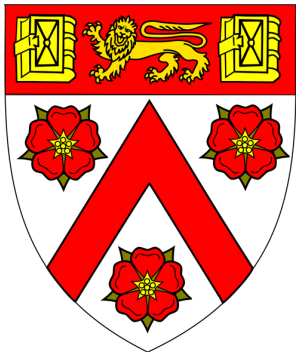
Statistical Techniques to Fine Map the Related Genetic Aetiology of Autoimmune Diseases



Mary Fortune

University of Cambridge

This dissertation is submitted for the degree of

Doctor of Philosophy

Trinity College

May 2017

Statistical Techniques to Fine Map the Related Genetic Aetiology of Autoimmune Diseases

Mary Fortune

Genome Wide Association Studies (GWAS) have uncovered many genetic regions which are associated with autoimmune disease risk. In this thesis, I present methods which I have developed to build upon these studies and enable the analysis of the causal variants of these diseases.

Colocalization methods disentangle whether potential causal variants are shared or distinct in related diseases, and enable the discovery of novel associations below the single-trait significance threshold. However, existing approaches require independent datasets to accomplish this. I extended two methods to allow for the shared-control design; one of these extensions also enables fine mapping in the case of shared variants. My analysis of four autoimmune diseases identified 90 regions associated with at least one disease, 33 of which were associated with 2 or more disorders; 14 of these had evidence of distinct causal variants.

Once associated variants have been identified, we may wish to test some aggregate property, such as enrichment within an annotation of interest. However, the null distribution of GWAS signals showing association with a trait and preserving expected correlation due to linkage disequilibrium is complicated. I present an algorithm which computes the expected output of a GWAS, given any arbitrary definition of "null", and hence can be used to simulate the null distribution required for such a test.

Commonly, GWAS report only summary data, and determining which genetic variants are causal is more difficult; the strongest signal may merely be correlated with the true causal variant. I have developed a statistical method for fine mapping a region, requiring only GWAS p-values and publicly available reference datasets. I sample from the space of potential causal models, rejecting those leading to expected summary data excessively different from that observed. This removes the need for the assumption of a single causal variant. In contrast to other summary statistic methods which allow for multiple causal variants, it does not depend upon availability of effect size estimates, or the allelic direction of effect and it can infer whether the pattern of association is likely caused by a non-genotyped SNP without requiring imputation. I discuss the effect of choice of reference dataset, and the implications for other summary statistics techniques.

Declaration

This dissertation is the result of my own work and includes nothing which is the outcome of work done in collaboration except as declared in the Preface and specified in the text.

It is not substantially the same as any that I have submitted, or, is being concurrently submitted for a degree or diploma or other qualification at the University of Cambridge or any other University or similar institution except as declared in the Preface and specified in the text. I further state that no substantial part of my dissertation has already been submitted, or, is being concurrently submitted for any such degree, diploma or other qualification at the University of Cambridge or any other University of similar institution except as declared in the Preface and specified in the text.

It does not exceed the prescribed word limit for the relevant Degree Committee.

Mary Fortune

May 2017

Acknowledgements

Foremost, I wish to thank my supervisor, Dr. Chris Wallace, for all her advice and support throughout my PhD.

I would also like to thank past and present members of the Wallace Group: Olly Burren; Nick Cooper; Marina Evangelou; Stasia Grinberg; Hui Guo; Steve Kiddle; James Liley; Niko Pontikos and Xin Yang, for their friendship, encouragement and assistance in increasing the prevalence of depictions of fish in the Department of Medicine.

My PhD was funded by the Wellcome Trust. The JDRF/Wellcome Trust Diabetes and Inflammation Laboratory, in which I spent the majority of my PhD, is funded by the Juvenile Diabetes Research Foundation, the Wellcome Trust and the National Institute for Health Cambridge Biomedical Research Centre.

A PhD is all about learning new skills and so, finally, I would like to thank my St John Ambulance Unit for giving me a better appreciation of medicine, letting me stage a motorcycle crash in a nightclub, and generally ensuring that my submission deadline has not been the most stressful situation I have been in over the last four years.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

An increasing focus in medicine is the integration of genetic information, from early prediction of disease risk to making more informed treatment choices. As genetic mechanisms are becoming better understood, and technology is being developed to enable increasingly larger scale and cheaper molecular phenotyping of patients, we are coming to understand the heterogeneity of genetics in complex diseases. As the volume of biological data increases, the challenges to be solved in understanding disease aetiology have progressively become statistical rather than biological.

The biological datasets available are often incomplete, and contain structures which are not fully understood, making analysis more difficult. In order to increase our power to detect disease causing genetic variants, it is necessary to develop methodologies which enable the integration of several datasets, be they cross-disease analysis or incorporating epigenetic information. In this thesis I address the statistical methods required for fine mapping causal variants and for examining whether causal variants are shared between two diseases, applying my methods to the area of autoimmune disease.

1.1 Autoimmune Diseases

Autoimmune diseases are caused by the immune system being reactive to self-tissue, resulting in damage to the organs or structures being targeted. Many autoimmune diseases are known, but my research focuses upon four in particular: Type 1 Diabetes; Rheumatoid Arthritis; Celiac Disease and Multiple Sclerosis.

1.1.1 Individual Diseases

1.1.1.1 Type 1 Diabetes

Type 1 Diabetes (T1D, formerly known as insulin-dependent diabetes mellitus) is caused by the autoimmune destruction of the insulin-producing β-cells within the pancreatic islets. Insulin is a peptide hormone which promotes the absorption of glucose by cells. In its absence, the body is unable to regulate blood glucose levels; this leads to both hyperglycemia and ketoacidosis, as cells are starved of glucose. These are potentially fatal if the underlying condition is not recognised and treated.

There is no cure for T1D, and, although there is some evidence that immunosuppressive agents can slow progression in newly diagnosed patients [Bluestone et al, 2010], the side effects of these drugs mean that they are not used; since age of onset is typically young, with the peak age of diagnosis around 14, long term safety of therapies is a particular concern. Instead, a life-long regime of insulin-replacement is required, and those whose disease is not well-controlled are at risk of complications such as cardiomyopathy, renal failure and retinopathy. Glucose intake must be carefully monitored so that the correct insulin dose can be given; hypoglycemia can have a swift onset, and is fatal if not corrected. T1D is associated with a significantly increased mortality rate [Soedamah-Muthu et al, 2006].

A symptomatically related disease is Type 2 Diabetes; however this is not thought

to be immune-related, but rather a metabolic disease in which the host cells become resistant to insulin.

1.1.1.2 Rheumatoid Arthritis

Rheumatoid Arthritis (RA) is caused by inflammation of the synovial membrane, the connective tissue lining the inner surface of many joints. The inflammatory environment in the synovium of patients with RA can damage the cartilage of the joint, which leads in turn to erosion of the bone. RA is a polyarthritis, typically initially presenting with pain and swelling in small joints, such as those in the hands, before progressing to larger joints. In addition, there are many extra-articular manifestations of the disease; it is associated with ischaemic heart disease and pulmonary fibrosis, leading to increased mortality.

The joint damage done is irreversible, and RA is frequently disabling. However, disease-modifying drugs, such as *TNF*-inhibitors, are sometimes able to slow disease progression; and, together with analgesia, are the primary treatments used.

1.1.1.3 Celiac Disease

Celiac Disease (CEL, also known as Coeliac Disease) is caused by an autoinflammatory reaction to small bowel tissue in the presence of gliadin, a gluten protein found in wheat and similar cereals. As well as causing pain, this inflammation results in atrophy of the villi lining the small intestine and consequent inability to properly absorb food. CEL is a pre-malignant condition, leading to an increased risk of both lymphoma and adenocarcinoma of the small bowel.

Since the autoantibodies in CEL are produced only in the presence of gliadin, the most effective treatment for this condition is a lifelong gluten-free diet.

1.1.1.4 Multiple Sclerosis

In Multiple Sclerosis (MS) the immune system targets antigens within the brain and spinal cord, resulting in the destruction of the myelin sheaths insulating neurons and the formation of lesions within the central nervous system. This results in a wide range of neurological symptoms, depending upon the location of the lesions, including sensory, motor and cognitive defects.

The disease has several distinct clinical manifestations. Patients with the relapsing-remitting form have a pattern of periods of increased disease activity followed by fading of symptoms, as demyelination occurs and then heals poorly. By contrast, other patients present a progressive pattern of the disease, leading to prolonged demyelination of the neurons and eventually to axiol loss. These patients experience a steady worsening over time. While the majority of those with the disease present with the relapsing-remitting phenotype, this typically converts to the progressive form.

Although there is currently no cure for MS, recently monoclonal antibodies such as alemtuzumab, which targets *CD52*, a protein expressed on the surface of lymphocytes, have shown promise in reducing the rate of relapses [Coles, 2012].

1.1.2 Biology of Autoimmune Diseases

The immune system is a collection of structures and processes in the body which protect against invading pathogens and promote host tissue integrity. An important constituent of the immune system are the white blood cells (leukocytes), which are produced in the bone marrow but found throughout the circulatory system and body tissues. Leukocytes encompass a variety of functionally-distinct cell types, which are characterised by a remarkable plasticity to recognise and respond to virtually any type of pathogen. However, two types are of particular interest for the study of autoimmune disease.

Many autoimmune diseases, including the four discussed above, occur when T-lymphocytes target self-antigens. There are several types of T-cells. Prior to differentiation, they undergo several rounds of selection; in the last round, "negative selection", they are presented with self-antigens on the Major Histocompatibility Complex (MHC) of medullary thymic epithelial cells. Those which react too strongly have the potential to cause autoimmune disease; these are generally induced to undergo apoptosis, however, some differentiate into regulatory T-cells (T_{REG}s). Although the mechanisms by which they do so are not fully understood, these suppress the responses of other T-cells, reducing reactions to self-antigens; increasing T_{REG} function has been suggested as a therapy strategy for autoimmune disease [Waldron-Lynch et al, 2014]. Other types of T-cells include cytotoxic T-cells (T_Cs) and helper T-cells (T_Hs). When they first differentiate, both T_Cs and T_Hs are "naive", and require specific antigen stimulation in the context of the MHC to promote their activation and clonal selection. T_Cs directly kill cells which express their target antigen on its class 1 MHC molecule by releasing cytokines and inducing apoptosis. In contrast, when T_Hs encounters a cell which expresses their target antigen on its class 2 MHC molecule, they release cytokines, which assist in the immune response. After an immune response has occurred, some antigen-experienced T-cells remain as long-lived memory T-cells, to enable a quicker response to be mounted against the same pathogen in the future.

B-lymphocytes secrete antibodies, proteins which bind to antigens on the pathogen and either impede their target or signal in order to mark their target out for destruction by other immune cells. B-cells play a role in some autoimmune diseases, although to a lesser extent than T-cells; rheumatoid factor is an auto-antibody which is often found in patients with RA. As with T-cells, B-cells go through a negative selection step in development, in order to prevent the differentiation of mature B cells, with the capacity to produce autoreactive antibodies, that can recognise self-antigens.

1.1.3 Genetics of Autoimmune Diseases

From looking at how disease prevalence clusters within families, there is strong evidence that autoimmune disease propensity is heritable. A standard estimate of familial aggregation is the sibling recurrence risk, how much more likely is it that a sibling of someone affected also has the disease than we would expect by random chance. This is high in autoimmune diseases; it has been reported that T1D has a sibling recurrence risk ratio as high as 12 [Risch, 1987]. This is highly suggestive of a significant genetic component to disease risk.

The strongest genetic association with autoimmune diseases (and the first to be discovered, via linkage studies in the 1970s) is the HLA region on chromosome 6. This region encodes the MHC, which is vital in immune system regulation and presents the key molecules that allow T-cells to recognise self and foreign antigens. However, common autoimmune diseases are strongly polygenic and, especially with the advent of genome-wide association studies (GWAS), a great many new disease-associated regions have been discovered. Figure 1.1 shows the location of all 51 currently known T1D-associated regions; these are distributed throughout the genome.

Genetic factors are known to play a role in all aspects of autoimmune disease: from susceptibility to self-reactive immune responses, determining the targets of autoimmune destruction and disease progression. However, there are also non-genetic factors at play. By examining disease concordance between pairs of monozygotic twins [Selmi et al, 2012] produce estimates of what proportion of autoimmune disease risk is attributable to genetic factors; these are significantly below one. The remainder is due to environmental factors.

Although autoimmune diseases are not modern diseases (we have records of some dating back to antiquity), in recent years their prevalence has been increasing at too great a rate to be explicable by genetic factors. There is also marked geographical

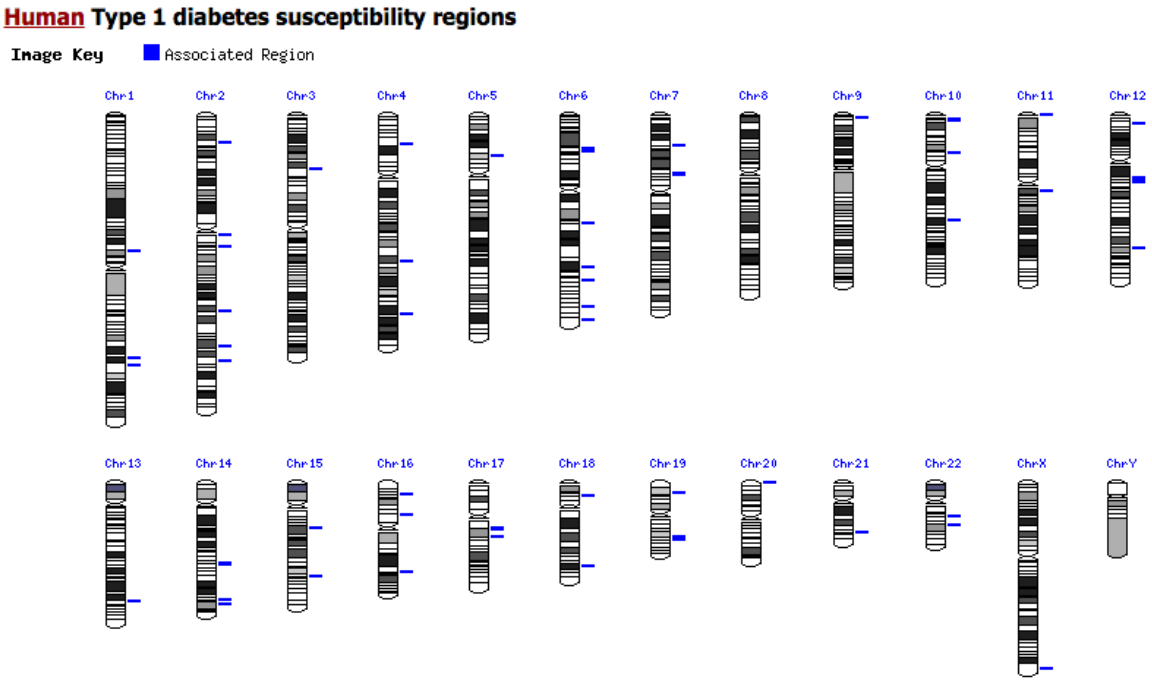


Figure 1.1 The distribution of known T1D-associated regions across the genome, taken from T1D base (`www.t1dbase.org`).

variation in prevalence, even when accounting for confounders such as prosperity, climate and exposure to pathogens. The hygiene hypothesis posits that a lack of early exposure to pathogens leads to an inappropriate education of the immune system and the generation of an immune repertoire that is more biased to autoimmune responses; Vitamin D deficiency has also been implicated in autoimmune disease development and has been proposed as a cause for the increased incidence of autoimmune diseases in Northern countries. It is thought that, though both genetic and environmental factors contribute to disease susceptibility, a trigger such as a viral infection is necessary for onset [Rodriguez-Calvo et al, 2016].

1.2 Genome Wide Association Studies

The first regions associated with autoimmune disease, such as *HLA* and the chr11 region containing the gene *INS* (associated with T1D), were found by linkage studies. These studies look at large families where members are affected by the disease over several generations, and try to find a region where shared inheritance patterns correspond to disease status. However, the erosion of LD is slow, and so given data from only a couple of generations, disease associations can only be narrowed down to a large genetic region (typically millions of base pairs). In addition, although linkage studies can work very well on traits which follow a Mendelian pattern of inheritance, if a disease is complex, with many variants having less than fully penetrant influence on disease susceptibility, only very strong associations will show in a linkage study.

The vast majority of the human genome, which contains around three billion base pairs, is conserved across humans. However, there is genetic variation across the population; any two individuals differ in about 0.5% of their DNA. The most common form of human genetic variation is the single nucleotide polymorphism (SNP). SNPs occur when there is a difference at a single nucleotide; in humans, these are estimated to occur about once every 300 base pairs. They vary from very common with no clinical implications to rare yet highly pathogenic; for instance, a SNP within the coding region of a gene may lead to an incorrect amino acid being used to form a protein, and hence changed function. The minor allele frequency (MAF) of a SNP is the frequency at which the less common allele occurs; this may vary considerably between populations (indeed, the alleles carried by an individual can be used to infer ancestry).

SNPs do not occur independently from each other; two alleles at different positions are in linkage disequilibrium (LD) if they occur together more frequently than would be expected by random chance. This has several causes. Most commonly, during the

chromosomal crossover phase of meiosis, alleles which are close together are more likely to be on the same chromatid, and hence be inherited together. A particular combination of SNP alleles may also confer some selective advantage and hence be frequently found together. A benign SNP may therefore appear to be associated with a disease if it is in high LD with a pathogenic SNP; such effects are very difficult to disentangle.

An advantage of linkage studies is that they enable discovery of genetic association without the requirement for knowledge about and determining of the specific causes of genetic variation in disease risk. More recent technology, however, has enabled the fast and cheap genotyping by microarray of large numbers of SNPs across the genome, and it is now possible to do genome wide association studies (GWAS). In these studies, independent univariate analysis of SNPs against some phenotype (such as disease status) are done for a genome-wide set of SNPs. The most common design, which is the one discussed in the rest of this thesis, is a case/control comparison, with SNP association with the disease being measured by means of a score test to determine whether allelic frequencies are significantly different between cases and controls. Typically only the summary statistics, such as unsigned p-values, are reported from a GWAS; these are often presented on a Manhattan plot, which plots position on the chromosome against $-\log_{10}$ p-value, enabling the strongest associations to be easily visible as peaks (for an example of an autoimmune disease Manhattan plot, see Figure 1.2).

As with any method which uses p-values to determine the significance of the effect found, a threshold must be chosen. By standard custom, $p < 0.05$ is used for a test of a single hypothesis. However, the number of tests being performed in a GWAS (not all of which are independent, due to the effect of LD) means that multiple testing must be corrected for. A threshold of 5×10^{-8} has been widely accepted by the field [Dudbridge and Gusnanto, 2008]. However, the true number of independent tests is population dependent, with an African cohort having more genetic variation than a European one;

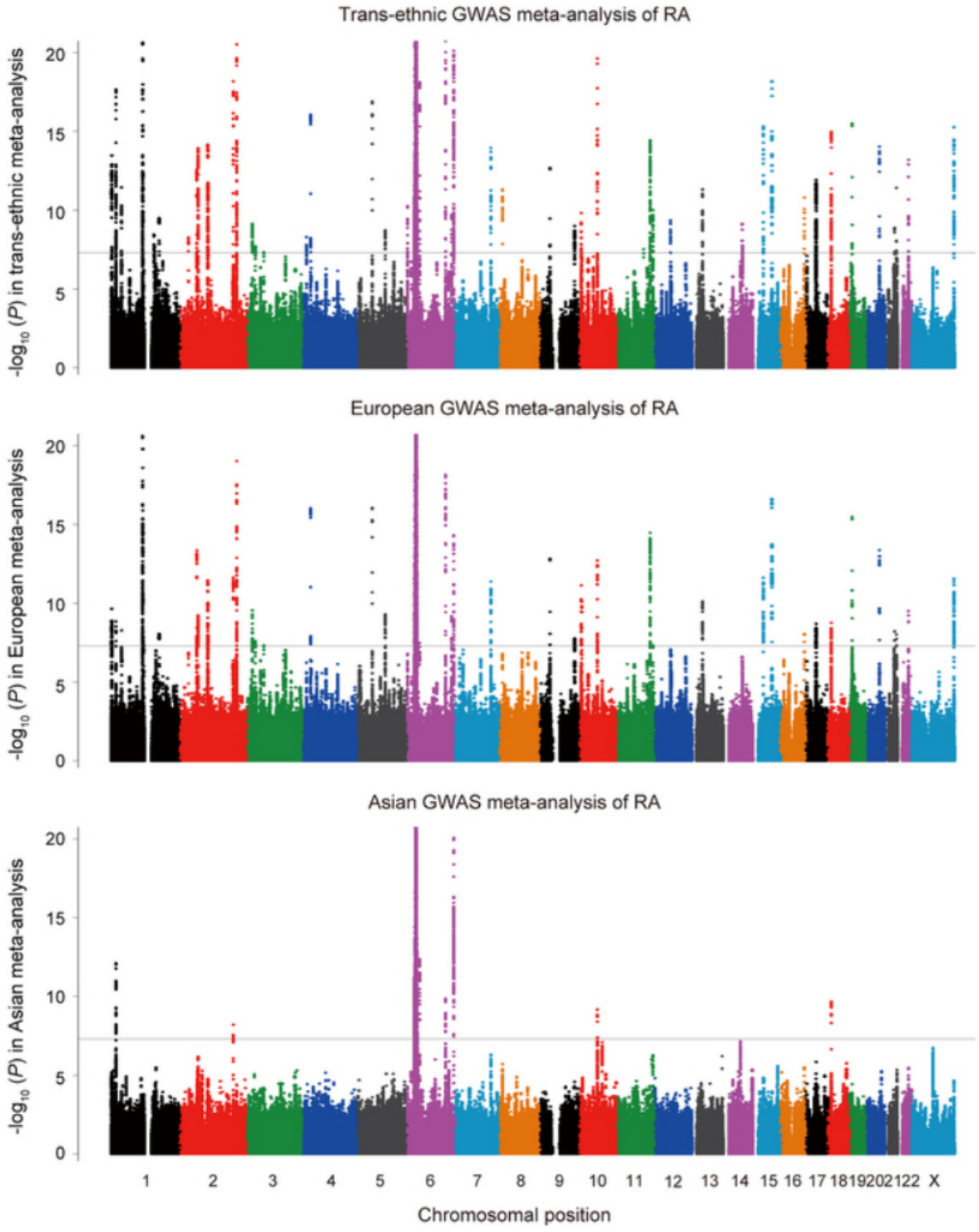


Figure 1.2 Manhattan plots of RA associations, from a meta-analysis of GWAS for trans-ethnic, European and Asian populations, coloured by chromosome. The high peak in chromosome 6 corresponds to *HLA*. Figure taken from [Okada et al, 2014].

the "correct" threshold for an African population is likely to be lower than 5×10^{-8}, and this is the subject of ongoing research.

Note that, in Figure 1.2, there are differences in the heights of the peaks of the European and Asian Manhattan plots. Many signals in the European analysis fail to reach genome-wide significance in the Asian analysis (mainly due to the European study having higher power). Due to genetic drift, allele frequency of SNPs varies between populations; a causal SNP which accounts for much disease-variation in one population may not be present in another. Since GWAS analysis is fundamentally about determining whether allele frequency differs between cases and controls, it is important to stratify by population, and ensure the use of an appropriate control dataset, to ensure any differences we detect are truly caused by disease-status. Typically a GWAS will analyse a single (often European) population in order to reduce this effect.

The first GWAS [Haines et al, 2005] used only 50 controls and 96 cases, however, as genotyping has become cheaper, the desire to increase the power to detect associations at SNPs with lower odds ratios, or greater rarity, has driven up the sample sizes, which are now typically in the tens of thousands. Rigorous quality control measures, many of them introduced by the Wellcome Trust Case Control Consortium [Burton et al, 2007] (at the time the largest GWAS ever performed, this analysed seven diseases for SNP association and found 21 significant loci, the vast majority of which have since been replicated), and an honesty about the number of hypotheses being considered are required to provide meaningful results. SNPs with low call rates are removed, since this may be indicative of DNA sample quality. Similarly, SNPs not being in Hardy-Weinberg equilibrium is suggestive of a genotyping or genotype calling error at this position, and so they are removed. In addition, the effects of population stratification must be controlled for in order to reduce the chance of systemic differences other than disease status between cases and controls.

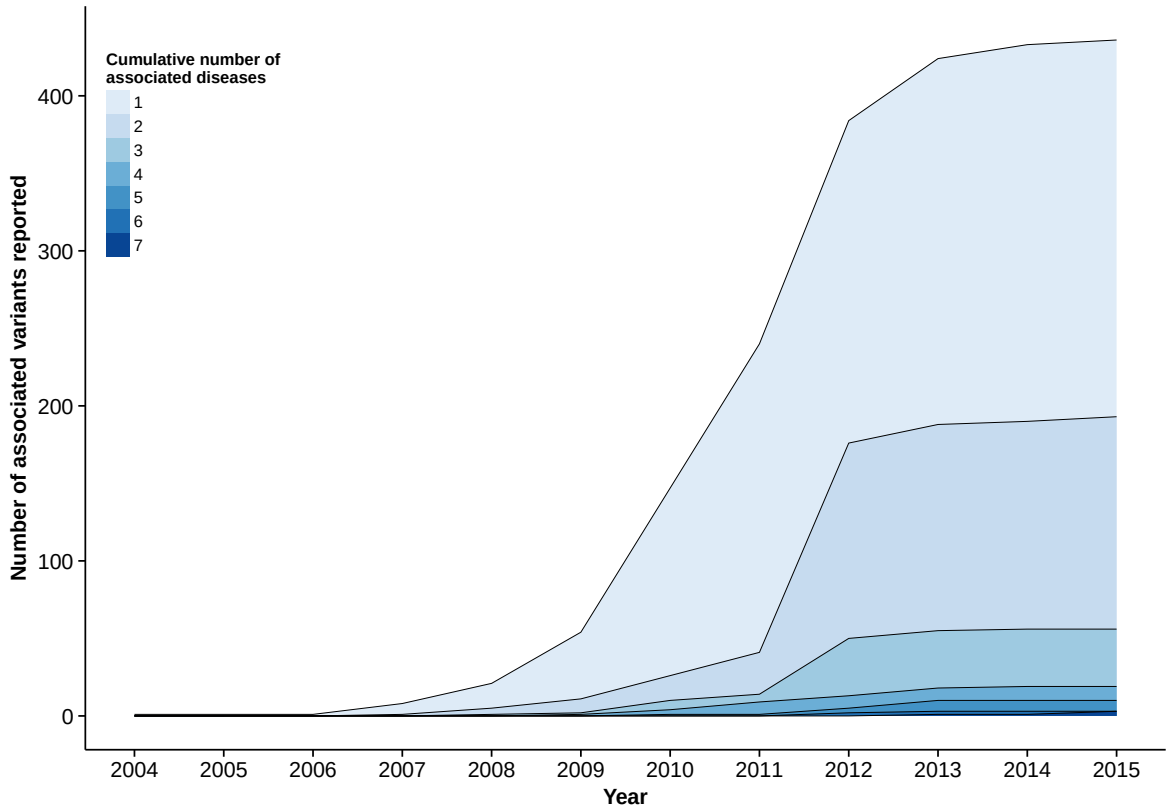


Figure 1.3 Number of published autoimmune disease associations found by GWAS, including the number of disease implicated. Figure produced by C Wallace, using data from `www.immunobase.org`.

Over the course of the past decade, the number of GWAS performed has exploded. Figure 1.3 shows the number of autoimmune disease associated variants reported by year, with increasing numbers being implicated in multiple diseases. The development of the GWAS methodology has enabled the analysis of complex polygenetic traits, and greatly improved our understanding of autoimmune disease aetiology.

1.2.1 Limitations of GWAS

Although called "genome wide", GWAS by no means test all known variants. They genotype a subset of common SNPs thought to explain a large proportion of genetic variation, but relatively few rare SNPs (indeed, for very rare SNPs, the sample sizes

required to obtain statistical power would be prohibitive). Coverage is not even across the genome; some regions of the genome have no coverage at all (see, for instance, the gaps in coverage within chromosome 1 and chromosome 9 corresponding to centromeres in Figure 1.2). Although SNPs are the most common form of genetic variation, and the easiest to analyse, others do exist; structural variations such as copy number variations and translocations may also be important in disease aetiology (though often tagged by SNPs).

For the purposes of determining whether a genetic region contains a trait-associated variant, such incomplete coverage often suffices. GWAS are able to explain significant proportions of the estimated autoimmune disease hereditary However, some heredity is unavoidably "missing"; for example, [Barrett et al, 2009b] estimate that the 32 loci identified as associated with Crohn's disease through GWAS explain only 20% of the genetic risk. Even if the true causal variant is not included, if it is in high LD with a sequenced SNP (that is, if it is "tagged" by one of the GWAS SNPs), we will still be able to see its effect upon disease status. It has also been suggested that GWAS will reveal the influence of very rare SNPs via a "synthetic association", by occurring more often in association with one of the alleles of a common GWAS SNP. However, for this signal to be statistically significant, the effect size of the rare SNP would have to be very large; in autoimmune disease, effect sizes tend to be modest, and it is unlikely that a GWAS would be powered to detect a rare causal variant of this form. Indeed, a well-powered study which searched directly for rare autoimmune associated variants found that they explained only 3% of the heritability explained by common variants [Hunt et al, 2013].

This reliance on tag-SNPs to find disease associations, however, makes the use of GWAS data for fine mapping the causal variants themselves difficult; from reported p-values alone, how do we determine whether a SNP is causal for a disease, or merely in high LD with the (possibly not genotyped) true causal variant? Fine mapping the

variants which underlie disease association is vital to understanding aetiology. Association to a region does not necessarily correspond to any given gene within that region being implicated in the disease process; it may be that the causal variant is, for instance, in a regulatory element which acts upon some distant gene. Identifying the causal variant can enable discovery of common aetiology between autoimmune diseases, and may also suggest novel treatment strategies. Fine mapping from GWAS data is a thread which runs throughout my thesis.

1.3 ImmunoChip

One solution to the problem of the true causal variants not being genotyped, especially if you have a prior belief about the identities or locations of these variants, is to create a custom genotyping platform specifically for the analysis of your disease of interest. The ImmunoChip [Cortes and Brown, 2011] is one such platform, designed to aid in the fine mapping of autoimmune disease associated signals. It contains all SNPs which had previously been associated with one of the 11 autoimmune diseases being studied, as well as all known SNPs at the time (February 2010) from the 1000 Genomes Project and European data in LD blocks surrounding these SNPs for which probes could be designed. In addition, for each disease, 3000 "wild-card" SNPs were included; these were typically either SNPs which had failed to reach genome-wide significance yet were deemed to be potentially interesting, or else further SNPs within a region believed to be disease-associated. In total, 186 loci believed to be associated with an autoimmune disease are densely covered by the ImmunoChip SNPs. In addition, to aid fine mapping of the true causal variants, SNPs were not filtered by LD or by spacing (in a GWAS, such filtering increases the number of independent signals which can be analysed, but at a cost of greatly complicating any fine mapping efforts).

Although autoimmune disease associated regions are well covered, ImmunoChip

contains only $\sim 200,000$ SNPs, much fewer than a typical GWAS. This, together with the bulk numbers in which the chips were produced, reduced the price of analysis, making running an autoimmune GWAS more attractive and enabling an increase in sample size and therefore power. The fact that analysis was done for so many diseases upon broadly the same set of SNPs makes discovering shared causal variants logistically much easier; this is also aided by the ImmunoChip Consortium providing common control data.

The ImmunoChip, however, does not completely negate the issues with use of GWAS data for fine mapping discussed in Section 1.2.1. While common variants are systematically accounted for, due to the inclusion of 1000 Genomes SNPs, the only rare variants included are those which had already been identified in an existing autoimmune disease GWAS, or which happen to have been chosen due to lying within a densely genotyped regions. This focus upon known regions means that the rest of the genome is sparsely covered; if a novel association happens to exist outside these regions (say for a different autoimmune disease), it is unlikely to be identified by a study using the ImmunoChip unless it happens to be one with a low p-value in an existing GWAS of one of the 11 diseases.

ImmunoChip contains probes for $195,806$ SNPs; the remaining 718 variants are small insertion-deletions. These variants are both the most common types and the easiest to genotype. However, this means that the effect of large structural variants is missing from discovery unless they happen to be tagged by a SNP or detectable from raw SNP intensity signals [Cooper et al, 2015].

In addition, the SNPs selected were chosen due to their association in European-only GWAS. Disease-causal variants which are found only in non-European populations will be under-represented upon the ImmunoChip, which may have implications for the analysis of such populations.

GWAS to analyse autoimmune disease genetics, including those using the ImmunoChip, are curated within ImmunoBase (`http://www.immunobase.org/`), and it is from here that I have sourced the majority of the summary data analysed in this thesis.

1.4 Statistical Methods for Assessing Shared Aetiology

Clinical studies have shown that having one autoimmune disease is a strong risk factor for developing others; for instance, many patients with ankylosing spondylitis go on to develop inflammatory bowel disorder, and vice versa [Laukens et al, 2010]. Further, family studies show clustering of multiple autoimmune diseases between relations. Some therapies, such as anti-*TNF* drugs, are effective against a wide range of autoimmune diseases. Together with the commonality of the mechanisms of the diseases, this is strong evidence for the presence of shared aetiology, and much of this sharing is likely to be genetic.

Many autoimmune-related genetic regions, including the most strongly associated, *HLA*, are associated with several diseases. In databases such as ImmunoBase (`http://www.immunobase.org/`), known association with one disease is considered grounds for lowering the threshold required for a genetic variant to be considered significantly associated with another. Figure 1.4 shows the proportion of risk associated variants which are shared between diseases. However, finding shared disease association to a genetic region does not automatically correspond to shared aetiology; instead, different variants within the region may lead to different disease processes.

One method of quantifying the shared aetiology between two traits is the genetic correlation, the correlation between the vectors of effect sizes [Bulik-Sullivan et al, 2015].

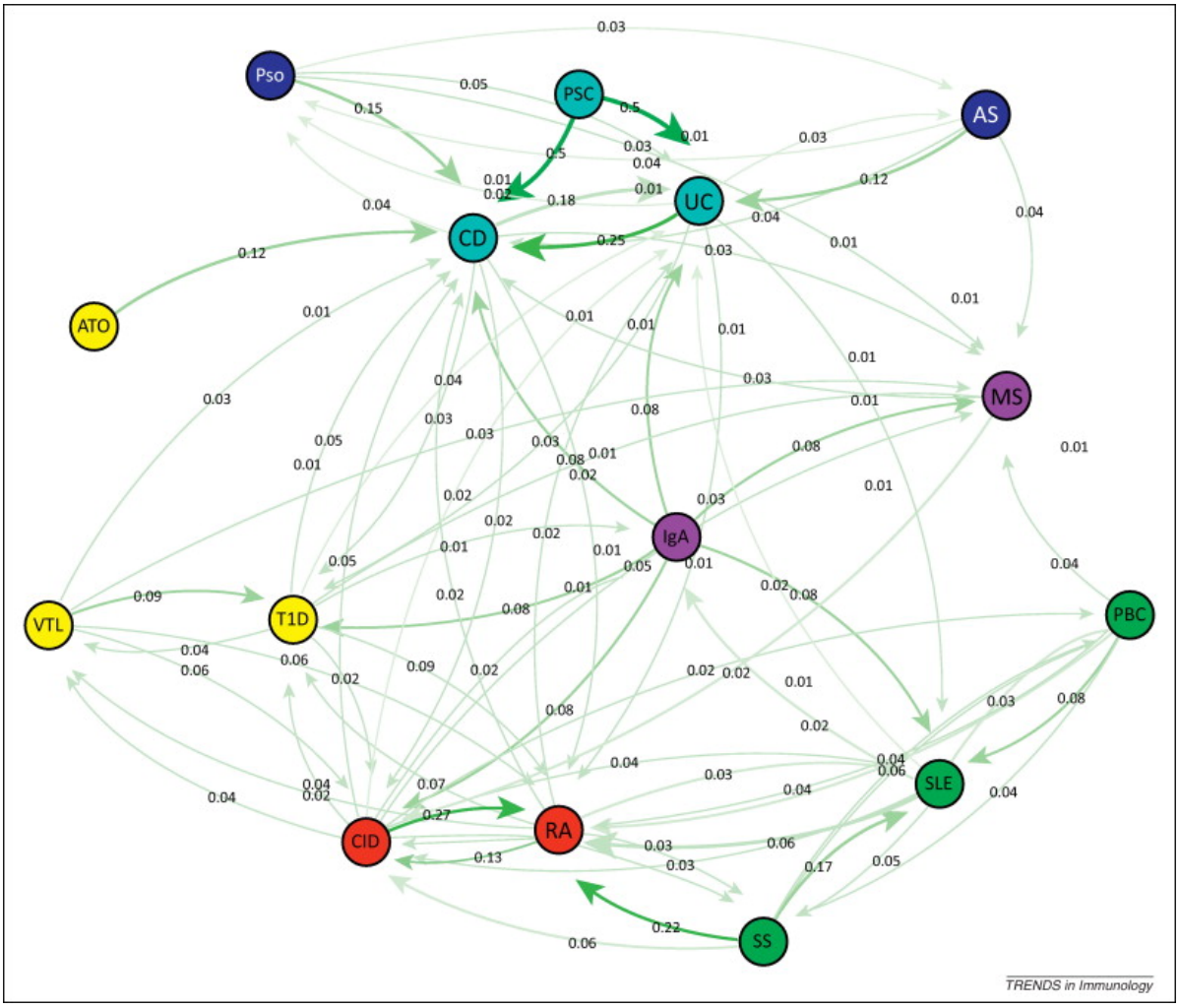


Figure 1.4 For each pair of diseases, the arrows show the proportion of variants identified as being causal for the first disease which are shared with the second disease. Figure taken from [Cotsapas and Hafler, 2013].

Let Y_1 and Y_2 denote phenotypes for disease 1 and disease 2 respectivly, and let $\mathbf{X}$ be the matrix of sample genotypes. Consider a set of SNPs of interest, $\mathcal{S}$ (typically, $\mathcal{S}$ contains all SNPs under study). Let $\boldsymbol{\gamma}$ be the zero-centred vector which satisfies $\text{argmax}_{\boldsymbol{\alpha}} Cor(Y_1, \mathbf{X}\boldsymbol{\alpha})$ (that is, the projection of Y_1 onto the space of $\mathbf{X}$), and let $\boldsymbol{\beta}$ be this value for disease 2. Then the genetic correlation between phenotypes at SNPs in $\mathcal{S}$ is computed as:

$$r_{\mathcal{S}} = \frac{\sum_{i \in \mathcal{S}} \gamma_i \beta_i}{\sqrt{\left(\sum_{i \in \mathcal{S}} \gamma_i^2\right)\left(\sum_{i \in \mathcal{S}} \beta_i^2\right)}}$$

However, although this provides a good measure for the amount of sharing between traits, it assesses this at a global level, and does not reveal whether causal variants are shared within a specific region, nor does it enable a fine mapping to determine which causal variants are shared. For this, we require colocalization techniques.

In this section, I summarise existing colocalization methods as published. By looking for colocalisation between trait-associated SNPs, these methods can be used to perform an analysis of association between two traits such as disease status. They can also be used to find colocalisation between trait-associated SNPs and gene expression Quantitative Trait Loci (eQTLs) in a cell type of interest. However, they require that for each of the trait datasets we have an independent control dataset.

1.4.1 Bayesian Approach

Here, I summarise a Bayesian approach to colocalization, which is given in [Giambartolomei et al, 2014].

1.4.1.1 Framework

Consider the case when each trait is influenced by at most one variant in a pre-specified region. Let there be $Q \geq 2$ SNPs in a region and let a configuration describe which SNP(s) in the region are causal for which trait, if any. Then there are

$(Q+1)^2$ configurations of possible causal SNPs, each of which can be assigned to one of five possible hypotheses:

$\mathbb{H}_0$: No SNP is associated with either trait.

$\mathbb{H}_1$: There is a SNP associated with trait 1, but no SNP is associated with trait 2.

$\mathbb{H}_2$: There is a SNP associated with trait 2, but no SNP is associated with trait 1.

$\mathbb{H}_3$: There is a SNP associated with trait 1, and a different SNP associated with trait 2.

$\mathbb{H}_4$: A single SNP is associated with both trait 1 and trait 2.

Colocalization is equivalent to hypothesis $\mathbb{H}_4$; a large posterior probability for this hypothesis provides evidence for colocalization.

1.4.1.2 Choice of Priors

In a Bayesian approach, we begin by defining our prior expectation of each of these hypotheses.

For each SNP in the region, we can define p_0, p_1, p_2 and p_{12} as:

p_0, the prior probability that the SNP is associated with neither disease

p_1, the prior probability that it is associated with the first disease only

p_2, the prior probability that it is associated with the second disease only

p_{12}, the prior probability that it is associated with both diseases.

In the absence of additional information about the SNPs, assume that all are equivalently likely a priori. p_1 and p_2 were assigned to be 10^{-4}, an estimate of the proportion of all SNPs expected to be associated with a given trait and equivalent to

expecting 50 detectable causal variants in a GWAS with 500,000 SNPs. p_{12} was assigned to be 10^{-6}; this was tested with a sensitivity analysis, comparing the results given when $p_{12} = 10^{-5}, 2 \times 10^{-6}, 10^{-6}$ with the original analysis of the blood lipid dataset given in [Teslovich, T. Musunuru, K. Smith, 2010]. As $p_0 = 1 - p_1 - p_2 - p_{12}$, $p_0 \simeq 1$.

Then, the prior for a configuration $\mathcal{M}$, $\Pi(\mathcal{M})$, is dependent only upon the hypothesis it corresponds to:

$$\mathcal{M} \in \mathbb{H}_0\text{: } \Pi(\mathcal{M}) = p_0^Q = p_0^2(p_0^{Q-2}) \simeq 1$$

$$\mathcal{M} \in \mathbb{H}_1\text{: } \Pi(\mathcal{M}) = p_0^{Q-1}p_1 = p_0p_1(p_0^{Q-2}) \simeq 10^{-4}$$

$$\mathcal{M} \in \mathbb{H}_2\text{: } \Pi(\mathcal{M}) = p_0^{Q-1}p_2 = p_0p_2(p_0^{Q-2}) \simeq 10^{-4}$$

$$\mathcal{M} \in \mathbb{H}_3\text{: } \Pi(\mathcal{M}) = p_0^{Q-2}p_1p_2 = p_1p_2(p_0^{Q-2}) \simeq 10^{-8}$$

$$\mathcal{M} \in \mathbb{H}_4\text{: } \Pi(\mathcal{M}) = p_0^{Q-1}p_{12} = p_0p_{12}(p_0^{Q-2}) \simeq 10^{-6}$$

1.4.1.3 Computation of Posterior Probabilities

In order to compare two models under a Bayesian framework, the Bayes Factor (BF) is frequently used; this is the ratio of how well two models predict $\mathcal{D}$, the data observed. In the case of colocalization, the BF for each model $\mathcal{M}$ against the null hypothesis of no association ($\mathbb{H}_0$) is:

$$BF(\mathcal{M}) = \frac{\mathbb{P}(\mathcal{D}|\mathcal{M})}{\mathbb{P}(\mathcal{D}|\mathbb{H}_0)}$$

Let BF_i^1 be the BF derived for the model {SNP i is causal for trait 1} and BF_j^2 be the BF derived for the model {SNP j is causal for trait 2}. Since there is an independent control dataset for each trait, the regression models are independent and hence the term $\mathbb{P}(\mathcal{D}|\mathcal{M})$ can be split into two independent terms, one for each trait, giving:

$$\mathcal{M} \in \mathbb{H}_0\text{: } BF(\mathcal{M}) = 1$$

$$\mathcal{M} \in \mathbb{H}_1\text{: } BF(\mathcal{M}) = \frac{\mathbb{P}(\mathcal{D}|\text{SNP } i \text{ is causal for trait 1})\mathbb{P}(\mathcal{D}|\text{no SNP is causal for trait 2})}{\mathbb{P}(\mathcal{D}|\text{no SNP is causal for trait 1})\mathbb{P}(\mathcal{D}|\text{no SNP is causal for trait 2})} = BF_i^1$$

$$\mathcal{M} \in \mathbb{H}_2\text{: } BF(\mathcal{M}) = \frac{\mathbb{P}(\mathcal{D}|\text{no SNP is causal for trait 1})\mathbb{P}(\mathcal{D}|\text{SNP } j \text{ is causal for trait 2})}{\mathbb{P}(\mathcal{D}|\text{no SNP is causal for trait 1})\mathbb{P}(\mathcal{D}|\text{no SNP is causal for trait 2})} = BF_j^2$$

$$\mathcal{M} \in \mathbb{H}_3\text{: } BF(\mathcal{M}) = \frac{\mathbb{P}(\mathcal{D}|\text{SNP } i \text{ is causal for trait 1})\mathbb{P}(\mathcal{D}|\text{SNP } j \text{ is causal for trait 2})}{\mathbb{P}(\mathcal{D}|\text{no SNP is causal for trait 1})\mathbb{P}(\mathcal{D}|\text{no SNP is causal for trait 2})} = BF_i^1 BF_j^2$$

$$\mathcal{M} \in \mathbb{H}_4\text{: } BF(\mathcal{M}) = \frac{\mathbb{P}(\mathcal{D}|\text{SNP } i \text{ is causal for trait 1})\mathbb{P}(\mathcal{D}|\text{SNP } i \text{ is causal for trait 2})}{\mathbb{P}(\mathcal{D}|\text{no SNP is causal for trait 1})\mathbb{P}(\mathcal{D}|\text{no SNP is causal for trait 2})} = BF_i^1 BF_i^2$$

for any SNPs i and j, $i \neq j$.

In practice, these Bayes Factors can be expensive to compute, and instead the Approximate Bayes Factors (ABF) are calculated, using the method described in [Wakefield, 2009], which enables the computation of an ABF from only summary statistics. Since often in GWAS only the summary statistics are reported, it is valuable to be able to perform colocalization on such data.

For each hypothesis $\mathbb{H}_i$, given data $\mathcal{D}$, we have:

$$\mathbb{P}(\mathbb{H}_i|\mathcal{D}) \propto \sum_{\mathcal{M}\in\mathbb{H}_i} \mathbb{P}(\mathcal{D}|\mathcal{M})\Pi(\mathcal{M})$$

and hence:

$$\begin{aligned}\frac{\mathbb{P}(\mathbb{H}_i|\mathcal{D})}{\mathbb{P}(\mathbb{H}_0|\mathcal{D})} &= \frac{\Pi(\mathcal{M}|\mathcal{M}\in\mathbb{H}_i)}{\Pi(\mathbb{H}_0)} \sum_{\mathcal{M}\in\mathbb{H}_i} \frac{\mathbb{P}(\mathcal{D}|\mathcal{M})}{\mathbb{P}(\mathcal{D}|\mathbb{H}_0)} \\ &= \frac{\Pi(\mathcal{M}|\mathcal{M}\in\mathbb{H}_i)}{\Pi(\mathbb{H}_0)} \sum_{\mathcal{M}\in\mathbb{H}_i} BF(\mathcal{M})\end{aligned}$$

Defining ABF_i^1 to be the ABF derived for the model {SNP i is causal for trait 1} and ABF_j^2 to be the ABF derived for the model {SNP j is causal for trait 2} as before, and using the priors defined in Section 1.4.1.2, we can write:

$$\frac{\mathbb{P}(\mathbb{H}_0|\mathcal{D})}{\mathbb{P}(\mathbb{H}_0|\mathcal{D})} = 1$$

$$\frac{\mathbb{P}(\mathbb{H}_1|\mathcal{D})}{\mathbb{P}(\mathbb{H}_0|\mathcal{D})} = p_1 \sum_{i=1}^{Q} \mathrm{ABF}_i^1$$

$$\frac{\mathbb{P}(\mathbb{H}_2|\mathcal{D})}{\mathbb{P}(\mathbb{H}_0|\mathcal{D})} = p_2 \sum_{i=1}^{Q} \mathrm{ABF}_i^2$$

$$\frac{\mathbb{P}(\mathbb{H}_3|\mathcal{D})}{\mathbb{P}(\mathbb{H}_0|\mathcal{D})} = p_1 p_2 \sum_{i \neq j} \mathrm{ABF}_i^1 \mathrm{ABF}_j^2$$

$$\frac{\mathbb{P}(\mathbb{H}_4|\mathcal{D})}{\mathbb{P}(\mathbb{H}_0|\mathcal{D})} = p_{12} \sum_{i=1}^{Q} \mathrm{ABF}_i^1 \mathrm{ABF}_i^2$$

1.4.2 Proportional Approach

Here, I describe the proportional approach to colocalization, as given in [Wallace et al, 2012] and [Wallace, 2013].

1.4.2.1 Test for Colocalization

Write $\mathbf{Y}_1$ and $\mathbf{Y}_2$ for the response vectors of the two traits of interest (for instance, disease status for two diseases with believed shared aetiology or disease status and gene expression data). Let $\hat{\beta}_1$ and $\hat{\beta}_2$ be the maximum likelihood estimators for β_1 and β_2, the coefficients obtained when the $\mathbf{Y}$s are individually regressed upon genotype data for a set of Q explanatory SNPs. Let β_1 and β_2 have covariance matrices $\mathbf{V}_1$ and $\mathbf{V}_2$ respectively.

The null hypothesis of colocalization is expressed as proportionality between the coefficient vectors; that is, there exists some constant η such that $\beta_1 = \frac{\beta_2}{\eta}$. By Fieller's theorem [Fieller, 1954], under this null hypothesis, if η is known:

$$\left(\hat{\beta}_1 - \frac{\hat{\beta}_2}{\eta}\right)^T \left(V_1 + \frac{V_2}{\eta^2}\right)^{-1} \left(\hat{\beta}_1 - \frac{\hat{\beta}_2}{\eta}\right) \sim \chi^2_Q \tag{1.1}$$

However, η is unknown. Instead, under a profile likelihood approach, the maximum likelihood estimator, $\hat{\eta}$, is used and the distribution in Equation 1.1 cannot be assumed.

One option would be to assume a χ^2_{Q-1} distribution, but discontinuities in the likelihood also pose a problem. Instead, a posterior predictive p-value is computed [Rubin, 1984].

Writing $\eta = \tan(\theta)$, the test statistic at a given value of θ is:

$$T(\theta) = \left(\sin(\theta)\hat{\beta}_1 - \cos(\theta)\hat{\beta}_2\right)^T \left(\sin^2(\theta)V_1 + \cos^2(\theta)V_2\right)^{-1} \left(\sin(\theta)\hat{\beta}_1 - \cos(\theta)\hat{\beta}_2\right) \sim \chi^2_Q$$

Write $\mathcal{P}(\theta)$ for the posterior distribution of θ given $\hat{\beta}_1$ and $\hat{\beta}_2$. Write $T^*(\theta)$ for the p-value derived at a specific value of θ. Then the posterior predictive p-value for testing the null hypothesis of colocalization is given by:

$$\int_0^\pi T^*(\theta)\mathcal{P}(\theta)d\theta$$

1.4.2.2 Computation of $\mathcal{P}(\theta)$, the Posterior Distribution of the Proportionality Constant

In the absence of any additional information, non-informative priors $\pi(\theta) \sim 1$ and $\pi(\beta) \sim \mathbf{1}$ are used.

Let $M = \left(\cos^2(\theta)V_1^{-1} + \sin^2(\theta)V_2^{-1}\right)^{-1}$ and $\mu = \left(\cos(\theta)\hat{\beta}_1 V_1^{-1} + \sin(\theta)\hat{\beta}_2 V_2^{-1}\right) M$

The likelihood of $\hat{\beta}_1, \hat{\beta}_2$ given θ, β, is given by:

$$\begin{aligned}
\mathcal{L}(\hat{\beta}_1, \hat{\beta}_2|\beta, \theta) &= \frac{1}{(2\pi)^{\frac{Q}{2}}\left(|V_1||V_2|\right)^{\frac{1}{2}}} \\
&\quad \times \exp\left(\frac{-\left[(\hat{\beta}_1 - \cos(\theta)\beta)^T V_1^{-1}(\hat{\beta}_1 - \cos(\theta)\beta) + (\hat{\beta}_2 - \sin(\theta)\beta)^T V_2^{-1}(\hat{\beta}_2 - \sin(\theta)\beta)\right]}{2}\right) \\
&= \frac{1}{(2\pi)^{\frac{Q}{2}}\left(|V_1||V_2|\right)^{\frac{1}{2}}} \\
&\quad \times \exp\left(\frac{-\left[(\beta - \mu)^T M^{-1}(\beta - \mu) - \mu^T M^{-1}\mu + b_1^T V_1^{-1}\hat{\beta}_1 + b_2^T V_2^{-1}\hat{\beta}_2\right]}{2}\right)
\end{aligned}$$

and the posterior distribution of θ is:

$$\begin{aligned}\mathcal{P}(\theta) &\propto \int_{-\infty}^{\infty} \mathcal{L}(\hat{\beta}_1, \hat{\beta}_2|\beta, \theta)\pi(\theta)\pi(\beta)d\beta \\ &\propto \frac{|M|^{\frac{1}{2}}}{(2\pi)^{\frac{Q}{2}}\left(|V_1||V_2|\right)^{\frac{1}{2}}}\exp\left(\frac{-\left[b_1^T V_1^{-1}\hat{\beta}_1 + b_2^T V_2^{-1}\hat{\beta}_2 - \mu^T M^{-1}\mu\right]}{2}\right)\end{aligned}$$

This posterior distribution can also be used to compute a (possibly disjoint) 95% credible interval for η.

1.4.2.3 Selection of Appropriate SNPs for the Proportional Approach

We expect that only a small proportion of the SNPs in the region are actually associated with the traits. Each additional SNP analyzed adds a degree of freedom, leading to a loss of power in our test; due to linkage disequilibrium, these SNPs are highly correlated, and hence an appropiately chosen much smaller subset of SNPs can capture almost the same information as the complete set of SNPs. In addition, this algorithm requires the computation of the inverses of the Q by Q matrices V_1 and V_2; it is seldomly computationally feasible to analyse all SNPs within a genetic locus. We therefore wish to choose a subset of these to examine.

One suggestion would be to choose the top SNPs based upon their relation to the traits being studied. However, due to the *Winner's Curse* effect [Ioannidis, 2008], these apparent top SNPs are likely to have inflated effect sizes compared to their true values: while asymptotically the expectation of the Maximum Likelihood Estimator, $\hat{\beta}$, is β, once we condition upon having passed some significance threshold γ, $\mathbb{E}\left(\hat{\beta}\middle|\left\{\left|\frac{\hat{\beta}}{SE(\hat{\beta})}\right| > \gamma\right\}\right) \neq \beta$. Instead, there are several selection strategies which do not result in biased estimators being passed to the colocalization analysis.

The first is Bayesian Model Averaging. In this technique, we treat the choice of SNPs in the model itself as a nuisance parameter. We consider all regression models of traits against two SNPs, considering each equally likely a priori, and for each compute a Bayes factor (and thus a posterior probability). In practice, in order to speed up computation, an Approximate

Bayes Factor is computed, using a Laplace approximation [Raftery, 1996]. For each model, we also compute the posterior predictive p-value (calculated as in 1.4.2.2). By summing over these posterior predictive p-values, weighted by the posterior probability of the model they correspond to, we are able to compute an overall posterior predictive p-value for testing the hypothesis of proportional effects for the two traits in the region.

Alternatively, Principal Component Analysis can be used. This procedure transforms the SNP data into an orthogonal set of linear sums of the original variables, in such a way that the first component explains the largest possible amount of the variance, the second component explains the largest possible amount of the remaining variance, and so on. By taking the first few principle components, we are able to run the analysis on a much reduced dataset which still explains the majority of variance in disease status. This has the advantage that the algorithm need only be run once. However, the principle components are linear sums, potentially involving all SNPs; this makes the results hard to interpret in the context of the effect of individual SNPs.

1.5 Statistical Methods for Fine Mapping

Ideally, full genotype data would be used to identify causal variants. However, this is often unavailable, and many fine mapping studies have been done in order to identify causal variants from GWAS summary data, such as p-values, SNP odds ratios and standard errors [Maller et al, 2012]. Frequently these methods assume that each genetic region contains a single causal variant, an assumption known to be false in many autoimmune-associated regions. A common strategy for dealing with multiple causal SNPs is to adopt a conditional approach, at each iteration finding the strongest signal remaining. However, this approach can lead to us discounting what could turn out to be the strongest model. In the *IL2RA* region, for instance, the best 2 SNP model for MS does not contain the top performing single SNP, and hence is not found by a forward stepwise search [Wallace et al, 2015].

In this section, I summarise two existing techniques which infer causal variants from GWAS

summary data without making assumptions about the number of causal variants. They do, however, require as input the direction of effects at each SNP, and are only able to analyse potential causal variants which have been genotyped.

1.5.1 PAINTOR

PAINTOR (Kichaev et al [2014]) assumes multiple causal variants are possible and allows for the integration of functional genomic annotation data such as transcription factor binding sites; these can be found from sources including ENCODE [ENCODE Project Consortium et al, 2012].

1.5.1.1 The Model

Let L be the number of fine-mapping loci under analysis. Let locus j contain N_j SNPs, have Z-Score vector $\mathbf{Z}_j$ and R^2 matrix Σ_j, where Σ_j is estimated from a reference dataset such as 1000 Genomes [Auton et al, 2015] if necessary. For SNP i within locus j, let C_{ij} be the indicator that i is causal, and let λ_{ij} be the non-centrality parameter of the standardised effect size of i.

