Simple Bayesian approaches to modelling pleiotropy in genetic association data

Holly Trochet

Pembroke College

University of Oxford

A thesis submitted for the degree of

Doctor of Philosophy

Trinity Term, 2017

Acknowledgments

This thesis would not have been possible without the enthusiasm and guidance of my supervisors, Chris Spencer and Gil McVean. Additional support was generously provided by Matti Pirinen and Luke Jostins. I hope Luke enjoyed our one-on-one meetings as much as I did.

I've made a number of friends at Oxford, but the support of Clare Bycroft, Na Cai, and Ryan Christ has been integral to my time here. I hope to see them all again, however scattered we are around the globe. Many thanks to Julie Hussin, as well, for her support, and for giving me something to look forward to in my post-Oxford life.

And last, but certainly not least, I thank my husband, Malcolm Fyfe, for his love, patience, and cooking, as well as for the proofreading he's done on this thesis (including these acknowledgments). I look forward to getting to spend some time with him again and hope that once I have, our marriage will still seem like a good idea.

Simple Bayesian approaches to modelling pleiotropy in genetic association data

Holly Trochet
Pembroke College

Doctor of Philosophy
Trinity Term 2017

Abstract

Genome-wide association studies (GWAS) have become one of the most common types of genetic studies, due to their success at finding markers associated with diseases and other traits of interest. With thousands of these studies performed in the last decade, a variety of methods have been developed to meta-analyze their results to search for variants that affect multiple traits. This thesis introduces one such method, based on approximate Bayes factors (ABFs), which relies only on effect size estimates and standard errors from GWAS. Through application to simulated data as well as three different datasets, we demonstrate the statistical properties, strengths, and limitations of our approach.

We show that only can this method be applied to the meta-analysis of a single trait across multiple studies, but because it does does not make strong assumptions about the similarity of effect sizes across traits, it can also be used to detect effects across multiple traits at a single marker. Additionally, it can account for confounding due to things like shared samples and can be applied exhaustively across all possible combinations of associations to determine the subset of traits or studies that are most likely to be associated with a given variant. This affords the opportunity to make statements about which traits explicitly are and are not associated with a marker, and these patterns can be explored over the whole genome to learn about the genetic relationships between different traits. We also discuss some of the individual markers highlighted by our analyses—some known, and some potentially novel—and the traits associated with them.

Contents

Chapter 1

Introduction

The proliferation of genome-wide association studies (GWAS) in the past decade has yielded many novel genotype-phenotype associations. In meta-analyzing these studies, there is an opportunity to expand our knowledge of biology even further by searching for loci that affect multiple traits. However, in order to combine studies across multiple data sources that have been collected for different purposes, we need the proper statistical tools. This thesis discusses some methods that have been proposed, and compares them to a novel method that uses approximate Bayes factors to perform both a genome-wide scan for variants associated with multiple phenotypes, and which can be used to find the subset of phenotypes that is the most likely to be associated with a given genetic marker.

The field of genetics has always had a close relationship to statistics. Gregor Mendel, an Augustinian friar, deduced his principles of inheritance in the mid 19th century from observing the phenotypes of various crosses of the pea *Pisum sativum* and understood that observed ratios would not match the expected ones exactly because of random sampling. His work went unnoticed until the early 20th century, when biologists became concerned with how new species could arise from the "blending" model of heredity favored by Darwin and his contemporaries [1]—if the traits of the offspring are some sort of midpoint of those of its parents, then over multiple generations, the variation in a population will decrease, making it less likely for speciation to occur. And yet, as Galton observed in his own experiments—during which he contributed to the development linear regression—the phenotype of offspring often shows reversion to the population mean, even when the phenotypes of its parents are

at the extremes [2]. Fisher resolved this debate in 1918 by showing that a trait controlled by multiple genes (or "factors") would display the blending observed by Galton and the other biometricians—including Pearson, another noted statistican—while still obeying Mendel's laws [3].

In this time, the first linkage studies were performed on *Drosophila* and the arrangement of genes on chromosomes had been identified as being part of the mechanism by which traits were transferred from one generation to the next [4]. However, it was not until 1952 that DNA was confirmed as the molecular basis of heredity [5], and its structure was characterized the following year [6]. It was nearly another 50 years before the complete human genome was successfully sequenced [7] and researchers could learn about how genomes differ on both an individual and population level through the HapMap and 1000 Genomes projects [8, 9]. Among other things, these projects showed that the linkage disequilibrium between variants across the genome could be leveraged to perform genome-wide studies without having to sequence every base pair. This insight led to the creation SNP (single nucelotide polymorphism) chips, which allow researchers to interrogate markers across the genome for association with a trait or phenotype (see Sections 1.1 and 1.4 for further discussion of genome-wide association studies).

Prior to this influx of information about the genome as a whole, researchers interested in the genetic basis of a trait had to rely on linkage studies, which require willing families and pedigrees. In this study design, participants are genotyped at a low-density set of markers in order to find a region of the genome that is associated with the trait of interest. In addition to often being unable to identify causal variants due to the low density of the markers, this type of study often lacks sufficient power to discover regions associated with traits that have a complex genetic architecture, due to small sample sizes and small effect sizes of associated markers. However, when the phenotype of interest is highly penetrant so that multiple relatives are affected, it is possible for the causal gene and variants to be identified, as it was for Huntington's disease [10].

Candidate gene studies were another popular study design, in which participants were sequenced at genes—or regions of the genome—whose function was thought to be relevant. There were some successes: for instance, the *BRCA1* and *BRCA2* genes

associated with breast cancer were found in this manner [11, 12]. However, results published from these types of studies were often not replicable and required the sort of knowledge that genetic research hoped to uncover: namely, gene function and disease etiology [13].

Indeed, the interest in genetics stems from a desire to further our understanding of biology. One's sequence of DNA nucleotides tells one's cells what proteins to make and how much. These proteins interact with one another to form the biological pathways within and between cells, eventually building up to the whole organism. If we can understand the genetic instructions that influence the end trait, then we can gain an insight into what proteins, and what pathways related to those proteins, result in a disease or phenotype. Using genetics is particularly attractive because we know the direction of causation—that is, we are certain that the sequence of germline DNA tells the cell what proteins to make and not that the cell makes proteins which then determine the DNA sequence.

This chapter will provide an overview of a commonly used study design—the genome-wide association study (GWAS)—and the concepts of pleiotropy and cross-trait or cross-phenotype associations. We then will discuss meta-analysis in the context of GWAS and cross-trait associations, and summarize the variety of methods used to discover these associations.

1.1 Genome-wide association studies

There are two ways of conceptualizing a genome wide association study (GWAS). The first says that if a marker is associated with the phenotype of interest, then the genotype frequencies at that marker will differ between people who have the phenotype and those who do not. Alternatively, one can understand this as the proportion of cases and controls differing across genotypes. These can be extended to continuous traits as well. In that case, we expect to see that individuals who have a low phenotype measurements tend to have one of the SNP's alleles more often than do people with high phentoype measurements, or—equivalently—that people with a particular genotype at the SNP will disproportionately have a high or low phenotype measurement.

One of the first genome-wide association studies was published in 2005 in *Science*. It compared the allele frequencies of 103,611 single nucleotide polymorphisms (SNPs) in 96 people with age-related macular degeneration and 50 people without. The authors found that people who were homozygous for one of the alleles of a SNP in an intron of the complement factor H gene were seven times as likely to be cases than they were to be controls compared to those who had at least one copy of the other allele [14]. While the study size and number of single nucleotide polymorphisms seem small by today's standards, much of the study design and analysis has remained the same 12 years on. See Section 1.4 for a fuller treatment of how a standard GWAS is conducted.

With genotyping made cheaper and easier thanks to commercially available genotyping arrays, such as Illumina and Affymetrix, the number of genome-wide association studies has exploded over the years (Figure 1.1), looking at everything from eye color to risk of schizophrenia.

1.1.1 Linkage disequilibrium

A natural consequence of the biology of gamete formation is that large sections of DNA are inherited together from one generation to the next. This is due to "crossing over" or "recombination" of chromosomes during the metaphase stage of meiosis, the process by which haploid gametes are formed in a diploid organism. A chromosome contained in the gamete is a mosaic of a pair of homologous chromosomes in the parent. In a given population, when genotypes at one variant are correlated with genotypes at another, the two variants are said to be in "linkage disequilibrium" (LD) with each other [15]. LD between markers is chiefly characterized by two different, but related measures: $D\prime$ and r^2. Both depend on D, which is the difference between the observed frequency of a given haplotype of markers and the expected frequency assuming the independence of the loci. Given two biallelic markers, the absolute value of D is the same across all possible haplotypes. $D\prime$ is D, scaled by the maximum theoretical value of D based on the allele frequencies at each marker, while r^2 is D divided by the product of the frequencies of all alleles at each marker, all squared [16]. When $D\prime = 1$, the loci in question are said to be in "complete LD", meaning that they are associated as they can be, given potentially differing allele frequencies at the two sites.

Meanwhile, when $r^2 = 1$, the loci are in "perfect LD", as this is only possible when the allele frequencies at the sites are the same, allowing the genotypes at the two markers to be perfectly correlated [17].

The existence of linkage disequilibrium across the genome allows genetic association studies to be performed without requiring that every base in the genome be genotyped [8, 9]. Population-specific patterns of LD can be leveraged to increase the number of markers in the GWAS beyond those that were genotyped directly (see Section 1.4.2 below). While LD means that genotyping can be done without whole genome sequencing—which is still too expensive to be performed with the desired accuracy on a large number of individuals at once [18]—it also means that the marker with the most evidence for association with a trait (the "lead marker") is not necessarily the one that is causal. It may be that the true causal variant has not been typed or that the lead marker is in strong LD with the causal marker (as well as others), which was not the lead SNP due to sampling variation reducing the estimated effect of the causal variant below that of one of the others. This uncertainty necessitates follow-up analyses in order to resolve the causal locus. This can involve genotyping the region at a denser set of markers [19] or sequencing just the area around the association signal [20], integrating GWAS associations with genome annotations [21], or using statistical methods to model different patterns of causality in the region [22, 23, 24].

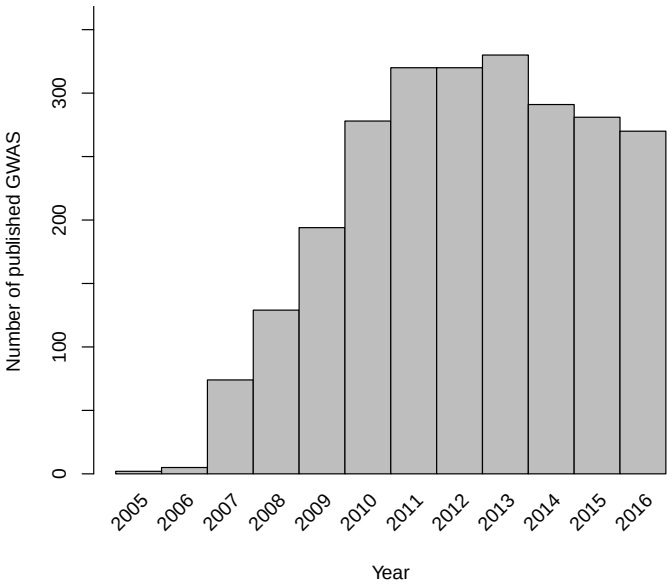


Figure 1.1: Plot of the number of genome-wide association studies published per year, recorded in the GWAS catalog [25].

The Wellcome Trust Case Control Consortium (WTCCC1) was formed in 2005

to take advantage of the SNP chip technology and this emerging GWAS study design. The traits chosen for this set of studies were bipolar disorder, coronary artery disease, Crohn's disease, hypertension, rheumatoid arthritis, type 1 diabetes, and type 2 diabetes. Each of the seven studies comprised about 2,000 cases and about 3,000 controls, the latter of which were shared across all studies. Half of the controls came from the 1958 British Birth Cohort and the other half were individuals who were recruited from donors to the UK Blood Services. A further study was done using 1,500 of the common controls with 1,000 cases of each of the following diseases: breast cancer, multiple sclerosis, ankylosing spondylitis, and autoimmune thyroid disease. Additionally, the WTCCC1 conducted a study on tuberculosis in The Gambia (1,500 cases and 1,500 controls) [26].

Despite the proliferation of GWAS, the SNPs found to be significantly associated with a disease or phenotype account for only a small proportion of the genetic variance observed from heritability studies; this is the "missing heritability" problem [27, 28]. One explanation for this phenomenon is that the heritability is not actually missing in the sense that the genetic markers responsible have not been typed, but that the traits—especially common and complex ones, such as height or complex diseases—are controlled by a large number of loci with very small effect sizes that a GWAS is not powerful enough to pick up [29]. Another hypothesis is that GWASs often exclude rare variants and so some of the missing heritability may come from rare variants (minor allele frequency $\leq 0.5\%$) with large (or even moderate) effects [30]. Still others claim that the problem might stem from overestimation of the heritability of complex traits, since most GWASs are performed with the assumption that all effects are additive [31].

