalized model:

- The error model I have introduced leads to a relative decline of the importance of long-range haplotype differences in terms of collapsing nodes: $|P_U(v_1, x) - P_U(v_2, x)|$ is decreased for x with large differences. This depends on d, the scaling parameter in the collapsing criterion, and m_B, the graph building error probability.

- The generalized model is potentially useful in other applications than the one considered here. For example, if a haplotype graph is to be constructed for a set of experimentally determined haplotypes (from single chromosome sequencing, say), the uncertainty model for the graph-building step I have introduced can be used to model read errors.

- The described algorithm can deal with missing data in the set of haplotypes in a straightforward way by defining a probability distribution on missing characters, e.g. a uniform distribution. This property allows us not having to guess genotypes for the first iteration of graph-building. Although the algorithm as described here imputes missing genotypes in the reference panel during the first sampling process, the missing data status could as well be preserved in the sampled haplotypes and could be carried over to later stages. As the reference panels which are used here are consistently typed on dense sets of markers, this route is not pursued here. However, under other circumstances, for example when SNP coverage in the reference panel varies strongly, not imputing missing SNP data may turn out to be beneficial [Howie et al., 2009].

- Treating HLA alleles as multiallelic SNPs leads to a couple of useful properties in learning and inference settings. The graph itself can reflect patterns of long-range linkage disequilibrium between HLA alleles – HLA and SNP genotypes are used to infer the graph structure in a combined manner, and there is no requirement that all individuals be typed at the same set of HLA loci. Consider, for example, an inference dataset with *HLA-DRB1*-typed individuals, but lacking information for *HLA-DQB1*. Providing the *HLA-DRB1* genotypes as well as the SNP genotypes enables the model to use partial HLA type information in inferring missing bits of the complete HLA type (depending on the particular structure of the graph used for inference, of course).

- Sampling new haplotype pairs from the diploid HMM, conditional on observed genotypes, automatically leads to an internal inconsistency correction by allowing for graph sampling error – the probability that a path for an individual does not traverse the nodes which correspond to the individual's classical HLA type is low, but not 0.

Parameter inference

Initial experiments have indicated that setting $m = 0.002$ (graph building and sampling error probabilities) and $d = 0.8$ (scaling parameter for similar function) leads to high-accuracy HLA type imputations from the model described here.

Choosing optimal parameters for building haplotype graphs and for inference is an important direction for further research. Although standard statistical techniques like ML and MCMC could be applied in theory, the computational costs to do so seem prohibitive at the moment. In the context of this thesis, the main aim is statistical HLA type imputation, and the appropriateness of model and parameterization is measured by the validation experiments presented later.

In order to justify the introduction of additional parameters, I have repeated some of the experiments presented in Leslie et al. [2008] and Dilthey et al. [2011] (presented here in Section 3.2). HapMap CEU data was used as a reference panel to impute HLA types into a subset of the BC58 (all data exactly as described in our earlier papers and in

Section 3.2). The results are summarized in Table 4.1. The row "HLA*IMP:02" refers to the full model as described here. The other rows refer to the seven other possible configurations, activating and deactivating parameters in turn. The full model, with localization deactivated at *HLA-B* and *HLA-DRB1*, yields good results. It is currently not clear why localization only seems to be beneficial at *HLA-A* and *HLA-DQB1* – problems with classical HLA typing at *HLA-B* and *HLA-DRB1* in either dataset may play a role here. Until further investigation, I recommend to deactivate localization for *HLA-B* and *HLA-DRB1*.

For the interested reader, Table 4.1 also presents accuracies for the BEAGLE model – which are lower than HLA*IMP:02 accuracies at 3 of 4 loci. As suggested (Sharon R. Browning, *personal communication*), the number of iterations for BEAGLE was increased. The number of iterations influenced the performance of BEAGLE, but in no evaluated scenario did the observed pattern qualitatively change (i.e. good results for BEAGLE at *HLA-DQB1*, but worse results than HLA*IMP:02 at the other loci).

Model/Locus	*HLA-A*	*HLA-B*	*HLA-DQB1*	*HLA-DRB1*
HLA*IMP:02	0.92	0.78	0.84	0.70
$m_B = 0$, $m_S = 0$, $L = 0$	0.69	0.55	0.53	0.46
$m_B = 0$, $m_S = 0$, $L = 1$	0.82	0.56	0.58	0.50
$m_B = 0$, $m_S = 1$, $L = 0$	0.71	0.60	0.57	0.45
$m_B = 0$, $m_S = 1$, $L = 1$	0.75	0.62	0.53	0.48
$m_B = 1$, $m_S = 0$, $L = 0$	0.92	0.76	0.83	0.75
$m_B = 1$, $m_S = 0$, $L = 1$	0.91	0.77	0.84	0.73
$m_B = 1$, $m_S = 1$, $L = 0$	0.90	0.83	0.83	0.74
BEAGLE	0.90	0.48	0.85	0.65

Table 4.1: Accuracies (PPV) for the HapMap-based BC58 validation, as described in Leslie et al. [2008] and Section 3.2. No call threshold is employed. To accommodate for the smaller size of the reference panel, $d = 0.3$ here. The row "HLA*IMP:02" refers to the full model with error parameters $m_B \neq 0$, $m_S \neq 0$ and localization activated. For the following seven rows, the column "Model" specifies which parameters are activated (1) or deactivated (0): m_B is the graph-building error parameter, m_S is the graph-sampling error parameter, L stands for localization. For the interested reader, the final row "BEAGLE" presents the accuracies the BEAGLE model [Browning and Browning, 2009] achieves in this experimental setting. As recommended by Sharon R. Browning (*personal communication*), I experimented with the number of iterations, and the model indeed seemed to experience convergence problems – the obtained overall results for BEAGLE, however, were always lower than those for HLA*IMP:02 (with the exception of *DQB1*, as shown here).

4.2 Availability

HLA*IMP:02 is made available to the biomedical research community using the HLA*IMP:01 web interface, including the specifically developed security concept – see Section 3.4 for details.

4.3 Validation

In the following, I present validation results for HLA*IMP:02, based on the same experimental setting as the validation of HLA*IMP:01 (see Section 3.5). I focus on 4-digit validation here – from the previous experiments, it is already clear that high accuracies at 4-digit resolution lead to good (in fact, slightly higher) accuracies when measured at 2-digit resolution.

4.3.1 Genotype validation

By sampling from a diploid HMM, HLA*IMP:02 produces *genotype* probabilities – that is, the haplotype validation methodology for HLA*IMP:01, presented in the previous chapter, cannot be applied to HLA*IMP:02-derived data.

All validation figures presented in this chapter are therefore based on genotype validation (see Section 3.2, page 72). Accuracies at the per-locus level are specified as concordance (PPV).

Note that genotype validation and haplotype validation do not fundamentally differ in what they measure – if anything, genotype validation can be expected to yield slightly higher accuracies, because haplotype switching errors are being accounted for. However, the set of validated individuals is not exactly identical between the two approaches, so that the figures presented here and in Section 3.5 should not be expected to be in perfect agreement.

4.3.2 2/3 - 1/3 cross validation

In this experiment, 2/3 (random split) of the haplotypes in the GS are used to impute the HLA types of the remaining 1/3 (note that this split is not identical to the one presented for HLA*IMP:01, and that the full set of SNPs is used here to build the graph). Table 4.2 summarizes the results. When no call threshold is applied, accuracy at 4-digit resolution is $\geq$95% for all loci but *HLA-DRB1*, where it is at 89%. A call threshold of T = 0.7 increases all accuracies to $\geq$94%, at call rates $\geq$ 92% for all loci. In comparison to HLA*IMP:01 (haplotype validation, as presented in Section 3.5), the un-thresholded numbers presented here are nearly identical – only accuracy for *HLA-DRB1* is 1% lower, but the subset of validated individuals is slightly different. The thresholded results look more different, but this is mainly due to the fact the the thresholds were not matched to obtain the same call rate.

Threshold	Locus	# Validated	Call Rate	Accuracy
T= 0.00	*HLA-A*	566	1.00	0.96
	HLA-B	1034	1.00	0.95
	HLA-C	582	1.00	0.96
	HLA-DQA1	46	1.00	0.96
	HLA-DQB1	728	1.00	0.98
	HLA-DRB1	734	1.00	0.89
T = 0.70	*HLA-A*	566	0.94	0.98
	HLA-B	1034	0.96	0.98
	HLA-C	582	0.98	0.97
	HLA-DQA1	46	0.98	0.96
	HLA-DQB1	728	0.99	0.98
	HLA-DRB1	734	0.92	0.94

Table 4.2: Non-thresholded and thresholded cross-validation results for HLA*IMP:02 (see Section 4.3.2): 2/3 of the GS are used to impute the remaining 1/3. Accuracy (PPV) is measured at 4-digit resolution. "# Validated" refers to the number of validated alleleles (pre-thresholding).

4.3.3 GSK validation

In order to realistically evaluate of the performance of HLA*IMP:02, the GSK_EU validation experiment, as presented in Section 3.5.2, was repeated as well. The Golden Set (see Section 3.3) was used to construct haplotype graphs for *HLA-A*, *HLA-B*, *HLA-*

C, *HLA-DQA1*, *HLA-DQB1* and *HLA-DRB1*, and imputation was carried out into the GSK_EU dataset (Section 3.5.2). The set of available SNPs in the validation datasets (for HLA*IMP:01 and HLA*IMP:02) was restricted to those that were available in a recent disease association study on multiple sclerosis [Sawcer et al., 2011]. Data for HLA*IMP:01 are presented at the genotype level and based on a slightly smaller validation SNP set than in the previous chapter; they are therefore not directly comparable with the results from Section 3.5.2.

As discussed in Chapter 3, note that this scenario reflects the accuracy that can be expected in cross-European imputation studies with a not perfectly matched imputation panel.

Table 4.3 summarizes the results in a comparative manner. Using no call threshold, average accuracy (PPV) over six loci is 93% for HLA*IMP:02 and 92% for HLA*IMP:01. Setting a call threshold of T = 0.7 for HLA*IMP:01 and a call threshold for HLA*IMP:02 that corresponds to the call rates obtained for HLA*IMP:01 at T = 0.7, the achieved accuracies increase slightly: average accuracy for HLA*IMP:02 is 94% at an average call rate of 97%. HLA*IMP:01 achieves an average accuracy of 93% at the same average call rate.

Threshold	Locus	# Validated	HLA*IMP:02			HLA:IMP:01		
			Call Rate	Accuracy	T	Call Rate	Accuracy	T
T= 0.00	*HLA-A*	574	1.00	0.96		1.00	0.90	
	HLA-B	2002	1.00	0.90		1.00	0.93	
	HLA-C	596	1.00	0.96		1.00	0.96	
	HLA-DQA1	446	1.00	0.87		1.00	0.87	
	HLA-DQB1	758	1.00	0.98		1.00	0.97	
	HLA-DRB1	1730	1.00	0.88		1.00	0.89	
T = Matched	*HLA-A*	574	0.96	0.96	0.55	0.94	0.91	0.700
	HLA-B	2002	0.98	0.92	0.40	0.98	0.94	0.700
	HLA-C	596	0.99	0.96	0.60	0.99	0.97	0.700
	HLA-DQA1	446	0.99	0.88	0.40	0.99	0.88	0.700
	HLA-DQB1	758	0.99	0.98	0.60	0.99	0.97	0.700
	HLA-DRB1	1730	0.90	0.93	0.60	0.90	0.93	0.700

Table 4.3: Non-thresholded and thresholded GSK validation results for HLA*IMP:02 and HLA*IMP:01 (see Section 4.3.3): the complete GS is used to impute GSK_EU samples. Accuracy (PPV) is measured at 4-digit resolution (genotype validation for both HLA*IMP:01 and HLA*IMP:02). "# Validated" refers to the number of validated alleleles (pre-thresholding). Note that the call threshold for HLA*IMP:02 was matched to obtain equal or higher call rates than with HLA*IMP:01; also, accuracy for HLA*IMP:01 is measured at the genotype level here.

Comparing the two models in detail at T = 0.00, we find that HLA*IMP:02 slightly outperforms HLA*IMP:01 on *HLA-A* and *HLA-DQB1*. Performance on *HLA-DQA1* and *HLA-C* is equal, and HLA*IMP:01 achieves a slightly higher accuracies at *HLA-B* and *HLA-DRB1*. The results at T = 0.7/Matched reflect a similar pattern: HLA*IMP:02 is slightly more accurate on *HLA-A* and *HLA-DQB1*. Performance on *HLA-DQB1* and *HLA-DRB1* achieves parity, and HLA*IMP:01 is slightly more accurate than HLA*IMP:02 on *HLA-C* and *HLA-B*.

In summary, HLA*IMP:02 very slightly outperforms HLA*IMP:01 in thresholded and non-thresholded scenarios on average, but the observed differences are mostly small and in the 1 - 2% range.

Per-allele analyses show nearly identical error profiles for HLA*IMP:01 and HLA*IMP:02 – the conclusion from Section 3.5.2, that well-represented alleles are more reliably imputed, holds for HLA*IMP:02 as well.

4.3.4 Imputing *DRB* structural variation and *HLA-DPB1*

Haplotype graphs, in theory, should be well suited to predict structural variation with simple history. The GSK_EU dataset contains information on presence/absence of *DRB* orthologs (*HLA-DRB3*, *HLA-DRB4*, *HLA-DRB5*) and their allelic state, as well as HLA type information for *HLA-DPB1*. In a 2/3 - 1/3 split experiment on the GSK_EU dataset, the performance of HLA*IMP:02 on these loci was investigated. The *HLA-DRB1* orthologs are only typed at 2-digit resolution, therefore only allowing for validation (and training) at this resolution.

Table 4.4 summarizes the results. At T = 0.00, accuracy at *HLA-DPB1* is at 90% and ≥94% for the DRB orthologs. Setting a threshold of T = 0.7 leads to increased accuracies of ≥98% and call rates of 88% (DPB1) / ≥96%.

Note that the reference sets for these loci, particularly for *HLA-DPB1*, contain only few individuals if compared to the reference sets for the other loci.

Locus	# Validated	Accuracy
HLA-DPB1	48	0.90
HLA-DRB3	190	0.94
HLA-DRB4	190	0.98
HLA-DRB5	190	0.99

Table 4.4: GSK_EU cross validation for additional loci and structural variation (see Section 4.3.4): 2/3 of the GSK_EU dataset are used as reference to impute the remaining 1/3. No call threshold is employed. Accuracy (PPV) for *HLA-DPB1* measured at 4-digit resolution, at 2-digit resolution (including one pseudo-allele for absence) for *DRB* orthologs.

4.4 Summary

In this chapter, I have presented a novel HLA type imputation method, implemented in the framework HLA*IMP:02. HLA*IMP:02 outperforms HLA*IMP:01 in the examined GSK_EU scenario by a small margin (i.e. accuracies and call rates $\geq$90% for all but one of the six examined classical loci, genotype validation) and achieves high accuracies when predicting structural variation (*DRB* orthologs) and *HLA-DPB1*. It has therefore been shown that haplotype graphs can effectively capture the haplotypic variation of the MHC, in a (path-based) way very different from the averaging of allelic backgrounds that HLA*IMP:01 is based on.

The problems encountered by HLA*IMP:01 when imputing into the GSK_EU dataset, i.e. mainly lacking allele coverage and distance between reference and imputation panel, apply – as expected – to HLA*IMP:02. In terms of the scenarios presented in the introduction, the conclusions for HLA*IMP:02 are identical to those for HLA*IMP:01: Scenario 1 can be effectively addressed. It remains to be seen how the two models behave under larger and more heterogeneous reference panels.

HLA*IMP:02 is based on a probabilistic generalization of existing haplotype graph models [Browning and Browning, 2007]. While building the graph, haplotypes are probabilistically attached to nodes. Two novel parameters are introduced, one to optimize for the long-range LD structure of the MHC ("localization") and the other one to model uncertainty in haplotype estimates. The model also naturally deals with missing data and, by separating model building and inference, addresses concerns that missing data imputation

could contaminate the model building procedure [Howie et al., 2009]. Compared with HLA*IMP:01, it has a couple of methodologically appealing advantages. HLA*IMP:02 does not require SNP selection, which renders it more tolerant against missing data (see next chapter) and leads to reduced computational demands (once graphs have been built). It does not require phased input data at any stage, generates allele pair (genotype) probabilities and offers quantitative estimates of consistency between classical typing data and the haplotype graph, which we will return to in the next chapter.

Like HLA*IMP:01, HLA*IMP:02 is available via a user-friendly web interface, enabling biomedical researchers to employ HLA type imputation.

Chapter 5

Heterogeneous reference and imputation panels

In the previous chapters, I have presented and validated two HLA type imputation models. Both models were evaluated in 2/3 - 1/3 cross validation scenarios and on an external dataset, the GSK_EU panel. That is, in terms of the utilized reference panels, the validation results presented so far cover Scenario 1 of the Introduction and partly Scenario 2: imputation from a single-population cohort into a single-population cohort[1] and into a cross-European cohort. Scenarios 2 (multi-population reference set into multi-population imputation panel) and 3 (multi-ethnic reference into multi-ethnic imputation cohort) have not been properly addressed yet.

In this chapter, I will explicitly address these scenarios, or, put differently, how the models behave when presented with more heterogeneous reference data. I will also examine whether matching imputation and reference panel will actually increase imputation accuracies, as predicted by population genetics theory. Note that HLA*IMP:01 requires all data to be phased – as described in previous chapters, HapMap reference panels (which are also included as HLA type imputation reference panels) were used to this end. HLA*IMP:02, in contrast, does not require phased input data.

[1] As noted earlier, the BC58 and CEU cohorts are so close in PCA space that it is a reasonable approximation to assume that they represent a "single" population, although this is of course not true technically. See Figure 5.1.