Let K be the number of functional genomic annotations obtained for these loci. Define vectors $\mathbf{A}_{ij}$ with:

$$A_{ijk} = \begin{cases} 0 & k = 0 \text{ (the baseline)} \\ 1 & \text{SNP } i \text{ in locus } j \text{ is part of annotation } k \\ 0 & \text{SNP } i \text{ in locus } j \text{ is not part of annotation } k \end{cases}$$

and let γ_k be the effect size of a causal SNP having annotation k.

Then the likelihood of observing $\mathbf{Z}$ is

$$\begin{aligned}\mathcal{L}(\mathbf{Z}|\boldsymbol{\gamma},\boldsymbol{\lambda},\mathbf{A}) &= \sum_{\mathbf{C}\in\mathcal{C}} \mathbb{P}(\mathbf{Z}\cap\mathbf{C}|\boldsymbol{\gamma},\boldsymbol{\lambda},\mathbf{A})) \\ &= \prod_j \sum_{C_j\in\mathcal{C}_j} \mathbb{P}(Z_j|C_j\lambda_j)\mathbb{P}(C_j|\boldsymbol{\gamma}\mathbf{A}_{*j})\end{aligned}$$

The effect of $\boldsymbol{\gamma}$ on causality is modelled as a logistic regression:

$$\mathbb{P}(C_j|\boldsymbol{\gamma}) = \prod_i \left(\frac{1}{1+exp(\boldsymbol{\gamma}^T\boldsymbol{A}_{ij})}\right)^{C_{ij}} \left(\frac{1}{1+exp(-\boldsymbol{\gamma}^T\boldsymbol{A}_{ij})}\right)^{1-C_{ij}}$$

and $\mathbf{Z}$ is modelled as a multivariate normal:

$$\mathbf{Z}|\lambda_j C_j \sim N(Z_j, \Sigma_j(\lambda_j \circ C_j)\Sigma_j) \tag{1.2}$$

where the vector $(\lambda_j \circ C_j)$ gives the elemental pairwise multiplication between λ_j and C_j.

1.5.1.2 Model Fitting

Since the model includes latent variables C_j, it is not possible to simply maximise the likelihood in order to fit data to the model. Instead, a Expectation Maximisation (EM) algorithm is used to maximise the likelihood over $\boldsymbol{\gamma}$. This involves two steps, iterated until convergence is achieved. In the first step (the 'E Step'), the expected value of the log-likelihood of C given the current values of $\boldsymbol{\gamma}$ is computed. In the second step (the 'M Step'), the equation computed in the previous E Step is maximised to update the estimate of $\boldsymbol{\gamma}$.

In order to prevent the model from being over-specified, the non-centrality parameters λ are fixed, and only γ is optimised over. The value of λ used is:

$$\lambda_j = \begin{cases} \mathbf{Z}_j & |\mathbf{Z}_j| > 3.7 \\ 3.7 \times \text{sign}(\mathbf{Z}_j) & \text{else} \end{cases}$$

(Note that a Z-Score of 3.7 corresponds to a p-value of 10^{-4}).

1.5.2 CAVIAR

CAVIAR (Hormozdiari et al [2014]) allows for multiple causal variants (although, for computational reasons, in practice we assume at most 6). As output, it generates a set of SNPs that with high probability contains all causal variants.

1.5.2.1 Single Causal Variant

Let n be the number of individuals and m be the number of SNPs genotyped, with SNP c being the sole causal SNP.

Write $\mathbf{y}$ for the vector of phenotypes and $\mathbf{x}_i$ for the vector of genotypes at SNP i. Without loss of generality, let $\mathbf{x}_i$ be normalised such that $\mathbf{1}^T\mathbf{x}_i = 0$ and $\mathbf{x}_i^T\mathbf{x}_i = n$.

Assume that the phenotypes can be modelled by a linear model:

$$\mathbf{y} = \mu\mathbf{1} + \beta_c\mathbf{x}_c + \boldsymbol{\epsilon} \qquad \boldsymbol{\epsilon} \sim N_n(\mathbf{0}, \sigma^2\mathbf{I})$$

Then the likelihood function is:

$$\mathcal{L}(\mathbf{y}|\mu, \beta_c, \sigma^2) = |2\pi\sigma^2\mathbf{I}|exp\left(-\frac{1}{2\sigma^2}(\mathbf{y} - \mu\mathbf{1} - \beta_c\mathbf{x}_c)^T(\mathbf{y} - \mu\mathbf{1} - \beta_c\mathbf{x}_c)\right)$$

and maximising this gives:

$$\begin{pmatrix}\hat{\mu}\\ \hat{\beta}_c\end{pmatrix} = \frac{1}{n}\begin{pmatrix}\mathbf{1}^T\mathbf{y}\\ \mathbf{x}_c^T\mathbf{y}\end{pmatrix} \qquad \begin{pmatrix}\hat{\mu}\\ \hat{\beta}_c\end{pmatrix} \sim N\left(\begin{pmatrix}\mu\\ \beta_c\end{pmatrix}, \frac{\sigma^2}{n}\begin{pmatrix}1 & 0\\ 0 & 1\end{pmatrix}\right)$$

$$\hat{\boldsymbol{\epsilon}} = \mathbf{y} - \hat{\mu}\mathbf{1} - \hat{\beta}_c\mathbf{x}_c \qquad \hat{\boldsymbol{\epsilon}}^T\hat{\boldsymbol{\epsilon}} \sim \sigma^2\chi^2_n \qquad \text{independent of } \hat{\mu},\ \hat{\beta}_c$$

Then, the association statistics for SNP c, $S_c = \hat{s}_c$,

$$\hat{s}_c = \frac{\frac{\sqrt{n}\hat{\beta}_c}{\sigma}}{\sqrt{\frac{1}{n}\frac{\sqrt{\hat{\boldsymbol{\epsilon}}^T\hat{\boldsymbol{\epsilon}}}}{\sigma}}} = \frac{n\hat{\beta}_c}{\sqrt{\hat{\boldsymbol{\epsilon}}^T\hat{\boldsymbol{\epsilon}}}}$$

has a $t_{\lambda_c,n}$ distribution, with non-centrality parameter $\lambda_c = \frac{\hat{\beta}_c}{\sigma}\sqrt{n}$. For sufficiently large n, approximate $S_c \sim N(\lambda_c, 1)$.

Now consider some other SNP, i, and let $r = \frac{1}{n}\mathbf{x}_i^T\mathbf{x}_c$. Now:

$$\hat{\beta}_i = \frac{\mathbf{x}_i^T\mathbf{y}}{n} \qquad \lambda_i = r\lambda_c$$

$$\frac{\hat{\beta}_i}{\sigma}\sqrt{n} \sim N\left(r\frac{\hat{\beta}_c}{\sigma}\sqrt{n}, 1\right) \qquad Cov\left(\frac{\hat{\beta}_c}{\sigma}\sqrt{n}, \frac{\hat{\beta}_i}{\sigma}\sqrt{n}\right) = \frac{1}{n\sigma^2}\mathbf{x}_c^T\sigma^2\mathbf{I}\mathbf{x}_i = r$$

Hence, for any two SNPs i and j

$$\begin{pmatrix} S_i \\ S_j \end{pmatrix} \sim N\left(\begin{pmatrix} \lambda_i \\ \lambda_j \end{pmatrix}, \begin{pmatrix} 1 & r_{ij} \\ r_{ij} & 1 \end{pmatrix}\right)$$

and SNP i is causal at the α-significance level if:

$$|\hat{s}_i| > \Phi^{-1}(1 - \frac{\alpha}{2})$$

where Φ^{-1} is the quantile function of the standard normal distribution.

1.5.2.2 Multiple Causal Variants - Identical Non-Centrality Parameters

Now assume that there may be multiple causal SNPs. Write $\mathbf{c}$ for the vector which indicates signed the causal status for each SNP: $\mathbf{c}_i = 1$ if SNP i has positive effect; $\mathbf{c}_i = -1$ if SNP i has negative effect and $\mathbf{c}_i = 0$ if SNP i has no effect. Assume that all causal SNPs share the same non-centrality parameter, λ_c.

Write $\boldsymbol{\Sigma}$ for the matrix of pairwise correlations between SNPs. Then, as in the previous

single variant case:

$$\mathbf{S} \sim N(\lambda_c \boldsymbol{\Sigma} \mathbf{c}, \boldsymbol{\Sigma})$$

Let γ be the probability that any given SNP is causal and has a positive effect; under the assumption that positive and negative effects are equally likely, this is also the probability that any given SNP is causal and has a negative effect. Then the prior probability that a particular vector $\mathbf{c}^*$ occurs is:

$$\mathbb{P}(\mathbf{c}^*) = \prod_{i=1}^{n} \gamma^{|c_i^*|}(1-2\gamma)^{(1-|c_i^*|)}$$

and hence the posterior probability of $\mathbf{c}^*$ given association statistics $\hat{\mathbf{s}}$ is:

$$\mathbb{P}(\mathbf{c}^*|\hat{\mathbf{s}}) = \frac{\mathbb{P}(\hat{\mathbf{s}}|\mathbf{c}^*)\mathbb{P}(\mathbf{c}^*)}{\sum_c \mathbb{P}(\hat{\mathbf{s}}|\mathbf{c})\mathbb{P}(\mathbf{c})}$$

Functional data can be incorporated by using a variant-specific probability, γ_i.

Now consider a set of SNPs $\mathcal{K}$, and write $\mathbf{C}_\mathcal{K}$ for the set of configurations with all causal SNPs being from $\mathcal{K}$ (including the null configuration). Then let ρ, the confidence level that $\mathcal{K}$ captures the true causal SNPs, be:

$$\rho = \mathbb{P}(\mathbf{C}_\mathcal{K}|\hat{\mathbf{s}}) = \sum_{c \in \mathbf{C}_\mathcal{K}} \mathbb{P}(\mathbf{c}|\hat{\mathbf{s}})$$

$\mathcal{K}$ is called a ρ-confidence-set of causal SNPs.

For a given minimum confidence threshold ρ^*, the best estimate of the causal model is then given by the smallest set $\mathcal{K}^*$ such that $\mathbb{P}(\mathbf{C}_{\mathcal{K}^*}|\hat{\mathbf{s}}) \geq \rho^*$.

1.5.2.3 Multiple Causal Variants - Different Non-Centrality Parameters

Now, allow causal SNPs to have different non-centrality parameters. Use the prior probability:

$$\boldsymbol{\lambda}_c|\mathbf{c} \sim N(0, \boldsymbol{\Sigma}_c) \qquad \boldsymbol{\Sigma}_c = \begin{cases} 0 & i \neq j \\ \sigma & i\text{causal} \\ \epsilon & i\text{not causal} \end{cases}$$

Write $f(\boldsymbol{\lambda}_c, 0, \boldsymbol{\Sigma}_c)$ for the probability density function of $\boldsymbol{\lambda}_c|\mathbf{c}$.

The prior then becomes:

$$\begin{aligned}\mathbb{P}(\mathbf{c}, \boldsymbol{\lambda}_c) &= \mathbb{P}(\mathbf{c})\mathbb{P}(\boldsymbol{\lambda}_c|\mathbf{c}) \\ &= \left(\prod_{i=1}^{n} \gamma^{|c_i|}(1-2\gamma)^{(1-|c_i|)}\right) f(\boldsymbol{\lambda}_c, 0, \boldsymbol{\Sigma}_c)\end{aligned}$$

Since the vector of non-centrality parameters for all SNPs, $\boldsymbol{\lambda}$, is given by $\boldsymbol{\Sigma}\boldsymbol{\lambda}_c$, the association statistics must follow a multivariate distribution:

$$\mathbf{S} \sim N(0, \boldsymbol{\Sigma}\boldsymbol{\Sigma}_c\boldsymbol{\Sigma}) \tag{1.3}$$

and one can proceed as before.

Since the denominator of the posterior probability of causal status is computationally intractable in the generalised case, it is assumed that the total number of causal SNPs in a region is at most 6. A greedy algorithm is used: at each iteration, the SNP which most increases the posterior probability is selected as causal, until the posterior probability of the causal set is at least ρ times the total posterior probability of the data.

1.6 Structure of Thesis

In this chapter, I have introduced the diseases and data-types I will analyse in my thesis, and summarised some of the existing techniques developed for the analysis of such data.

In Chapter 2, I extend the colocalization techniques discussed in Section 1.4 to the case where there is a single common control dataset, a common approach for GWAS of related diseases, since it provides a more efficient use of samples, and hence increases the power avaliable, and use them to better understand causal structure within regions, and to find novel disease associations. By extending the Bayesian approach, I also show how multinomial models can be used to fine map variants shared between two diseases using full genotype data and under the assumption of a common single causal variant.

Chapter 3 describes a method which simulates the summary data from a GWAS. I discuss the theory behind my approach, and suggest its use for estimating the null distribution in a SNP set enrichment analysis.

Full genotype data of the sort used in Chapter 2 is often unavailable. In Chapter 4, I propose a method which, by simulating the output from GWAS under different causal models, enables the inference of causal models from summary data consisting only of p-values and which avoids the single causal variant assumption.

Finally, in Chapter 5, I discuss the future relevance of the methods presented in this thesis, and outline my thoughts regarding future directions for the genomic analysis of complex disease.

Chapter 2

Colocalization for Common Controls

2.1 Author Contributions

Work in this chapter has been published [Fortune et al, 2015], and parts of the text in the Results and Discussion sections closely mirror those in the published paper, which was jointly edited by Chris Wallace and myself. All development, coding and analysis described is entirely my own.

2.2 Motivation

As discussed in Section 1.4, there is substantial overlap in genetic regions showing association with autoimmune disease; this is strongly indicative of shared aetiology. I wanted to find out whether these overlaps correspond to truly shared variants, or whether they are due to the action of distinct variants which happen to be in physical proximity. Knowing that two diseases share a common causal variant is indicative of shared causal mechanism, and might also suggest investigation of similar treatment strategies, particularly if a known treatment for one of the diseases targets a gene product in this region related to the causal variant. Conversely, evidence of distinct causal variants may indicate that the region is associated with a divergence in pathological processes.

It may also be possible to leverage the knowledge that certain diseases have shared genetic association in order to identify novel causal variants. A known association with other related diseases gives us prior plausibility when assessing the evidence that a variant is causal for a disease of interest, even if the signal falls below the usual significance threshold in a single-disease analysis.

However, showing that a variant is associated with two traits does not demonstrate that it is causal for both, as this effect may be due to distinct causal variants in linkage disequilibrium with each other. Instead, a more formal framework which takes into account this possibility is required; this is the task of colocalization.

2.3 Application of Method: Association of *FADS2* with Crohn's Disease

I start with an example of an application of existing colocalization method to find a causal gene for Crohn's Disease (CD).

Inflammatory Bowel Disease (IBD) is a group of autoimmune diseases characterised by inflammation of the colon and small intestine. The two most common types of IBD are CD and Ulcerative Colitis (UC). eQTLs are genetic variants which affect the expression of genes. These effects are often cell-type specific, and understanding them can help us understand disease aetiology.

My collaborators, [Peters et al, 2016], mapped eQTLs in five primary immune cell types: $CD4^+$; $CD8^+$; $CD14^+$; $CD16^+$ and $CD19^+$, for patients with various autoimmune disease including IBD. They found that rs102275 appears to both be associated with CD and an eQTL for *FADS2*, a gene with known prior plausibility from mouse studies to be associated with IBD. However, this association is not sufficient to demonstrate causality, and so I carried out a colocalization test upon this region, comparing the results from the eQTL mapping to the IBD association. My results are shown in Table 2.1. In all of the immune cell types analysed, for both the CD and IBD datasets, my posterior probability of colocalization was > 0.98. By

contrast, for the UC-only dataset, my posterior probability of colocalization was < 0.005 for all cell types. This provides strong evidence that this eQTL is also causal for CD, and supports a causal role for *FADS2* in CD.

		Cell Type				
		$CD4^+$	$CD8^+$	$CD14^+$	$CD16^+$	$CD19^+$
Disease	CD	0.982	0.982	0.989	0.992	0.991
	IBD	0.993	0.993	0.987	0.991	0.992
	UC	0.00247	0.00471	0.00419	0.00145	0.000983

Table 2.1 Posterior probabilities of colocalization ($\mathbb{H}_4$) between eQTL data for given immune cell types and association with IBD for the *FADS2* region, containing candidate causal SNP rs102275. These strongly suggest causality for CD.

2.4 The Complication of a Common Control Dataset

Both the techniques described in Section 1.4 require that it be possible to model each trait using an independent regression. In order to do this, for each of our datasets, we need an independent control dataset to perform this regression upon. However, due to the cost of genotyping, it is common for GWAS of multiple diseases to maximise their power by using a single common control dataset, against which each disease dataset is compared. The Wellcome Trust Case Control Consortium [Burton et al, 2007], for instance, analysed ~ 2000 individuals with each of seven major diseases with a shared set of ~ 3000 controls. The use of a common control dataset introduces dependency between the results of each regression analysis and hence violates the underlying assumptions. One possibility is to split the controls into several independent but smaller datasets; however, this sacrifices power.

In the next section, I describe how I extended both colocalization methods to allow for the use of a common control dataset, as presented in [Fortune et al, 2015]. The code used is given in the colocCommonControl R package, which can be found at `https://github.com/mdfortune/colocCommonControl`. I then applied my colocalization

method to four autoimmune diseases: T1D; RA; CEL and MS, using data from the ImmunoChip, which provides dense coverage of 188 regions associated with at least one autoimmune disease.

2.5 Extending Methods to the Case of a Common Control

2.5.1 Common Control Extension to the Bayesian Approach

2.5.1.1 Framework

As in [Giambartolomei et al, 2014] (see Section 1.4.1), I consider the case when each trait is caused by at most one variant. Let there be $Q \geq 2$ SNPs in a region. Then there are $(Q+1)^2$ configurations of possible causal SNPs, each of which can be assigned to one of five possible hypotheses (Figure 2.1).

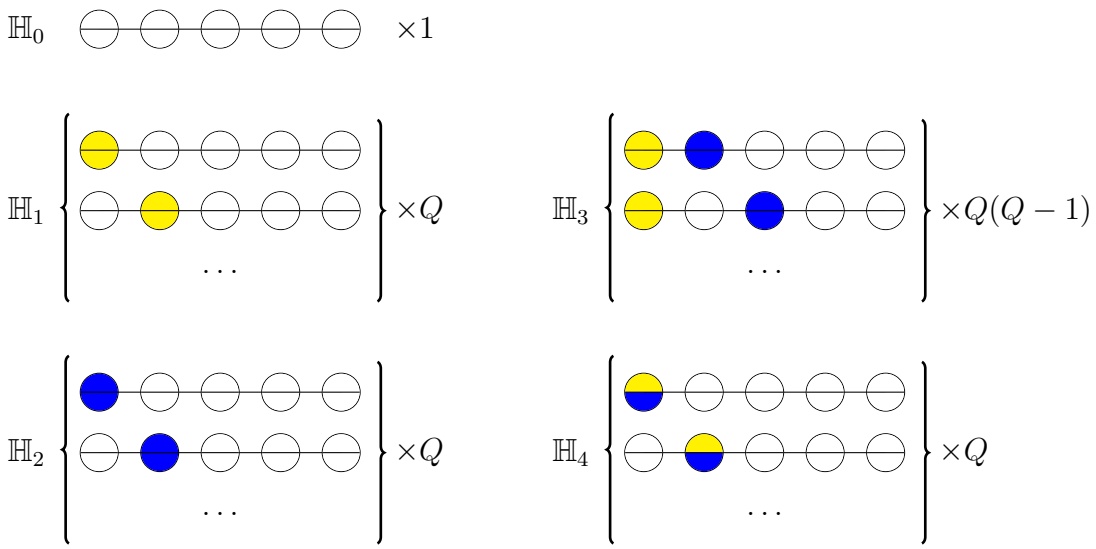


Figure 2.1 The hypotheses being tested by the Bayesian approach are represented as collections of configurations. Each configuration is represented by a line, and each circle represents one of the Q SNPs in a region under consideration. Yellow circles represent SNPs that are causal for disease 1; blue circles represent SNPs that are causal for disease 2 and yellow/blue circles represent SNPs that are causal for both diseases. We assume that at most one SNP within the region can be causal for each disease.

2.5.1.2 Choice of Priors

In the absence of any other information about a region, I considered that a prior of the same form as that in [Giambartolomei et al, 2014] should be used. Write p_1 and p_2 for the probability that a SNP is associated exclusivly with each one of the two traits. Write p_{12} for the probability that one SNP is associated with both traits and $p_0 = 1 - p_1 - p_2 - p_{12}$ be the probability that a SNP is not associated with any trait. Then our prior for each model, $\mathcal{M}$, is as in Section 1.4.1.

Following [Giambartolomei et al, 2014], I set $p_1 = p_2 = 10^{-4}$, which implies, conservatively, that ImmunoChip contains around 20 causal variants for any autoimmune disease. Consider a region with Q SNPs. Then, since the number of models $|\mathcal{M} \in \mathbb{H}_3| = \frac{Q(Q-1)}{2}$ and $|\mathcal{M} \in \mathbb{H}_4| = Q$,

we have:

$$\begin{aligned}\tau &= \mathbb{P}(\mathbb{H}_4|\mathbb{H}_3 \text{ or } \mathbb{H}_4) \\ &= \frac{Qp_0p_{12}}{Qp_0p_{12} + \frac{Q(Q-1)}{2}p_1p_2} \\ &= \frac{2p_0p_{12}}{2p_0p_{12} + (Q-1)p_1p_2}\end{aligned}$$

Since colocalization requires association with two diseases, the probability of colocalization must be less than the probability of one of the two diseases being associated (p_1). However, since autoimmune diseases are known to share genetics, colocalization must occur more frequently than random chance would predict from two independent diseases (p_1p_2). Hence, I require $p_1 = 10^{-4} > p_{12} > p_1p_2 = 10^{-8}$. Note that the above also highlights that τ decreases with increasing Q. This makes sense because with more SNPs, there is more chance for close but distinct causal variants to occur by random chance.

Of the 416 regions associated to two traits analysed in [Parkes et al, 2013], 45% were concordant (that is, $\mathbb{H}_4$ with the same direction of effect), 14% discordant (that is, $\mathbb{H}_4$ with opposite directions of effect) and 42% were not correlated (that is, $\mathbb{H}_3$). This agrees with the result of surveying colleagues, where the median suggestion was 0.5. Hence, I concluded that a sensible estimate of τ is around 0.5.

Figure 2.2 plots the value of τ generated for each of $p_{12} = 10^{-5}, p_{12} = 10^{-6}$ and $p_{12} = 10^{-7}$, together with the distribution of Q for the 126 regions studied. From this, we can see that $p_{12} = 10^{-6}$ comes closest to achieving my desired value of τ, and so it is this I chose to use in my analysis. Hence, my prior becomes:

$$\mathbb{P}(\mathcal{M}) = (1, 10^{-4}, 10^{-4}, 10^{-8}, 10^{-6})$$

Changing Priors to Reflect Information from External Publications The prior above was computed to be an appropriate choice for an arbitary region present in the ImmunoBase. It is a good default prior, allowing for comparative analysis of many regions,

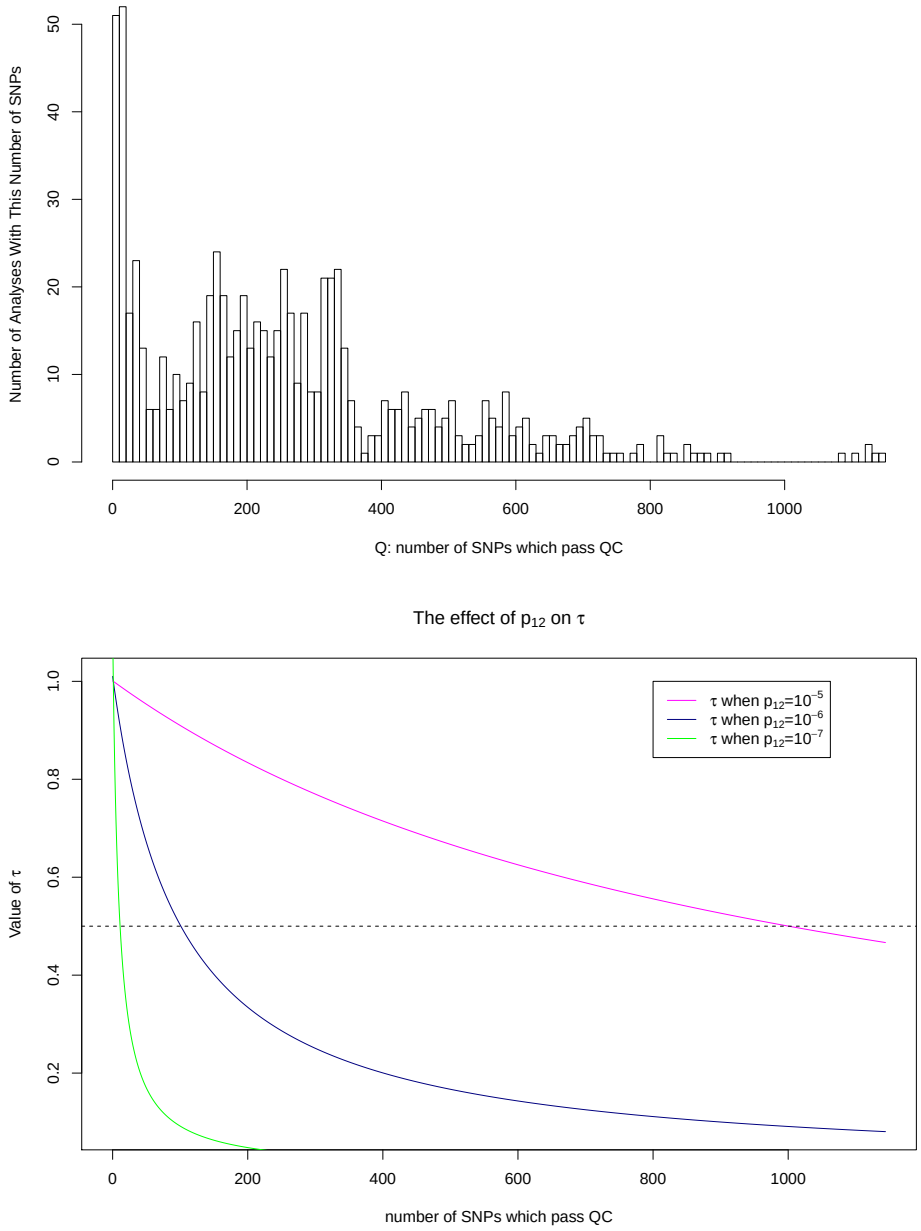


Figure 2.2 τ, the probability of colocalization, given that both traits are associated with a region. τ can be expressed as a function of Q, the number of SNPs in the region, and p_{12}, the probability of any given SNP being associated to two traits (we assume that the probability of a SNP being associated to the first trait only is held constant at 10^{-4}). The top plot gives a histogram showing the distribution of Q over all regions analysed. The lower plot shows τ plotted against $Q$for each of $p_{12} = 10^{-5}, p_{12} = 10^{-6}, p_{12} = 10^{-7}$. The dotted line shows $\tau = 0.50$, which I believe to be a reasonable average value. From this, I conclude that $p_{12} = 10^{-6}$ is the most appropriate value to use.

however, it may be improved upon, particularly for the analysis of a single, well studied, region. If a region contains only a few SNPs of interest (that is, Q is small) then we would expect $\mathbb{P}(\mathbb{H}_4|\mathbb{H}_3 \text{ or } \mathbb{H}_4)$ to be inflated, and this could be reflected in the prior. Alternativly we could have additional information, independent from the data to be analysed, regarding the likelihood of disease association with the region, which we may wish to incorporate into our prior.

For instance, my analysis was restricted to UK samples only, enabling me to assume equal linkage disequilibrium between different case cohorts (a requirement of the simple multinomial model I will use in Section 2.5.1.3). However, in the case of RA and MS, this meant that I analysed only a fraction of samples originally used; for these two traits, the published results curated in ImmunoBase, `http://www.immunobase.org`, give important additional information about the regions which I wish to incorporate into our priors. Denote this information by $\mathcal{A}$. I write $\mathbb{P}(\mathcal{A}|\mathcal{M})$ as a function of $q_- = \mathbb{P}(\text{The region is considered associated by ImmunoBase when it is not})$ and $q_+ = \mathbb{P}(\text{The region is not considered associated by ImmunoBase when it is})$. Then, I can adjust Π', the original prior for $\mathcal{M}$, in light of $\mathcal{A}$ according to:

$$\mathbb{P}(\mathcal{M}|\mathcal{A}) \propto \mathbb{P}(\mathcal{A}|\mathcal{M})\Pi'(\mathcal{M})$$

For instance, if in a region where Trait 1 is considered to be associated by ImmunoBase, and Trait 2 is not considered associated by ImmunoBase then the following priors are appropriate:

$\mathcal{M} \in \mathbb{H}_0$: $\mathbb{P}(\mathcal{M}|\mathcal{A}) \propto q_-(1-q_-)\Pi'(\mathcal{M})$

$\mathcal{M} \in \mathbb{H}_1$: $\mathbb{P}(\mathcal{M}|\mathcal{A}) \propto (1-q_+)(1-q_-)\Pi'(\mathcal{M})$

$\mathcal{M} \in \mathbb{H}_2$: $\mathbb{P}(\mathcal{M}|\mathcal{A}) \propto q_-q_+\Pi'(\mathcal{M})$

$\mathcal{M} \in \mathbb{H}_3$: $\mathbb{P}(\mathcal{M}|\mathcal{A}) \propto (1-q_+)q_+\Pi'(\mathcal{M})$

$\mathcal{M} \in \mathbb{H}_4$: $\mathbb{P}(\mathcal{M}|\mathcal{A}) \propto (1-q_+)q_+\Pi'(\mathcal{M})$

and, if in a region where Trait 1 is considered to be associated by ImmunoBase, and nothing is known about Trait 2, then the following priors are appropriate:

$\mathcal{M} \in \mathbb{H}_0$: $\mathbb{P}(\mathcal{M}|\mathcal{A}) \propto q_- \Pi'(\mathcal{M})$

$\mathcal{M} \in \mathbb{H}_1$: $\mathbb{P}(\mathcal{M}|\mathcal{A}) \propto (1-q_+)\Pi'(\mathcal{M})$

$\mathcal{M} \in \mathbb{H}_2$: $\mathbb{P}(\mathcal{M}|\mathcal{A}) \propto q_- \Pi'(\mathcal{M})$

$\mathcal{M} \in \mathbb{H}_3$: $\mathbb{P}(\mathcal{M}|\mathcal{A}) \propto (1-q_+)\Pi'(\mathcal{M})$

$\mathcal{M} \in \mathbb{H}_4$: $\mathbb{P}(\mathcal{M}|\mathcal{A}) \propto (1-q_+)\Pi'(\mathcal{M})$

In my analysis, additional information exists in ImmunoBase for RA and MS, whereas nothing is known for T1D and CEL. The criteria for inclusion of a region in ImmunoBase is quite stringent; either it has to reach genome-wide significance in a study of the trait ($p < 5 \times 10^{-8}$) or be strongly associated with the trait ($p < 10^{-5}$) and already be included for another autoimmune disease. Hence, the chance of a false positive is small, and I shall use $q_- = 0.01$. By contrast, all ImmunoBase regions were selected to have some autoimmune association. Hence, even if a region is not currently considered associated with a specific trait, due to the colocalization between diseases, the chance that there is association is quite high. Hence, I shall use $q_+ = 0.5$. Substituting these values into the equations above, I used the following priors for the RA-MS analysis:

If $\mathcal{A}$ = {No association with either trait}, then:

$\mathbb{P}(\mathcal{M}|\mathcal{A}) = (1, 5.05 \times 10^{-5}, 5.05 \times 10^{-5}, 1.02 \times 10^{-12}, 1.02 \times 10^{-10})$

If $\mathcal{A}$ = {Association with only one trait (Trait 1)}, then:

$\mathbb{P}(\mathcal{M}|\mathcal{A}) = (0.995, 4.98 \times 10^{-3}, 5.03 \times 10^{-5}, 2.51 \times 10^{-7}, 2.51 \times 10^{-5})$

If $\mathcal{A}$ = {Association with both traits}, then:

$\mathbb{P}(\mathcal{M}|\mathcal{A}) = (0.988, 4.94 \times 10^{-3}, 4.94 \times 10^{-3}, 2.47 \times 10^{-5}, 2.47 \times 10^{-3})$

And the following priors for, for instance, the T1D-RA analysis:

If $\mathcal{A}$ = {No association with RA}, then:

$\mathbb{P}(\mathcal{M}|\mathcal{A}) = (1.00, 1.00 \times 10^{-4}, 5.05 \times 10^{-5}, 5.05 \times 10^{-9}, 5.05 \times 10^{-7})$

If $\mathcal{A}$ = {Association with RA}, then:

$\mathbb{P}(\mathcal{M}|\mathcal{A}) = (0.995, 9.95 \times 10^{-5}, 4.98 \times 10^{-3}, 4.98 \times 10^{-7}, 4.98 \times 10^{-5})$

Similar adjustments are made for comparison of RA or MS with either T1D or CEL.

2.5.1.3 Computation of Posterior Probabilities

Since there is a shared control dataset, I cannot model the two traits with independent logistic regression models. Instead, let n be the total number of samples. Let S be the n by Q genotype matrix. Let Y be the vector of length n giving the disease association for each sample, (with $Y_k = 0$ corresponding to individual k being from the control group, $Y_k = 1$ corresponding to trait 1, and $Y_k = 2$ corresponding to trait 2. Then I model the configuration corresponding to SNP i being causal for disease 1 and SNP j being causal for disease 2 using the multinomial model:

$$\frac{\ln(\mathbb{P}(Y_k = 1))}{\ln(\mathbb{P}(Y_k = 0))} = \beta_0^1 + \beta_1^1 S_{ki} \qquad \frac{\ln(\mathbb{P}(Y_k = 2))}{\ln(\mathbb{P}(Y_k = 0))} = \beta_0^2 + \beta_1^2 S_{kj} \qquad \forall k$$

Using the Begg and Gray approximation [Begg and Gray, 1984], this can be converted into a binary logistic regression, as implemented in the R package *mlogitBMA* (`http://cran.r-project.org/web/packages/mlogitBMA/index.html`). I then use a Laplace approximation to compute approximate Bayes Factors for each configuration [Raftery, 1996], using the implementation in the R package *BMA* (`https://cran.r-project.org/web/packages/BMA/BMA.pdf`).

Let $\mathcal{C}_i$ be a configuration, and let $\mathcal{C}_0$ denote the configuration in which no SNP is causal for either trait. Let $\mathcal{D}$ denote the entire dataset, and let $\mathcal{D}_i$ denote the dataset restricted to only those SNPS in $\mathcal{C}_i$. Note that the Bayes Factor for $\mathcal{C}_i$ is computed using only $\mathcal{D}_i$. However, following the technique used in [Maller et al, 2012], we have that BF_i, the Bayes Factor for $\mathcal{C}_i$

is:

$$\begin{aligned} BF_i \text{ (as computed)} &= \frac{\mathbb{P}(\mathcal{D}_i|\mathcal{C}_i)}{\mathbb{P}(\mathcal{D}_i|\mathcal{C}_0)} \\ &= \frac{\mathbb{P}(\mathcal{D}_i|\mathcal{C}_i)\mathbb{P}(\mathcal{D}|\mathcal{C}_0)}{\mathbb{P}(\mathcal{D}|\mathcal{C}_0)\mathbb{P}(\mathcal{D}_i|\mathcal{C}_0)} \\ &= \frac{\mathbb{P}(\mathcal{D}_i|\mathcal{C}_i)\mathbb{P}(\mathcal{D}_{-i}|\mathcal{D}_i,\mathcal{C}_0)\mathbb{P}(\mathcal{D}_i|\mathcal{C}_0)}{\mathbb{P}(\mathcal{D}|\mathcal{C}_0)\mathbb{P}(\mathcal{D}_i|\mathcal{C}_0)} \\ &= \frac{\mathbb{P}(\mathcal{D}_i|\mathcal{C}_i)\mathbb{P}(\mathcal{D}_{-i}|\mathcal{D}_i,\mathcal{C}_0)}{\mathbb{P}(\mathcal{D}|\mathcal{C}_0)} \\ &= \frac{\mathbb{P}(\mathcal{D}_i|\mathcal{C}_i)\mathbb{P}(\mathcal{D}_{-i}|\mathcal{D}_i,\mathcal{C}_i)}{\mathbb{P}(\mathcal{D}|\mathcal{C}_0)} \\ &= \frac{\mathbb{P}(\mathcal{D}|\mathcal{C}_i)}{\mathbb{P}(\mathcal{D}|\mathcal{C}_0)} \end{aligned}$$

And hence, the Bayes Factor computed using the restricted dataset is identical to that computed using the entire dataset.

I then compute the posterior probabilities of a hypothesis $\mathbb{H}$ made of configurations $\mathcal{C}_1, ..., \mathcal{C}_n$ as follows:

$$\begin{aligned} \mathbb{P}(\mathbb{H}|\mathcal{D}) &= \sum_{i=1}^{n} \mathbb{P}(\mathcal{C}_i|\mathcal{D}) \\ &\propto \sum_{i=1}^{n} \mathbb{P}(\mathcal{D}|\mathcal{C}_i)\Pi(\mathcal{C}_i) \qquad \text{(Bayes' Theorem)} \\ &\propto \Pi(\mathcal{C}|\mathcal{C} \in \mathbb{H}) \sum_{i=1}^{n} BF(\mathcal{C}_i) \\ &\quad \text{(since each configuration within the Hypothesis has an identical prior)} \end{aligned}$$

Use of Different Priors By computing the Bayes factors for each configuration, and then multiplying by the prior, I am able to apply many different priors without increasing the computational time of the algorithm, enabling exploration of sensitivity to prior.

The results presented here reflect my prior beliefs about the relative hypotheses in the case of autoimmune disease; these may not be appropriate for a different set of traits. However,

so long as each configuration within a hypothesis has an equal prior probability, it is simple to derive posterior probabilities for alternative priors without re-computing the Bayes Factors. Write Π for the original prior, and Π^a for the alternative prior. If the original posterior probabilities are $\mathbb{P}(\mathbb{H}|\mathcal{D})$, then we can compute the alternative posterior probabilities using:

$$\begin{aligned}
\mathbb{P}^a(\mathbb{H}|\mathcal{D}) &\propto \Pi^a(\mathcal{C}|\mathcal{C} \in \mathbb{H}) \sum_{i=1}^{n} BF(\mathcal{C}_i) \\
&\propto \frac{\Pi^a(\mathcal{C}|\mathcal{C} \in \mathbb{H})\Pi(\mathcal{C}|\mathcal{C} \in \mathbb{H})}{\Pi(\mathcal{C}|\mathcal{C} \in \mathbb{H})} \sum_{i=1}^{n} BF(\mathcal{C}_i) \\
&\propto \frac{\Pi^a(\mathcal{C}|\mathcal{C} \in \mathbb{H})}{\Pi(\mathcal{C}|\mathcal{C} \in \mathbb{H})} \mathbb{P}(\mathbb{H}|\mathcal{D})
\end{aligned}$$

and hence

$$\mathbb{P}^a(\mathbb{H}|\mathcal{D}) = \frac{\frac{\Pi^a(\mathcal{C}|\mathcal{C}\in\mathbb{H})}{\Pi(\mathcal{C}|\mathcal{C}\in\mathbb{H})}\mathbb{P}(\mathbb{H}|\mathcal{D})}{\sum_{\mathbb{H}'}\left(\frac{\Pi^a(\mathcal{C}|\mathcal{C}\in\mathbb{H}')}{\Pi(\mathcal{C}|\mathcal{C}\in\mathbb{H}')}\mathbb{P}(\mathbb{H}'|\mathcal{D})\right)}$$

Use of this Method for Fine Mapping To find evidence of genetic association between the SNPs at a region and a trait is suggestive that one of the genes near the SNP is causal. However, this need not be the case. Even if the locus contains a causal SNP, it could be, for instance, in a regulatory element which modifies the expression of a gene some distance away. However, if we know the identity of the causal variants in the region, we can incorporate information from other datasets to make inferences about the causal genes. For instance, a chromosome conformation capture analysis of a causal variant enables us to map the genetic regions it interacts with [Davison et al, 2012; Martin et al, 2015]. Hence, we are interested in fine mapping any region which shows evidence of disease association in search of likely causal SNPs. The Bayesian approach enables such analysis with minimal additional computational time and borrows power by using information from both traits. Since:

$$\mathbb{P}(\mathcal{C}_i|\mathbb{H}) \propto BF(\mathcal{C}_i|\mathbb{H})$$

the probability of an individual SNP being causal given a hypothesis being true is proportional to the contribution of the corresponding configuration to the summed Bayes factor.

2.5.1.4 The Use of Tagging to Speed Computational Time

To speed computation, I used tagging; SNPs are represented by others with which they have $r^2 > 0.99$.

Write $\{SNP_i, SNP_j\}$ for the model where trait 1 is caused by SNP_i and trait 2 is caused by SNP_j. Write $\{SNP_i, 0\}$ for the model where trait 1 is caused by SNP_i and trait 2 is not caused by any SNP. If SNP_1 and SNP_2 are in the same tag set, I need only compute the Bayes factor for model $\{SNP_1, SNP_1\}$, B_{11}, and I assume that the Bayes Factors for models $\{SNP_1, SNP_2\}$, $\{SNP_2, SNP_1\}$ and $\{SNP_2, SNP_2\}$ can be approximated by B_{11}. Although this decreases the number of models we need to analyze, it increases the complexity of associating models with hypotheses. If Tag_i is of size n_i, then the model $\{Tag_i, Tag_i\}$ corresponds to n_i models in $\mathbb{H}_4$ and $n_i(n_i - 1)$ models in $\mathbb{H}_3$.

2.5.1.5 Extension to More than Two Traits

Conceptually, the framework of the Bayesian analysis is easy to extend to more than two traits, or to allowing multiple causal variants. However, this greatly increases the hypothesis space, making computation and interpretation significantly more complex. For instance, in the case of three traits, there are fifteen possible hypotheses to consider:

$\mathbb{H}_0$: No SNP is associated with any trait.

$\mathbb{H}_1$: Only trait 1 is associated with any SNP.

$\mathbb{H}_2$: Only trait 2 is associated with any SNP.

$\mathbb{H}_3$: Only trait 3 is associated with any SNP.

$\mathbb{H}_4$: Trait 1 and trait 2 are associated with different SNPs; trait 3 is not associated with any SNP.

$\mathbb{H}_5$: Trait 1 and trait 3 are associated with different SNPs; trait 2 is not associated with any SNP.

$\mathbb{H}_6$: Trait 2 and trait 3 are associated with different SNPs; trait 1 is not associated with any SNP.

$\mathbb{H}_7$: Trait 1 and trait 2 are colocated; trait 3 is not associated with any SNP.

$\mathbb{H}_8$: Trait 1 and trait 3 are colocated; trait 2 is not associated with any SNP.

$\mathbb{H}_9$: Trait 2 and trait 3 are colocated; trait 1 is not associated with any SNP.

$\mathbb{H}_{10}$: All traits are associated, but with different SNPs.

$\mathbb{H}_{11}$: Trait 1 and trait 2 are colocated; trait 3 is not, but is associated with a different SNP.

$\mathbb{H}_{12}$: Trait 1 and trait 3 are colocated; trait 2 is not, but is associated with a different SNP.

$\mathbb{H}_{13}$: Trait 2 and trait 3 are colocated; trait 1 is not, but is associated with a different SNP.

$\mathbb{H}_{14}$: All traits are colocated.

In practice, I used pairwise analysis of traits.

2.5.1.6 Conditional Extension to the Bayesian Approach

The Bayesian method assumes that each trait is caused by at most one variant; in some regions, this is not a realistic assumption. Hence, I developed an extension of the Bayesian approach which allows us to analyse a region where some SNPs are already known to be associated with the traits. The multinomial model is used as before, however each configuration analysed now contains all the known causal SNPs, and we investigate the effect of including at most one additional causal SNP for each trait. Hence, using the notation above, if SNPs $a_1, ...a_A$ are already (or assumed) known to be causal for trait 1, and SNPs $b_1, ..., b_B$ are already known (or assumed) to be causal for trait 2, then we test the configuration that SNP_i is additionally causal for trait 1, and SNP_j is additionally causal for trait 2, using the model:

$$\frac{\ln(\mathbb{P}(Y_k = 1))}{\ln(\mathbb{P}(Y_k = 0))} = \beta_0^1 + \beta_1^1 S_{ki} + \gamma_1^1 S_{ka_1} + ... + \gamma_A^1 S_{ka_A}$$

$$\frac{\ln(\mathbb{P}(Y_k = 2))}{\ln(\mathbb{P}(Y_k = 0))} = \beta_0^2 + \beta_1^2 S_{kj} + \gamma_1^2 S_{kb_1} + ... + \gamma_B^2 S_{kb_B}$$

$$\forall k$$

This method is run iteratively for each region until the configuration containing no additional SNPs (i.e. $\mathbb{H}_0$) is the configuration preferred.

2.5.2 Common Control Extension to the Proportional Approach

2.5.2.1 Test for Colocalization

As in the case of independent controls (Section 1.4.2), I write $\mathbf{Y}_1$ and $\mathbf{Y}_2$ for the phenotypes of the two traits and $\hat{\beta}_1$, $\hat{\beta}_2$ for the maximum likelihood estimators of β_1 and β_2, the coefficients obtained when the $\mathbf{Y}$s are regressed upon Q explanatory SNPs. Let β_1 and β_2 have covariance matrices $\mathbf{V}_{11}$ and $\mathbf{V}_{22}$ respectively.

The null hypothesis of proportionality can be expressed as:

$$\beta_1 = \frac{\beta_2}{\eta} \Leftrightarrow \beta_1 - \frac{\beta_2}{\eta} = 0$$

Since there is a shared control, the regressions of the two traits are no longer independent and so β_1 and β_2 have some non-zero covariance matrix $\mathbf{V}_{12}$. Write $\mathbf{V}_{21}$=$\mathbf{V}_{12}^T$. By asymptotic normality of maximum likelihood estimators:

$$\begin{pmatrix} \hat{\beta}_1 \\ \hat{\beta}_2 \end{pmatrix} \sim N\left(\begin{pmatrix} \beta_1 \\ \beta_2 \end{pmatrix}, \begin{pmatrix} V_{11} & V_{12} \\ V_{21} & V_{22} \end{pmatrix} \right)$$

A linear transformation gives:

$$\hat{\beta}_1 - \frac{\hat{\beta}_2}{\eta} \sim N\left(\beta_1 - \frac{1}{\eta}\beta_2, V_{11} - \frac{1}{\eta}V_{12} - \frac{1}{\eta}V_{21} + \frac{1}{\eta^2}V_{22}\right)$$

Hence, under the null hypothesis:

$$\left(\hat{\beta}_1 - \frac{1}{\eta}\hat{\beta}_2\right)^T \left(V_{11} - \frac{1}{\eta}V_{12} - \frac{1}{\eta}V_{21} + \frac{1}{\eta^2}V_{22}\right)^{-1} \left(\hat{\beta}_1 - \frac{1}{\eta}\hat{\beta}_2\right) \sim \chi^2_Q$$

Or, writing $\eta = \tan(\theta)$:

$$\begin{aligned} T(\theta) &= \left(\sin(\theta)\hat{\beta}_1 - \cos(\theta)\hat{\beta}_2\right)^T \\ &\quad \left(\sin^2(\theta)V_{11} - \sin(\theta)\cos(\theta)V_{12} - \sin(\theta)\cos(\theta)V_{21} + \cos^2(\theta)V_{22}\right)^{-1} \\ &\quad \left(\sin(\theta)\hat{\beta}_1 - \cos(\theta)\hat{\beta}_2\right) \\ &\sim \chi^2_Q \end{aligned}$$

giving us a test statistic for a given value of θ.

Since the true value of θ is unknown, I am unable to directly compute a p-value for this test statistic. Instead, write $\mathcal{P}(\theta)$ for the posterior distribution of θ given $\hat{\beta}_1$ and $\hat{\beta}_2$. Write $T^*(\theta)$ for the p-value derived at a specific value of θ. Then the posterior predictive p-value for testing the null hypothesis of colocalization is given by:

$$\int_0^\pi T^*(\theta)\mathcal{P}(\theta)d\theta$$

2.5.2.2 Computation of $\mathcal{P}(\theta)$, the Posterior Distribution of the Proportionality Constant

In the absence of any additional information, uninformative priors $\pi(\theta) \sim 1$ and $\pi(\beta) \sim \mathbf{1}$ are used.

Let $V = \begin{pmatrix} V_{11} & V_{12} \\ V_{21} & V_{22} \end{pmatrix}$ and let $W = \begin{pmatrix} W_{11} & W_{12} \\ W_{21} & W_{22} \end{pmatrix} = V^{-1}$.

Under proportionality, there exists β such that $\beta_1 = \beta\cos(\theta)$ and $\beta_2 = \beta\sin(\theta)$, and hence I can write the likelihood of $\hat{\beta}_1, \hat{\beta}_2$, given θ, β, as:

$$\mathcal{L}(\hat{\beta}_1, \hat{\beta}_2|\beta, \theta) = \frac{1}{(2\pi)^{\frac{Q}{2}}(|V|)^{\frac{1}{2}}} \exp\left[-\frac{1}{2}\begin{pmatrix} \hat{\beta}_1 - \beta\cos(\theta) \\ \hat{\beta}_2 - \beta\sin(\theta) \end{pmatrix}^T \begin{pmatrix} W_{11} & W_{12} \\ W_{21} & W_{22} \end{pmatrix} \begin{pmatrix} \hat{\beta}_1 - \beta\cos(\theta) \\ \hat{\beta}_2 - \beta\sin(\theta) \end{pmatrix}\right]$$

Write:

$$x = \cos(\theta)W_{11}\hat{\beta}_1 + \cos(\theta)W_{12}\hat{\beta}_2 + \sin(\theta)W_{21}\hat{\beta}_1 + \sin(\theta)W_{22}\hat{\beta}_2$$

$$X = \cos^2(\theta)W_{11} + \sin(\theta)\cos(\theta)W_{12} + \sin(\theta)\cos(\theta)W_{21} + \sin^2(\theta)W_{22}$$

$$\mu = X^{-1}x$$

Then:

$$\begin{aligned} \mathcal{L}(\hat{\beta}_1, \hat{\beta}_2|\beta, \theta) &= \frac{1}{(2\pi)^{\frac{Q}{2}}(|V|)^{\frac{1}{2}}} \exp\left[-\frac{1}{2}\left(\begin{pmatrix} \hat{\beta}_1 \\ \hat{\beta}_2 \end{pmatrix}^T W \begin{pmatrix} \hat{\beta}_1 \\ \hat{\beta}_2 \end{pmatrix} - \beta^T x - x^T\beta + \beta^T X\beta\right)\right] \\ &= \frac{1}{(2\pi)^{\frac{Q}{2}}(|V|)^{\frac{1}{2}}} \exp\left[-\frac{1}{2}\left(\begin{pmatrix} \hat{\beta}_1 \\ \hat{\beta}_2 \end{pmatrix}^T W \begin{pmatrix} \hat{\beta}_1 \\ \hat{\beta}_2 \end{pmatrix} + (\beta-\mu)^T X(\beta-\mu) - \mu^T X\mu\right)\right] \end{aligned}$$

And so we can compute:

$$\begin{aligned} \mathcal{P}(\theta) &\propto \int_\beta \mathcal{L}(\hat{\beta}_1, \hat{\beta}_2|\beta, \theta)\pi(\theta)\pi(\beta)d\beta \\ &\propto \frac{\pi(\theta)}{(2\pi)^{\frac{Q}{2}}(|V|)^{\frac{1}{2}}} \exp\left[-\frac{1}{2}\left(\begin{pmatrix} \hat{\beta}_1 \\ \hat{\beta}_2 \end{pmatrix}^T W \begin{pmatrix} \hat{\beta}_1 \\ \hat{\beta}_2 \end{pmatrix} - \mu^T X\mu\right)\right] \\ &\qquad \times \int_\beta \exp\left[-\frac{1}{2}\left((\beta-\mu)^T X(\beta-\mu)\right)\right] d\beta \\ &\propto \frac{\pi(\theta)}{(2\pi)^{\frac{Q}{2}}(|V|)^{\frac{1}{2}}} \exp\left[-\frac{1}{2}\left(\begin{pmatrix} \hat{\beta}_1 \\ \hat{\beta}_2 \end{pmatrix}^T W \begin{pmatrix} \hat{\beta}_1 \\ \hat{\beta}_2 \end{pmatrix} - \mu^T X\mu\right)\right] \end{aligned}$$

2.5.2.3 Test Statistic Used

Hence, the posterior predictive p-value to test the null hypothesis of proportionality is given by:

$$\int_0^{\pi} T^*(\theta) \frac{\pi(\theta)}{(2\pi)^{\frac{Q}{2}}(|V|)^{\frac{1}{2}}} \exp\left[-\frac{1}{2}\left(\begin{pmatrix}\hat{\beta}_1\\ \hat{\beta}_2\end{pmatrix}^T W \begin{pmatrix}\hat{\beta}_1\\ \hat{\beta}_2\end{pmatrix} - \mu^T X \mu\right)\right] d\theta$$

where $T^*(\theta)$ is the p-value obtained when

$$\left(\sin(\theta)\hat{\beta}_1 - \cos(\theta)\hat{\beta}_2\right)^T \left(\sin^2(\theta)V_{11} - \sin(\theta)\cos(\theta)V_{12} - \sin(\theta)\cos(\theta)V_{21} + \cos^2(\theta)V_{22}\right)^{-1} \left(\sin(\theta)\hat{\beta}_1 - \cos(\theta)\hat{\beta}_2\right)$$

is compared to a χ^2_Q distribution.

2.5.3 Comparison of the Bayesian and Proportional Approaches

A limitation of the Bayesian approach is its assumption of single causal variants for each trait. By contrast, the proportional approach allows the assumption of an arbitrary number of causal variants (although in practice this is restricted, as discussed in Section 1.4.2.3). However, the proportional null hypothesis does not correspond to $\mathbb{H}_4$ from the Bayesian approach; it corresponds to colocalization, single-disease association or association with neither disease. A failure to reject the null hypothesis could also be caused by insufficient power. However, in the proportional approach, having the power to reject this null hypothesis corresponds to strong evidence for separate SNP effects ($\mathbb{H}_3$ in the notation of the Bayesian approach).

2.6 Details of Datasets

2.6.1 Samples

All samples included in this analysis were gathered in the UK and have reported or self-declared European ancestry. Informed consent was obtained from all subjects after approval from

the ethics committee or institutional review board of all participating institutions. Detailed summaries of the sample cohorts are given in the ImmunoChip reports for celiac disease [Trynka et al, 2012], rheumatoid arthritis [Eyre et al, 2012], multiple sclerosis [Beecham et al, 2013] and T1D [Onengut-Gumuscu et al, 2015]. For rheumatoid arthritis and multiple sclerosis, we used the subset of cases from the UK. Sample exclusions were applied as described in each paper; in total, 6,691 T1D, 3,870 rheumatoid arthritis, 7,987 celiac disease, 5,112 multiple sclerosis and 12,370 control samples were analysed. SNPs were filtered to meet the following criteria: call rate > 0.99; minor allele frequency > 0.01; Hardy–Weinberg $|Z| < 5$. SNPs that passed these thresholds in controls and any specific pair of cases were used for that pairwise analysis.

I excluded low frequency variants (MAF$< 1\%$), both to reduce the number of models to be considered and because genotyping errors are more common amongst this group of SNPs, and I did not have cluster plots available for all diseases. Although GWAS typically have sufficient power to detect association only with more common SNPs, some rarer variants (for example, in *TYK2* [Mero et al, 2010]) have been reported with these diseases which will be missed in my analysis.

Using only UK cases and controls means that I expect any effects of population stratification to be very limited, as evidenced by the low genomic inflation factors in published UK ImmunoChip analyses [Cooper et al, 2012], and I did not take any further specific actions to limit effects from population stratification.

2.6.2 Selection of Regions for Analysis

I considered all regions annotated in ImmunoBase (accessed 11 December 2013) to be associated with at least one of the four diseases we studied. Where regions overlapped, I formed the union. Regions containing fewer than 10 SNPs or with a SNP density of < 1 SNP/kb were excluded. The major histocompatibility complex (MHC) region (chr6:29,797,978–33,606,563, hg18) was removed from the analysis because this region is known to have complex multi-SNP effects. A full list of the 126 regions analysed, together with our resulting associations, can be found in Appendix A.

2.6.3 Identification of Disease-Specific Regions

To examine evidence for GWAS association with other, non-autoimmune, traits, I took the index SNP with the smallest p-values in a region and then identified proxy SNPs on the basis of LD ($r^2 \geq 0.9$) using 1000 Genomes Project EUR (European) data. I used these SNPs as a query set to examine associations annotated in the US National Institutes of Health GWAS catalog (accessed 10 July 2014).