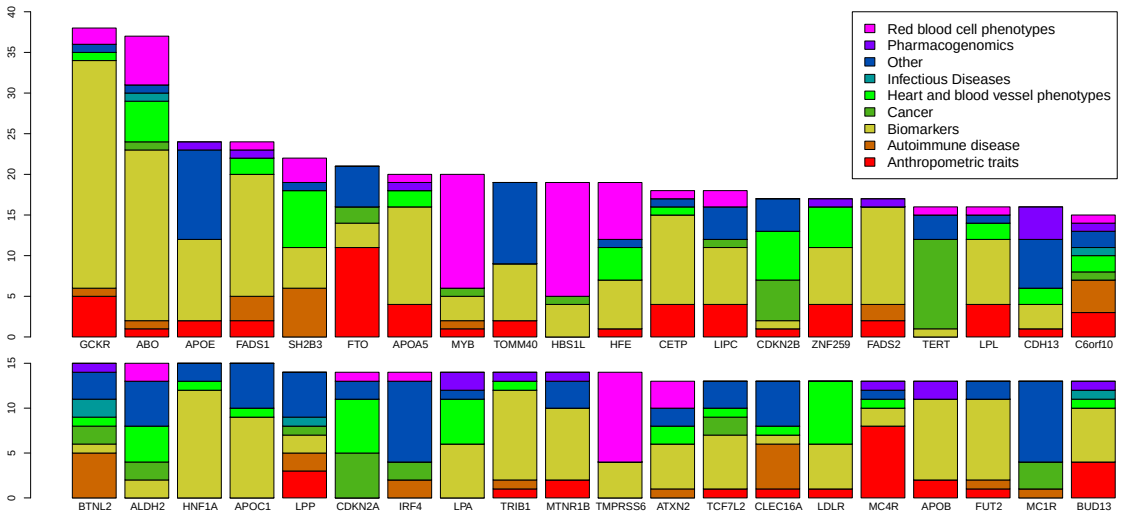
Even though genome-wide association studies have failed to provide full explanations for the genetic architecture of complex traits, they have uncovered very useful and important clues about disease etiology and human biology. Examples range from the original GWAS, which showed the link between complement factor H in age-related macular degeneration [14], to the discovery of the associations of variants in *FTO* with obesity [32, 33]. Furthermore, Okada *et al.* [34] showed that even when effect sizes are not large, genes linked to rheumatoid arthritis in GWAS correlate with previously established drug targets for the disease, suggesting that novel findings from

genome-wide association studies may pave the way to new pharmaceutical treatments.

1.2 Cross-phenotype associations and pleiotropy

Genome-wide association studies tend to look at a single trait, or small number of highly related traits—for instance, intraocular pressure and glaucoma, or high- and low-density lipoprotein levels and cardiovascular disease. However, with the large number of studies that have been performed in the last 10 years, there is now an opportunity to look across studies for variants that are associated across traits. Indeed, given finite number of genes in the human genome—current estimates put the figure at somewhere between 19,000 and 21,000 [35, 36] and the nearly limitless number of phenotypes, logic dictates that some genes must affect more than one trait. We see from Figure 1.2 that many genes have associations with multiple classes of traits, never mind separate traits within a class.

Figure 1.2: Barplot of the number of traits associated with the top 40 non-HLA genes, colored by the class of traits to which they belong. Genes are ranked by the number of associated traits. The data were taken from the GWAS catalog [25] in October 2014.



Often markers associated with a trait are assumed to affect the gene nearest to them; however, this may not be the case. For example, the gene *SH2B3* has many associations with blood pressure phenotypes in Europeans. However, it is in a region of high linkage disequilibrium that includes other genes, one of which is *ALDH2* [37], whose putative variants have been associated with blood pressure and alcohol intake in East Asians [38]. It may be that the gene affecting blood pressure does not

actually differ between the two populations, and that the different patterns of LD in East Asians and Europeans are responsible for observed associations with different markers and genes.

Similarly, *FTO* is associated with many traits, especially those related to obesity. However, there is evidence that markers which are contained in an intron of this gene—and which would therefore usually be assumed to be affecting it—actually affect the expression of *IRX3*, which is about half a megabase downstream [39]. This is not to imply that none of the associations with *FTO* (or any other gene) are true. We only highlight this and the *SH2B3*/*ALDH2* cases to show that some of the associations in Figure 1.2 are spurious because it is not always the case that the gene affected by a marker is the one closest to it. There is much that is unknown about the way different parts of the genome interact with one another.

Pleiotropy has many definitions depending on context. The most general one, given by Paaby and Rockman [40], is that "pleiotropy implies a mapping from one thing at the genetic level to multiple things at a phenotypic level." This is both true and too vague to be useful. For our purposes, we will define it as the phenomenon where a genetic variant influences more than one trait or phenotype. This can be subdivided into whether or not the variant affects associated traits through different metabolic pathways (called mosaic or horizontal pleiotropy, see 1.3a), in contrast to relational or vertical pleiotropy, where the variant acts on multiple traits through a single pathway (Figure 1.3b) [40, 41]. Some feel very strongly that the latter type of pleiotropy is not true pleiotropy. Because genome-wide association studies link only genetic markers and traits, without functional follow-up analyses in cell lines or model organisms, it is impossible to differentiate between these two types of pleiotropy from association studies alone.

1.2.1 Colocalization

Another reason to exercise caution is that markers found to be associated with a trait in GWAS may not be causal for the phenotype, but in linkage disequilibrium (LD, see Section 1.1.1) with an untyped causal variant; this need not be another SNP, but could be some sort of structural variant. If a SNP happens to be in LD with two such variants that are causal for different phenotypes, then there may be the appearance

Figure 1.3

(a) Schematic of horizontal/mosaic pleiotropy. The marker affects two different metabolic pathways, which in turn, affect two different phenotypes.

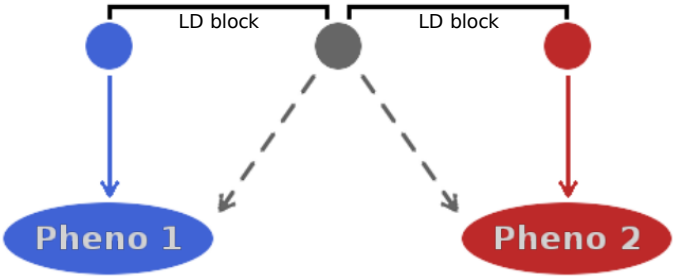
(b) Schematic of vertical/relational pleiotropy. The marker affects one metabolic pathway, which in turn, affects two different phenotypes.

of pleiotropy at the SNP when in fact none exists (see Figure 1.4 for a diagram). In recent years, methods have been developed to determine whether two signals from different GWASs "colocalize"—that is, to see if their respective association signals are the result of the same causal variant [42, 43, 44]. It is also possible that the reverse could happen—a pleiotropic variant may be tagged by different markers that in different studies. If there is no reason to suspect that these markers are tagging the same causal variant, then the true underlying pleiotropy at the locus will not be discovered.

This thesis will avoid declaring "pleiotropy" between traits when the only evidence is from GWAS SNPs, preferring the terms "cross-phenotype association" or "cross-trait association" instead. The connection between pleiotropy and the other two terms is obvious, but without fuller understanding of the path from genotype to phenotype, which would require analyses well beyond the scope of this thesis, the latter two terms are more accurate.

Because there are so many genome-wide associations of single traits published, now is the perfect time to use this wealth of information to look for cross-phenotype associations that could indicate pleiotropy, as there is reason to believe it is prevalent throughout the genome [45]. In fact, Boyle, Li, and Pritchard [46] argue for an "omnigenic" model; namely, that for complex traits, variants show a detectable effect in

Figure 1.4: Schematic of how a cross-phenotype association may appear when no real pleiotropy exists. One SNP has been typed (symbolized by the grey circle) and is in LD (black lines) with two variants that have not been (the red and blue circles). The blue and red variants are causal for two different phenotypes (solid lines), however because they have not been typed, genome-wide assocation studies cannot pick up on an association between them and their respective phenotypes. Instead, they each associate the phenotypes with the grey SNP (dotted lines), causing the appearance of pleiotropy at that SNP.



a "core set" of genes that are part of etiologically relevant pathways, but also in all genes throughout the genome, though their effects will be too small to be detectable. The authors suggest that these smaller effects may act through perturbations in gene regulatory networks active in cells and tissues relevant to the trait of interest. An obvious question that emerges from this hypothesis is whether effect sizes at a single locus are correlated across traits; that is, if a variant has a large enough effect at one trait that it is detectable by a GWAS, is it likely to have detectable effects in others? We note that heritable diseases, such as autoimmune disorders, are often comorbid with other heritable diseases, suggesting a shared genetic basis [47, 48]. Genetic variants shared between diseases have already been found, for instance between between metabolic disorder and schizophrenia [49, 50], ischemic stroke and migraines [51], and the manifold relationships between autoimmune diseases [52, 53, 54].

Knowledge of genetic relationships between traits can aid researchers in discovering novel genetic associations. In 2013, Andreassen *et al.* [50] used the epidemiological relationship between schizophrenia and cardiovascular disease as a starting point to investigate whether there was evidence for a genetic correlation between the two phenotypes. After establishing this, they performed a GWAS, conditioning on whether or not SNPs were associated with cardiovascular disease risk factors (triglyceride levels, waist-hip ratio, body mass index, type 2 diabetes, blood pressure, and cholesterol) and were able to increase their power to find variants associated with schizophrenia.

1.2.2 Examples of pleiotropy

Frequently, pleiotropy is discussed alongside the observation that a single gene appears to influence a number of traits. One of the most well-known genes shown in Figure 1.2 is *FTO*, which is noted for its association with a wide range of obesity-related traits like BMI and waist-hip ratio [32, 55], but has also shown an association with melanoma [56, 57]. While one may argue that BMI and waist-hip ratio may be different measures of the same underlying phenotype, it is hard to see how they could be measures of melanoma risk, especially as Iles *et al.* found the gene's effect on melanoma risk to be independent of BMI, and that the variants associated with cancer risk are not in LD with those associated with obesity, suggesting that this gene may be acting on these traits through different pathways [58].

However, single variants may also be associated with multiple phenotypes. One such variant is rs11209026, located in the gene *IL23R*. Carriers of the G allele of this SNP are at an increased risk of inflammatory bowel disease (both ulcerative colitis and Crohn's disease [59, 60]) as well as for ankylosing spondylitis [61] and psoriasis [62]. This is a known missense variant, with the G allele increasing the number of interleukin 23 (IL-23) receptors on the cell surface compared to the A allele [63]. IL-23 is a known mediator of inflammatory response in the body and has been associated with autoimmune diseases even before the discovery of associations with *IL23R* [64, 65]. It is therefore not surprising that so many autoimmune diseases are associated with this variant and it seems likely that it is the same pathway underlying the effect in all associated diseases.

Other well-known cross-phenotype associations in autoimmune diseases include a variant in *PTPN22* that has opposite effects in Crohn's disease than to those in rheumatoid arthiritis and type 1 diabetes—that is, the same allele that reduces one's risk of Crohn's disease increases one's risk of the other two. In a similar manner, there is a variant in *TNFRSF1A* that increases the risk of multiple sclerosis, but decreases the risk of ankylosing spondylitis [28].

In contrast to the example of probable vertical pleiotropy, Chami *et al.* [66] found apparent horizontal pleiotropy at five markers. The study focussed largely on red blood cell traits and the markers in question were associated with red blood cell distribution width or hemoglobin levels, but were also found to be associated with lipid

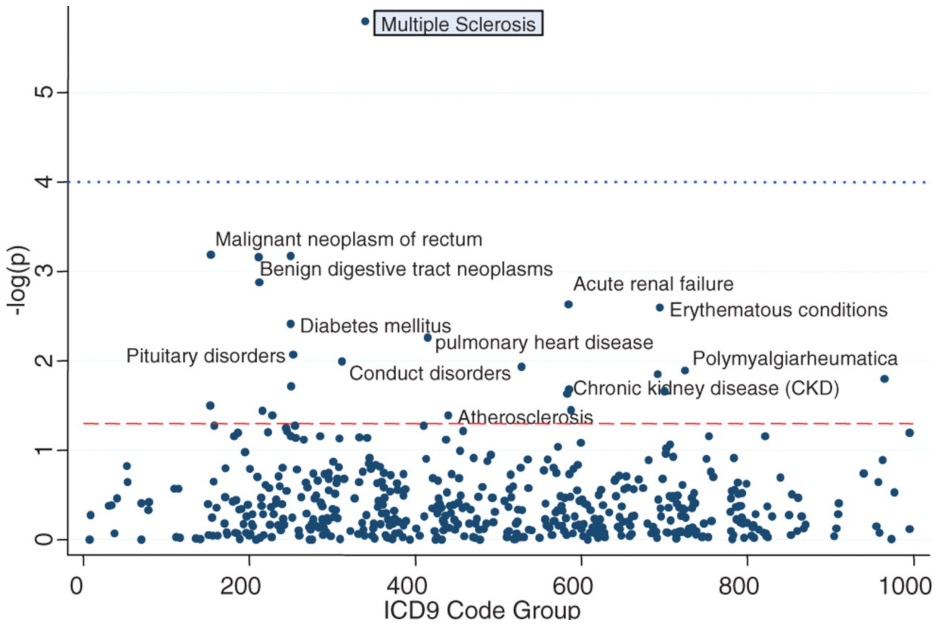


Figure 1.5: Plot of $-\log_{10}(p\text{-values})$ of associations (y-axis) of different diseases (plotted on the x-axis, using numeric billing codes from electronic medical records) at the SNP rs3135388. The lower, dashed line is at the log-scale equivalent of $p = 0.05$ and the upper, dotted line is the Bonferroni-corrected threshold. This is Figure 1 in Denny *et al.*'s 2010 PheWAS paper [68].

traits: total cholesterol, triglyceride levels, and high-density lipoprotein cholesterol. Effect sizes for the red blood cell traits remained the same after adjustment for lipid traits, suggesting that the association with one trait is not mediated through the others.

