Imputation aided analysis of the association between autoimmune diseases and the MHC

Loukas Moutsianas
Department of Statistics and Hertford College
University of Oxford

Supervisor:

Professor Gil McVean

Trinity Term, 2011

Thesis Submitted in Partial Fulfilment of the Requirements for the Degree of Doctor of Philosophy

Imputation aided analysis of the association between autoimmune diseases and the MHC

Loukas Moutsianas
Hertford College, University of Oxford
D.Phil. Thesis, Trinity Term, 2011

Abstract

The Major Histocompatibility Complex (MHC) is a genomic region in chromosome 6 which has been consistently found to be associated with the risk of developing vrtually all common autoimmune diseases. Although its importance in disease pathogenesis has been known for decades, efforts to disentangle the roles of the classical human leukocyte antigens (HLA) and other variants responsible for the susceptibility to disease have often met with limited success, owing to the complex structure and extreme heterogeneity of the region.

In this thesis, I interrogate the MHC for association with three common autoimmune diseases, ankylosing spondylitis, psoriasis and multiple sclerosis, with the aim of confirming the previously-reported associations and of identifying novel ones. To do so, I employ a systematic, joint analysis of single nucleotide polymorphism (SNP) and HLA allele data, in a logistic regression framework, using a recently developed algorithm to predict the HLA alleles for samples where such information is unavailable. To ensure the reliability of the analysis, I apply stringent quality control procedures and integrate over the uncertainty of the HLA allele predictions. Moreover, I resolve the haplotype phase of individuals from the HapMap project to create reliable reference panels, used in both HLA prediction and in quality control procedures.

By directly testing HLA subtypes for association with the disease, the power to detect such associations is increased. I present the results of the analysis on the three disease phenotypes and discuss the evidence for important novel findings amongst both SNPs and HLA alleles in two of the diseases.

In the final part of this thesis, I introduce a novel, model-based approach to detect inconsistencies in the data and show how it can be used to flag problematic SNPs which conventional quality control procedures may fail to identify.

Contents

Chapter 1

Background and Introduction

1.1 Historical Overview

The earliest documented usage of the term *genetics* to describe the science of inheritance is in a letter[1] written by the British biologist William Bateson in 1905. Four years later, the Danish botanist Wilhelm Johannsen, whose research was mainly supported by funds donated by the owner of the Carlsberg breweries to produce improved varieties of barley for the brewing industry [1], coined the term *gene* to describe, in his own words (translated in [2]), the

> "special conditions, foundations and determiners which are present [in the gametes] in unique, separate and thereby independent ways [by which] many characteristics of the organism are specified".

The definition of the term *gene* has been continuously evolving ever since, reflecting our increasing understanding of the complexity of genomic architecture. A nice historical overview of these changes is given by Gerstein *et al.* in their 2007 review [2], together with a proposal for a new definition of *gene* as *"a union of genomic sequences encoding a coherent set of potentially overlapping functional products"*. Both words (*gene* and *genetics*) are etymologically derived from the Greek word γένος, which means race/family descent.

Inspired by the seminal work of Mendel, geneticists in the early 20th century had realised that the way some traits were inherited was being governed by the same rules of inheritance. Moreover,

[1] `http://www.jic.ac.uk/corporate/about/bateson.htm`

Chapter 2

Phasing HapMap3 and other datasets

The importance of constructing a dense and reliable reference panel has been highlighted throughout Chapter 1, and in section 1.4 in particular. In this chapter, phasing details are discussed for the HapMap3 datasets. This should be considered a methods description section, rather than one discussing novel statistical methods or analysis. Instead, it is hoped that the haplotype data made available by us for this project will aid investigators in conducting improved quality control, phasing and imputation of their respective datasets, and therefore facilitate analysis of genetic variation across the human genome. Although the dataset was only recently published [20], it was extensively used by the 1000 Genomes Consortium [31] and is recommended as the reference resource to be used (complementary to the data from the 1000 Genomes Project) by the developers of one of the most commonly used imputation platforms[1] [119].

I have been responsible for the preparation and execution of the phasing process for all HapMap3 populations, as well as for continuous support and recommendations on its usage. Moreover, and despite the fact that a publicly available algorithm was employed for the stochastic part of the process, I implemented custom routines for the deterministic part, as well as for

[1] In the words of Marchini and Howie, "We recently posted the latest haplotypes from the 1,000 Genomes Project, which were released in June 2010. There are 120 CEU haplotypes, 120 CHB+JPT haplotypes, and 118 YRI haplotypes in the new dataset. You can download the official release haplotypes or a set of haplotypes tailored to work with HapMap 3 below; we recommend using the latter (1,000 Genomes + HapMap 3) reference set for most imputation tasks." Extract found in the IMPUTE2 website: `http://mathgen.stats.ox.ac.uk/impute/impute_v2.html#Using_Impute2_with_Public_Reference_Data`, information correct as of January, 2011.