I defined disease-specific regions as those for which (i) the posterior probability of single-SNP association was >0.5; (ii) the posterior probability of association with any other disease was <0.2; (iii) the region was not annotated as associated with any other autoimmune disease in ImmunoBase; and (iv) no proxies for the index SNP were associated with any other autoimmune disease in the US National Institutes of Health GWAS catalog.

2.6.4 Type 2 Diabetes Data

Summary from a T2D GWAS meta–analysis [Morris et al, 2012] was downloaded from the DIAGRAM website (accessed 20/10/14).

2.7 Results

2.7.1 Overview of Results

The Bayesian approach assumes a single causal variant per trait in any region. To allow for multiple causal variants, I used a stepwise method. In the overwhelming majority of cases (740 out of 756 pairwise comparisons, or 98%) data were consistent with at most one causal variant per trait in a region. Ninety of the 126 regions (71%) showed association with at least one disease; in 33 regions, the association was shared between at least two diseases (Figure 2.3). Complete results can be found in Appendix A, and in Supplementary Table 2 and Supplementary Table 3 of [Fortune et al, 2015] (`http://bit.ly/2d7KvUO`). For fifty–seven regions, the greatest support was for association with precisely one of the four diseases; in 21

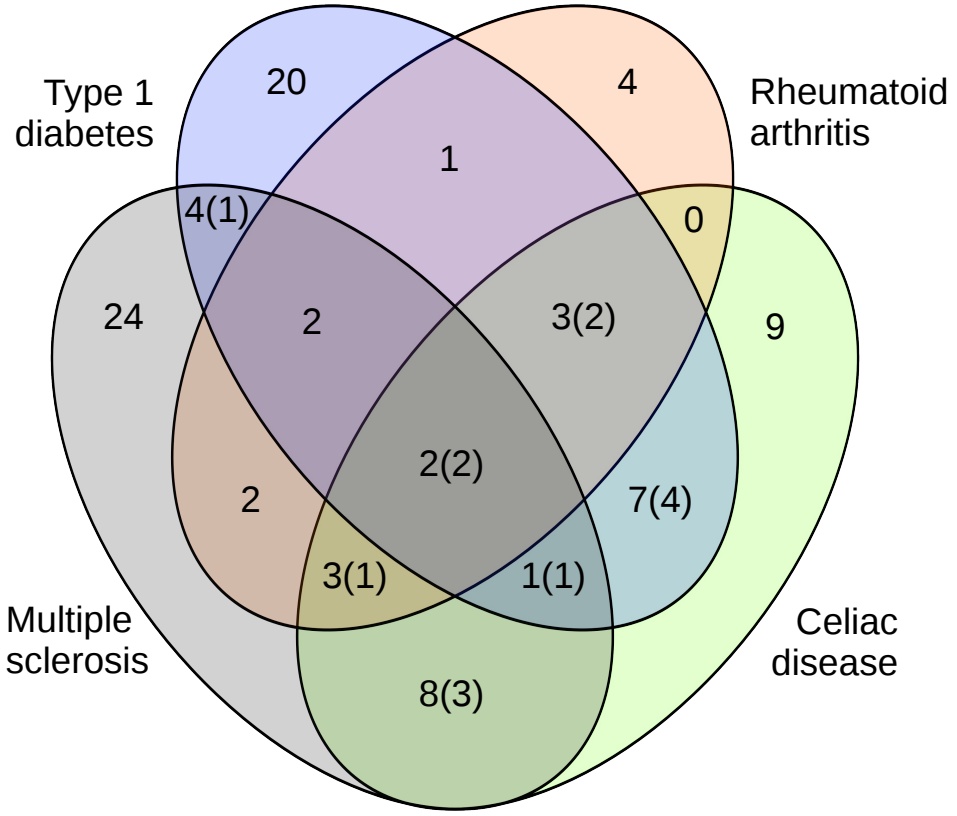


Figure 2.3 A Venn diagram a showing summary of disease assignments to 90 regions which showed association to at least one disease, based upon the results of the Bayesian analysis. In cases where assignment was uncertain, the assignment most supported by the posterior probabilities was used. The numbers in brackets correspond to how many of these regions show evidence of distinct causal variants. Thirty six regions analysed did not demonstrate association to any disease within my available data, and so are not included in this figure.

cases, I know of no other immune–mediated diseases that have reported association to these regions and therefore hypothesize these may be disease specific among autoimmune diseases (Table 2.2).

In the Bayesian approach, when the posterior probability of a hypothesis is close to 0.5, assignment cannot be made with confidence to any single hypothesis. However, in the 30 instances in which both diseases showed very strong evidence of association ($\mathbb{P}(\mathbb{H}_3$ or $\mathbb{H}_4) > 0.9$), the Bayesian and proportional approaches produced consistent results. For these 30 cases, the proportional null was rejected only in cases in which the Bayesian analysis favoured $\mathbb{H}_3$, and

not rejected in cases where $\mathbb{H}_4$ was favoured. Focusing on these, the data strongly supported that the same causal variants underlie all diseases in ten cases, while seven showed strong evidence for distinct variants, suggesting that just under half, 42%, of overlapping association signals reflect distinct causal variants. In total, fourteen regions showed evidence of separate SNP effects ($\mathbb{P}(\mathbb{H}_3 > 0.5)$, see (Table 2.3).

Chromosome	Position (bp)	Disease Association	Posterior Probability of Single Association	Candidate Causal Gene(s) (Gene(s) in Region)
1p22.1	92023171-93311800	Multiple sclerosis	0.9981321	*EVI5*
1p21.2	100982239-101455699	Multiple sclerosis	0.57339	*EXTL2 VCAM1 SLC30A7*
1p13.1	116831830-116911865	Multiple sclerosis	0.9998239	*CD58*
3p24.1	28015774-28105476	Multiple sclerosis	0.9937012	(*CMC1*)
3q13.33	122818149-123329522	Multiple sclerosis	0.9998044	*IQCB1 SLC15A2 CD86*
5q21.1	102062861-102777130	Rheumatoid arthritis	0.5821915	*C5orf30*
6q23.3	137348296-137587799	Multiple sclerosis	0.9958746	*IL22RA2*
7p12.2	50337180-50662811	Type 1 Diabetes	0.9718656	3' *IKZF1* region *
7p12.2	50866661-51640000	Type 1 Diabetes	0.9951803	*COBL*
8q21.12	79575897-79914680	Multiple sclerosis	0.9981433	*ZC2HC1A*
8q24.21	129187117-129368419	Multiple sclerosis	0.5095271	*PVT1 MIR1208*
9p24.2	4218549-4311558	Type 1 Diabetes	0.9964941	*GLIS3*
10q23.31	89998026-90268360	Type 1 Diabetes	0.8667183	*RNLS*
11p15.5	2024999-2264880	Type 1 Diabetes	0.9997444	*INS*
12q24.31	121926103-122574026	Multiple sclerosis	0.5895478	*PITPNM2*
14q32.2	100357783-100398492	Type 1 Diabetes	0.9812239	*DLK1*
16q23.1	73760230-74086012	Type 1 Diabetes	0.9986112	*BCAR1*
19p13.3	6564831-6636304	Multiple sclerosis	0.9999355	*TNFSF14*
19p13.11	16300497-16612240	Multiple sclerosis	0.9995059	(*EPS15L1*, *CALR3*, *MED26*, *C19orf44*, *CHERP*, *SLC35E1*)
19p13.11	17905598-18272802	Multiple sclerosis	0.8722256	*MPV17L2 IFI30*
20p13	1444472-1707590	Type 1 Diabetes	0.9906156	(*SIRPD*, *SIRPB1*, *SIRPG*)

Table 2.2 Twenty–one regions which are most likely disease specific under my analysis and for which I know of no other immune–mediated diseases (from the 15 diseases curated in ImmunoBase) that have reported association to these regions (as curated in ImmunoBase, accessed July 9th 2014, and NIHR GWAS catalog, accessed 07/10/2014). Regions required posterior probability of single disease association > 0.5 in at least one pairwise analysis (SNP coverage varies between analyses) and posterior probability of association to any other disorder < 0.2. Candidate causal genes are given. In the case where no candidate causal genes are known, I have given, in brackets, the genes in and around the region.
*There are two ImmunoChip regions which overlap *IKZF1* and are separated by a recombination hotspot. The region towards the 5' end has colocalizing associations with MS and T1D while the region towards the 3' end appears specific to T1D, as shown in Figure 2.11. Note I provide coordinates of the region, and not an index SNP as is conventional in GWAS because the method synthesises information across the whole region and does not, in most cases, highlight a single SNP responsible for the association.

Chromosome	Position	Associations	Evidence	Candidate Causal Gene
2p16.1	60722116-61952276	C|M	CM:H3∼0.65	*REL*
2q32.2	191412527-191739472	RC|M	RM:H3∼0.51	*STAT1 STAT4*
2q33.1	202920548-204528303	D|C|R	DR:H3∼0.98 RC:H3∼0.91	*CD28 CTLA4 ICOS*
3p21.31	45812888-46633741	D|C	DC:H3∼0.92	*CCR3 CCR1 CCR5*
3q25.33	160950948-161389020	C|M	CM:H3∼0.96	*IL12A*
4q27	123121079-124497235	D|C	DC:H3∼1.00	*IL2 IL21*
6q23.3	137914792-138345363	DRC|M	RM:H3∼0.75 CM:H3∼0.85	*TNFAIP3*
10p15.1	6068495-6237542	D|M	DM:H3∼1.00	*IL2RA*
11q23.3	117805448-118403529	C|M	CM:H3∼0.82	*CXCR5*
13q32.3	98723872-99034738	D|C	DC:H3∼0.67	*GPR183*
16p13.13	10831557-11408130	DM|C	DC:H3∼0.51	*DEXI SOCS1*
18p11.21	12407903-12919721	D|C	DC:H3∼0.58	*PTPN2*
19p13.2	10081000-11019034	DRM|C	DC:H3∼0.53	*ICAM1 ICAM3 TYK2*
21q22.3	42681877-42771181	D|R|C	DR:H3∼0.77 DC:H3∼0.99 RC:H3∼0.69	*UBASH3A*

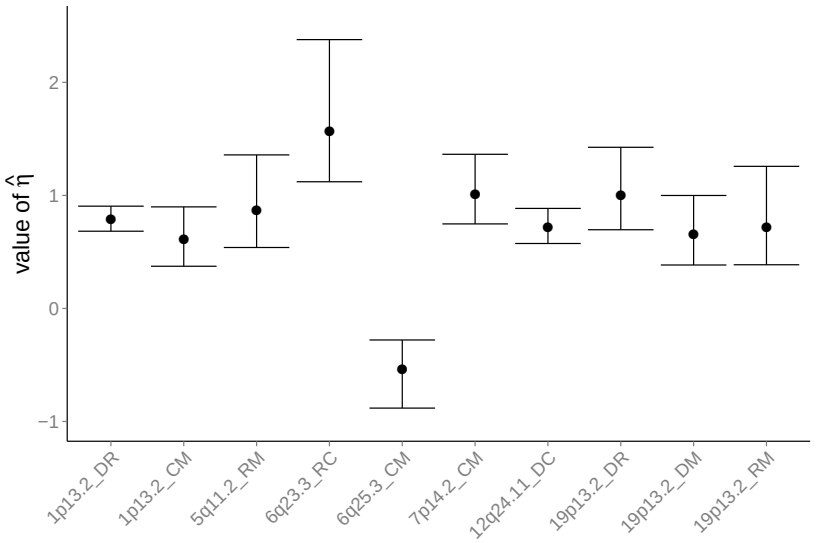
Table 2.3 Fourteen regions showing evidence of separate SNP effects ($\mathbb{P}(\mathbb{H}_3) > 0.5$).D corresponds to T1D, R to RA, C to CEL and M to MS. Candidate causal genes are as associated across all curated diseases by ImmunoBase. Distinct signals are indicated by '| '. Many of these regions are associated with other diseases (see ImmunoBase). For instance, the 2q32.2 region is additionally associated with Ulcerative Colitis, Crohn's Disease, Primary Biliary Cirrhosis, Systemic Lupus Erythematosus and Juvenile Idiopathic Arthritis. The 6q23.3 region is additionally associated with Ulcerative Colitis, Systemic Lupus Erythematosus and Psoriasis. Note that in some regions, such as 10p15.1, the conditional analysis supports the existence of multiple associated variants: if none of these overlap, then I consider the region to have separate SNP effects. Note I provide coordinates of the region, and not an index SNP as is conventional in GWAS studies because the method synthesises information across the whole region and does not, in most cases, highlight a single SNP responsible for the association.

2.7.2 Disentangling Patterns of Association

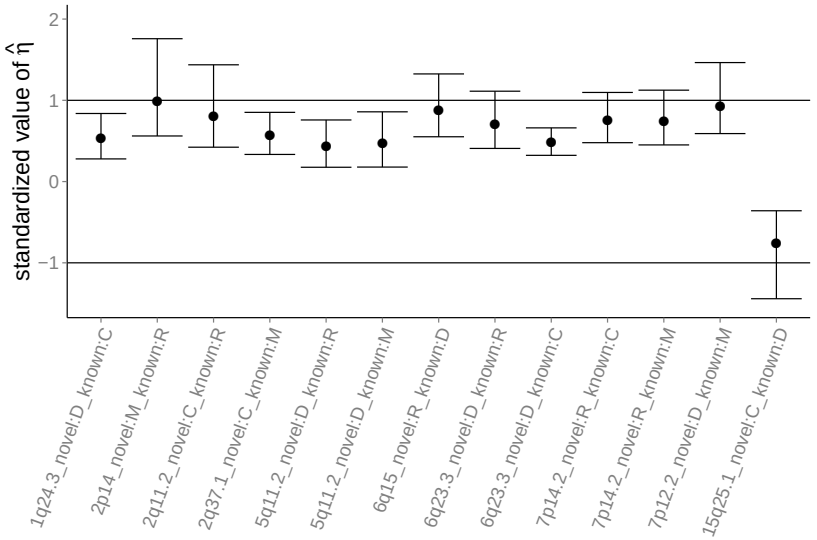
For colocalized disease regions, the two diseases generally have consistent directions of effect (Figure 2.4) with the exception of the 6q25.3 region containing candidate gene *TAGAP*, which is associated in my analysis with CEL and MS only: the risk allele for CEL is protective for MS and vice versa (Figure 2.5). This opposing effect of *TAGAP* alleles has been previously described for T1D and CEL [Smyth et al, 2008], although the region did not provide sufficient evidence for association with T1D in the data available to us. A similar effect for the 2q12.1 region containing candidate gene *IL18RAP* has also been reported [Smyth et al, 2008]. However, later data [Barrett et al, 2009a] have not offered support for T1D association to 2q12.1, and, in my analysis, the posterior support is concentrated on CEL association alone.

Patterns of association with multiple diseases can be complex. In the 2q33 region containing established candidate gene *CTLA4*, as well as the equally strong functional candidate genes, *CD28* and *ICOS*, three potential causal variants appear to be partially shared between T1D, RA and CEL. The strongest association with T1D is at rs3087243 (which has previously been called CT60), while the strongest association with CEL is with rs231775 (which alters the amino acid at position 17 of *CTLA4*, Ala17Thr, and has previously been called CT42). The two SNPs have $r^2 = 0.5$, and haplotype analysis has previously suggested CT60 and not CT42 is causal for Graves' disease [Ueda et al, 2003]. For RA, the strongest single SNP signal is at rs1980422, which is not in LD with either CT42 or CT60 ($r^2 < 0.1$). I fitted the 512 possible standard multinomial models involving these three SNPs for the three diseases, and computed approximate Bayes factors for each. Assuming each model to be equally likely a priori, the model with highest posterior probability has rs1980422/rs3087243 (CT60) signals for CEL and rs231775 (CT42)/rs1980422 for both T1D and RA, although whilst rs231775 (CT42) is the strongest effect for T1D, rs1980422 is strongest for RA (Fig. 2.6). I note that my analysis is based on SNPs selected through a stepwise process and that without fine mapping analysis I cannot claim that any one of these models correctly reflects the causal variants for any disease. These results do, however, clearly illustrate the different patterns of association

with the three disorders and emphasize the potential complexity that can arise in regions of multiple association signals. They motivate the future extension of the colocalization approach developed here to allow model search strategies that do not require stepwise assumptions.



(a)



(b)

Figure 2.4 The distribution of $\hat{\eta}$, the estimated proportionality coefficient together with its 95% confidence interval. In the case of colocalization, η is the ratio of the effects the region exerts upon the two traits. $|\eta| > 1$ corresponds to a stronger effect in Trait 2 than Trait 1. I estimate η by $\hat{\eta}$. Labels on the x-axis give the traits and regions analysed; D for T1D, R for RA, C for CEL and M for MS. Note that in some regions, the conditional analysis supports the existence of multiple associated variants: if none of these overlap, then I consider the region to have separate SNP effects. (a) Regions with strong evidence of colocalization ($\mathbb{P}(\mathbb{H}_4) > 0.9$). As expected, $\hat{\eta}$ is distributed about 1, which corresponds to the regions having equal effects on each trait. Note that 6q25.3, containing the candidate causal gene *TAGAP*, has $\hat{\eta} < 0$, indicating opposite effects on the two diseases. Trait 1 is listed first, and trait 2 second. (b) Regions with novel evidence of disease association, in which I believe there to be colocalization present between the novel association and at least one of the existing associations. Regions have been ordered such that $\hat{\eta}$ estimates the effect size for the novel trait divided by the effect size for the known association. The disease at which the novel association is found is given first in the labels. It can be seen that the effect size tends to be smaller in the new disease.

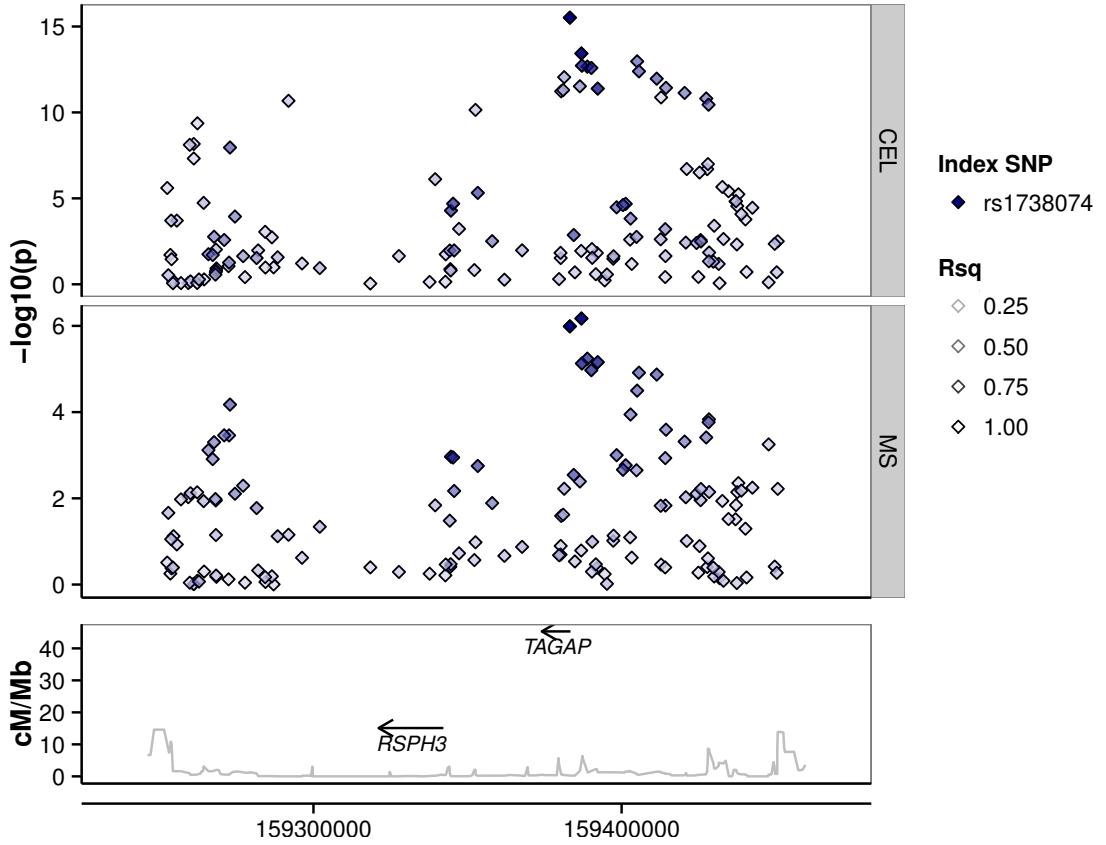
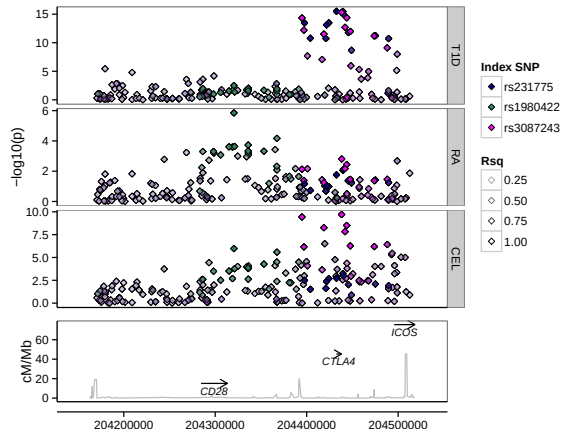


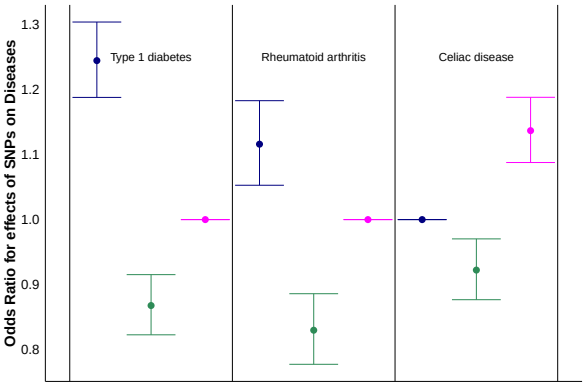
Figure 2.5 The 6q25.3 region containing candidate causal gene *TAGAP*. There is strong evidence of colocalization between CEL and MS ($\mathbb{P}(\mathbb{H}_4) \sim 0.94$). However, the proportional approach reveals that the risk allele for CEL is protective for MS and vice versa.



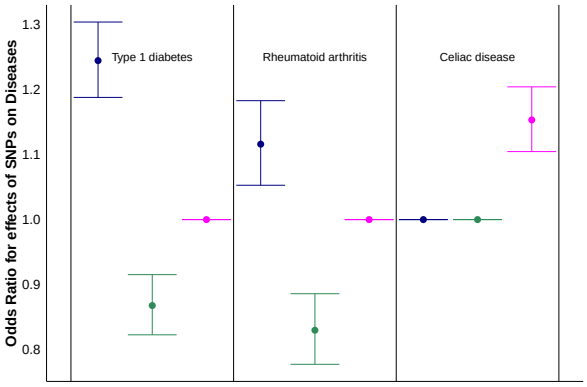
(a)

Posterior Probability	T1D SNPs	RA SNPs	CEL SNPs
0.45	rs231775, rs1980422	rs231775, rs1980422	rs1980422, rs3087243
0.35	rs231775, rs1980422	rs231775, rs1980422	rs3087243
0.07	rs231775, rs1980422	rs1980422	rs1980422, rs3087243
0.06	rs231775, rs1980422	rs1980422	rs3087243

(b)



(c)



(d)

Figure 2.6 (a) A Manhattan plot of the 2q33.1 region containing the candidate gene *CTLA4*. Three potential causal variants are partially shared between T1D, RA and CEL; the blue signal corresponds to the tag rs231775, the green to rs1980422 and the red to rs3087243. All other SNPs are coloured according to their linkage disequilibrium with these three SNPs. SNPs rs231775 and rs3087243 have $r^2 = 0.50$; all other pairwise $r^2 < 1$. (b) Each possible model involving these three SNPs was tested; the four models with highest posterior probabilities, which together encompass over 90% of the total posterior probability, are shown. (c) Effect size estimates (including 95% confidence intervals) of each SNP for each disease for the most likely model. (d) Effect size estimates (including 95% confidence intervals) of each SNP for each disease for the second most likely model.

2.7.3 Discovery of Novel Associations

Two regions were associated with all four diseases (Figure 2.3). One was the 6q23.3 region containing candidate gene *TNFAIP3*, known to be associated with RA and CEL. There has been some published evidence that T1D is associated with this region [Fung et al, 2009], although not at genome–wide significant levels. My results identify a T1D signal, colocalized with that for RA and CEL, suggesting a single shared causal variant affecting the three diseases. There is also evidence of MS association, driven by a distinct causal variant (in the CEL-MS analysis, $\mathbb{P}(\mathbb{H}_3) = 0.83$, Figure 2.7).

The second region was 19p13.2, known to be associated with T1D, RA and MS, containing the strong functional candidate gene *TYK2*, although immune adhesion genes *ICAM1* and *ICAM3* are also good candidate genes. My analysis supports these associations, with a posterior probability of colocalization approaching 1. I also find evidence for a novel CEL association. In each of the pairwise analyses involving CEL, the probability of both diseases being associated $\simeq$ 0.88, although this could be a distinct signal: we have $\mathbb{P}(\mathbb{H}_4|\mathbb{H}_3 \cap \mathbb{H}_4) \simeq 0.5$ (Figure 2.8). In total, eleven regions showed strong evidence of novel association with $\mathbb{P}(\mathbb{H}_3 \cap \mathbb{H}_4) > 0.5$ (Table 2.4).

In regions with colocalizing novel associations, effect sizes tended to be smaller in the new disease (Figure 2.4). This could indicate that the stronger effect is in the previously known association, or it could be due to the Winner's Curse effect [Ioannidis, 2008], with the previously known associations displaying inflated effect size estimates. In general for colocalized signals, the coefficient of proportionality is centred about 1.

One novel association found was in the chromosome 1q24.3 region, known to be associated with CEL and containing candidate gene *FASLG*. Pathway analysis also produced evidence for a T1D-associated variant here [Evangelou et al, 2014], although no SNP has reached the genome-wide significance threshold. My results support a shared causal variant for T1D and CEL (posterior probability 0.71). The Bayesian approach also enables fine-mapping when dense genotyping data are available, as is the case here. I identified a single likely causal variant lying in a region with strong evidence of predicted regulatory activity, rs78037977 (Figure 2.9), with

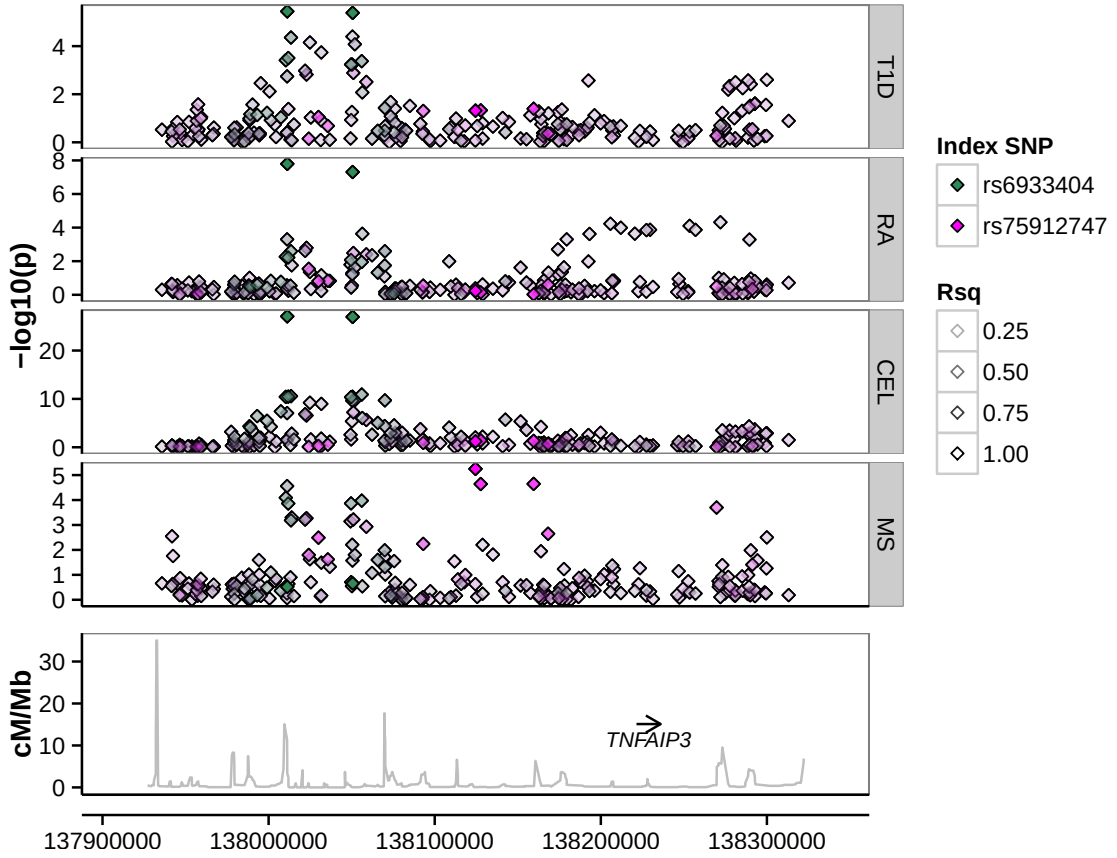


Figure 2.7 The 6q23.3 region containing candidate causal gene *TNFAIP3*. My results show that T1D, RA and CEL all colocalize, suggesting a single shared causal variant affecting the three diseases; rs6933404 being the most likely SNP. There is also evidence of MS association, driven by a distinct causal variant. Note that this region was associated with MS at genome-wide significant levels in the analysis of the international MS dataset [Beecham et al, 2013]. SNPs are highlighted according to their LD with the SNPs considered most likely to be causal by my analysis.

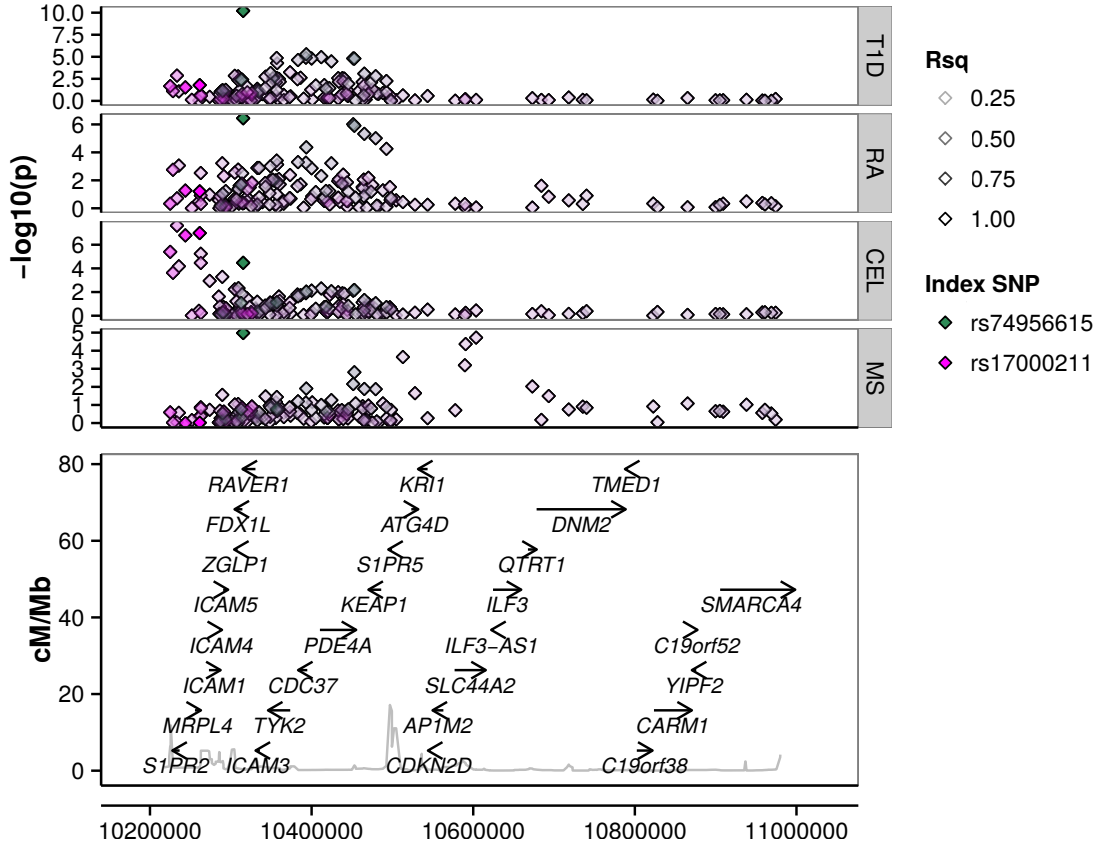


Figure 2.8 A Manhattan plot of the 19p13.2 region containing the candidate causal genes *ICAM1*, *ICAM3* and *TYK2*. SNPs are highlighted according to their LD with the SNPs considered most likely to be causal by my analysis. The green signal is shared across all diseases, whereas the red signal is unique to CEL.

a posterior probability of being causal amongst all genotyped variants, given the colocalization hypothesis, of 0.99. Note that rs78037977 was removed from the CEL data in the original analysis [Trynka et al, 2012] owing to failing a missingness check (the call rate of 99.942% was just below the 99.95% cut-off). Plots of the signal clouds for the samples at this SNP are given in Figure 2.10. The clustering shown here is of good quality, implying that the rs78037977 genotype can be considered reliable.

Figure 2.9 Information from the UCSC Genome Browser for the 1q24.3 *FASLG* region (a region which shows association with T1D and CEL); note that there is strong evidence of regulatory activity in the region of rs78037977, suggesting that this SNP may be significant.

Chromosome	Position	Prior Associations	Associations Found	Post prob both diseases are associated $\mathbb{P}(\mathbb{H}_3 \cup \mathbb{H}_4)$	Post prob shared causal variant given joint association $\mathbb{P}(\mathbb{H}_4 \mid \mathbb{H}_3 \cup \mathbb{H}_4)$	Candidate Causal Genes/ Genes in Region
1q24.3	170882016-171208336	C	**D**C	DC:0.75	DC:0.95	*FASLG*
2p14	65246601-65570598	R	R**M**	RM:0.86	RM:0.72	*SPRED2*
2q11.2	99883120-100415547	DR	DR**C**	DC:0.98 RC:1.00	DC:0.57 RC:0.90	*AFF3*
2q37.1	230758228-230962304	M	**C**M	CM:0.94	CM:0.90	*SP140*
5q11.2	55450712-55492884	RM	**D**RM	DR:0.71 DM:0.71	DR:1.00 DM:1.00	*ANKRD55*
6q23.3	137914792-138345363	RCM	**D**RC|M	DR:0.80 DC:0.77	DR:0.94 DC:0.93	*TNFAIP3*
7p14.2	37323488-37406978	CM	**R**CM	RC:0.80 RM:0.77	RC:0.84 RM:0.83	*ELMO1*
7p12.2	50222360-50335957	M	**D**M	DM:0.73	DM:0.70	5' *IKZF1* region*
13q32.3	98723872-99034738	DM	D|**C**	DC:0.67	DC:0.00	*GPR183*
15q25.1	76773859-77050416	DM	D**C**	DC:0.82	DC:0.99	*CTSH*
19p13.2	10081000-11019034	DRM	DRM|**C**	DC:0.87 RC:0.87 CM:0.88	DC:0.40 RC:0.46 CM:0.57	*ICAM1* *ICAM3* *TYK2*

Table 2.4 Eleven regions showing strong evidence of novel association ($\mathbb{P}(\mathbb{H}_3 \cup \mathbb{H}_4) > 0.5$) for an analysis involving a previously non-associated trait. D corresponds to T1D, R to RA, C to CEL and M to MS. Novel associations are underlined and indicated by the use of bold font. Candidate causal genes are as associated across all curated diseases by ImmunoBase. Note that in the case of *TNFAIP3*, there is strong evidence that MS is caused by a distinct causal variant compared to the other traits. Distinct signals are separated by a "|". Since I only have a subset of the genotype data, not all of the prior (previously published) associations are seen.
*An association of T1D in a region 3' of *IKZF1*, for which it is hypothesised that *IKZF1* is the candidate causal gene is already known [Swafford et al, 2011] (see Table 2.3). The novel association I report here is in a region 5' of *IKZF1*, and independent of the established association. Note I provide coordinates of the region, and not an index SNP as is conventional in GWAS because the method synthesises information across the whole region and does not, in most cases, highlight a single SNP responsible for the association.

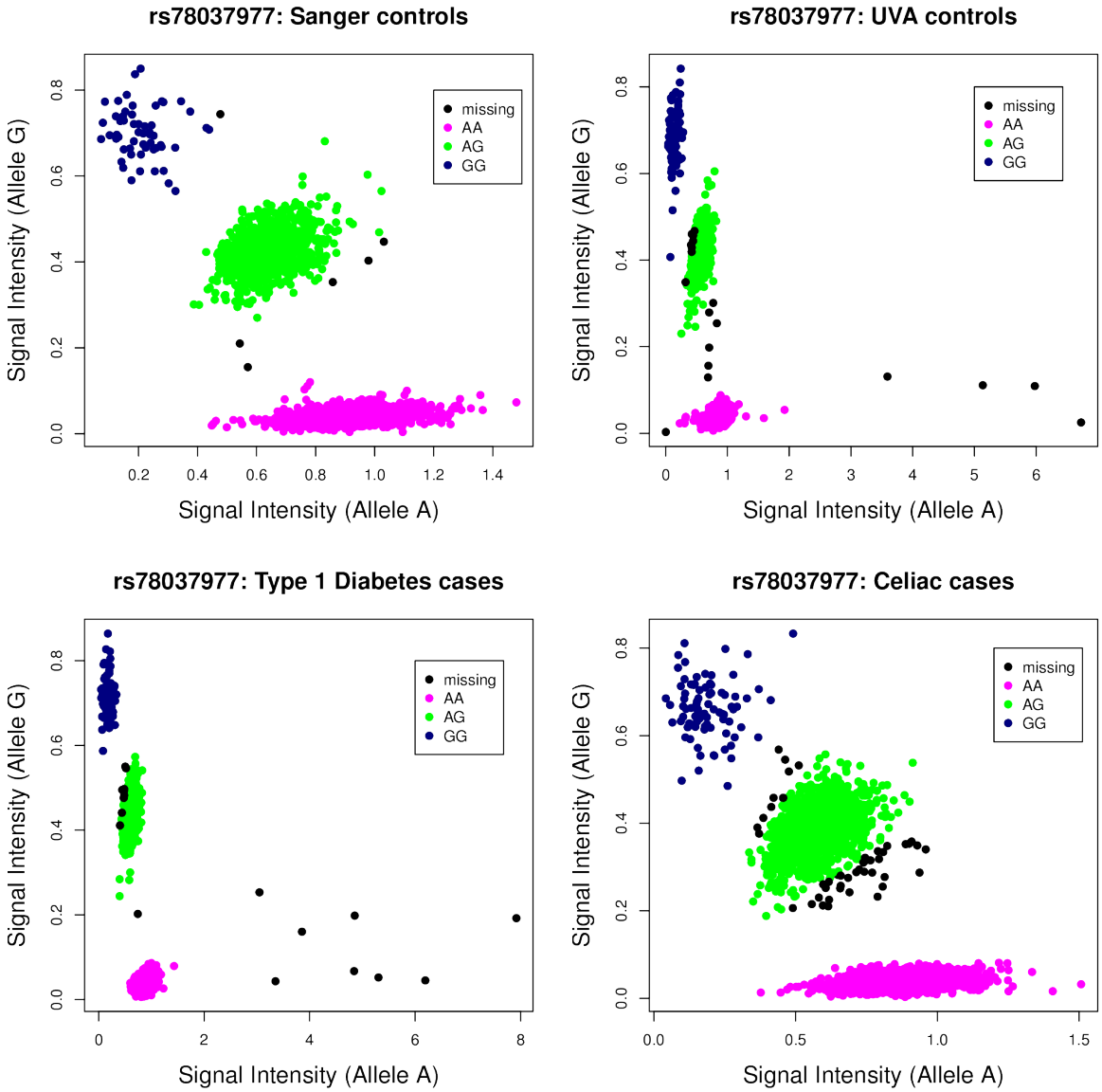


Figure 2.10 Signal clouds for rs78037977, a SNP within the 1q24.3 region containing candidate causal gene *FASLG*. This SNP was removed from the celiac disease data in the original analysis owing to failing a missingness check. However, the clustering shown here is of good quality, implying that the rs78037977 genotype can be considered reliable.

2.7.4 Prior Sensitivity

We tested prior sensitivity by varying p_{12} (the probability that an arbitrary SNP is associated with both diseases) from $p_{12} = 10^{-5}$ to 10^{-7}, while keeping p_1 and p_2 (the probability that this SNP is associated with only trait 1, or only trait 2) constant at 10^{-4} (Table 2.5). Whether a region is disease specific is largely unaffected by choice of p_{12} and, for the five regions discussed in detail in this paper (1q24.3/*FASLG*; 2q33.1/*CTLA4*; 6q23.3/*TNFAIP3*; 6q25.3/*TAGAP* and 19p13.2/*TYK2*), the prior does not change which diseases are associated. However, the posterior odds for $\mathbb{H}_4 : \mathbb{H}_3$ does vary with p_{12}. Under $p_{12} = 10^{-7}$, neither 1q24.3/*FASLG* nor 6q23.3/*TNFAIP3* had strong posterior support as a novel T1D region since the evidence for novel association in these regions comes about due to colocalization with the stronger previously known association. This dependence on prior belief is a strength of Bayesian methods, but they require that priors be carefully calibrated. Whilst my prior belief is that about 50% of regions associated with two immune–mediated diseases are likely to correspond to a shared causal variant, others may disagree. The results given in Supplementary Table 2 of [Fortune et al, 2015] (`http://bit.ly/2d7KvUO`) can be used to calculate the posterior under any alternative p_{12} using the formula given in Section 2.5.1.3.

Details of the results from the Bayesian and proportional analyses, for regions discussed in detail in this Chapter, included in Table 2.3 or in Table 2.4, are given in Appendix B.

choice of p_{12} in prior	number of regions associated with $\mathbb{H}_3$	number of regions associated with $\mathbb{H}_4$	number of disease specific regions
10^{-5}	10	30	20
5×10^{-6}	10	25	20
10^{-6}	14	20	21
5×10^{-7}	17	15	22
10^{-7}	19	8	22

(a)

Candidate Casual Gene	Region	$p_{12} = 10^{-5}$	$p_{12} = 5 \times 10^{-6}$	$p_{12} = 10^{-6}$	$p_{12} = 5 \times 10^{-7}$	$p_{12} = 10^{-7}$
FASLG	1q24.3	T1D CEL	T1D CEL	T1D CEL	T1D CEL	CEL
CTLA4	2q33.1	T1D CEL|RA	T1D CEL|RA	T1D CEL|RA	T1D CEL|RA	T1D|RA|CEL
TNFAIP3	6q23.3	T1D RA CEL|MS	T1D RA CEL|MS	T1D RA CEL|MS	T1D RA CEL|MS	RA CEL|MS
TAGAP	6q25.3	CEL MS	CEL MS	CEL MS	CEL MS	CEL MS
TYK2	19p13.2	T1D RA CEL MS	T1D RA CEL MS	T1D RA MS|CEL	T1D RA MS|CEL	T1D RA MS|CEL

(b)

Table 2.5 The effect of prior choice upon the results of the Bayesian method. Five values of p_{12} (the probability that one SNP is associated with both traits) are tested: $p_{12} = 10^{-5}$; $p_{12} = 5 \times 10^{-6}$; $p_{12} = 10^{-6}$; $p_{12} = 5 \times 10^{-7}$ and $p_{12} = 10^{-7}$. p_1 and p_2 are kept constant at 10^{-4}. (a) The effect of varying p_{12} on the number of regions associated with $\mathbb{H}_3$ ($\mathbb{P}(\mathbb{H}_3) > 0.5$), the number of regions associated with $\mathbb{H}_4$ ($\mathbb{P}(\mathbb{H}_4) > 0.5$) and the number of disease specific regions (posterior probability of single disease association >0.5 in at least one pairwise analysis and posterior probability of association to any other disorder <0.2). As we would expect, varying p_{12} changes whether marginal regions are assigned to $\mathbb{H}_3$ or $\mathbb{H}_4$. However, which regions are disease specific is largely unaffected. (b) Five regions are discussed in detail in this paper: the 1q24.3 region containing candidate causal gene *FASLG*; the 2q33.1 region containing candidate causal gene *CTLA4*; the 6q23.3 region containing candidate causal gene *TNFAIP3*; the 6q25.3 region containing candidate causal gene *TAGAP* and the 19p13.2 region containing candidate causal gene *TYK2*. In the majority of cases, the prior does not change which diseases are associated to one of these regions. However, under $p_{12} = 10^{-7}$, neither *FASLG* or *TNFAIP3* is a novel T1D region.

2.8 Discussion

Colocalization methods so far have allowed for the simultaneous analysis of only two traits: a potential weakness when considering more than two diseases, as investigated here. The Bayesian approach could be extended to arbitrarily many traits, at the cost of increased computational complexity and spreading the posterior over an exponentially increasing hypothesis space, potentially making it difficult to draw firm conclusions. [Flutre et al, 2013], in their description of an alternative method for partitioning the association of a single SNP amongst multiple related quantitative traits, suggest dealing with this complexity by considering only the extremes - a SNP is associated with all traits, exactly one, or none. Such reduction is impractical when analysing regions, since it does not allow for overlapping but distinct signals. Although I have extended my software to consider three diseases simultaneously, I have chosen for practical reasons to focus on pairwise analyses with manual curation of the 11 cases (9%) for which more than two diseases showed association.

[Giambartolomei et al, 2014] showed that inference is consistent when the causal variant is directly genotyped or well imputed. The decision was taken when the ImmunoChip was designed not to thin by LD, but instead target all SNPs and small insertions/deletions known at that time in 1000 Genomes European samples and it has since been shown that common variants may be very accurately imputed using ImmunoChip [Deelen et al, 2014]. Therefore I am likely to be very close to the situation where causal variants are directly genotyped. The application of my method to the less complete coverage provided by genome–wide SNP arrays would require an imputation step to allow consistent inference to be made. The Bayesian colocalization analysis assumes a single causal variant per region, which could be restrictive, and I addressed this using a stepwise approach, attempting to colocalize the individual signals for each disease where there was evidence for more than one. The agreement between my results with this approach and using the proportional colocalization approach which does not make this assumption confirms the appropriateness of the stepwise approach in the cases I consider.

I identified 21 regions that appeared to be associated to only one autoimmune disease.

One challenge in interpretation when defining disease unique signals is exemplified by a region on chromosome 7p12.2 which contains the candidate causal gene *IKZF1*. This gene overlaps two ImmunoChip regions separated by a recombination hotspot, one 5' of *IKZF1* and one 3' of *IKZF1*. The 5' region contains a colocalized signal for MS and T1D, whilst the 3' region contains only a T1D signal (Figure 2.11). My analysis has been based on regions, as defined in the design of the ImmunoChip and based on recombination hot spots. However, whilst the T1D signals in these regions are independent and the 3' region of *IKZF1* appears unique to T1D, it is plausible that the causal variants in both regions act through the same gene, *IKZF1*. Another challenge is to deal with the effects of power, given the established influence of sample size on power to detect associations [Visscher et al, 2012]. Many of the regions in Table 2.2 contain genes linked to immune function, and I expect a number of apparent disease–specific results to associate with other diseases as sample sizes for each disease continue to increase. Indeed, the chromosome 19p13.11 region, associated only with MS in my analysis, has previously been associated with lymphocyte count [Nalls et al, 2011], with high LD between the peak MS SNP (rs1870071) and the lymphocyte count SNP (rs11878602, $r^2 = 0.99$), suggesting an immune mechanism for the association.

However, in the case of T1D, three disease–unique regions overlap known type 2 diabetes (T2D) regions. Chromosome 9p24.2, containing the candidate gene *GLIS3*, has been associated with T2D [Morris et al, 2012] and fasting glucose [Dupuis et al, 2010] with high LD between the peak SNP for T1D (rs10814914) and these other traits (rs7041847, $r^2 > 0.9$). *GLIS3* and its causal allele alter disease risk by altering pancreatic beta–cell function, probably by increasing beta–cell apoptosis [Nogueira et al, 2013]. Chromosome 16q23.1, containing the candidate gene *BCAR1*, is associated with T1D in my analysis and T2D [Morris et al, 2012], and the T2D alleles in this region have been associated with reduced beta cell function [Harder et al, 2013], again with high LD between the peak SNPs for T1D (rs8056814) and T2D (rs7202877, r2 = 0.81). Inspecting the distribution of T2D GWAS p-values at the peak SNPs in my T1D associated regions (Figure 2.12), I note that the peak SNP in the T1D associated region 6q22.32, rs17754780, also shows association to T2D ($p = 7.9 \times 10^{-5}$) and is in tight LD with peak T2D

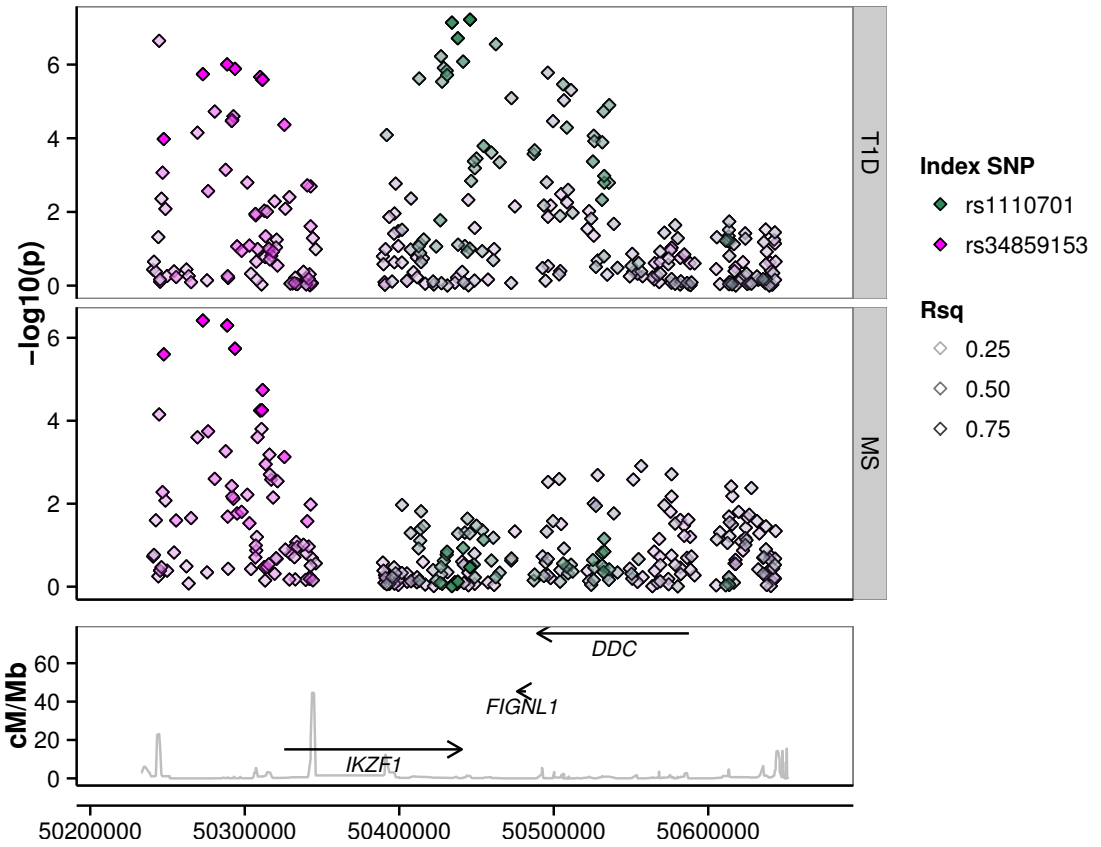


Figure 2.11 The 7p12.2 region containing candidate causal gene *IKZF1*. This gene overlaps two ImmunoChip regions separated by a recombination hotspot, one at the 5' end, one at the 3' end. The 5' region contains a colocalized signal for multiple sclerosis (MS) and type 1 diabetes (T1D), while the 3' end contains only a T1D signal.

SNP in the region (rs9385400, $r^2 = 0.97$). This region has been reported as associated with T2D at genome–wide significance in a larger study [Scott et al, 2014]. Chromosome 6q22.3 is not uniquely associated to T1D in my analysis because it overlaps an established Crohn's disease region [Jostins et al, 2012], but the lead Crohn's SNP (rs9491697) is not in LD with the T1D SNP ($r^2 = 0.03$). This is then likely to be a third shared signal between T1D and T2D. The nearest genes are *MIR588*, about which little appears to be known, and *CENPW* (centromere protein W) which has no obvious functional candidacy. This genetic overlap between T1D and T2D (Table 2.6) emphasizes that T1D results from an interaction between the immune system and beta cells, and it is probable that some of my other apparent disease unique regions will also prove to be specific to the target of autoimmune destruction in MS and RA.

By analyzing regions known to associate with one disease, I was able to link 11 of them to additional disorders: in most cases (8/11) the novel disease association was clearly colocalized with a previously known signal, whilst in one case, *GPR183*, the evidence supported a distinct causal variant for the novel association. In others (3/11) the evidence for colocalization was more equivocal, even with evidence for pairwise association. My results have been incorporated into the online resource ImmunoBase (`www.immunobase.org`).

In a standard GWAS analysis, a p–value significance threshold of 5×10^{-8} is used in the absence of replication data, due to a desire to minimise reporting of false positive results, although a relaxation of this threshold has been suggested [Panagiotou and Ioannidis, 2012]. However, since autoimmune diseases are known to share aetiology, conditioning upon association for one autoimmune disease, I should require a less stringent threshold to believe it significant for another. Indeed, whilst the question of whether the ImmunoChip significance threshold should be somewhat relaxed remains [Parkes et al, 2013], examination of p–values in the regions in which I observe novel associations (Figure 2.13) suggests that a threshold between 10^{-5} and 10^{-6} for SNPs that are confirmed index SNPs for another disease might be more appropriate. This logic was extended to call novel T1D associations conditional on other genome wide significance associations [Onengut-Gumuscu et al, 2015]. I

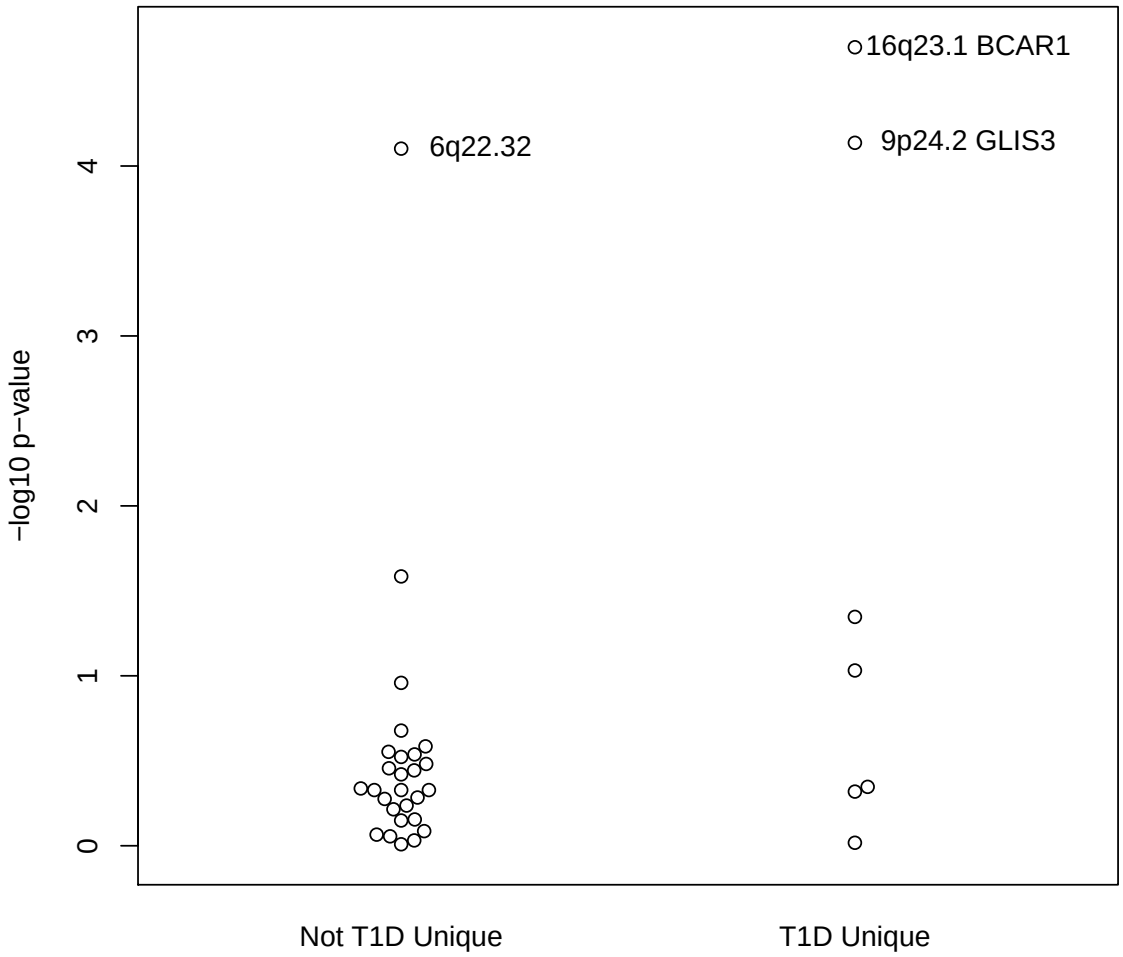


Figure 2.12 P-values for Type 2 Diabetes at the peak SNP for all T1D-associated regions. These regions are divided into those associated with T1D only and those associated with other autoimmune diseases. Those associated with no other autoimmune disease tend to have lower T2D p-values. T2D data was taken from the Stage 1 GWAS and Stage 2 Metabochip: Summary Statistics downloaded from `http://diagram-consortium.org`.

estimate that 42% of overlapping and genome–wide significant immune–mediated disease signals relate to distinct causal variants. In these regions, therefore, there appear to be distinct causal variants for two or more autoimmune diseases which are physically proximal but in low LD. I suggest that physical proximity to a known associated variant in a related disease, and not only LD with it, may prove to be an appropriate criterion with which to alter interpretation of a small but not genome–wide significance threshold. Variants meeting such thresholds might be prioritised for genotyping in replication samples. I note, also, that the four diseases I studied are all characterized by the presence of autoantibodies. Had I included autoantibody negative diseases I might have found a higher proportion of discordant associations as reported in a previous manual curation of ImmunoChip studies [Parkes et al, 2013], given that there remains considerable overlap in location of association signals. Although a careful and detailed manual curation of several studies has been conducted [Parkes et al, 2013], the ability of colocalization methods to distinguish shared from distinct causal variants allows clearer interpretation of genetic results.