1.3 PheWAS

One obvious way to look for pleiotropy is to test the effect of a single genetic variant on multiple phenotypes. This is the intuition behind Phenome-wide association studies (PheWAS) [67, 68]. In the 2010 paper that coined the name, individuals with genotype data and electronic medical records were analyzed. The authors deduced patient disease phenotypes from billing codes and systematically tested each code for associaton with five SNPs to determine whether or not allelic distributions differed between cases and controls, similar to a GWAS. We see a plot of their results for the SNP rs3135388 in Figure 1.5.

The outcome of these studies may be the discovery of cross-phenotype associations at a given marker. However, like GWAS, PheWAS results must be adjusted for multiple testing; a standard Bonferroni correction would have to account for the

number of phenotypes tested, as well as the number of SNPs. This is bound to be too conservative because many of the phenotypes included in PheWAS are correlated, which introduces the problem of determining what the appropriate adjustment is in this case. Furthermore, SNPs selected for a PheWAS are often those discovered in a GWAS [69]. This potentially misses pleiotropic SNPs that have not yet achieved genome-wide significance in a study of a single phenotype, but which would easily achieve phenome-wide significance in multiple traits if studied. A mixed model approach for performing a full genome-wide analysis (while also controlling for population structure) has been suggested [70], but the approximation used in the provided software to avoid re-estimating variance components at each SNP leads to a loss of power to detect effects [71].

PheWASs necessarily require a large amount of data: the phenotype of every individual for every trait, as well as the genotype information for each individual at each of the SNPs under investigation. Fortunately, recent years have seen the proliferation of biobanks, such as the UK Biobank [72] or the China Kadoorie Biobank [73], as well as electronic medical records [74], meaning that the appropriate datasets are being created. It should also be noted that while PheWASs are commonly performed testing for association with genetic markers, this need not be the case. The study design has also been used to establish the association of phenotypes with levels of various biomarkers. This can reveal important information about the correlation structure of different phenotypes in the PheWAS dataset, as well as highlight biologically-relevant associations that GWAS would not necessary have been able to discover [75].

1.4 Performing a genome-wide association study

The work detailed in the rest of this thesis pertains to summary statistics from genome-wide association studies, so it is necessary to explain in some detail how these studies are performed and from where these summary statistics originate, starting from after the study subjects have been ascertained, measured for the phenotype, and genotyped. The work that needs to be done to ensure valid results will differ slightly from study to study, depending on the design and the data available. We do not attempt to cover all the variations in how a GWAS can be performed, but follow

the summary given by Anderson *et al.* [76] to give a broad outline of the quality control process, after the genotyping and phenotyping of study subjects is complete.

1.4.1 Quality control

The first part of quality control concerns ensuring that genotyping has been done correctly. Generally, these checks are performed by the company that has performed the genotyping. Once the data have been sent to the researchers, they may perform additional checks on genotype calling and take steps to ensure that the genotyping is correct. Often, the data are filtered so that SNPs that fail to genotype a certain proportion of the time (using a cut-off of 5% or lower) are removed. Similarly, individuals whose genotypes failed to be captured on a certain proportion of all markers (again, with a cut-off for 5% or less) are filtered out of the analysis. Furthermore, unless the study design is explicitly family-based, most GWASs will remove people who are closely related (first or second degree relatives) to other people in the study. SNPs may also be filtered for things such as low minor allele frequency (MAF) and for being out of Hardy-Weinberg equilibrium (usually measured by a χ_1^2 test or Fisher's exact test), since these may be the result of genotyping errors. In case-control studies, differential call-rate between cases and controls may also be a filter.

1.4.2 Imputation

It is also common practice to impute one's data to a reference panel, usually 1000 Genomes [77] or HapMap [8]. Informally, this means using the information at SNPs that were typed to estimate subjects' genotypes at SNPs that were not typed. This usually involves a process of phasing the data—that is, determining which of each person's alleles at a set of SNPs were inherited together on the same haplotype. Since we cannot discern this from the genotypes alone, we must use information from databases of sequenced individuals and from the other genotypes in the study. Once phases are determined, imputation occurs by comparing the haplotypes of the individuals in the studies to haplotypes in the reference panel and, at SNPs that were not genotyped for the study, using the known linkage disequilibrium structure and known haplotypes from the reference panel to fill in the genotyping gaps.

1.4.3 Covariates and confounding

Thus far, we have ignored the fact that an individual's phenotype is usually the product of both his or her genotype and environment. The latter includes both the environment external to the individual—the place where she or he lives, for example—or the individual's internal environment. For example, a number of traits show a sexual dimorphism. Height is a well-known example of this, with men being taller than women on average. Thus, any GWAS of height would have to account for sex as a covariate—a variable that is correlated with the trait measurement—to ensure that all subjects' phenotypes were being compared to the appropriate distribution.

Another example of a frequently used covariate is tobacco use, which is a well-known risk factor for both cancer and cardiovascular disease [78] and thus whether or not someone uses tobacco products (or did previously) would be an important variable to account for in a GWAS of one of these traits. Failing to account for it may cause the GWAS to discover variants associated with tobacco use rather than those associated with the traits under investigation.

Investigators also need to worry about confounders, which are unmeasured variables that affect both the independent and dependent variables [79]. One common potential confounder is population stratification, discussed below, which acts on both genotype and phenotype frequencies.

Population stratification

Allele frequencies often differ between different populations, so most studies try to ensure that the individuals gentoyped come from the same ethnic or racial group—for example, people of European descent or Han Chinese—however these populations will contain subpopulations that do not mate randomly, due to things like physical distance between the subpopulations, or social mores (for instance, people from different social classes). This can cause population structure, where allele frequencies, as well as phenotype prevalence, differ among subpopulations. When this population structure or stratification exists, a GWAS that does not account for the differences among these populations will find highly significant, but spurious results associating the phenotype with any locus whose allele frequencies differ significantly among the subpopulations. Additionally, even when the phenotype is expressed at the same

rate in all subpopulations, the underlying genetic causes may be different in different subpopulations [80, 81].

Genomic control

One way to test and account for population structure is through genomic control, where the test statistic λ_{GC} is calculated by taking the ratio of the median χ^2_1 test statistic calculated in the data and the median expected by the null (this is ≈ 0.455) [82]. If $\lambda_{GC} > 1$, then this is a sign that population stratification may be inflating the test statistics. One way of correcting this is to divide all χ^2 values by λ_{GC}. While genomic control is often used as a test of whether or not the data are affected by population structure, it cannot distinguish between markers that are associated with different populations and ones that are associated with the phenotype. As a result dividing by λ_{GC} in a GWAS leads to a conservative test and can hide true signals of association [83].

LD score regression

LD score regression was introduced to improve on genomic control, and claims to separate the influence of polygencity from population structure in inflated χ^2 statistics. It can also be used to estimate the genetic correlations between pairs of traits [84, 85].

The LD score of variant j, ℓ_j, given that the correlations, $r^2_{j,k}$ between j and the variants around it (calculated within a 1 centiMorgan window), variants $1, \ldots, M$ is simply

$$\ell_j = \sum_{k=1}^{M} r^2_{j,k}.$$

This gives a measure of how much variation the marker tags in a given population (LD scores need to be calculated in the same population that the GWAS was performed in for the results of the regression to be valid), with higher scores indicating higher amounts of tagging. LD score regression—at its most basic—is the regression of the LD scores of the markers in a GWAS onto the resulting χ^2 statistics. The intercept of the regression line (or, more accurately, the intercept $- 1$) gives a measure of the amount of inflation in the test statistic due to structure or cryptic relatedness in the

samples and the slope is an estimate of the heritability.

This method can be modified to get a measure of the genetic covariance between two phentoypes. Instead of regressing LD scores on the χ^2 statistics, the LD scores (multiplied by the square root of the product of the study sizes) are regressed onto the product of the z scores, with some adjustment for shared samples between the studies. Here, the slope multiplied by the number of markers gives the estimate of the genetic covariance between the two traits, ρ_g. When normalized by the SNP heritabilities of the two studies (h_1^2, h_2^2), we have

$$r_g = \frac{\rho_g}{\sqrt{h_1^2 h_2^2}}.$$

Principal Component Analysis

Most GWASs now rely on principal component analysis (PCA) of the genetic data to account for population structure. In general, PCA is a way of remapping data—whose variables may be correlated—onto independent variables called principal components. The process reduces the dimensionality of the data, while maintaining as much of the variability of the data within each principal component as possible [86, 81].

Principal components are listed in decreasing order of variability described (i.e. the first principal component explains the most variability, the second PC explains the second most, and so on). It is left largely up to the researcher how many principal components she or he would like to include in the regression analysis as covariates in order to account for the potential confounding due to population structure. Usually, λ_{GC} is calculated both before and after applying PCA to make sure that it is having its intended effect and that λ_{GC} is closer to 1 after the PCA than it was before it [86].

1.4.4 Statistical analysis

To determine which markers are associated with the trait, researchers can make a contingency table of genotypes at a given marker and phenotype class, then perform a χ^2 test or Fisher's exact test to determine whether there is a statistically significant association between the phenotype and the marker. This can be adapted from an additive model, where people who are heterozygous show an intermediate phenotype compared to those who are homozygous, to a dominant/recessive model by grouping

heterozygotes with one of the homozygote classes. However, these simple tests of allele frequency cannot account for covariates that may be necessary to include.

As a result of this, most genome-wide association studies use some form of generalized linear model, usually a logistic regression for binary traits (i.e. case-control data) or a linear regression for quantitative traits. Generalized linear models allow researchers to add to the model the covariates that are known to affect the trait under investigation, such as age, gender, whether or not the participant smokes, principal components (see Section 1.4.3), or any other relevant measurement. When applied correctly, this allows the effect of the contribution of the marker to the phenotype to be isolated from other factors.

Once the analysis is run, SNPs with strong evidence for association with the trait are noted and usually subject to some sort of follow-up analysis to determine biological plausibility. The generally accepted genome-wide significant threshold is $p < 5 \times 10^{-8}$ [28], though individual studies may modify this, depending on the desired power to detect true effects [26]. It is standard practice to perform a replication study in another cohort to ensure the result is a true positive. This usually involves genotyping a larger number of people on a smaller number of SNPs to verify the regions of association from the first (or "discovery") round [28]. Sometimes, some of the SNPs chosen for replication will have p-values that did not reach genome-wide significance in the discovery round, but were chosen because there was some evidence of association (usually on the order of $p < 10^{-6}$) and researchers have reason to believe they might be associated with the trait.

Power

In statistics, power is the ability to detect a true effect when it is present. It is a function of the Type I error rate (α), which is when one rejects the null hypothesis when it is true. Conversely, failing to the reject the null hypothesis when it is false is a Type II error, with rate β. Power is simply $1 - \beta$. It is defined for some α threshold and its corresponding test statistic threshold.

Given a p-value threshold, power in GWAS is dependent on the the study size and the minor allele frequency of a given marker [87, 28]. It can vary throughout the genome, depending on the density of markers for which there is information within

a given region. For instance, a region containing a causal variant that is not well tagged by any of the typed SNPs will likely fail to be statistically significant in a GWAS scan. On the other hand, in the same dataset, there may be another region containing a causal SNP that was typed, along with several other SNPs in high LD with it, allowing it to be detected—though it is not necessarily a given the causal SNP will be the lead SNP for the region (see Section 1.1.1). Additionally, errors from the genotyping process can lead to false positives or negatives. These may be the result of simply calling genotypes wrong, for a large subset of cases or controls, or may be due to something more subtle, such as systematic differences in the ability to call genotypes in cases and controls, to cases and controls being genotyped on different chips, or on the same chip, but processed separately, leading to batch effects.

1.4.5 Linear and logistic regression

We briefly review the forms of linear and logistic regression in genome-wide association studies. Assuming a quantitative trait, let $\boldsymbol{P} = \{P_1, P_2, \ldots, P_n\}$ be the vector of phenotype measurements in n individuals. For a given SNP, let $\boldsymbol{G} = \{g_1, g_2, \ldots, g_n\}$ be the vector of each of the n individual's genotypes. This is usually coded as number of copies of one of the SNP's alleles (called the "effect allele"). The regression algorithm calculates the maximum likelihood estimate of the effect of the "dosage" of the effect allele compared to the other allele. The true effect is β, and we denote the MLE calculated from the linear regression as $\hat{\beta}$.

Now, let $\boldsymbol{C_i} = \{c_{i,1}, c_{i,2}, \ldots, c_{i,n}\}$ be the vector of covariate measurements (or statuses, if the covariate is a factor, like sex or smoking exposure) for the ith of l covariates. Finally, we define $\boldsymbol{\epsilon}$, the vector of residual errors, which are approximately normally distributed with a mean of zero. The full model includes an intercept, $\hat{\beta}_0$ and the maximum likelihood estimates the effect sizes, $\hat{\beta}_i$, on the various vectors (whose true underlying values are β_0 and β_i, respectively). Hence,

$$\boldsymbol{P} = \hat{\beta}_0 + \hat{\beta}\boldsymbol{G} + \hat{\beta}_1\boldsymbol{C_1} + \ldots + \hat{\beta}_l\boldsymbol{C_l} + \boldsymbol{\epsilon}. \tag{1.1}$$

A more compact form of Equation 1.1 can be obtained if we collapse $\boldsymbol{G}$ and the $C_1, \ldots, C_l$ into an $n \times (l+1)$ matrix, $\boldsymbol{X}$, and let $\hat{\boldsymbol{\beta}}$ be the vector of effect sizes on

each predictor variable, we have

$$P = \hat{\beta}_0 + \hat{\boldsymbol{\beta}}\boldsymbol{X} + \boldsymbol{\epsilon}. \tag{1.2}$$

In either case, the parameter of interest—at least in a GWAS context—is $\hat{\beta}$, the maximum likelihood estimate of the effect size of the SNP's effect allele on the phenotype. This is interpreted as the amount by which the measurement increases (or decreases, if $\hat{\beta} < 0$) with the number of copies of the effect allele. Associated with each effect size is a standard error, SE, which gives a measure of how much the data varies around the predicted values and thus, how accurate the estimate of $\hat{\beta}$ is. To assess whether $\hat{\beta}$ is statistically significant at a given SNP—that is, whether or not one rejects the null hypothesis that $\hat{\beta} = 0$—one can perform a Wald Test, where

$$W = \frac{\hat{\beta}^2}{SE^2} \sim \chi_1^2 \tag{1.3}$$

or, alternatively, a Student's t-test, where

$$T = \frac{\hat{\beta}}{SE} \sim t_{n-2}. \tag{1.4}$$

If n is large ($n > 1000$), $t_{n-2} \approx \mathrm{N}(0, 1)$.