The experiments presented here are based on different combinations of the GS and GSK datasets. I will first join GS and GSK_EU and then carry out a 2/3 - 1/3 split, resulting in cross-European reference and imputation panels. I refer to this as a situation of medium heterogeneity in the reference panel (Scenario 2). To create a scenario of high heterogeneity (Scenario 3), I will merge the GS with the full (multi-ethnic) GSK dataset and the HapMap YRI samples, and then again carry out a 2/3 - 1/3 split to obtain a training and a validation panel.

For the same reasons as in the previous chapters, I will focus on 4-digit validation, unless stated otherwise. All accuracy figures are based on genotype validation (see Section 3.2, page 72).

5.1 Data

5.1.1 The cross-European panel

Note that the European dataset, GSK_EU, was already described in Section 3.5.2. The GS and GSK_EU datasets are joined into a dataset denoted as GS&GSK_EU. Only SNPs present in both datasets are kept in the joint set. Table 3.2 gives a summary of the datasets in terms of SNPs, HLA typing and HLA alleles, and Table 3.4 describes the distribution of countries the individuals in GSK_EU come from. Note that all individuals in GS&GSK_EU are of (self-declared) White ethnicity.

5.1.2 The multi-ethnic panel

To create a highly heterogeneous multi-ethnic reference panel, the GS, GSK and YRI datasets are joined into a dataset denoted as "GS&GSK_ALL". In contrast to the previous paragraph, the union of SNPs (and not the intersection) is used for the joint dataset. The YRI dataset is the Yoruban HapMap cohort [Consortium et al., 2007], and we applied the same QC protocols as for the other cohorts (see Section 3.3). Table 3.4 summarizes the distributions of countries and ethnicities in the full GSK dataset. Table 3.2 presents a

summary of the SNP, HLA type and HLA allele characteristics of GS&GSK_ALL.

5.1.3 Heterogeneity in the GSK datasets

It is well known that Principal Component Analysis can be used to control for population stratification [Price et al., 2006], and that it is informative of ancestry and sample coalescent times [McVean, 2009; Novembre et al., 2008; Patterson et al., 2006]. GlaxoSmithKline have analyzed the GSK dataset with Eigenstrat [Price et al., 2006] and made the PCA data available. The PCA analysis is based on $\tilde{7}$87,000 SNPs after quality control (carried out by GSK, aimed at removing SNPs with inverted strandedness and missing data $>$ 5%). The xMHC was not excluded. Two sets of analyses were carried out by GSK: a separate analysis of samples of European ancestry and a combined analysis of all samples (which are here used for experiments with mainly European and multi-ethnic datasets, respectively). Both sets included samples from the relevant HapMap populations.

As we introduce additional heterogeneity into the HLA type imputation panels by adding in the GSK samples, heterogeneity in the GSK dataset as measured by PCA is a useful metric to consider (I will later provide an additional analysis of imputation accuracy in terms of the samples' position in PCA space).

Figure 5.1 visualizes the first two principal components for the GSK_EU dataset. It is apparent from the figures that the first two PCs help to separate the samples of European origin; specifically, samples of Eastern European, Southern European and Central European origin tend to cluster together. Most New World samples share ancestry with the Central Europeans. The notion of "heterogeneity" refers to the fact that there are many samples – in particular of Southern European and Eastern European descent – which do not cluster around the Central European or British samples, the population groups which are used in the reference panel.

Figure 5.2 visualizes the first two principal components for the full GSK dataset. As expected, the differentiation between ethnicities is much more pronounced than the differentiation observed between samples of European origin – the first two PCs clearly separate Black, White, Hispanic and Asian samples. Note that there are two outlier groups of White

origin, which seem to share ancestry with some Asian and maybe Black samples. It is not clear whether there are problems with the self-declared ethnicity of these samples.

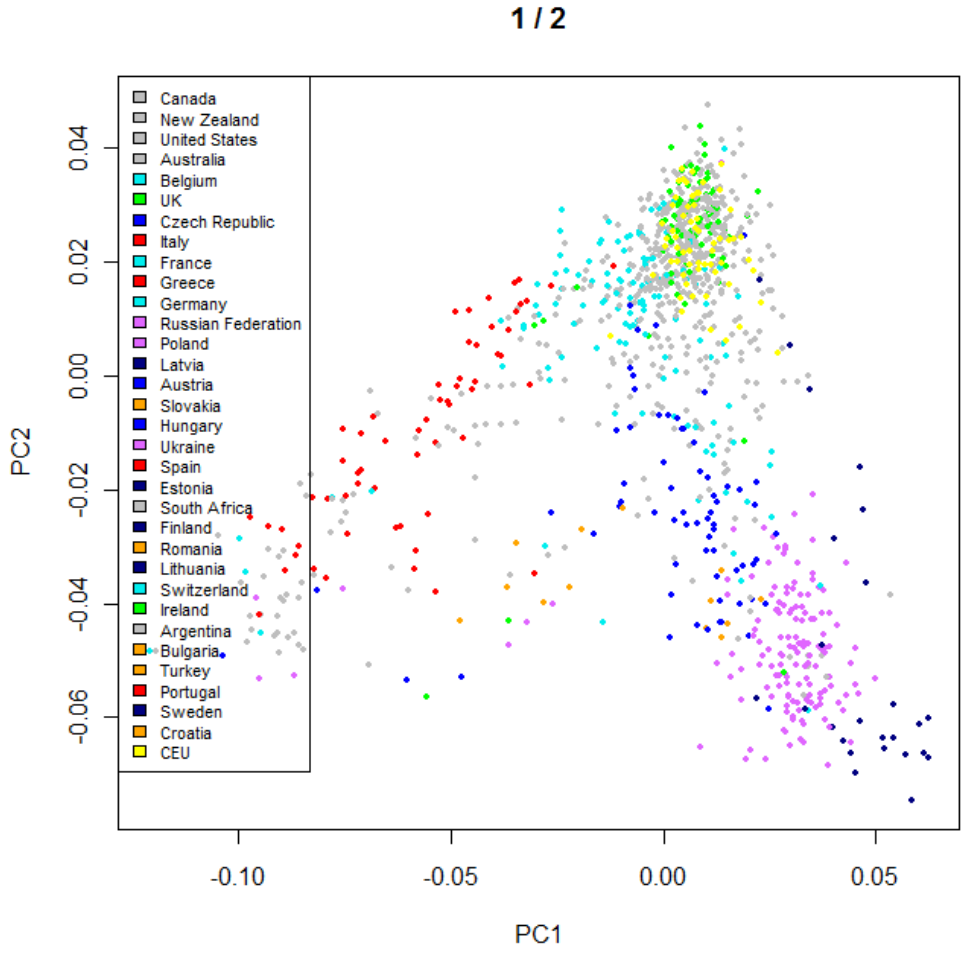

Figure 5.1: Principal Component Analysis (PCA) of the samples in GSK_EU. Shown here: components 1 and 2.

5.2 Validation experiments: Medium heterogeneity

In the experiments with medium heterogeneity in the reference panel, I use 2/3 of the GS&GSK_EU dataset to impute the remaining 1/3. Table 5.1 summarizes the results in a comparative manner at a call threshold of T = 0.00 and T = 0.70. All figures are, unless stated differently, based on genotype validation at 4-digit resolution.

The differences between HLA*IMP:01 and HLA*IMP:02 are more pronounced than in the

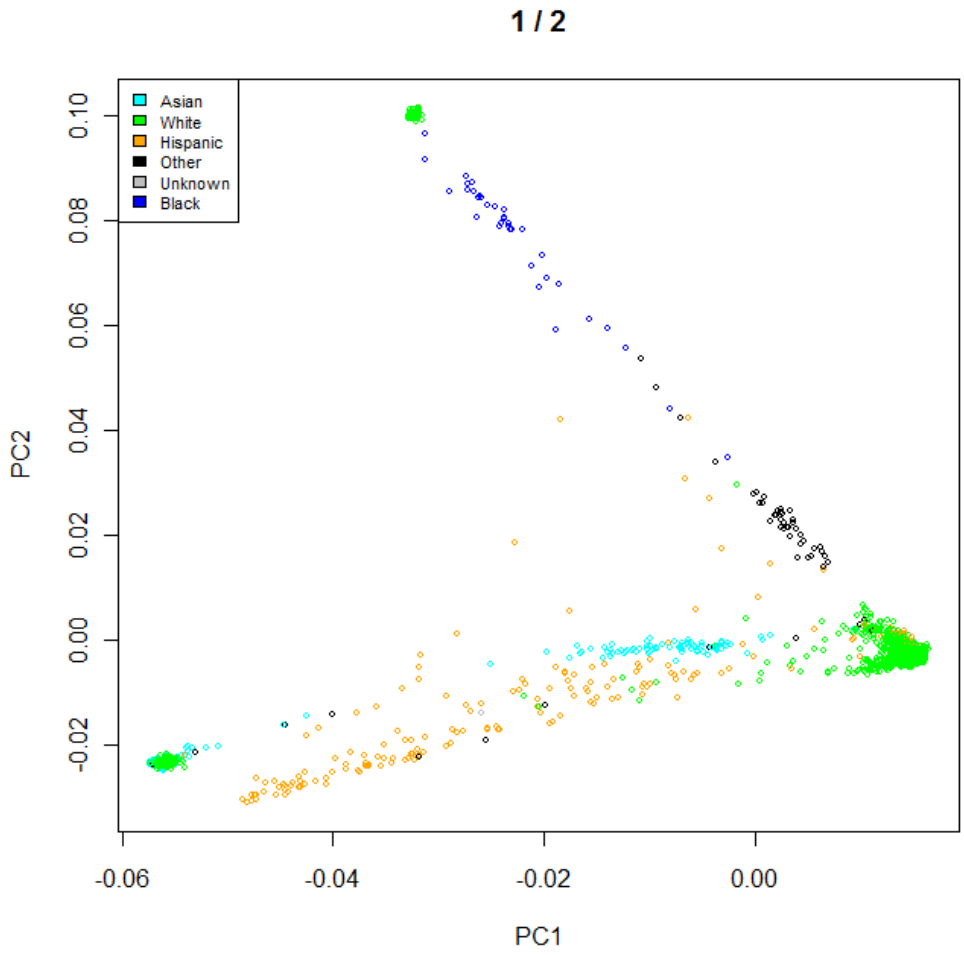

Figure 5.2: Principal Component Analysis (PCA) of the samples in GSK_ALL. Shown here: components 1 and 2. Note that there are two outlier groups of "White" origin, sharing ancestry with Asian and possibly Black samples. It is likely that this is an artefact, possibly arising from wrong self-declaration.

previous validation experiments.

First, note that HLA*IMP:01 still achieves accuracies ≥87% at all loci at T = 0.00, but also note that the accuracy at most examined loci is lower than HLA*IMP:01 accuracy in the GSK validation experiment presented in Section 4.3 (note that I reference the HLA*IMP:02 chapter here, because genotype accuracies are presented only there). This behaviour is clearly unexpected: moving one part of the GSK dataset into the training data should increase, not decrease, performance. This expectation is underlined by the greater diversity of HLA alleles represented in GS&GSK 2/3, as compared to GS alone (see Table 3.2).

HLA*IMP:02, in contrast, achieves an average accuracy of 96% at T = 0.00 (versus 90% for HLA*IMP:01). It outperforms HLA*IMP:01 at every locus, the locus-specific differences ranging from 4% to 10%. At T = 0.70, HLA*IMP:02 achieves an average accuracy of 97% at an average call rate of 97%. The most problematic locus for HLA*IMP:02 is *HLA-DRB1*, with an achieved accuracy/call rate of 91%/100% at T = 0.00 and 95%/91% at T = 0.70.

Figure 5.3 shows that HLA*IMP:02 is generally well-calibrated, the confidence intervals generally including the expected value.

For the interested reader, in Table 5.2, I also provide some results on BEAGLE [Browning and Browning, 2009]. BEAGLE is no explicit HLA type imputation method and therefore not in the focus of this thesis. In general, it can be expected that HLA type imputation for BEAGLE is "challenging" (Sharon R. Browning, personal communication). The presented results are based on 20 iterations – as suggested (Sharon R. Browning, personal communication), different numbers of iterations were tried out, the results for 10 and 50 iterations being very similar to 20 iterations (data not shown).

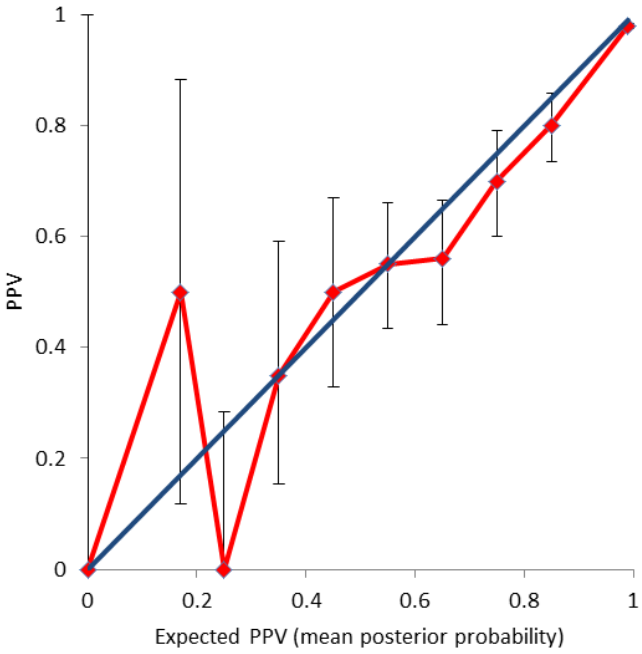

Figure 5.3: Calibration plot HLA*IMP:02, medium heterogeneity (see Section 5.2). The red points show expected (x-axis) and achieved mean accuracies (y-axis) in each bin of step size 0.1, and the blue line is a plot of $x = y$. Note that the first four data points (bins 0 - 3) are only based on 37 individuals.

5.2.1 PCA analysis

If we compare the GS&GSK_EU results for HLA*IMP:02 obtained here to the results obtained from the 2/3 - 1/3 GS cross validation experiment (see Section 4.3.2), we find that the accuracies are at a similarly high level. As shown in Table 3.2, the allelic diversity in GS&GSK_EU 1/3 is higher than in the full Golden Set (GS) – that is, imputation into GS&GSK_EU 1/3 is certainly harder than imputation into GS 1/3, and by extending the reference panel we achieve similar accuracies for the harder scenario as previously for the easier one.

Is it possible to relate these results to the increased heterogeneity, in particular in terms of population structure, in reference and imputation panels? The hypothesis we would like to test is whether matching the reference to the imputation panel in terms of population structure actually increases accuracies. If this hypothesis holds, we would expect a) that accuracies for the already well-matched samples do not increase by extending the reference panel and b) that imputation accuracy for distant samples (in terms of the non-matched reference panel) increases by improving the match. We can examine these predictions by restricting our attention to the samples with PCA data available (i.e. the GSK_EU samples, of which roughly 1/3 are present in GS&GSK_EU 1/3) and by stratifying accuracy by sample position in PCA space. In this framework, we now compare the accuracies observed based on the complete GS&GSK_EU 2/3 (well matched) reference panel with accuracies achieved based on a GS&GSK_EU 2/3 variant with all non-GS samples removed (less well matched, as leaving only British and Central European samples in the reference panel).

The data is in good agreement with the two expectations formulated above. For *HLA-B*, *HLA-C*, *HLA-DQA1* and *HLA-DRB1*, the mean accuracy over all quadrants increases. This effect is driven by the periphery (see Figures 5.4, 5.5 and 5.6, which compare the achieved accuracies for *HLA-B*, *HLA-C* and *HLA-DRB1*), whereas accuracies around the centre of the GS training change only slightly, if at all. For *HLA-DQB1* and *HLA-A*, we observe a small decrease in mean accuracy (0.4% and 1%, respectively). The decrease at *DQB1* is related to a single quadrant. At *HLA-A*, the drop is caused by two quadrants

close to the center to the GS training data. It is not clear whether this is an example of very subtle population structure, or simply a sampling effect. The overall effect of improving the match of the reference panel is, however, clearly positive.

To rule out the possibility that the different sizes of the compared reference panels could account for the observed difference, I repeated the experiment, this time comparing GS&GSK_EU 2/3 and the full GS. The essential results still hold: mean accuracy is constant for *HLA-A*, slightly worse for *HLA-C* and *HLA-DQB1*, slightly increased for *HLA-DRB1* and *HLA-B* and markedly increased for *HLA-DQA1* (approx. 7%). The results for *HLA-DQB1* are driven by two outlier quadrants, and most of the difference for *HLA-C* occurs in the proximity of the GS reference data, indicating that the GS probably better represents the Central European/British structure of *HLA-C* than GS&GSK_EU 2/3.

Threshold	**Locus**	**# Validated**	**HLA*IMP:02**		**HLA:IMP:01**	
			Call Rate	Accuracy	Call Rate	Accuracy
T= 0.00	*HLA-A*	808	1.00	0.97	1.00	0.91
	HLA-B	1646	1.00	0.95	1.00	0.89
	HLA-C	752	1.00	0.96	1.00	0.91
	HLA-DQA1	194	1.00	0.97	1.00	0.87
	HLA-DQB1	934	1.00	0.98	1.00	0.92
	HLA-DRB1	1358	1.00	0.91	1.00	0.87
T = 0.70	*HLA-A*	808	0.98	0.97	0.94	0.94
	HLA-B	1646	0.96	0.97	0.93	0.92
	HLA-C	752	0.99	0.97	0.94	0.94
	HLA-DQA1	194	0.96	0.98	0.93	0.90
	HLA-DQB1	934	0.99	0.98	0.94	0.94
	HLA-DRB1	1358	0.91	0.95	0.89	0.92

Table 5.1: Medium heterogeneity non-thresholded and thresholded cross-validation results for HLA*IMP:02 and HLA*IMP:01 (see Section 5.2): GS&GSK_EU 2/3 is used to impute GS&GSK_EU 1/3. Accuracy (PPV) is measured at 4-digit resolution (genotype validation for both HLA*IMP:01 and HLA*IMP:02). "# Validated" refers to the number of validated alleleles (pre-thresholding).