In summary, I have developed a methodology for examining shared genetic aetiology between diseases in the case of common control datasets, extending previous work [Plagnol et al, 2009; Wallace et al, 2012; Giambartolomei et al, 2014]. This enables the discovery of new disease associations and the exploration of complex association patterns. Although these methods have been presented in this paper to analyse autoimmune diseases, the prior is user-defined, and could be used to analyse any pair of related diseases.

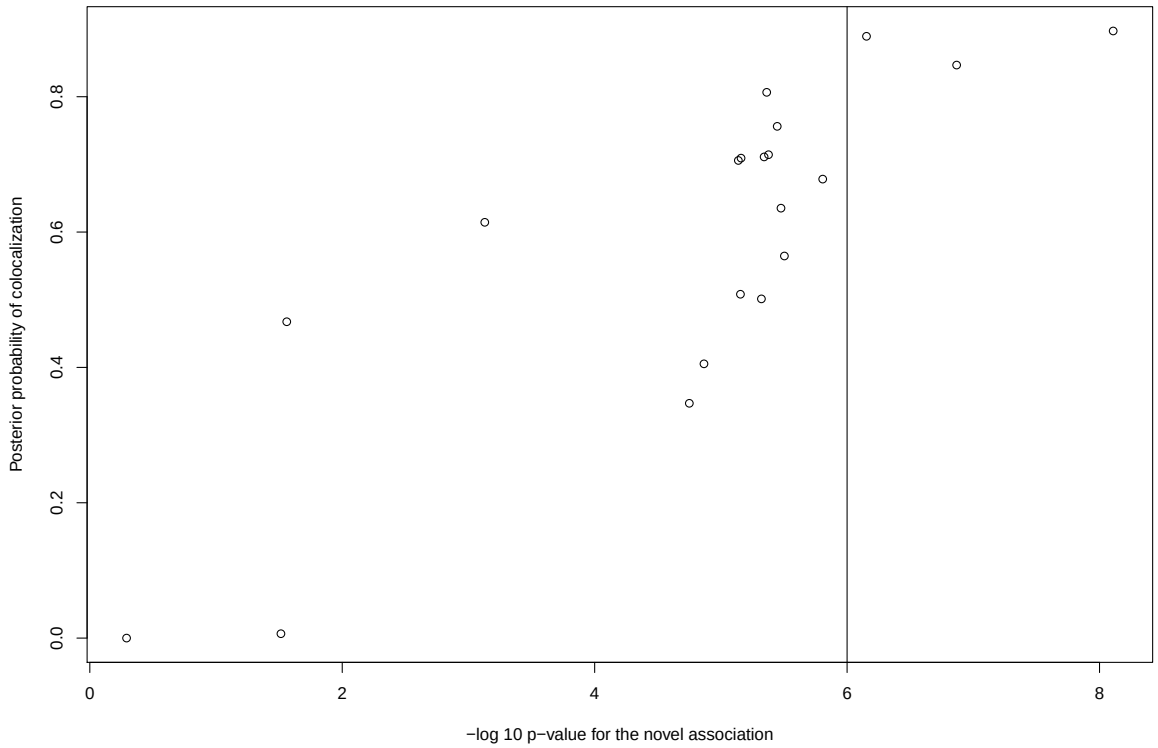


Figure 2.13 P-value and colocalization data from the regions with newly identified associations. The most significant SNP for the known association is found, and its p-value for the newly identified association is computed. This is plotted against the posterior probability of colocalization (as computed using the Bayesian colocalization approach).

Region	Candidate Gene	T1D Peak SNP	Proxy SNP	r^2	Position	MAF	T1D association			T2D association		
							Risk	OR	P	Risk	OR	P
6q22.32	–	rs17754780	rs6919397:T>G	1	126659043	0.44	G	1.108	6.76 (-08)	G	1.06	7.9 (-05)
9p24.2	*GLIS3*	rs10814914	rs10758591:G>C	0.98	4285986	0.49	C	1.092	2.97 (-06)	G	1.05	7.3 (-05)
16q23.1	*BCAR1*	rs8056814	rs9921586:G>T	0.86	75245003	0.08	T	1.266	3.47 (-14)	G	1.12	2.0 (-05)

Table 2.6 Three regions show association with T1D and T2D. I list the peak T1D SNP in my analysis, together with a proxy SNP used to look up evidence for association in the T2D dataset. Risk alleles (Risk), odds ratios (OR) and p-values (P) are presented for the proxy SNP for T1D and T2D. MAF denotes minor allele frequency. r^2 between T1D peak SNP and proxy SNP were calculated in T1D controls. Position is according to NCBI36.

Chapter 3

Simulating GWAS Summary Statistics

3.1 Motivation

In a typical statistical hypothesis test, we have a null hypothesis of no association and quantify the significance of the result with the p-value: the probability of seeing data at least as extreme as that observed under the null hypothesis. Often, this is relatively simple to compute. For instance, in a GWAS, we perform single-SNP based tests, and under the null hypothesis that the SNP in question has no association with the trait, the test generates a Z Score which has distribution $N(0, 1)$ from which a p-value can be computed. However, when the null hypothesis is more complex, the distribution of the test statistic under the null hypothesis may become difficult to compute.

Consider for instance SNP set enrichment analysis. Here, an entire set of GWAS statistics is analysed to see whether the GWAS signals or causal SNPs appear to be enriched in areas of interest; see for instance [Trynka et al, 2015]. In a network analysis, we might be interested in whether the causal SNPs occur disproportionately nearer genes within the network, implicating the network in the process of the disease [Carbonetto and Stephens, 2012]. Alternatively, we might wish to integrate functional annotations with GWAS data, and see whether, for instance,

the causal variants occur within binding sites of the transcription factors which regulate some process of interest. In this sort of analysis, the null hypothesis is not that of no association: instead, it is that there is association, but that association is not enriched within the features of interest. It is not trivial to compute the null distribution for any enrichment test statistic without assumptions of independence between SNPs and feature locations which do not hold.

A common technique in such studies, both to enable the generation of an appropriate null and to ensure that multiple testing is properly corrected for, is permutation testing. Here, the underlying dataset is used, but the labels are permuted to create a new dataset with equivalent trait-association but no correlation to the feature of interest. If full genotype data is known, case/control can be permuted between samples, but this removes any genotype-trait association. If only summary statistics are given, the Z Scores can be permuted between SNPs, or the location of the functional annotations can also be moved. By computing the output statistic from this new dataset and repeating such permutations, one hopes to converge upon the null distribution of outputs.

However, by permuting, we often destroy the genomic structure within the region, leading to an estimate of the null distribution which does not accurately reflect the true behaviour. For instance, SNPs which are in high LD will have similar p-values, but permutation testing does not preserve this behaviour. Such permutation also relies on an underlying assumption that causal variation is evenly distributed throughout the region. However, this is not the case. In a GWAS, we would typically expect to see lower p-values in areas around genes, compared to those p-values observed in intergenic regions. When we are interested in features that also locate near genes, proximity to genes becomes a confounding factor.

One approach, which is similar conceptually to permutation testing, but which controls for genomic structure, is GoShifter [Trynka et al, 2015], which tests for enrichment of functional annotations at causal SNPs. Rather than applying a random permutation, the method shifts annotations by a random distance (with annotations shifted beyond the region re-emerging on the opposite side), thereby preserving the majority of the local structure. It also enables controlling for a second, potentially causal, annotation which colocalizes with the annotation

under analysis, by partitioning the region into those sections with the second annotation and those without, and shifting within these two sections independently. However, this method is computationally complex. It also only preserves structure within individual genetic loci, not ideal for pathway analysis, which is frequently conducted over larger regions.

All these methods rely upon somehow simulating an approximation of the null distribution. This need not be necessary if it is possible instead to use a competitive null. This approach is taken in the VSEAMS package [Burren et al, 2014]. In this method, a control set of genes, not believed to have any association with the trait but selected to match the structure around the test set of genes, are chosen, and p-values near each set are compared. If the control is well chosen, this category of methods works well. However, they rely upon the existence of a large body of prior knowledge about the regions being analysed; it is also necessary to very carefully specify the biological hypothesis being tested.

In this chapter, I present a method which takes as input a list of causal SNPs, $\mathbf{W}$, and a list of odds ratios of effect, $\boldsymbol{\gamma}$, and returns the expected output from a GWAS of the region in question with N_0 control samples and N_1 case samples, assuming the causal model holds. The only additional input required is the LD structure of the region(s) under analysis, which can be inferred from an appropriate reference dataset; hence, it can be used in cases where raw genotype data is not available. By selecting appropriate causal SNPs, multivariate normal sampling given this expected Z Score can then be used to form outputs which could have arisen under a competitive null. By repeatedly computing such expected Z Scores for random causal models rather than perturbing the observed GWAS output in some way, the true null distribution may be estimated while preserving genomic structure. The code used can be found at `https://github.com/mdfortune/simGWAS`.

3.2 Estimation of Expected Z Scores from a GWAS

In this section, I show how expected GWAS Z Scores may be computed given only data on allele frequencies in controls, reference haplotype frequencies, and a model of which SNPs are

causal and their effects on disease odds.

3.2.1 Value of the Z Score

For a GWAS dataset, let $Y_i \in \{0, 1\}$ denote the indicator of disease status at the ith sample. Let there be a total of N samples selected, with N_1 having been chosen from disease cases ($Y_i = 1$) and N_0 having been chosen from disease controls ($Y_i = 0$). Since this sampling is conditional upon case/control status, genotype frequencies may differ between our N samples and the whole population at disease associated SNPs. I therefore need to distinguish between which datasets my genotype probabilities are from; write $\mathbb{P}_{sam}$ for probabilities computed for the samples (i.e. $\mathbb{P}_{sam}(Y_i = 1) = \frac{N_1}{N}$)

Let n be the total number of SNPs. For any SNP X, write G_i^X for its genotype coding $\in \{0, 1, 2\}$ at sample i .

For the commonly used Cochran-Armitage test, the Z-Score at SNP X is computed as:

$$Z_X = \frac{U_X}{\sqrt{V}}$$

Where:

$$U_X = \sum_{i=1}^{N}((G_i^X - \overline{G^X})(Y_i - \bar{Y}))$$

$$V = (N-1)V_X V_Y$$

and V_X, V_Y are the variance of G^X and Y respectively:

$$V_X = \frac{N}{N-1}\frac{\sum_{i=1}^{N}(G_i^X - \overline{G^X})^2}{N}$$

$$V_Y = \frac{N_0 N_1}{N(N-1)}$$

i.e.:

$$V = \frac{N_0 N_1}{N(N-1)}\sum_{i=1}^{N}(G_i^X - \overline{G^X})^2$$

Under the null hypothesis of no association at SNP X, Z_X is distributed as a standard

normal. Hence the two-sided p-value at X is given by:

$$p_X = 2(1 - \Phi(|Z_X|))$$

where Φ is the cumulative distribution function of the standard normal distribution. Conversely, given the unsigned p-value at X, the absolute value of the Z-Score is:

$$-\Phi^{-1}\left(\frac{p}{2}\right)$$

3.2.2 The Causal Model

Write $\mathbf{W} = W_1, ..., W_m$ for the vector of causal SNPs. Since they are causal, the allele frequencies at these SNPs will vary between cases and controls. From publicly available reference datasets such as UK10K [Walter et al, 2015], it is possible to estimate the haplotype frequencies, $\mathbb{P}(G^{\mathbf{W}} = \mathbf{w}|Y = 0)$ using SNPHAP (Clayton D, `http://www-gene.cimr.cam.ac.uk/clayton/software/`), which implements an EM Algorithm. Note that, since sampling is independent of anything but case/control status, I can assume:

$$\mathbb{P}_{sam}(G^{\mathbf{W}} = \mathbf{w}|Y = 0) = \mathbb{P}(G^{\mathbf{W}} = \mathbf{w}|Y = 0)$$

Write $\gamma_1, ..., \gamma_m$ for the log odds ratios of effect for the causal SNPs in the population. I assume that Y given $G^{\mathbf{W}}$ can be modelled as a binomial logistic regression. Then, from [Prentice and Pyke, 1979], the sample-specific odds ratios are the same as those at the population-level, and I can write:

$$\mathbb{P}_{sam}(Y_i = 1|G_i^{\mathbf{W}} = \mathbf{w}) = \frac{e^{\gamma_0+\gamma_1 w_1+...+\gamma_m w_m}}{1 + e^{\gamma_0+\gamma_1 w_1+...+\gamma_m w_m}}$$

where γ_0 is a parameter which corresponds to the overall proportion of cases. Since GWAS sampling is retrospective, this proportion is fixed at $\frac{N_1}{N}$, constraining γ_0, which can be computed

as follows:

$$\begin{aligned}
\mathbb{P}_{sam}(Y_i = 1) &= \frac{N_1}{N} \\
&= \sum_{\mathbf{w}\in\mathbb{Z}_3^m} \mathbb{P}_{sam}(Y_i = 1|G_i^{\mathbf{W}} = \mathbf{w})\mathbb{P}_{sam}(G_i^{\mathbf{W}} = \mathbf{w}) \\
&= \sum_{\mathbf{w}\in\mathbb{Z}_3^m} \mathbb{P}_{sam}(Y_i = 1|G_i^{\mathbf{W}} = \mathbf{w})\frac{\mathbb{P}_{sam}(Y_i = 0|G_i^{\mathbf{W}} = \mathbf{w})}{\mathbb{P}_{sam}(Y_i = 0|G_i^{\mathbf{W}} = \mathbf{w})}\mathbb{P}_{sam}(G_i^{\mathbf{W}} = \mathbf{w}) \\
&= \sum_{\mathbf{w}\in\mathbb{Z}_3^m} \frac{\mathbb{P}_{sam}(Y_i = 1|G_i^{\mathbf{W}} = \mathbf{w})}{\mathbb{P}_{sam}(Y_i = 0|G_i^{\mathbf{W}} = \mathbf{w})}\mathbb{P}_{sam}(Y_i = 0)\mathbb{P}(G_i^{\mathbf{W}} = \mathbf{w}|Y_i = 0) \\
&= \frac{N_0}{N}\sum_{\mathbf{w}\in\mathbb{Z}_3^m} e^{\gamma_0+\gamma_1 w_1+\ldots+\gamma_m w_m}\mathbb{P}(G_i^{\mathbf{W}} = \mathbf{w}|Y_i = 0)
\end{aligned}$$

$$\gamma_0 = \ln\left(\frac{N_1}{N_0\sum_{\mathbf{w}\in\mathbb{Z}_3^m} e^{\gamma_1 w_1+\ldots+\gamma_m w_m}\mathbb{P}(G_i^{\mathbf{W}} = \mathbf{w}|Y_i = 0)}\right)$$

Hence I can compute:

$$\mathbb{P}_{sam}(G_i^{\mathbf{W}} = \mathbf{w}) = \frac{\mathbb{P}(G_i^{\mathbf{W}} = \mathbf{w}|Y_i = 0)\mathbb{P}_{sam}(Y_i = 0)}{\mathbb{P}_{sam}(Y_i = 0|G_i^{\mathbf{W}} = \mathbf{w})}$$

And also:

$$\mathbb{P}_{sam}(G_i^{\mathbf{W}} = \mathbf{w}|Y_i = 1) = \mathbb{P}(G_i^{\mathbf{W}} = \mathbf{w}|Y_i = 1) = \frac{\mathbb{P}_{sam}(Y_i = 1|G_i^{\mathbf{W}} = \mathbf{w})\mathbb{P}_{sam}(G_i^{\mathbf{W}} = \mathbf{w})}{\mathbb{P}_{sam}(Y_i = 1)}$$

I assume that LD structures do not differ between cases and controls, and hence the correlation between $\mathbf{W}$ and X is independent of disease status, or of our sampling. Thus:

$$\mathbb{P}_{sam}(G_i^X = x|G_i^{\mathbf{W}} = \mathbf{w}) = \mathbb{P}(G_i^X = x|G_i^{\mathbf{W}} = \mathbf{w})$$

and I can estimate, for both the whole population, and for my sample:

$$\mathbb{E}((G_i^X)^a|G_i^{\mathbf{W}} = \mathbf{w}) = 2^a\frac{\mathbb{P}(G_i^X = 2 \cap G_i^{\mathbf{W}} = \mathbf{w}|Y_i = 0)}{\mathbb{P}(G_i^{\mathbf{W}} = \mathbf{w}|Y_i = 0)} + \frac{\mathbb{P}(G_i^X = 1 \cap G_i^{\mathbf{W}} = \mathbf{w}|Y_i = 0)}{\mathbb{P}(G_i^{\mathbf{W}} = \mathbf{w}|Y_i = 0)}$$

from my reference dataset, for any constant a. From this, I compute:

$$\begin{aligned}\mathbb{E}((G_i^X)^a|Y_i=1) &= \sum_{\mathbf{w}\in\mathbb{Z}_3^m}\mathbb{E}((G_i^X)^a|G_i^{\mathbf{W}}=\mathbf{w})\mathbb{P}(G_i^{\mathbf{W}}=\mathbf{w}|Y_i=1)\\ &= \sum_{\mathbf{w}\in\mathbb{Z}_3^m}\frac{\mathbb{P}(G_i^{\mathbf{W}}=\mathbf{w}|Y_i=1)}{\mathbb{P}(G_i^{\mathbf{W}}=\mathbf{w}|Y_i=0)}\left[2^a\mathbb{P}(G_i^X=2\cap G_i^{\mathbf{W}}=\mathbf{w})+\mathbb{P}(G_i^X=1\cap G_i^{\mathbf{W}}=\mathbf{w})\right]\end{aligned}$$

$$\begin{aligned}\mathbb{E}((G_i^X)^a|Y_i=0) &= \sum_{\mathbf{w}\in\mathbb{Z}_3^m}\mathbb{E}((G_i^X)^a|G_i^{\mathbf{W}}=\mathbf{w})\mathbb{P}(G_i^{\mathbf{W}}=\mathbf{w}|Y_i=0)\\ &= \sum_{\mathbf{w}\in\mathbb{Z}_3^m}2^a\mathbb{P}(G_i^X=2\cap G_i^{\mathbf{W}}=\mathbf{w})+\mathbb{P}(G_i^X=1\cap G_i^{\mathbf{W}}=\mathbf{w})\end{aligned}$$

By expanding out the numerator in terms of probabilities within the sample dataset, I see that:

$$\begin{aligned}\frac{\mathbb{P}(G_i^{\mathbf{W}}=\mathbf{w}|Y_i=1)}{\mathbb{P}(G_i^{\mathbf{W}}=\mathbf{w}|Y_i=0)} &= \frac{\mathbb{P}_{sam}(Y_i=1|G_i^{\mathbf{W}}=\mathbf{w})\mathbb{P}_{sam}(G_i^{\mathbf{W}}=\mathbf{w})}{\mathbb{P}(G_i^{\mathbf{W}}=\mathbf{w}|Y_i=0)\mathbb{P}_{sam}(Y_i=1)}\\ &= \frac{\mathbb{P}_{sam}(Y_i=1|G_i^{\mathbf{W}}=\mathbf{w})\mathbb{P}_{sam}(G_i^{\mathbf{W}}=\mathbf{w}|Y_i=0)\mathbb{P}_{sam}(Y_i=0)}{\mathbb{P}(G_i^{\mathbf{W}}=\mathbf{w}|Y_i=0)\mathbb{P}_{sam}(Y_i=1)\mathbb{P}_{sam}(Y_i=0|G_i^{\mathbf{W}}=\mathbf{w})}\\ &= \frac{N_0}{N_1}e^{\gamma_0+\gamma_1w_1+\ldots+\gamma_mw_m}\end{aligned}$$

And hence:

$$\begin{aligned}\mathbb{E}_{sam}((G_i^X)^a) &= \frac{N_1}{N}\mathbb{E}((G_i^X)^a|Y_i=1)+\frac{N_0}{N}\mathbb{E}((G_i^X)^a|Y_i=0)\\ &= \frac{N_0}{N}\sum_{\mathbf{w}\in\mathbb{Z}_3^m}e^{\gamma_0+\gamma_1w_1+\ldots+\gamma_mw_m}\left[2^a\mathbb{P}(G_i^X=2\cap G_i^{\mathbf{W}}=\mathbf{w})+\mathbb{P}(G_i^X=1\cap G_i^{\mathbf{W}}=\mathbf{w})\right]\\ &\quad+\frac{N_0}{N}\sum_{\mathbf{w}\in\mathbb{Z}_3^m}\left[2^a\mathbb{P}(G_i^X=2\cap G_i^{\mathbf{W}}=\mathbf{w})+\mathbb{P}(G_i^X=1\cap G_i^{\mathbf{W}}=\mathbf{w})\right]\\ &= \frac{N_0}{N}\sum_{\mathbf{w}\in\mathbb{Z}_3^m}\left(e^{\gamma_0+\gamma_1w_1+\ldots+\gamma_mw_m}+1\right)\\ &\qquad\left[2^a\mathbb{P}(G_i^X=2\cap G_i^{\mathbf{W}}=\mathbf{w})+\mathbb{P}(G_i^X=1\cap G_i^{\mathbf{W}}=\mathbf{w})\right]\end{aligned}$$

3.2.3 Estimation of Z Score for the causal model given by W and γ

Finding the true expectation of $\frac{U_X}{\sqrt{V}}$ is intractable, so instead I compute a first order approximation by assuming independence:

$$\mathbb{E}(Z_X) = \mathbb{E}\left(\frac{U_X}{\sqrt{V}}\right) \approx \mathbb{E}(U_X) \times \mathbb{E}(\frac{1}{\sqrt{V}})$$

These terms can be computed as shown in the following sections.

3.2.3.1 Estimation of U_X, the covariance between G^X and Y, for the causal model given by W and γ

I compute the expectation of U_X in my sample as follows:

$$\begin{aligned}
\mathbb{E}_{sam}(U_X) &= \mathbb{E}_{sam}\left[\sum_{i=1}^{N}(G_i^X - \overline{G^X})(Y_i - \bar{Y})\right] \\
&= \mathbb{E}_{sam}\left[N\left(\sum_{i=1}^{N} G_i^X Y_i\right) - \frac{1}{N}\left(\sum_{i=1}^{N} G_i^X\right)\left(\sum_{i=1}^{N} Y_i\right)\right] \\
&= N\mathbb{E}_{sam}(G_i^X Y_i) - \frac{1}{N}\left[N\mathbb{E}_{sam}(G_i^X Y_i) + N(N-1)\mathbb{E}_{sam}(G_i^X Y_j)\right] \qquad i \neq j \\
&= (N-1)\left[\mathbb{E}_{sam}(G_i^X Y_i) - \mathbb{E}_{sam}(G_i^X Y_j)\right] \\
&= (N-1)\left[\mathbb{E}_{sam}(G_i^X | Y_i = 1)\mathbb{P}_{sam}(Y_i = 1)\right] - \\
&\quad -(N-1)\mathbb{E}_{sam}(Y_j)\left[\mathbb{E}_{sam}(G_i^X | Y_i = 1)\mathbb{P}_{sam}(Y_i = 1) + \mathbb{E}_{sam}(G_i^X | Y_i = 0)\mathbb{P}_{sam}(Y_i = 0)\right] \\
&= \frac{(N-1)N_0N_1}{N^2}\left[\mathbb{E}_{sam}(G_i^X | Y_i = 1) - \mathbb{E}_{sam}(G_i^X | Y_i = 0)\right]
\end{aligned}$$

Using the expressions for $\mathbb{E}_{sam}(G_i^X|Y_i)$ given in Section 3.2.2, this becomes:

$$\begin{aligned}\mathbb{E}_{sam}(U_X) &= \frac{(N-1)N_0N_1}{N^2}\sum_{\mathbf{w}\in\mathbb{Z}_3^m}\Bigg[\left(\frac{N_0}{N_1}e^{\gamma_0+\gamma_1w_1+\ldots+\gamma_mw_m}-1\right)\\ &\qquad\Big[2\mathbb{P}(G_i^X=2\cap G_i^{\mathbf{W}}=\mathbf{w})+\mathbb{P}(G_i^X=1\cap G_i^{\mathbf{W}}=\mathbf{w})\Big]\Bigg]\end{aligned}$$

3.2.3.2 Estimation of V_X, the variance of G^X, for the causal model given by W and γ

Recall:

$$\begin{aligned}V_X &= \frac{1}{(N-1)}\sum_{i=1}^{N}(G_i^X-\overline{G^X})^2\\ &= \frac{1}{(N-1)}\left[\left(\sum_{i=1}^{N}(G_i^X)^2\right)-\frac{1}{N}\left(\sum_{i=1}^{N}G_i^X\right)^2\right]\end{aligned}$$

However, I need to find $\mathbb{E}\left(\frac{1}{\sqrt{V_X}}\right)$.

V_X is the variance of a normal, and so I model it as an Inverse Gamma (α,β) distribution. Then V_X^{-1} has a $\Gamma(\alpha,\beta^{-1})$ distribution, and $\sqrt{V_X^{-1}}$ has a generalised gamma distribution with parameters $p=2, d=2\alpha, a=\sqrt{\beta^{-1}}$. If $V_X\sim$ Inverse Gamma (α,β), then

$$\mathbb{E}(V_X)=\frac{\beta}{\alpha-1}\qquad Var(V_X)=\frac{\beta^2}{(\alpha-1)^2(\alpha-2)}$$

Assuming I can compute computed $\mathbb{E}_{sam}(V_X)$ and $\mathbb{E}_{sam}(V_X^2)$, α and β are completely specified as:

$$\alpha=\frac{2\mathbb{E}(V_X^2)-(\mathbb{E}(V_X))^2}{\mathbb{E}(V_X^2)-(\mathbb{E}(V_X))^2}\qquad \beta=\frac{\mathbb{E}(V_X)\mathbb{E}(V_X^2)}{\mathbb{E}(V_X^2)-(\mathbb{E}(V_X))^2}$$

and $\mathbb{E}\left(\frac{1}{\sqrt{V_X}}\right)$ may be simply computed using:

$$\mathbb{E}\left(\frac{1}{\sqrt{V_X}}\right)=a\frac{\Gamma(\frac{d+1}{p})}{\Gamma(\frac{d}{p})}=\frac{1}{\sqrt{\beta}}\frac{\Gamma(\frac{2\alpha+1}{2})}{\Gamma(\alpha)}$$

Expectation of V_X

$$\begin{aligned}
\mathbb{E}_{sam}(V_X) &= \frac{1}{(N-1)}\left[N\mathbb{E}_{sam}((G_i^X)^2) - \frac{1}{N}\left(N\mathbb{E}_{sam}((G_i^X)^2) + N(N-1)\mathbb{E}_{sam}(G_i^X G_j^X)\right)\right] \\
&= \frac{1}{(N-1)}\left[(N-1)\mathbb{E}_{sam}((G_i^X)^2) - (N-1)\mathbb{E}_{sam}(G_i^X G_j^X)\right] \\
&= \mathbb{E}_{sam}((G_i^X)^2) - (\mathbb{E}_{sam}(G_i^X))^2
\end{aligned}$$

Expectation of V_X^2

$$\mathbb{E}_{sam}(V_X^2) = \left(\frac{1}{(N-1)}\right)^2 \mathbb{E}_{sam}\left[\left(\sum_{i=1}^{N}(G_i^X)^2\right)^2 - \frac{2}{N}\left(\sum_{i=1}^{N}(G_i^X)^2\right)\left(\sum_{i=1}^{N}G_i^X\right)^2 + \frac{1}{N^2}\left(\sum_{i=1}^{N}G_i^X\right)^4\right]$$

Let $E_n = \mathbb{E}_{sam}((G_i^X)^n)$. Breaking this down into terms, for (i, j, k, l) representing different indices, I have:

$$\begin{aligned}
&\mathbb{E}_{sam}\left[\left(\sum_{i=1}^{N}(G_i^X)^2\right)^2\right] \\
&= N\mathbb{E}_{sam}((G_i^X)^4) + N(N-1)\mathbb{E}_{sam}((G_i^X)^2(G_j^X)^2) \\
&= NE_4 + N(N-1)
\end{aligned}$$

$$\begin{aligned}
&\mathbb{E}_{sam}\left[\left(\sum_{i=1}^{N}(G_i^X)^2\right)\left(\sum_{i=1}^{N}G_i^X\right)^2\right] \\
&= N\mathbb{E}_{sam}((G_i^X)^4) + 2N(N-1)\mathbb{E}_{sam}((G_i^X)^3(G_j^X)) + N(N-1)\mathbb{E}_{sam}((G_i^X)^2(G_j^X)^2) + \\
&+ N(N-1)(N-2)\mathbb{E}_{sam}((G_i^X)^2(G_j^X)(G_k^X)) \\
&= NE_4 + 2N(N-2)E_3E_1 + N(N-1)E_2^2 + N(N-1)(N-2)E_2E_1^2
\end{aligned}$$

$$
\begin{aligned}
& \mathbb{E}_{sam}\left[\left(\sum_{i=1}^{N} G_i^X\right)^4\right] \\
= \;& N\mathbb{E}_{sam}((G_i^X)^4) + 4N(N-1)\mathbb{E}_{sam}((G_i^X)^3(G_j^X)) + 6N(N-1)\mathbb{E}_{sam}((G_i^X)^2(G_j^X)^2) + \\
+ \;& 6N(N-1)(N-2)\mathbb{E}_{sam}((G_i^X)^2(G_j^X)(G_k^X)) + \\
+ \;& N(N-1)(N-2)(N-3)\mathbb{E}_{sam}((G_i^X)(G_j^X)(G_k^X)(G_l^X)) \\
= \;& NE_4 + 4N(N-1)E_3E_1 + 6N(N-1)E_2^2 + 6N(N-1)(N-2)E_2E_1^2 + \\
+ \;& N(N-1)(N-2)(N-3)E_1^4
\end{aligned}
$$

Giving:

$$
\mathbb{E}_{sam}(V_X^2) = \frac{1}{N}E_4 - \frac{4}{N}E_3E_1 + 2\frac{N^2-2N+6}{N(N-1)}E_2^2 - 2\frac{(N-2)(N-3)}{N(N-1)}E_2E_1^2 + \frac{(N-2)(N-3)}{N(N-1)}E_1^4
$$

3.2.4 Summary

Thus, given only a choice of which SNPs are causal ($\mathbf{W}$), their effect sizes ($\boldsymbol{\gamma}$), sample sizes (N_0, N_1) and a reference dataset from which I can derive allele frequencies ($\mathbb{E}(G_i^X|Y_i=0)$) and the relationships between SNPs ($\mathbb{E}(G_i^X|G_i^{\mathbf{W}}=\mathbf{w})$), I can derive an expected Z Score, $\mathbf{Z}^{EXP}$ at any SNP, causal or not. This can then either be used directly, or in order to compute the simulated output from such a GWAS, $\mathbf{Z}^{SIM}$, which will be distributed:

$$
\mathbf{Z}^{SIM} \sim \mathbf{N}(\mathbf{Z}^{EXP}, \mathbf{LD})
$$

where $\mathbf{LD}$ is the LD matrix of SNPs within the region.

3.3 Choice of Parameters for Simulation

In order to compute a null distribution, we may wish to simulate the Z Scores from many causal models $\{\mathbf{W}, \boldsymbol{\gamma}\}$, requiring a choice of sampling distribution over m, $\mathbf{W}$ and $\boldsymbol{\gamma}$. It may be that the context of the analysis constrains the choice of causal models. However, in the absence of such context, I suggest the following null distributions:

3.3.1 Choice of m: number of causal SNPs

m is the number of causal SNPs in an LD-defined region in our model. It is difficult to obtain an appropriate prior for m. Although a great many GWAS have been done (see, for instance, the GWAS catalogue, `www.ebi.ac.uk/gwas/`), these typically report only the top hit per region. When additional associations are sought, a typical approach is to work conditionally, at each iteration finding the strongest signal which explains the remaining trait variance; this is the strategy I took in my colocalization analysis (see Section 2.5.1.6). By doing this, we potentially miss a set of multiple causal variants, none of which is individually the top performing SNP [Wallace et al, 2015]. In addition, conditional approaches tend to be conservative about the number of causal variants, with strong, often genome-wide significant, evidence of an additional signal required, although this is not universal and more relaxed thresholds are used too [Trynka et al, 2012]. This means that our estimates of m are likely under-estimates.

It is also likely that the number of causal SNPs within a genetic locus varies with the disease being studied. However, the biological focus of my analysis is upon autoimmune diseases. Hence I present here my suggestion for the prior of m when testing autoimmune association.

GUESSFM [Wallace et al, 2015] is a fine mapping algorithm which uses a Bayesian Stochastic Search to find the most likely set of causal variants in a region. As part of its output, it gives its posterior belief about the number of causal SNPs in that region. I obtained my prior on m by taking the GUESSFM results from 918 analyses (of 95 genetic regions and for 10 autoimmune diseases) and averaging over these posterior probabilities

m	Probability
1	0.892
2	0.0856
3	0.0105
4	0.00754
5	3.59×10^{-3}
6	3.49×10^{-4}
7	1.20×10^{-5}
8	4.08×10^{-8}
9	2.98×10^{-11}

Table 3.1 Prior on m, the number of causal SNPs in a model.

(Chris Wallace, unpublished data). The numeric values of my prior are given in Table 3.1. Although these results are highly dependent on the priors input into GUESSFM, the stochastic search algorithm performs better than a conventional stepwise approach, and more accurately incorporates knowledge about m in regions where the causal model does not include the top hit.

3.3.2 Choice of W: causal SNPs

$W_1, ..., W_m$ are the causal SNPs. In the absence of any information about the region, the most appropriate prior to use is a flat prior. Since causal variants are not distributed evenly throughout the genome, if additional functional data about a region are available, it may be appropriate to incorporate this to create a more informative prior.

Typically in GWAS, causal variants are found in non-coding but functionally active regions. We also expect to see disease-specific effects; in a study of causal autoimmune disease variants, $\sim 60\%$ were found to map to immune-cell enhancers [Farh et al, 2015]. There is evidence that causal variants are disproportionally found in DNase 1 Hypersensitivity sites, regions where the structure of chromatin exposes DNA to cleavage by the DNase enzyme [Maurano et al, 2012]. A variable prior for **W** could be found by application of a fine mapping tool which incorporates this information, such as fgwas [Pickrell, 2014].

3.3.3 Choice of γ: SNP effect sizes

The values of γ_i give the log odds ratio of the effect of each causal SNP i upon disease risk. These can be user specified, and if a null is being generated, previous knowledge about the specific circumstance may suggest values of γ_i. However, for those who do not have such constraints, I suggest the following prior:

In order to estimate an appropriate prior distribution, I generated an empirical distribution for γ_i by sampling ~ 6000 studies of a similar design from the genome catalogue (*http://www.ebi.ac.uk/gwas/*) and took the odds ratios for the potential causal SNPs found. I then took the absolute values of the log odds, since sign corresponds to whether the allele under study is the protective or deleterious allele, and this should be independent of the absolute value of the effect. A plot of their distribution is given in Figure 3.1. This distribution appears to be shaped like a gamma distribution, with the best fit being $\Gamma(0.81, 1.15)$.

Hence, I will use a $\Gamma(0.8, 1.2)$ distribution as the prior for the absolute value of γ_i, with the sign being positive or negative with equal probability (that is, I assume each direction of effect is equally likely). Note that, this distribution is slightly flatter than the empirical distribution of γ_i. However, this may be beneficial in cases where I am simulating multiple values of $\boldsymbol{\gamma}$ for the same causal variants (see Chapter 4), as it enables me to search the space of odds ratios more quickly.

As discussed in Section 3.2.2, once $\gamma_1, ..., \gamma_m$ have been chosen, γ_0 is constrained to be:

$$\gamma_0 = \ln\left(\frac{N_1}{N_0 \sum_{\mathbf{w}\in\mathbb{Z}_3^m} e^{\gamma_1 w_1 + ... + \gamma_m w_m} \mathbb{P}(G_i^{\mathbf{W}} = \mathbf{w} | Y_i = 0)}\right)$$

Note, however, that selecting $\boldsymbol{\gamma}$ from this distribution does not guarantee that the resulting simulated output will be GWAS significant, even if the same odds ratios gave a significant result on a different study. Z_X is a function of both U_X and V_X; when the MAF of X is low, V_X is large and hence the resulting Z Score is lower than it would be for a more common SNP.

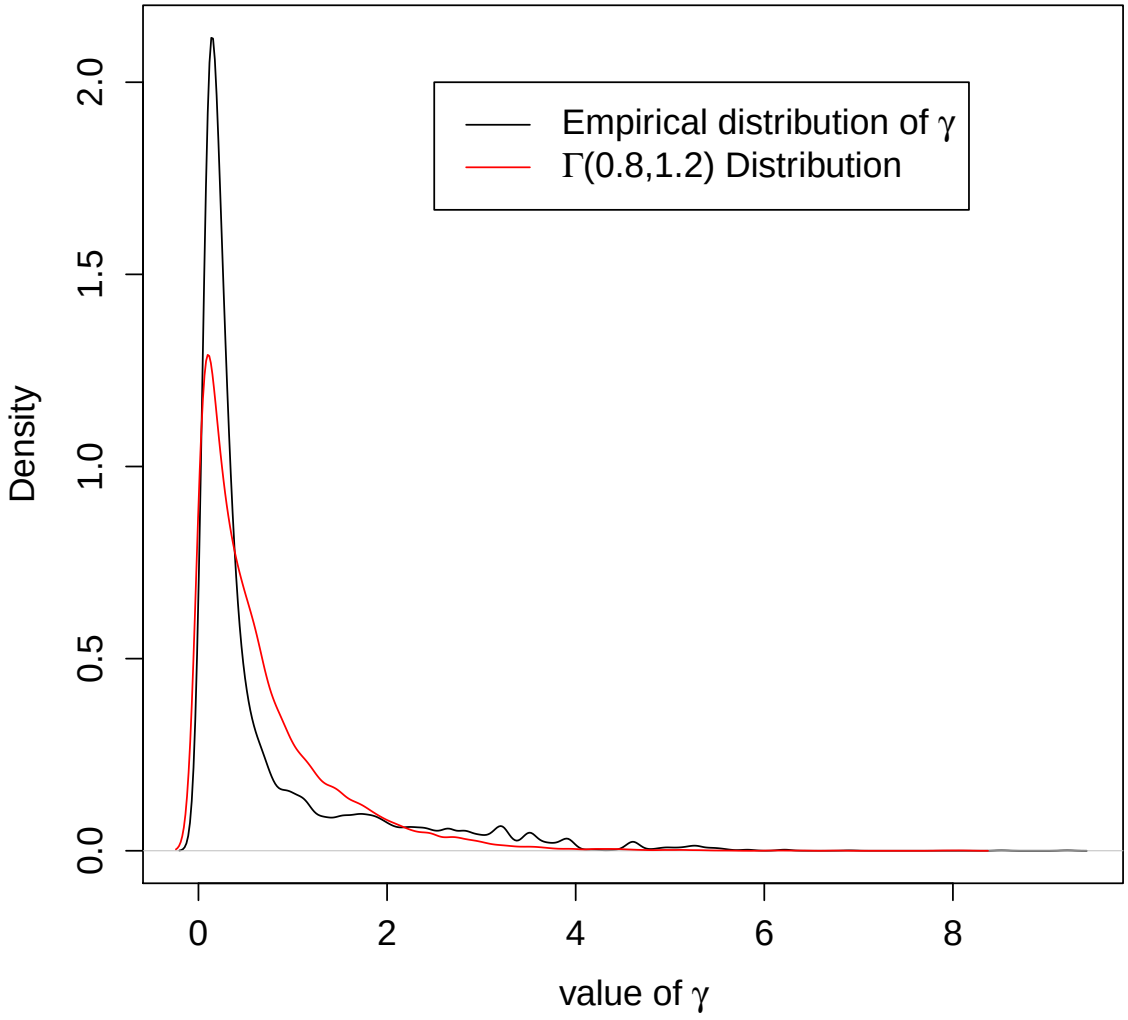


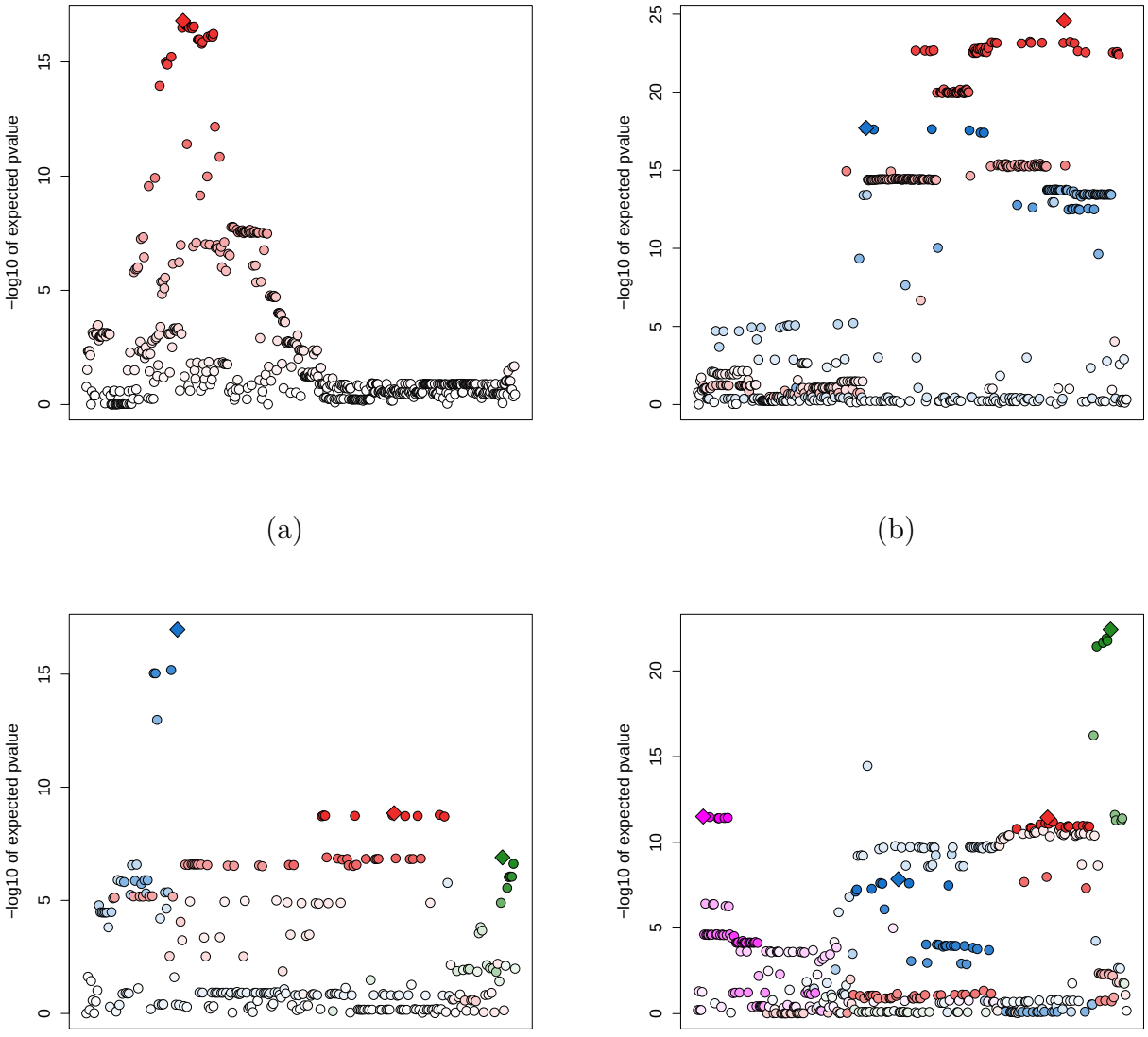
Figure 3.1 In black is given an estimated density of the empirical distribution of the absolute value of the log odds ratio of effect for potential causal SNPs in GWAS, as taken from the genome catalogue. In red is given the distribution of a $\Gamma(0.8, 1.2)$ random variable, which is an acceptable approximation.

3.4 Validation of Method

3.4.1 Output from Method

Manhattan plots giving the p-values obtained from simulated GWAS (for one, two, three and four causal variants) are shown in Figure 3.2. All SNPs are coloured according to their correlation with the causal SNPs (which are shown by diamonds rather than circles). From this, I see that the output appears to take the correct form, with causal SNPs having high association and the p-value of other SNPs corresponding to their LD with the causal SNPs.

However, in order to validate my method, I must compare the outputs it gives to the true Z Scores computed directly from a dataset where the causal model is known.



(a) (b)

(c) (d)

Figure 3.2 Example outputs from simulated GWAS, for one, two, three and four causal SNP models. Causal SNPs are designated by a coloured diamond. Non-causal SNPs are designated by a circle, coloured according to their LD with their most correlated causal SNP.
(a) A single causal variant, with $\gamma_0 = -0.19, \boldsymbol{\gamma} = 0.25$
(b) Two causal variants, with $\gamma_0 = -0.84, \boldsymbol{\gamma} = (0.4, 0.2)$
(c) Three causal variants, with $\gamma_0 = -0.37, \boldsymbol{\gamma} = (0.2, 0.2, 0.2)$
(d) Four causal variants, with $\gamma_0 = -2.30, \boldsymbol{\gamma} = (0.4, 0.4, 0.4, 0.4)$

3.4.2 Construction of Datasets for Testing

I wish to generate a dataset of case/control genotypes with known causal SNPs and odds ratios of effect. Control data can be sampled from a reference dataset used to estimate genotype frequency, since by definition it has the desired $\mathbb{P}(G^{\mathbf{W}} = \mathbf{w}|Y = 0)$. Sampling case data, however, is harder; reference datasets for the disease under study may not be available, and, in any case, unless the model $\{\mathbf{W}, \boldsymbol{\gamma}\}$ is correct, true case data will not have the desired $\mathbb{P}_{sam}(G^{\mathbf{W}} = \mathbf{w}|Y = 1)$. Hence, I also sample case data from the reference dataset, with subject i, where $G_i^{\mathbf{W}} = \mathbf{w}$, being sampled with weight $\zeta_{\mathbf{w}}$. Since under this sampling scheme:

$$\begin{aligned}\mathbb{P}_{\text{case data}}(G^{\mathbf{W}} = \mathbf{w}) &= \zeta_{\mathbf{w}}\mathbb{P}(G^{\mathbf{W}} = \mathbf{w} \text{ in reference dataset}) \\ &= \zeta_{\mathbf{w}}\mathbb{P}(G^{\mathbf{W}} = \mathbf{w}|Y = 0)\end{aligned}$$

by sampling with (non-normalised) weights $\zeta_{\mathbf{w}} = \frac{\mathbb{P}(G^{\mathbf{W}}=\mathbf{w}|Y=1)}{\mathbb{P}(G^{\mathbf{W}}=\mathbf{w}|Y=0)}$ I obtain a dataset with the desired features.

3.4.3 Comparision between Observed and Expected Z Scores

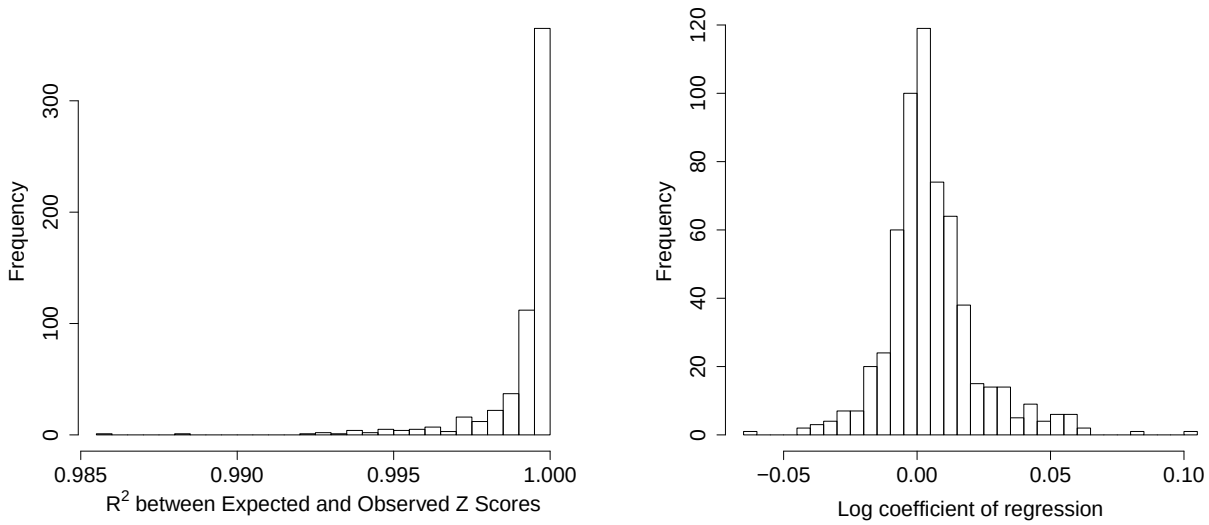
Using the control datasets from my colocalization analysis (see Section 2.6 for details of samples and QC) as reference datasets, I computed expected Z Scores for 600 GWAS with $N_0 = N_1 = 5000$, using the methods outlined in Section 3.2; causal models $\{m, \mathbf{W}, \boldsymbol{\gamma}\}$ were generated using the priors discussed in Section 3.3 and selected so that at least one GWAS-significant p-value was expected.

For each of these causal models, I generated 100 case/control testing datasets using the strategy detailed above. I then estimated an empirical expected observed Z Score by averaging over the Z Scores computed directly from the datasets; by computing the mean, I reduce the effect of randomness in dataset generation and hence have a better basis of comparison with my theoretical expected Z Scores.

Ideally, my observed and expected Z Scores would be identical. In order to compare them, I regressed the expected Z Score onto the observed Z Score for each of the 600 GWAS scenarios,

with the intercept set to one. Figure 3.3 shows the results of this regression. On the left is the distribution of R^2, a measure of how collinear the two vectors are. The values from my analyses are very close to one, indicating that expected and observed Z Scores are very proportional. The coefficient of the regression indicates whether there is a difference of a constant factor between the vectors; I have plotted the log transform of this for ease of interpretation. This log coefficient is distributed tightly around zero, indicating that the expected and observed Z Scores have almost identical values.

Taken together, these outputs demonstrate that my method produces an expected Z Score which very closely matches the empirical expected Z Score from full data simulations. Outlying points on Figure 3.3 typically correspond to datasets where even the top hit barely reaches genome-wide significance. In these cases, the variance in the distribution of Z Scores around zero, from those SNPs which are not associated with the disease, causes the linear fit to be less good.



(a) (b)

Figure 3.3 Results when the expected Z Score, as generated by my algorithm, is regressed upon the observed Z Score, with no intercept included.
(a) The R^2, or coefficient of determination. Values close to 1 indicate that almost all the variance in the observed Z Score can be predicted by the expected Z Score.
(b) The log of the regression coefficient. This is close to 0, corresponding to the observed and expected Z Scores having almost identical values.

3.5 Discussion

I present in this chapter an algorithm for computing the expected output of a GWAS analysis under a specified causal model. The only input required is a reference dataset of control samples; genotyping of case samples is not needed. My results very closely match the summary statistics generated from direct analysis of fully genotyped case/control datasets.

By generating many such GWAS and applying whichever enrichment analysis is of interest, I can now estimate the null distribution for any test of whether the set of causal SNPs appears to be significantly enriched within a particular annotation of interest. Unlike existing methods such as GoShifter [Trynka et al, 2015], this fully preserves the genomic structure, even when analysis is scaled over multiple regions.

In principle, this algorithm could be used to simulate output for any disease and in any

region. However, it assumes that we can accurately estimate $\mathbb{E}(G_i^X|Y_i = 0)$ from the reference dataset used; this is, it assumes that the reference data, which may come from a cohort collected for other purposes, and for which phenotype information is not available, contains no cases. For rare diseases, such as autoimmune diseases, this may be an accurate assumption. While there may be occasional disease cases within the reference dataset, the effect should be small enough to not significantly change our results. However, this assumption is not justifiable for a more common trait, such as obesity, and this algorithm is not appropriate for the case of a common disease unless care is taken to ensure that the reference dataset used does accurately reflect the genotypes of control samples.

With the benefit of more time, I would apply this method to integrate GWAS outputs with epigenetic information. By using ChIP-seq data, I could determine whether disease-associated variants occur within protein-binding sites, thereby implicating the protein in the disease process. Integrating promoter capture Hi-C analysis would enable me to discover whether causal variants are significantly enriched within promoter interacting regions for genes of interest; this may also enable me to find tissue-specific effects.

However, this algorithm also has the potential to be used in fine mapping for datasets in which only GWAS summary data are available. I can evaluate how well a causal model fits the data given by computing the expected Z Scores, assuming that the model is true, and comparing them to the Z Scores observed in the real data. Given the limitations in the availability of full genotyping data I have encountered, this seemed a more pressing need. This extension is the subject of my next chapter.

Chapter 4

Fine Mapping using GWAS Summary Statistics

4.1 Motivation

Although GWAS have enabled the identification of many regions of association with complex traits such as autoimmune disease, the top signals reported do not necessarily correspond to the underlying causal variants (CVs); instead a low p-value may be caused by the apparent lead SNP being in high LD with the true CV, and therefore "tagging", it. Similarly, a gene located in, or close to, a region showing disease association need not have any role in disease aetiology; the CVs may instead act upon a different gene, some distance away from themselves. In order to understand disease aetiology then, once a GWAS has reported association to a region, fine mapping algorithms must be used in order to infer which variants are causal. These can then be integrated with functional annotation data (for instance by performing a SNP set enrichment analysis such as those discussed in Chapter 3 upon the set of causal SNPs) to find disease-associated genes.

An obstacle often encountered in fine mapping is a lack of genotype data for the original samples analysed in GWAS. GWAS datasets are very large, possibly consisting of a meta-analysis of datasets from different sources; together with privacy concerns, this makes accessing

raw data from studies difficult. Results from GWAS are typically reported as the summary statistics from single-SNP regressions upon the trait of interest. For eQTL studies, even these are often trimmed, only reporting nominally significant results. It is rarely reported whether missing SNPs have p-values above the significance threshold, or whether they failed quality control. While odds ratios are reported, or inferable, without specification of reference and alternative alleles, and strand information in the case of A/T and C/G SNPs, it is not possible to know the direction of effects of SNPs, and this information is not always given. There is therefore need for fine mapping methods based solely upon p-values, with the use of a publicly available reference dataset when required to estimate correlation between SNPs.

There are many approaches to fine mapping. One can directly use the p-values, and consider all SNPs which reach a certain threshold, or which are in high LD with the top GWAS signals, as potentially causal. Alternatively, a Bayesian approach can be taken, and posterior probabilities for each SNP being the CV computed; these can then be combined to obtain a credible set of SNPs which contain the true CV with some desired probability. However, these methods assume that there is a single causal variant: a common simplifying assumption, but not one which is biologically plausible in many regions. An iterative approach can be used, repeatedly running such techniques until no additional CVs are found. However, this risks making incorrect inferences in the case where no SNP in the true (multi-SNP) CV set is individually the top performing SNP. These methods also require that all SNPs to be considered are analysed (or the results imputed) in the GWAS.

In Section 1.5 I summarised two existing techniques, PAINTOR [Kichaev et al, 2014] and CAVIAR [Hormozdiari et al, 2014] which perform fine mapping, using only marginal test statistics from GWAS and the correlation between SNPs. They both allow for multiple-CV models; PAINTOR also integrates functional genomic annotation data. However these methods are likelihood based, requiring directions of effects for each SNP.