Logistic regression begins in much the same way as its linear counterpart, except that instead of $\boldsymbol{P}$ being a vector of measurements taken on a continuous scale, $\boldsymbol{P}$ is now a vector of 1s and 0s, which typically indicate case and control status, respectively. Let $\mathrm{P}(X)$ denote the probability of X. Then,

$$\log\left(\frac{\mathrm{P}(\boldsymbol{P}=1)}{\mathrm{P}(\boldsymbol{P}=0)}\right) = \beta_0 + \hat{\beta}\boldsymbol{G} + \hat{\beta}_1\boldsymbol{C_1} + \ldots + \hat{\beta}_l\boldsymbol{C_l}, \tag{1.5}$$

or, more compactly,

$$\log\left(\frac{\mathrm{P}(\boldsymbol{P}=1)}{\mathrm{P}(\boldsymbol{P}=0)}\right) = \hat{\beta}_0 + \hat{\boldsymbol{\beta}}_{\boldsymbol{X}}. \tag{1.6}$$

This models the odds of being a case as an exponential function of one's genotype and covariates. Here, the effect $e^{\hat{\beta}}$ tells us the odds ratio for each copy of the effect allele. If the prevalence of the phenotypes is low ($< 10\%$) or if the study population

is not enriched for the phenotype relative to the general population (i.e. the study is being carried out in a cohort), then the odds ratio can be a good approximation of the true underlying relative risk of being a case given one's genotype at the SNP [88].

Statistical significance is assessed in much the same way as before. In both cases, we have defined an additive model, where the phenotype risk of the heterozygote is intermediate to those of the two homozygotes. As in the case with the tests involving contingency tables, it is trivial to turn these into dominant/recessive models by changing the allele dose into an indicator of whether or not one carries the effect allele (so possible genotypes become 0, 1, and 1 instead of 0, 1, and 2) or whether one is homozygous for the effect allele (0, 0, 1).

Alternatively, one can perform a likelihood ratio test to determine whether $\hat{\beta} \neq 0$. We perform this on the log-likelihoods of the data. The test statistic, Λ is

$$\Lambda = 2 \log \left(\frac{\mathrm{P}\left(data|\beta = \hat{\beta}\right)}{\mathrm{P}\left(data|\beta = 0\right)} \right). \tag{1.7}$$

This is distributed approximately χ^2_1 , which allows p-values to be generated from this test. We can also set some threshold, T, for Λ such that if $\Lambda \geq T$, the null hypothesis is rejected in favor of the alternative.

Linear Mixed Models

Another way of accounting for population stratification is by using a linear mixed model approach. This involves including a random effects term (or vector of random effects terms, $\boldsymbol{r}$) in the linear regression model (see Section 1.4.5 and Equation 1.1). This term will usually be assumed to be normally distributed, centered on 0 with a variance dependent on the genetic relationship matrix (GRM, denoted $\boldsymbol{K}$) and its underlying variance, σ^2_r between all of the subjects in the study. That is, if

$$\boldsymbol{r} \sim \mathrm{N}\left(0, \sigma^2_r \boldsymbol{K}\right), \tag{1.8}$$

then the linear mixed model (see paragraph above Equation 1.2 for variable definitions) becomes

$$\boldsymbol{P} = \beta_0 + \boldsymbol{\beta X} + \boldsymbol{r} + \boldsymbol{\epsilon}. \tag{1.9}$$

The challenge then becomes estimating $\beta_G, \beta_1, \ldots, \beta_l, \sigma_r^2$, and σ_ϵ^2, which is the variance on the residual error term. This is a non-trivial problem and many algorithms exist to compute solutions accurately and efficiently [89, 90, 91, 92]. Additionally, $\boldsymbol{r}$ need not be the only random effect included in the model. Other random effects with their own design matrices may be included in a linear mixed model of this sort.

1.5 Bayesian approaches to GWAS

Bayesian approaches have become popular in statistics in the last few decades due to increased computational power, which makes it possible to approximate posterior distributions using sampling methods. These advances have naturally been applied to genome-wide association studies [93, 94]. The advantages these approaches have over their frequentist counterparts is that they allow for the incorporation of information drawn from other sources through the use of priors, and they naturally account for power by testing the null model directly [95, 93]. The downside is that it is often unclear what the appropriate prior distribution on a given parameter is appropriate, and a poorly chosen prior may produce spurious results. Furthermore, until recently, it was too computationally expensive to use Bayesian approaches, and it is only now that researchers are becoming familiar with them.

1.5.1 Bayes factors

Bayes factors are a natural consequence of Bayes' theorem, defined below in Equation 1.10:

$$\mathrm{P}\left(M|D\right) = \frac{\mathrm{P}\left(D|M\right)\mathrm{P}\left(M\right)}{\mathrm{P}\left(D\right)}\ . \tag{1.10}$$

If we have two models of the observed data, M_0 and M_1, then from Equation 1.10, we have

$$\underbrace{\frac{\mathrm{P}\left(M_1|data\right)}{\mathrm{P}\left(M_0|data\right)}}_{\text{Posterior odds}} = \underbrace{\frac{\mathrm{P}\left(data|M_1\right)}{\mathrm{P}\left(data|M_0\right)}}_{\text{Bayes factor}} \times \underbrace{\frac{\mathrm{P}\left(M_1\right)}{\mathrm{P}\left(M_0\right)}}_{\text{Prior odds}} \tag{1.11}$$

A Bayes Factor (BF) > 1 indicates evidence that M_1 describes the data better

than M_0 and, while BF < 1 indicates that M_0 describes the data better. Bayes factors are useful in comparing models of the data and can be thought of as the Bayesian version of the likelihood ratio test (Equation 1.7). However, unlike the LRT, which is the ratio of two likelihood functions evaluated at the values of each model that maximize it, a Bayes factor is the ratio of two likelihoods integrated over the parameter spaces of the their respective models. That is, for θ_0 and θ_1, corresponding to the continuous variables parameterizing M_0 and M_1, respectively. Then,

$$BF = \frac{\mathrm{P}\left(data|M_1\right)}{\mathrm{P}\left(data|M_0\right)} = \frac{\int \mathrm{P}\left(data|\theta_1, M_1\right)\mathrm{P}\left(\theta_1|M_1\right)d\theta_1}{\int \mathrm{P}\left(data|\theta_0, M_0\right)\mathrm{P}\left(\theta_0|M_0\right)d\theta_0} \tag{1.12}$$

These integrals are difficult—and in some cases impossible—to evaluate, hence an approximation is needed. Previous work has been done to calculate Bayes factors using Markov chain Monte Carlo (MCMC) [96, 97] or performing a Laplace approximation [98], however Jon Wakefield [87] developed an approximation of Bayes factors for GWASs, described in Section 1.5.2 below, which is easier and faster to calculate than either of these methods, and gives a similar answer a Laplace approximation [93].

Bayes factors vs. p-values

The short-comings of p-values is the motivating factor of Jon Wakefield's paper [87] introducing Bayes factors as a potential measure of evidence in GWAS. One advantage of Bayes factors is that they can explicitly test the evidence for one hypothesis over another. So instead of rejecting or not rejecting a null hypothesis of no association between the SNP and the trait, as in p-value based analyses, there is an opportunity to measure the strength of evidence in favor of an association versus no association.

Additionally, using p-values to determine statistical significance in a GWAS will result in a loss of power to detect effects when sample sizes are small and when minor allele frequencies are low, while Bayes factors do not depend on these features of the data. In short, p-values do not naturally account for power the way Bayes factors do. By specifying the alternative hypothesis, one can account for power, but this requires a specific alternative hypothesis to be stated, instead of simply rejecting a null hypothesis or not. When the prior for the alternative model is the same across all SNPs, Bayes factors give identical rankings to p-values of the evidence of association,

so it is possible to recover the frequentist results from a Bayesian analysis [93, 87].

1.5.2 Wakefield's approximation

One practical drawback of Bayes factors is that they are often difficult to calculate due to the need to take integrals of likelihoods that may not have closed forms, as is the case with logistic regression.

A Bayes factor is defined as the ratio of two likelihoods. Let M_0 be the null hypothesis and M_1 be an alternative hypothesis (note that M_1 need not be the only alternative hypothesis available). Recall from Equation 1.12 that the Bayes factor, BF, of the evidence in favor of M_1 against the null is simply

$$BF = \frac{\mathrm{P}\left(data|M_1\right)}{\mathrm{P}\left(data|M_0\right)}.$$

Because genome-wide association studies often have sample sizes in the thousands, Wakefield shows that asymptotically, the maximum likelihood estimate of the effect size of the SNP on the phenotype, $\hat{\beta}$, is normally distributed about the true effect size, β, with its variance as simply the square of the standard error of $\hat{\beta}$, $SE^2_{\hat{\beta}}$. Using this with a prior on the true effect size of

$$\beta \sim \mathrm{N}\left(0, \sigma^2\right) \tag{1.13}$$

yields an approximate Bayes factor (ABF) for the alternative model that there is an association between the SNP and the phenotype against the null. This is defined using the probability density function of the normal distribution, $f\left(\mu, \sigma^2\right)$ (where μ is the mean and σ is the standard deviation). Thus,

$$ABF = \frac{f\left(0, SE^2_{\hat{\beta}} + \sigma^2\right)}{f\left(0, SE^2_{\hat{\beta}}\right)} \tag{1.14}$$

or equivalently,

$$ABF = \sqrt{\frac{SE^2_{\hat{\beta}}}{SE^2_{\hat{\beta}} + \sigma^2}} \exp\left\{\frac{1}{2}\left(\frac{\hat{\beta}^2}{SE^2_{\hat{\beta}}} - \frac{\hat{\beta}^2}{SE^2_{\hat{\beta}} + \sigma^2}\right)\right\}. \tag{1.15}$$

This approach has the advantage of being easy to calculate and relying only

on summary statistics, which are more readily available and easier to share than genotype-level data.

1.5.3 Posterior probability of association

Once a Bayes factor has been calculated (or approximated), we can start investigating the posterior distribution. If we choose a prior π on the probability of a given marker's association with a phenotype and calculate a Bayes factor (BF) for each SNP, we see from Equation 1.12 that we can calculate the posterior odds (PO) on the alternative hypothesis [93]. That is,

$$\text{PO} = \text{BF} \times \frac{\pi}{1-\pi}. \tag{1.16}$$

From here, we can calculate the posterior probability of association (PPA) as

$$\text{PPA} = \frac{\text{PO}}{1+\text{PO}}. \tag{1.17}$$

Stephens and Balding (2009) [93] argue that, unlike p-values, the PPA has the advantage of being unaffected by power, sample size, or multiple testing burden. Additionally, they explain how non-additive models of association can be applied in this framework and suggest averaging the Bayes factors of each of these different models, weighted according to prior belief that a given model being the true one underlying the data. They further suggest using a mixture of normal distributions as the prior on the true effect size, which is equivalent to the weighted average of Bayes factors under each individual normal distribution.

1.6 Meta-analysis in GWAS

Due to the need for large sample sizes and replication cohorts, genome-wide association studies often involve combining the results of many studies of the same or similar phenotypes. An overlapping group of SNPs are chosen and then carefully analyzed to determine the "true" genetic effect of the marker on the phenotype. Two of the chief ways this is done are through fixed effects meta analysis and random effects meta-analysis. It is important to note that these two examples assume that the studies in

the meta-analysis are totally independent of one another—that is, there is no overlap in study subjects.

Often, the results from individual studies can be visualized as a forest plot. Figure 1.6 shows an example of this on imaginary data: here we imagine that we are visualizing the results of four different GWASs at the same SNP. The boxes represent the estimates of the effect allele on the phenotype from each individual GWAS, and the bars show the 95% confidence intervals of these estimates, as is standard. In our example, we see straight away that there is a genome-wide significant association in Study 1. We also see a possible association with Study 3, however this association does not quite achieve genome-wide significance. If we have reason to believe that there should be a correlation between Study 1 and Study 3—perhaps they were studies of similar or even the same phenotype—we may be willing to accept that there is a true association with Study 3 at this marker.

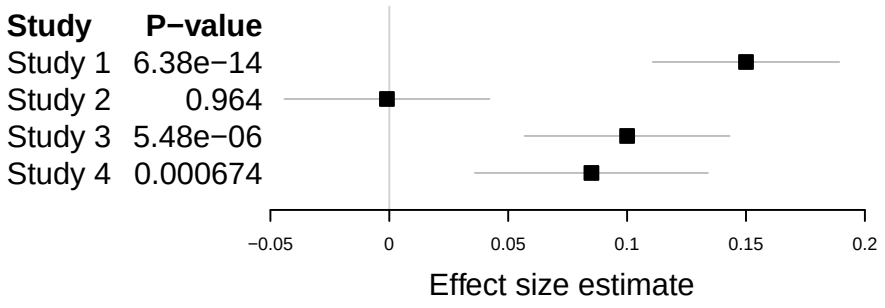


Figure 1.6: Forest plot of effect size estimates (boxes) and the 95% confidence interval (lines) of the effect of the same marker in four different studies.