5.2.2 Robustness to missing data

HLA*IMP:02 does not require SNP selection and can utilize all available SNP genotype information. However, many imputation panels, in particular from older SNP genotyping chips, comprise a limited set of SNPs in the xMHC region. In two variations of the

Threshold	Locus	# Validated	HLA*IMP:02		BEAGLE	
			Call Rate	Accuracy	Call Rate	Accuracy
T= 0.00	*HLA-A*	808	1.00	0.97	1.00	0.88
	HLA-B	1646	1.00	0.95	1.00	0.95
	HLA-C	752	1.00	0.96	1.00	0.91
	HLA-DQA1	194	1.00	0.97	1.00	0.92
	HLA-DQB1	934	1.00	0.98	1.00	0.94
	HLA-DRB1	1358	1.00	0.91	1.00	0.91

Table 5.2: Medium heterogeneity non-thresholded and thresholded cross-validation results for HLA*IMP:02 and BEAGLE [Browning and Browning, 2009] (see Section 5.2): GS&GSK_EU 2/3 is used to impute GS&GSK_EU 1/3. Accuracy (PPV) is measured at 4-digit resolution (genotype validation). "# Validated" refers to the number of validated alleleles. As recommended by Sharon R. Browning (*personal communication*), BEAGLE was run with 10, 20 (data shown) and 50 iterations. There was no indication for convergence problems and no qualitative difference between the runs in terms of the achieved accuracies.

medium heterogeneity scenario, I therefore investigated how robust the model is towards larger amounts of missing data. The results of these experiments are presented in Table 5.3. In the first scenario (70% missing data), the achieved per-locus accuracy is reduced by at maximum 1%. Note that each individual in the 70% scenario is left with approximately 1800 SNPs in the xMHC. This is on the order of the xMHC coverage of many ~500K genotyping arrays. Even in the second scenario (90% missing data), accuracy is relatively stable – the maximum loss in accuracy is 5% for all loci but *HLA-DQA1* (probably related to the smaller amount of reference data for this locus), where it is 7%. In the second scenario, each individual is left with approximately 600 SNPs in the xMHC region – this is substantially less than than even older ~300K genotyping arrays provide.

Locus	# Validated	Accuracy 70%	Accuracy 90%
HLA-A	808	0.96	0.94
HLA-B	1646	0.95	0.93
HLA-C	752	0.95	0.94
HLA-DQA1	194	0.96	0.90
HLA-DQB1	934	0.97	0.95
HLA-DRB1	1358	0.90	0.86

Table 5.3: 4-digit resolution accuracies (PPV) when 70% and 90% of the inference panel SNP genotypes (GS&GSK_EU 1/3) in the medium heterogeneity experiment are randomly set to missing (see Section 5.2.2). No call threshold is employed. "# Validated" refers to the number of validated alleleles.

5.2.3 Errors in the reference dataset

The effectiveness of the model's internal HLA type quality control is assessed by introducing artificial errors into a reference genotype set and measuring how many of these are detected during the graph-building stage. Let S_1 denote the reference set without artificially introduced errors and S_2 the set with artificial errors. S_2 is generated by copying S_1 and assuming an error rate of 0.02 for alleles in individuals with two alleles typed at 4-digit accuracy. If an allele is hit by an "error" event, a new allele is drawn from the population distribution at random (here estimated from S_1) and assigned to the corresponding individual in S_2. We can then build haplotype graphs for S_1 and S_2, resulting in two sets of posterior probability estimates of individual consistency with the specified classical HLA types (this immediately follows from the sampled haplotypes, conditional on the current graph estimate and the observed genotypes, as noted above). Filtering these sets for individuals with a posterior HLA type inconsistency probability of ≥ 0.001 results in two sets E_1 and E_2 of individuals with potential classical typing problems. We compare E_1 (based on the graph for S_1) and E_2 (based on S_2). Note that E_1 should not necessarily be empty, as one would expect a certain error probability in lab-based HLA genotyping. Define $E_{\text{new}} = E_2 \setminus E_1$ and $E_{\text{missing}} = E_1 \cap \overline{E_2}$. All elements of E_{new} are defined as "detected errors", and we measure sensitivity and minimum PPV accordingly by comparing E_{new} with the set of artificially introduced errors. Introducing artificial errors could also distort the graph in a way that reduces sensitivity for pre-existing errors in S_1. Therefore, we also measure the potential loss of sensitivity, defined as $|E_{\text{missing}}|/|E_1|$.

Artificial errors were introduced into the GS&GSK_EU 2/3 reference panel (i.e. S_1 = GS&GSK_EU 2/3) for *HLA-C* and *HLA-DQB1*.

At *HLA-C*, 25 of 30 artificially introduced errors are detected (i.e. sensitivity = 0.83). 27 individuals in total were flagged as problematic (i.e. minimum PPV = 0.93). Of the 45 individuals flagged as problematic in the non-modified dataset, 11 are lost in the dataset with artificial errors (i.e. maximum detection loss = 0.24). The corresponding results for *HLA-DQB1* are broadly comparable: sensitivity is at 0.76, minimum PPV at 0.94 and maximum detection loss at 0.12.

These results demonstrate that it is generally possible to spot errors in the reference dataset, although more work needs to be done to develop a more detailed understanding of the strengths and shortcomings of this approach. Clearly, the posterior probabilities for an error in the classically typed data are not well-calibrated – the results presented here assume that a posterior error probability of more than 0.001 is high enough to classify an individual as problematic. Also, it was not attempted to estimate an overall classical HLA typing error rate, and to link m (sampling and graph building error) for HLA loci to such an estimate. Note that the two loci considered here were selected on the basis that they are comparably easy to impute; it can be expected that the results for *HLA-DRB1* and *HLA-B* would look worse. However, extending the experiment to the more complicated loci seems promising only after having developed (and validated on the less complicated loci) a way to specifically estimate m for potentially strongly contaminated reference panels.

5.3 Validation experiments: High heterogeneity

To investigate whether the good performance of HLA*IMP:02 in scenarios with medium heterogeneity translates to scenarios of high heterogeneity, I created a merged GS&GSK_ALL dataset (comprising samples of multiple ethnicities, described above), and carried out a 2/3 - 1/3 split experiment. It has to be emphasized that this experiment is limited by the amount of available non-European reference data. A priori, it is clear that this data will not be sufficient to build a high-accuracy reference panel for other ethnicities, and therefore we cannot examine how HLA*IMP:02 would behave on such an (even larger) panel. However, positive results on the smaller scenario examined here would be encouraging and warrant further research into highly heterogeneous reference panels.[2]

The results for the high-heterogeneity experiment are presented in Table 5.4. Note that accuracies are broken down by imputation sample ancestry, i.e. by whether a sample is European or not. For the European samples, compared to the medium heterogeneity

[2]For these reasons, I do not present results for BEAGLE here – the non-European proportion of the reference panel is too small to allow for a substantial performance comparison of different models with generalizable results.

experiment, we find that accuracies and call rates (for T = 0.7) are usually, though not always, minimally decreased, mostly in the 1% range. The un-thresholded accuracies for non-European samples are much lower: *HLA-A*, *HLA-B* and *HLA-DRB1* are in the 59 - 66% range. *HLA-C*, *HLA-DQA1* and *HLA-DQB1* are more reliably imputed: accuracies range from 81 - 89%. Setting a call threshold of T = 0.7 has the expected effect: accuracies increase, to ≥95% (call rates ≥95%, 85% for DRB1) for the European samples, but only to ≥78% (call rates ≥71%) for the non-European samples.

How should we interpret these results? First, note that the slightly decreased accuracies for the European samples might be related to a smaller chromosomal region being captured in this experiment: due to the increased xMHC SNP density in this experiment, the 300 SNP (standard settings were not changed) windows around the classical loci capture a smaller region in terms of genetic distance than in the previous experiments. However, the depression of results is small, so that it is not investigated in further detail here.

The observed accuracies for the non-European samples are low, but how low exactly? In order to address this question, I removed all non-European samples from GS&GSK_ALL, and repeated the experiment. The results are presented in Table 5.5: while the results for the Europeans fluctuate a bit, accuracies for the non-European samples are significantly (8% on average) lower than when utilizing the complete reference panel.

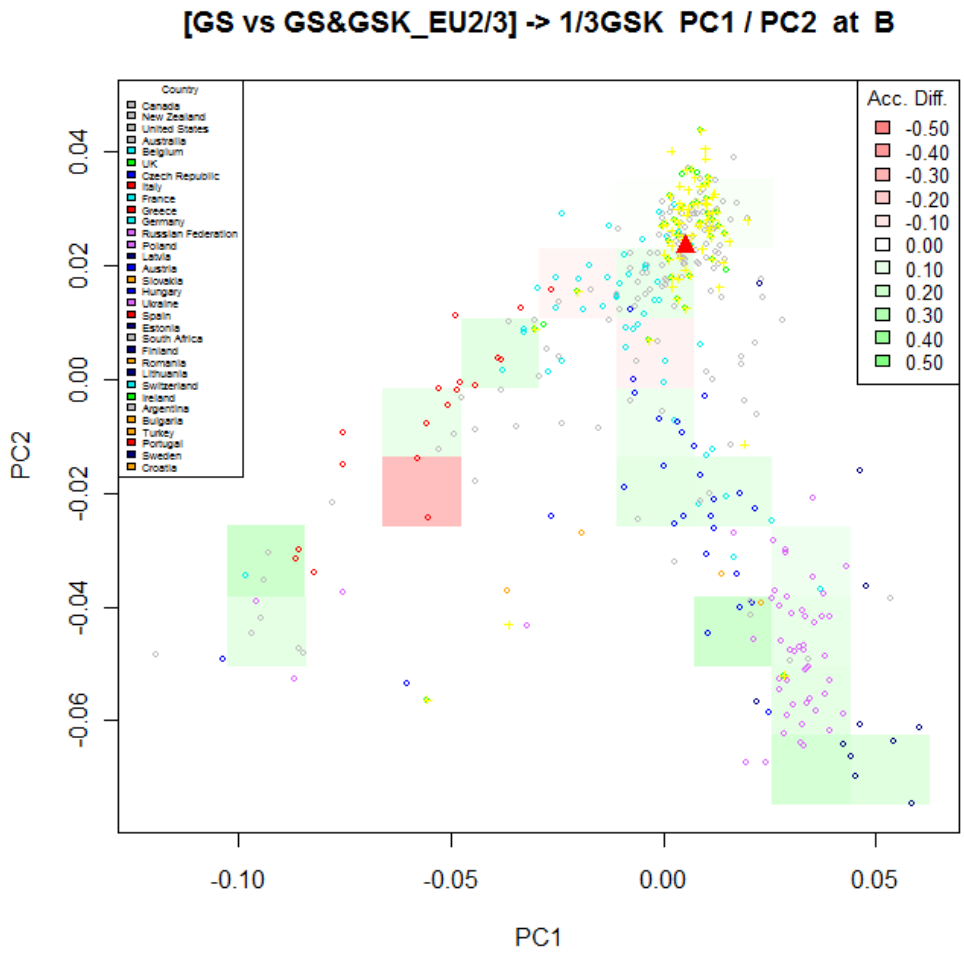

Figure 5.4: PCA-stratified accuracy comparison (*HLA-B*) between the complete reference panel and a GS-restricted reference panel for the medium heterogeneity scenario (Section 5.2). In each quadrant, the difference in mean accuracy (PPV) between the two reference panel scenarios is indicated by a color. Green indicates that the complete reference panel leads to higher accuracies in a quadrant. Yellow points indicate the position of the CEU cohort, and the red triangle is in the (approximate) center of the GS panel.

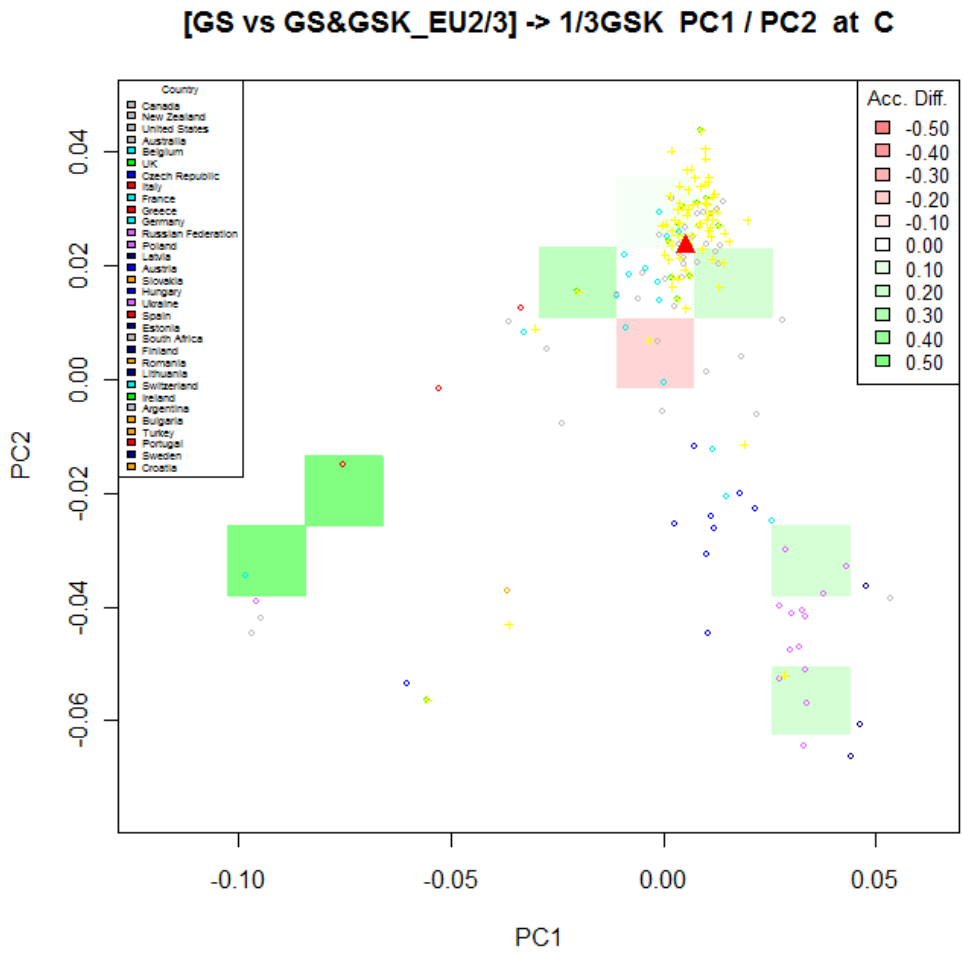

Figure 5.5: PCA-stratified accuracy comparison (*HLA-C*) between the complete reference panel and a GS-restricted reference panel for the medium heterogeneity scenario (Section 5.2). In each quadrant, the difference in mean accuracy (PPV) between the two reference panel scenarios is indicated by color. Green indicates that the complete reference panel leads to higher accuracies in a quadrant. Yellow points indicate the position of the CEU cohort, and the red triangle is in the (approximate) center of the GS panel.

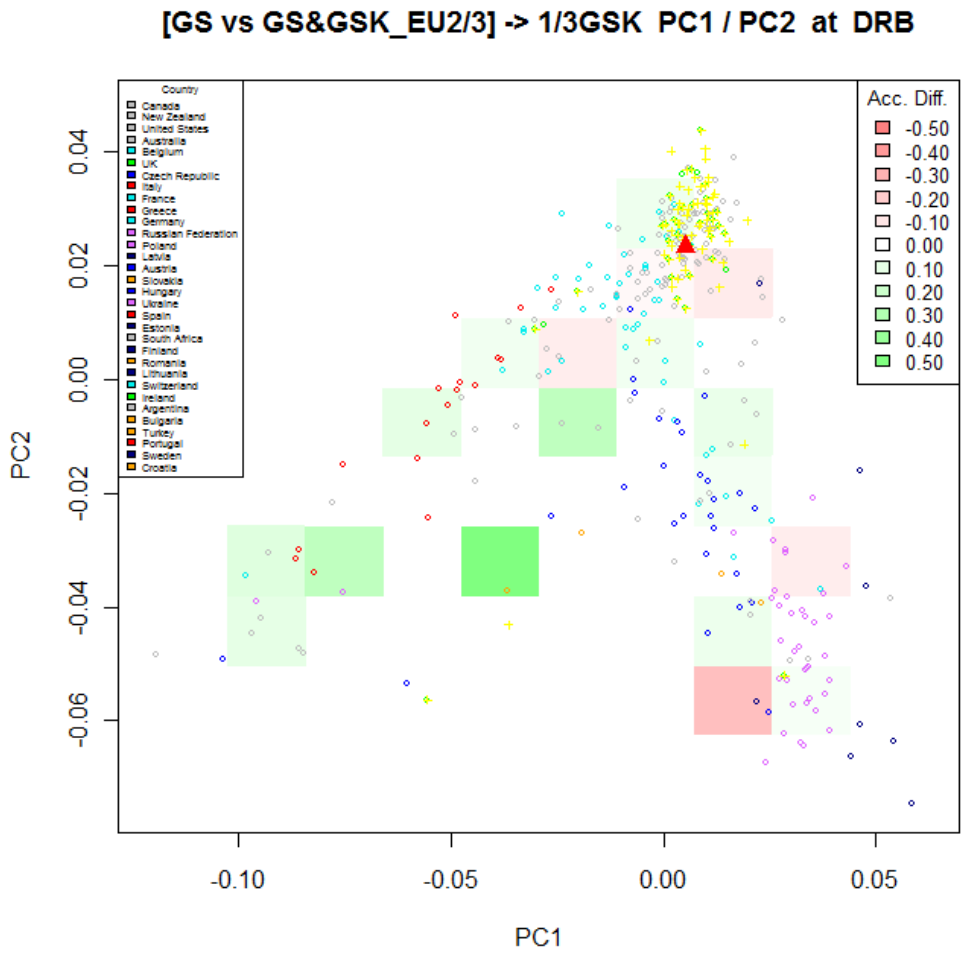

Figure 5.6: PCA-stratified accuracy comparison (*HLA-DRB1*) between the complete reference panel and a GS-restricted reference panel for the medium heterogeneity scenario (Section 5.2). In each quadrant, the difference in mean accuracy (PPV) between the two reference panel scenarios is indicated by color. Green indicates that the complete reference panel leads to higher accuracies in a quadrant. Yellow points indicate the position of the CEU cohort, and the red triangle is in the (approximate) center of the GS panel.