In this Chapter, I present a fine mapping technique I have developed which allows for multiple causal variants while requiring only summary GWAS p-values and publicly available reference datasets. It does not depend upon availability of effect size estimates or the allelic

direction of effect, and it can infer whether the pattern of association is likely to be caused by a non-genotyped SNP without requiring imputation of summary statistics at this position. Using the method outlined in Chapter 3, I am able to compute the expected Z Score from a GWAS assuming the true causal model to be parameterized by {$\mathbf{W}$, the causal SNPs, and $\boldsymbol{\gamma}$, their odds ratios of effect}. I can then compare this to the absolute Z Score I have from my GWAS derived from the observed p-value; if the model is true, I would expect these two to be similar. By considering many such models $\{\mathbf{W}, \boldsymbol{\gamma}\}$, and finding those which best fit with the observed data, I am able to propose a set which likely contains the true causal model.

4.2 Development of my Approach

4.2.1 Likelihood Based Approaches

Using the same framework as in Section 3.2.1, the Z-Score at SNP X is given by:

$$Z_X = \frac{\sum_{i=1}^{N}((G_i^X - \overline{G^X})(Y_i - \bar{Y}))}{\sqrt{\frac{N_0 N_1}{N(N-1)}\sum_{i=1}^{N}(G_i^X - \overline{G^X})^2}}$$

Let $\mathbf{Z}^{OBS}$ denote the Z Score derived from a GWAS p-value. Writing $\mathbf{W} = W_1, ..., W_m$ for the causal SNPs, and $\boldsymbol{\gamma} = (\gamma_0, \gamma_1, ..., \gamma_m)$ for their odds ratios of effect, as before, I can decompose:

$$\mathbb{P}(\mathbf{G^X}_i = x | Y_i = 0) \quad = \quad \sum_{\mathbf{w} \in \mathbb{Z}_3^m} \mathbb{P}(\mathbf{G^X}_i = x | \mathbf{G^W}_i = \mathbf{w}) \mathbb{P}(\mathbf{G^W}_i = \mathbf{w} | Y_i = 0)$$

$$\begin{aligned}
\mathbb{P}(\mathbf{G^X}_i = x|Y_i = 1) &= \sum_{\mathbf{w}\in\mathbb{Z}_3^m} \mathbb{P}(\mathbf{G^X}_i = x|\mathbf{G^W}_i = \mathbf{w})\mathbb{P}(\mathbf{G^W}_i = \mathbf{w}|Y_i = 1) \\
&= \sum_{\mathbf{w}\in\mathbb{Z}_3^m} \mathbb{P}(\mathbf{G^X}_i = x|\mathbf{G^W}_i = \mathbf{w})\frac{\mathbb{P}_{sam}(Y_i = 1|\mathbf{G^W}_i = \mathbf{w})\mathbb{P}_{sam}(\mathbf{G^W}_i = \mathbf{w})}{\mathbb{P}_{sam}(Y_i = 1)} \\
&= \sum_{\mathbf{w}\in\mathbb{Z}_3^m} \mathbb{P}(\mathbf{G^X}_i = x|\mathbf{G^W}_i = \mathbf{w})\mathbb{P}(\mathbf{G^W}_i = \mathbf{w}|Y_i = 0) \\
&\qquad \left(\frac{\mathbb{P}_{sam}(Y_i = 1|\mathbf{G^W}_i = \mathbf{w})}{\mathbb{P}_{sam}(Y_i = 0|\mathbf{G^W}_i = \mathbf{w})}\frac{\mathbb{P}_{sam}(Y_i = 0)}{\mathbb{P}_{sam}(Y_i = 1)}\right) \\
&= \sum_{\mathbf{w}\in\mathbb{Z}_3^m} \frac{N_0}{N_1} e^{\gamma_0+\gamma_1 w_1+\ldots+\gamma_m w_m}\mathbb{P}(\mathbf{G^X}_i = x|\mathbf{G^W}_i = \mathbf{w})\mathbb{P}(\mathbf{G^W} = \mathbf{w}|Y_i = 0)
\end{aligned}$$

Using this, I could compute the likelihood of seeing the data observed, $\mathcal{L}(\mathbf{W}, \boldsymbol{\gamma}|\mathbf{Z}^{OBS}) = \mathbb{P}(\mathbf{Z}^{OBS} = \mathbf{z}^{OBS}|\mathbf{W}, \boldsymbol{\gamma})$, using a simular approach to that used in Chapter 3 to compute $\mathbb{E}(\mathbf{Z}^{OBS}|\mathbf{W}, \boldsymbol{\gamma})$, and hence proceed with a likelihood based inference for $\mathbf{W}$ and $\boldsymbol{\gamma}$.

A disease associated SNP can either be protective (that is, with a negative Z Score) or deleterious (that is, with a positive Z Score) for the disease in question. However, in practice, it is disease association which studies report, rather than direction of effect, and so summary data is typically given as unsigned p-values; see, for instance, [International Consortium for Blood Pressure Genome-Wide Association Studies, 2012]. From these, it is only possible to compute the absolute value of the Z Scores, $|\mathbf{Z}^{OBS}|$.

Hence, the likelihood function of $\mathbf{W}$ and $\boldsymbol{\gamma}$ is actually the sum of 2^n terms:

$$\begin{aligned}
\mathcal{L}\left(\mathbf{W}, \boldsymbol{\gamma} \,\middle|\, |\mathbf{Z}^{OBS}|\right) &= \mathbb{P}\left(|\mathbf{Z}^{OBS}| = \boldsymbol{\zeta} \,\middle|\, \mathbf{W}, \boldsymbol{\gamma}\right) \\
&= \sum_{\mathbf{s}\in\{-1,1\}^n} \mathbb{P}\left(\mathbf{Z}^{OBS} = \mathbf{s} \circ \boldsymbol{\zeta} \,\middle|\, \mathbf{W}, \boldsymbol{\gamma}\right)
\end{aligned}$$

where $\mathbf{s} \circ \boldsymbol{\zeta}$ gives the elemental pairwise multiplication between the sign vector $\mathbf{s}$ and the observed absolute Z Scores $\boldsymbol{\zeta}$.

As n is typically in the 100s, a direct calculation of this likelihood is computationally intractable. I considered a number of strategies to manage this:

SNPs which are in very high LD with each other are very unlikely to have different directions of effect. Hence, the SNPs could be grouped into blocks of high LD, and contributions to the likelihood which have different signs within these blocks could be discarded.

If a SNP has a Z Score very close to 0, its contribution to the likelihood sum is unlikely to differ much by sign; for any combinations of signs over the other SNPs, the likelihood contribution when our SNP is assumed to be protective can be assumed to be identical to the contribution when our SNP is assumed to be deleterious, and hence only one of these terms needs to be computed.

However, these techniques are unlikely to reduce n down to a value n^* such that $2^{(n^*)}$ is tractable; they also risk losing a significant fraction of our information. Instead, then, I used an approach which does not require the evaluation of the likelihood function.

4.2.2 The theory of ABC

Approximate Bayesian Computation (ABC) [Tavaré et al, 1997] is a technique which enables parameter inference without the need for evaluating the likelihood function. Instead, the likelihood is approximated by simulating data from the prior distribution of the parameters and comparing simulated data to observed.

Consider the case where we have data $\mathcal{D}$ from some model with parameter θ, and we wish to make an inference about the value of θ. We have some belief, $\pi(\theta)$, about the prior distribution of θ. We sample $\tilde{\theta}_1, ..., \tilde{\theta}_p$ from $\pi(\theta)$ and for each $i = 1, \ldots, p$, we simulate $\tilde{\mathcal{D}}_i$ from the model parameterized by $\tilde{\theta}_i$. We compare $\tilde{\mathcal{D}}_i$ with the true dataset $\mathcal{D}$, using some distance metric ρ which we must specify (possibly a function of summary statistics rather than the complete data). If $\rho(\mathcal{D}, \tilde{\mathcal{D}}_i) \leq \epsilon$, for some appropriate value of ϵ, we accept the simulation; otherwise reject it. The posterior distribution of $\theta|\mathcal{D}$ can then be approximated from those $\tilde{\theta}_i$ corresponding to accepted simulations. See Figure 4.1 for an illustration of the algorithm.

As the number of parameters, $\tilde{\theta}_i$ sampled from $\pi(\theta)$, the posterior distribution of θ will converge towards the value which would be obtained by directly evaluating the likelihood, However, there are several limitations when it is applied in practice. The tolerance parameter,

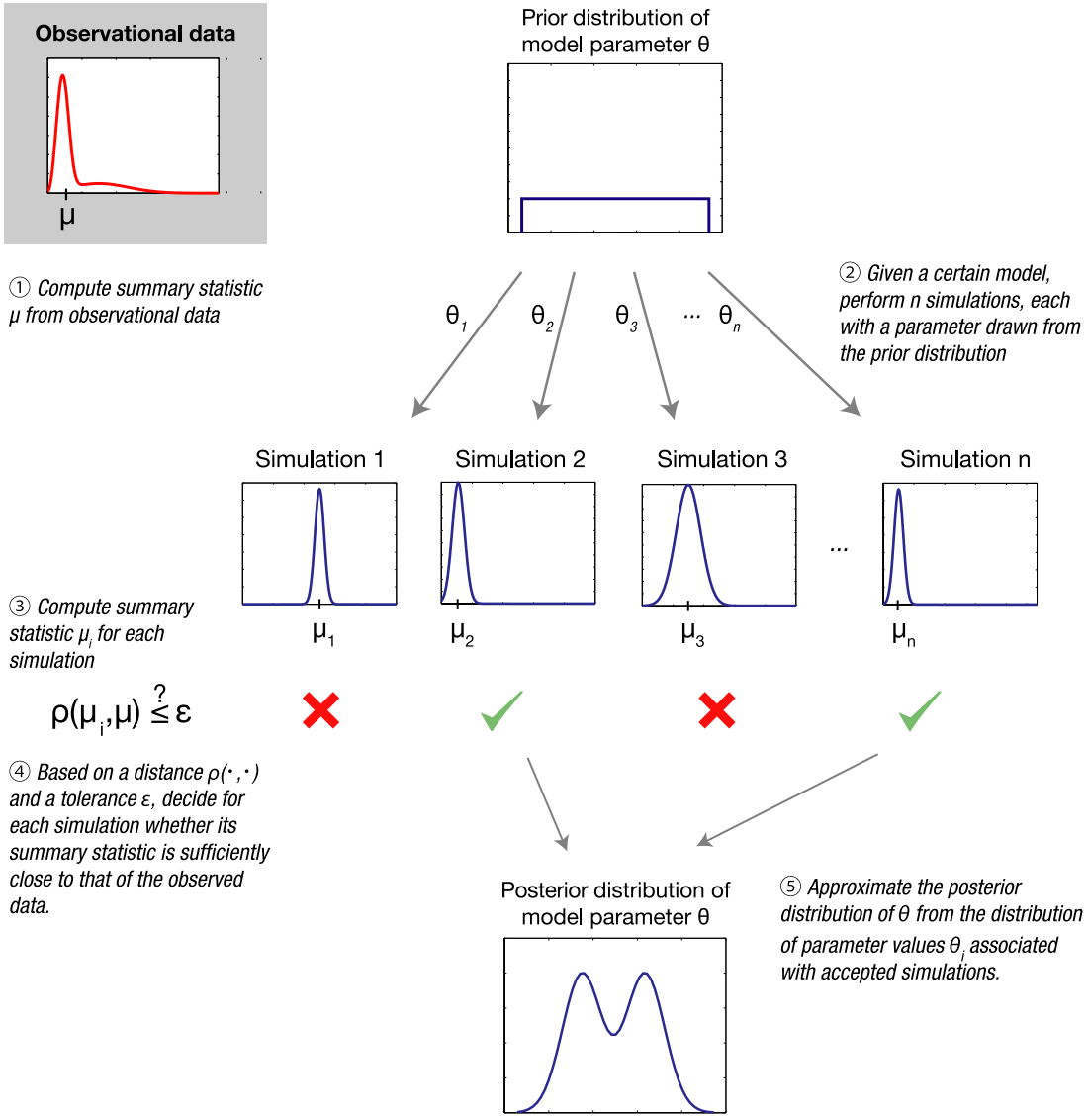


Figure 4.1 The ABC Algorithm, showing the stages of sampling from the prior distribution, computing a summary statistic for each simulation and comparing this to the observed data using some threshold ϵ. This figure is taken from [Sunnåker et al, 2013].

ϵ, must be specified, and the use of a non-zero ϵ introduces a source of bias. As in any Bayesian technique, the results are dependent upon the choice of $\pi(\theta)$. Additionally, in ABC, we must simulate from this prior; in the case of high dimensional θ, it may not be computationally feasible to sample enough datapoints to properly cover the parameter space.

4.3 Testing Goodness of Fit For a Causal Model

Given a causal model parametrised by $\{\mathbf{W}, \boldsymbol{\gamma}\}$, I need to compute how well it compares to the vector of GWAS Z Scores, $\mathbf{Z}^{OBS}$.

Section 3.2 explains how to compute Z_X^{EXP}, the expected Z Score at SNP X if the model parametrized by $\{\mathbf{W}, \boldsymbol{\gamma}\}$ is true. Then the observed Z Score at X is distributed $N(Z_X^{EXP}, 1)$. However, due to the LD between SNPs within a region, these distributions are not independent. To simulate a Z Score, given $\mathbf{Z}^{EXP}$, I take $\mathbf{LD}$ to be the matrix of correlations between SNPs, where $LD_{ii} = 1$ and $LD_{ij}, i \neq j$, is the signed R statistic between the genotypes of SNP i and SNP j. Then I compute $\mathbf{Z}^{SIM}$, the simulated output if $\{\mathbf{W}, \boldsymbol{\gamma}\}$ were true, from:

$$\mathbf{Z}^{SIM} \sim \mathbf{N}(\mathbf{Z}^{EXP}, \mathbf{LD})$$

Note that $\mathbf{Z}^{OBS}$ is itself sampled from some $\mathbf{N}(\mathbf{Z}^{TRUE}, \mathbf{LD})$.

For my purposes, $\boldsymbol{\gamma}$ is essentially a nuisance parameter. I aim to fine map a region and discover the likely causal SNPs; the exact values of their odds ratios of effect are not of particular interest. However, especially when the model contains many CVs, changing the relative values of $\boldsymbol{\gamma}$ greatly changes the pattern of $\mathbf{Z}^{EXP}$; hence, $\boldsymbol{\gamma}$ must be included. However, if a distance metric can be found which reduces the degree to which $\boldsymbol{\gamma}$-space must be sampled to verify a $\mathbf{W}$, it would be a valuable speed-up.

4.3.1 Choice of Distance Metric for Comparison

I considered two approaches to compute a measure of the goodness of fit.

4.3.1.1 First Approach: *pmcc*

When the causal SNPs $\mathbf{W}$ are correct, and $\boldsymbol{\gamma}$ exhibits the right behaviour, we would expect to see peaks of association at the same SNPs in both the observed and simulated Z Scores. If there are several such peaks, we would also expect the ratios of their heights to be similar. This corresponds to a requirement of collinearity between the observed Z Scores and the simulated Z Scores under the model, and hence to a requirement of collinearity between the absolute values of these Scores. This suggests that searching for collinearity rather than equality may suffice to find the best $\mathbf{W}$.

Hence, I tried the Pearson product moment correlation coefficient (*pmcc*) as my distance metric between the observed and expected data. This measures the linear correlation between two variables and takes values in $[-1, 1]$, with 1 corresponding to perfect collinearity, and hence perfect model fit. It is computed as:

$$pmcc(|\mathbf{Z}^{OBS}|, |\mathbf{Z}^{SIM}|) = \frac{\sum_{X \in \text{SNPs}}(|Z_X^{OBS}| - \overline{|Z^{OBS}|})(|Z_X^{SIM}| - \overline{|Z^{SIM}|})}{\sqrt{\sum_{X \in \text{SNPs}}(|Z_X^{OBS}| - \overline{|Z^{OBS}|})^2}\sqrt{\sum_{X \in \text{SNPs}}(|Z_X^{SIM}| - \overline{|Z^{SIM}|})^2}}$$

The *pmcc* provides a good measure of the linear dependence between two vectors when the individual values within these vectors are independent. However, this is not the case for a vector of Z Scores. The correlation structure within a region results in LD blocks of SNPs with similar Z Scores.

If the observed and simulated Z Scores were strongly collinear, save for at a single SNP, we would still conclude that the model fitted well, and correspondingly, the effect of the outlying SNP would not distort the *pmcc* by much. However, if that SNP happened to be in high LD with many other SNPs, the weight of that LD block would cause the *pmcc* to be low, leading us to conclude that the Z Scores were not collinear despite this behaviour truly only occurring at a single signal. Similarly, not adjusting for LD could cause false positive results; if there are very few associations significantly greater than 0, then a single significant LD block could lead to a high *pmcc*.

I used LDAK [Speed et al, 2012] to compute weightings reflecting the correlation patterns

of SNPs in my region. I then used these to compute the weighted *pmcc* between the observed and simulated Z Scores, which gives a more accurate measure of whether the Z Scores at signals are collinear. If SNP X has weight $\mathcal{W}_X$ then it is computed as:

$$pmcc(|\mathbf{Z}^{OBS}|, |\mathbf{Z}^{SIM}|) = \frac{\sum_{X \in \text{SNPs}} \mathcal{W}_X (|Z_X^{OBS}| - \overline{|\mathbf{Z}^{OBS}|})(|Z_X^{SIM}| - \overline{|\mathbf{Z}^{SIM}|})}{\sqrt{\sum_{X \in \text{SNPs}} \mathcal{W}_X (|Z_X^{OBS}| - \overline{|\mathbf{Z}^{OBS}|})^2} \sqrt{\sum_{X \in \text{SNPs}} \mathcal{W}_X (|Z_X^{SIM}| - \overline{|\mathbf{Z}^{SIM}|})^2}}$$

As in a standard ABC algorithm, I simulate $\mathbf{Z}^{SIM}$ from the model parameterized by the values I have sampled from their prior, and compare this to the observed data $\mathbf{Z}^{OBS}$. However, as an intermediate step, I compute $\mathbf{Z}^{EXP}$. If the causal model were true, then $\mathbf{Z}^{OBS} \sim \mathbf{N}(\mathbf{Z}^{EXP}, \mathbf{LD})$. By working directly with $\mathbf{Z}^{EXP}$, it might be possible to avoid having to account for the extra error in simulating $\mathbf{Z}^{SIM}$.

However, unlike the $\mathbf{Z}^{SIM}$ case, we would not expect $|\mathbf{Z}^{EXP}|$ and $|\mathbf{Z}^{OBS}|$ to be collinear in the case of a correct causal model. The majority of SNPs in most regions are not disease-associated; these will have $|\mathbf{Z}^{EXP}| \sim 0$. However, if $Z \sim N(0,1)$, then $\mathbb{E}(Z) = 0$ but $\mathbb{E}(|Z|) = \sqrt{\frac{2}{\pi}} \approx 0.8$, and so these SNPs will have $|\mathbf{Z}^{OBS}| >> 0$, resulting in the collinear behaviour being disrupted around 0. Any comparison used would have to take account of this behaviour; the simplest solution would be to discount points around 0, but this risks losing important information.

4.3.1.2 Flaw in the *pmcc* Approach

The *pmcc* initially appears to be a good metric. When tested upon datasets with known causal models, it performs well, and the number of $\boldsymbol{\gamma}$ tested before the result starts to converge is computationally manageable.

Recall that there is some true $\mathbf{Z}^{TRUE}$, and the observed Z Score $\mathbf{Z}^{OBS} \sim \mathbf{N}(\mathbf{Z}^{TRUE}, \mathbf{LD})$. Similarly, the simulated Z Score $\mathbf{Z}^{SIM} \sim \mathbf{N}(\mathbf{Z}^{EXP}, \mathbf{LD})$.

Consider two cases. In Case (a), the model being tested is the correct one, and $\mathbf{Z}^{EXP} = \mathbf{Z}^{TRUE}$. However, in Case (b), while the model being tested is substantially correct, its odds

ratio is inflated, causing $\mathbf{Z}^{EXP} = \alpha \mathbf{Z}^{TRUE}$ for some constant $\alpha > 1$.

In Case (b), $\sum_{X \in \text{SNPs}} \mathcal{W}_X (|Z_X^{OBS}| - \overline{|\mathbf{Z}^{OBS}|})^2$ is unchanged from Case (a). However, since $\mathbf{Z}^{EXP}$ has changed, the distribution of $\mathbf{Z}^{SIM}$ has also changed. In Case (b), the expectation of $\sum_{X \in \text{SNPs}} \mathcal{W}_X (|Z_X^{OBS}| - \overline{|\mathbf{Z}^{OBS}|})(|Z_X^{SIM}| - \overline{|\mathbf{Z}^{SIM}|})$ is equal to α times the value obtained in Case (a). However, while the mean in Case (b) has changed, the variance of $\mathbf{Z}^{SIM}$ about $\mathbf{Z}^{EXP}$ is still 1, and so the expectation of $\sum_{X \in \text{SNPs}} \mathcal{W}_X (|Z_X^{SIM}| - \overline{|\mathbf{Z}^{SIM}|})^2$ is less than α^2 times the value obtained in Case (a).

Hence, the *pmcc* in Case (a) is less than the *pmcc* in Case (b), despite Case (a) being the correct model (See Figure 4.2 for an illustration of this effect). For any model, we can improve the *pmcc* by sampling a gamma which results in the same $\mathbf{Z}^{EXP}$ multiplied by a constant factor; any deviation away from the line of best fit being $y = x$ will improve our apparent fit. It follows that a $\boldsymbol{\gamma}$ which results in an extremely large $\mathbf{Z}^{EXP}$ will result in a high *pmcc* regardless of true model fit.

Note that this effect is not due to some peculiarity in the *pmcc*, but will also hold for any other measure of collinearity (for instance, the Adjusted R Squared from the best fitting linear model between $\mathbf{Z}^{OBS}$ and $\mathbf{Z}^{SIM}$ will behave identically).

This means that, although in general a high *pmcc* will correspond to a good model fit, *pmcc* is not a valid proxy for model fit. It is particularly dangerous to use *pmcc* to compare two models which both appear to fit well, since the slight improvement in the "better" model may well only be due to it having an inflated $\boldsymbol{\gamma}$.

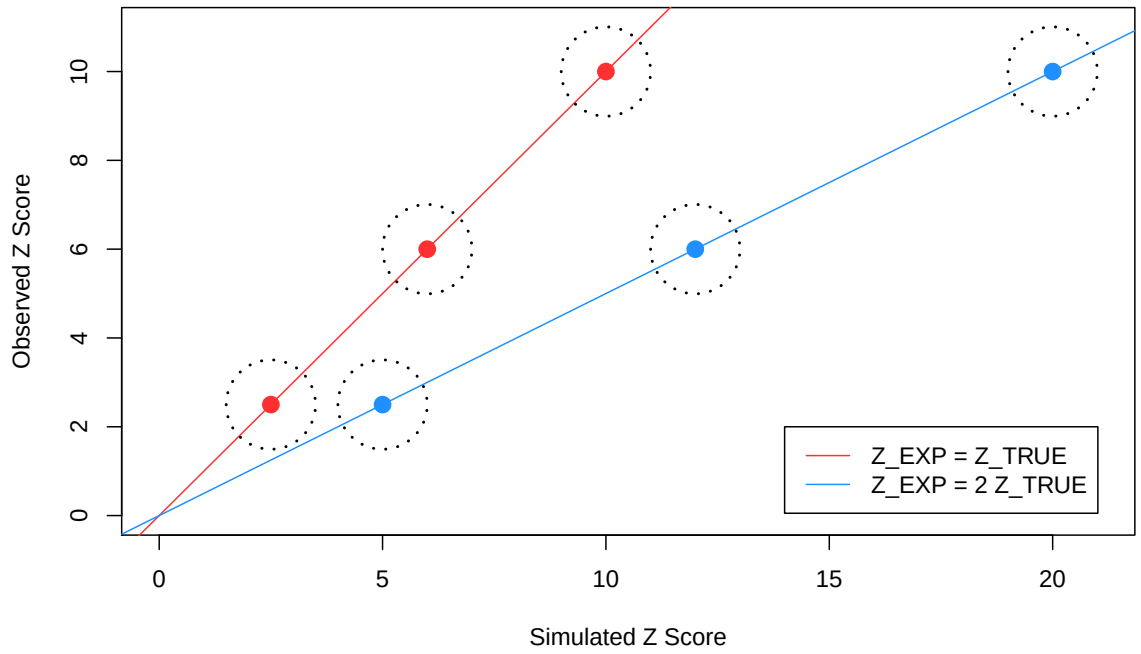


Figure 4.2 The red line shows the case when the model is correct, with dotted circles showing the errors in the points. The blue line shows a case when the causal SNPs are correct, but the odds ratios sampled cause the $\mathbf{Z}_{EXP}$ to be inflated to twice the true values. Since the variance of $\mathbf{Z}_{SIM}$ around $\mathbf{Z}_{EXP}$ is the same in each case, sampled points about the blue line are more collinear than sampled points about red line, and will perform better under distance metrics such as *pmcc*.

4.3.1.3 New Approach: Weighted Sum of Squares

Since any test for collinearity will have the flaw described above, instead I must use a distance measure which tests for equality. Although this unavoidably requires sampling a larger number of $\boldsymbol{\gamma}$ in order to have a good chance of verifying whether a given $\mathbf{W}$ fits well, it does mean that any $\boldsymbol{\gamma}$ which performs well is likely to be close to the correct one; hence such a distance metric also enables us to make inference about the odds ratios of the causal SNPs if desired.

Any measure of whether two vectors are identical is fundamentally a function of the distance of the points from the $y = x$ line. The simplest of these measures the sum of squares of the difference between the observed Z Score and the Z Score under my model, and it is this I shall use for my distance metric. As in the *pmcc* case, I shall weight this by the values computed by LDAK, in order to control for the effect of large LD blocks.

Although standard ABC uses $|\mathbf{Z}^{SIM}|$, simulating such values produces extra error, requiring additional runs of the algorithm to converge to an optimal solution, which is not necessary in this case. Since under the true model $\mathbf{Z}^{OBS} \sim \mathbf{N}(\mathbf{Z}^{EXP}, \mathbf{LD})$, I shall directly compare $|\mathbf{Z}^{OBS}|$ and $|\mathbf{Z}^{EXP}|$. (I did not do this in the *pmcc* case, since when $\mathbf{Z}^{EXP}$ is close to zero, we expect $|\mathbf{Z}^{OBS}| > |\mathbf{Z}^{EXP}|$ and so collinear behaviour is disrupted even in the case of the true causal model. However, the sum of squared differences does not take account of in which direction deviance from the $y = x$ line has occurred, and so it is valid to compare $|\mathbf{Z}^{OBS}|$ and $|\mathbf{Z}^{EXP}|$.)

Hence, my measure of model fit is:

$$sumsq(|\mathbf{Z}^{OBS}|, |\mathbf{Z}^{EXP}|) = \frac{\sum_{X \in \text{SNPs}} \mathcal{W}_X (|Z_X^{OBS}| - |Z_X^{EXP}|)^2}{\sum_{X \in \text{SNPs}} \mathcal{W}_X}$$

Note that, while the true causal model will have a low *sumsq*, we would expect spurious results in cases where $\mathbf{Z}^{OBS}$ is close to zero. The variance in the true model at any SNP X between Z_X^{OBS} and Z_X^{TRUE} is 1; if all Z Scores are small, there will be many other non-significant and non-associated models which still have their $\mathbf{Z}^{EXP}$ within the bounds of variance 1 from $\mathbf{Z}^{OBS}$. However, this is a feature of the variance in the observed Z Score, and we would see this effect regardless of distance measure used. It is not appropriate to perform a fine mapping

analysis upon a region which has no evidence of disease association.

4.3.2 Choice of Tolerance Parameter

Given a value of *sumsq*, we wish to know whether it is consistent with the model being analysed being the true causal model. In the standard ABC approach, this would be done by accepting a model if $sumsq < \epsilon$ for some tolerance parameter.

4.3.2.1 True Distribution of $sumsq(|\mathbf{Z}^{OBS}|, |\mathbf{Z}^{EXP}|)$

If the model is true, then $\mathbf{Z}^{OBS} \sim N(\mathbf{Z}^{EXP}, \mathbf{LD})$; equally $\mathbf{Z}^{EXP} \sim N(\mathbf{Z}^{OBS}, \mathbf{LD})$. (It is this second formulation I shall use, since it more accurately reflects the nature of my task; $\mathbf{Z}^{OBS}$ is a fixed input, and I wish to find a model whose resulting $\mathbf{Z}^{EXP}$ is consistent with $\mathbf{Z}^{EXP} \sim N(\mathbf{Z}^{OBS}, \mathbf{LD})$.) Hence, at any SNP X, $|Z_X^{OBS}|$ has a folded normal $(Z_X^{EXP}, 1)$.

The folded normal $(\mu, 1)$ distribution has moments:

$$\begin{aligned}
\mathbb{E}Z &= \sqrt{\frac{2}{\pi}} e^{\frac{-\mu^2}{2}} + \mu\left[1 - 2\Phi(-\mu)\right] \\
\mathbb{E}Z^2 &= \mu^2 + 1 \\
\mathbb{E}Z^3 &= (\mu^2 + 2)\mathbb{E}z - \mu\left[1 - 2\Phi(-\mu)\right] \\
\mathbb{E}Z^4 &= \mu^4 + 6\mu^2 + 3
\end{aligned}$$

Hence I can compute:

$$\begin{aligned}
\mathbb{E}\left[\left(|Z_X^{OBS}| - |Z_X^{EXP}|\right)^2\right] &= \mathbb{E}\left(|Z_X^{EXP}|^2\right) - 2|Z_X^{OBS}|\mathbb{E}\left(|Z_X^{EXP}|\right) + |Z_X^{OBS}|^2 \\
&= 1 + 2|Z_X^{OBS}|^2 - 2|Z_X^{OBS}|\left(\sqrt{\frac{2}{\pi}} e^{\frac{-(Z_X^{OBS})^2}{2}} + Z_X^{OBS}\left[1 - 2\Phi(-Z_X^{OBS})\right]\right)
\end{aligned}$$

And thence:

$$\mathbb{E}\left(sumsq(|\mathbf{Z}^{OBS}|,|\mathbf{Z}^{EXP}|)\right) = \frac{\sum_{X\in\text{SNPs}} \mathcal{W}_X \mathbb{E}\left(|Z_X^{OBS}| - |Z_X^{EXP}|\right)^2}{\sum_{X\in\text{SNPs}} \mathcal{W}_X}$$

Similarly, I can compute

$$Var\left(|Z_X^{OBS}| - |Z_X^{EXP}|\right)^2 = \mathbb{E}\left(|Z_X^{OBS}| - |Z_X^{EXP}|\right)^4 - \left(\mathbb{E}\left(|Z_X^{OBS}| - |Z_X^{EXP}|\right)^2\right)^2$$

With these values, I could in theory compute some appropriate tolerance parameter ϵ, which would enable me to select or reject causal models, with the selected models being used to approximate the true distribution of $\{\mathbf{W}, \boldsymbol{\gamma}\}$. However, this distribution is complex, and not easily tractable. Is it possible instead to approximate its behaviour and thereby obtain a practical form for ϵ?

4.3.2.2 Approximating the Distribution of $sumsq(|\mathbf{Z}^{OBS}|,|\mathbf{Z}^{EXP}|)$

Under the true model, $\left(Z_X^{OBS} - Z_X^{EXP}\right)^2 \sim \chi_1^2$, which has mean 1 and variance 2. When Z_X^{OBS} and Z_X^{EXP} have the same sign,

$$\left(|Z_X^{OBS}| - |Z_X^{EXP}|\right)^2 = \left(Z_X^{OBS} - Z_X^{EXP}\right)^2$$

Hence, if the majority of observed Z Scores are sufficiently greater than 0, we have:

$$\mathbb{E}sumsq = \frac{\sum_{X\in\text{SNPs}} \mathcal{W}_X}{\sum_{X\in\text{SNPs}} \mathcal{W}_X} = 1$$

However, when the Z Scores are ~ 0,

$$\left(|Z_X^{OBS}| - |Z_X^{EXP}|\right)^2 < \left(Z_X^{OBS} - Z_X^{EXP}\right)^2$$

Since, in practice, the majority of SNPs in a region are not disease associated, we can expect the true mean of *sumsq* to be below 1; any more precision will require considering the behaviour

of the folded normal distribution around the origin.

Assuming all SNPs to be independent, then:

$$\text{Var}\left(\frac{\sum_{X\in\text{SNPs}}\mathcal{W}_X(Z_X^{OBS}-Z_X^{EXP})^2}{\sum_{X\in\text{SNPs}}\mathcal{W}_X}\right)=\frac{\sum_{X\in\text{SNPs}}\mathcal{W}_X^2}{(\sum_{X\in\text{SNPs}}\mathcal{W}_X)^2}$$

Although the SNPs are not independent, the SNP weightings computed by LDAK ensure that no two SNPs in high LD are included in the sum, and hence this is a valid assumption to a first approximation. As in the $\mathbb{E}$ case, the effect of non-associated SNPs means that this is likely to be an overestimate for the variance of *sumsq*.

4.3.2.3 Use of $sumsq(|\mathbf{Z}^{OBS}|,|\mathbf{Z}^{EXP}|)$ for Comparison

Hence, while I expect well fitting models to have $sumsq(|\mathbf{Z}^{OBS}|,|\mathbf{Z}^{EXP}|)<1$, there is no obvious way to compute an appropriate tolerance parameter ϵ. Any such ϵ would need to be region specific; certainly depending upon the weights $\mathcal{W}_X$, but possibly also upon the values of $\mathbf{Z}^{OBS}$. However, it might be possible to use simulations to choose and appripiate ϵ for a specific instance.

If the model being testesd is truely causal, then $\mathbf{Z}^{OBS}\sim N(\mathbf{Z}^{EXP},\mathbf{LD})$, and so $\mathbf{Z}^{EXP}\sim N(\mathbf{Z}^{OBS},\mathbf{LD})$. Hence, by simulating from a $N(|\mathbf{Z}^{OBS}|,\mathbf{LD})$ distribution and taking absolute values we can estimate the space where we will see $|\mathbf{Z}^{EXP}|$ for the true causal model and hence the empirical distribution of $sumsq(|\mathbf{Z}^{OBS}|,|\mathbf{Z}^{EXP}|)$ for the true causal model. ϵ can then be taken to be some appropiate quantile of this distribution.

Future work will explore choices of ϵ. However, within a region, *sumsq* provides a very good proxy for how close each proposed causal model comes to the observed data relative to other proposed models. For the rest of this chapter, I shall use *sumsq* to provide a ranking of models within a region, but shall not attempt to compare *sumsq* values between regions.

4.4 Implementation of Algorithm

4.4.1 Efficient Search of Model Space

Implementing this method using a standard ABC framework would proceed as in Algorithm 1. The priors discussed in Section 3.3 are used as the priors for the causal model, $\Pi_{\mathbf{W}}$ and $\Pi_{\boldsymbol{\gamma}}$.

Algorithm 1 Inferring Causal Models using ABC

1: **procedure** IMPLEMENTATION1
2: load $|\mathbf{Z}_{OBS}|$
3: load RefData
4: load N_0, N_1
5: load LDAK weights
6: compute **LD** from RefData
7: **for** iterations in NumIterations **do**
8: sample m from Π_m, $\mathbf{W}$ from $\Pi_{\mathbf{W}}$, $\boldsymbol{\gamma}$ from $\Pi_{\boldsymbol{\gamma}}$
9: compute γ_0
10: extract relationships between $\mathbf{W}$ and other SNPs from RefData
11: compute $\mathbf{Z}_{EXP}$
12: compute $\mathbf{Z}_{SIM} \sim N(\mathbf{Z}_{EXP}, \mathbf{LD})$
13: compute $sumsq_{\mathbf{W},\boldsymbol{\gamma}}$, the weighted $sumsq$ between $|\mathbf{Z}_{SIM}|$ and $|\mathbf{Z}_{OBS}|$
14: **if** $sumsq_{\mathbf{W},\boldsymbol{\gamma}} < \epsilon$ **then**
15: accept the model $\{\mathbf{W}, \boldsymbol{\gamma}\}$
16: **return** Distribution inferred from all models $\{\mathbf{W}, \boldsymbol{\gamma}\}$ with $sumsq_{\mathbf{W},\boldsymbol{\gamma}} < \epsilon$

However, as discussed in Section 4.3, I am not using a standard ABC framework. Rather than sampling $\mathbf{Z}_{SIM}$, I am directly using $\mathbf{Z}_{EXP}$ in my comparison. Also, rather than having a single tolerance parameter ϵ, I am using my distance metric, $sumsq$ to rank all models sampled; I can then return either the ranking of all models, or the identities of the top N_{mod} models. Hence, my method proceeds as in Algorithm 2.

Note that in these algorithms, the step on Row 9 of Algorithm 2, "extract relationships between $\mathbf{W}$ and other SNPs from RefData" must be repeated for each iteration. This computation is complex, requiring the inference of genotype probabilities from haplotype data, and the requirement to perform it NumIterations times results in a procedure which is prohibitively slow to run on typical genomic regions. Many times this is repeated computation, since in order to reasonably sample the causal model space, many $\boldsymbol{\gamma}$ must be

Algorithm 2 My Method: Inferring Causal Models using the Weighted Sums of Squares as Distance Metric

procedure IMPLEMENTATION2
    load $|\mathbf{Z}_{OBS}|$
    load RefData
    load N_0, N_1
    load LDAK weights
    for iterations in NumIterations **do**
        sample m from Π_m, $\mathbf{W}$ from $\Pi_{\mathbf{W}}$, $\boldsymbol{\gamma}$ from $\Pi_{\boldsymbol{\gamma}}$
        compute γ_0
        extract relationships between $\mathbf{W}$ and other SNPs from RefData
        compute $\mathbf{Z}_{EXP}$
        compute $sumsq_{\mathbf{W},\boldsymbol{\gamma}}$, the weighted sum of squared differences between $|\mathbf{Z}_{OBS}|$ and $|\mathbf{Z}_{EXP}|$
        save $\{\mathbf{W}, \boldsymbol{\gamma}, sumsq_{\mathbf{W},\boldsymbol{\gamma}}\}$
    return Models $\{\mathbf{W}, \boldsymbol{\gamma}\}$ with top N_{mod} $sumsq_{\mathbf{W},\boldsymbol{\gamma}}$

sampled for each $\mathbf{W}$, and this step is dependent only upon $\mathbf{W}$.

One option might be to compute this data once for each $\mathbf{W}$ at the start of the procedure. This could then be saved and retrieved when needed. However, although this is computationally more efficient, it is unrealistic in terms of memory usage. Consider the regions analysed in Chapter 2 as an example of typical region size. After QC, these have in expectation 261 SNPs; some have many more. This corresponds to 33930 2 SNP models, and over 9 billion 5 SNP models (although with the prior on $m = 5$ being 2.98×10^{-11}, not many 5 SNP models will be sampled); it is infeasible to store results for all of these.

Instead, rather than only considering a single $\boldsymbol{\gamma}$ at each iteration, I propose to sample from $\mathbf{W}$, and then test many values of $\boldsymbol{\gamma}$ before moving onto the next $\mathbf{W}$, as in Algorithm 3. For each $\mathbf{W}$, I can save the top $sumsq$, and compare these.

Algorithm 3 My Method: Inferring Causal Models using the Weighted Sums of Squares as Distance Metric

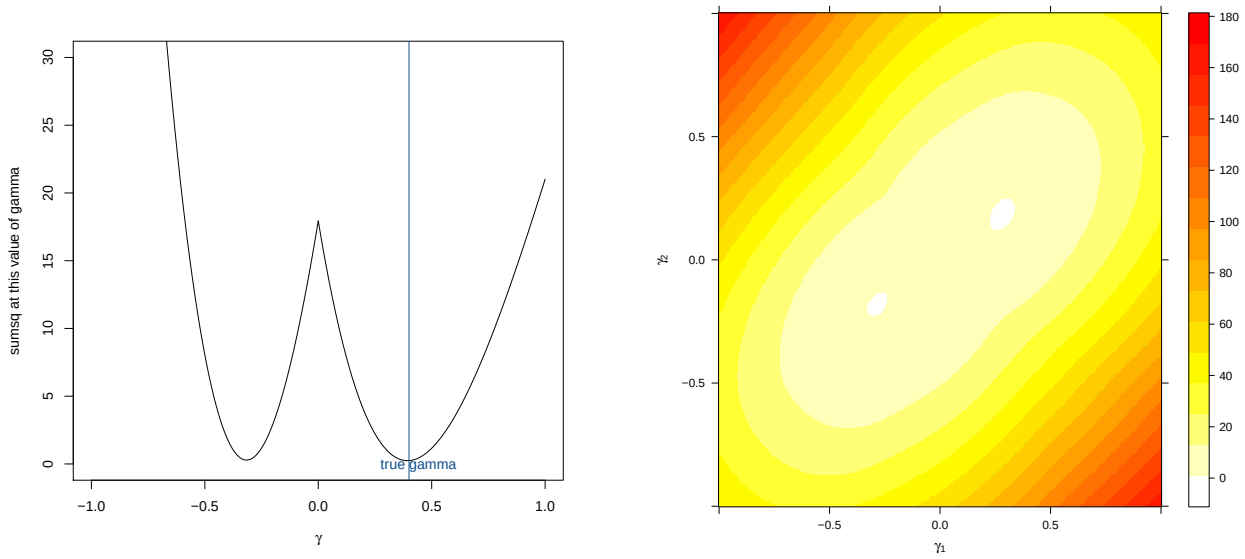
procedure IMPLEMENTATION2
  load $|\mathbf{Z}_{OBS}|$
  load RefData
  load N_0, N_1
  load LDAK weights
  for iterations in N_W **do**
    sample m from Π_m, $\mathbf{W}$ from $\Pi_{\mathbf{W}}$
    extract relationships between $\mathbf{W}$ and other SNPs from RefData
    for iterations in N_γ **do**
      sample $\boldsymbol{\gamma}$ from $\Pi_{\boldsymbol{\gamma}}$
      compute γ_0
      compute $\mathbf{Z}_{EXP}$
      compute $sumsq_{\mathbf{W},\boldsymbol{\gamma}}$
      save $\{\mathbf{W}, \boldsymbol{\gamma}, sumsq_{\mathbf{W},\boldsymbol{\gamma}}\}$
    find $sumsq_{\mathbf{W}}$, the top $sumsq_{\mathbf{W},\boldsymbol{\gamma}}$ for this $\mathbf{W}$, which occurs at $\boldsymbol{\gamma}^*_{\mathbf{W}}$
    save $\{\mathbf{W}, \boldsymbol{\gamma}^*_{\mathbf{W}}, sumsq_{\mathbf{W}}\}$
  return Models $\{\mathbf{W}, \boldsymbol{\gamma}^*_{\mathbf{W}}\}$ with top N_{mod} $sumsq_{\mathbf{W}}$

4.4.2 Number of Samples of Gamma

Recall from above that N_γ is the number of times we sample $\boldsymbol{\gamma}$ for a particular value of $\mathbf{W}$. It is not obvious what the value of N_γ should be; it may well be that a more efficient approach is instead to use some probabilistic technique such as simulated annealing to optimise the *sumsq* over $\boldsymbol{\gamma}$. However, in this section, I concern myself merely with the question: if $\mathbf{W}$ is indeed the true set of causal variants, how many $\boldsymbol{\gamma}$ must I sample from $\Pi_{\mathbf{W}}$ to demonstrate this? Since the size of the search space is a subset of $\mathbb{R}^m$, the appropriate N_γ will increase as m does.

Figure 4.3 shows a typical distribution of *sumsq* when plotted against $\boldsymbol{\gamma}$, in both the 1-CV and 2-CV cases, for simulated data (generated using the method discussed in Section 3.4.2) with a known $\mathbf{W}$ also used to calculate *sumsq*. We see that the plots are smooth, dipping to a minimum around 1 at both $\boldsymbol{\gamma}^{TRUE}$ and $-\boldsymbol{\gamma}^{TRUE}$ (since we have only the absolute value of the Z Score, it is not likely to be possible to distinguish between these two cases). The gradients of these dips are such that, especially around $\boldsymbol{\gamma}^{TRUE}$, if two values of $\boldsymbol{\gamma}$ are close to each other, the resulting *sumsq* statistics will also be close to each other. Hence, sampling a $\boldsymbol{\gamma}$ which is

"too close" to an already-analysed value of $\boldsymbol{\gamma}$ adds very little useful information, and is wasted computation time. Rather than sample independently from $\Pi_{\boldsymbol{\gamma}}$, I propose instead to reject new samples if they are within some distance $\delta_{\boldsymbol{\gamma}}$ of an already existing $\boldsymbol{\gamma}$ (with distance defined as the Euclidean distance in the multi-dimensional cases). Since the width of the "dip" in *sumsq* appears similar for different values of $\boldsymbol{\gamma}$ and for different value of m, I shall determine $\delta_{\boldsymbol{\gamma}}$ from the 1-CV case and then use it for all analyses.



(a) (b)

Figure 4.3 Isolated examples to show the shape of the distribution of the *sumsq* statistic when $\boldsymbol{\gamma}$ varies, for the true value of $\mathbf{W}$. These are results from simulated data, generated using the method discussed in Section 3.4.2.
(a) A single causal variant. This region had $N_0 = 11750$, $N_1 = 6500$ and $\boldsymbol{\gamma} = 0.4$. The observed p-value at the causal SNP was 8.04×10^{-19}, corresponding to an observed Z Score of 8.86.
(b) Two causal variants. This region had $N_0 =$, $N_1 =$ and $\boldsymbol{\gamma} = (0.25, 0.2)$. The observed p-values at the causal SNPs were $(2.50 \times 10^{-20}, 1.57 \times 10^{-5})$, corresponding to an observed Z Score of $(9.24, 4.32)$.

I am satisfied if I sample a $\boldsymbol{\gamma}$ with a *sumsq* within 0.1 of that generated at the $\boldsymbol{\gamma}^{TRUE}$. 0.1 is a small enough difference that I shall still categorise the model as well fitting; it is large enough that I will not have to set my $\delta_{\boldsymbol{\gamma}}$ to an impractically low value. If at the end of my algorithms I have found a small number of models which all perform very well, and have *sumsq*

within ~ 0.1 of each other, making it impossible to rank them, then I could run a second-pass analysis of these models alone, trying more $\boldsymbol{\gamma}$ in order to hone in upon the true relative fits.

I calculated the size of this "dip" in 500 simulated regions with true causal model containing a single CV. The median width of the area within 0.1 of the *sumsq* corresponding to $\boldsymbol{\gamma}^{TRUE}$ is 0.1. Hence, I shall take $\delta_{\boldsymbol{\gamma}} = 0.025$, a quarter of this distance, to ensure that with a reasonable sample size $N_{\boldsymbol{\gamma}}$, I sample within this area. Note also that the median width of the area within 0.01 of the *sumsq* corresponding to $\boldsymbol{\gamma}^{TRUE}$ is 0.06, so it is likely that at least one of my models tested will be in this area also.

Given a set of true causal variants $\mathbf{W}$, I wish to know how many $\boldsymbol{\gamma}$, $N_{\boldsymbol{\gamma}}$, I must sample (pruning all within distance $\delta_{\boldsymbol{\gamma}} = 0.025$ of a value of $\boldsymbol{\gamma}$ already analysed) in order to achieve convergence to within 0.1 of the *sumsq* obtained at the true model. I therefore simulated 600 datasets with known 1-CV models, and 600 datasets with known 2-CV models, using the method given in Section 3.4.2, and tested different values of $\boldsymbol{\gamma}$ within models containing the true $\mathbf{W}$. I counted the number of $\boldsymbol{\gamma}$ tested until the resulting minimum *sumsq* came to within 0.1 of the *sumsq* obtained at the true causal model, for 1-CV and 2-CV cases. Figure 4.4 shows the empirical cumulative distribution of this value. From this, I shall take $N_{\boldsymbol{\gamma}} = 100$ in the 1-CV case, and $N_{\boldsymbol{\gamma}} = 2000$ in the 2-CV case. In order to obtain $\sim N_{\boldsymbol{\gamma}}$ values 0.025 apart, I shall sample $N^*_{\boldsymbol{\gamma}} > N_{\boldsymbol{\gamma}}$, and filter by distance.

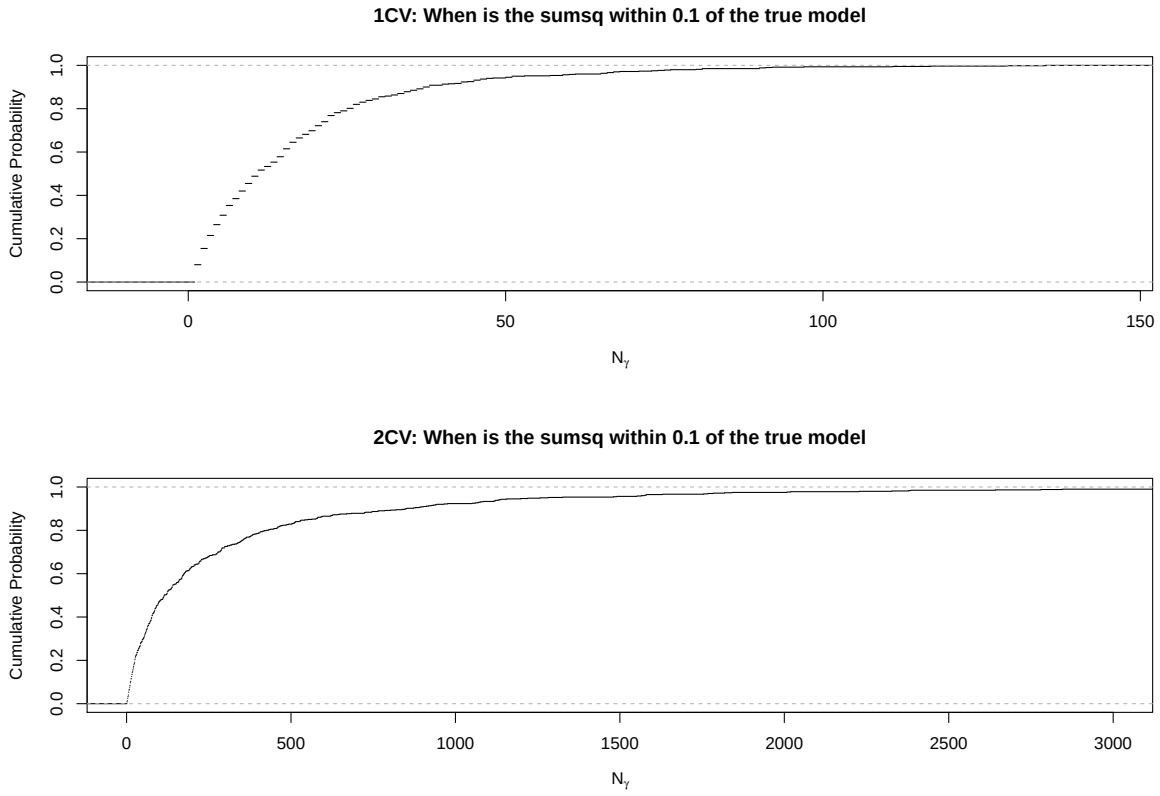


Figure 4.4 The value of N_γ required to converge within 0.1 of the true value, when analysing with the true $\mathbf{W}$ for 1 CV and 2 CV models, taking only values of $\boldsymbol{\gamma}$ which are distance greater than 0.025 away from any which have already been sampled. Distribution is estimated using the results from 600 simulations for each of the 1-CV and 2-CV models.

4.4.3 Sampling from $\Pi_{\mathbf{W}}$: Random Sampling or Exhaustive Search?

In a typical application of ABC, causal models are randomly sampled from their priors; in the limit, resulting posterior distribution will converge towards the value which would be obtained by directly evaluating the likelihood. However, I am not preforming a true random choice of causal models independent of those previously sampled. Instead, each time I sample a $\mathbf{W}$, I test enough $\boldsymbol{\gamma}$ to hopefully confirm whether it is a plausible candidate for the true causal SNPs. Following this, there is little further value in testing $\mathbf{W}$ again; hence, for efficiency, I should sample from $\Pi_{\mathbf{W}}$ without replacement.

That being the case, it often makes sense to perform an exhaustive search of all models of interest, rather than sample from $\Pi_{\mathbf{W}}$ (indeed, using the priors given in Section 3.3, it is likely that sampling without replacement will result in an exhaustive search of all 1-CV models regardless). Rather than doing a slow sampling of the entire model space, I could perform a relatively fast test of all 1CV models, to determine whether data in a given region is compatible with having a single causal SNP, as a first-pass analysis of a large number of regions. By testing all 1-CV and 2-CV models, I gain a good understanding of the causal structure within the region, and with a better appreciation of the limitations of my conclusions than had I performed a random search. The results of my analysis may also suggest avenues to explore even if none of the models considered appear to adequately fit the observed Z Score.

One case in which exhaustive testing is certainly appropriate is when potential causal SNPs have been narrowed down by prior research to a small number, and the question of interest is which specific SNPs are truly causal. Indeed, in this case, the number of possible $\mathbf{W}$ being tested is so few that I would recommend using a higher value of N_γ than that suggested in Section 4.4.2, in order to more accurately rank the goodness of fit of the models. If the region has complex association, several of the models may perform well.

4.5 This Approach Allows Direct Evaluation of Non-Genotyped SNPs without any Imputation

GWAS typically provide sparse coverage of the genome. It may often be the case that there are SNPs we are interested in considering as potential causal variants, and for which we have reference data, but for which we do not have p-values to analyse. Possibly the GWAS contains only common SNPs, and we wish to consider a denser set of causal models in a region. Possibly a SNP which has elsewhere been implicated in the disease process failed QC in the GWAS, but we nevertheless wish to include it in my fine mapping analysis. If I had access to full genotype data from the GWAS, I could use imputation to infer the genotypes of the missing SNPs from reference data, and hence approximate their true p-values. However, if this data is not available, then imputation is impossible.

However, my method computes the expected Z Score at any SNP as a function of the correlation between it and the causal variants within the model being analysed, and compares this to the vector of observed Z Scores. If the true causal variants are not genotyped, a comparison of observed and expected Z Scores at a subset of SNPs containing some in LD with the causal SNPs will suffice as a measure of model fit. (If no SNPs in LD with the true causal variants are genotyped, the region will appear to have no disease association, and hence any fine mapping approach must necessarily fail). In this section, I explore the application of my method when the most likely causal variants are "hidden" in two exemplar regions.

4.5.1 1p13.2 Region Containing *PTPN22*

As an example of this, consider the 1p13.2 region containing candidate causal gene *PTPN22*. This region shows very strong evidence of T1D-association (minimum p-value $< 10^{-80}$). In addition, the association appears to be caused by a single causal SNP. Figure 4.5 shows the Manhattan plot of T1D-associated p-values in this region; from this we see that there are clearly two top SNPs, rs6679677 and rs2476601. They have an R^2 of 0.995, and so it would take a much larger sample size than is realistically feasible to disentangle their effect in a fine

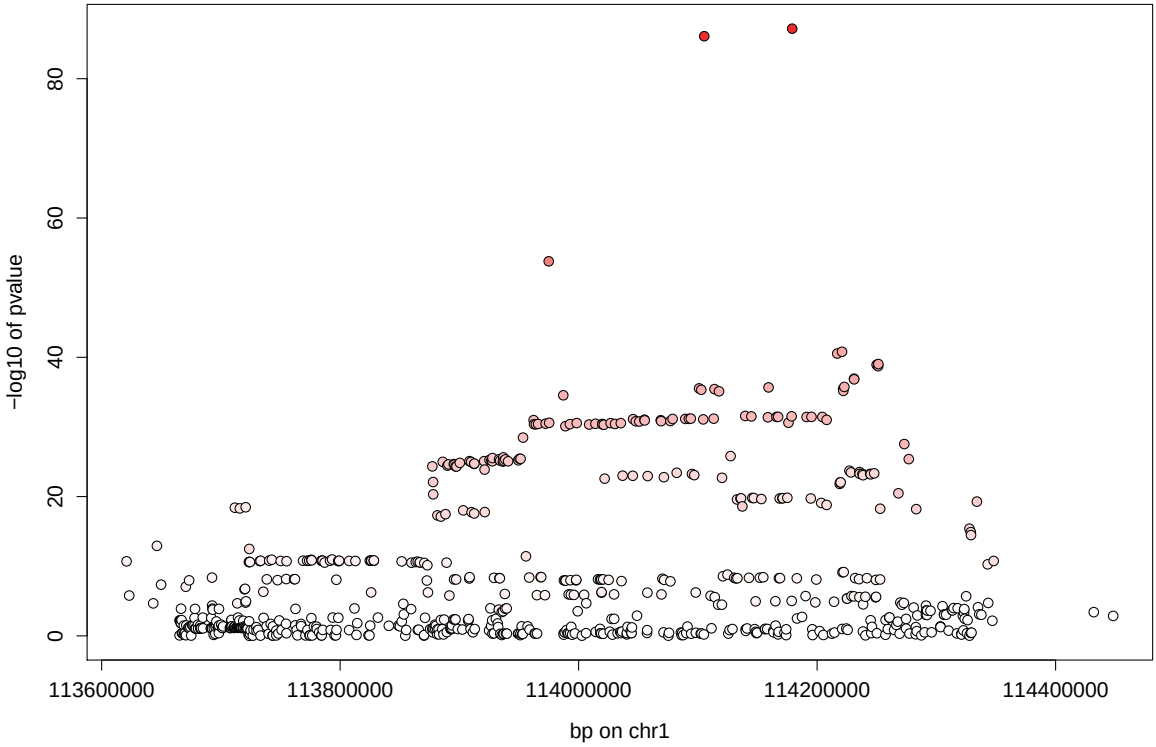


Figure 4.5 Manhattan plot of T1D Association in the 1p13.2 region containing candidate causal gene *PTPN22*. SNPs are coloured according to their LD with rs2476601, the probable causal SNP.

mapping analysis. However, rs2476601 is a non-synonymous protein coding SNP, and hence is generally accepted to be the causal variant.