Chapter 3

Datasets, preparation, and uncertainty-aware analysis

In this chapter I describe the datasets used in the association studies presented in this thesis, as well as the quality control procedures applied to them. The UK control datasets are common to all three association studies, and are described first. The disease datasets for PS and AS follow next. The datasets employed for the MS association study are being presented in a separate section, in line with the presentation of results from the three studies in two separate chapters, one for AS and PS (Chapter 4) and another for MS (Chapter 5). One of the main reasons for the joint presentation of the AS and PS results, with more being given at the beginning of Chapter 4, is the similarity of the data in hand for the two diseases.

The potential sources of error in association studies can be extremely heterogeneous. Errors may have been introduced in any of the experimental, technical or computational steps, from sample collection to the phasing of genotype data. Sources of error include, but are not limited to, bias in the selection of samples, sample contamination, cryptic relatedness or other causes of sample ascertainment, faulty genotyping (e.g. batch effects), imperfect genotype calling and phasing. For that reason, the explicit modelling of error in the analysis would be an extremely challenging process. Therefore, a set of tests and filtering procedures have been commonly applied to genotype data in the setting of association studies, to limit the effect errors may have in the analysis and to minimise false associations which could occur as a result.

Chapter 4

Studying the association of the MHC with two common autoimmune diseases

In this chapter I present an analysis of the association of the Major Histocompatibility Complex with two common, complex autoimmune diseases, Ankylosing Spondylitis (AS) and Psoriasis (PS). A brief overview of the MHC has been provided elsewhere (§1.3). The data were analysed as part of my involvement in the Wellcome Trust Case Control Consortium 2 (WTCCC2)[1]. The two analyses are presented together in a single chapter due to similarities in (*i*) the genetic architecture of the conditions and (*ii*) the methods used for preparation and analysis of the data. Specifically,

i. Both conditions have a consistently replicated primary association in the class I HLA region of the MHC, which is widely believed to be a direct association of a specific class I HLA allele with the disease, with a dominant effect on the log-odds of developing the disease. Moreover, non-MHC genes involved in inflammatory pathways have been associated with both diseases (*e.g.* *ERAP*1 [151, 175]).

ii. The procedures followed for the quality control and phasing of the SNP data,

[1] `https://www.wtccc.org.uk/ccc2/`

Chapter 5

The association of the MHC with multiple sclerosis

In this chapter I present an analysis of the association of the Major Histocompatibility Complex with Multiple Sclerosis (MS), based on a large dataset consisting of samples from multiple cohorts across Europe and US. This analysis was conducted as part of my involvement in the Wellcome Trust Case Control Consortium 2 (WTCCC2)[1]. I start by giving an overview of the disease and an outline of the environmental and genetic factors associated with it, both in the MHC and in the rest of the genome. I then present and discuss the most important results from my analysis on the disease outcome, as well as on sub-phenotypes such as severity and clinical course. The chapter ends with a discussion on the analysis and its main findings.

5.1 Pathogenesis of MS and the effect of genetic and environmental components

Multiple sclerosis is a common autoimmune disease of unknown, possibly heterogeneous aetiology. It is a severe neurological disability characterised by axonal damage (and loss, in the later stages) and demyelination in the central nervous system (CNS). There is substantial variability in the course of the disease between patients, with speculations that it could partially be an outcome

[1] `https://www.wtccc.org.uk/ccc2/`

Chapter 6

Detection of sites with inconsistent LD structure using a model-based approach

In this chapter I present an algorithm for the estimation of a per-site weighting scheme for phased SNP and HLA alleles, to be employed as a means of assessing the quality of haplotype data. The algorithm builds on a framework which is commonly employed for computationally tractable statistical inference on population genetic data, the Li & Stephens model [115]. Therefore, this model is discussed in some detail below, together with some of its limitations. I then explain the motivation which led to my extension to the model, followed by its main characteristics and a detailed description of the algorithm. Finally, I present results on simulated and real data, assess the model's performance and discuss its strength and weaknesses, as well as its potential application as an additional QC metric for haplotype data.

6.1 The Li & Stephens Model

The Li & Stephens model explicitly relates observed genetic variation patterns, and the LD between bi-allelic markers in particular, to underlying recombination rates. It was originally developed to identify positions with increased recombination activity, commonly referred to as

Chapter 7

General discussion and perspectives

Association studies rely on the LD structure between *observed* variation in polymorphic markers genotyped across the genome, and *potentially unobserved* nearby variation, which directly affects the phenotype of interest. Using the former as proxies, they indirectly study the effects of the latter on a trait. The marked drop in genotyping costs during the last decade, together with sustained efforts of international collaborations (e.g. [18, 19, 20]) to build dense maps of human genetic variation, enabled the design of standard commercial genotyping chips which contain hundreds of thousands of markers spanning the whole genome. The use of such chips to study variation in allelic differences at a population level led in turn to a major increase in the number of loci found to be associated with various diseases [27, 289]. However, studies using typed HLA data had already shown evidence for the association of the MHC with multiple sclerosis [48], ankylosing spondylitis [182], and psoriasis [208], almost thirty years before such resources became available. Therefore, the main aim of this thesis was not to discover a novel region of interest with respect to the trait in question, but to exploit the availability of dense genotype data in order to fine-map the signals which are driving the observed associations in the MHC, and to interrogate the region for additional, secondary ones.

In contrast to the GWAS's working principle, the LD structure in the MHC region made this fine-mapping and exploratory effort more challenging, rather than aiding it. The complex