Threshold	Locus	Combined			European			Non-European		
		# Validated	Call Rate	Accuracy	# Validated	Call Rate	Accuracy	# Validated	Call Rate	Accuracy
T= 0.00	*HLA-A*	910	1.00	0.94	824	1.00	0.96	86	1.00	0.78
	HLA-B	1942	1.00	0.91	1662	1.00	0.95	280	1.00	0.66
	HLA-C	846	1.00	0.96	752	1.00	0.97	94	1.00	0.9
	HLA-DQA1	284	1.00	0.94	206	1.00	0.96	78	1.00	0.87
	HLA-DQB1	1030	1.00	0.97	924	1.00	0.97	106	1.00	0.91
	HLA-DRB1	1620	1.00	0.88	1356	1.00	0.90	264	1.00	0.77
T = 0.70	*HLA-A*	910	0.95	0.96	824	0.95	0.97	86	0.98	0.80
	HLA-B	1942	0.92	0.95	1662	0.95	0.97	280	0.78	0.78
	HLA-C	846	0.99	0.97	752	0.99	0.97	94	0.95	0.92
	HLA-DQA1	284	0.95	0.96	206	0.97	0.97	78	0.88	0.91
	HLA-DQB1	1030	0.98	0.97	924	0.99	0.98	106	0.90	0.94
	HLA-DRB1	1620	0.84	0.94	1356	0.87	0.95	264	0.71	0.87

Table 5.4: High heterogeneity non-thresholded and thresholded cross-validation results for HLA*IMP:02 (see Section 5.3): GS&GSK_ALL 2/3 is used to impute GS&GSK_ALL 1/3. Accuracy (PPV) is measured at 4-digit resolution and stratified ("European", "Non-European") by the ethnicity of the samples in the imputation dataset. The column "Combined" presents un-stratified accuracies. "# Validated" refers to the number of validated alleleles (pre-thresholding).

Reference	Locus	European			Non-European		
		# Validated	Call Rate	Accuracy	# Validated	Call Rate	Accuracy
Complete	*HLA-A*	824	1.00	0.96	86	1.00	0.78
	HLA-B	1662	1.00	0.95	280	1.00	0.66
	HLA-C	752	1.00	0.97	94	1.00	0.90
	HLA-DQA1	206	1.00	0.96	78	1.00	0.87
	HLA-DQB1	924	1.00	0.97	106	1.00	0.91
	HLA-DRB1	1356	1.00	0.90	264	1.00	0.77
European	*HLA-A*	824	1.00	0.96	86	1.00	0.66
	HLA-B	1662	1.00	0.94	280	1.00	0.59
	HLA-C	752	1.00	0.97	94	1.00	0.85
	HLA-DQA1	206	1.00	0.97	78	1.00	0.81
	HLA-DQB1	924	1.00	0.98	106	1.00	0.89
	HLA-DRB1	1356	1.00	0.88	264	1.00	0.59

Table 5.5: High heterogeneity non-thresholded and thresholded cross-validation results for HLA*IMP:02 (see Section 5.3), comparing performance on the complete GS&GSK_ALL 2/3 to performance on a European-only reduced version of GS&GSK_ALL 2/3. Accuracy (PPV) is measured at 4-digit resolution and stratified ("European", "Non-European") by the ethnicity of the samples in the imputation dataset. "# Validated" refers to the number of validated alleleles (pre-thresholding).

5.3.1 PCA analysis

Again, we can analyze the difference between the complete GS&GSK_ALL reference panel and the European-only variant in terms of Principal Components. Figures 5.7 and 5.8 show the results for *HLA-B* and *HLA-DRB1*, two particularly problematic loci in the high heterogeneity experiment (data for other loci similar, but less pronounced). It is even more clear than in the European heterogeneity experiments that adding in matched training data improves performance in the more distant quadrants, whereas performance in the already well-covered quadrant around the European training data remains essentially unchanged.

5.3.2 2-digit analysis

In some studies, 2-digit HLA genotyping may be sufficient to obtain meaningful results. I therefore provide 2-digit validation figures for the non-European samples in the high heterogeneity scenario. Table 5.6 summarizes the results: even without setting a call threshold, 2-digit accuracy is $\geq 93\%$ for all loci but *HLA-B*, where it is at 86%.

Locus	# Validated	Call Rate	Accuracy (2-digit)
HLA-A	94	1.00	0.93
HLA-B	284	1.00	0.86
HLA-C	104	1.00	0.97
HLA-DQA1	94	1.00	0.97
HLA-DQB1	114	1.00	0.96
HLA-DRB1	274	1.00	0.97

Table 5.6: High heterogeneity cross-validation results (PPV) for non-European samples, measured at 2-digit resolution for HLA*IMP:02 (see Section 5.3). No call threshold is employed. "# Validated" refers to the number of validated alleleles.

5.4 Discussion

I have demonstrated that HLA*IMP:02 achieves high accuracies ($\geq$95% range for all loci but *HLA-DRB1*, 91% for *HLA-DRB1*, without a call threshold) when applied to a medium-heterogeneity multi-population reference panel. Maybe paradoxically, HLA*IMP:01 produces less accurate imputations on the same data set, even when compared to results from a reference panel with less allele coverage. The PCA analysis clearly shows that HLA*IMP:02 benefits from improving the match between reference and imputation panels. It also indicates that PCs can, to an extent, be used to predict imputation success, but I do not attempt to formalize this notion in this thesis.

The high-heterogeneity multi-ethnicity experiment indicates that the ability of HLA*IMP:02 to exploit heterogeneity in the reference panel may hold in even more diverse situations, but the limitations of the data do not allow for any definitive statements. The fact that haplotypic relationships between alleles can be picked up at the 2-digit level for quite a diverse set of samples justifies cautious optimism. The PCA analysis results from the high-heterogeneity experiment are consistent with those from the medium-heterogeneity experiment. Taken together, these results are encouraging with respect to future extensions of the reference data.

Note that the medium heterogeneity experiment refers to Scenario 2 from the Introduction, and that the high heterogeneity experiment refers to Scenario 3. Scenario 2 – in which HLA*IMP:02 produces high call rates and accuracies – is relevant to many current and future multi-country disease association studies, and the data presented here suggest that

HLA*IMP:02 can contribute to such studies in the future by providing more accurate HLA type information.

In the context of medium heterogeneity, I have also demonstrated that HLA*IMP:02 is tolerant of missing data up to between 70 and 90% missingness of individual genotypes, a highly desirable property in practical applications. To some extent, it seems also possible to identify errors in the reference panel, but this issue needs to be studied in more detail.

I will try to provide some intuition as to why HLA*IMP:02 is superior to HLA*IMP:01 in dealing with heterogeneous reference panels. For a given SNP haplotype with unknown HLA type, HLA*IMP:01 first computes the L&S emission probabilities separately for each allelic group, and then normalizes these probabilities to obtain a distribution over possible HLA types. To some extent, the group-specific emission probabilities measure the degree of similarity between a group and a given haplotype, in a manner that assigns equal weight to each chromosome in a group. If an allele occurs on different SNP haplotype backgrounds, if there are technical typing differences between member chromosomes (systematically different SNP genotypes, say), the result will inevitably be a decrease of the "correct" group's emission probability. The SNP selection function of HLA*IMP:01 tries to avoid these problems by identifying a consensus set of informative SNPs, but the higher the heterogeneity of the dataset, the less successful this strategy will be (in particular if this heterogeneity includes alleles which appear on genuinely different haplotypic backgrounds).

HLA*IMP:02, in contrast, bases inference on *single path* probabilities: systematically different haplotypic backgrounds can be represented as different paths through the model, and increasing the number of such paths will not compromise the model's ability to correctly follow already existing paths.

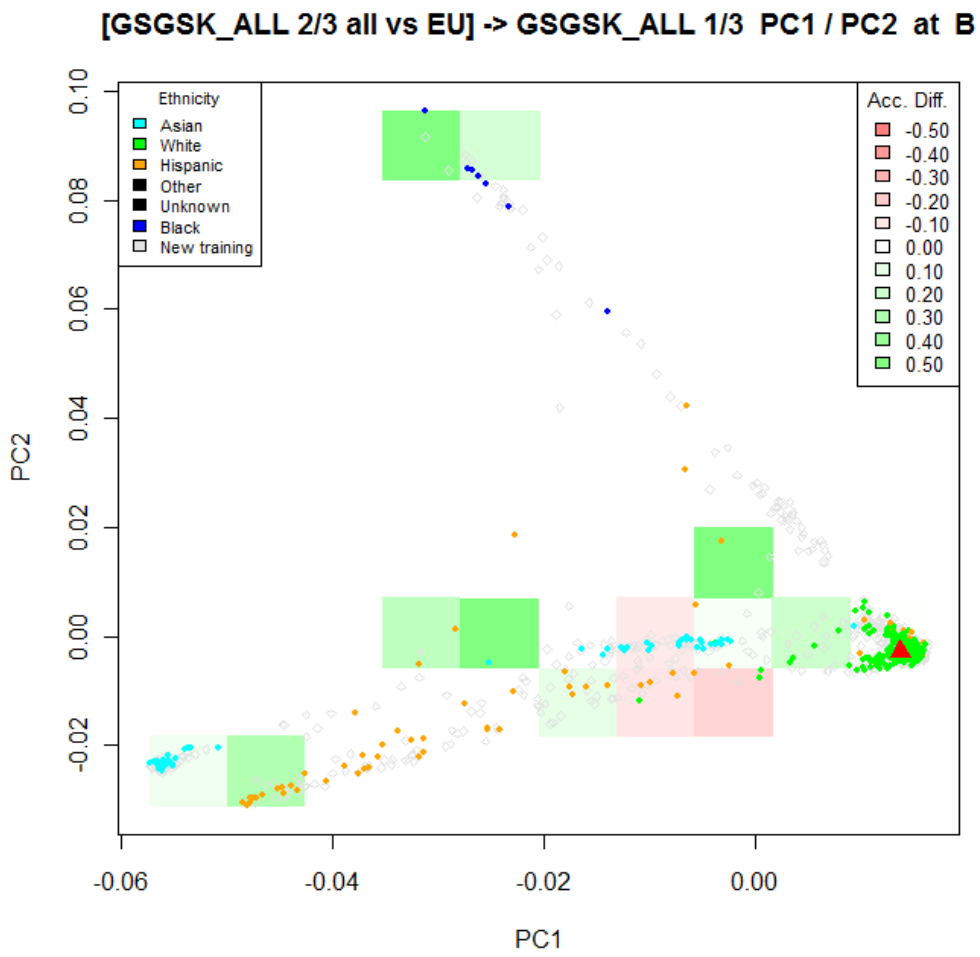

Figure 5.7: PCA-stratified accuracy comparison (*HLA-B*) between the complete reference panel (GS&GSK_ALL) and a European-restricted reference panel for the high heterogeneity scenario (see Section 5.3). In each quadrant, the difference in mean accuracy (PPV) between the two reference panel scenarios is indicated by color. Green indicates that the complete reference panel leads to higher accuracies in a quadrant. Gray points indicate the positions of training samples which were removed for the Europe-restricted analysis, and the red triangle is in the (approximate) centre of the European reference data.

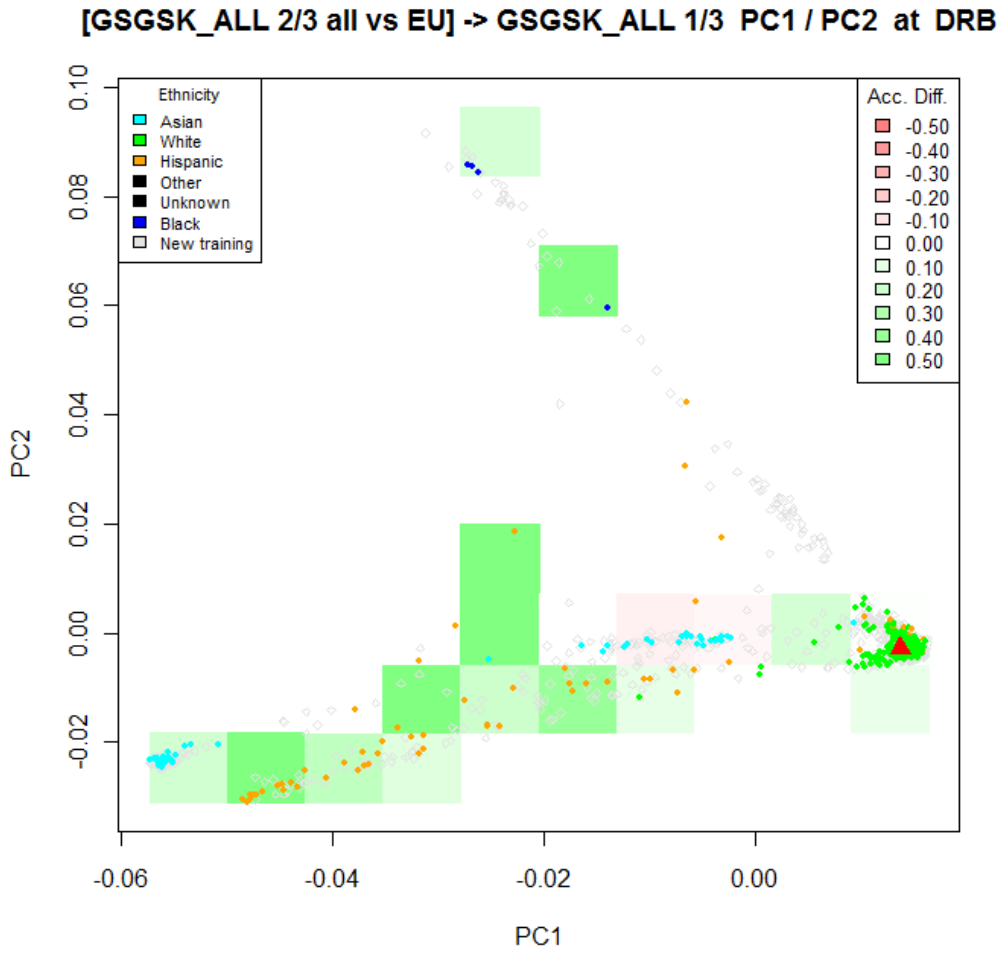

Figure 5.8: PCA-stratified accuracy comparison (*HLA-DRB1*) between the complete reference panel (GS&GSK_ALL) and a European-restricted reference panel for the high heterogeneity scenario (see Section 5.3). In each quadrant, the difference in mean accuracy (PPV) between the two reference panel scenarios is indicated by color. Green indicates that the complete reference panel leads to higher accuracies in a quadrant. Gray points indicate the positions of training samples which were removed for the Europe-restricted analysis, and the red triangle is in the (approximate) centre of the European reference data.

Chapter 6

Bayesian integration of HLA type imputation in GWAS

Bayesian statistical methods are commonly employed in genome-wide association studies [Consortium, 2007; Stephens and Balding, 2009]. In this chapter, I describe how to analyse HLA type imputations in a Bayesian framework for disease associations. It complements the work of my colleague Loukas Moutsianas, who has described an analogous solution for the frequentist case [Moutsianas, 2011].

To specify the Bayesian framework, I begin with introducing Bayes Factors, an essential measure for model comparisons. I then go on to describe retrospective and prospective likelihoods, integrating the imputation-derived uncertainties, and describe how to use the Laplace approximation to approximate the integral of the posterior probability function.

In the second part of this chapter, I present a limited validation study on the potential benefits derived from using HLA*IMP:02 in this framework, and I show some uncertainty-aware results from a GWAS recently published [Sawcer et al., 2011].

6.1 Bayesian measures of association

Bayes Factors (BFs) measure whether a model M_1 is preferred over a model M_0, conditional on some observed data D, by integrating the probability of the data over both models' parameter space:

$$BF := \frac{\int \Pr_{M_1}(\theta) \times \mathrm{P}_{M_1}(D|\theta)\,\mathrm{d}\theta}{\int \Pr_{M_0}(\theta) \times \mathrm{P}_{M_0}(D|\theta)\,\mathrm{d}\theta}\,.$$

In this notation, a large BF indicates that M_1 is preferred over M_0. Bayes Factors are sometimes more intuitively interpreted than p-values, allowing for direct within-study and across-study comparisons of the strength of an observed association [Stephens and Balding, 2009].

In association studies, M_0 is usually a model with no effect at a particular locus, and M_1 allows for a risk-modifying effect of particular alleles. The specification of M_1 determines how alleles at a locus change risk, i.e. in a (partially) dominant, (partially) recessive or additive manner, and how this effect is parameterized – in models which are not fully additive, two parameters are needed to specify the risk of heterozygotes and homozygotes.

In order to use Bayesian methods to analyze HLA type imputation data, two fundamental issues must be addressed: i) How to take into account uncertainty when specifying the likelihood of the data, and ii) how to carry out the actual Bayes Factor computation (the underlying integral is not necessarily trivial)?

6.1.1 The likelihood

Prospective and retrospective approaches

The retrospective likelihood Most genome-wide association datasets are sampled conditional on disease status (*retrospectively*): i.e. a random sample of individuals is taken from the population of cases, and another sample of individuals is taken from a unaffected control population. All individuals are subsequently genotyped. To address the question of

association, a statistical test is carried out which compares the distribution of genotypes between cases and controls. Any alleles with risk-increasing effects are expected to be enriched in the case sample, and vice versa.