Using p-values obtained from the ImmunoChip T1D study discussed in Section 2.6 I conducted a fine mapping analysis of the region. I computed Approximate Bayes Factors for each SNP using Wakefield's approximation [Wakefield, 2009] and used these to compute posterior probabilities that the SNP is the single causal SNP in the region, following [Bowes et al, 2015]. This is now a standard fine mapping technique, but it assumes that there is a single causal variant, and requires that all SNPs have been assigned a p-value. The results are shown in Figure 4.6.

In the complete analysis, rs2476601 is assigned posterior probability of being causal 0.993, while the posterior probability of rs6679677 being causal is 0.007. Although rs2476601 does

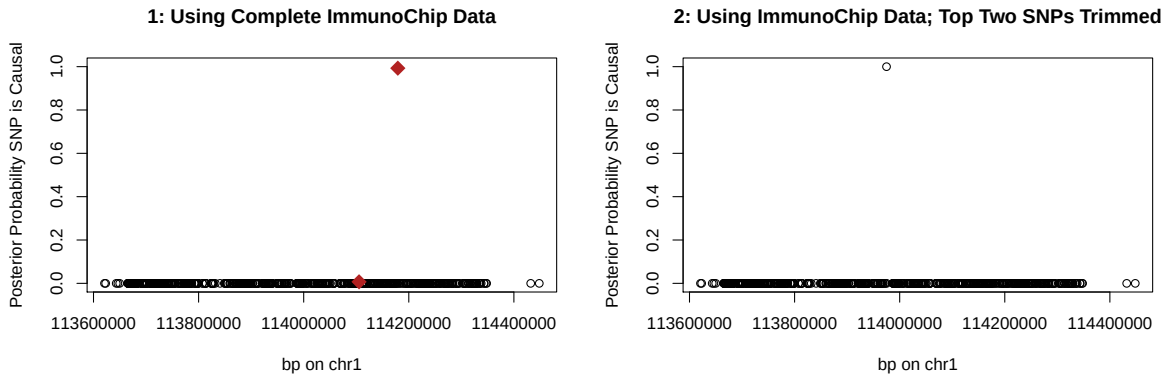


Figure 4.6 Posterior probabilities of being causal, from a standard single CV summary statistic fine mapping of the *PTPN22* region for T1D, using Wakefield's approximation to computhe Bayes Factors. The two SNPs highlighted in red in the left-hand plot are rs6679677 and rs2476601. These have been trimmed from the p-values analysed in the right-hand plot.

have a slightly lower p-value, since both SNPs have p-value $< 10^{-80}$, and are in high LD, this assignment of probability of causality seems overconfident given the data it is based on. When rs6679677 and rs2476601 are trimmed, the fine mapping confidently assigns causality to the third signal in the region, rs1230666. Hence, by removing the top SNPs, my inference changes completely.

I ran my method upon single SNP models on this region, using Z Scores taken from the following sources:

1. An ImmunoChip analysis of T1D for UK samples, data as discussed in Section 2.6 (containing 687 SNPs)

2. The same ImmunoChip analysis, but with the p-values for the top pair of SNPs removed (containing 685 SNPs)

3. A GWAS analysis of T1D for UK samples [Barrett et al, 2009a], containing a subset of those SNPs present in the ImmunoChip analysis (containing 127 SNPs)

4. The same GWAS analysis, but with the p-values for the top pair of SNPs removed (containing 125 SNPs)

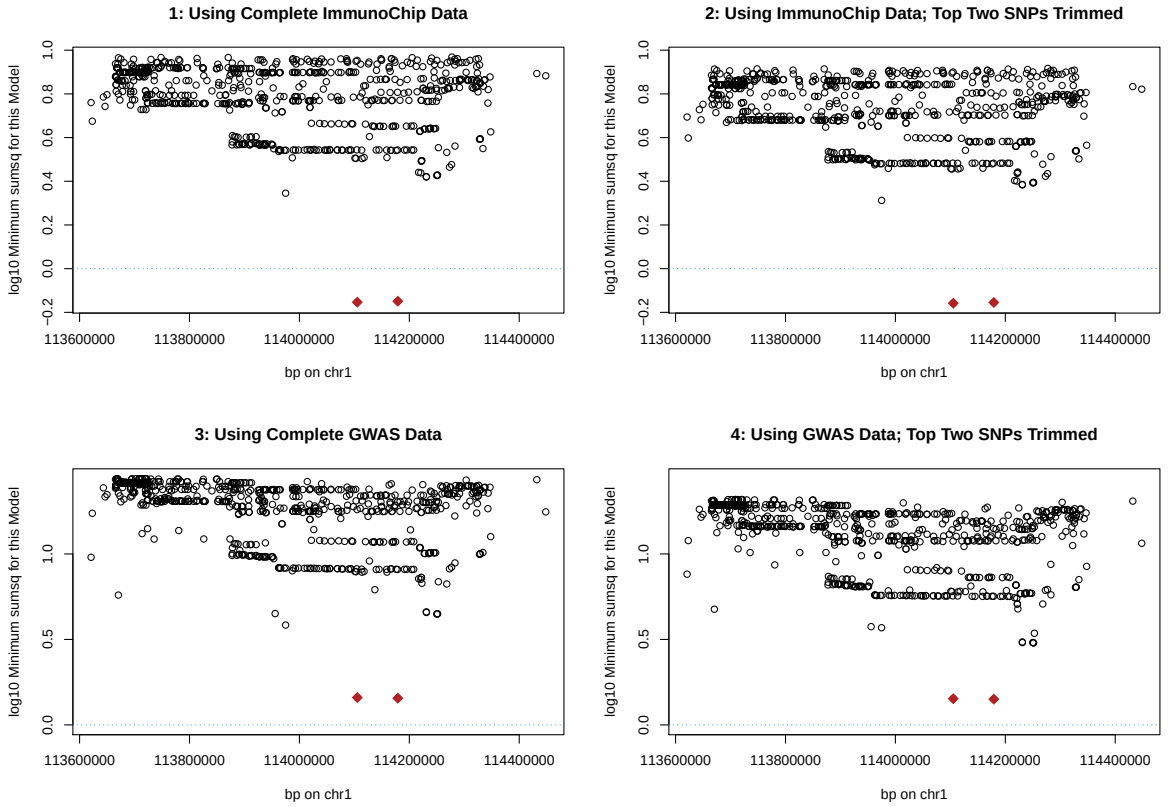


Figure 4.7 Results from testing single CV models for T1D association in the 1p13.2 region containing candidate causal gene *PTPN22*, using four different sources of observed Z Scores but the same reference dataset. The results for SNPs rs6679677 and rs2476601 are highlighted in red. The blue line denotes $\log_{10}(sumsq) = 0$, which corresponds to $sumsq = 1$. Notice that my conclusions are the same in all four analyses: even if I analyse a thinned dataset, with the Z Scores for rs6679677 and rs24766 removed, I still conclude that one of these two SNPs is causal.

Since I have the full genotype data from the ImmunoChip analysis, I used the control samples from this for my reference dataset. I analysed all SNPs contained in this dataset as potential causal SNPs, regardless of whether or not they had a corresponding observed Z Score. The resulting minimum *sumsq* obtained for each SNP are shown in Figure 4.7.

For each set of Z Scores, my conclusions are the same; even with rs6679677 and rs2476601 removed, the best models, with $sumsq \approx 0.7$ in the ImmunoChip analyses and ≈ 1.4 in the GWAS analyses, are those corresponding to these SNPs. These *sumsq* are well within the tolerance for belief that a model is causal; no other SNPs are plausible by comparison. Hence,

even with a much trimmed dataset, and the top SNPs themselves not genotyped, my algorithm comes to the same conclusion about the likely causal variants.

Note that removing rs6679677 and rs2476601 does not appear to greatly affect my *sumsq* for any model. However, moving from the ImmunoChip Z Scores to the GWAS Z Scores can have a dramatic effect.

This may be due to the trimming of SNPs; for instance, the third signal in the region, rs1230666, which is in high LD with rs6679677 and rs2476601 ($R^2 = 0.61$), is not present in the GWAS data. Another plausible explanation is that the GWAS data is partially imputed, which will result in the expected Z Score not matching the observed Z Score as well, even for the true model. However, it may also be caused by subtle changes between the control populations used in the analyses, resulting in my reference dataset not being quite the appropriate one for the GWAS; for more discussion of this effect, see Section 4.7.

4.5.2 1q24.3 Region Containing *FASLG*

Although *PTPN22* demonstrates that finding non-genotyped CVs is possible with my method, the strength of the T1D-association means it is not a typical region. In order to see how well my method performs upon non-genotyped SNPs in a region with more subtle effects, I will analyse the 1q24.3 region containing candidate causal gene *FASLG*.

Recall from Section 2.7.3 that this region has been associated to CEL. There is some evidence of T1D association, however, none of the T1D SNPs reach genome-wide significance. My colocalization analysis suggested that there is a shared CV for T1D and CEL, and furthermore using the colocalization results to perform fine mapping identified SNP rs78037977 as being this CV, with a posterior probability of 0.99. However, this SNP failed a QC step, and hence was removed from the CEL analysis. I wish to see whether my fine mapping method will also recover this SNP using only data from SNPs which have passed QC (those shown in the CEL plot of Figure 4.8).

Using the p-values obtained from ImmunoChip T1D and CEL studies, I first conducted a fine mapping analysis of the region for genotyped SNPs only. As in my analysis of the *PTPN22*

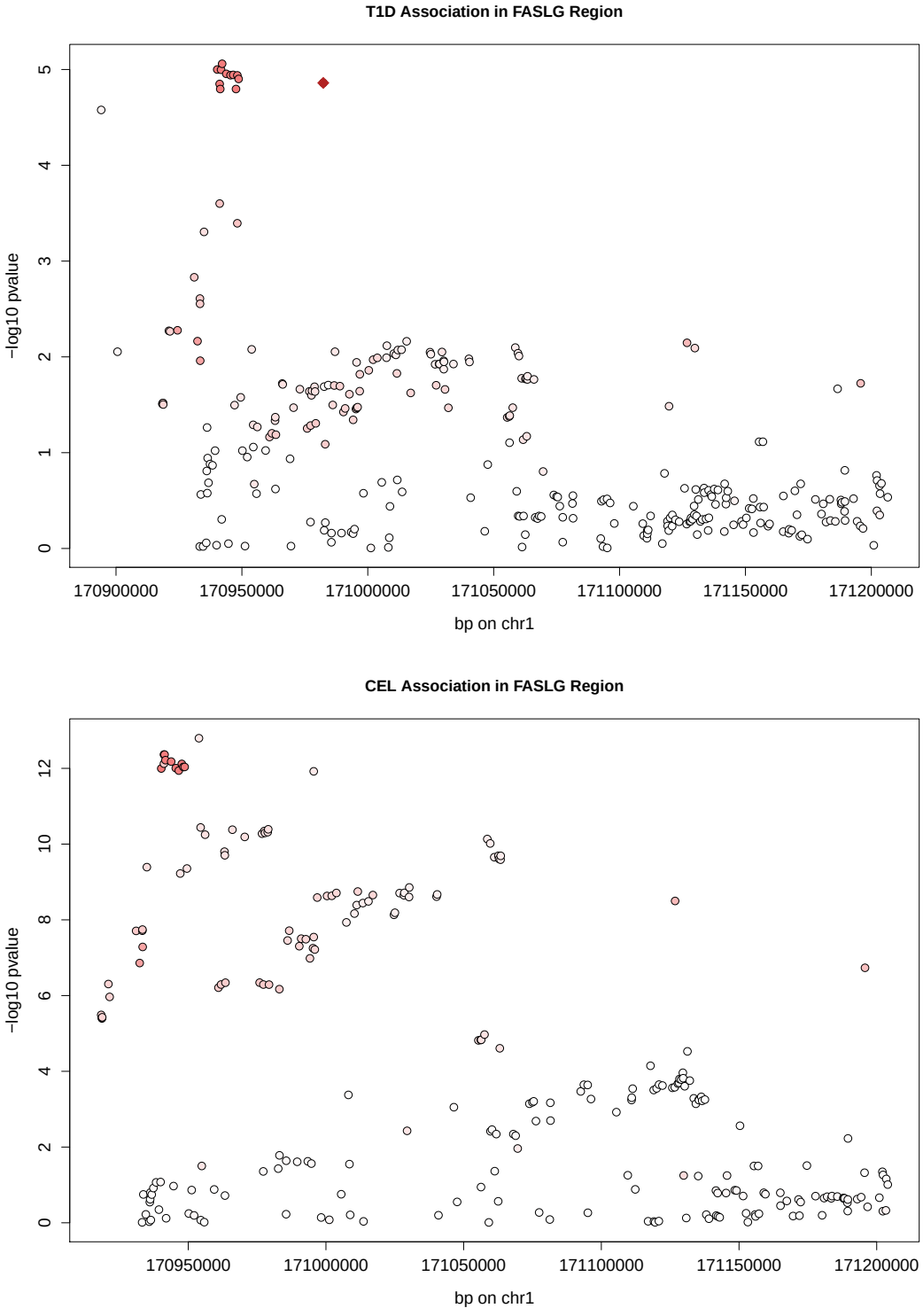


Figure 4.8 Manhattan plot of T1D and CEL Association in the 1q24.3 region containing candidate causal gene *FASLG*, using p-values from ImmunoChip studies. SNPs are coloured according to their LD with rs78037977 (designated by a red diamond), which I identified as a potential common causal SNP in section 2.7.3 but which was dropped from the published CEL ImmunoChip dataset due to QC conerns.

region above, I followed the method in [Bowes et al, 2015]. Recall that this technique assumes that there is a single causal variant, and requires that all SNPs have been assigned a p-value. The results are shown in Figure 4.9.

In the T1D fine mapping (Figure 4.9), although rs78037977 does come within the top set of 13 models with non-zero posterior probabilities, I cannot choose a single SNP. By contrast, from the CEL analysis (Figure 4.9), I can clearly see rs2157477 is the top performing causal model. Since it is not genotyped, it is impossible to come to any conclusions about rs78037977 using this method. However, it and rs2157477 have high LD, with $R^2 = 0.62$.

I therefore ran my method upon the *FASLG* region, using three sets of Z Scores:

1. An ImmunoChip analysis of T1D for UK samples (containing 291 SNPs)
2. The same T1D ImmunoChip analysis, but with the p-value for rs78037977 removed (containing 290 SNPs)
3. ImmunoChip analysis of CEL for UK samples; rs78037977 does not have a p-value (containing 244 SNPs)

using the genotypes from the ImmunoChip controls as my reference dataset. The resulting minimum *sumsq* obtained for the T1D analyses are shown in Figure 4.10, and for CEL in Figure 4.11.

From the CEL analysis, we clearly see that rs78037977 is the top performing SNP; with a *sumsq* of 1.72, it is the only model within the range we would expect to contain the true causal model. The T1D analysis is less clear cut; there are a number of SNPs, including rs78037977, which cluster with $sumsq < 1$, and though this set is likely to contain the true CV(s), it is not obvious how to distinguish between them. Since the T1D association in this region is relatively small (minimum observed p-value = 8.71×10^{-6}, maximum observed Z Score = 4.45), it is possible that some models with low association will appear to perform well by random chance.

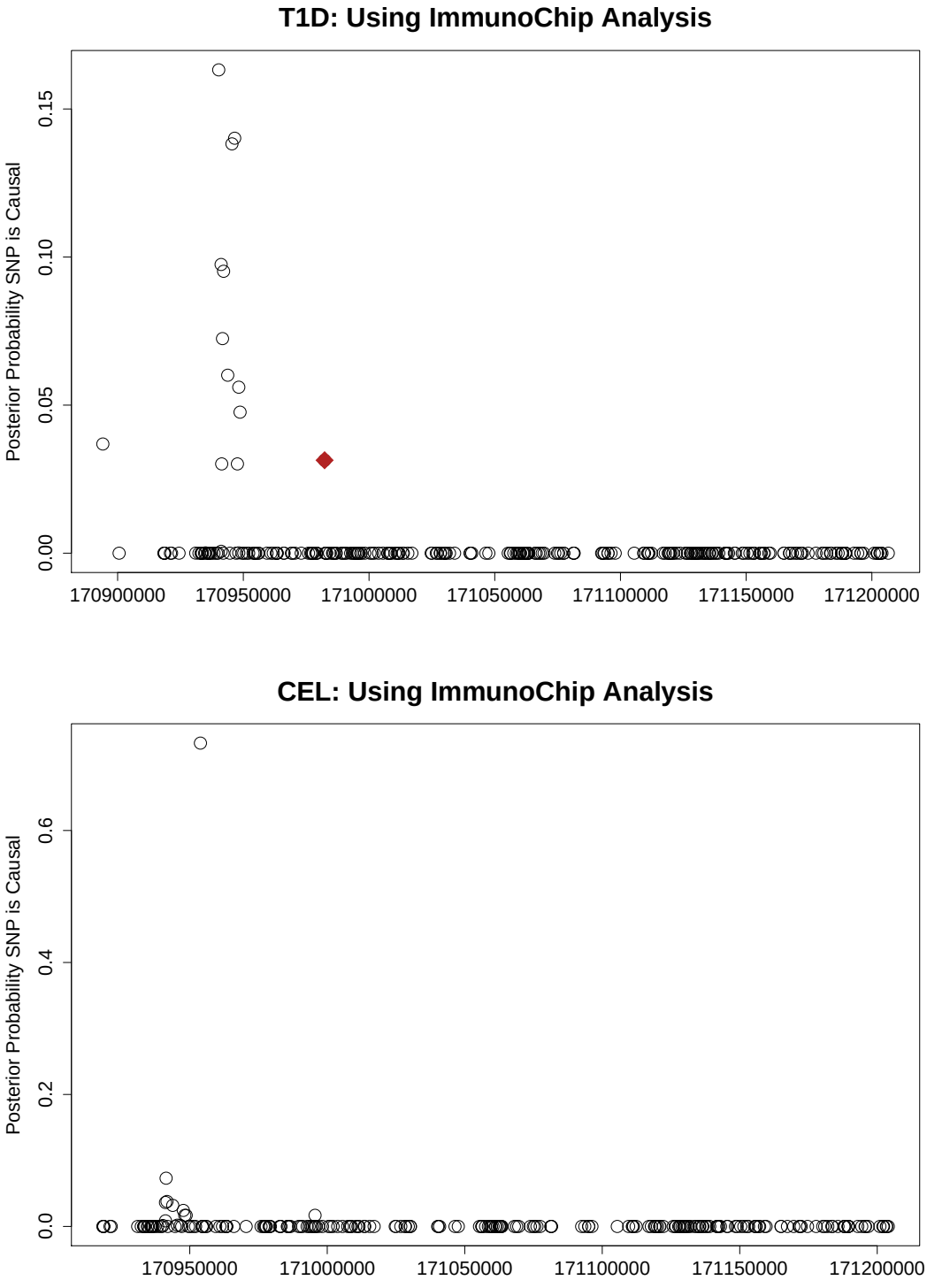


Figure 4.9 Posterior probabilities of being causal, from a standard single CV summary statistic fine mapping of the *FASLG* region for T1D and CEL, using Wakefield's approximation to compute Bayes Factors. The SNP highlighted in red is rs78037977, which I identified as a potential common causal SNP in section 2.7.3 but which was dropped from the published ImmunoChip CEL dataset due to QC concerns. The top SNP in the CEL analysis is rs2157477.

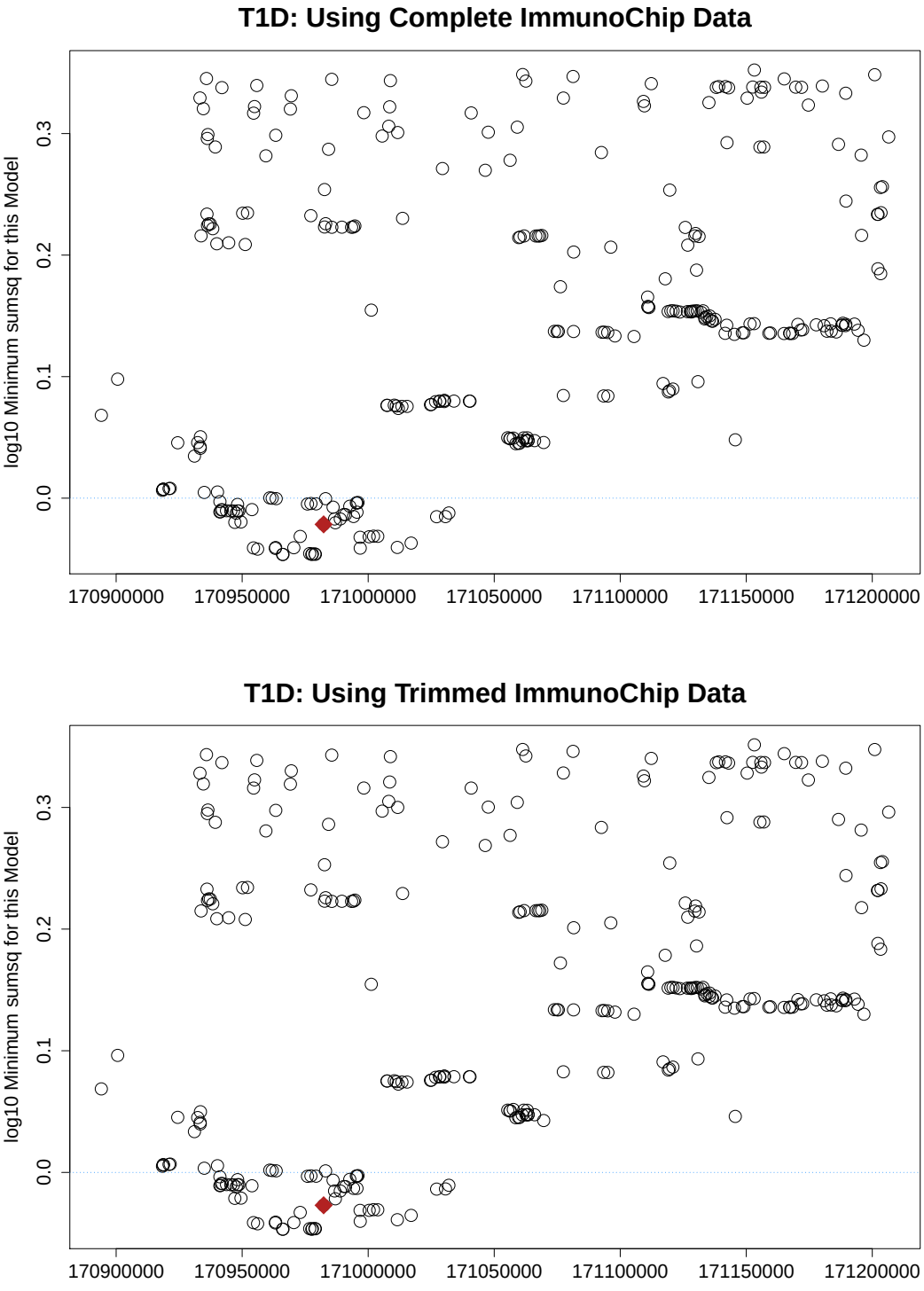


Figure 4.10 Results from testing single CV models for T1D association in the 1q24.3 region containing candidate causal gene *FASLG*. The results for SNP rs78037977 are highlighted in red. The blue line denotes $\log_{10}(sumsq) = 0$, which corresponds to $sumsq = 1$.

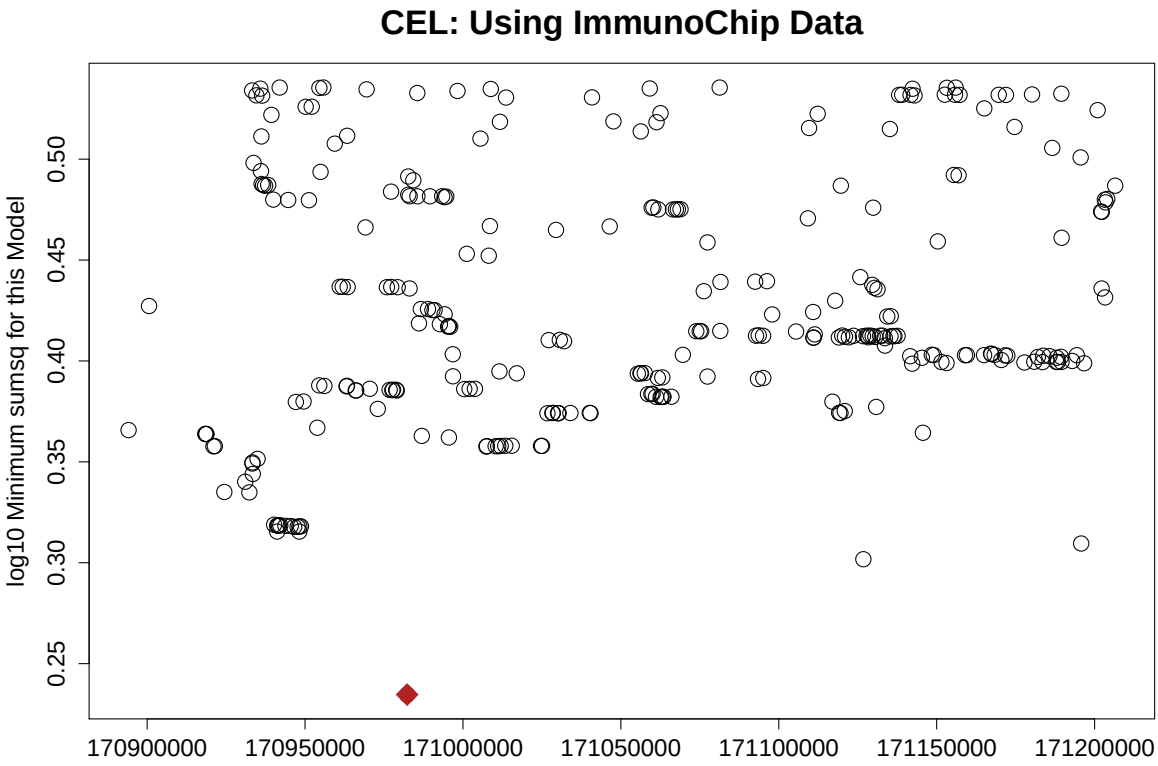


Figure 4.11 Results from testing single CV models for CEL association in the 1q24.3 region containing candidate causal gene *FASLG*. The result for SNP rs78037977 (not in the summary data, but captured through LD) is highlighted in red.

4.6 Comparison of Top Models for a Complex Region

The 10p15.1 region containing candidate causal gene *IL2RA* has been associated with multiple autoimmune diseases. In ImmunoBase, it is associated with the diseases: Autoimmune Thyroid Disease (ATD); Crohn's Disease (CD); Juvenile Idiopathic Arthritis (JIA); MS; RA; T1D and Ulcerative Colitis (UC). I analysed this region in Chapter 2, and found strong evidence of association with both T1D and MS; however, I also found evidence that this association is caused by distinct causal variants. In the Bayesian analysis (which assumes at most one causal variant per disease), I obtained a posterior probability of 1.00 for $\mathbb{H}_3$ (there is a SNP associated with trait 1, and a different SNP associated with trait 2), while in the proportional analysis, I obtained a posterior predictive p-value of 5.04×10^{-9} against the null hypothesis of proportionality. I also found evidence for multiple SNP association with T1D from the conditional Bayesian analysis. Figure 4.12 shows Manhattan plots of association in this region with the seven autoimmune diseases listed above; from this, the pattern of association across the diseases is heterogeneous.

A recent fine mapping analysis of this region [Wallace et al, 2015], using Bayesian model comparison with a stochastic search applied to full genotyping data, identified a four causal variant model for T1D and two competing causal models for MS, one containing a single SNP and one containing two causal SNPs. It divides potential causal SNPs up into six LD blocks (see Figure 4.13), labelled Group A, ..., Group F. The patterns of potential causal variants are very different between the two diseases; while MS appears to be associated with SNPs in Groups A and D, or in group B alone, T1D appears to be associated with SNPs in Groups A, C, E and F.

The two causal variant model for MS (A and D) is notable since neither of the SNPs it contains comes top on an analysis assuming a single causal variant (in which B alone is selected); hence, a conventional stepwise approach to fine mapping allowing for multiple causal variants would neglect this model. Nonetheless, a haplotype analysis of the regional association

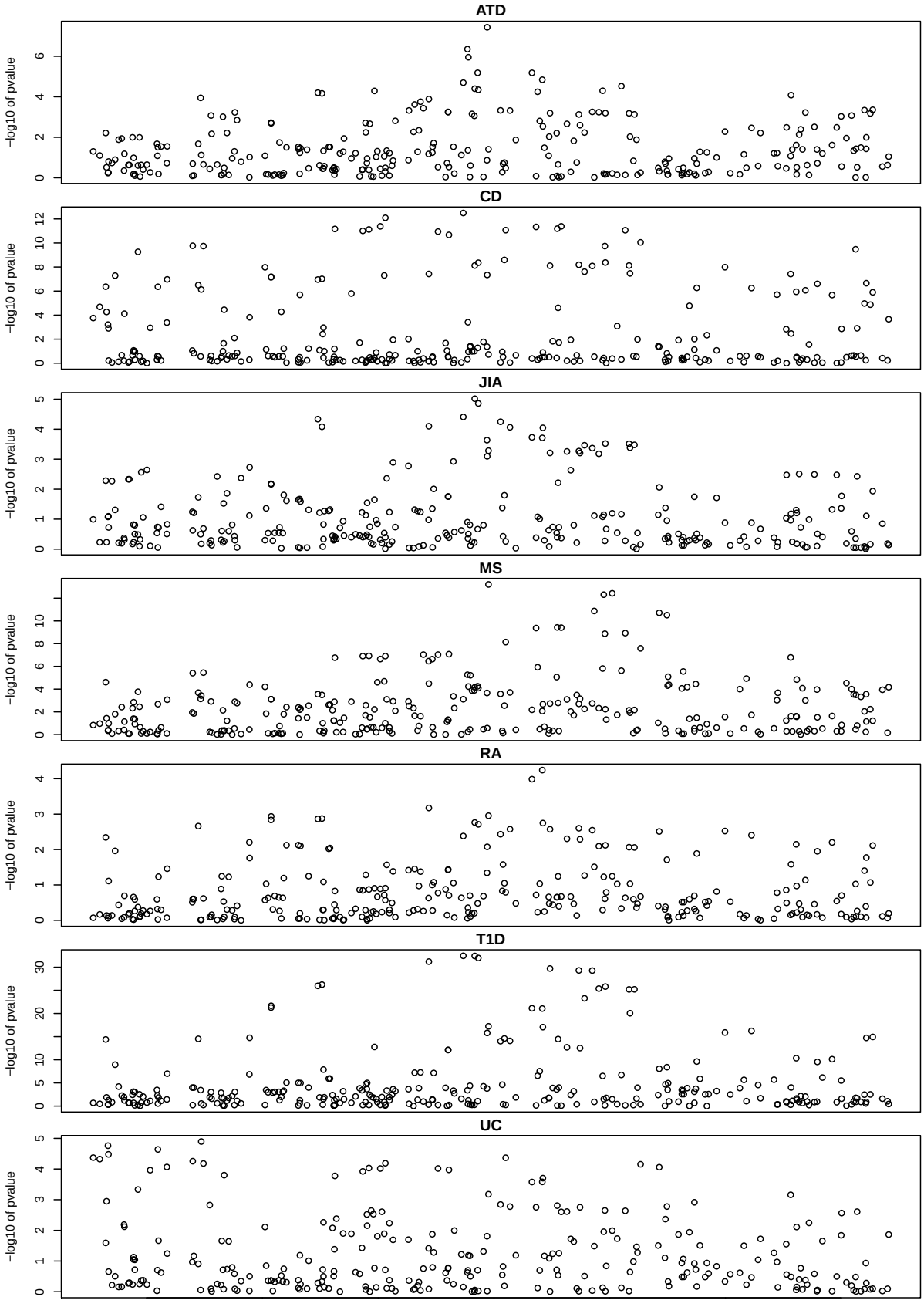


Figure 4.12 Manhattan plots of association with seven autoimmune diseases in the 10p15.1 region containing candidate causal gene *IL2RA*.

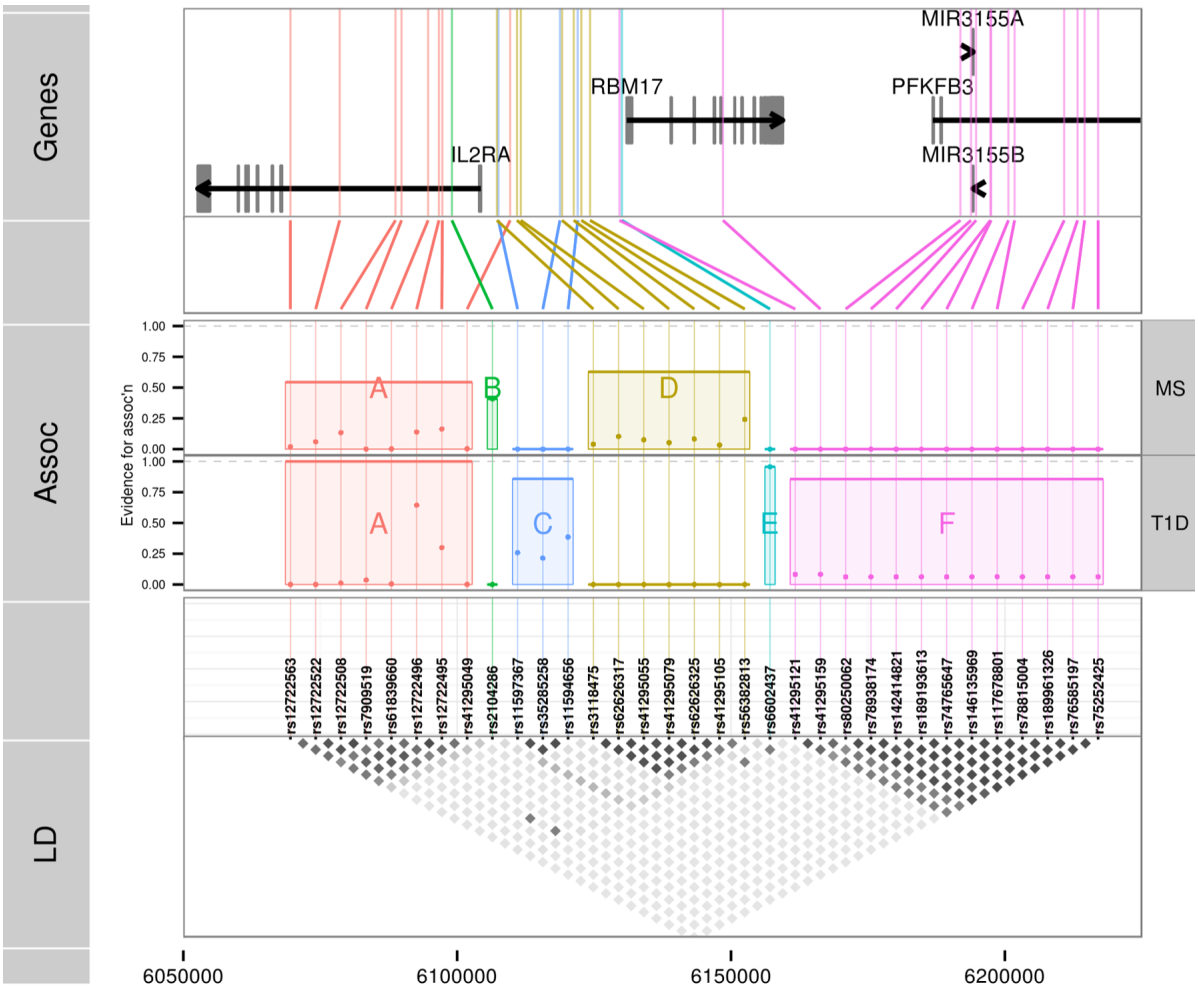


Figure 4.13 The locations of MS and T1D associated variants within the 10p15.1 region containing candidate causal gene *IL2RA*. Variants seem to fall into six groups, labelled Group A, ..., Group F; R^2 between these SNPs is shown in the lowest of the figures. Figure is taken from [Wallace et al, 2015].

(see Figure 4.14) demonstrates that, while the two SNP model is preferred, rs2104286 (group B) is selected by a univariate model, since it tags the common disease susceptible haplotypes formed by A and D.

Given that this fine mapping analysis suggests an overlap between causal variants for MS and T1D not evident from univariate, or stepwise, analyses, it suggests multi-SNP causal models might perform well when fitted to the other associated diseases, for which full genotype data is not available.

As discussed in Section 4.4.3, when the potential causal SNPs within a region have been

rs12722496	rs56382813	rs2104286	Fq (%)	OR	CI–95%	p
A	G	T	69.135	1.000	–	–
A	**A**	**C**	15.053	0.804	0.769–0.840	$< 2 \times 10^{-16}$
G	G	**C**	9.996	0.823	0.781–0.868	4.66×10^{-13}
A	**A**	T	5.594	0.854	0.797–0.915	8.06×10^{-06}

doi:10.1371/journal.pgen.1005272.t004

Figure 4.14 Results from a haplotype analysis of MS-associated SNPs in the *IL2RA* region. Two models perform well for MS, tagged by {rs2104286} (group B) and {rs12722496, rs56382813} (groups A and D respectivly). This demonstrates that, while the two SNP model is preferred, rs2104286 is selected by a univariate model, since it tags the common disease susceptible haplotypes. Figure is taken from [Wallace et al, 2015].

narrowed down to a small set, my fine mapping approach, with a high N_γ, can be deployed in order to distinguish between the fit of models containing subsets of these SNPs. After discussion with a collaborator (Linda Wicker, an expert in *IL2RA*), I considered the following SNPs as potential causal variants, selected from the MS and T1D fine mapping paper with additional "top SNPs" in the region for other diseases, as found in ImmunoBase:

- rs4147359 (associated with UC)
- rs10795791 (most associated SNP in region for RA)
- rs3118470 (most associated SNP in region for Alopecia Areata)
- rs706779 (most associated SNP in region for Vitiligo and Thyroiditis)
- rs41295055 (Group D)
- rs11594656 (Group C)
- rs2104286 (Group B)
- rs61839660 (Group A)
- rs6602437 (Group E)

I considered models containing one, two, three, four or five of these SNPs. I did not consider larger causal models, since with six or more causal SNPs, the search space of $\boldsymbol{\gamma}$ becomes arduously large. In order to cut down upon the necessary computation, I also did not consider

models containing two or more of rs4147359, rs10795791 and rs3118470; these SNPs are in high LD (pairwise $R^2 \sim 0.7$) and hence little information would be added by these models. I do not have p-values for rs3118470; it was not genotyped in all the GWAS whose results I analysed. However, following the demonstration in Section 4.5 above that my method is able to detect causal models containing non-genotyped SNPs, I left rs3118470 in the analysis.

In total, I considered 233 causal models. These I analysed for fit against the Z Scores obtained from UK-only ImmunoChip studies for the seven diseases listed above; I also tested them against the Z Scores obtained from international ImmunoChip studies (with all subjects having self-reported European ancestry) for MS and RA. I sampled 1000 $\boldsymbol{\gamma}$ at a time until N_γ was large enough that the 5-CV models converged to within 0.1 of their final minimum value.

I first explored the results for MS, using both the UK and the international data. Figure 4.15 gives heatmaps of the optimised γ for each of the 233 models, ordered by the resulting minimum *sumsq*. The 2-SNP model, {rs12722496, rs56382813}, identified in [Wallace et al, 2015] is tagged by the SNPs {rs61839660, rs41295055} in my analysis. The 1 SNP model identified in [Wallace et al, 2015] is {rs2104286}, which is included in my analysis.

Since all the models considered in the MS UK analysis have *sumsq* between 0.678 and 1.82, the vast majority are within the bounds where we would expect the true causal model to fall. The variation due to the inherent randomness of the observed Z Score is much greater than the difference between many of the models. My choice of N_γ resulted in convergence within only 0.1 of the best possible *sumsq* for a model containing five causal variants. Hence, on the results of this analysis, it is not possible to come to strong conclusions about our preference for a specific model. However, by viewing the heatmap of SNP effect sizes, arranged by goodness of fit of model, some patterns emerge.

The top models appear to contain rs41295055, with a weaker secondary signal in rs11594656 or rs61839660; this effect is particularly strong in the international analysis. rs41295055 is the SNP with most support for being individually causal; by contrast, rs2104286 does not appear to adequately explain the pattern of association across all SNPs in this region.

Note that models perform differently between the analyses; for instance, model

{rs61839660, rs41295055}, corresponding to the 2-SNP model identified in [Wallace et al, 2015], emerges more strongly in the international analysis. This may be due to the increase in power achieved by analysing more samples, or an effect not found in the UK population. However, it may also be due to the use of an inappropriate UK-only reference dataset; this possibility is examined in more depth in Section 4.7.

For T1D, if I fit the preferred model from [Wallace et al, 2015] ({Group A, Group C, Group E, Group F}, tagged in my analysis by SNPs {rs61839660, rs11594656, rs6602437, rs41295121 rs41295121 is from Group F), I get a *sumsq* of 1.61. However, the Group F signal appears to be T1D-specific, and not associated with the other six diseases; by contrast, the effects of Groups A-E appear to be cross-disease. Hence, the SNPs selected above are taken from groups A-E only, and no models were fitted including Group F. By contrast with the UK analyses of the other six diseases, each of which contained many models with *sumsq* consistent with being the true causal variants, very few of the minimum *sumsq* obtained for the T1D analysis were consistent with being the true model (Figure 4.16). This confirms that the signal found within Group F is required to explain the T1D association in the region.

Nonetheless, all other diseases achieved *sumsq* below or close to 1 in the UK analysis using only SNPs within Groups A-E, suggesting that the signal within Group F of SNPs is unique to T1D (Figures 4.17, 4.18 and 4.19 give results for the other traits).

From these heatmaps, we see some patterns. The results from the CD analysis are consistent with a single SNP, rs6602437, being causal; the presence of rs6602437 within the model is a strong predictor of whether that model will have $sumsq < 1$. No other disease displays this behaviour; indeed, in UC, where we would expect to see very similar behaviour to CD due to the similarities in the disease processes, it is not possible to come to any strong conclusions about preferred models. However, both ATD and JIA show evidence of being associated with the same 2-CV model: {rs706779, rs6602437}; all top models in these diseases contain these two SNPs, with rs706779 consistently having the larger effect size.

Figure 4.15 Heatmap of optimised $\boldsymbol{\gamma}$ from the analysis of UK MS and international MS summary data in the *IL2RA* region. The 233 models analysed have been sorted by their goodness of fit (that is, the smallness of the resulting minimum *sumsq*, which is given in the blue column to the left of the plot) so that the best fitting models are at the top. Each row of the heatmap then shows the value of $\boldsymbol{\gamma}$ at which the model obtained this minimum *sumsq*, with SNPs not present in the model having their corresponding γ set to 0.

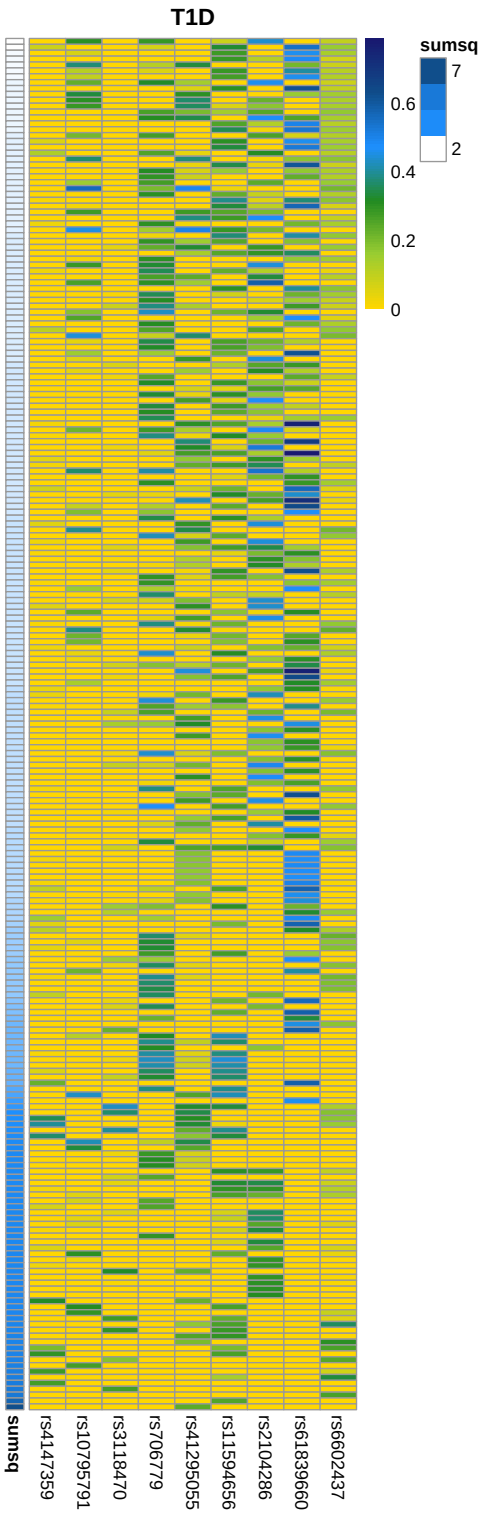


Figure 4.16 Heatmap of optimised $\boldsymbol{\gamma}$ from the analysis of UK T1D summary data in the *IL2RA* region. The 233 models analysed have been sorted by their goodness of fit (that is, the smallness of the resulting minimum *sumsq*, which is given in the blue column to the left of the plot) so that the best fitting models are at the top. Each row of the heatmap then shows the value of $\boldsymbol{\gamma}$ at which the model obtained this minimum *sumsq*, with SNPs not present in the model having their corresponding γ set to 0.

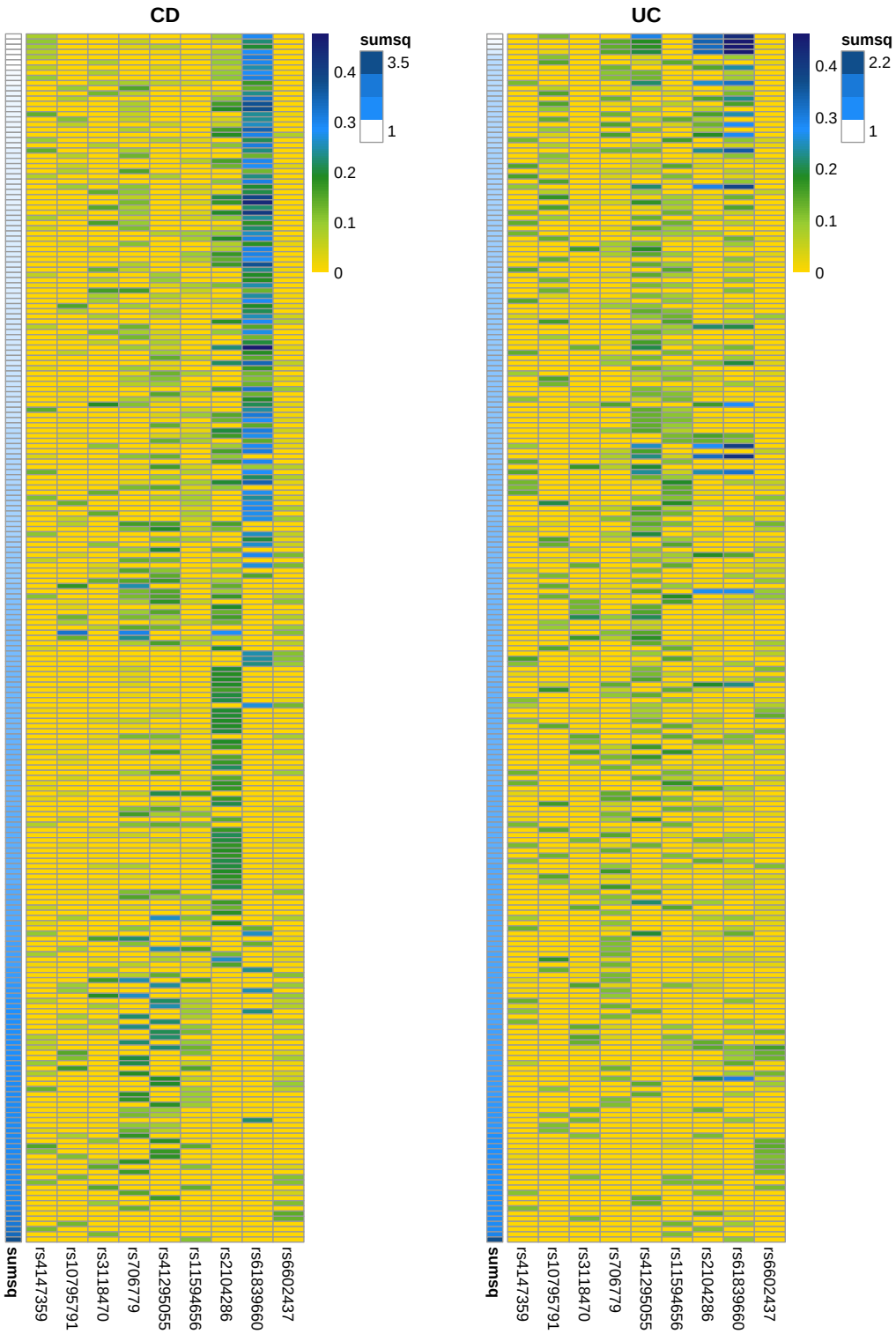


Figure 4.17 Heatmap of optimised $\boldsymbol{\gamma}$ from the analysis of UK CD and UK UC summary data in the *IL2RA* region. The 233 models analysed have been sorted by their goodness of fit (that is, the smallness of the resulting minimum *sumsq*, which is given in the blue column to the left of the plot) so that the best fitting models are at the top. Each row of the heatmap then shows the value of $\boldsymbol{\gamma}$ at which the model obtained this minimum *sumsq*, with SNPs not present in the model having their corresponding γ set to 0.

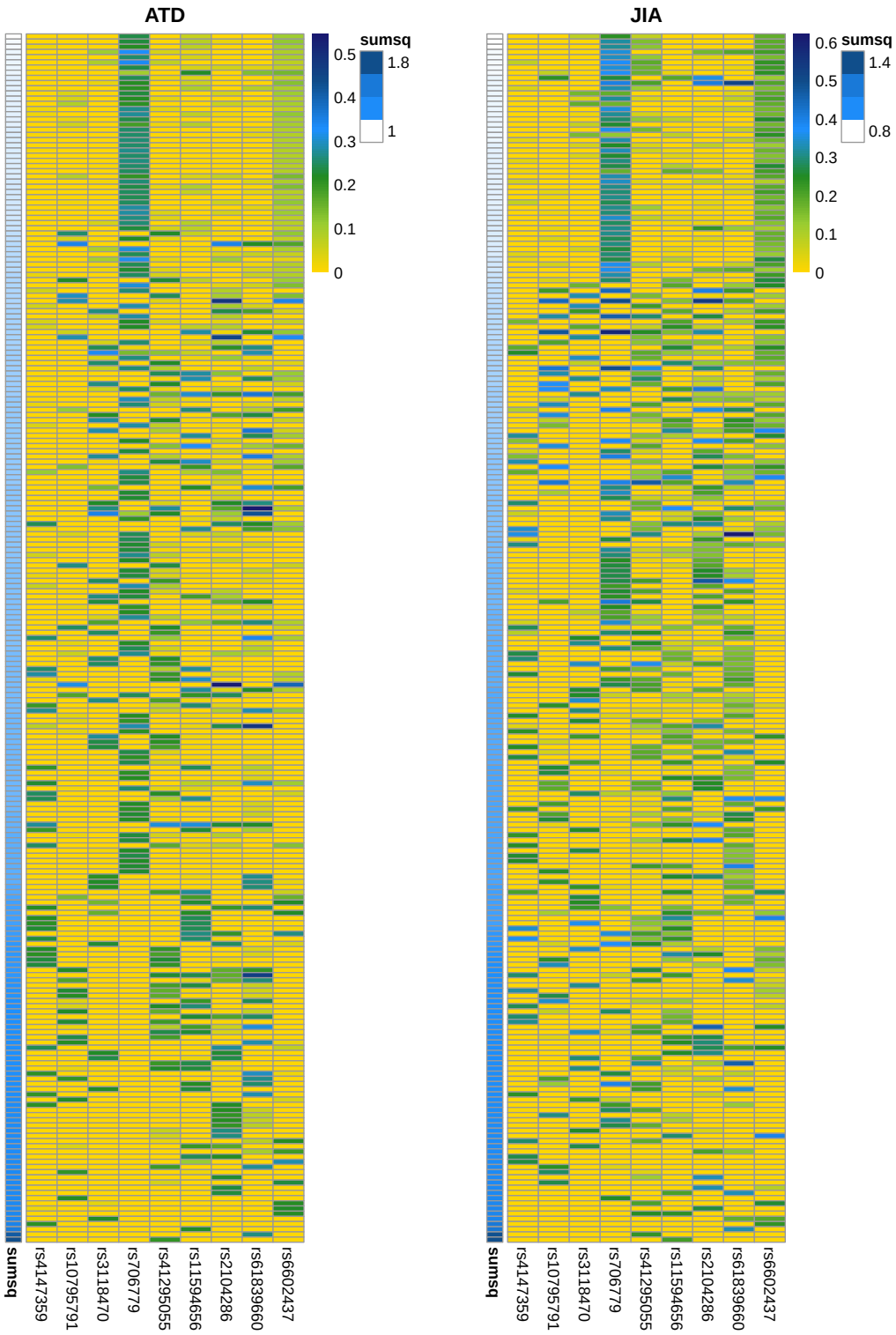


Figure 4.18 Heatmap of optimised $\boldsymbol{\gamma}$ from the analysis of UK ATD and UK JIA summary data in the *IL2RA* region. The 233 models analysed have been sorted by their goodness of fit (that is, the smallness of the resulting minimum *sumsq*, which is given in the blue column to the left of the plot) so that the best fitting models are at the top. Each row of the heatmap then shows the value of $\boldsymbol{\gamma}$ at which the model obtained this minimum *sumsq*, with SNPs not present in the model having their corresponding γ set to 0.

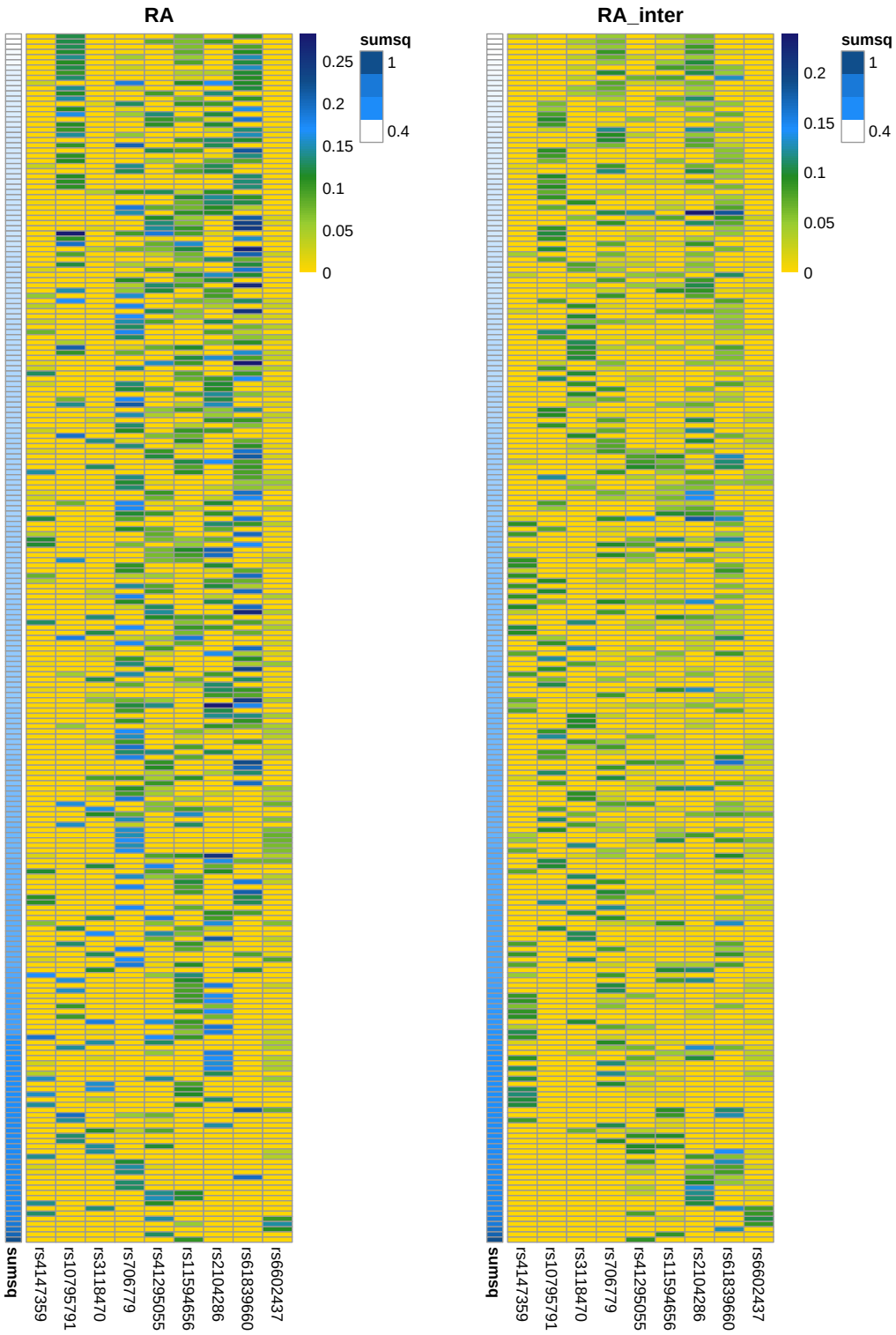


Figure 4.19 Heatmap of optimised $\boldsymbol{\gamma}$ from the analysis of UK RA and International RA summary data in the *IL2RA* region. The 233 models analysed have been sorted by their goodness of fit (that is, the smallness of the resulting minimum *sumsq*, which is given in the blue column to the left of the plot) so that the best fitting models are at the top. Each row of the heatmap then shows the value of $\boldsymbol{\gamma}$ at which the model obtained this minimum *sumsq*, with SNPs not present in the model having their corresponding γ set to 0.