In a similar way, there is a possible association for Study 4 with this marker; however, due to the usual correction for multiple testing in GWAS, its p-value does not come close to the required threshold. Again, if we believe that there is a relationship between Study 4 and Study 1, or between Study 4 and Study 3, we might be willing to accept that there is evidence of association in Study 4, even if it is not genome-wide. Ideally, any meta-analysis we perform would allow us to take into account what is known about the relationships between studies and automatically adjust significance thresholds based on these. That is, if one study achieves genome-wide significance at a marker, and other studies of the same or similar phenotypes show evidence of

association at the same marker, we would like a meta-analysis method that encodes our intuition that the non genome-wide significant associations are probably real as well.

1.6.1 Quality control of meta-analysis

As with GWAS, the data for meta-analyses must be checked and harmonized to ensure the accuracy of the results. The details will depend on the source datasets and the type of meta-analysis being performed. Some common tasks include converting the source data so that everything is reported in the same manner and on the same scale. For instance, some studies report effect size estimates ($\hat{\beta}$s) and standard errors ($SE_{\hat{\beta}}$s), while others will report z-scores, p-values, or 95% confidence intervals. Effect sizes may be reported as the $\hat{\beta}$ from the regression model, or as the relative risk of the effect allele versus the alternate allele. Researchers will need to convert these data to the summary statistics they need before proceeding.

Additionally, some studies may analyze a trait as binary while others analyze the same trait as quantitative. Consider glaucoma: GWAS of this disease have been performed on both intra-ocular pressure (a quantitative trait) [99] and as a binary trait, comparing those with diagnoses to healthy controls [100]. Before meta-analyzing such studies, researchers need to be satisfied that the traits in the analysis are either similar enough that the meta-analysis can be performed, or that the meta-analysis method is flexible enough to handle different types of data. Additionally, it is important that a researcher be certain that relevant covariates were appropriately accounted for in each of the individual studies, and that the effects being analyzed are purely genetic effects. Finally, if the meta-analysis is not being performed on the same set of markers from all studies, then researchers either need a method that looks at evidence across a genomic region, which is flexible enough to handle missing data, or be satisfied that they can legitimately use proxy SNPs in high LD with the target SNP and that these are all tagging the same causal variant (see Section 1.2.1).

Unless the meta-analysis only involves p-values or some other summary statistic that is agnostic to the direction of a marker's effect, researchers must check that the markers are aligned—that is, that the effect allele is the same allele across all studies. Otherwise, effects in opposite directions may cancel each other out and prevent a true

signal from being discovered. This can be achieved by choosing some database, such as 1000 Genomes (which we use in our data alignment in Chapter 4, Section 4.2.2), and aligning all markers to the those reported in the database. This can also act as a check against potentially multi-allelic markers and genotyping errors in one or more of the datasets, and problematic markers can be removed. Further checks can be done by comparing allele frequencies at the markers in the dataset to the frequencies reported in the dataset and removing those that differ too greatly.

1.6.2 Fixed effects meta-analysis

A fixed effects meta-analysis assumes that the true underlying effect of the SNP on the phenotype is the same across all studies. Thus, any differences that arise between studies in their effect size estimates are the result of sampling noise. Because larger studies are assumed to produce a better estimate of the true effect size, studies are usually weighted by their sample size, or—more frequently—by the inverse of the estimated variance of the marker's effect size.

More formally, for the s studies in the meta-analysis, each with estimated effect size $\hat{\beta}_i$ and standard error SE_i (for $i \in 1, \ldots, s$), we have weights, w_i

$$w_i = \frac{1}{SE_i^2} \tag{1.18}$$

Now with the weights defined, we can obtain the meta-analyzed effect-size, $\hat{\beta}_f$ as

$$\hat{\beta}_f = \frac{\sum_{i=1}^{s} w_i \hat{\beta}_i}{\sum_{i=1}^{s} w_i}, \tag{1.19}$$

and the meta-analyzed standard error SE_f is

$$SE_f = \sqrt{\frac{1}{\sum_{i=1}^{s} w_i}}. \tag{1.20}$$

A fixed effects meta-analysis is easy to perform and its results are easy to interpret. When using studies that are unambiguously measuring the same thing, it can be a very useful way of collating information across cohorts and improving the

power to detect an effect. It is often used to combine evidence between discovery and replication cohorts in a GWAS to determine the "true" effect size of the marker on the phenotype.

1.6.3 Random effects meta-analysis

A random effects meta-analysis is appropriate when there is reason to believe that the true effect size may not be the same for all studies. In this type of meta-analysis we assume that effect sizes are all drawn from the same distribution, whose mean effect size, μ and variance, σ^2 we wish to estimate.

Unlike fixed effects, random effects assumes two sources of error: one is the error due to sampling the true effect size of the study, β (as in fixed effects), and the other from the variation in true effect sizes between studies—that is, from the fact that $\beta \neq \mu$. We will call the measurements of these σ_i and τ respectively (and, as before, $i \in 1, \ldots, s$).

Estimating τ is not trivial and there have been several methods proposed, some of which involve maximization or other iterative processes [101]. Once this value is estimated, we can continue with the rest of the meta-analysis, which is almost identical to the fixed effects version. We calculate new weights for each study, v_i such that

$$v_i = \frac{1}{SE_i^2 + \hat{\tau}^2}. \tag{1.21}$$

And with these v_i, we can reuse Equations 1.19 and 1.20 to obtain the meta-analyzed estimates of the mean, $\hat{\beta}_r$, and variance, SE_r^2 , of the distribution from which the effect sizes for each study were drawn. That is,

$$\hat{\beta}_r = \frac{\sum_{i=1}^{s} v_i \hat{\beta}_i}{\sum_{i=1}^{s} v_i}, \tag{1.22}$$

and

$$SE_r = \sqrt{\frac{1}{\sum_{i=1}^{s} v_i}}. \tag{1.23}$$

Compared to SE_f in the fixed effects framework, the random effects SE_r will

be larger because $\hat{\beta}_r$ and SE_r are describing a distribution as opposed to a point estimate. This feature also makes the resulting p-values for random effects models larger than those of fixed effects models. Han and Eskin [102] make this observation and demonstrate it through simulation. They note that the null model in a traditional random effects meta-analysis assumes the same heterogeneity in the effect sizes of the studies as is calculated for the alternative hypothesis. This is despite the fact that, by definition, the null model assumes that the effect sizes are zero in all the studies. To solve this, they suggest a modification that assumes no heterogeneity in effect sizes under the null and uses a likelihood ratio test to assess the evidence.

Thus, the null model assumes that $\tau^2 = 0$ and $\mu = 0$. Then if $\hat{\tau}^2$ and $\hat{\mu}$ are the maximum likelihood estimates for these parameters (the authors suggest that this be obtained through an iterative procedure), the likelihood ratio test statistic, S_{New} is

$$S_{New} = \sum_{i=1}^{s} \log\left(\frac{v_i}{v_i + \hat{\tau}^2}\right) + \sum_{i=1}^{s} \frac{\hat{\beta}_i^2}{v_i} - \sum_{i=1}^{s} \frac{\left(\hat{\beta}_i - \hat{\mu}\right)^2}{v_i + \hat{\tau}}.$$

The statistical significance of S_{New} can be assessed by a permutation test (randomly permute each studies' data), sampling $\hat{\beta}_i$ from a normal distribution centered on zero and with variance v_i to obtain a null distribution, or using the asymptotic distribution, which is a mixture of the χ_1^2 and χ_2^2 distributions [102, 103]. Additionally, the authors show that this method has better power than fixed effects or the traditional random effects approaches to detect true effects among heterogeneous effect sizes while controlling appropriately for for Type I errors. Furthermore, the test statistic can be decomposed into its constituent null and alternative parts, which can be adjusted separately for different levels of genomic control, if necessary.

Traditionally meta-analysis has been done to combine results across different cohorts; this is still an active area of research. We see from Shi and Lee [104] and Morris [105] that the idea of effect size estimates across studies being distributed multivariate normal (as in several of the methods described below, including ours) is not new and there is much that those interested in cross-phenotype associations can learn from the methodologies used in trans-ethnic meta-analyses, particularly in dealing with heterogeneous effects.

1.6.4 Fisher's method

R.A. Fisher also proposed a method of aggregating evidence across independent studies. For n studies with p-values $p_1, p_2, \ldots, p_n$, he proposed a χ^2_{2n} test statistic of

$$\chi^2_{2n} = -2 \sum_{i=1}^{n} \log (p_i) . \tag{1.24}$$

From here the meta-analyzed p-value can be calculated [106].

We note that because this method takes only p-values as its input, there is no assumption about the size and direction of effects underlying the studies, similar to a random effects meta-analysis. It is a test of a global null hypothesis that $\beta_i = 0, \forall i \in \{1, 2, \ldots, n\}$. However, unlike fixed and random effects meta analyses, it cannot provide estimates of the global effect size across or its standard error.

1.7 Methods for detecting cross-phenotype associations

While not as extreme as the explosion of genome-wide association studies, there has been an increase in interest in finding variants that affect multiple phenotypes. In the last five years, there have been a number of papers proposing methods for how to go about this. In Table 1.1, we highlight the methods that try to find associations across phenotypes at a SNP or in a region of the genome and discuss them in more detail in the coming subsections.

Broadly speaking, these methods divide into two categories: those that search for a gene or region of the genome that is associated with multiple phenotypes and those that search for specific variants associated with multiple phenotypes. Thus, some methods do not attempt a SNP by SNP comparison, but use linkage disequilibrium to divide the genome into regions for analysis [107, 108]. Another division occurs between methods that are concerned with quantifying the correlation and the heterogeneity between phenotypes [108, 109, 110], and those that aim to provide a definitive test for multiple phenotypic associations at a single marker [52, 111]. We review some of these methods to provide context for our own work, but the rate at which new methods are published prohibits a systematic and comprehensive review of all methods.

Table 1.1: Table of GWAS meta-analysis methods discussed in this chapter. We list the name of the method, its bibliographic reference number, what summary statistics it uses (if any), whether or not the method scales to use in genome-wide datasets (a ✕ means that it does), whether or not it can incorporate prior information (a ✕ means that it can), and whether it can accommodate covariance in the data due to study design, for example shared controls (a ✕ means that it can).

Method	Ref	Summary stats	Scales	Prior info	Cryptic Cov.
CPMA	[52]	p-values	✕		
PRIMe	[107]	p-values, LD structure between SNPs	✕		
Binary effects model	[112]	effect size estimates and standard errors		✕	
ASSET	[111]	GWAS z-scores	✕	✕	✕
TATES	[113]	p-values, LD structure between SNPs	✕		
Bayes factors with heterogeneous subgroups	[109]	z-scores or effect size estimates and standard errors	✕	✕	
Univariate Bayes factor with cost thresholds	[110]	effect size estimates and standard errors	✕	✕	
Test based on the Fisher combination function	[114]	GWAS p-values	✕		
Trinculo	[115]	Needs genotype data	✕	✕	
Pairwise Bayes factors	[108]	effect size estimates and standard errors	✕	✕	✕
BUHMBOX	[54]	Needs genotype data	✕		
Bayesian meta-analysis using MCMC	[116]	effect size estimates and standard errors	✕		✕
META-SCOPA	[117]	Regression coefficients and covariance matrices from SCOPA analyses	✕		
MTAG	[118]	GWAS z-scores	✕		✕

1.7.1 Cross-phenotype meta-analysis [52]

One of the earliest and most well-cited methods is that of Cotsapas *et al.*, which proposes a cross-phenotype meta-analysis (CPMA) test. There will be more discussion of this test in Chapter 3 (see Section 3.3.1), but briefly, it works by testing $-\log(p\text{-values})$ for the associations of a collection of phenotypes at a given marker, with the value for the phenotype originally associated with the SNP removed. The test is whether the distribution of these p-values deviates significantly from the null model that the p-values (when sorted in decreasing order) are decaying exponentially at a rate (λ) of 1. The CPMA test statistic is simply

$$CPMA = -2 \times \frac{\mathrm{P}\left(data|\lambda = 1\right)}{\mathrm{P}\left(data|\lambda = \hat{\lambda}\right)}.$$

This test statistic is distributed χ^2_1, and so we can get a measure of how extreme the distribution of p-values is against a null model of no association with the marker. Both the test statistic and corresponding p-value are quickly and easily calculated and most people are familiar with p-values and have some idea of how to interpret them. However, there are downsides to this test. While it may show evidence for further phenotypic associations at the marker, it cannot say how many or which phenotypes are associated. As discussed in 1.5.1 GWAS p-values are sensitive to minor allele frequency and study size, which means that meta-analysis methods that use them as their inputs will inherit this sensitivity.

1.7.2 PRIMe [107]

The Pleiotropic Region Identification Method (PRIMe), in addition to taking p-values from genome-wide association studies, requires the LD structure between the SNPs under investigation. From there, the method identifies the SNP with the lowest p-value over all phenotypes in a region of LD as the "driver" and all other SNPs that have also reached genome-wide significance in that LD block as "passengers". Regions of the genome are defined as those that have a driver. Next, the number of phenotypes associated with a region are counted and a p-value, P_T, is calculated as the probability of observing the number of associated phenotypes in a given region. Because LD structure is different for each region, there is no universal null distribution

for the number of associated phenotypes. A region is considered pleiotropic if its $P_T < 5 \times 10^{-8}$.