To formalize these notions, let G_i denote the count of control individuals with genotype i, and C_i the number of case individuals with genotype i. G and C denote the total count of cases and controls, and individuals are ordered in a way that the first G individuals are all controls. n is the number of genotypes (as this notation implies, "genotype" is used in a broad sense here, and may refer to the allelic state of multiple loci – each unique combination would then be assigned an individual index i). Genotype frequencies can be assumed to follow a multinomial distribution, parameterized by the vector θ (each element of θ specifies the expected frequency of the corresponding genotype). Under M_1, we allow the multinomial genotype distribution in cases to differ from that in controls (θ_C and θ_G), according to some assumed risk effect of genotypes. Under M_0, the multinomial genotype distributions in cases and controls are identical. The likelihood of the data under M_0 and M_1 is then defined as

$$\mathrm{L}_{M_1}(\theta_C, \theta_G) := \mathrm{mult}(C_1, .., C_n | C, \theta_C) \times \mathrm{mult}(G_1, .., G_n | G, \theta_G)$$

and

$$\mathrm{L}_{M_0}(\theta) := \mathrm{mult}(C_1, .., C_n | C, \theta) \times \mathrm{mult}(G_1, .., G_n | G, \theta)\,.$$

The difference between θ_C and θ_G is usually parameterized in terms of *Odds Ratios* (ORs). Define the *odds* of disease for a particular genotype i as $\mathrm{odds}_i := \mathrm{P}(\mathrm{disease}|\mathrm{genotype} = i)/(1 - \mathrm{P}(\mathrm{disease}|\mathrm{genotype} = i))$ and the odds ratio of genotype i against genotype j as $\mathrm{OR}_{i,j} := \mathrm{odds}_i/\mathrm{odds}_j$. Usually, the genotype with the lowest risk is assigned the index 1, and the ORs of all other genotypes are specified against genotype 1: $\mathrm{OR}_i := \mathrm{odds}_i/\mathrm{odds}_1$, and $\mathrm{OR}_i = 1$.

Note that odds ratios can be estimated from a case-control sample. For example, $\mathrm{OR}_i \sim$

$(C_i/G_i)/(C_1/G_1)$. Is is important to appreciate that the expected value for the odds ratios observed in a case-control sample is equal to the population odds ratios. Odds ratios can be used to quantify the risk of particular genotypes, without requiring an estimate of a disease's population incidence (which cannot be estimated from case control samples).

Parameterizing the difference between θ_C and θ_G follows from these properties. Conditional on θ_G and odds ratios for each genotype,

$$\begin{aligned} \theta_{C,i} &= \frac{\theta_{G_i}}{\theta_{G_1}} \times \theta_{C,1} \times \mathrm{OR}_i \quad \forall i \in \{2..n\} \\ \theta_{C,1} &= (1 + [\textstyle\sum_{i=2}^{n} \frac{\theta_{G_i}}{\theta_{G_1}} \times \mathrm{OR}_i])^{-1} . \end{aligned}$$

Finally, we can go on to express the odds ratios in terms of individual allelic contributions. For each genotype i, define an *allele count vector* z_i with A columns. $z_{i,j}$ specifies the number of copies of allele j carried by each individual with genotype i. It is common to specify odds ratios as log-additive contributions β of alleles (i.e. β is a vector with A columns, and the j-th column specifies the multiplicative increase in an individual's odds ratio with every copy of a j allele):

$$\mathrm{OR}_i := z_i * \beta_i .$$

Various biological risk mechanisms can be modeled by modifying z_i and β in this framework. For dominance, replace each 2 in z with a 1. For independent effects of homozygotes and heterozygotes, model them by independent columns in z. For interactions, introduce an additional column into z which is 1 only in the presence of the interacting factors.

We are now at a point where we can parameterize the likelihood function in terms of β and θ_G. θ_G specifies the frequency of each genotype, i.e. of all considered allelic combinations. If desired, we can instead specify the frequencies of the corresponding alleles in the control dataset, and use Hardy-Weinberg equilibrium to populate θ_G with the expected genotype frequencies – note however that this is only valid if the underlying loci are unlinked.

Specifying priors on β and on θ_G completes the Bayesian inferential framework. Note that the numbers of parameters to be fit is quite substantial in most realistic scenarios (for example, 11 parameters in a two-locus scenario with directly specified θ_G frequencies and a two-component β). θ_G is, in terms of testing for disease association, a *nuisance parameter*: its assumed values are usually not supposed to be informative about association status (although it is possible that there is actually information in the genotype frequencies themselves, for example signals of selection [Guan and Stephens, 2008]). Standard methods like Markov Chain Monte Carlo can be used to sample from the posterior distribution of all parameters.

The prospective likelihood Fitting the retrospective model can be cumbersome and is often complicated by the number of parameters. If it was possible to treat a retrospectively collected dataset in a *prospective* way, i.e. as though disease status had been determined a defined period of time after random genotyping of healthy individuals, the data could be modeled in terms of more easily handled logistic regression frameworks. In practice, this would eliminate the need to estimate the nuisance parameter θ_G.

As it turns out, it has long been known that retrospective and prospective analyses on retrospectively collected case-control datasets yield asymptotically identical results, at least in the frequentist context [Prentice and Pyke, 1979]. More recent works have shown that a very similar result holds for the Bayesian case, if Dirichlet priors are used for θ_G in the retrospective analysis and an (improper) uniform prior is used for the baseline odds of disease in the prospective analysis [Seaman and Richardson, 2004].

It should be noted that this result holds for retrospective analyses in which θ_G is specified on a per-genotype level, i.e. not in cases where assumptions of HWE are used to derive θ_G based on allele frequencies. However, it does hold for complex genetic modes of risk modification, for example interaction effects, as these can be modeled in terms of z_i.

Most genome-wide association studies treat their data in a prospective way and model disease risk in a logistic regression framework [Consortium, 2007; Stephens and Balding, 2009], and so does this thesis proceed. The results mentioned above justify this prac-

tice; however, it should be noted that they only hold asymptotically, and that differences between prospective and retrospective likelihoods may occur on medium-sized samples. Stephens and Balding [2009] present some simulation results and show that both approaches usually yield similar, though not identical, results.

Incorporating uncertainty To analyze HLA type imputation data, the imputation-derived uncertainty needs to be taken into account. To be specific, let $\mathrm{P}_j(z_i)$ denote the probability that individual j carries the allele count vector z_i. z_i can refer to alleles from HLA loci or SNP alleles, and the probability that a particular individual carries z_i can be computed from the posterior HLA type probabilities produced by HLA*IMP (which are, conditional on phasing, independent). Note that two columns k and l in z_i are *not* independent if they refer to alleles from the same locus.

In the notation of the previous section, the disease status of individual j under M_1 can now be modeled as

$$\begin{aligned}\mathrm{P}(\text{disease}_j|\mu,\beta) &:= \sum_{i=1}^{n}[\mathrm{P}(\text{disease}|z_i)\times \mathrm{P}_j(z_i)] \\ &= \sum_{i=1}^{n}[\frac{\exp(\mu+z_i*\beta)}{1+\exp(\mu+z_i*\beta)}\times \mathrm{P}_j(z_i)]\,.\end{aligned}$$

The log likelihood of the complete data D is then

$$\log \mathrm{P}(D|\mu,\beta) = \sum_{j=1}^{G}[\log(1-\mathrm{P}(\text{disease}_j|\mu,\beta))] + \sum_{j=G+1}^{G+C}[\log \mathrm{P}(\text{disease}_j|\mu,\beta)]\,.$$

Under M_0, β disappears or is reduced.

In practice, normal distributions are often used as prior distributions for the coefficients in prospective disease models; for example, N(0,1) for the baseline disease coefficient and N(0,0.2) for the components of β [Consortium, 2007]. In the context of studying the HLA, the variance for the priors for β could arguably be increased. Analogous to the retrospective case, modifications of z and β allow for modeling different genetic risk models

or interaction effects. In order to assess evidence that a *locus* contributes to disease risk (as opposed to single-allele estimates), it is possible to either aggregate alleles into classes (according to 2-digit HLA types, for example), and/or specify coefficients for multiple alleles (or allele groups) from one locus. In the latter case, it might be adequate to specify a prior which assigns risk effects to alleles in a non-independent way, but this option is not explored here.

It should be noted that it is not clear whether the exact asymptotic equivalence between prospective and retrospective Bayesian contexts holds for the uncertainty-aware likelihood. An alternative approach with well-understood properties would be to sample from the distribution of possible genotypes for all individuals, analyze each sample in a standard logistic regression framework, and average over the result.

6.1.2 Calculating a Bayes Factor

Calculating a Bayes Factor for model comparison purposes is not as trivial as may be initially expected. Generally speaking, three popular approaches are present in the literature:

- introduce a "model indicator" variable into an MCMC sampler, and the proportion of the time that the sampler spends in one model can immediately be used to calculate a Bayes Factor. However, because of the usually different parameter space of the two models, approaches like rjMCMC are necessary to obtain a properly defined Markov Chain. Also, assuring satisfactory mixing behaviour is not trivial; sometimes, it is recommended to pre-run a separate chain for each model, and use these results to inform a subsequent combined chain run. Han and Carlin [2001] reviews some of the approaches and arising issues. Whether the "model indicator" approaches are suited to a genome-wide association context, where one wants to examine many loci in a largely automated way, is not clear.

- importance sampling, with the posterior as importance distribution; the harmonic mean of the sampled likelihoods can then be used to estimate the marginal likelihood

[Newton and Raftery, 1994]. However, the harmonic mean sometimes exhibits stability problems, which then requires additional measures to quantify the uncertainty in the estimated marginal likelihood [Drummond and Rambaut, 2007].

- use the Laplace approximation to directly obtain a value for the Bayes Factor integral [Lewis and Raftery, 1997]. This approach was applied by large disease association study projects [Consortium, 2007] and is pursued here.

The quantity that we want to approximate is

$$f(\mu, \beta) := \int F(\mu, \beta)\, d\mu, \beta := \int \mathrm{P}(D|\mu, \beta) \times \Pr(\mu) \times \Pr(\beta)\, d\mu, \beta\,.$$

The Laplace approximation of this quantity is

$$f(\mu, \beta) \sim (2\pi)^{P/2} \times |\mathrm{H}(\mu^*, \beta^*)|^{1/2} \times F(\mu^*, \beta^*)\,,$$

where P is the combined dimensionality of μ and β, $\mathrm{H}(\mu^*, \beta^*)$ is minus the inverse Hessian of $\log F$ evaluated at (μ^*, β^*), and μ^* and β^* are the values at which $F(\mu, \beta)$ attains its maximum.

Taking logarithms, we have

$$\log f(\mu, \beta) \sim \frac{P}{2} \log(2\pi) + \frac{1}{2} log|\mathrm{H}(\mu^*, \beta^*)| + \log F(\mu^*, \beta^*)\,.$$

Note that this is the general form of M_1, but M_0 follows immediately from reducing the β parameter space.

To compute the Laplace approximation, we need a couple of quantities:

- the mode of $\log F$ – we can either maximize this function, using standard library routines from for example the GNU Scientific Library (GSL), or use MCMC to fit the model and use the output to estimate the mode.
- the Hessian $\mathrm{H}(\mu^*, \beta^*)$ of $\log F$ at the mode – see below.

The Laplace approximation becomes more accurate as more samples are added (in the notation of the previous section, the error term of the approximation is in $O((G+C)^{-1})$. For the approximation to be valid, we have to be able to find the global maximum of the likelihood function. In the uncertainty-aware setting presented here, this might sometimes be more difficult – ensuring convergence of the numerical optimization routines (from the GSL, for example) by starting from different points, or carefully inspecting the posterior parameter distributions for multiple possible maxima, is therefore highly advisable.

Calculating the Hessian of $\log F$ is straightforward calculus. For $x, y \in \mu, \beta_1..\beta..$, we have:

$$\begin{aligned} & \frac{\partial \log F}{\partial x} \\ = & \quad \frac{\partial}{\partial x} \log \Pr(\mu) + \frac{\partial}{\partial x} \log \Pr(\beta) \\ + & \quad \sum_{j=1}^{G} [\frac{\partial}{\partial x} \log(1 - (\mathrm{P}(\mathrm{disease}_j|\mu,\beta)))] \\ + & \quad \sum_{j=G+1}^{G+C} [\frac{\partial}{\partial x} (\log \mathrm{P}(\mathrm{disease}_j|\mu,\beta))] \end{aligned}$$

$$\frac{\partial}{\partial x} \log \mathrm{P}(\mathrm{disease}_j|\mu,\beta) = \frac{\frac{\partial}{\partial x}\mathrm{P}(\mathrm{disease}_j|\mu,\beta))}{\mathrm{P}(\mathrm{disease}_j|\mu,\beta))}$$

$$\frac{\partial}{\partial x} \log(1 - \mathrm{P}(\mathrm{disease}_j|\mu,\beta)) = \frac{-\frac{\partial}{\partial x}\mathrm{P}(\mathrm{disease}_j|\mu,\beta))}{1 - \mathrm{P}(\mathrm{disease}_j|\mu,\beta))}$$

$$\begin{aligned} & \frac{\partial^2}{\partial y \partial x} \log \mathrm{P}(\mathrm{disease}_j|\mu,\beta) \\ = & \quad \frac{[\frac{\partial^2}{\partial y \partial x}\mathrm{P}(\mathrm{disease}_j|\mu,\beta)] \times \mathrm{P}(\mathrm{disease}_j|\mu,\beta) - [\frac{\partial}{\partial x}\mathrm{P}(\mathrm{disease}_j|\mu,\beta)] \times [\frac{\partial}{\partial y}\mathrm{P}(\mathrm{disease}_j|\mu,\beta)]}{\mathrm{P}(\mathrm{disease}_j|\mu,\beta)^2} \end{aligned}$$

$$\frac{\partial^2}{\partial y \partial x} \log(1 - \mathrm{P}(\mathrm{disease}_j|\mu, \beta))$$

$$= \frac{[-\frac{\partial^2}{\partial y \partial x}\mathrm{P}(\mathrm{disease}_j|\mu,\beta)]\times[1-\mathrm{P}(\mathrm{disease}_j|\mu,\beta)]-[\frac{\partial}{\partial x}\mathrm{P}(\mathrm{disease}_j|\mu,\beta)]\times[\frac{\partial}{\partial y}\mathrm{P}(\mathrm{disease}_j|\mu,\beta)]}{[1-\mathrm{P}(\mathrm{disease}_j|\mu,\beta)]^2}$$

$$\frac{\partial}{\partial x}\mathrm{P}(\mathrm{disease}_j|\mu, \beta)$$

$$= \sum_{i=1}^{n} \frac{\frac{\partial}{\partial x}\exp(\mu+z_i*\beta)}{[1+\exp(\mu+z_i*\beta)]^2} \times \mathrm{P}_j(z_i)$$

$$\frac{\partial^2}{\partial y \partial x}\mathrm{P}(\mathrm{disease}_j|\mu, \beta)$$

$$= \sum_{i=1}^{n} \frac{[\frac{\partial^2}{\partial y \partial x}\exp(\mu+z_i*\beta)]\times[1+\exp(\mu+z_i*\beta)]-2[\frac{\partial}{\partial x}\exp(\mu+z_i*\beta)]\times[\frac{\partial}{\partial y}\exp(\mu+z_i*\beta)]}{[1+\exp(\mu+z_i*\beta)]^3}$$

$$\times \mathrm{P}_j(z_i)$$

Calculating the Laplace approximation of $\int F(\mu, \beta)\, d\mu, \beta$ follows immediately from these equations.

6.2 Power comparisons: a small simulation study

To evaluate whether the increased accuracy of HLA*IMP:02 can lead to any practical advantages in disease association studies, I present results from a limited simulation study.

6.2.1 Simulation of case-control datasets

The simulation study is based on 100 simulated case-control datasets of 2000 cases and 2000 controls, assuming a simple disease model: the allele HLADRB1*0401 influences disease risk with a log odds ratio of 2, and the allele HLAB*0702 is protective with a log odds ratio of -1 (in terms of absolute magnitude, these coefficients are well within the

range spanned by known HLA coefficients in autoimmune diseases like multiple sclerosis, ankylosing spondylitis and psoriasis). The two alleles were chosen on the basis that their allele-specific sensitivities are slightly higher under HLA*IMP:02. The probability of an individual carrying the disease is

$$\mathrm{P(disease)} := \frac{\exp(\mu + I_{\mathrm{HLAB*0702}} * (-1) + I_{\mathrm{HLADRB1*0401}} * 2)}{1 + \exp(\mu + I_{\mathrm{HLAB*0702}} * (-1) + I_{\mathrm{HLADRB1*0401}} * 2)},$$

where I_{allele} is a function counting the copies of allele *allele*. μ is arbitrarily fixed at 1.

The GS&GSK_EU 1/3 dataset (see Section 5.1.1) serves as a population that the disease phenotype is simulated into; note that each suitable population dataset has to comprise "true" (classically typed) HLA genotypes as well as imputation results from both models for each individual in order to enable the comparison between the two models.

The algorithm for simulating a case-control dataset is

1. Remove all individuals from GS&GSK_EU 1/3 which have missing data at *HLA-DRB1* or *HLA-B*; or which carry a 2-digit allele which is compatible with any of the two (4-digit) risk alleles. All remaining individuals have a defined disease risk according to our model.
2. Sample an individual i from GS&GSK_EU 1/3 (with replacement).
3. Simulate disease status of i, according to its "true" HLA genotype.
4. If i is a case and there are less than 2000 stored cases, store corresponding imputation results from HLA*IMP:01 and HLA*IMP:02 in the case group.
5. If i is a control and there are less than 2000 stored controls, store corresponding imputation results from HLA*IMP:01 and HLA*IMP:02 in the control group.
6. If there are less than 2000 cases or less than 2000 controls, go to step 2; otherwise terminate.

6.2.2 Evaluation

Each of the 100 iterations comprises one case-control group with HLA*IMP:01 imputations and one case-control group with HLA*IMP:02 imputations (the individuals in both groups and their disease status are identical). We use the methods from the first part of the chapter to fit the coefficients of a prospective logistic regression model based on the two known risk alleles (acting additively; N(0,1) priors for all coefficients) and calculate an uncertainty-aware Bayes Factor (BF), for imputations from both models. For each iteration, we store the two BFs and the means of the estimated coefficients.