4.7 The Impact of the Reference Dataset

The work done so far assumes that the reference dataset, which is used to estimate the between-SNP relations $\mathbb{P}(G_i^X = x \cap G_i^{\mathbf{W}} = \mathbf{w})$, and as input to the LDAK algorithm to estimate weightings, is a correct reflection of the correlations between SNPs in the control samples analysed in the original GWAS. For the majority of work presented in this thesis, this assumption is valid; most of my p-values come from ImmunoChip studies curated in ImmunoBase, and the control data I have used as a reference is the common control data used by the ImmunoChip consortium in their analyses. However, my method aims to enable fine mapping of GWAS output in the absence of genotype data. Hence, in practice, the reference dataset used will be at most population-matched to the true control data. This is a requirement for all such fine mapping techniques using summary data; both methods discussed in Section 1.5, PAINTOR and Caviar, suggest using 1000 Genomes data for this purpose. However, it is not obvious how much impact the choice of reference dataset has upon the results obtained.

The frequencies of, and correlations between, SNPs vary greatly across populations; their presence or absence can be used in genetic ancestry testing. Hence, in order for my reference dataset to give an accurate reflection of the SNP correlation structure found in the control sample of the GWAS whose summary results I am fine mapping, it must come from the same ethnic population. In the case of the analysis presented in this thesis, that population is individuals of European descent in the UK.

Recall in Section 4.5, I analysed the 1p13.2 region containing candidate causal gene *PTPN22* using p-values obtained from a T1D UK-only ImmunoChip analysis, and from a much downsampled T1D GWAS analysis where, while all the control samples were of self-reported European ancestry, some were taken from a US study. There was a substantial difference between the results. Although this may be explainable by the SNPs downsampled including one in high LD with the likely causal SNPs, it may also be the effect of analysing the GWAS output using an inappropriate reference dataset.

Similarly, in Section 4.6, I analysed the 10p15.1 region containing candidate causal gene

IL2RA. Although the main body of my analysis was upon the p-values obtained from UK-only ImmunoChip analyses, and I used UK-only reference datasets, for MS and RA I also analysed the p-values from the ImmunoChip analysis of an international (although of self-reported European ancestry) cohort. Comparison of the results obtained for MS can be found in Figure 4.15 and for RA can be found in Figure 4.19. The MS results show similar patterns, with the two-SNP model identified in a previous analysis of the region having more evidence of association; any discrepancies could be the result of an increase in the power to detect associated signals. However, there are no clear patterns of association in the RA results, and our preferred models change greatly between the UK-only and international analysis. This does not appear to be the result of a power difference (a concentration of the same pattern in the larger sample), and suggests instead that this may reflect differences between the GWAS and reference datasets.

The smaller the number of samples in our reference dataset, the greater the effect of sampling variation will be on estimates of SNP structure computed from them. My method requires computing the $(m+1)$-way correlation between each SNP and the m causal SNPs in $\mathbf{W}$. Thus, the larger m, the greater number of samples required to obtain an accurate estimate. 1000 Genomes, the reference dataset often suggested by summary statistic fine mapping papers, contains only 91 British samples; this is surely too small to accurately estimate the relationships between any but the most common SNPs. One solution would be to integrate 1000 Genomes data from across Europe, taking us up to 503 samples, and hope that the reduction in sampling variance outweighs any bias due to different population structures.

In order to assess the potential effect of choice of reference dataset upon the results presented in this chapter, I reran them using five different reference datasets, to see whether my conclusions might be changed. In order to disentangle the effects of down sampling and different populations, I used the following reference datasets:

1. All UK ImmunoChip controls, as discussed in Section 2.6

2. UK ImmunoChip controls, downsampled to 503 samples

3. UK ImmunoChip controls, downsampled to 91 samples

4. EUR cohort from 1000 Genomes, containing 503 samples

5. GBR cohort from 1000 Genomes, containing 91 samples

4.7.1 Analysis of the *PTPN22* Region

Recall that the 1p13.2 region containing candidate causal gene *PTPN22* has strong evidence of being a single-CV region for T1D, with two disease associated SNPs (rs6679677 and rs2476601 with r^2 of 0.995) outperforming all other models in my fine mapping analysis (Section 4.5). Figure 4.20 shows the results from using my five reference datasets to analyse all one-CV models in the region. Each line gives the minimum *sumsq* obtained for one model under the different reference datasets. The bold blue line corresponds to rs6679677, while the bold red line corresponds to rs2476601. Regardless of reference datasets chosen, we conclude that these are the top pair of SNPs. In the ImmunoChip downsampled to 91 analysis, however, they obtain a minimum *sumsq* of > 2.4, which is too high to be consistent with being the true causal model.

The bold green line on Figure 4.20 corresponds to SNP rs1230666. This is in LD with the top SNPs ($R^2 = 0.61$ with each of them), and on a Manhattan plot of the region (Figure 4.5) it is clearly the third most associated SNP (p-value$= 1.64 \times 10^{-54}$ in the ImmunoChip study), with a substantial gap between it and the next SNPs. When a standard single CV summary statistic fine mapping analysis is performed upon this region with rs6679677 and rs2476601 excluded (figure 4.6), rs1230666 is clearly the preferred SNP, with a posterior probability of being causal = 1. In my fine mapping analysis, using the correct reference dataset, rs1230666 is the third preferred model (although its minimum *sumsq* of 2.21 is not consistent with being the true causal model). By contrast, in the analyses using 1000 Genomes data, a set of SNPs with ($r^2 \sim 0.41$ with rs6679677 and rs2476601) outperform rs1230666.

This analysis shows that, even in a highly associated region with a clear causal SNP, changing the reference dataset can affect the resulting minimum *sumsq*, although not our

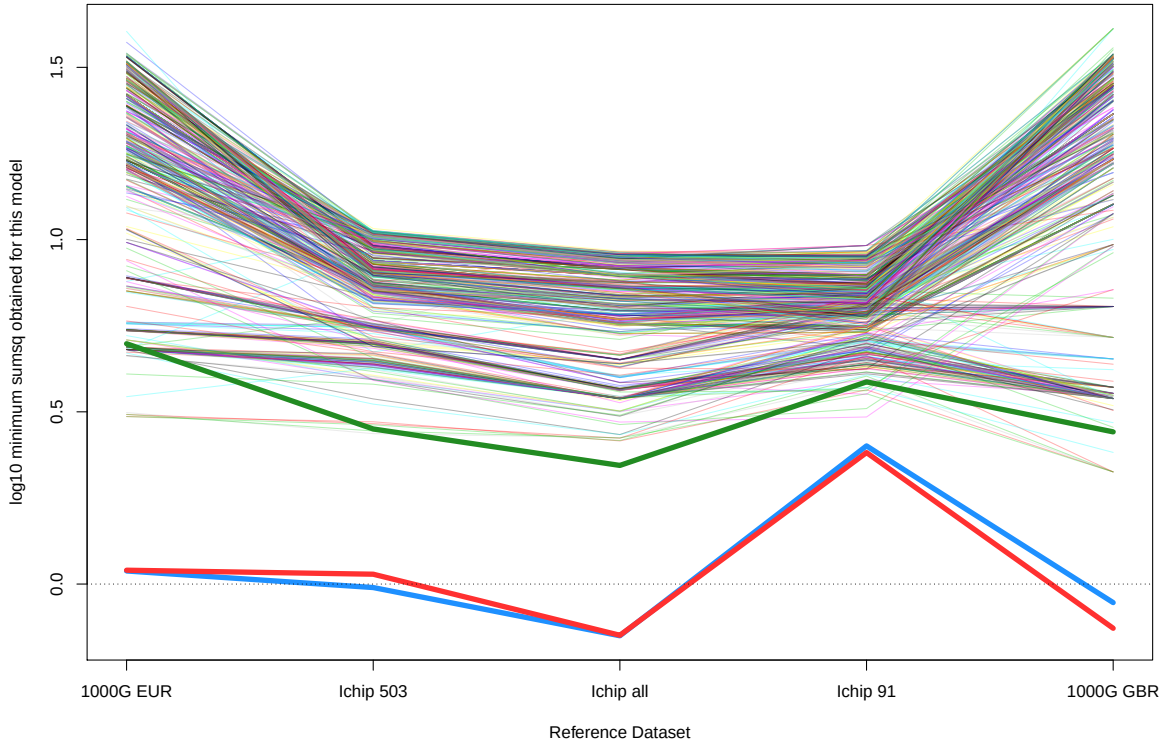


Figure 4.20 The impact of choice of reference dataset on single CV models for T1D association in the 1p13.2 region containing *PPTPN22*. Each line corresponds to the minimum *sumsq* obtained for each model when the five different reference datasets, represented by position on the horizontal axis, are used. The bold blue, red and green lines correspond to the results for rs6679677, rs2476601 and rs1230666 respectively, which are the three most strongly associated SNPs in the region. The dotted line denotes $\log_{10}(sumsq) = 0$, which corresponds to $sumsq = 1$.

conclusions about the best performing models. Looking at the models overall, while downsampling the reference dataset does have some effects upon the results in this region, the largest change occurs when we switch to a different population of samples.

4.7.2 Analysis of the *FASLG* Region

Recall that the 1q24.3 region containing candidate causal gene *FASLG* has a potential causal variant, rs78037977, which was not genotyped in the original CEL ImmunoChip analysis. However, my fine mapping approach, using the complete ImmunoChip reference dataset,

identifies this variant as performing well in the T1D analysis (Figure 4.10) and as the top model in the CEL analysis (Figure 4.11). Here, I investigate the effect of choice of reference dataset upon my conclusions for this region.

Figure 4.21 shows the T1D analysis, giving the minimum *sumsq* obtained for each model/ reference dataset pair. The bold red line corresponds to rs78037977, which performs well regardless of the reference dataset; for all analyses, its minimum *sumsq* is within the bounds where we would expect the *sumsq* from the true causal model to lie.

Looking over all models tested for T1D analysis, the choice of reference dataset appears to have affected their relative rankings much more than was the case for the *PTPN22* region. This is not surprising. The observed Z Score for this region is relatively low (it has a maximum value of 4.45; T1D association in this region is not genome-wide significant). While we would expect the true causal model to perform well upon this region, we would also expect some models with expected Z Scores clustered around the origin to do well by random chance. Which models have this spurious low *sumsq* will be heavily dependent upon their MAF and LD structure; our estimate of this varies with choice of reference dataset, and hence our well performing near-zero models will also vary. Consider for instance rs74844118 (r^2 of 0.00319 with rs78037977), which is denoted in Figure 4.21 by the bold green line. While it performed poorly in four of the analyses, it is the top model when the reference dataset is the ImmunoChip data downsampled to 91.

Figure 4.22 shows the CEL analysis, giving the minimum *sumsq* obtained for each model/ reference dataset pair. The bold red line corresponds to rs78037977. This SNP is ranked highly regardless of reference dataset, however, it is no longer the top model in the ImmunoChip downsampled to 91, 1000 Genomes EUR and 1000 Genomes GBR analyses; these are the reference datasets whose estimates are likely to be furthest from the true values. Looking over all models analysed, both downsampling and changing to the 1000 Genomes population change our conclusions, and the rankings, of many of the models. The CEL association in this region is GWAS significant (maximum Z Score = 7.38), and we see fewer potentially spurious top models than in the T1D analysis.

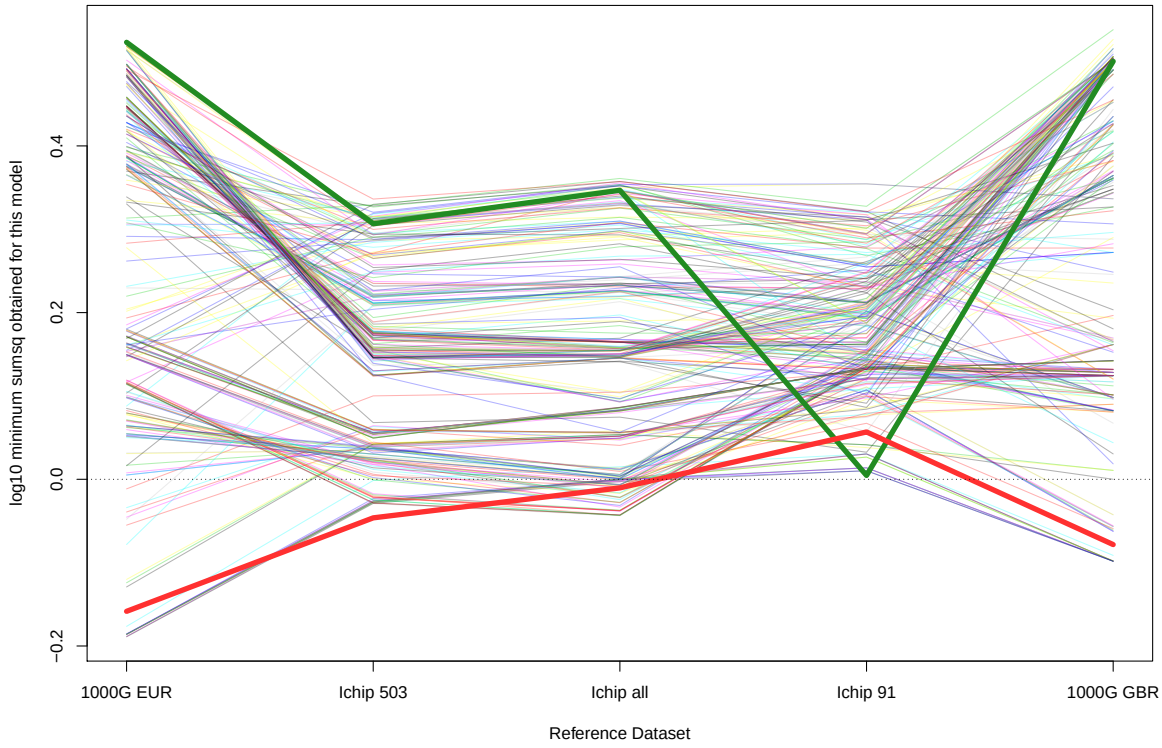


Figure 4.21 The impact of choice of reference dataset on single CV models for T1D association in the 1q24.3 region containing *FASLG*. Each line corresponds to the minimum *sumsq* obtained for each model when the five different reference datasets, represented by position on the horizontal axis, are used. The bold red line corresponds to the results for rs78037977. The dotted line denotes $\log_{10}(sumsq) = 0$, which corresponds to $sumsq = 1$.

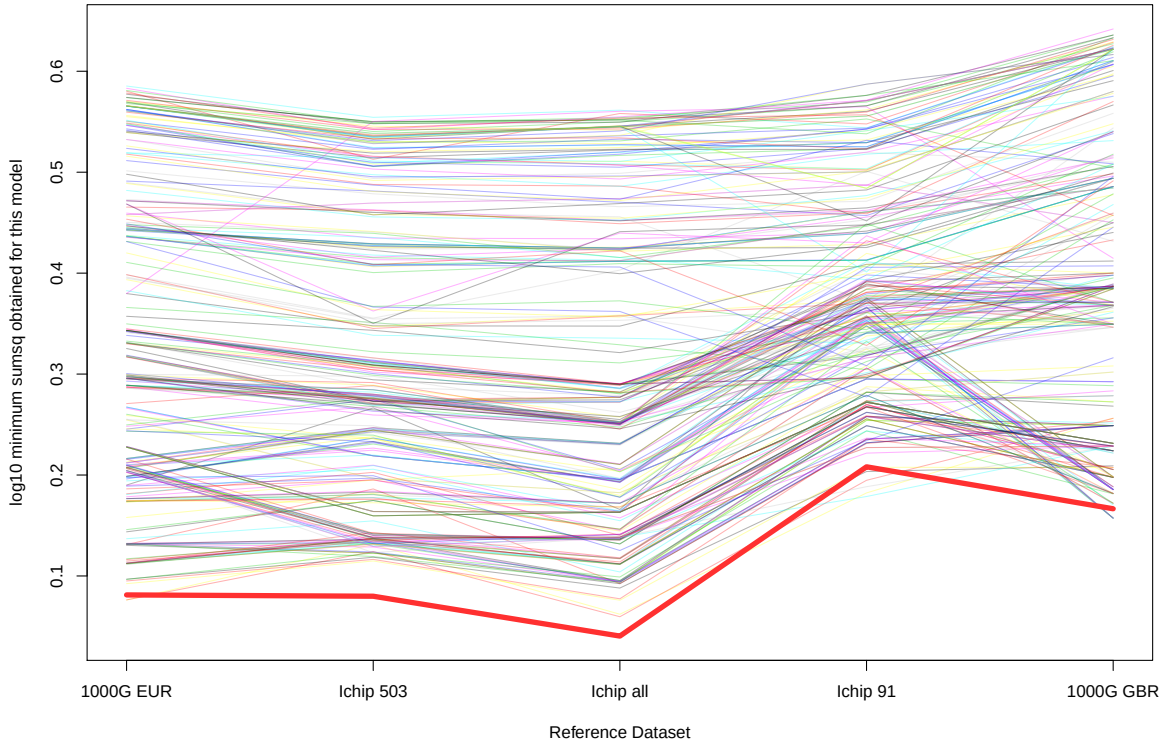


Figure 4.22 The impact of choice of reference dataset on single-CV models for CEL association in the 1q24.3 region containing *FASLG*. Each line corresponds to the minimum *sumsq* obtained for each model when the five different reference datasets, represented by position on the horizontal axis, are used. The bold red line corresponds to the results for rs78037977.

4.7.3 Analysis of the *IL2RA* Region

Recall that while the 10p15.1 region containing candidate causal gene *IL2RA* is associated with many autoimmune diseases the patterns of association are complex and vary between diseases. A fine mapping analysis of this region [Wallace et al, 2015] found a 4-CV model for T1D, whereas in MS there were two competing causal models, one with 1 CV ({rs2104286}) and one with 2 CV (tagged in my analysis by {rs61839660, rs41295055}). In Section 4.6 I compared models containing a subset of nine SNPs with evidence of being causal; the results of my T1D and MS analyses were consistent with [Wallace et al, 2015]. In this section, I investigated the effect of choice of reference dataset on all 1-CV and 2-CV models for T1D and MS association in this region.

Figure 4.23 shows the results of these analyses, giving the minimum *sumsq* obtained for each model/reference dataset pair. This shows that while changing the reference data does change the *sumsq* values for individual models, it does not appear to change our conclusions in this region. None of the models in the T1D analysis have $sumsq \sim 1$; from this, we conclude that this region is not consistent with having only 1 or 2 CVs, as found in [Wallace et al, 2015]. By contrast, in the MS analysis, a great many of the models fall within the *sumsq* bounds where we would expect to see the true causal model. The two models identified in [Wallace et al, 2015] for MS association are denoted by the bold yellow and red lines in Figure 4.23. Regardless of choice of reference dataset, while neither is ever the top performing model, both have *sumsq* consistent with being causal; the exact ordering of potentially causal models with $sumsq < 1$ is likely due to random chance.

However, although Figure 4.23 gives an overview of our inference for the region, the number of models being analysed makes it difficult to see the effect of changing the reference dataset on a specific choice of $\mathbf{W}$.

Figure 4.24 uses a heatmap to represent the *sumsq* for each of the five reference datasets in the T1D analysis. Each row and column of each heatmap corresponds to a SNP. SNPs have been ordered according to their minimum *sumsq* in the single causal variant models when analysed using the complete ImmunoChip reference dataset, starting from the bottom left

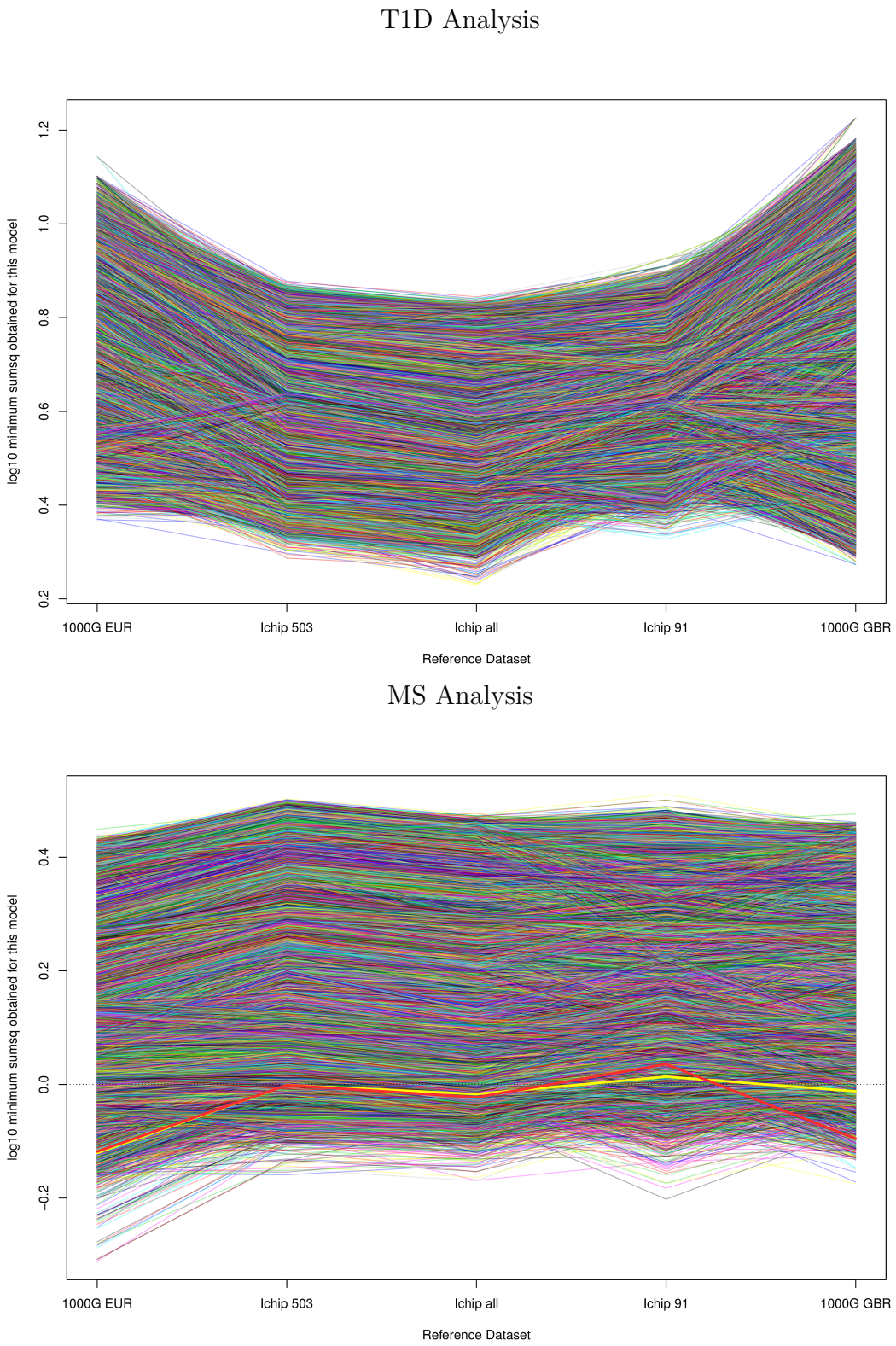


Figure 4.23 The impact of choice of reference dataset on T1D and MS 1-CV and 2-CV models in the 10p15.1 region containing *IL2RA*. Each line corresponds to the minimum *sumsq* obtained for each model when the five different reference datasets, represented by position on the horizontal axis, are used. In the MS plot, the bold yellow line corresponds to the results for {rs2104286}. The bold red line corresponds to the results for {rs61839660, rs41295055}. The dotted line denotes $\log_{10}(sumsq) = 0$, which corresponds to $sumsq = 1$.

hand corner. Each cell is coloured according to the minimum *sumsq* obtained at the model where **W** contains the two SNPs corresponding to the cell's row and column; the diagonals give the minimum *sumsq* obtained by the 1-CV model containing that SNP. Shading goes from blue (low *sumsq* for this region) to red (high *sumsq* for this region).

There are notable differences between each of the five heatmaps. However, the most striking effect is how much of a difference the choice of reference population makes. The trend is for *sumsq* computed using the correlation structures from 1000 Genomes reference datasets to be higher than those computed using ImmunoChip reference datasets. However, in the 1000 Genomes analysis, a subset of SNPs, corresponding to the thin blue lines visible upon the right hand heatmaps, form an exception. Despite not performing well in the original, ImmunoChip reference dataset, analysis, with the 1000 Genomes data, these SNPs have among the lowest *sumsq* in the region when included in **W**. This effect is consistent in both the EUR and GBR datasets. This demonstrates that choice of reference population can dramatically change the ranking of models in this analysis.

Figure 4.25 gives the same five heatmaps, but for the MS analysis. Here, the effect of choice of reference dataset is not so dramatic as it was for the T1D analysis. However, in all four of the smaller reference datasets, we do see a number of cases of SNPs which perform much better than in the original analysis. Consider for instance, rs41294917, corresponding to the blue line seen towards the top right of the ImmunoChip 503 heatmap. While **W** = rs41294917 has minimum $sumsq = 2.96$ when using the complete ImmunoChip genotypes for reference data, when this is downsampled to 503, the minimum *sumsq* obtained is 1.45. When the 1000 Genomes GBR reference data is used, rs41294917 has minimum $sumsq = 1.32$. As in the T1D case, such differences affect the relative ranking of the models. In addition, a change of *sumsq* from 2.96 to 1.32 is large enough that it moves the model from unlikely to be causal to the region where we would expect to see the true causal model. This highlights that changing the reference dataset may change our inferences about where the causal model is likely to be found.

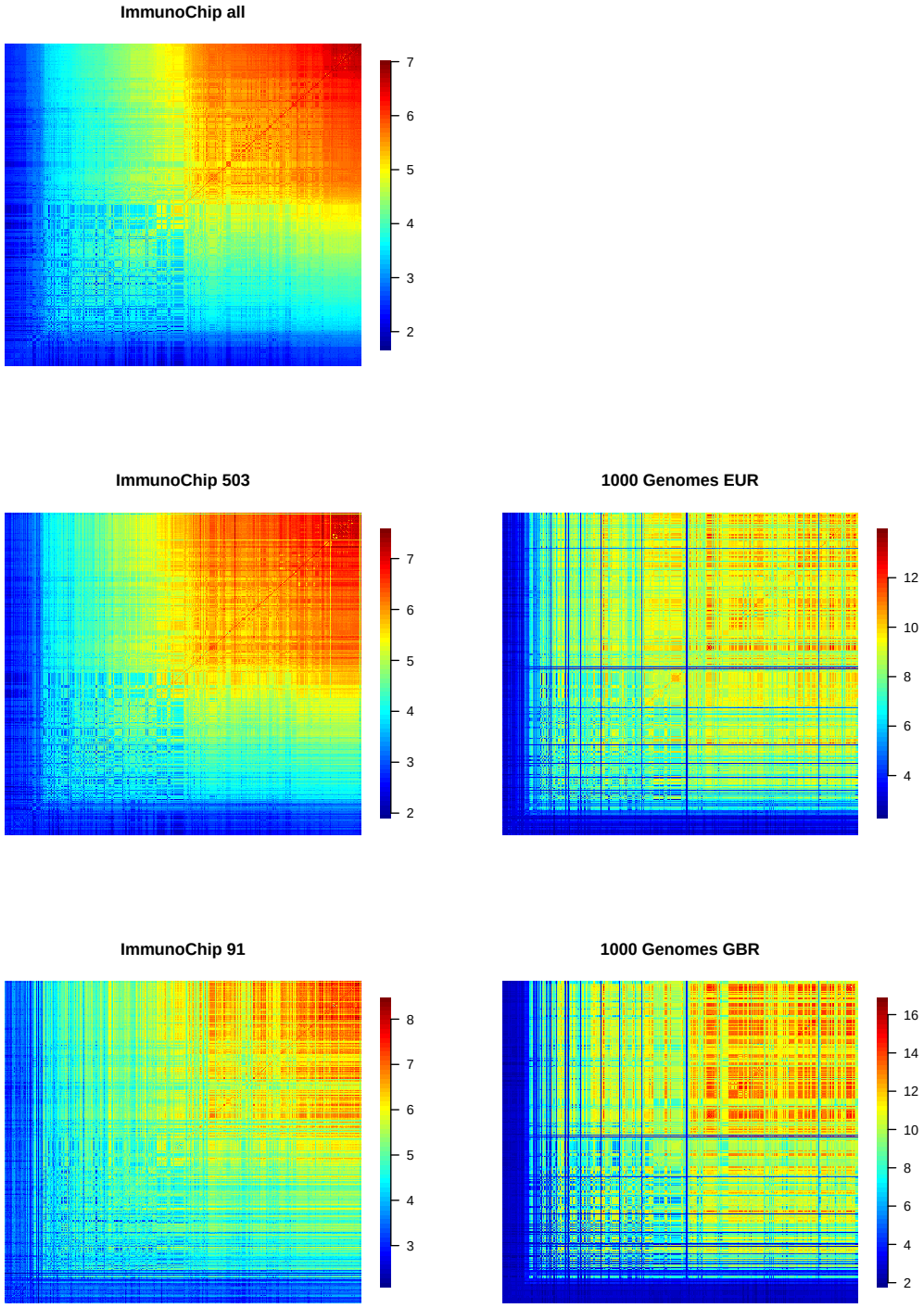


Figure 4.24 Heatmaps showing the impact of choice of each of the five reference datasets used upon T1D association found in the 10p15.1 region containing *IL2RA*. Each row and column of each heatmap corresponds to a SNP. SNPs have been ordered according to their minimum *sumsq* in the single causal variant models when analysed using the complete ImmunoChip reference dataset, starting from the bottom left hand corner. Each cell is coloured according to the minimum *sumsq* obtained at the model where **W** contains the two SNPs corresponding to the cell's row and column; the diagonals give the minimum *sumsq* obtained by the 1-CV model containing that SNP.

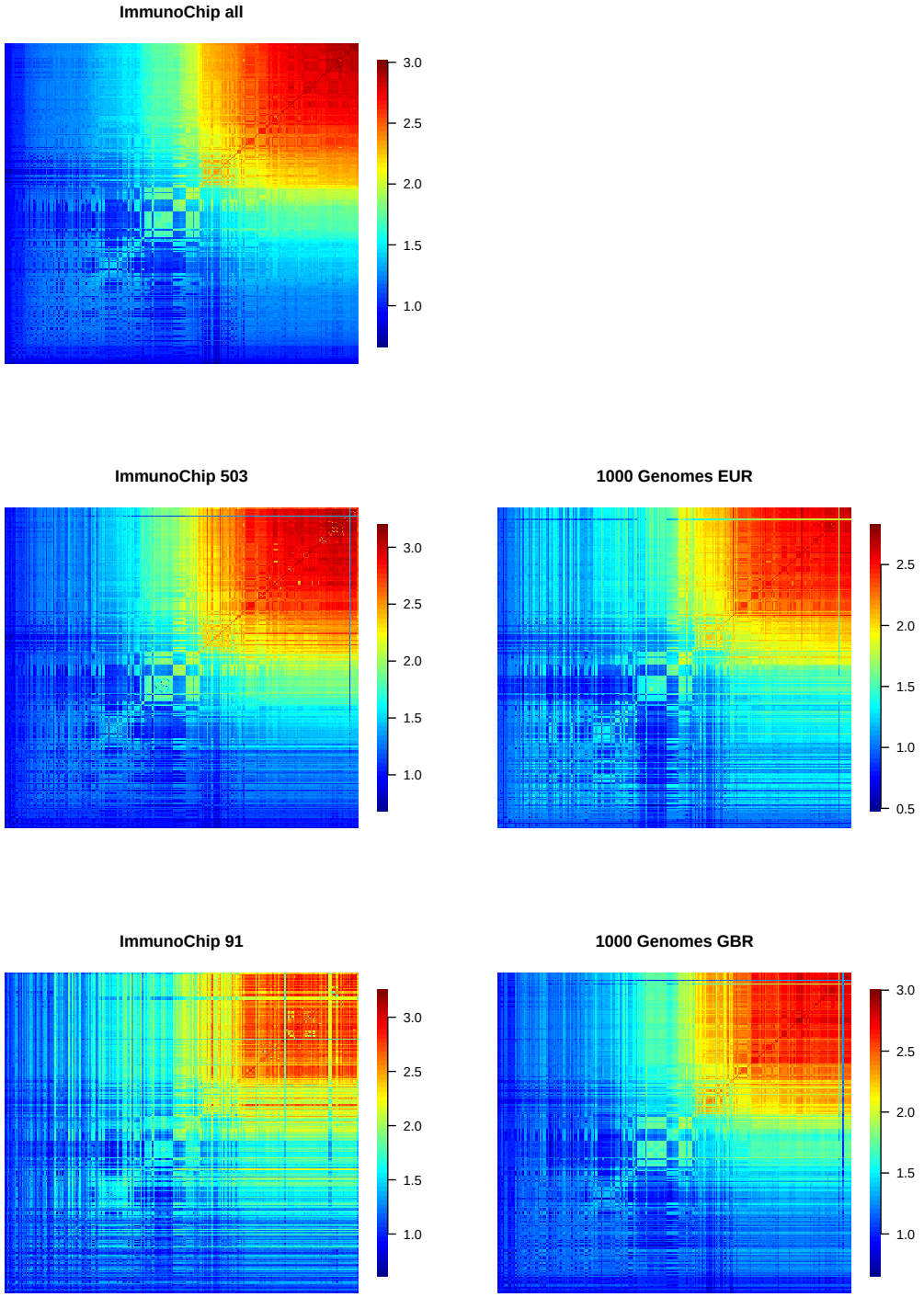


Figure 4.25 Heatmaps showing the impact of choice of each of the five reference datasets used upon MS association found in the 10p15.1 region containing *IL2RA*. Each row and column of each heatmap corresponds to a SNP. SNPs have been ordered according to their minimum *sumsq* in the single causal variant models when analysed using the complete ImmunoChip reference dataset, starting from the bottom left hand corner. Each cell is coloured according to the minimum *sumsq* obtained at the model where **W** contains the two SNPs corresponding to the cell's row and column; the diagnoals give the minimum *sumsq* obtained by the 1-CV model containing that SNP.

4.8 Discussion

In this chapter, I demonstrate a method for fine mapping which builds upon the algorithm for computing the expected output of a GWAS presented in Chapter 3. Using a publicly available reference dataset to estimate the between-SNP correlation structure, I can compute the expected summary statistics under a given causal model. By comparing this to the (unsigned) summary statistics reported by a GWAS, I am able to compute a measure of model-fit. Using an ABC-like framework, I sample from many such models, and am able to rank them. With an appropriate threshold, it may also be possible to determine a set which is likely to contain the true model, if the true model has been sampled.

Note that this method for fine mapping could also be employed when the direction of effect at each SNP is known; the observed signed Z Score can be directly compared to the expected signed Z Score. However, in such a case, the likelihood is tractable, and hence other fine mapping approaches, such as those discussed in Section 1.5 may be used instead.

4.8.1 Comparisons to other Fine Mapping Approaches

The simplest techniques for performing fine mapping using summary data use only the results of the single-SNP tests. They do not require the use of a reference dataset, and are computationally very quick to perform. However, since they approximate the complete data likelihood under a single-CV model, they are restricted to testing single CV models only. They also assume all possible CVs have been genotyped, and hence require imputation if the summary data is not known for all SNPs. In the absence of complete genotyping they have the potential to give incorrect results with high confidence; Figure 4.6 shows an example of a region with the likely causal variants removed; a standard single-CV summary statistic fine mapping results in a posterior probability of being causal of 1 for a SNP which mostly likely represents only LD with the CV. By contrast, my method allows for the analysis of arbitrarily large (subject to restrictions on computation time) potential causal models. It is able to recover causal variants even when those SNPs have been trimmed from the output

presented.

Other tests do exist which enable fine mapping of multiple CV models from summary data. I summarise two, PAINTOR and CAVIAR, in Section 1.4. However, both these approaches are likelihood based; as I discussed in Section 4.2.1, this requires knowing the signed p-values; that is, knowing the direction of effects for each SNP. However, many GWAS which report only summary data also fail to report these values or fail to clearly state which allele the direction corresponds to. Without these directions of effect, directly computing the likelihood is computationally intractable. By sampling from the prior distribution of causal models, I measure model fit by distance between absolute Z scores rather than by probability of observed data under full likelihood, and hence am able to perform fine mapping from the unsigned p-values.

In addition, both PAINTOR and CAVIAR assume that the effect at any SNP can be represented as a linear sum over the effects upon the causal variants (see Equations 1.2 and 1.3). However, this assumes that only pairwise LD affects the interaction between any SNP and the causal SNPs, and there are not any higher order effects. This is not a realistic assumption; if SNPs X_1, X_2 and X_3 are in LD, it is not possible to decompose $\mathbb{P}(G_i^{X_1} = x_1 \cap G_i^{X_2} = x_2 \cap G_i^{X_3} = x_3)$ into second-order terms. In the extreme case when SNP X occurs only in the presence of two or more of the causal SNPs, then Equations 1.2 and 1.3 will not be valid models for the effect at X. For this reason, rather than assuming linear sums, I have computed the complete probability $\mathbb{P}(G_i^X = x \cap G_i^{\mathbf{W}} = \mathbf{w})$ for all possible $\mathbf{w} \in \mathbb{Z}_3^m$. This will result in a more accurate estimate of the expected Z Score, however it also increases the computational time required, since I must compute the result in all 3^m cases.

4.8.2 Impact of the Reference Dataset

In common with any summary statistic fine mapping algorithm which does more than single SNP analysis, my method requires a reference dataset of genotype information to infer correlation structure in the control data. In practice, this data will not precisely match the control population used in the GWAS. In addition, in order to calculate the expected Z

Score, we must first compute $\mathbb{P}(G_i^X = x \cap G_i^{\mathbf{W}} = \mathbf{w})$, the joint probability of the genotype at which we wish to compute the Z Score, X and the genotypes of the causal SNPs, $\mathbf{W}$. This has a multivariate binomial distribution, and is estimated from the reference dataset of control samples. However, this reference dataset is of finite size, n. For a univariate binomial distribution, $Bin(2, \pi)$, the standard error is $\sqrt{\frac{2\pi(1-\pi)}{n}}$. Similarly, in the multivariate case, the standard error is proportional to $\sqrt{\frac{1}{n}}$, and becomes large as the joint probabilities of events become small. Hence, in models with many causal SNPs, or where causal SNPs have low MAF, many samples are required in order to accurately estimate the reference probabilities.

I show in Section 4.7 above that changing the reference dataset (whether that be by downsampling, or by using data from a subtly different population) has the potential to change my conclusions about a region. Although a strongly preferred model will remain the top model, its minimum *sumsq* value is dependant on the reference dataset. Whether any models fall within the *sumsq* bounds where we would expect to see a true causal model, and the relative rankings of models, is also subject to change.

This variability is to be expected, and is almost certainly present to some extent in all similar fine mapping algorithms. The Z Score at a non-causal SNP X must be some function of its correlation with the causal SNPs $\mathbf{W}$. Even within a relatively similar population, taking different samples as references will result in different estimated relationships between X and $\mathbf{W}$, and hence different expected Z Scores. A common source suggested for a reference dataset is the 1000 Genomes project. However, this contains only 91 samples from the GBR population; this is insufficient power to accurately estimate the MAF of a rare SNP, let alone the correlations between a group of rare SNPs. Aggregating 1000 Genomes datasets across many populations would increase the power to estimate such variables, but at a cost of the estimates no longer being appropriate to the population in the original GWAS. I therefore suggest the use of larger reference datasets, such as UK10K [Walter et al, 2015] if UK specific samples are sought, for this class of methods.

It is likely that the results of my method are more sensitive to choice of reference dataset

than either PAINTOR and CAVIAR. As discussed in 4.8.1 above, these methods require only estimation of the LD matrix. By contrast, my method requires the estimation of the correlation between the SNP at which we are currently estimating the Z Score and all causal SNPs in the model. For large models, the sampling variance in this estimation, and hence the variance in my results, will be much larger than the sampling variance when merely computing LD.

4.8.3 Extensions to my Fine Mapping Method

I have so far applied my fine mapping method to the analysis of results from a single population. Due to the difference in LD structures between populations, cross-ethnic studies cannot be performed simply by concatenating samples into a single large dataset, and instead a meta-analysis must be performed. However, the results from such a multi-population meta-analysis are simply a function of the Z Scores from the individual GWAS. It will be theoretically simple to apply my method to simulate the output of each of these GWAS and then combine them to estimate the expected meta-analysis output under a causal model of interest (which can potentially include population-specific causal SNPs or effect sizes), and hence perform cross-population fine mapping.

Similarly, I have so far applied this method to the fine mapping of a single disease. In Chapter 2, however, I discuss colocalization techniques, which disentangle whether two related diseases share causal variants or whether apparent common association to a region is a result of distinct causal variants. I extend existing approaches to the case of a shared control dataset; this is a common study design, since it increases power, but the resulting correlation between the analysis of the two diseases makes it more difficult to analyse. My colocalization techniques require complete genotype data for both sets of cases and controls; as is discussed in this chapter, such information is often not available.

It would be possible to extend my fine mapping approach to fine mapping two diseases simultaneously, and hence providing information about regions where colocalization occurs. Rather than considering causal models $\{\mathbf{W}, \boldsymbol{\gamma}\}$, the fit of joint models $\{\mathbf{W}^{\text{trait 1}}, \mathbf{W}^{\text{trait 2}}, \boldsymbol{\gamma}^{\text{trait 1}}, \boldsymbol{\gamma}^{\text{trait 2}}\}$ would be analysed. In the case of separate controls, no

additional method development would be required, and the expected output for each trait under the model being considered would be computed independently. However, methods already exist to perform colocalization from summary data only [Giambartolomei et al, 2014] (although they often assume at most one causal variant per trait per region, an assumption not required in my approach). There would be greater value in extending my method to allow for colocalization from summary statistic data when the control data is shared, which could be done by extending the algebra in Section 3.2 to the case where disease status is modelled by multinomial, rather than binomial, logistic regression.

Note, however, that due to the number of $\boldsymbol{\gamma}$ which must be sampled in order to converge to the minimum *sumsq* at a set of potential causal SNPs $\mathbf{W}$ (particularly in the case where the size of $\mathbf{W}$ is large), my fine mapping method is computationally expensive even when applied to a single population or disease. This effect will be magnified by the number of studies being simulated, and the size of the models being analysed.

Chapter 5

Discussion

In this thesis, I develop methods for the analysis of causal variants and causal processes in complex diseases. I extend methods which investigate whether variants within a region are shared between related diseases or are distinct. I also develop a method which enables the simulation of the output from a GWAS; this can be used to estimate the null distribution in a SNP set enrichment analysis, or to provide the backbone of a technique for fine mapping causal variants from only summary statistic data, using an approach similar to ABC.

5.1 Future Relevance of Methods Presented Here

In Chapter 2, I present extensions of colocalization algorithms to the study design where disease case samples are compared to a common control. Such a study design is common, especially by consortia which are analysing many diseases. Genotyping is expensive, and by genotyping only a single control set, the power of the study can be maximised. However, with data increasingly becoming cheaper to generate, or even being made freely available from repositories, will this design be less used?

The UK Biobank (`www.ukbiobank.ac.uk`) is a biorepository containing health information, including genotypes and disease statuses from 500,000 participants. It was designed to study multiple diseases with onset in later age, but such a large publicly available

resource is obviously an attractive source of population controls for many studies, including colocalization. However, there are complicating factors. For instance, unlike the ImmunoChip for autoimmune-associated regions, Biobank samples are not densely genotyped. In colocalization tests, the likelihood is partitioned according to the probability that each included variant is causal; an implicit assumption in interpreting the results is that, if there is a causal variant, it is genotyped. Hence, these tests require densely genotyped controls, and, by downsampling to the density found in Biobank, we would lose accuracy.

One solution might be to impute genotype data for those SNPs not genotyped by Biobank, using resources such as the Haplotype Reference Consortium to estimate haplotype probabilities (`www.haplotype-reference-consortium.org/`). This is not desirable, since it is preferable to avoid having to compare an imputed control dataset to a non-imputed case set. However, the use of imputation in the control data alone will result in differential measurement error; in addition, the choice of reference dataset used for imputation can lead to bias, particularly in SNPs with low MAF. This has the potential to cause an inflated Type 1 error rate in the subsequent analysis. While additional QC steps may be able to control this, they will likely result in the exclusion of many of the SNPs we wish to study [Sinnott and Kraft, 2012]. However, in such a case, the only other option is to downsample the ImmunoChip cases and then impute them back up to full density; a needless loss of information.

More fundamentally, methods for comparing case/control cohorts assume that there is no systematic difference other than disease status between the datasets being analysed. With large sample sizes taken from similar populations, extracted and stored in similar conditions, and genotyped upon the same chip, at the same centre, this assumption is probably valid, although detecting variants with systematically different errors between cases and controls remains a key step in every GWAS. When comparing an ImmunoChip case group to a Biobank control group such similarity conditions do not hold, and any results obtained might easily be artefacts of the differences in chips used for genotyping. Even when comparing two Biobank groups (which might not be possible for a sufficiently rare disease) care must be taken, since two different

arrays have been used for genotyping.

Hence it is likely that, particularly in the case of rare diseases or diseases where a specialised chip is appropriate, the common control design will continue to be used. Regardless of data availability, even though this design does make the downstream analysis more difficult due to the correlation structure induced, it does have a number of methodological advantages. It increases the power to detect associations (if only marginally as control data sizes increase relative to case data sizes) and negates the issue of error due to systematic differences between finite samples of separate controls.

Another common design likely to see continued use is that of overlapping controls. This may come about due to varying choice of standard reference cohorts, or even due to substantial difference in QC, between two studies. Just as in the common control case, the partial sharing of controls between studies leads to correlations which must be accounted for in comparative analyses. For instance, [Pickrell et al, 2016] estimate the correlation between effect sizes under the null model and use this to correct effect size estimates. The methods I present in Chapter 2 could also be extended to the case of overlapping controls.

In Chapter 4, I developed a technique which fine maps a genetic region given only summary data. Even if the use of freely available genotypes from biorepositories becomes standard within the field, there will still be a requirement for summary statistic methods. For instance, we may wish to analyse the results from a historical GWAS for which full genotype data is not available, or to integrate such a study into a meta-analysis. Making genotype data available requires ethical approval and consent from the donors: due to patient privacy concerns, summary data may be all it is possible to release from a study.

As datasets become increasingly large, using full genotype data becomes computationally infeasible. In such cases, a divide-and-conquer approach may be employed; genotype data is segmented, and the analysis is done within these segments, all outputs then being merged to obtain a final result. An approach using even a fraction of the true genotype data is likely to be more accurate than a summary statistic approach, which unavoidably relies upon a reference dataset to estimate SNP correlations. However, if a summary statistic approach could be

developed which is sufficient for the analysis in question, it might increase computational efficiency in such cases.

5.2 Future Directions

In Section 5.1 above, I discuss the issues with integrating data (for example, a case cohort and a control cohort) which have come from different sources. A fundamental assumption of the methods used to analyse such data is that there are no systematic differences between the cohorts other than that being studied. As the amount of data available increases, and sample sizes become larger, the statistical theory behind these methods will assume that the only source of error, sampling variation, becomes negligible, and will find differences between cohorts which would not be expected by random chance. However, practically in analysis of biological data, there are invariably errors which are systemic, and which do not decrease as sample size increases; for instance the cases and controls will frequently be genotyped at different centres. There is a need for the development of fine mapping analyses which allow for such errors.

However, the ultimate aim of disease analysis is for the results to translate to clinical impact. By finding the true causal variants, linking these to a gene or gene network and hence identifying a target, such as a protein or RNA molecule, we may be able to find a therapeutic which modulates this target and hence, potentially, treats the disease. This requires collaboration between many different disciplines, and projects such as Open Targets (`https://www.targetvalidation.org/`) aim to assist target validation by integrating information about potential drug targets.

GWAS, while they identify disease associated regions, are not able to perform causal variant identification. Instead, fine mapping is run as a secondary analysis, often on an ad-hoc basis, and with incomplete input data, possibly only on summary statistics, as reported from the GWAS. In order to better understand disease aetiology, I believe fine mapping should be part of the initial aim of a GWAS, with a causal variant identification technique run systematically

upon each region which shows significant evidence of disease association.

In addition, once these causal variants have been found, it still remains to characterise the effect they have upon genes. By integrating GWAS and fine mapping results with functional annotation data, we hope to identify such effects and hence also novel candidate causal genes. There are many effects which could potentially have a role in disease processes; variants in coding regions may affect the function of a protein, or a variant might result in splicing errors. However, in the case of complex traits such as autoimmune disease, one of the largest effects is expected to be from variants changing gene expression; hence, many approaches focus on eQTL data to identify causal genes. These effects may well be cell-type specific, or even cell-state specific (in autoimmune disease, for instance, we might look for differences in gene expression between non-activated and activated $CD4^+$ T Cells).

Although I extended it to the analysis of two traits in the case of a common control dataset, the Bayesian colocalization method described in Section 1.4.1 can be used to infer whether disease associated variants appear to colocalize with eQTL signals by simultaneously fine mapping disease and eQTL signals, and integrating the likelihood of shared versus distinct causal variants over the two probabilistic maps. This is illustrated in Section 2.3, where I ran this method upon the region containing candidate causal gene *FADS2*. I found strong evidence of colocalization between Crohn's Disease and an eQTL for *FADS2* in five cell types ($CD4^+$, $CD8^+$, $CD14^+$, $CD16^+$, $CD19^+$), implicating *FADS2* in the aetiology of CD. Functional data such as eQTLs can also be directly analysed for disease association. In a Transcription Wide Association Study (TWAS) [Gusev et al, 2016] the eQTLs are mapped in a separate dataset, and then used to impute the gene expression for a set of control samples and for a set of case samples (thus increasing the power available, since the cost of measuring gene expression would lead to small samples sizes without imputation). A GWAS-like analysis of dependence between expression and disease status is then performed in order to identify genes which have significant differential expression between cases and controls. Alternatively, [Zhu et al, 2016] uses Mendelian Randomisation style techniques to test for association between disease expression and disease status, taking as input summary data from GWAS and eQTL studies.

Studies combining eQTLs with autoimmune disease GWAS have found evidence of disease-associated genes. [Guo et al, 2015] identified evidence of disease association at six genes in monocytes and/or B cells, including some cell-type specific effects. However, out of the 125 genes showing evidence of potential overlapping disease association and eQTL signals, only 28 showed some support for colocalization. [Huang et al, 2015] performed a fine mapping of 94 loci associated with IBD, and found that the overlap between eQTLs and disease associated credible sets of SNps was no more than would be expected by random chance. While enrichment of eQTLs in the disease credible sets were found in cell types including $CD4^+$ and ileum cells, the majority of cell types tested, all of which were plausibly associated with IBD, showed no significant overlap. Based upon our current understanding of gene expression as an important mechanism in the aetiology of complex diseases, we would expect to see overlap of causal variants and eQTLs more frequently than these studies suggest is the case. What is causing these studies to be less successful than we might expect?

One explanation might be that, while the effects of eQTLs upon disease status are present, they are so tissue-context specific that, unless a study surveys precisely the correct tissue, in the correct state, no association will be found. However, the space of potential tissue contexts is large; for instance, gene expression differs between a T-cell which has just been activated, and that same T-cell an hour post activation, as well as depending on the means of T-cell activation, so it may not be appropriate to consider a single analysis of eQTLs in activated T-cells. This suggests that, for any study into the effect of gene expression upon disease, care must be taken to select the appropriate analysis, and a range of cell types in a range of conditions should probably be considered.

In addition, the odds ratios of effect found in GWAS for complex diseases tend to be small; see Section 3.3.3 for an empirical distribution of effects sizes reported in the GWAS catalog. Due to the expense, many eQTL studies have low sample sizes; it could simply be that they are underpowered to detect the effects which are present. As technology becomes cheaper, and larger datasets become available, it is to be hoped that studies will become powered to detect smaller effects due to gene expression.

However, a single gene will have many eQTLs, and there is evidence that their effect upon the expression of their target is additive. Hence, it is also possible that we do not tend to see significant association between individual eQTLs and disease status since genes are able to compensate for a variant affecting a single enhancer via the actions of their other eQTLs.

While I believe that the development of methods which identify differential gene expression in disease is important in the search for potential drug targets, there is still an important place for fine mapping techniques such as those presented in this thesis. Although gene expression being associated with disease status is suggestive of a causal role, it does not constitute verification of causal status. eQTL studies are unable to distinguish between "driver" genes, which contribute to the disease process, and "passenger" genes, where the differential expression is a downstream effect of the disease. It is by investigating the genetic variants which cause the disease that we can determine in which direction this causality runs.

Appendix A

Regions Analysed in Chapter 2

This Appendix gives a list of all 126 regions analysed by my colocalization analyses in Chapter 2, their locations and the candidate causal genes within them. Also given are which of the four diseases (T1D, RA, CEL, MS) have existing associations with the region, and which of the four diseases my colocalization analysis found to be associated with the region. Novel disease associations are indicated by the use of a bold font, while distinct signals are indicated by a "|".