If the different GWAS used in this analysis are completely independent of one another, then the null distribution for a region may be estimated using a binomial distribution. For correlated GWASs, they turned to simulation to assess P_T. This is because the correlation between studies necessitates looking at z scores (which can be calculated from p-values), which should be distributed as a multivariate normal. However, calculating P_T in this environment would require evaluating the multiple integral of a multivariate normal distribution over M SNPs, so simulating z scores under the null that no phenotypes are associated with the M SNPs in this region is the faster approach.

Though this method would be useful for scanning genomes for regions that are associated with many phenotypes, as before, P_T cannot tell us which phenotypes are associated. Furthermore, because it uses p-values from studies as its input statistics, we encounter the same problems that we did with the CPMA. In particular, it is troubling that regions are defined according to the lowest p-value associated with the LD block, as the lowest p-value may not correspond to the most biologically relevant marker or the most extreme effect on any phenotype, both of which are equally valid and justifiable means by which to define a region of the genome.

1.7.3 Binary effects model [112]

The binary effects model was introduced to deal with heterogeneity among effect sizes. The authors point out that some genetic effects only appear to affect one population. Thus they suggest a method for conducting a meta-analysis in two steps: first by determining what they call the m-value, which is the posterior probability that each study has a true association with the phenotype at a given marker, and then using this probability as an added weight in a fixed effects meta-analysis (see Section 1.6).

Let $\hat{\boldsymbol{\beta}}$ be the vector of effect size estimates from a GWAS. Let T be the binary string showing the pattern of association of across all n studies, and t be a particular pattern of association. Finally, U is the set of all possible patterns of association and U_i is the set of all binary strings whose ith element is 1. Then the m-value, m_i for study i is

$$m_i = \frac{\sum\limits_{t \in U_i} \mathrm{P}\left(\hat{\boldsymbol{\beta}} | T = t\right) \mathrm{P}\left(T = t\right)}{\sum\limits_{t \in U} \mathrm{P}\left(\hat{\boldsymbol{\beta}} | T = t\right) \mathrm{P}\left(T = t\right)}.$$

The prior $\mathrm{P}\left(T = t\right)$ is the ratio of two beta distributions, while the likelihood $\mathrm{P}\left(\hat{\boldsymbol{\beta}} | T = t\right)$ is the product of multivariate normal densities.

The m value can be calculated directly by summing through all the elements of U or—if this n is large ($n \geq 20$, say)—by using a Markov chain Monte Carlo approach to estimate the distribution. Once this is done, let w_i be the weights as in Equation 1.18 and Z_i be the z score of study i. That is,

$$Z_i = \frac{\hat{\beta}_i}{SE_i}.$$

Then the binary effects test statistic, S_{BE} is

$$S_{BE} = \frac{\sum\limits_{i=1}^{n} m_i Z_i \sqrt{w_i}}{\sum\limits_{i=1}^{n} \sqrt{m_i^2 w_i}}.$$

1.7.4 Subset based analysis [111]

This method involves looking across subsets of studies included in the meta-analysis. Unlike previous methods, this uses z scores (the effect size estimate of the SNP on the phenotype divided by its standard error) and the study sizes as their inputs. The study sizes are used to calculate the weights, w_k, of each of the K studies in the chosen subset, S. From there, they calculate $Z\left(S\right)$, which is the weighted sum of the z-scores from the K studies. Next, they calculate their final test statistic, $Z_{meta-max}$, which is the maximum $|Z\left(S\right)|$ for all possible subsets, $\boldsymbol{S}$. That is

$$Z_{meta-max} = \max_{S \in \boldsymbol{S}} |Z\left(S\right)|.$$

Rather than investigating all subsets exhaustively, they use the discrete local maximum method [119] to find the subset that maximizes this statistic and gives a corresponding p-value for $Z_{meta-max}$. To allow for a two-sided test (which also accounts for combinations of studies where the effects of the SNP on the traits may

not be in the same direction), they introduce an adjustment where they split the studies by direction of effect to find $Z_{max,-}$ and $Z_{max,+}$ and estimate the respective p-values for them, ($\tilde{P}^{-}_{DLM}$ and $\tilde{P}^{+}_{DLM}$, respectively). Then

$$Z^2_{meta-max} = -2\left[\log\left(\tilde{P}^{+}_{DLM}\right) + \left(\tilde{P}^{-}_{DLM}\right)\right],$$

which has a χ^2_4 distribution. If the effect is only ever in one direction, then $Z^2_{meta-max} = -2\log\left(\tilde{P}^{+}_{DLM}\right)$ or $Z^2_{meta-max} = -2\log\left(\tilde{P}^{-}_{DLM}\right)$.

This method is flexible enough to cope with covariance between the different studies (from shared controls, for instance). It has the advantage of not only being able to show that a SNP affects multiple traits, but of reporting exactly which traits are affected. The study's creators have written an R package [120] to implement this method, called ASSET [121], which is further discussed in Chapter 3, Section 3.5.

1.7.5 Multivariate phenotype analyses (TATES and Trinculo)

One area of particular interest is in searching for regions of the genome associated with a particular trait by combining associations from studies of phenotypes associated with the trait. For instance, a study of cardiovascular disease (CVD) may be subdivided into studies of the various risk factors associated with CVD, such as levels of high- and low-density lipoproteins, serum triglyceride levels, body mass index, etc. Having performed a GWAS on these different phenotypes, the researcher may wish to pool them together again at the end in the hopes of increasing his or her power to find a genetic association with the overall trait of interest. It may also be of interest to see the extent to which each component contributes to the overall phenotype.

Among others, Yang *et al.* (2010) [122], van der Sluis *et al.* [113], and Yang *et al.* (2016) [114] all propose methods for performing this type of meta-analysis, each involving some way of combining test statistics from an ordinary GWAS into a single statistic with a known null distribution. They have been grouped together here because while their methods are interesting, they cannot be used in a more flexible manner to search for regions associated with phenotypes that might not ordinarily be put together. While it certainly makes sense to search for cross-phenotype associations among closely related traits, this thesis is more concerned with approaches that treat

each phenotype as self-contained rather than as one part of a larger, predetermined whole.

Trinculo [115]

Additionally, Jostins and McVean [115] have introduced a speedy way of implementing, among other things, multinomial regression, which is the more general form of logistic regression (see Section 1.4.5). In multinomial regression, the number of discrete outcomes can be more than two.

This allows for the creation of outcome classes that can be tested via a LRT for association across multiple phenotypes, as well as heterogeneity between effect sizes for each outcome [123]. This, of course, requires genotyping data and measurements on multiple phenotypes, and so is most frequently used to look for associations across multivariate phenotypes, rather than disparate ones.

1.7.6 SCOPA and META-SCOPA [117]

The standard regression model may be inverted, so that the genotype at a given SNP becomes the dependent variable, while the phenotype becomes the independent variable. That is for phenotypes $y_1, y_2, \ldots, y_k$, for a given SNP, the genotype of the ith individual in the study, G_i can be denoted

$$G_i = a + \sum_{j=1}^{k} \beta_j y_{ij} + \epsilon_i$$

where, as usual, β_j is the regression coefficient of the jth phenotype on the genotype, and ϵ_i is the residual error term. This is implemented in Software for COrrelated Phenotype Analysis (SCOPA) [117]. The authors advise that the phenotypes be residuals of the measured phenotypes after adjusting for covariates like age, sex, and population structure. From there, the maximized log-likelihood of the full model is compared to the the maximized log-likelihood of the null model (where $\beta_j = 0, \forall j \in 1, \ldots, k$) in a likelihood ratio test, which can be converted into a p-value.

An advantage of this approach is that for SNPs that show a genome-wide significant association between phenotype and genotype, a model selection can be used to determine the most likely subset of phenotypes that are associated with the SNP. The

authors systematically performed the regression using every possible combination of phenotypes, calculated the log-likelihood for each model, and chose the likeliest subset as the one with the lowest Bayesian information criterion. Furthermore, because the effects of the phenotypes are being estimated jointly, the correlation structure between them is accounted for.

The authors also implemented a way to meta-analyze the results of multiple SCOPA analyses (META-SCOPA), which uses a multivariate generalized least squares method to aggregate the regression coefficients across all phenotypes [124] and performs the equivalent of a fixed-effects meta-analysis for this study design. Additionally, the Bayesian information criteria for each subset model at a SNP can be combined across all studies to determine the most likely model using the full meta-analyzed dataset.

A disadvantage of this method is that the regression coeffients from both the original analysis and the meta-analysis are difficult to interpret. The estimate of β_j cannot represent the effect of the phenotype on genotype, because the causation is safely assumed to be in the opposite direction.

1.7.7 BUHMBOX [54]

Sometimes, the underlying genetic architecture for a disease shows heterogeneity in subgroups, as is the case with ischemic stroke [125]. Specifically, variants were found that appeared to affect an individual's risk of cardioembolic stroke without appearing to affect the risk of large or small vessel stroke. When the variants causing the disease in one of the subgroups have a pleiotropic effect and affect the risk of a different disease, the genetic correlation induced between the two diseases overall can suggest pleiotropy. BUHMBOX—Breaking Up Heterogeneous Mixture Based On cross(X)-locus correlations—searches for genetic heterogeneity among subgroups of a disease in order to distinguish between these two scenarios.

This method compares the allele frequencies of variants known to be associated with the two diseases of interest (D_A and D_B) in cases of both studies. If there is subgroup heterogeneity in one of the diseases—say D_A—then the risk allele frequencies of markers associated with D_B are expected to be increased in only a subset of of cases for D_A. If the variants are shared between the two diseases more generally, then the

risk variants affecting D_B are expected to show increased frequency in cases of D_A in general. The BUHMBOX test statistic is designed to return a statistically significant result only in the case where subset heterogeneity exists.

The most obvious drawback of this approach is that risk variants for D_B must be known ahead of time. The method cannot find loci that are novel for both D_A (or a subset of it) and D_B. It also requires the genotype data of the participants in both disease studies in order to calculate a correlation matrix of the effects of the variants being used in the test on both diseases. This creates challenges in sharing data across labs—both due to needing ethical approval to do so and also due to the challenges that come with sharing and then storing large amounts of data.

1.7.8 Univariate Bayes factor with cost thresholds [110]

Asimit *et. al* [110] suggest an approach of calculating Wakefield's approximate Bayes factor and then testing whether the Bayes factors across multiple traits (the paper focuses on the case where only two traits are under investigation) are high at more SNPs than would be expected by chance. The threshold for whether an approximate Bayes factor is considered "high" or "low" depends on the prior probability, π_0, of no association between the SNP and the phenotype and R, the ratio between the cost of a Type II error and a Type I error—that is,

$$R = \frac{\text{cost of not rejecting the null hypothesis when it is false}}{\text{cost of rejecting the null hypothesis when it is true}}.$$

Then the significance threshold for the Bayes factors is

$$\frac{\pi_0}{R\,(1-\pi_0)}.$$

Using this threshold, the study's authors use a procedure to thin the number of SNPs while prioritizing those with high ABFs so that they are left with a set of SNPs that are approximately independent. From there, they construct a contingency table and then perform the paired equivalence of the χ^2 test, the McNemar mid-p test [126], which allows them to test for more genetic overlap between traits than one might expect by chance. There is no requirement for this method to use Wakefield's ABF—a fact they acknowledge by describing a pruning method that uses p-values

rather than ABFs. However, it is unlikely that using ABFs would cause them to exclude SNPs with p-values that were otherwise genome-wide significant.

1.7.9 Bayes factors with heterogeneous subgroups [109]

Wen and Stephens (2014) [109] use Bayes factors (though not necessarily calculated using Wakefield's approach) to perform a version of a random effects meta-analysis (see Section 1.6) and test for association between a given SNP and multiple traits (or a subsets of traits) against a global null hypothesis of no association in any of them. They consider both the case where the effects across studies have been standardized (that is, the effect sizes have been transformed into z scores) and the case where they have not, but the models for analysis under both these frameworks are almost identical.

In this analysis, the set of effect sizes (standardized or not) are assumed to be distributed multivariate normal about a mean vector of zeroes, with a variance-covariance matrix of Σ. The hyperparameters ϕ^2 and ω^2 describe the prior variances on the distribution of effect sizes and the distribution of mean effect sizes, respectively. Thus, the diagonal of Σ is $\phi^2 + \omega^2$ and the off-diagonal is ω^2. Here, ω gives a measure of heterogeneity of effect sizes across the traits. Thus, the null hypothesis is that $\omega = \phi = 0 \rightarrow$ all effect sizes are zero.

They performed their analysis on multiple values of ϕ and ω and take the weighted mean Bayes factor over these models to obtain a single meta-analyzed Bayes factor over all models of heterogeneity. This led to an increase in power to detect effects over performing the analysis in subgroups and also allowed for the discovery of novel associations.

1.7.10 Pairwise Bayes factors [108]

Pickrell *et al.* [108] use Wakefield's ABF to test for the association of a SNP with pairs of phenotypes. They initially tested three models:

Model 1: The SNP is associated with Trait 1 only,

Model 2: The SNP is associated with Trait 2 only, and

Model 3: The SNP is associated with Trait 1 and Trait 2.

These correspond to calculating the ABF for Trait 1, the ABF for Trait 2, and

multiplying the $ABFs$ for Traits 1 and 2—this is permitted under the assumption that cohorts in the studies are independent. From here, the genome is split into regions, each with their own Bayes factors. For a region r with K SNPs (with pre-calculated Bayes factors $BF_i : i \in 1, \ldots, K$), the regional Bayes factor for model l ($RBF^{(l)}$) is, given weight π_i^l:*

$$RBF_r^{(l)} = \sum_{i=1}^{K} \pi_i^{(l)} BF_i^{(l)}.$$

They then introduce a fourth model, which allows for the region to contain two causal SNPs—one for each trait. Thus, for an indicator function where $I = 1$ if $i \neq j$ and $I = 0$ otherwise,

$$RBF_r^{(4)} = \sum_{i=1}^{K} \sum_{j=1}^{K} \pi_i^1 \pi_j^2 BF_i^1 BF_j^2 I\left(i, j\right).$$

As with PRIMe (Section 1.7.2), this allows for a by-region investigation of the genome for each pair of traits.