6.2.3 Results

	HLA*IMP:01	**HLA*IMP:02**	**Truth**
HLADRB1*0401 (mean coefficient)	1.63	1.93	2
HLAB*0702 (mean coefficient)	-0.98	-0.97	-1
log10(BF)	100.27	118.81	-

Table 6.1: Results from 100 simulated case-control datasets, comprising 2000 cases and 2000 controls, based on GS&GSK_EU 1/3. A disease phenotype based on HLADRB1*0401 and HLAB*0702 is simulated according to classically typed HLA genotypes, and logistic regression models for the causal alleles are fitted based on the imputations from both models. The values are averaged over 100 iterations.

Table 6.1 summarizes the results. HLA*IMP:02 clearly outperforms HLA*IMP:01, in terms of the achieved Bayes Factor (by 18 orders of magnitude) as well as in terms of the estimated coefficient for the risk-increasing allele. The estimated coefficient for the protective allele is relatively similar under both models.

The simulation study presented here is limited: most importantly in that the causal alleles were selected with prior knowledge on the superior performance of HLA*IMP:02 on these alleles; and in that only the model comprising the causal alleles was fitted, and no other models were considered (as would be the case when analyzing non-simulated data).

Elsewhere in this thesis, I have discussed the relationship between power and imputation accuracy (see page 32): power increases as imputations become more accurate. The results from the small study presented here are consistent with this expectation. Together with the validation results presented in previous chapters they suggest that using HLA*IMP:02

instead of HLA*IMP:01 can be expected to lead to increased power in GWAS.

6.3 Real data case study: MS

HLA*IMP:01 has been applied in three recent WTCCC2 case studies: on MS [Sawcer et al., 2011], AS [Evans et al., 2011], and psoriasis [Strange et al., 2010].

My role in these studies entailed producing HLA type imputations and interpreting signals from the HLA region in a Bayesian framework. MS was the study that I was most involved in. Therefore I present a sketch of the Bayesian analysis of the observed signals from the xMHC in the UK cohort (after it had become clear that we would only publish frequentist results, we decided not to replicate the full frequentist analysis, and therefore didn't go on to apply methods to correct for population stratification in the Bayesian context).

I will also discuss findings from the frequentist analysis and the biological implications of our results.

6.3.1 Methods

Different models of association were assessed by calculating BFs as previously described. N(0, 1) priors were used for all coefficients. Qualitatively identical results were obtained when varying the variance parameter of the priors.

To discover associations, we applied a stepwise model building algorithm: starting from the simplest model (i.e. no associated locus), we tested the effect of the remaining available alleles (HLA alleles and SNPs genotypes from the xMHC region), conditional on the already included alleles, and included the most strongly associated remaining allele (sometimes including biological considerations and evidence from other cohorts). At each step, we assessed the evidence for interactions between the included alleles and deviations from additivity; however, all the results we obtained were consistent with additive models without interaction effects.

In the process of analysing associations in the xMHC region, rigorous quality control

turned out to be essential: we discovered abundant SNP genotyping problems, as evident from cluster plots.

6.3.2 Results

HLA-DRB1*15:01 is the allele most strongly associated with MS in the xMHC region (log10(BF) ∼116.6, mean posterior log odds ratio ∼1.09). HLA-DQB1*06:02 and HLA-DQA1*01:02 achieve log10(BF) values of ∼108 and ∼89, and both alleles are known to frequently occur on the same haplotype as DRB1*15:01. The strongest SNP signal comes from rs9271366 (log10(BF) ∼ 116.0). All of these signals disappear upon conditioning on HLA-DRB1*15:01.

The first conditioning step reveals an independent peak in the *HLA-A* region. HLA-A*02:01 achieves a BF of ∼13.5, and the most strongly associated SNP (rs38233355) achieves a BF of ∼14.0. The log odds ratios are estimated to be around -0.4 in both cases (posterior mean). As A*02:01 exhibited stronger association than the SNP in the combined dataset (consistent with frequentist analyses; data now shown), we decided to include A*02:01 as next allele.

A last step of conditioning shows an effect of HLA-DQB1*02:01 and HLA-DRB1*03:01 (strongly linked and very similar coefficients; log10(BF) ∼8.9 for DQB1*02:01; posterior mean of the log odds ratio ∼ -0.36; results based on allowing deviations from additivity for A*02:01).

6.3.3 Further frequentist results

As our frequentist analysis was based on the same uncertainty-aware likelihood that was discussed earlier, it seems appropriate to discuss some of the additional results in this context (compare Moutsianas [2011], as the frequentist analysis was carried out by Loukas Moutsianas, and Sawcer et al. [2011] for a more compact presentation).

First, the results from the frequentist analysis are entirely consistent with the Bayesian results presented so far. In the frequentist framework, evidence from different cohorts

could be combined in a fixed effect meta-analysis setting. This showed that the secondary effect is most likely to be driven by A*02:01.

Also, we were not able to reliably disentangle the effects from HLA-DQB1*02:01 and HLA-DRB1*03:01. That is, our findings are consistent with a disease model in which the class II effect is entirely driven by alleles at *HLA-DRB1*; however, we cannot reject a model which involves *HLA-DQB1* either.

Finally, there is an independent significant effect from a SNP in the *HLA-DPB1* region. It is not clear whether this is driven by LD with an *HLA-DPB1* allele, or whether this constitutes an effect independent of classical alleles.

6.3.4 Biological implications

Pursuing an imputation-driven HLA analysis approach has led to biologically relevant results.

First, we were able to confirm the existence of secondary risk-increasing effects from the class II region. An interesting follow-up experiment would be to examine the presentability of Myelin Basic Protein (MBP) on these, as MBP is the putative auto-antigen in MS [Oksenberg et al., 2008].

Furthermore, our study has produced strong evidence for a protective effect exerted by A*02:01. This renders likely an involvement of CD8+ T cells in the pathogenesis of MS, as CD8+ T cells exclusively react to antigens presented on HLA class I molecules like HLA-A.

It is also tempting to speculate that there might be a connection between the protective effect of A*02:01 and the Epstein-Barr-Virus (EBV). There is evidence that EBV might play a role in the pathogenesis of MS [Oksenberg et al., 2008]; specifically, it has been shown that there are T cell receptors which recognize both EBV and MBP fragments under particular circumstances [Lang et al., 2002]. Intriguingly, A*02:01 is protective against EBV-positive Hodgkin's lymphoma [Hjalgrim et al., 2010]. An working hypothesis could therefore be that A*02:01 is particularly effective at presenting EBV peptides; testing this

hypothesis experimentally seems at least not impossible.

Chapter 7

Discussion

7.1 Developments and limitations

HLA genotyping is an important task in many areas of biomedical research. In this thesis, I have presented extended and novel methods which allow for accurate statistical imputation of HLA types, based on SNP genotype data.

The development of HLA-specific genotype imputation methods was initially motivated by earlier experiments that had demonstrated that the Li & Stephens approximation did not perform well in the HLA region (see page 61 and Moutsianas [2011, p. 194]). These experiments were evaluated at the level of 4-digit HLA types, which each refer to a number of possible underlying chromosomal sequences. That is, the accuracy measure applied was more generous than in normal SNP genotype imputation, and results measured at the SNP level (DNA sequence at the relevant positions) would necessarily have been worse.

The LDMhc algorithm [Leslie et al., 2008] first demonstrated that the Li & Stephens approximation can be useful in HLA type imputation, at least if the model is configured in a way that allows for a more effective capturing of longer-range LD relationships. My contributions to HLA type imputation started when it became necessary to apply LDMhc to large datasets in an efficient and reliable way and resulted in HLA*IMP:01. Combining multiple populations then requires a model than can deal with multiple SNP haplotype

backgrounds for individual alleles, which inspired the development of HLA*IMP:02.

HLA*IMP:01 is based on an existing model, LDMhc, but was modified to yield higher call rates and was implemented in a parallelized framework. The modified algorithm can therefore deal with large reference panels. I have shown that HLA*IMP:01 produces accurate HLA type imputations when applied to a large homogeneous reference panel (in this thesis, comprising British and central European samples).

HLA*IMP:02 implements a novel HLA type imputation method described in this thesis. It is based on a probabilistic generalization of haplotype graph construction algorithms, which probabilistically as opposed to deterministically attaches haplotypes to nodes while building the graph. It allows for uncertainty and missing data in the set of haplotype estimates and "localization" to the known long-range LD patterns of the MHC region. I have shown that HLA*IMP:02 is as accurate as HLA*IMP:01 on homogeneous reference panels and that it clearly outperforms HLA*IMP:01 on multi-country medium heterogeneity reference panels. I have also presented some evidence that the model of HLA*IMP:02 may be suited to deal with multi-ethnicity reference panels, but the available data is too limited to allow for definitive conclusions. Returning to the scenarios I have presented in the introduction, Scenario 1 can be regarded as solved; Scenario 2 can be regarded as solved; Scenario 3 remains problematic, but the data I have presented allows for cautious optimism that HLA*IMP:02 may be up for the task, if a suitable reference panel can be assembled.

From a practical perspective, having developed a model to effectively deal with Scenario 2 is probably the most important result of this thesis: it allows for more accurate HLA type imputations in multi-country genome-wide association studies, which will become increasingly more important.

HLA*IMP:02 has a couple of other advantages over HLA*IMP:01: it is highly tolerant of missing data and does not require phased input data. There is some limited ability to detect errors in the reference panel. It does not require SNP selection, and one graph, built including all available reference data, can be used for all imputation datasets.

From a slightly more theoretical perspective, this thesis provides two main developments: a

novel, more probabilistic way to construct haplotype representation models and the insight that so-constructed models can be used to capture the MHC's haplotypic complexity over multiple populations, including structural variants.

7.2 Lessons learnt from HLA type imputation

HLA type imputation has already had a measurable impact on our understanding of the pathogenesis of important diseases. As described in Section 6.3.4, our understanding of the genetic variants influencing MS risk has improved through the application of HLA*IMP:01, and our findings inspire some interesting and far-reaching questions regarding the role of CD8+ T cells and EBV.

HLA*IMP:01 did also have a role in characterizing the recently discovered statistical interaction between HLA alleles and the ERAP1 protein in psoriasis and AS (although the initial discovery was not driven by HLA*IMP:01).

More recently, another imputation-based study has been able to link the MHC-driven risk in rheumatoid arthritis to particular amino acid residues in the HLA proteins [Raychaudhuri et al., 2012].

It is certainly the case that our functional understanding of the general workings of the immune system (parts of which I tried to expose in the second chapter) is still much more advanced than our understanding of the specific parts that go wrong in autoimmune diseases; nevertheless, important progress has been made recently, and HLA type imputation is part of that story.

7.2.1 Limitations

The work presented here is limited in a variety of ways. First, it would be desirable to carry out further high heterogeneity experiments. This is subject to new datasets becoming available. Second, the theoretical properties of the model presented here are not well understood: it is not clear whether the APFA learnability proof presented by

Ron et al. [1998] holds for the error-aware construction algorithm (see page 54 for a short discussion on applicability to haplotype graph models). The population genetics properties of haplotype graphs in general are not well understood, except for the fact that they provide an effective means to represent haplotypes. This extends to the novel algorithms presented in this thesis. In particular, a more formally justified method to fit the model's parameters would be valuable.

7.3 Possible future applications of haplotype graph models

HLA*IMP:02 is a specialized model which performs well in an important, but restricted region in the human genome. How well would HLA*IMP:02 perform if it was applied in other areas of the human genome, or to other kinds of genetic variation?

HLA*IMP:02 has three main distinctive attributes, which probably contribute to its good imputation performance in the HLA: it can accommodate long-range linkage structures, in particular if being localized to particular loci of interest; it can deal with heterogeneity in the reference panel; and it allows for geno- and haplotype error while building the graph. These features, however, come at a computational cost, to an extent that the model seems less well-suited to deal with hundreds of thousands of individual SNPs.

Natural targets for HLA*IMP:02 therefore share some of the HLA's attributes: an unconventional haplotype structure, genotype uncertainty and an investigator's elevated interest in a (limited) set of predefined loci. The KIR region of the human genome, also heavily involved in regulating immune responses, meets these criteria and provides an interesting test case for HLA*IMP:02 (less alleles, but more genes and more extensive copy number variation than in the HLA) – a collaboration with the Immunochip consortium to develop a KIR imputation technology is already under way.

As our knowledge of haplotype structure and variation in the human genome increases, it seems entirely possible that more regions which resist standard approaches to imputation are discovered; applying HLA*IMP:02 to these would be an interesting exercise.

7.4 The impact of sequencing

With next-generation sequencing technologies becoming more affordable and better (for example in terms of read length), the long-term future of SNP genotyping and SNP-based imputation seems uncertain.

There are at least two plausible scenarios: Sequencing may come to dominate the realm of biomedical research to an extent that it becomes the world's favourite and unrivaled genotyping technology. Imputation of complex attributes then becomes unnecessary, because their genotypes are readily available from the read data.

In the other scenario, SNP genotyping remains more cost-effective than sequencing by at least a factor of 10. Although competing for resources in some areas, important synergistic effects develop between the two technologies: high-quality long reads are used to make imputation more effective, making it possible to impute most common and many rare variants with a high degree of accuracy.

One important, and unresolved, factor in evaluating these scenarios is the relative importance of rare variants versus interactions between common variants. If interactions of common variants are going to become a focal point of research, genotyping ten times more samples at an imputation-induced modest loss of genotype accuracy is probably an attractive option, leading to increased power to detect small-effect interactions. If, on the other hand, rare variants turn out to be all-important, the prospects for imputation technology are rather bleak.

Without doubt, next-generation sequencing projects will lead to a massive increase in the number of known human polymorphisms, and then to the question of how this catalogue of variation can be useful in improving data analysis. At some point, the notion of a single human reference genome will be substituted by another data structure, which will enable the coherent representation of different types of variants – for example, a SNP only occurring in the sequence of a particular insertion, or a polymorphic deletion in an insertion. Interestingly, haplotype graph models seem well-suited to deal with these challenges: by allowing for the emission of "gap" symbols, it is straightforward to construct

a “dense” haplotype graph model (referring to the property that every haplotype is a walk through the graph) from a multiple sequence alignment. By modifying the criteria of when to collapse nodes, one can influence the complexity of the graph and the extent to which it preserves LD (a haplotype graph may lose linkage information, but the algorithms described in this thesis can be easily modified to make sure that no identified variant is ever lost in the construction process). The most extreme (though clearly intuitive) manifestation of this idea is the “variation graph”, in which LD is collapsed as soon as possible, only retaining the local LD information that is necessary to coherently represent complex variants.

7.5 Final frontiers – clinical applications

Transplantation medicine, without doubt the most important field of application for classical HLA typing, has been carefully avoided in this thesis. Why is this? It is clear that the reported accuracies are not high enough to inform a final decision on whether two people are HLA-compatible or not. This, however, is often only the last step in a series of preliminary examinations. The NMDP (the national bone marrow donor registry of the United States), for example, has a database of >9 million registered potential donors, only a small fraction of which will ever advance to the stage of being considered as a potential match for a patient needing a bone marrow transplant. By carefully curating and extending the available reference panels, it should be possible to further increase the accuracy of HLA type imputation. This could, in combination with a cost-effective dense MHC SNP genotyping chip, lead to an HLA genotyping technology which is both accurate enough for preliminary screening of potential donors (for an NMDP database entry, say) and less expensive than classical typing technologies, maybe by up to a factor of 5 or 10. This would facilitate the further expansion of donor databases, delivering a real benefit to public health.

Bibliography

A. Albrechtsen, I. Moltke, and R. Nielsen. Natural selection and the distribution of identity-by-descent in the human genome. *Genetics*, 186(1):295–308, 2010.

A. Alcami and U. H. Koszinowski. Viral mechanisms of immune evasion. *Trends Microbiol*, 8(9):410–8, 2000.

D. Altshuler, M. J. Daly, and E. S. Lander. Genetic mapping in human disease. *Science*, 322(5903):881–8, 2008.

J. Amiel. *Study of leucocyte phenotypes in Hodgkin's disease*, pages 79–81. Munksgaard, Copenhagen, 1967.

L. Andersson and S. Mikko. Generation of MHC class II diversity by intra- and intergenic recombination. *Immunol Rev*, 143:5–12, 1995.

A. M. Andres, M. J. Hubisz, A. Indap, D. G. Torgerson, J. D. Degenhardt, A. R. Boyko, R. N. Gutenkunst, T. J. White, E. D. Green, C. D. Bustamante, A. G. Clark, and R. Nielsen. Targets of balancing selection in the human genome. *Mol Biol Evol*, 26(12):2755–64, 2009.

T. F. Bergstrom, A. Josefsson, H. A. Erlich, and U. Gyllensten. Recent origin of HLA-DRB1 alleles and implications for human evolution. *Nat Genet*, 18(3):237–42, 1998.

J. M. Blackwell, S. E. Jamieson, and D. Burgner. HLA and infectious diseases. *Clin Microbiol Rev*, 22(2):370–85, Table of Contents, 2009.

W. Bodmer and C. Bonilla. Common and rare variants in multifactorial susceptibility to common diseases. *Nat Genet*, 40(6):695–701, 2008.

W. F. Bodmer. Genetic factors in Hodgkin's disease: association with a disease-susceptibility locus (DSA) in the HL-A region. *Natl Cancer Inst Monogr*, 36:127–34, 1973.

R. M. Brennan and S. R. Burrows. A mechanism for the HLA-A*01-associated risk for EBV+ Hodgkin lymphoma and infectious mononucleosis. *Blood*, 112(6):2589–90, 2008.

B. L. Browning and S. R. Browning. A unified approach to genotype imputation and haplotype-phase inference for large data sets of trios and unrelated individuals. *Am J Hum Genet*, 84(2):210–23, 2009.