Locus (hg18)	Band	Existing associations	Associations found	Candidate causal genes and genes in region
chr1:2353185-2786479	1p36.32	RA, CEL, MS	RA, CEL, MS	*PLCH2 TNFRSF14 FAM213B MMEL1 TTC34*
chr1:25096906-25180863	1p36.11	CEL	None	*RUNX3*
chr1:85376325-85713887	1p22.3	MS	None	*BCL10 DDAH1*
chr1:92023171-93311800	1p22.1	MS	MS	*EVI5*
chr1:100982239-101455699	1p21.2	MS	MS	*EXTL2 SLC30A7*
chr1:113619999-114460000	1p13.2	T1D, RA	T1D, RA	*PTPN22*
chr1:116831830-116911865	1p13.1	MS	MS	*CD58*
chr1:152574287-152933315	1q21.3	RA	RA	*ATP8B2 IL6R*

Continued on next page

Table A.1 – continued from previous page

Locus (hg18)	Band	Existing associations	Associations found	Candidate causal genes and genes in region
chr1:155746666-156085174	1q23.1	MS	None	*FCRL3 FCRL1*
chr1:158947387-159200000	1q23.3	MS	None	*CD48 SLAMF7 CD244 ITLN1*
chr1:170882016-171208336	1q24.3	CEL	**T1D**, CEL	*FASLG*
chr1:190728935-190815166	1q31.2	CEL, MS	CEL, MS	*RGS1*
chr1:199110000-199320000	1q32.1	T1D, CEL, MS	CEL, MS	*C1orf106 KIF21B*
chr1:204869062-205116454	1q32.1	T1D	T1D	*MAPKAPK2 IL10 IL19 IL20*
chr2:24539944-25341162	2p23.3	T1D, MS	None	*CENPO ADCY3 EFR3B DNMT3A*
chr2:43165703-43240464	2p21	MS	MS	
chr2:60722116-61952276	2p16.1	RA, CEL, MS	CEL|MS	*AC010733.4 REL PUS10 KIAA1841 C2orf74 AHSA2*
chr2:65246601-65570598	2p14	RA	RA, **MS**	*SPRED2*
chr2:68388948-68711822	2p14	CEL, MS	None	*PLEK*
chr2:99883120-100415547	2q11.2	T1D, RA	T1D, RA, **CEL**	*AFF3*
chr2:102169652-102670082	2q12.1	CEL	CEL	*IL1RL2 IL1RL1 IL18R1 IL18RAP*
chr2:162669118-163101007	2q24.2	T1D	T1D	*IFIH1 KCNH7*
chr2:181022069-181977071	2q31.3	CEL	CEL	*UBE2E3*
chr2:191412527-191739472	2q32.2	T1D, RA, CEL, MS	RA, CEL|MS	*STAT1 STAT4*
chr2:202920548-204528303	2q33.1	T1D, RA, CEL	T1D, CEL|RA	*CD28 CTLA4 ICOS*
chr2:230758228-230962304	2q37.1	MS	**CEL**, MS	*SP140*
chr3:18582795-18831864	3p24.3	MS	MS	
chr3:27656007-27811049	3p24.1	MS	None	*EOMES*
chr3:28015774-28105476	3p24.1	MS	MS	
chr3:32873208-33063377	3p22.3	CEL, MS	None	*CCR4*

Continued on next page

Table A.1 – continued from previous page

Locus (hg18)	Band	Existing associations	Associations found	Candidate causal genes and genes in region
chr3:45812888-46633741	3p21.31	T1D, CEL	T1D|CEL	*CCR3 CCR1 CCR5 LTF*
chr3:120581991-120788414	3q13.33	CEL, MS	MS	*ARHGAP31 TMEM39A POGLUT1 TIMMDC1 CD80*
chr3:122818149-123329522	3q13.33	MS	MS	*IQCB1 SLC15A2 CD86*
chr3:160950948-161389020	3q25.33	CEL, MS	CEL|MS	*IL12A*
chr3:189504161-189629875	3q28	CEL	CEL	
chr4:25637284-25745871	4p15.2	T1D, RA	RA	
chr4:103607587-104383056	4q24	MS	None	*NFKB1 MANBA*
chr4:106143093-106702164	4q24	MS	None	*TET2*
chr4:123121079-124497235	4q27	T1D, RA, CEL	T1D|CEL	*KIAA1109 ADAD1 IL2 IL21*
chr5:35831493-36107254	5p13.2	MS	**T1D**, MS	*SPEF2 IL7R CAPSL UGT3A1*
chr5:40322722-40723788	5p13.1	MS	MS	*PTGER4*
chr5:55450712-55492884	5q11.2	RA, MS	**D**RM	*ANKRD55*
chr5:102062861-102777130	5q21.1	RA	RA	*GIN1 C5orf30*
chr5:141392811-141620603	5q31.3	MS	None	*NDFIP1*
chr5:158451344-158758888	5q33.3	MS	None	*IL12B*
chr5:176439335-176780625	5q35.2	MS	None	*RGS14*
chr6:315547-412533	6p25.3	CEL	CEL	*IRF4*
chr6:36452190-36721790	6p21.31	MS	MS	*PXT1*
chr6:90863554-91103018	6q15	T1D, CEL, MS	T1D, **RA**, MS	*BACH2*
chr6:126479721-127461527	6q22.32	T1D	T1D	*CENPW*
chr6:127876526-128385456	6q22.33	CEL, MS	CEL	*THEMIS PTPRK*
chr6:135630625-136228061	6q23.3	MS	None	*AHI1*
chr6:137348296-137587799	6q23.3	MS	MS	*IL22RA2*

Continued on next page

Table A.1 – continued from previous page

Locus (hg18)	Band	Existing associations	Associations found	Candidate causal genes and genes in region
chr6:137914792-138345363	6q23.3	RCM	**T1D**, RA, CEL|MS	*TNFAIP3*
chr6:159237498-159464567	6q25.3	T1D, RA, CEL, MS	CEL, MS	*TAGAP*
chr6:167268496-167467944	6q27	RA	RA	*RNASET2 FGFR1OP CCR6*
chr7:26624486-27436525	7p15.2	T1D, MS	T1D, MS	*SKAP2*
chr7:28086237-28228851	7p15.1	MS	None	*JAZF1*
chr7:37323488-37406978	7p14.2	CEL, MS	**RA**, CEL, MS	*ELMO1*
chr7:50222360-50335957	7p12.2	MS	**T1D**, MS	*IKZF1*
chr7:50337180-50662811	7p12.2	T1D	T1D	*IKZF1 FIGNL1*
chr7:50866661-51640000	7p12.2	T1D	T1D	*COBL*
chr7:128338975-128564756	7q32.1	RA	RA	*IRF5 TNPO3*
chr8:11375792-11389894	8p23.1	RA	None	*BLK*
chr8:79575897-79914680	8q21.12	MS	MS	*PKIA ZC2HC1A*
chr9:4218549-4311558	9p24.2	T1D	T1D	*GLIS3*
chr9:34638417-34986014	9p13.3	RA	**CEL**	*CCL21*
chr10:6068495-6237542	10p15.1	T1D, RA, MS	T1D|MS	*IL2RA*
chr10:6428075-6585110	10p15.1	T1D, RA, CEL	None	*PRKCQ*
chr10:31172479-31520710	10p11.23	MS	None	*ZNF438*
chr10:35080006-35590006	10p11.21	T1D	None	*CREM CCNY*
chr10:80658841-80774414	10q22.3	CEL, MS	CEL	*ZMIZ1*
chr10:89998026-90268360	10q23.31	T1D	T1D	*RNLS*
chr10:94189315-94491883	10q23.32	MS	None	*HHEX*
chr11:2024999-2264880	11p15.5	T1D	T1D	*INS*

Continued on next page

Table A.1 – continued from previous page

Locus (hg18)	Band	Existing associations	Associations found	Candidate causal genes and genes in region
chr11:46304899-49088571	11p11.2	MS	None	*AGBL2*
chr11:60482183-60617465	11q12.2	MS	MS	*CD6*
chr11:63600519-63980103	11q13.1	MS	None	*FLRT1 TRPT1 ESRRA PRDX5 CCDC88B RPS6KA4*
chr11:117805448-118403529	11q23.3	RA, CEL, MS	CEL|MS	*TREH DDX6 CXCR5*
chr11:127754640-128010703	11q24.3	CEL	CEL	*ETS1*
chr12:6291754-6334123	12p13.31	MS	MS	*TNFRSF1A SCNN1A*
chr12:6353046-6393510	12p13.31	MS	None	*SCNN1A LTBR*
chr12:9407874-9867423	12p13.31	T1D, MS	**CEL**	*CLECL1 CD69*
chr12:54637612-55091576	12q13.2	T1D	T1D	*IKZF4 ERBB3 IL23A STAT2*
chr12:56127370-56774934	12q13.3	T1D, MS	MS	*AGAP2 CYP27B1 TSFM*
chr12:109772108-111723111	12q24.11	T1D, RA, CEL	T1D|CEL	*SH2B3 ATXN2 BRAP*
chr12:121926103-122574026	12q24.31	MS	MS	*PITPNM2*
chr13:98723872-99034738	13q32.3	T1D, MS	T1D	*GPR183*
chr14:68231082-68387815	14q24.1	T1D, CEL, MS	MS	*RAD51B ZFP36L1*
chr14:75012674-75107858	14q24.3	MS	None	*BATF*
chr14:87372049-87716867	14q31.3	MS	MS	*GALC GPR65*
chr14:97427666-97601359	14q32.2	T1D	T1D	
chr14:100357783-100398492	14q32.2	T1D	T1D	
chr15:36603999-36786000	15q14	T1D ,RA	T1D	*RASGRP1*
chr15:72389033-73270664	15q24.1	CEL	None	*CLK3 CSK*
chr15:76773859-77050416	15q25.1	T1D, MS	T1D, **CEL**	*CTSH*
chr15:88612805-89221004	15q26.1	MS	None	*IQGAP1 CRTC3*

Continued on next page

Table A.1 – continued from previous page

Locus (hg18)	Band	Existing associations	Associations found	Candidate causal genes and genes in region
chr16:10831557-11408130	16p13.13	T1D, CEL, MS	T1D, MS|CEL	*CIITA DEXI CLEC16A RMI2 SOCS1 PRM3 PRM2 PRM1*
chr16:28191235-28944416	16p11.2	T1D	T1D	*IL27 NUPR1 SULT1A2 SULT1A1 EIF3C SH2B1 RABEP2 CD19 LAT*
chr16:29753185-30627501	16p11.2	MS	None	*MAPK3 ITGAL*
chr16:66887501-67407338	16q22.1	MS	None	*ZFP90 CDH3*
chr16:73760230-74086012	16q23.1	T1D	T1D	
chr16:84539746-84581605	16q24.1	RA, MS	RA, MS	
chr17:34629755-35508018	17q12	T1D, RA, MS	T1D	*PNMT RP11-94L15.2 IKZF3 ZPBP2 GSDMB ORMDL3 GSDMA*
chr17:35990899-36132000	17q21.2	T1D	None	
chr17:37562258-38298988	17q21.2	MS	MS	*STAT5B STAT5A STAT3 PTRF MLX*
chr17:42664102-43231021	17q21.32	MS	MS	*NPEPPS TBKBP1 TBX21*
chr18:12407903-12919721	18p11.21	T1D, CEL	T1D|CEL	*PTPN2*
chr18:65630494-65722590	18q22.2	T1D	None	*CD226*
chr19:6564831-6636304	19p13.3	MS	MS	*TNFSF14*
chr19:10081000-11019034	19p13.2	T1D, RA, MS	T1D, RA, MS|**CEL**	*PPAN-P2RY11 PPAN ICAM1 ICAM3 TYK2 CDC37 SLC44A2 ILF3 CARM1*
chr19:16300497-16612240	19p13.11	MS	MS	
chr19:17905598-18272802	19p13.11	MS	MS	*IFI30 MPV17L2*
chr19:51843217-52015224	19q13.32	T1D	T1D	*PRKD2 STRN4*
chr19:53784241-53969894	19q13.32	T1D	T1D	*SPHK2 DBP FUT2 MAMSTR RASIP1 IZUMO1*
chr20:1444472-1707590	20p13	T1D	T1D	
chr20:43965660-44217558	20q13.12	RA, MS	**CEL**, MS	*PLTP MMP9 NCOA5 CD40*
chr20:47840533-48095989	20q13.13	MS	None	*SLC9A8 RNF114*

Continued on next page

Table A.1 – continued from previous page

Locus (hg18)	Band	Existing associations	Associations found	Candidate causal genes and genes in region
chr20:52207832-52256247	20q13.2	MS	None	*CYP24A1*
chr20:61650000-61959471	20q13.33	MS	None	*STMN3 RTEL1 RTEL1-TNFRSF6B TNFRSF6B ZGPAT LIME1 SLC2A4RG ZBTB46*
chr21:42681877-42771181	21q22.3	T1D, RA, CEL	T1D|RA|CEL	*UBASH3A*
chr21:44414408-44528088	21q22.3	CEL	None	*ICOSLG*
chr22:20042414-20686540	22q11.21	RA, CEL, Ms	CEL	*UBE2L3 YDJC CCDC116 MAPK1*
chr22:28137854-28999883	22q12.2	T1D	T1D	*MTMR3 LIF OSM*
chr22:35898615-35996732	22q12.3	T1D	T1D	*IL2RB C1QTNF6*

Appendix B

Results from Colocalization Analysis for Regions Mentioned in Chapter 2

This Appendix gives the results from my colocalization analysis, for 24 of the 126 regions tested. Included are regions which were discussed in detail in Chapter 2, regions which were given in Table 2.3 (regions showing evidence of seperate SNP effects) and regions which were given in Table 2.4 (regions showing strong evidence of novel association).

For each region is given the location, the candidate causal genes within the region, which of the four autoimmune diseases (T1D, RA, CEL, MS) it had existing associations with, and which of the four autoimmune diseases my analysis concluded it has associations with. For each of the six pairwise analyses of the diseases, from the Bayesian approach, I give the posterior probabilities of $\mathbb{H}_0$, $\mathbb{H}_1$, $\mathbb{H}_2$, $\mathbb{H}_3$, $\mathbb{H}_4$. From the proportional approach I give the posterior predictive p-value, and the estimate $\hat{\eta}$, the constant of proportionality.

Location		Candidate Causal Genes		Existing Associations		Associations Found	
1q24.3		*FASLG*		CEL		T1D, CEL	
Analysis	PP.H0	PP.H1	PP.H2	PP.H3	PP.H4	eta.hat	P-Value
T1D-RA	8.30E-01	1.64E-01	2.15E-03	3.98E-04	2.66E-03	1.04E+00	3.13E-01
T1D-CEL	5.21E-08	8.72E-09	2.51E-01	3.49E-02	7.14E-01	1.88E+00	6.22E-01
T1D-MS	5.87E-01	4.11E-01	1.05E-03	7.33E-04	1.56E-05	3.33E+00	6.87E-02
RA-CEL	1.28E-06	2.17E-09	9.88E-01	1.58E-03	1.00E-02	2.18E+00	5.00E-01
RA-MS	9.96E-01	2.64E-03	1.85E-03	4.91E-06	7.00E-07	8.18E-01	5.44E-02
CEL-MS	4.19E-08	9.98E-01	7.35E-11	1.74E-03	6.65E-05	2.65E-01	5.43E-02

Location		Candidate Causal Genes		Existing Associations		Associations Found	
2p16.1		*REL*		RA, CEL, MS		CEL | MS	
Analysis	PP.H0	PP.H1	PP.H2	PP.H3	PP.H4	eta.hat	P-Value
T1D-RA	9.43E-01	1.87E-03	5.54E-02	1.10E-04	4.13E-05	2.87E+00	2.01E-01
T1D-CEL	1.28E-02	1.76E-05	9.86E-01	1.35E-03	3.01E-04	3.21E+00	1.11E-01
T1D-MS	2.57E-01	5.31E-04	7.41E-01	1.53E-03	4.11E-04	4.02E+00	2.67E-01
RA-CEL	6.25E-03	7.12E-04	8.86E-01	1.01E-01	6.48E-03	3.16E+00	3.61E-02
RA-MS	3.21E-01	2.33E-02	5.87E-01	4.25E-02	2.63E-02	1.15E+00	1.58E-01
CEL-MS	1.52E-04	2.72E-01	3.60E-04	6.45E-01	8.19E-02	6.34E-01	1.42E-02

Location		Candidate Causal Genes		Existing Associations		Associations Found	
2p14		*SPRED2*		RA		RA, MS	
Analysis	PP.H0	PP.H1	PP.H2	PP.H3	PP.H4	eta.hat	P-Value
T1D-RA	6.25E-02	1.28E-04	9.35E-01	1.91E-03	1.68E-04	3.58E+00	2.23E-01
T1D-CEL	9.97E-01	1.11E-03	1.80E-03	2.00E-06	3.28E-07	5.75E-01	3.77E-01
T1D-MS	4.33E-01	7.06E-04	5.66E-01	9.22E-04	7.38E-05	3.08E+00	7.48E-01
RA-CEL	5.59E-02	9.42E-01	9.72E-05	1.63E-03	7.77E-04	3.98E-01	6.13E-01
RA-MS	6.91E-03	1.22E-01	1.41E-02	2.43E-01	6.14E-01	9.89E-01	4.07E-01
CEL-MS	7.54E-01	2.00E-03	2.43E-01	6.42E-04	1.92E-04	3.14E+00	2.69E-01

Location		Candidate Causal Genes		Existing Associations		Associations Found	
2q11.2		*AFF3*		T1D, RA		T1D, RA, CEL	
Analysis	PP.H0	PP.H1	PP.H2	PP.H3	PP.H4	eta.hat	P-Value
T1D-RA	1.18E-06	7.69E-05	1.94E-03	1.18E-01	8.80E-01	1.25E+00	3.84E-01
T1D-CEL	1.33E-04	1.18E-02	4.76E-03	4.19E-01	5.65E-01	9.43E-01	1.38E-01
T1D-MS	4.28E-02	9.43E-01	4.33E-04	9.49E-03	4.50E-03	5.57E-01	2.09E-01
RA-CEL	1.89E-07	5.25E-04	4.01E-05	1.02E-01	8.97E-01	8.00E-01	2.37E-01
RA-MS	5.88E-04	9.58E-01	2.50E-05	4.07E-02	6.90E-04	2.85E-01	2.06E-01
CEL-MS	2.59E-02	9.53E-01	5.16E-04	1.90E-02	1.20E-03	7.34E-01	2.77E-01

Location	Candidate Causal Genes	Existing Associations	Associations Found
2q32.2	*STAT1 STAT4*	T1D, RA, CEL, MS	RA, CEL | MS

Analysis	PP.H0	PP.H1	PP.H2	PP.H3	PP.H4	eta.hat	P-Value
T1D-RA	2.86E-01	2.45E-02	6.26E-01	5.34E-02	1.03E-02	9.73E-01	5.25E-01
T1D-CEL	4.73E-01	4.71E-02	3.53E-01	3.42E-02	9.22E-02	1.05E+00	2.95E-01
T1D-MS	3.80E-02	5.01E-03	8.45E-01	1.11E-01	2.07E-04	-7.32E-01	3.38E-02
RA-CEL	2.68E-01	2.99E-01	1.85E-01	2.07E-01	4.08E-02	1.22E+00	2.56E-01
RA-MS	1.32E-02	1.50E-02	4.51E-01	5.11E-01	9.23E-03	-1.04E+00	5.74E-02
CEL-MS	1.99E-02	1.99E-02	4.79E-01	4.79E-01	1.57E-03	-7.08E-01	1.51E-03

Location	Candidate Causal Genes	Existing Associations	Associations Found
2q33	*CD28, CTLA4, ICOS*	T1D, RA, CEL	T1D, RA, CEL

Analysis	PP.H0	PP.H1	PP.H2	PP.H3	PP.H4	eta.hat	P-Value
T1D-RA	1.18E-11	1.29E-02	8.97E-10	9.85E-01	2.52E-03	5.10E-01	1.69E-01
T1D-CEL	1.10E-14	2.47E-05	5.67E-11	1.18E-01	8.82E-01	7.32E-01	1.62E-02
T1D-MS	1.78E-09	9.99E-01	1.01E-12	5.69E-04	6.22E-05	-4.72E-02	2.71E-01
RA-CEL	7.77E-07	9.94E-06	7.12E-02	9.12E-01	1.66E-02	1.62E+00	1.47E-01
RA-MS	2.68E-01	7.31E-01	1.26E-04	3.44E-04	1.02E-04	1.79E-01	1.72E-01
CEL-MS	4.24E-05	9.99E-01	1.88E-08	4.42E-04	7.52E-05	-1.09E-01	2.44E-01

Location	Candidate Causal Genes	Existing Associations	Associations Found
2q37.1	*SP140*	MS	CEL, MS

Analysis	PP.H0	PP.H1	PP.H2	PP.H3	PP.H4	eta.hat	P-Value
T1D-RA	6.64E-01	3.36E-01	1.02E-04	5.13E-05	2.34E-05	-1.74E-01	1.93E-01
T1D-CEL	3.76E-01	2.07E-02	3.73E-01	1.84E-02	2.12E-01	9.96E-01	5.16E-01
T1D-MS	2.61E-05	2.29E-06	6.71E-01	5.62E-02	2.73E-01	1.57E+00	3.37E-01
RA-CEL	4.40E-01	9.94E-05	5.60E-01	1.25E-04	9.60E-05	-3.18E+00	2.58E-01
RA-MS	1.00E-07	9.44E-11	9.99E-01	9.40E-04	7.34E-05	-1.48E+01	2.67E-01
CEL-MS	3.18E-09	5.12E-09	6.19E-02	9.13E-02	8.47E-01	1.76E+00	6.54E-01

Location	Candidate Causal Genes	Existing Associations	Associations Found
3p21.31	*CCR3, CCR1, CCR5*	T1D, CEL	T1D | CEL

Analysis	PP.H0	PP.H1	PP.H2	PP.H3	PP.H4	eta.hat	P-Value
T1D-RA	1.68E-01	8.29E-01	3.63E-04	1.79E-03	2.18E-04	1.43E+00	2.81E-01
T1D-CEL	2.13E-09	2.70E-08	7.29E-02	9.22E-01	4.87E-03	2.12E+00	2.65E-01
T1D-MS	6.45E-01	3.46E-01	5.51E-03	2.95E-03	3.08E-04	1.54E+00	4.75E-01
RA-CEL	9.01E-08	3.15E-10	9.96E-01	3.47E-03	3.68E-04	2.13E+00	3.74E-01
RA-MS	9.90E-01	3.23E-03	7.18E-03	2.33E-05	1.03E-05	1.23E+00	1.61E-01
CEL-MS	1.63E-08	8.53E-01	2.61E-09	1.36E-01	1.04E-02	6.42E-01	8.07E-01

Location	Candidate Causal Genes	Existing Associations	Associations Found
3q25.33	*IL12A*	CEL, MS	CEL | MS

Analysis	PP.H0	PP.H1	PP.H2	PP.H3	PP.H4	eta.hat	P-Value
T1D-RA	9.99E-01	4.56E-04	3.34E-04	1.51E-07	1.16E-07	-1.22E-01	2.04E-01
T1D-CEL	5.01E-15	2.01E-18	9.99E-01	3.97E-04	1.51E-04	8.00E+01	4.97E-01
T1D-MS	9.52E-03	7.73E-06	9.90E-01	8.03E-04	7.37E-05	-1.76E+01	1.14E-01
RA-CEL	1.26E-14	5.83E-18	9.99E-01	4.60E-04	2.72E-04	-5.54E+00	4.09E-01
RA-MS	1.08E-02	4.67E-06	9.89E-01	4.26E-04	4.79E-05	8.84E+00	3.01E-01
CEL-MS	3.05E-18	3.30E-02	8.86E-17	9.59E-01	7.59E-03	-4.75E-01	1.10E-01

Location	Candidate Causal Genes	Existing Associations	Associations Found
4q27	*IL2 IL21*	T1D, RA, CEL	T1D | CEL

Analysis	PP.H0	PP.H1	PP.H2	PP.H3	PP.H4	eta.hat	P-Value
T1D-RA	1.31E-05	8.15E-01	2.83E-06	1.77E-01	8.85E-03	4.74E-01	1.61E-01
T1D-CEL	9.46E-26	1.89E-20	5.01E-06	1.00E+00	4.50E-08	2.05E+00	6.40E-09
T1D-MS	1.19E-05	1.00E+00	1.82E-09	1.52E-04	5.35E-05	1.94E-01	2.27E-01
RA-CEL	5.51E-22	1.03E-22	5.84E-01	1.06E-01	3.09E-01	3.07E+00	2.91E-02
RA-MS	8.25E-01	1.75E-01	2.68E-04	5.62E-05	5.96E-05	5.71E-01	4.83E-01
CEL-MS	2.23E-21	9.99E-01	4.72E-25	2.08E-04	3.20E-04	2.70E-01	4.02E-01

Location	Candidate Causal Genes	Existing Associations	Associations Found
5q11.2	*ANKRD55*	RA, MS	T1D, RA, MS

Analysis	PP.H0	PP.H1	PP.H2	PP.H3	PP.H4	eta.hat	P-Value
T1D-RA	1.40E-07	3.41E-09	2.91E-01	4.56E-131	7.09E-01	2.30E+00	5.01E-01
T1D-CEL	9.51E-01	4.88E-02	3.15E-04	1.52E-05	1.01E-04	8.36E-01	4.09E-01
T1D-MS	8.26E-06	2.00E-07	2.94E-01	6.76E-05	7.06E-01	2.09E+00	3.43E-01
RA-CEL	1.43E-07	9.98E-01	4.49E-11	2.93E-04	1.98E-03	2.82E-01	2.61E-01
RA-MS	9.30E-14	2.94E-07	3.17E-09	8.08E-05	1.00E+00	8.64E-01	2.39E-01
CEL-MS	3.48E-05	9.04E-09	9.98E-01	2.40E-04	1.96E-03	2.75E+00	1.41E-01

Location	Candidate Causal Genes	Existing Associations	Associations Found
6q23.3	*TNFAIP3*	RA, CEL, MS	T1D, RA, CEL | MS

Analysis	PP.H0	PP.H1	PP.H2	PP.H3	PP.H4	eta.hat	P-Value
T1D-RA	1.45E-04	3.83E-05	1.99E-01	4.50E-02	7.56E-01	1.43E+00	3.74E-01
T1D-CEL	2.22E-21	6.14E-22	2.32E-01	5.70E-02	7.11E-01	2.06E+00	1.14E-01
T1D-MS	3.51E-02	2.09E-02	5.88E-01	3.50E-01	6.50E-03	-1.47E+00	1.59E-03
RA-CEL	2.16E-25	3.27E-22	3.19E-05	3.88E-02	9.61E-01	1.56E+00	2.48E-01
RA-MS	7.38E-04	2.44E-01	2.26E-03	7.49E-01	3.47E-03	-4.24E-01	1.70E-02
CEL-MS	2.06E-21	1.51E-01	1.16E-20	8.48E-01	1.32E-03	-7.76E-02	1.04E-03

Location	Candidate Causal Genes	Existing Associations	Associations Found
6q25.3	*TAGAP*	T1D, RA, CEL, MS	CEL, MS

Analysis	PP.H0	PP.H1	PP.H2	PP.H3	PP.H4	eta.hat	P-Value
T1D-RA	9.36E-01	4.43E-03	5.82E-02	2.67E-04	7.75E-04	1.13E+00	7.94E-01
T1D-CEL	2.78E-09	1.32E-11	9.83E-01	4.54E-03	1.23E-02	-2.97E+00	2.59E-01
T1D-MS	1.22E-02	3.21E-05	9.77E-01	2.50E-03	8.09E-03	1.76E+00	6.15E-01
RA-CEL	3.44E-08	2.12E-09	9.28E-01	5.71E-02	1.44E-02	-5.23E+00	4.27E-01
RA-MS	1.42E-02	1.03E-03	8.59E-01	6.18E-02	6.43E-02	2.93E+00	1.98E-01
CEL-MS	1.97E-12	1.66E-03	8.56E-11	6.26E-02	9.36E-01	-5.42E-01	4.59E-02

Location	Candidate Causal Genes	Existing Associations	Associations Found
7p14.2	*ELMO1*	CEL, MS	RA, CEL, MS

Analysis	PP.H0	PP.H1	PP.H2	PP.H3	PP.H4	eta.hat	P-Value
T1D-RA	6.45E-01	2.59E-03	3.49E-01	1.39E-03	1.52E-03	1.77E+00	4.25E-01
T1D-CEL	8.95E-10	2.17E-11	9.22E-01	2.18E-02	5.57E-02	2.01E+00	3.00E-01
T1D-MS	3.45E-09	1.37E-10	8.49E-01	3.24E-02	1.18E-01	2.22E+00	3.57E-01
RA-CEL	3.02E-10	1.95E-10	1.99E-01	1.23E-01	6.78E-01	1.32E+00	5.59E-01
RA-MS	1.11E-07	6.71E-08	2.31E-01	1.34E-01	6.35E-01	1.34E+00	5.84E-01
CEL-MS	1.58E-17	1.08E-08	4.89E-11	2.40E-02	9.76E-01	1.01E+00	2.93E-01

Location	Candidate Causal Genes	Existing Associations	Associations Found
7p12.2	*IKZF1*	MS	T1D, MS

Analysis	PP.H0	PP.H1	PP.H2	PP.H3	PP.H4	eta.hat	P-Value
T1D-RA	1.85E-01	8.10E-01	2.17E-04	9.17E-04	3.31E-03	8.08E-01	4.34E-01
T1D-CEL	5.44E-01	4.53E-01	1.38E-03	1.14E-03	1.48E-04	3.27E-01	9.53E-02
T1D-MS	4.52E-03	3.82E-03	2.65E-01	2.19E-01	5.08E-01	1.08E+00	1.96E-01
RA-CEL	9.97E-01	7.78E-04	2.34E-03	1.82E-06	9.76E-07	9.37E-01	1.02E-01
RA-MS	7.11E-03	5.17E-06	9.90E-01	7.02E-04	1.88E-03	1.57E+00	1.57E-01
CEL-MS	1.65E-02	2.82E-05	9.81E-01	1.66E-03	1.14E-03	2.12E+00	1.56E-01

Location	Candidate Causal Genes	Existing Associations	Associations Found
7p12.2	*IKZF1, FIGNL1*	T1D	T1D

Analysis	PP.H0	PP.H1	PP.H2	PP.H3	PP.H4	eta.hat	P-Value
T1D-RA	2.84E-02	9.71E-01	6.51E-06	2.20E-04	2.03E-04	-2.31E-01	2.74E-01
T1D-CEL	2.68E-02	9.72E-01	1.78E-05	6.37E-04	6.46E-04	3.26E-01	3.24E-01
T1D-MS	3.44E-02	9.64E-01	4.57E-05	1.28E-03	5.20E-05	-1.45E-01	1.30E-01
RA-CEL	9.99E-01	2.58E-04	5.93E-04	1.52E-07	1.46E-07	-5.75E-01	4.94E-01
RA-MS	9.98E-01	2.55E-04	1.31E-03	3.30E-07	4.15E-07	1.90E-01	1.98E-01
CEL-MS	9.99E-01	5.51E-04	9.11E-04	5.01E-07	2.03E-07	-1.64E+00	4.11E-01

Location	Candidate Causal Genes	Existing Associations	Associations Found
10p15.1	*IL2RA*	T1D, RA, MS	T1D|MS

Analysis	PP.H0	PP.H1	PP.H2	PP.H3	PP.H4	eta.hat	P-Value
T1D-RA	6.69E-27	6.85E-01	1.69E-27	1.72E-01	1.43E-01	4.27E-01	7.38E-01
T1D-CEL	1.06E-25	8.64E-01	1.65E-26	1.35E-01	1.13E-03	9.41E-02	6.50E-04
T1D-MS	5.84E-33	8.94E-08	6.57E-26	1.00E+00	2.54E-07	4.38E-01	5.04E-09
RA-CEL	6.18E-01	3.18E-01	3.61E-02	1.85E-02	9.32E-03	9.66E+00	1.93E-01
RA-MS	9.79E-08	4.72E-08	5.41E-01	2.58E-01	2.00E-01	2.10E+00	3.93E-01
CEL-MS	2.35E-09	9.60E-10	7.10E-01	2.90E-01	1.52E-04	-6.52E+00	9.38E-03

Location	Candidate Causal Genes	Existing Associations	Associations Found
11q23.3	*CXCR5*	RA, CEL, MS	CEL | MS

Analysis	PP.H0	PP.H1	PP.H2	PP.H3	PP.H4	eta.hat	P-Value
T1D-RA	9.92E-01	1.40E-03	6.75E-03	9.42E-06	1.34E-05	-4.08E-01	4.77E-01
T1D-CEL	1.53E-03	2.31E-06	9.97E-01	1.50E-03	1.20E-04	1.70E+01	2.70E-01
T1D-MS	2.85E-01	7.53E-04	7.12E-01	1.88E-03	1.08E-04	-2.36E+00	2.93E-01
RA-CEL	5.33E-03	7.83E-05	9.57E-01	1.38E-02	2.33E-02	4.14E+00	4.25E-01
RA-MS	2.41E-01	5.59E-03	7.29E-01	1.68E-02	8.02E-03	-1.34E-01	4.88E-01
CEL-MS	4.06E-05	1.70E-01	1.98E-04	8.27E-01	2.61E-03	3.06E-01	2.34E-02

Location	Candidate Causal Genes	Existing Associations	Associations Found
13q32.3	*GPR183*	T1D, MS	T1D | CEL

Analysis	PP.H0	PP.H1	PP.H2	PP.H3	PP.H4	eta.hat	P-Value
T1D-RA	4.70E-02	9.53E-01	3.40E-06	6.71E-05	1.84E-04	-3.51E-01	4.05E-01
T1D-CEL	7.97E-03	3.04E-01	1.75E-02	6.70E-01	3.92E-05	-2.68E-01	1.40E-05
T1D-MS	1.22E-02	8.14E-01	2.50E-03	1.67E-01	4.55E-03	-3.19E-02	4.12E-03
RA-CEL	4.47E-01	2.95E-05	5.53E-01	3.63E-05	2.97E-05	5.29E+00	3.79E-01
RA-MS	8.61E-01	8.06E-05	1.39E-01	1.29E-05	1.04E-05	3.56E+01	4.66E-01
CEL-MS	3.57E-01	4.97E-01	5.87E-02	8.16E-02	6.01E-03	1.20E-02	1.33E-02

Location	Candidate Causal Genes	Existing Associations	Associations Found
15q25.1	*CTSH*	T1D, MS	T1D | CEL

Analysis	PP.H0	PP.H1	PP.H2	PP.H3	PP.H4	eta.hat	P-Value
T1D-RA	3.35E-07	1.00E+00	1.14E-10	3.38E-04	1.50E-04	1.00E-01	4.55E-01
T1D-CEL	2.99E-08	1.83E-01	3.00E-09	1.02E-02	8.06E-01	-7.55E-01	4.33E-01
T1D-MS	1.36E-07	9.46E-01	6.47E-09	4.51E-02	9.03E-03	-3.51E-02	4.36E-02
RA-CEL	8.94E-01	1.98E-04	1.06E-01	2.33E-05	1.60E-05	-2.43E+01	3.17E-01
RA-MS	9.31E-01	1.56E-04	6.93E-02	1.16E-05	5.61E-06	5.45E+00	4.24E-01
CEL-MS	6.09E-01	3.49E-01	2.43E-02	1.38E-02	3.53E-03	2.05E-01	9.44E-02

Location	Candidate Causal Genes	Existing Associations	Associations Found
16p13.13	*DEXI, SOCS1*	T1D, CEL, MS	T1D, MS | CEL

Analysis	PP.H0	PP.H1	PP.H2	PP.H3	PP.H4	eta.hat	P-Value
T1D-RA	3.92E-07	9.99E-01	3.77E-10	9.60E-04	7.42E-05	-4.26E-02	4.02E-01
T1D-CEL	1.36E-10	4.88E-01	1.43E-10	5.12E-01	6.66E-05	-3.75E-01	1.85E-03
T1D-MS	6.28E-15	7.85E-06	3.29E-10	4.05E-01	5.95E-01	8.21E-01	2.47E-01
RA-CEL	4.65E-01	3.48E-04	5.34E-01	3.99E-04	3.16E-05	3.83E+00	1.39E-01
RA-MS	3.86E-05	2.38E-08	9.99E-01	6.17E-04	5.30E-05	5.56E+00	2.18E-01
CEL-MS	4.53E-06	2.58E-06	6.37E-01	3.63E-01	1.93E-04	-1.48E+00	3.76E-03

Location	Candidate Causal Genes	Existing Associations	Associations Found
18p11.21	*PTPN2*	T1D, CEL	T1D | CEL

Analysis	PP.H0	PP.H1	PP.H2	PP.H3	PP.H4	eta.hat	P-Value
T1D-RA	1.43E-10	9.99E-01	8.81E-14	6.10E-04	6.33E-04	4.54E-01	2.62E-02
T1D-CEL	1.83E-11	2.10E-02	5.11E-10	5.83E-01	3.96E-01	6.18E-01	8.90E-01
T1D-MS	2.93E-09	1.00E+00	4.79E-13	1.63E-04	6.45E-05	-2.38E-03	7.84E-01
RA-CEL	2.41E-02	1.18E-04	9.71E-01	4.78E-03	1.53E-04	1.41E+00	1.34E-02
RA-MS	9.99E-01	6.40E-04	8.46E-05	5.34E-08	7.96E-08	2.03E-01	3.24E-01
CEL-MS	1.76E-02	9.82E-01	1.99E-06	1.11E-04	5.60E-05	5.64E-02	2.30E-01

Location	Candidate Causal Genes	Existing Associations	Associations Found
19p13.2	*ICAM1, ICAM3, TYK2*	T1D, RA, MS	T1D, RA, CEL, MS

Analysis	PP.H0	PP.H1	PP.H2	PP.H3	PP.H4	eta.hat	P-Value
T1D-RA	2.63E-14	2.11E-07	1.21E-09	8.99E-146	1.00E+00	1.00E+00	8.22E-01
T1D-CEL	7.62E-09	1.28E-01	3.13E-08	5.25E-01	3.47E-01	8.84E-01	3.16E-03
T1D-MS	2.75E-11	1.50E-03	2.72E-10	5.22E-03	9.93E-01	6.57E-01	5.08E-01
RA-CEL	1.26E-06	1.25E-01	4.72E-06	4.69E-01	4.05E-01	8.91E-01	6.93E-02
RA-MS	7.23E-09	1.54E-03	7.36E-08	6.01E-03	9.92E-01	7.18E-01	4.78E-01
CEL-MS	7.51E-03	3.76E-02	7.61E-02	3.77E-01	5.01E-01	5.06E-01	2.80E-03

Location	Candidate Causal Genes	Existing Associations	Associations Found
21q22.3	*UBASH3A*	T1D, RA, CEL	T1D | RA | CEL

Analysis	PP.H0	PP.H1	PP.H2	PP.H3	PP.H4	eta.hat	P-Value
T1D-RA	1.12E-07	2.00E-01	4.33E-07	7.70E-01	3.00E-02	5.13E-01	2.24E-02
T1D-CEL	1.77E-10	7.78E-04	2.25E-07	9.88E-01	1.11E-02	8.16E-01	6.93E-03
T1D-MS	2.20E-07	1.00E+00	2.70E-11	1.22E-04	7.31E-05	1.26E-03	1.71E-01
RA-CEL	3.49E-04	8.43E-04	2.87E-01	6.93E-01	1.92E-02	1.53E+00	4.80E-02
RA-MS	4.29E-01	5.71E-01	3.79E-05	4.99E-05	6.12E-05	2.27E-01	4.84E-01
CEL-MS	2.37E-04	1.00E+00	1.83E-08	7.69E-05	3.51E-05	-1.96E-02	3.66E-01

Bibliography

Auton A, Abecasis GR, Altshuler DM, Durbin RM, Bentley DR, Chakravarti A, Clark AG, Donnelly P, Eichler EE, Flicek P, et al. 2015. A global reference for human genetic variation. *Nature* **526**: 68–74.

Barrett JC, Clayton DG, Concannon P, Akolkar B, Cooper JD, Erlich HA, Julier C, Morahan G, Nerup J, Nierras C, et al. 2009a. Genome-wide association study and meta-analysis find that over 40 loci affect risk of type 1 diabetes. *Nature genetics* **41**: 703–7.

Barrett JC, Hansoul S, Nicolae DL, Cho JH, Duerr RH, Rioux JD, Brant SR, Silverberg MS, Taylor KD, Michael M, et al. 2009b. Genome-wide association defines more than thirty distinct susceptibility loci for Crohn's disease. *Nat Genet.* **40**: 955–962.

Beecham AH, Patsopoulos NA, Xifara DK, Davis MF, Kemppinen A, Cotsapas C, Shah TS, Spencer C, Booth D, Goris A, et al. 2013. Analysis of immune-related loci identifies 48 new susceptibility variants for multiple sclerosis. *Nature genetics* **45**: 1353–60.

Begg CB and Gray R. 1984. Calculation of polychotomous logistic regression parameters using individualized regressions. *Biometrika* **71**: 11–18.

Bluestone Ja, Herold K, and Eisenbarth G. 2010. Genetics, pathogenesis and clinical interventions in type 1 diabetes. *Nature* **464**: 1293–1300.

Bowes J, Budu-Aggrey A, Huffmeier U, Uebe S, Steel K, Hebert HL, Wallace C, Massey J, Bruce IN, Bluett J, et al. 2015. Dense genotyping of immune-related susceptibility loci reveals new insights into the genetics of psoriatic arthritis. *Nature communications* **6**: 6046.

Bulik-Sullivan B, Finucane HK, Anttila V, Gusev A, Day FR, Consortium R, Genomics Consortium P, of the Wellcome Trust Consortium GCfA, Perry JR, Patterson N, et al. 2015. An Atlas of Genetic Correlations across Human Diseases and Traits. *bioRxiv* **47**: 1–44.

Burren OS, Guo H, and Wallace C. 2014. VSEAMS: a pipeline for variant set enrichment analysis using summary GWAS data identifies IKZF3, BATF and ESRRA as key transcription factors in type 1 diabetes. *Bioinformatics (Oxford, England)* **30**: 3342–8.

Burton PR, Clayton DG, Cardon LR, Craddock N, Deloukas P, Duncanson A, Kwiatkowski DP, McCarthy MI, Ouwehand WH, Samani NJ, et al. 2007. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature* **447**: 661–678.

Carbonetto P and Stephens M. 2012. Scalable variational inference for bayesian variable selection in regression, and its accuracy in genetic association studies. *Bayesian Analysis* **7**: 73–108.

Coles AJ. 2012. Alemtuzumab Therapy for Multiple Sclerosis. *Neurotherapeutics : the journal of the American Society for Experimental NeuroTherapeutics* **10**: 0–4.

Cooper JD, Simmonds MJ, Walker NM, Burren O, Brand OJ, Guo H, Wallace C, Stevens H, Coleman G, Franklyn JA, et al. 2012. Seven newly identified loci for autoimmune thyroid disease. *Human molecular genetics* **21**: 5202–8.

Cooper NJ, Shtir CJ, Smyth DJ, Guo H, Swafford AD, Zanda M, Hurles ME, Walker NM, Plagnol V, Cooper JD, et al. 2015. Detection and correction of artefacts in estimation of rare copy number variants and analysis of rare deletions in type 1 diabetes. *Hum Mol Genet* **24**: 1774–1790.

Cortes A and Brown Ma. 2011. Promise and pitfalls of the Immunochip. *Arthritis research & therapy* **13**: 101.

Cotsapas C and Hafler DA. 2013. Immune-mediated disease genetics: The shared basis of pathogenesis. *Trends in Immunology* **34**: 22–26.

Davison LJ, Wallace C, Cooper JD, Cope NF, Wilson NK, Smyth DJ, Howson JMM, Saleh N, Al-Jeffery A, Angus KL, et al. 2012. Long-range DNA looping and gene expression analyses identify DEXI as an autoimmune disease candidate gene. *Human Molecular Genetics* **21**: 322–333.

Deelen P, Menelaou A, van Leeuwen EM, Kanterakis A, van Dijk F, Medina-Gomez C, Francioli LC, Hottenga JJ, Karssen LC, Estrada K, et al. 2014. Improved imputation quality of low-frequency and rare variants in European samples using the 'Genome of The Netherlands'. *European journal of human genetics : EJHG* **22**: 1321–6.

Dudbridge F and Gusnanto A. 2008. Estimation of significance thresholds for genomewide association scans. *Genetic epidemiology* **32**: 227–34.

Dupuis J, Langenberg C, Prokopenko I, Saxena R, Soranzo N, Jackson AU, Wheeler E, Glazer NL, Bouatia-Naji N, Gloyn AL, et al. 2010. New genetic loci implicated in fasting glucose homeostasis and their impact on type 2 diabetes risk. *Nature genetics* **42**: 105–16.

ENCODE Project Consortium, Bernstein BE, Birney E, Dunham I, Green ED, Gunter C, and Snyder M. 2012. An integrated encyclopedia of DNA elements in the human genome. *Nature* **489**: 57–74.

Evangelou M, Smyth DJ, Fortune MD, Burren OS, Walker NM, Guo H, Onengut-Gumuscu S, Chen WM, Concannon P, Rich SS, et al. 2014. A Method for gene-based pathway analysis using genomewide association study summary statistics reveals nine new type 1 diabetes associations. *Genetic Epidemiology* **38**: 661–670.

Eyre S, Bowes J, Diogo D, Lee A, Barton A, Martin P, Zhernakova A, Stahl E, Viatte S, McAllister K, et al. 2012. High-density genetic mapping identifies new susceptibility loci for rheumatoid arthritis. *Nature genetics* **44**: 1336–40.

Farh KKH, Marson A, Zhu J, Kleinewietfeld M, Housley WJ, Beik S, Shoresh N, Whitton H, Ryan RJH, Shishkin AA, et al. 2015. Genetic and epigenetic fine mapping of causal autoimmune disease variants. *Nature* **518**: 337–43.

Fieller EC. 1954. Some Problems in Interval Estimation. *Journal of the Royal Statistical Society. Series B (Methodological)* **16**: 175–185.

Flutre T, Wen X, Pritchard J, and Stephens M. 2013. A Statistical Framework for Joint eQTL Analysis in Multiple Tissues. *PLoS Genetics* **9**: e1003486.

Fortune MD, Guo H, Burren O, Schofield E, Walker NM, Ban M, Sawcer SJ, Bowes J, Worthington J, Barton A, et al. 2015. Statistical colocalization of genetic risk variants for related autoimmune diseases in the context of common controls. *Nature genetics* **47**: 839–846.

Fung EYMG, Smyth DJ, Howson JMM, Cooper JD, Walker NM, Stevens H, Wicker LS, and Todd JA. 2009. Analysis of 17 autoimmune disease-associated variants in type 1 diabetes identifies 6q23/TNFAIP3 as a susceptibility locus. *Genes and immunity* **10**: 188–91.

Giambartolomei C, Vukcevic D, Schadt EE, Franke L, Hingorani AD, Wallace C, and Plagnol V. 2014. Bayesian test for colocalisation between pairs of genetic association studies using summary statistics. *PLoS genetics* **10**: e1004383.

Guo H, Fortune MD, Burren OS, Schofield E, Todd Ja, and Wallace C. 2015. Integration of disease association and eQTL data using a Bayesian colocalisation approach highlights six candidate causal genes in immune-mediated diseases. *Human molecular genetics* **24**: 3305–3313.

Gusev A, Ko A, Shi H, Bhatia G, Chung W, Penninx BWJH, Jansen R, de Geus EJC, Boomsma DI, Wright FA, et al. 2016. Integrative approaches for large-scale transcriptome-wide association studies. *Nature genetics* **48**: 245–52.

Haines JL, Hauser Ma, Schmidt S, Scott WK, Olson LM, Gallins P, Spencer KL, Kwan SY, Noureddine M, Gilbert JR, et al. 2005. Complement factor H variant increases the risk of age-related macular degeneration. *Science (New York, N.Y.)* **308**: 419–421.

Harder MN, Ribel-Madsen R, Justesen JM, Sparsø T, Andersson EA, Grarup N, Jørgensen T, Linneberg A, Hansen T, and Pedersen O. 2013. Type 2 diabetes risk alleles near BCAR1 and in ANK1 associate with decreased β-cell function whereas risk alleles near ANKRD55 and GRB14 associate with decreased insulin sensitivity in the Danish Inter99 cohort. *The Journal of clinical endocrinology and metabolism* **98**: E801–6.

Hormozdiari F, Kostem E, Kang EY, Pasaniuc B, and Eskin E. 2014. Identifying causal variants at loci with multiple signals of association. *Genetics* **198**: 497–508.

Huang H, Fang M, Jostins L, Mirkov MU, Boucher G, Anderson CA, Andersen V, Cleynen I, Cortes A, Crins F, et al. 2015. Association mapping of inflammatory bowel disease loci to single variant resolution. *bioRxiv* p. 028688.

Hunt KA, Mistry V, Bockett NA, Ahmad T, Ban M, Barker JN, Barrett JC, Blackburn H, Brand O, Burren O, et al. 2013. Negligible impact of rare autoimmune-locus coding-region variants on missing heritability. *Nature* **498**: 232–235.

International Consortium for Blood Pressure Genome-Wide Association Studies. 2012. Genetic Variants in Novel Pathways Influence Blood Pressure and Cardiovascular Disease Risk. *Nature* **478**: 103–9.

Ioannidis JPA. 2008. Why most discovered true associations are inflated. *Epidemiology (Cambridge, Mass.)* **19**: 640–648.

Jostins L, Ripke S, Weersma RK, Duerr RH, McGovern DP, Hui KY, Lee JC, Schumm LP, Sharma Y, Anderson CA, et al. 2012. Host-microbe interactions have shaped the genetic architecture of inflammatory bowel disease. *Nature* **491**: 119–24.

Kichaev G, Yang WY, Lindstrom S, Hormozdiari F, Eskin E, Price AL, Kraft P, and Pasaniuc B. 2014. Integrating Functional Data to Prioritize Causal Variants in Statistical Fine-Mapping Studies. *PLoS Genetics* **10**: e1004722.

Laukens D, Georges M, Libioulle C, Sandor C, Mni M, Cruyssen BV, Peeters H, Elewaut D, and de Vos M. 2010. Evidence for Significant Overlap between Common Risk Variants for Crohn's Disease and Ankylosing Spondylitis. *PLoS ONE* **5**: e0013795.

Maller JB, McVean G, Byrnes J, Vukcevic D, Palin K, Su Z, Howson JMM, Auton A, Myers S, Morris A, et al. 2012. Bayesian refinement of association signals for 14 loci in 3 common diseases. *Nature genetics* **44**: 1294–301.

Martin P, Mcgovern A, Orozco G, Duffus K, Yarwood A, Cooper N, Barton A, Wallace C, Fraser P, Worthington J, et al. 2015. Chromosome interaction analysis of risk loci in related autoimmune diseases reveals complex , long-range promoter interactions implicating novel candidate genes. *Nature Communications* **6**: 1–17.

Maurano MT, Humbert R, Rynes E, Thurman RE, Haugen E, Wang H, Reynolds AP, Sandstrom R, Qu H, Brody J, et al. 2012. Systematic Localization of Common Disease-Associated Variation in Regulatory DNA. *Science* **337**: 1190–1195.

Mero IL, Lorentzen AR, Ban M, Smestad C, Celius EG, Aarseth JH, Myhr KM, Link J, Hillert J, Olsson T, et al. 2010. A rare variant of the TYK2 gene is confirmed to be associated with multiple sclerosis. *European journal of human genetics : EJHG* **18**: 502–4.

Morris AP, Voight BF, Teslovich TM, Ferreira T, Segrè AV, Steinthorsdottir V, Strawbridge RJ, Khan H, Grallert H, Mahajan A, et al. 2012. Large-scale association analysis provides insights into the genetic architecture and pathophysiology of type 2 diabetes. *Nature genetics* **44**: 981–90.

Nalls MA, Couper DJ, Tanaka T, van Rooij FJA, Chen MH, Smith AV, Toniolo D, Zakai NA, Yang Q, Greinacher A, et al. 2011. Multiple loci are associated with white blood cell phenotypes. *PLoS Genetics* **7**: e1002113.

Nogueira TC, Paula FM, Villate O, Colli ML, Moura RF, Cunha DA, Marselli L, Marchetti P, Cnop M, Julier C, et al. 2013. GLIS3, a susceptibility gene for type 1 and type 2 diabetes, modulates pancreatic beta cell apoptosis via regulation of a splice variant of the BH3-only protein Bim. *PLoS genetics* **9**: e1003532.

Okada Y, Wu D, Trynka G, Raj T, Terao C, Ikari K, Kochi Y, Ohmura K, Suzuki A, Yoshida S, et al. 2014. Genetics of rheumatoid arthritis contributes to biology and drug discovery. *Nature* **506**: 376–81.

Onengut-Gumuscu S, Chen WM, Burren O, Cooper NJ, Quinlan AR, Mychaleckyj JC, Farber E, Bonnie JK, Szpak M, Schofield E, et al. 2015. Fine mapping of type 1 diabetes susceptibility loci and evidence for colocalization of causal variants with lymphoid gene enhancers. *Nature genetics* **47**: 381–6.

Panagiotou OA and Ioannidis JPA. 2012. What should the genome-wide significance threshold be? Empirical replication of borderline genetic associations. *International journal of epidemiology* **41**: 273–86.

Parkes M, Cortes A, van Heel DA, and Brown MA. 2013. Genetic insights into common pathways and complex relationships among immune-mediated diseases. *Nat Rev Genet* **14**: 661–673.

Peters JE, Lyons PA, Lee JC, Richard AC, Fortune MD, Newcombe PJ, Richardson S, and Smith KGC. 2016. Insight into Genotype-Phenotype Associations through eQTL Mapping in Multiple Cell Types in Health and Immune-Mediated Disease. *PLoS genetics* **12**: e1005908.

Pickrell J, Berisa T, Segurel L, Tung JY, and Hinds D. 2016. Detection and interpretation of shared genetic influences on 40 human traits. *Nature genetics* **48**: 709–717.

Pickrell JK. 2014. Joint analysis of functional genomic data and genome-wide association studies of 18 human traits. *American journal of human genetics* **94**: 559–73.

Plagnol V, Smyth DJ, Todd JA, and Clayton DG. 2009. Statistical independence of the colocalized association signals for type 1 diabetes and RPS26 gene expression on chromosome 12q13. *Biostatistics* **10**: 327–334.

Prentice RL and Pyke R. 1979. Logistic disease incidence models and case-control studies. *Biometrika* **66**: 403–411.

Raftery AE. 1996. Approximate Bayes factors and accounting for model uncertainty in generalised linear models. *Biometrika* **83**: 251 –266.

Risch N. 1987. Assessing the role of HLA-linked and unlinked determinants of disease. *American journal of human genetics* **40**: 1–14.

Rodriguez-Calvo T, Sabouri S, Anquetil F, and von Herrath MG. 2016. The viral paradigm in type 1 diabetes: Who are the main suspects? *Autoimmunity Reviews* **15**: 964–969.

Rubin DB. 1984. Bayesianly Justifiable and Relevant Frequency Calculations for the Applies Statistician. *The Annals of Statistics* **12**: 1151–1172.

Scott RA, Magi R, Morris AP, Marullo L, Gaulton K, Boehnke M, Dupuis J, McCarthy MI, Scott LJ, Prokopenko I, et al. 2014. Genome-wide association study imputed to 1000 Genomes reveals 18 novel associations with type 2 diabetes. In *American Society of Human Genetics*.

Selmi C, Lu Q, and Humble MC. 2012. Heritability versus the role of the environment in autoimmunity. *Journal of Autoimmunity* **39**: 249–252.

Sinnott JA and Kraft P. 2012. Artifact due to differential error when cases and controls are imputed from different platforms. *Human Genetics* **131**: 111–119.

Smyth DJ, Plagnol V, Walker NM, Cooper JD, Downes K, Yang JHM, Howson JMM, Stevens H, McManus R, Wijmenga C, et al. 2008. Shared and distinct genetic variants in type 1 diabetes and celiac disease. *The New England journal of medicine* **359**: 2767–2777.

Soedamah-Muthu SS, Fuller JH, Mulnier HE, Raleigh VS, Lawrenson RA, and Colhoun HM. 2006. All-cause mortality rates in patients with type 1 diabetes mellitus compared with a non-diabetic population from the UK general practice research database, 1992-1999. *Diabetologia* **49**: 660–666.

Speed D, Hemani G, Johnson MR, and Balding DJ. 2012. Improved heritability estimation from genome-wide SNPs. *American Journal of Human Genetics* **91**: 1011–1021.

Sunnåker M, Busetto AG, Numminen E, Corander J, Foll M, and Dessimoz C. 2013. Approximate Bayesian Computation. *PLoS Computational Biology* **9**: e1002803.

Swafford ADE, Howson JMM, Davison LJ, Wallace C, Smyth DJ, Schuilenburg H, Maisuria-Armer M, Mistry T, Lenardo MJ, and Todd JA. 2011. An allele of IKZF1 (Ikaros) conferring susceptibility to childhood acute lymphoblastic leukemia protects against type 1 diabetes. *Diabetes* **60**: 1041–4.

Tavaré S, Balding DJ, Griffiths RC, and Donnelly P. 1997. Inferring coalescence times from DNA sequence data. *Genetics* **145**: 505–518.

Teslovich, T Musunuru, K Smith aEA. 2010. Biological, clinical and population relevance of 95 loci for blood lipids. *Nature* **466**: 707–713.

Trynka G, Hunt KA, Bockett NA, Romanos J, Castillejo G, Concha EGD, and Almeida RCD. 2012. Dense genotyping identifies and localizes multiple common and rare variant association signals in celiac disease. *Nature Genetics* **43**: 1193–1201.

Trynka G, Westra HJ, Slowikowski K, Hu X, Xu H, Stranger B, Klein R, Han B, and Raychaudhuri S. 2015. Disentangling the Effects of Colocalizing Genomic Annotations to Functionally Prioritize Non-coding Variants within Complex-Trait Loci. *The American Journal of Human Genetics* **97**: 139–152.

Ueda H, Howson JMM, Esposito L, Heward J, Snook H, Chamberlain G, Rainbow DB, Hunter KMD, Smith AN, Di Genova G, et al. 2003. Association of the T-cell regulatory gene CTLA4 with susceptibility to autoimmune disease. *Nature* **423**: 506–11.

Visscher PM, Brown MA, McCarthy MI, and Yang J. 2012. Five years of GWAS discovery. *American journal of human genetics* **90**: 7–24.

Wakefield J. 2009. Bayes factors for genome-wide association studies: comparison with P-values. *Genetic epidemiology* **33**: 79–86.

Waldron-Lynch F, Kareclas P, Irons K, Walker NM, Mander A, Wicker LS, Todd JA, and Bond S. 2014. Rationale and study design of the Adaptive study of IL-2 dose on regulatory T cells in type 1 diabetes (DILT1D): a non-randomised, open label, adaptive dose finding trial. *BMJ open* **4**: e005559.

Wallace C. 2013. Statistical testing of shared genetic control for potentially related traits. *Genetic Epidemiology* **37**: 802–813.

Wallace C, Cutler AJ, Pontikos N, Pekalski ML, Burren OS, Cooper JD, Rubio Garcia A, Ferreira RC, Guo H, Walker NM, et al. 2015. Dissection of a complex disease susceptibility region using a Bayesian stochastic search approach to fine mapping. *bioRxiv* **56382813**: 015164.

Wallace C, Rotival M, Cooper JD, Rice CM, Yang JHM, McNeill M, Smyth DJ, Niblett D, Cambien F, Tiret L, et al. 2012. Statistical colocalisation of monocyte gene expression and genetic risk variants for type 1 diabetes. *Human molecular genetics* **44**: 1–35.

Walter K, Min JL, Huang J, Crooks L, Memari Y, McCarthy S, Perry JRB, Xu C, Futema M, Lawson D, et al. 2015. The UK10K project identifies rare variants in health and disease. *Nature* **526**: 82–90.

Zhu Z, Zhang F, Hu H, Bakshi A, Robinson MR, Powell JE, Montgomery GW, Goddard ME, Wray NR, Visscher PM, et al. 2016. Integration of summary data from GWAS and eQTL studies predicts complex trait gene targets. *Nature genetics* **48**: 481–7.