While the initial description of the various models assumes the independence of the cohorts in which the traits were studied, this need not be the case. The Bayes factors for Model 3 can be calculated in non-independent studies with some correlation, C between the z scores of the two studies.

Applying this method across 42 traits, they were able to highlight pairs of traits that seemed particularly correlated (positively and negatively) and to find 341 regions of the genome that influenced multiple traits. This study demonstrated a potential strength of a Bayes factor approach in dealing with non-independent studies—the covariance between the studies can be included as part of the model.

This is reminiscent of the work done using LD Score Regression (Section 1.4.3), which was applied across 25 traits and found many pairs of correlated traits known to epidemiologists, and suggested some new pairs of genetic correlations that were novel. This could potentially act as a way of checking genome-wide correlations found by methods like the pairwise Bayes factor, since LD score regression cannot say which regions or markers are driving the genetic correlation.

*In practice, $\pi_i^l = \frac{1}{K}, \forall l$, but this need not be the case.

1.7.11 Bayesian meta-analysis using MCMC [116]

CPBayes, an R package, also performs meta-analysis on GWAS summary statistics. Here, the effect sizes, $\hat{\beta}_1, \ldots, \hat{\beta}_K$ are the observed effects in K studies, and $s_1, \ldots, s_K$ are their corresponding standard errors.

$$\hat{\beta}_i | \beta_i \sim \mathrm{N}\left(\beta_i, s_i\right) \forall i \in 1, \ldots, K.$$

Alternatively, if S is the covariance matrix of

$$\hat{\boldsymbol{\beta}} = \hat{\beta}_1, \ldots, \hat{\beta}_K,$$

and

$$\boldsymbol{\beta} = \beta_1, \ldots, \beta_K,$$

we can formulate the above as

$$\hat{\boldsymbol{\beta}} | \boldsymbol{\beta} \sim \mathrm{MVN}\left(\boldsymbol{\beta}, S\right).$$

The authors use a spike and slab prior [127, 128] on β_i. This depends on the latent variables $Z = z_1, \ldots, z_K$, where an arbitrary element $z_i = 1$ when the ith phenotype is associated with a given SNP and $z_i = 0$ otherwise. Their values depend on the parameter q, which is the proportion of phenotypes that are truly associated with the SNP. The variable q, in turn, is assumed to have a beta distribution with shape parameters c_1, c_2. Additionally, they introduce the parameters τ and d, which—like q—follow a beta distribution with shape parameters e_1, e_2. In practice, the authors chose $c_1 = c_2 = e_1 = e_2$, so that q and d would have uniform distributions over the interval (0,1). Thus,

$$\beta_i | z_i = 1, \tau, d \sim \mathrm{N}\left(0, \left(\frac{\tau}{d}\right)^2\right).$$

Conversely,

$$\beta_i | z_i = 0, \tau, d \sim \mathrm{N}\left(0, \tau^2\right).$$

They use a Markov chain Monte Carlo to generate values of β, Z, q, and d and

sample from the posterior distribution of these parameter values to calculate Bayes factors for a model where $z_i = 1$ for at least one $i \in 1, \ldots, K$ against the null model where $z_i = 0 \forall i \in 1, \ldots, K$. They also calculate the posterior probability of no association (PPNA) of the SNP on each individual phenotype. This is based on the posterior probability of association (PPA) (see 1.5 and Equation 1.17 [93]), and both are based on the posterior odds, which is, in turn, the product of the Bayes factor and the prior odds. That is, PPNA $= 1 -$ PPA.

This method can account for covariance due to things like shared controls with the covariance matrix S. It can also select the subset of phenotypes most likely to be associated with a SNP. However, while it might be useful to set a prior on the correlation between true effects, this method does not, which may result in a loss of power compared to those that do. The initial GWASs were done on a cohort of 22 phenotypes in 53,809 individuals of European descent who had been genotyped at 601,175 autosomal SNPs and this method identified only nine independent loci associated with more than one phenotype. There may also be problems in scaling to large numbers of phenotypes and large numbers of SNPs; the analysis of 1000 SNPs across the 22 phenotypes took 4.5 hours of computation time.

1.7.12 Multi-Trait Analysis of GWAS (MTAG) [118]

The authors of Mulit-Trait Analysis of GWAS (MTAG) describe it as "a generalization of standard, inverse-variance-weighted meta-analysis" [118]. The authors start with a set of T correlated traits, which have all been normalized to have a mean of 0 and a variance of 1.

Like many of the other methods discussed in this chapter, as well as our own, this method assumes that the vector of true effect sizes of the jth SNP on the T traits, $\boldsymbol{\beta}_j$, is distributed multivariate normal with a mean of $\mathbf{0}$ (the zero vector) and variance-covariance matrix $\boldsymbol{\Omega}$.[†] Similarly, the observed effect sizes, the vector $\hat{\boldsymbol{\beta}}_j$, are assumed to come from a multivariate normal distribution with mean vector $\boldsymbol{\beta}_j$ and variance-covariance matrix $\boldsymbol{\Sigma}$, which is calculated using the number of samples in the study and the intercept of the pairwise LD score regression (see Section 1.4.3).

[†]The notation used by the MTAG authors is very similar to the notation we use when describing our method in Chapter 2, however our definitions of the matrices $\boldsymbol{\Omega}$ and $\boldsymbol{\Sigma}$ reversed from the ones described here.

Because $\mathbf{\Sigma}$ represents the covariance due to sampling error, this matrix may vary from SNP to SNP due to things like variation in sample size at each marker. So for the jth SNP, the covariance matrix of $\hat{\boldsymbol{\beta}}_j|\boldsymbol{\beta}_j$ is $\mathbf{\Sigma}_j$.

The matrix $\mathbf{\Omega}$ is estimated using $\mathbf{\Sigma}_j$ and the fact that the marginal density of $\hat{\boldsymbol{\beta}}_j \sim MVN\left(\mathbf{0}, \mathbf{\Omega} + \mathbf{\Sigma}_j\right)$. The corresponding log-likelihood $\ell\left(\mathbf{\Omega}; \hat{\boldsymbol{\beta}}_j, \mathbf{\Sigma}_j\right)$ is maximized either numerically or using a closed form that assumes a constant sample size at each of the SNPs.

The authors then define weights based on the variance of the tth study, and show that $\hat{\boldsymbol{\beta}}_j|\beta_{j,t}$ is distributed multivariate normal about a mean of $\beta_{j,t}$ scaled by the weights and with a variance based on a linear combination of $\mathbf{\Omega}$, $\mathbf{\Sigma_j}$, and a quantity representing the vector of weights scaled by the variance of $\hat{\beta}_{j,t}$. By maximizing this marginal likelihood with respect to $\beta_{j,t}$, they define their MTAG estimator, $\hat{\beta}_{\text{MTAG},j,t}$. This is the weighted sum of $\hat{\boldsymbol{\beta}}_\boldsymbol{j}$ and is constructed to account for covariance due to shared genetic effects as well as covariance due to shared controls. When the studies in the meta-analysis are of the same trait, $\hat{\beta}_{\text{MTAG},j,t}$ and its standard error simplify to the fixed effects meta-analysis test statistic and standard error (see Equations 1.19 and 1.20).

The use of LD score regression to calculate $\mathbf{\Sigma}_j$ allows this method to account for cryptic covariance between studies, which can occur due to things like sharing of samples or population structure. Another strength of this method is the ability to estimate the true effect size of a given study using information from studies of related traits. By combining data in this way, true signals that fail to achieve genome-wide significance in their original studies may be discovered.

1.8 Our work

We see that while there are many methods to find cross-phenotype associations, few of them are able to do so on a SNP by SNP basis or make predictions about the pattern of phenotypic association at a given SNP. Furthermore, most of these methods require that the studies in the meta-analysis be completely independent. As more and more work is carried out by consortia and using data from biobanks, this requirement becomes less realistic, as these large groups often share controls and biobanks store

data on multiple phenotypes in the same set of individuals.

This thesis proposes a statistical method to search for cross-phenotype associations in a Bayesian setting using only summary statistics from genome-wide association studies. While we do not claim that this method is vastly better than any other method[‡], there are some appealing features of our method that we highlight here. For one, summary statistics are easily more shared than raw, since there is no risk of identifying study participants from summary data. Furthermore, our method allows for GWAS SNPs—as opposed to regions—to be investigated in a systematic way across as many theoretical phenotypes as we wish, though computational expediency may impose a limit in reality. Furthermore, there is no requirement for our phenotypes to be related at all, though phenotypes with expected genetic correlations will increase our power to detect associations. Unlike previous methods, our method provides the opportunity to say something about the probability of any individual phenotype being associated with a given SNP, as well as the probability of a given subset of phenotypes being associated.

Those studying cross-phenotype associations do so in the hope of learning more about the genetic relationship between different phenotypes, while the analogous relationships in transethnic studies—the genetic relationship between study populations—is something that is more easily calculated, if not understood. Because of this, the tests may not always be interchangeable because transethnic studies aim to identify a robust causal variant rather than to learn something new about the relationships between the ethnicities used. In fact, the relationship between ethnicities may already be well-characterized and included in the model [104, 105].

So while our method and many others may be applicable to transethnic meta-analyses, and vice versa, we will not be discussing these applications, as transethnic meta-analyses on a single phenotype are beyond the scope of this thesis. We note, however, that the meta-analysis performed on the Wellcome Trust Case Control Consortium in Chapter 4 searches for cross-phenotype associations across several continental populations.

We introduce our method in Chapter 2 and investigate its statistical properties in simulated data. We first apply it—along with four other methods described in

[‡]Indeed, the metric for defining what method is the "best" is far too dependent on the aims of the researcher and the data available for there to be one universal best method.

this chapter—to a small dataset in Chapter 3, allowing us to compare our method to others in terms of both speed and results in real data. In Chapter 4, we perform a meta-analysis on data from 14 Wellcome Trust Case Control Consortium 2 (WTCCC2) studies to investigate the method's scalability to diverse, genome-wide data. In Chapter 5, we apply it to more autoimmune disease data from Immunobase, and investigate the possibility of using an empirical prior on models of associations. Additionally, we use the datasets in Chapters 3 and 5 to investigate the question raised by the omnigenic model in Section 1.2: do we see that a large effect on one trait correlates to more effects detected in others? In each of our applications to real data, we examine the properties of the datasets we use and their effects on the method's performance and conclusions. Finally, we summarize the pros, cons, and future directions for this kind of research in Chapter 6.

This thesis would not have been possible without input from several people, whose contributions to this work are summarized in Table 1.2 below.

Table 1.2: Table of collaborators and their contributions to the work discussed in this thesis.

Collaborator	Contributions
Luke Jostins	Performed the alignment, data quality control, and summary statistic imputation on the Immunobase data before sharing with me. Shared his previous work on liability models with me and wrote initial code to simulate data under these models.
Matti Pirinen	Developed the method of calculating the p-value for the approximate Bayes factor analytically, using quadratic form of the matrix of effect size estimates, $\hat{\boldsymbol{\beta}}^{\mathsf{T}}$, the study matrix $\boldsymbol{\Omega}$ and the prior matrix $\boldsymbol{\Sigma}$. He also provided the R code for this calculation.
Chris Spencer	Performed the initial analysis of the WTCCC2 autoimmune diseases, which I recreated. The explanation of Wakefield's approximate Bayes factor is based on summaries Chris has given throughout the course of my DPhil studies.

Chapter 2

The Method

2.1 Introduction

As we saw in Chapter 1, there are a number of approaches to meta-analyzing genome-wide association studies. Though many approaches make use of summary statistics—which are shared and stored more easily than raw genetic data—none of them make much use of prior information that we might have about our variants, studies, or traits. Additionally, several methods lack closed-form solutions and require iterative procedures to arrive at their final result. In this thesis, we introduce a method that overcomes some of these drawbacks in terms of speed of calculation and the interpretation of results. There will never be one perfect meta-analysis method, and there will be situations when another method is more appropriate, but we believe that our method can be used in meta-analyses to discover associations that might otherwise have been overlooked and to discover possible cross-trait associations.

We build on the work of Wakefield to create our own methods of assessing patterns of association at a single SNP across multiple phenotypes. We introduce two main ways of testing for multiple phenotypic associations at a given SNP: the all-correlated approximate Bayes factor and the subset exhaustive approach. We use simulated data to to investigate the properties of these two approaches. We will also discuss how we generated data using a full logistic regression and compare it to a much faster method that generates the summary statistics from a normal distribution directly. We describe the different sets of simulated data we generated and use them to investigate the power of our methods to detect effects compared to standard fixed

and random effects meta-analyses.

The subset exhaustive framework affords us the opportunity to select the model that best fits the data. We look at how well this approach matches the true underlying data, even when priors that match the data poorly are used. This is important to understand since our results in subsequent chapters involve trying to learn about the relationships among diseases, so we need to be confident in our initial inference about patterns of association.