B. L. Browning and Z. Yu. Simultaneous genotype calling and haplotype phasing improves genotype accuracy and reduces false-positive associations for genome-wide association studies. *Am J Hum Genet*, 85(6):847–61, 2009.

S. R. Browning. Multilocus association mapping using variable-length Markov chains. *Am J Hum Genet*, 78(6):903–13, 2006.

S. R. Browning and B. L. Browning. Rapid and accurate haplotype phasing and missing-data inference for whole-genome association studies by use of localized haplotype clustering. *Am J Hum Genet*, 81(5):1084–97, 2007.

S. R. Browning and B. L. Browning. Haplotype phasing: existing methods and new developments. *Nat Rev Genet*, 12(10):703–14, 2011.

K. Cao, J. Hollenbach, X. Shi, W. Shi, M. Chopek, and M. A. Fernandez-Vina. Analysis of the frequencies of HLA-A, B, and C alleles and haplotypes in the five major ethnic groups of the United States reveals high levels of diversity in these loci and contrasting distribution patterns in these populations. *Hum Immunol*, 62(9):1009–30, 2001.

K. Cao, A. M. Moormann, K. E. Lyke, C. Masaberg, O. P. Sumba, O. K. Doumbo, D. Koech, A. Lancaster, M. Nelson, D. Meyer, R. Single, R. J. Hartzman, C. V. Plowe, J. Kazura, D. L. Mann, M. B. Sztein, G. Thomson, and M. A. Fernandez-Vina. Differentiation between African populations is evidenced by the diversity of alleles and haplotypes of HLA class I loci. *Tissue Antigens*, 63(4):293–325, 2004.

N. Cereb, A. L. Hughes, and S. Y. Yang. Locus-specific conservation of the HLA class I introns by intra-locus homogenization. *Immunogenetics*, 47(1):30–6, 1997.

J. M. Chapman, J. D. Cooper, J. A. Todd, and D. G. Clayton. Detecting disease associations due to linkage disequilibrium using haplotype tags: a class of tests and the determinants of statistical power. *Hum Hered*, 56(1-3):18–31, 2003.

H. Y. Chen, G. Hayashi, O. Y. Lai, A. Dilthey, P. J. Kuebler, T. V. Wong, M. P. Martin, M. A. F. Vina, G. McVean, M. Wabl, K. S. Leslie, T. Maurer, J. N. Martin, S. G. Deeks, M. Carrington, A. M. Bowcock, D. F. Nixon, and W. Liao. Psoriasis patients are enriched for genetic variants that protect against HIV-1 disease. *Plos Genetics*, 8(2), 2012.

W. H. Chung, S. I. Hung, and Y. T. Chen. Human leukocyte antigens and drug hypersensitivity. *Curr Opin Allergy Clin Immunol*, 7(4):317–23, 2007.

D. A. Compston, J. R. Batchelor, and W. I. McDonald. B-lymphocyte alloantigens associated with multiple sclerosis. *Lancet*, 2(7998):1261–5, 1976.

International HapMap Consortium, K. A. Frazer, D. G. Ballinger, D. R. Cox, D. A. Hinds, L. L. Stuve, R. A. Gibbs, J. W. Belmont, A. Boudreau, P. Hardenbol, S. M. Leal, S. Pasternak, D. A. Wheeler, T. D. Willis, F. Yu, H. Yang, C. Zeng, Y. Gao, H. Hu, W. Hu, C. Li, W. Lin, S. Liu, H. Pan, X. Tang, J. Wang, W. Wang, J. Yu, B. Zhang, Q. Zhang, H. Zhao, H. Zhao, J. Zhou, S. B. Gabriel, R. Barry, B. Blumenstiel, A. Camargo, M. Defelice, M. Faggart, M. Goyette, S. Gupta, J. Moore, H. Nguyen, R. C. Onofrio, M. Parkin, J. Roy, E. Stahl, E. Winchester, L. Ziaugra, D. Altshuler, Y. Shen, Z. Yao, W. Huang, X. Chu, Y. He, L. Jin, Y. Liu, Y. Shen, W. Sun, H. Wang, Y. Wang, Y. Wang, X. Xiong, L. Xu, M. M. Waye, S. K. Tsui, H. Xue, J. T. Wong, L. M. Galver, J. B. Fan, K. Gunderson, S. S. Murray, A. R. Oliphant, M. S. Chee, A. Montpetit, F. Chagnon, V. Ferretti, M. Leboeuf, J. F. Olivier, M. S. Phillips, S. Roumy, C. Sallee, A. Verner, T. J. Hudson, P. Y. Kwok, D. Cai, D. C. Koboldt, R. D. Miller, L. Pawlikowska, P. Taillon-Miller, M. Xiao, L. C. Tsui, W. Mak, Y. Q. Song, P. K. Tam, Y. Nakamura, T. Kawaguchi, T. Kitamoto, T. Morizono, A. Nagashima, et al. A second

generation human haplotype map of over 3.1 million SNPs. *Nature*, 449(7164):851–61, 2007.

Wellcome Trust Case Control Consortium. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature*, 447(7145):661–78, 2007.

R. W. Davies, G. A. Wells, A. F. Stewart, J. Erdmann, S. H. Shah, J. F. Ferguson, A. S. Hall, S. S. Anand, M. S. Burnett, S. E. Epstein, S. Dandona, L. Chen, J. Nahrstaedt, C. Loley, I. R. Konig, W. E. Krauss, C. B. Granger, J. C. Engert, C. Hengstenberg, H. E. Wichmann, S. Schreiber, W. H. Tang, S. G. Ellis, D. J. Rader, S. L. Hazen, M. P. Reilly, N. J. Samani, H. Schunkert, R. Roberts, and R. McPherson. A genome wide association study for coronary artery disease identifies a novel susceptibility locus in the major histocompatibility complex. *Circ Cardiovasc Genet*, 2012.

P. I. de Bakker, G. McVean, P. C. Sabeti, M. M. Miretti, T. Green, J. Marchini, X. Ke, A. J. Monsuur, P. Whittaker, M. Delgado, J. Morrison, A. Richardson, E. C. Walsh, X. Gao, L. Galver, J. Hart, D. A. Hafler, M. Pericak-Vance, J. A. Todd, M. J. Daly, J. Trowsdale, C. Wijmenga, T. J. Vyse, S. Beck, S. S. Murray, M. Carrington, S. Gregory, P. Deloukas, and J. D. Rioux. A high-resolution HLA and SNP haplotype map for disease association studies in the extended human MHC. *Nat Genet*, 38(10):1166–72, 2006.

C. M. Deighton, D. J. Walker, I. D. Griffiths, and D. F. Roberts. The contribution of HLA to rheumatoid arthritis. *Clin Genet*, 36(3):178–82, 1989.

A. T. Dilthey, L. Moutsianas, S. Leslie, and G. McVean. HLA*IMP – an integrated framework for imputing classical HLA alleles from SNP genotypes. *Bioinformatics*, 27(7):968–72, 2011.

A. J. Drummond and A. Rambaut. BEAST: Bayesian evolutionary analysis by sampling trees. *BMC Evol Biol*, 7:214, 2007.

P. P. Dunn. Human leucocyte antigen typing: techniques and technology, a critical appraisal. *Int J Immunogenet*, 38(6):463–73, 2011.

D. M. Evans, C. C. Spencer, J. J. Pointon, Z. Su, D. Harvey, G. Kochan, U. Opperman, A. Dilthey, M. Pirinen, M. A. Stone, L. Appleton, L. Moutsianis, S. Leslie,

T. Wordsworth, T. J. Kenna, T. Karaderi, G. P. Thomas, M. M. Ward, M. H. Weisman, C. Farrar, L. A. Bradbury, P. Danoy, R. D. Inman, W. Maksymowych, D. Gladman, P. Rahman, A. Morgan, H. Marzo-Ortega, P. Bowness, K. Gaffney, J. S. Gaston, M. Smith, J. Bruges-Armas, A. R. Couto, R. Sorrentino, F. Paladini, M. A. Ferreira, H. Xu, Y. Liu, L. Jiang, C. Lopez-Larrea, R. Diaz-Pena, A. Lopez-Vazquez, T. Zayats, G. Band, C. Bellenguez, H. Blackburn, J. M. Blackwell, E. Bramon, S. J. Bumpstead, J. P. Casas, A. Corvin, N. Craddock, P. Deloukas, S. Dronov, A. Duncanson, S. Edkins, C. Freeman, M. Gillman, E. Gray, R. Gwilliam, N. Hammond, S. E. Hunt, J. Jankowski, A. Jayakumar, C. Langford, J. Liddle, H. S. Markus, C. G. Mathew, O. T. McCann, M. I. McCarthy, C. N. Palmer, L. Peltonen, R. Plomin, S. C. Potter, A. Rautanen, R. Ravindrarajah, M. Ricketts, N. Samani, S. J. Sawcer, A. Strange, R. C. Trembath, A. C. Viswanathan, M. Waller, P. Weston, P. Whittaker, S. Widaa, N. W. Wood, G. McVean, J. D. Reveille, B. P. Wordsworth, M. A. Brown, and P. Donnelly. Interaction between ERAP1 and HLA-B27 in ankylosing spondylitis implicates peptide handling in the mechanism for HLA-B27 in disease susceptibility. *Nat Genet*, 43(8):761–767, 2011.

M. F. Flajnik and M. Kasahara. Comparative genomics of the MHC: glimpses into the evolution of the adaptive immune system. *Immunity*, 15(3):351–62, 2001.

M. F. Flajnik and M. Kasahara. Origin and evolution of the adaptive immune system: genetic events and selective pressures. *Nat Rev Genet*, 11(1):47–59, 2010.

G. Froeschke and S. Sommer. MHC class II DRB variability and parasite load in the striped mouse (Rhabdomys pumilio) in the Southern Kalahari. *Mol Biol Evol*, 22(5):1254–9, 2005.

E. Gambineri, T. R. Torgerson, and H. D. Ochs. Immune dysregulation, polyendocrinopathy, enteropathy, and X-linked inheritance (IPEX), a syndrome of systemic autoimmunity caused by mutations of FOXP3, a critical regulator of T-cell homeostasis. *Curr Opin Rheumatol*, 15(4):430–5, 2003.

J. Gay, S. Myers, and G. McVean. Estimating meiotic gene conversion rates from population genetic data. *Genetics*, 177(2):881–94, 2007.

J. A. Gebe, E. Swanson, and W. W. Kwok. HLA class II peptide-binding and autoimmunity. *Tissue Antigens*, 59(2):78–87, 2002.

L. H. Glimcher and C. J. Kara. Sequences and factors: a guide to MHC class-II transcription. *Annu Rev Immunol*, 10:13–49, 1992.

J. W. Gregersen, K. R. Kranc, X. Ke, P. Svendsen, L. S. Madsen, A. R. Thomsen, L. R. Cardon, J. I. Bell, and L. Fugger. Functional epistasis on a common MHC haplotype associated with multiple sclerosis. *Nature*, 443(7111):574–7, 2006.

R. C. Griffiths and P. Marjoram. Ancestral inference from samples of DNA sequences with recombination. *J Comput Biol*, 3(4):479–502, 1996.

Y. Guan and M. Stephens. Practical issues in imputation-based association mapping. *PLoS Genet*, 4(12):e1000279, 2008.

C. Han and B. P. Carlin. Markov chain Monte Carlo methods for computing Bayes factors: A comparative review. *Journal of the American Statistical Association*, 96(455):1122–1132, 2001.

P. W. Hedrick. Pathogen resistance and genetic variation at MHC loci. *Evolution*, 56(10):1902–8, 2002.

P. W. Hedrick and G. Thomson. Evidence for balancing selection at HLA. *Genetics*, 104(3):449–56, 1983.

S. M. Hedrick. The acquired immune system: a vantage from beneath. *Immunity*, 21(5):607–15, 2004.

S. Hetherington, A. R. Hughes, M. Mosteller, D. Shortino, K. L. Baker, W. Spreen, E. Lai, K. Davies, A. Handley, D. J. Dow, M. E. Fling, M. Stocum, C. Bowman, L. M. Thurmond, and A. D. Roses. Genetic variations in HLA-B region and hypersensitivity reactions to abacavir. *Lancet*, 359(9312):1121–2, 2002.

A. V. Hill. The genomics and genetics of human infectious disease susceptibility. *Annu Rev Genomics Hum Genet*, 2:373–400, 2001.

H. Hjalgrim, K. Rostgaard, P. C. Johnson, A. Lake, L. Shield, A. M. Little, K. Ekstrom-Smedby, H. O. Adami, B. Glimelius, S. Hamilton-Dutoit, E. Kane, G. M. Taylor, A. McConnachie, L. P. Ryder, C. Sundstrom, P. S. Andersen, E. T. Chang, F. E. Alexander, M. Melbye, and R. F. Jarrett. HLA-A alleles and infectious mononucleosis suggest a critical role for cytotoxic T-cell response in EBV-related Hodgkin lymphoma. *Proc Natl Acad Sci U S A*, 107(14):6400–5, 2010.

K. Hogstrand and J. Bohme. Gene conversion can create new MHC alleles. *Immunol Rev*, 167:305–17, 1999.

R. Holdsworth, C. K. Hurley, S. G. Marsh, M. Lau, H. J. Noreen, J. H. 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(2):95–170, 2009.

R. Horton, R. Gibson, P. Coggill, M. Miretti, R. J. Allcock, J. Almeida, S. Forbes, J. G. Gilbert, K. Halls, J. L. Harrow, E. Hart, K. Howe, D. K. Jackson, S. Palmer, A. N. Roberts, S. Sims, C. A. Stewart, J. A. Traherne, S. Trevanion, L. Wilming, J. Rogers, P. J. de Jong, J. F. Elliott, S. Sawcer, J. A. Todd, J. Trowsdale, and S. Beck. Variation analysis and gene annotation of eight MHC haplotypes: the MHC Haplotype Project. *Immunogenetics*, 60(1):1–18, 2008.

R. Horton, L. Wilming, V. Rand, R. C. Lovering, E. A. Bruford, V. K. Khodiyar, M. J. Lush, S. Povey, Jr. Talbot, C. C., M. W. Wright, H. M. Wain, J. Trowsdale, A. Ziegler, and S. Beck. Gene map of the extended human MHC. *Nat Rev Genet*, 5(12):889–99, 2004.

F. J. Hosking, S. Leslie, A. Dilthey, L. Moutsianas, Y. Wang, S. E. Dobbins, E. Papaemmanuil, E. Sheridan, S. E. Kinsey, T. Lightfoot, E. Roman, J. A. Irving, J. M. Allan, M. Taylor, M. Greaves, G. McVean, and R. S. Houlston. MHC variation and risk of childhood B-cell precursor acute lymphoblastic leukemia. *Blood*, 117(5):1633–40, 2011.

B. N. Howie, P. Donnelly, and J. Marchini. A flexible and accurate genotype imputa-

tion method for the next generation of genome-wide association studies. *PLoS Genet*, 5(6):e1000529, 2009.

A. L. Hughes. Natural selection and the diversification of vertebrate immune effectors. *Immunol Rev*, 190:161–8, 2002.

A. L. Hughes and M. Nei. Pattern of nucleotide substitution at major histocompatibility complex class I loci reveals overdominant selection. *Nature*, 335(6186):167–70, 1988.

A. L. Hughes and M. Nei. Nucleotide substitution at major histocompatibility complex class II loci: evidence for overdominant selection. *Proc Natl Acad Sci U S A*, 86(3):958–62, 1989.

Charles Janeway. *Immunobiology.The immune system in health and disease.* Garland Pub., New York, 5th edition, 2001. [electronic resource] : Charles A. Janeway, Jr. ... [et al.]. Title from caption (viewed Apr. 25, 2006) Also issued in print. Mode of access: World Wide Web.

J. Kelley, L. Walter, and J. Trowsdale. Comparative genomics of major histocompatibility complexes. *Immunogenetics*, 56(10):683–95, 2005.

J. F. Kingman. Origins of the coalescent. 1974-1982. *Genetics*, 156(4):1461–3, 2000.

J. M. Korn, F. G. Kuruvilla, S. A. McCarroll, A. Wysoker, J. Nemesh, S. Cawley, E. Hubbell, J. Veitch, P. J. Collins, K. Darvishi, C. Lee, M. M. Nizzari, S. B. Gabriel, S. Purcell, M. J. Daly, and D. Altshuler. Integrated genotype calling and association analysis of SNPs, common copy number polymorphisms and rare CNVs. *Nat Genet*, 40(10):1253–60, 2008.

A. Kumanovics, T. Takada, and K. F. Lindahl. Genomic organization of the mammalian MHC. *Annu Rev Immunol*, 21:629–57, 2003.

H. L. Lang, H. Jacobsen, S. Ikemizu, C. Andersson, K. Harlos, L. Madsen, P. Hjorth, L. Sondergaard, A. Svejgaard, K. Wucherpfennig, D. I. Stuart, J. I. Bell, E. Y. Jones, and L. Fugger. A functional and structural basis for TCR cross-reactivity in multiple sclerosis. *Nat Immunol*, 3(10):940–3, 2002.

A. M. Lazaro, J. Henry, J. Ng, C. K. Hurley, and P. E. Posch. Increased HLA class I and II diversity as 72 novel alleles are identified in volunteers for the National Marrow Donor Program Registry in 2010. *Tissue Antigens*, 2011.

A. J. Leslie, K. J. Pfafferott, P. Chetty, R. Draenert, M. M. Addo, M. Feeney, Y. Tang, E. C. Holmes, T. Allen, J. G. Prado, M. Altfeld, C. Brander, C. Dixon, D. Ramduth, P. Jeena, S. A. Thomas, A. St John, T. A. Roach, B. Kupfer, G. Luzzi, A. Edwards, G. Taylor, H. Lyall, G. Tudor-Williams, V. Novelli, J. Martinez-Picado, P. Kiepiela, B. D. Walker, and P. J. Goulder. HIV evolution: CTL escape mutation and reversion after transmission. *Nat Med*, 10(3):282–9, 2004.

S. Leslie, P. Donnelly, and G. McVean. A statistical method for predicting classical HLA alleles from SNP data. *Am J Hum Genet*, 82(1):48–56, 2008.

S. M. Lewis and A. E. Raftery. Estimating Bayes factors via posterior simulation with the Laplace-Metropolis estimator. *Journal of the American Statistical Association*, 92(438):648–655, 1997.

N. Li and M. Stephens. Modeling linkage disequilibrium and identifying recombination hotspots using single-nucleotide polymorphism data. *Genetics*, 165(4):2213–33, 2003.

G. W. Litman, J. P. Rast, and S. D. Fugmann. The origins of vertebrate adaptive immunity. *Nat Rev Immunol*, 10(8):543–53, 2010.

J. Lohm, M. Grahn, A. Langefors, O. Andersen, A. Storset, and T. von Schantz. Experimental evidence for major histocompatibility complex-allele-specific resistance to a bacterial infection. *Proc Biol Sci*, 269(1504):2029–33, 2002.

J. Marchini, D. Cutler, N. Patterson, M. Stephens, E. Eskin, E. Halperin, S. Lin, Z. S. Qin, H. M. Munro, G. R. Abecasis, and P. Donnelly. A comparison of phasing algorithms for trios and unrelated individuals. *Am J Hum Genet*, 78(3):437–50, 2006.

J. Marchini and B. Howie. Genotype imputation for genome-wide association studies. *Nat Rev Genet*, 11(7):499–511, 2010.

J. Marchini, B. Howie, S. Myers, G. McVean, and P. Donnelly. A new multipoint method for genome-wide association studies by imputation of genotypes. *Nat Genet*, 39(7):906–13, 2007.

A. J. McMichael, T. Sasazuki, H. O. McDevitt, and R. O. Payne. Increased frequency of HLA-Cw3 and HLA-Dw4 in rheumatoid arthritis. *Arthritis Rheum*, 20(5):1037–42, 1977.

G. McVean. A genealogical interpretation of principal components analysis. *PLoS Genet*, 5(10):e1000686, 2009.

G. A. McVean and N. J. Cardin. Approximating the coalescent with recombination. *Philos Trans R Soc Lond B Biol Sci*, 360(1459):1387–93, 2005.

C. A. Mein, L. Esposito, M. G. Dunn, G. C. Johnson, A. E. Timms, J. V. Goy, A. N. Smith, L. Sebag-Montefiore, M. E. Merriman, A. J. Wilson, L. E. Pritchard, F. Cucca, A. H. Barnett, S. C. Bain, and J. A. Todd. A search for type 1 diabetes susceptibility genes in families from the United Kingdom. *Nat Genet*, 19(3):297–300, 1998.

D. Middleton. History of DNA typing for the human MHC. *Rev Immunogenet*, 1(2):135–56, 1999.

L Moutsianas. *Imputation aided analysis of the association between autoimmune diseases and the MHC*. Ph.D. thesis, 2011.

L. Moutsianas, V. Enciso-Mora, Y. P. Ma, S. Leslie, A. Dilthey, P. Broderick, A. Sherborne, R. Cooke, A. Ashworth, A. J. Swerdlow, G. McVean, and R. S. Houlston. Multiple Hodgkin lymphoma-associated loci within the HLA region at chromosome 6p21.3. *Blood*, 118(3):670–4, 2011.

B. Mueller-Hilke and N. A. Mitchison. The role of HLA promoters in autoimmunity. *Curr Pharm Des*, 12(29):3743–52, 2006.

S. Myers, L. Bottolo, C. Freeman, G. McVean, and P. Donnelly. A fine-scale map of recombination rates and hotspots across the human genome. *Science*, 310(5746):321–4, 2005.

M. Nei and A. P. Rooney. Concerted and birth-and-death evolution of multigene families. *Annu Rev Genet*, 39:121–52, 2005.

J. Nerup, P. Platz, O. O. Andersen, M. Christy, J. Lyngsoe, J. E. Poulsen, L. P. Ryder, L. S. Nielsen, M. Thomsen, and A. Svejgaard. HL-A antigens and diabetes mellitus. *Lancet*, 2(7885):864–6, 1974.

M. A. Newton and A. E. Raftery. Approximate Bayesian inference with the weighted likelihood bootstrap. *Journal of the Royal Statistical Society Series B-Methodological*, 56(1):3–48, 1994.

J. Novembre, T. Johnson, K. Bryc, Z. Kutalik, A. R. Boyko, A. Auton, A. Indap, K. S. King, S. Bergmann, M. R. Nelson, M. Stephens, and C. D. Bustamante. Genes mirror geography within Europe. *Nature*, 456(7218):98–101, 2008.

J. R. Oksenberg, S. E. Baranzini, S. Sawcer, and S. L. Hauser. The genetics of multiple sclerosis: SNPs to pathways to pathogenesis. *Nat Rev Genet*, 9(7):516–26, 2008.

N. Patterson, A. L. Price, and D. Reich. Population structure and eigenanalysis. *PLoS Genet*, 2(12):e190, 2006.

D. G. Pellicci, A. J. Clarke, O. Patel, T. Mallevaey, T. Beddoe, J. Le Nours, A. P. Uldrich, J. McCluskey, G. S. Besra, S. A. Porcelli, L. Gapin, D. I. Godfrey, and J. Rossjohn. Recognition of beta-linked self glycolipids mediated by natural killer T cell antigen receptors. *Nat Immunol*, 12(9):827–33, 2011.

D. Piancatelli, A. Canossi, A. Aureli, K. Oumhani, T. Del Beato, M. Di Rocco, G. Liberatore, A. Tessitore, K. Witter, R. El Aouad, and D. Adorno. Human leukocyte antigen-A, -B, and -Cw polymorphism in a Berber population from North Morocco using sequence-based typing. *Tissue Antigens*, 63(2):158–72, 2004.

J. K. Pickrell, G. Coop, J. Novembre, S. Kudaravalli, J. Z. Li, D. Absher, B. S. Srinivasan, G. S. Barsh, R. M. Myers, M. W. Feldman, and J. K. Pritchard. Signals of recent positive selection in a worldwide sample of human populations. *Genome Res*, 19(5):826–37, 2009.

R. L. Prentice and R. Pyke. Logistic disease incidence models and case-control studies. *Biometrika*, 66(3):403–411, 1979.

A. L. Price, N. J. Patterson, R. M. Plenge, M. E. Weinblatt, N. A. Shadick, and D. Reich. Principal components analysis corrects for stratification in genome-wide association studies. *Nat Genet*, 38(8):904–9, 2006.

F. Prugnolle, A. Manica, M. Charpentier, J. F. Guegan, V. Guernier, and F. Balloux. Pathogen-driven selection and worldwide HLA class I diversity. *Curr Biol*, 15(11):1022–7, 2005.

S. Purcell, B. Neale, K. Todd-Brown, L. Thomas, M. A. Ferreira, D. Bender, J. Maller, P. Sklar, P. I. de Bakker, M. J. Daly, and P. C. Sham. PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am J Hum Genet*, 81(3):559–75, 2007.

N. Qutob, F. Balloux, T. Raj, H. Liu, S. Marion de Proce, J. Trowsdale, and A. Manica. Signatures of historical demography and pathogen richness on MHC class I genes. *Immunogenetics*, 2011.

L. R. Rabiner. A tutorial on hidden Markov models and selected applications in speech recognition. *Proceedings of the Ieee*, 77(2):257–286, 1989.

S. Raychaudhuri, C. Sandor, E. A. Stahl, J. Freudenberg, H. S. Lee, X. Jia, L. Alfredsson, L. Padyukov, L. Klareskog, J. Worthington, K. A. Siminovitch, S. C. Bae, R. M. Plenge, P. K. Gregersen, and P. I. de Bakker. Five amino acids in three HLA proteins explain most of the association between MHC and seropositive rheumatoid arthritis. *Nat Genet*, 44(3):291–6, 2012.

J. Robinson, M. J. Waller, P. Parham, N. de Groot, R. Bontrop, L. J. Kennedy, P. Stoehr, and S. G. Marsh. IMGT/HLA and IMGT/MHC: sequence databases for the study of the major histocompatibility complex. *Nucleic Acids Res*, 31(1):311–4, 2003.

D. Ron, Y. Singer, and N. Tishby. On the learnability and usage of acyclic probabilistic finite automata. *Journal of Computer and System Sciences*, 56(2):133–152, 1998.

S. Sakaguchi, M. Miyara, C. M. Costantino, and D. A. Hafler. FOXP3+ regulatory T cells in the human immune system. *Nat Rev Immunol*, 10(7):490–500, 2010.

S. Sakaguchi, K. Wing, and M. Miyara. Regulatory T cells - a brief history and perspective. *Eur J Immunol*, 37 Suppl 1:S116–23, 2007.

S. Sawcer, G. Hellenthal, M. Pirinen, C. C. Spencer, N. A. Patsopoulos, L. Moutsianas, A. Dilthey, Z. Su, C. Freeman, S. E. Hunt, S. Edkins, E. Gray, D. R. Booth, S. C. Potter, A. Goris, G. Band, A. B. Oturai, A. Strange, J. Saarela, C. Bellenguez, B. Fontaine, M. Gillman, B. Hemmer, R. Gwilliam, F. Zipp, A. Jayakumar, R. Martin, S. Leslie, S. Hawkins, E. Giannoulatou, S. D'Alfonso, H. Blackburn, F. M. Boneschi, J. Liddle, H. F. Harbo, M. L. Perez, A. Spurkland, M. J. Waller, M. P. Mycko, M. Ricketts, M. Comabella, N. Hammond, I. Kockum, O. T. McCann, M. Ban, P. Whittaker, A. Kemppinen, P. Weston, C. Hawkins, S. Widaa, J. Zajicek, S. Dronov, N. Robertson, S. J. Bumpstead, L. F. Barcellos, R. Ravindrarajah, R. Abraham, L. Alfredsson, K. Ardlie, C. Aubin, A. Baker, K. Baker, S. E. Baranzini, L. Bergamaschi, R. Bergamaschi, A. Bernstein, A. Berthele, M. Boggild, J. P. Bradfield, D. Brassat, S. A. Broadley, D. Buck, H. Butzkueven, R. Capra, W. M. Carroll, P. Cavalla, E. G. Celius, S. Cepok, R. Chiavacci, F. Clerget-Darpoux, K. Clysters, G. Comi, M. Cossburn, I. Cournu-Rebeix, M. B. Cox, W. Cozen, B. A. Cree, A. H. Cross, D. Cusi, M. J. Daly, E. Davis, P. I. de Bakker, M. Debouverie, B. D'Hooghe M, K. Dixon, R. Dobosi, B. Dubois, D. Ellinghaus, I. Elovaara, F. Esposito, et al. Genetic risk and a primary role for cell-mediated immune mechanisms in multiple sclerosis. *Nature*, 476(7359):214–9, 2011.

J. Schad, J. U. Ganzhorn, and S. Sommer. Parasite burden and constitution of major histocompatibility complex in the Malagasy mouse lemur, Microcebus murinus. *Evolution*, 59(2):439–50, 2005.

K. Schoenberg, M. Sribar, J. Enczmann, J. C. Fischer, and M. Uhrberg. Analyses of HLA-C-specific KIR repertoires in donors with group A and B haplotypes suggest a ligand-instructed model of NK cell receptor acquisition. *Blood*, 117(1):98–107, 2011.

S. R. Seaman and S. Richardson. Equivalence of prospective and retrospective models in the Bayesian analysis of case-control studies. *Biometrika*, 91(1):15–25, 2004.

T. Shiina, H. Inoko, and J. K. Kulski. An update of the HLA genomic region, locus information and disease associations: 2004. *Tissue Antigens*, 64(6):631–49, 2004.

D. P. Singal and M. A. Blajchman. Histocompatibility (HL-A) antigens, lymphocytotoxic antibodies and tissue antibodies in patients with diabetes mellitus. *Diabetes*, 22(6):429–32, 1973.

M. Sospedra and R. Martin. Immunology of multiple sclerosis. *Annu Rev Immunol*, 23:683–747, 2005.

C. C. Spencer, Z. Su, P. Donnelly, and J. Marchini. Designing genome-wide association studies: sample size, power, imputation, and the choice of genotyping chip. *PLoS Genet*, 5(5):e1000477, 2009.

B. Star, A. J. Nederbragt, S. Jentoft, U. Grimholt, M. Malmstrom, T. F. Gregers, T. B. Rounge, J. Paulsen, M. H. Solbakken, A. Sharma, O. F. Wetten, A. Lanzen, R. Winer, J. Knight, J. H. Vogel, B. Aken, O. Andersen, K. Lagesen, A. Tooming-Klunderud, R. B. Edvardsen, K. G. Tina, M. Espelund, C. Nepal, C. Previti, B. O. Karlsen, T. Moum, M. Skage, P. R. Berg, T. Gjoen, H. Kuhl, J. Thorsen, K. Malde, R. Reinhardt, L. Du, S. D. Johansen, S. Searle, S. Lien, F. Nilsen, I. Jonassen, S. W. Omholt, N. C. Stenseth, and K. S. Jakobsen. The genome sequence of Atlantic cod reveals a unique immune system. *Nature*, 477(7363):207–10, 2011.

P. Stastny. Association of the B-cell alloantigen DRw4 with rheumatoid arthritis. *N Engl J Med*, 298(16):869–71, 1978.

M. Stephens and D. J. Balding. Bayesian statistical methods for genetic association studies. *Nat Rev Genet*, 10(10):681–90, 2009.

M. Stephens and P. Scheet. Accounting for decay of linkage disequilibrium in haplotype inference and missing-data imputation. *Am J Hum Genet*, 76(3):449–62, 2005.

A. Strange, F. Capon, C. C. Spencer, J. Knight, M. E. Weale, M. H. Allen, A. Barton, G. Band, C. Bellenguez, J. G. Bergboer, J. M. Blackwell, E. Bramon, S. J. Bumpstead, J. P. Casas, M. J. Cork, A. Corvin, P. Deloukas, A. Dilthey, A. Duncanson, S. Edkins,

X. Estivill, O. Fitzgerald, C. Freeman, E. Giardina, E. Gray, A. Hofer, U. Huffmeier, S. E. Hunt, A. D. Irvine, J. Jankowski, B. Kirby, C. Langford, J. Lascorz, J. Leman, S. Leslie, L. Mallbris, H. S. Markus, C. G. Mathew, W. H. McLean, R. McManus, R. Mossner, L. Moutsianas, A. T. Naluai, F. O. Nestle, G. Novelli, A. Onoufriadis, C. N. Palmer, C. Perricone, M. Pirinen, R. Plomin, S. C. Potter, R. M. Pujol, A. Rautanen, E. Riveira-Munoz, A. W. Ryan, W. Salmhofer, L. Samuelsson, S. J. Sawcer, J. Schalkwijk, C. H. Smith, M. Stahle, Z. Su, R. Tazi-Ahnini, H. Traupe, A. C. Viswanathan, R. B. Warren, W. Weger, K. Wolk, N. Wood, J. Worthington, H. S. Young, P. L. Zeeuwen, A. Hayday, A. D. Burden, C. E. Griffiths, J. Kere, A. Reis, G. McVean, D. M. Evans, M. A. Brown, J. N. Barker, L. Peltonen, P. Donnelly, and R. C. Trembath. A genome-wide association study identifies new psoriasis susceptibility loci and an interaction between HLA-C and ERAP1. *Nat Genet*, 2010.

P. Szabolcs, M. Cavazzana-Calvo, A. Fischer, and P. Veys. Bone marrow transplantation for primary immunodeficiency diseases. *Pediatr Clin North Am*, 57(1):207–37, 2010.

P. I. Terasaki, M. S. Park, G. Opelz, and A. Ting. Multiple sclerosis and high incidence of a B lymphocyte antigen. *Science*, 193(4259):1245–7, 1976.

J. N. Torimiro, J. K. Carr, N. D. Wolfe, P. Karacki, M. P. Martin, X. Gao, U. Tamoufe, A. Thomas, E. M. Ngole, D. L. Birx, F. E. McCutchan, D. S. Burke, and M. Carrington. HLA class I diversity among rural rainforest inhabitants in Cameroon: identification of A*2612-B*4407 haplotype. *Tissue Antigens*, 67(1):30–7, 2006.

E. Trachtenberg, M. Vinson, E. Hayes, Y. M. Hsu, K. Houtchens, H. Erlich, W. Klitz, Y. Hsia, and J. Hollenbach. HLA class I (A, B, C) and class II (DRB1, DQA1, DQB1, DPB1) alleles and haplotypes in the Han from southern China. *Tissue Antigens*, 70(6):455–63, 2007.

J. Trowsdale. "Both man & bird & beast": comparative organization of MHC genes. *Immunogenetics*, 41(1):1–17, 1995.

M. Uhrberg. The KIR gene family: life in the fast lane of evolution. *Eur J Immunol*, 35(1):10–5, 2005.

P. J. van den Elsen, T. M. Holling, H. F. Kuipers, and N. van der Stoep. Transcriptional regulation of antigen presentation. *Curr Opin Immunol*, 16(1):67–75, 2004.

C. Vilches and P. Parham. KIR: diverse, rapidly evolving receptors of innate and adaptive immunity. *Annu Rev Immunol*, 20:217–51, 2002.

J. A. Villadangos and P. Schnorrer. Intrinsic and cooperative antigen-presenting functions of dendritic-cell subsets in vivo. *Nat Rev Immunol*, 7(7):543–55, 2007.

John Wakeley. *Coalescent theory: an introduction*. Roberts & Co. Publishers, Greenwood Village, Colo., 2009. John Wakeley. ill. ; 24 cm.

G. Zangenberg, M. M. Huang, N. Arnheim, and H. Erlich. New HLA-DPB1 alleles generated by interallelic gene conversion detected by analysis of sperm. *Nat Genet*, 10(4):407–14, 1995.