Additionally, we investigate the calibration of posterior probabilities on each possible model of phenotypic association calculated from the approximate Bayes factors. Posterior probabilities are more readily interpreted than Bayes factors, especially across SNPs, however we need to be certain that a posterior probability of, for example 50% on a model (or as a marginal probability on a disease's association with a given marker) does in fact mean that 50% of all models with that probability are the true underlying model (or that 50% of all diseases with that marginal probability are truly associated with the marker). If these probabilities are not well calibrated, then the calculated posterior probability may over- or under-estimate the true probability of a model being the true model or of a disease being truly associated with the marker.

2.2 Our approach

In Chapter 1, Section 1.5.1, we defined Bayes factors and Wakefield's approximation for genome-wide association studies [87]. Briefly, a Bayes factor is defined as the ratio of two likelihoods. Let M_0 be the null hypothesis and M_1 be an alternative hypothesis (note that M_1 may be one of several alternative hypotheses available). Then the Bayes factor, BF, of the evidence in favor of M_1 against the null is simply

$$BF = \frac{\mathrm{P}\left(data|M_1\right)}{\mathrm{P}\left(data|M_0\right)}. \tag{2.1}$$

Because the likelihoods in Bayes factors involve integrals of functions that may not have closed-forms, they are often approximated by iterative procedures, such as Markov chain Monte Carlo. For genome-wide association studies, Wakefield introduces an approximation, the ABF, defined as the ratio of two normal densities. Let $f\left(a;\mu,\tau^2\right)$ be the normal density function about mean μ and variance τ^2, evaluated

at a. We define $\hat{\beta}$ to be the observed effect size from a genome-wide association study and V to be the square of the standard error of $\hat{\beta}$. We define some prior on the true effect size, σ^2, and

$$ABF = \frac{f\left(\hat{\beta}; 0, V + \sigma^2\right)}{f\left(\hat{\beta}; 0, V\right)}. \tag{2.2}$$

2.2.1 Extending to the multivariate case

We can easily extend Wakefield's method to the multivariate case. We simply allow $\hat{\boldsymbol{\beta}}$ to be the vector of effect sizes at the same SNP across n studies. Similarly, $SE^2_{\hat{\beta}}$ becomes the covariance matrix $\boldsymbol{\Omega}_{\hat{\boldsymbol{\beta}}}$. In this thesis, $\boldsymbol{\Omega}_{\hat{\boldsymbol{\beta}}}$ is assumed to represent the cryptic relatedness between studies due to shared controls. Thus, when the studies in the meta-analysis do not share samples, this matrix is diagonal. However, it is possible that it could also capture environmental effects shared between the traits and other covariates not accounted for by the regression model. A basic assumption of our method is that effect size estimates reported in the summary statistics are good approximations of the true effect sizes, which implies that all covariates have been appropriately accounted for. While this chapter explores the effect of misspecifying the prior, we do not explore the results of the violation of the assumption about the accuracy of the summary statistics.

We choose the prior standard deviation, σ_i, for each study as well as $\rho_{i,j} : i, j \in \{1, \ldots, n : i \neq j\}$, the pairwise correlations between the effect size estimates of each study to create a prior covariance matrix, $\boldsymbol{\Sigma}$, on the effect sizes. Finally, instead of evaluating these parameters in the normal density function f, we use the multivariate normal density function, MVN. This extension of the approach allows us to compare SNP effects across studies.

That is, for $\upsilon_{i,j}$ the pairwise correlation between studies i and j (due to factors such as shared samples), we have

$$\boldsymbol{\Omega}_{\hat{\boldsymbol{\beta}}} = \begin{bmatrix} SE_1^2 & \upsilon_{1,2}SE_1SE_2 & \cdots & \upsilon_{1,n}SE_1SE_n \\ \upsilon_{1,2}SE_1SE_2 & SE_2^2 & \cdots & \upsilon_{2,n}SE_2SE_n \\ \vdots & \vdots & \ddots & \vdots \\ \upsilon_{1,n}SE_1SE_n & \upsilon_{2,n}SE_2SE_n & \cdots & SE_n^2 \end{bmatrix} \tag{2.3}$$

and for $\rho_{i,j}$, the pairwise prior correlation between studies i and j (due to expected genetic relationship between phenotypes), and σ_i, the prior standard deviation on the true effect size β_i we have

$$\boldsymbol{\Sigma} = \begin{bmatrix} \sigma_1^2 & \rho_{1,2}\sigma_1\sigma_2 & \cdots & \rho_{1,n}\sigma_1\sigma_n \\ \rho_{1,2}\sigma_1\sigma_2 & \sigma_2^2 & \cdots & \rho_{2,n}\sigma_2\sigma_n \\ \vdots & \vdots & \ddots & \vdots \\ \rho_{1,n}\sigma_1\sigma_n & \rho_{2,n}\sigma_2\sigma_n & \cdots & \sigma_n^2 \end{bmatrix} \tag{2.4}$$

Thus, for $\mathbf{0}$, the vector of zeroes, the multivariate extension of the ABF may be written more formally as

$$ABF = \frac{MVN(\hat{\boldsymbol{\beta}}; \mathbf{0}, \boldsymbol{\Omega}_{\hat{\boldsymbol{\beta}}} + \boldsymbol{\Sigma})}{MVN(\hat{\boldsymbol{\beta}}; \mathbf{0}, \boldsymbol{\Omega}_{\hat{\boldsymbol{\beta}}})}, \tag{2.5}$$

or, equivalently,

$$ABF = \sqrt{\frac{|\boldsymbol{\Omega}|}{|\boldsymbol{\Omega} + \boldsymbol{\Sigma}|}} \exp\left\{\frac{1}{2}\left(\hat{\boldsymbol{\beta}}^{\intercal}\boldsymbol{\Omega}^{-1}\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}^{\intercal}\left(\boldsymbol{\Omega} + \boldsymbol{\Sigma}\right)^{-1}\hat{\boldsymbol{\beta}}\right)\right\}. \tag{2.6}$$

2.2.2 The All-Correlated Bayes Factor

Our first approach tests a model of no shared effects against a model of shared effects across all phenotypes. The advantage of this approach is that we can convert the resulting Bayes factor back into a p-value for better comparison to other methods.

The conversion works by comparing the the Bayes factor against a null distribution. Alternatively, Matti Pirinen developed a way of approximating the p-value using the distribution function of the quadratic form of our variables, $\hat{\boldsymbol{\beta}}$, $\boldsymbol{\Omega}$, and $\boldsymbol{\Sigma}$. That is,

$$Q\left(\hat{\boldsymbol{\beta}}; \boldsymbol{\Omega}, \boldsymbol{\Sigma}\right) = \hat{\boldsymbol{\beta}}^{\intercal}\left(\boldsymbol{\Omega}^{-1} - \left(\boldsymbol{\Omega} + \boldsymbol{\Sigma}\right)^{-1}\right)\hat{\boldsymbol{\beta}}. \tag{2.7}$$

From here, the n eigenvalues, λ_i of the matrix $\boldsymbol{I} - \left(\left(\boldsymbol{\Omega} + \boldsymbol{\Sigma}\right)^{-1} \times \boldsymbol{\Omega}\right)$ are calculated, and $Q\left(\hat{\boldsymbol{\beta}}; \boldsymbol{\Omega}, \boldsymbol{\Sigma}\right) \sim \sum_i^n \lambda_i \chi_1^2$. The p-values associated with this mixture of χ_1^2 distributions can be evaluated in R [120], using the 'CompQuadForm' package [129].

2.2.3 Priors

The prior σ encodes a belief in the true effect sizes. The prior assumption is that these are drawn from a normal distribution centered on 0, and with variance σ^2, so the larger σ is, the lower the probability density is on true effect sizes near zero, and the higher it is on more extreme ones. We note that when observed effects are large, ABFs are high, even using a low value of σ, because the alternative model is still a better fit to the data than the null; however neither is as good a fit as the model with a larger value for σ. As a result, choosing σ is a matter of deciding what the smallest believable non-zero effect size is, for a given dataset. Naturally, this needs to be balanced against a potential increase in false positives, as associations are called for variants whose effect sizes are closer to zero, some of which may be due to sampling noise around zero rather than a true non-zero effect.

In the literature, the values of σ commonly used are 0.2 and 0.4 [94, 93]. Setting $\sigma = 0.2$ means that 95% of true effects are expected to have an odds ratio between 0.68 and 1.48, and 99% will be between 0.60 and 1.67. When σ is raised to 0.4, 95% of true effect sizes are expected to have odds ratios between 0.46 and 2.19, and 99% of odds ratios are expected to fall between 0.36 and 2.80.

The prior correlation matrix used in these methods can reflect a number of different models which might be more or less desirable depending on the data. We discuss these below.

Fixed effects

If we believe the assumptions of a standard, fixed-effects meta-analysis, then the appropriate prior says that the effect sizes are all completely correlated with each other. That is, for n studies $\rho_{i,j} = 1, \forall i, j \in \{1, \ldots, n\}$.

Suppose that for a given prior σ, one calculated the all-correlated Bayes factor for n studies using this fixed-effects prior. Now suppose one calculated the standard fixed-effects standard error and effect size for these n studies and then used these values to calculate Wakefield's univariate Bayes factor (equation 1.15), using the same value of σ as before. These two Bayes factors would be equal.

Independent effects

Going to the other extreme, where $\rho_{i,j} = 0, i \neq j$ indicates a model of independent effects—that is, the effect of the SNP in a given study is completely independent of its effects in the others. Calculating the all-correlated Bayes Factor with this prior is equivalent to calculating Wakefield's approximate Bayes factor for each study and then multiplying these ABFs together. A proof of this can be found in the appendix, Section 2.B.

Correlated and subset priors

Additionally, the prior matrix $\mathbf{\Sigma}$ can contain correlation coefficients, $\rho_{i,j}$, that vary between -1 and 1 as well as differing from each other. In practice, we often set $\rho_{i,j} = c$ for some constant c, however there is no reason that each correlation coefficient, $\rho_{i,j}$ cannot be unique. These can be estimated from the data directly or chosen based on prior belief in the similarity of effect sizes. The advantage of these priors over traditional meta-analysis is that they should have better power to detect true effects, even when the individual studies fail to pass traditional frequentist significance thresholds.

Subset priors are simply those which encode the belief that a proper subset of the studies show an effect at a particular SNP. This is achieved by setting to zero the prior variances on effect sizes of studies that are assumed not to show an effect. That is for the full set of studies, $S = \{S_1, S_2, \ldots, S_n\}$, and the subset $T \subset S$, which is the set of studies assumed to be truly associated with the marker, then the prior variances on effect sizes for $S \setminus T$ is uniformly $\sigma_{S \setminus T} = 0$.

2.2.4 Subset-exhaustive ABF approach

For n phenotypes in the meta-analysis, the subset-exhaustive ABF approach involves 2^n calculations of the approximate Bayes factor. For a given set of studies, we compare the models of every possible combination of affected phenotypes against the null model. This necessarily includes the all-correlated model. We briefly consider the various models we test, M_j. Here, the models are simply the possible patterns of association of a set of phenotypes to a SNP. Given n phenotypes, we can write all 2^n

possible patterns of association as binary strings, where a 1 in the ith position indicates an association with the ith phenotype. We provide an example of all possible strings for $n = 3$ in Table 2.1. In this table, the first row represents the null model and the last row represents the all-correlated model.

0	0	0
0	0	1
0	1	0
0	1	1
1	0	0
1	0	1
1	1	0
1	1	1

Table 2.1: The full set of binary strings for $n = 3$.

We may also incorporate a prior weight on each model—for instance, we might want to put a high weight on the null model. We then can multiply the Bayes factors corresponding to each model by their respective weights. From there, we can choose the model with the highest Bayes Factor as the most likely and see by how much it outperforms the second and third best models, as well as normalize the Bayes factors to get a posterior distribution on the set of models (see Sections 2.3 and 2.6 for more information). We can also take the mean over these Bayes factors and use that as a test statistic, similar to the all-correlated Bayes factor. That is, for n phenotypes, there are 2^n possible models. Let weight w_i correspond to model i with Bayes factor BF_i, then the mean is

$$BF_{mean} = \frac{\sum_{i=1}^{n} w_i BF_i}{\sum_{i=1}^{n} w_i}. \tag{2.8}$$

2.3 Posterior probabilities and model averaging

If we have performed the exhaustive Bayes factor analysis of the associations of n phenotypes with a SNP, then it is possible to obtain the posterior probabilities $p_1, \ldots, p_n$ on each model. Associated with these models is some sort of prior probability or

weight $(w_1, \ldots, w_n)$. This may be flat, that is

$$w_i = \frac{1}{2^n} : i \in \{1, \ldots, n\},$$

since there are 2^n possible models of association for n phenotypes. Alternatively, we might want there to be an equal probability on the number of associated phenotypes, in which case, for k associated phenotypes in model i we have

$$w_i = \frac{1}{\binom{n}{k}}.$$

Other priors we might put on the models are ones that assume that there is some probability q of a phenotype being associated with the SNP. Therefore, for k associated phenotypes in model i, we have

$$w_i = q^k (1-q)^{n-k}.$$

When $q = 0.5$, this prior is equivalent to the flat prior, above. We discuss and investigate the above three priors when applied to the dataset in Chapter 3—see Sections 3.3 and 3.4.2 of that chapter.

Finally, we might simply believe *a priori* that there is a high probability that no phenotype is associated with the SNP, but that there is a flat prior on all the other models. This is because there is a bound on the value that Bayes factors in favor of the null model can take under normal priors. Johnson and Rossell (2010) [130] explore alternatives that allow these to be unbounded, as they are for Bayes factors in favor of an alternative hypothesis. To achieve this, we can simply apply a weight of v on the null model to reflect that it is v times more likely to be the true model than any of the others and a leave $w_i = 1$ for all the other models. This also demonstrates that the weights over all models need not sum to 1, though they do for all of the previously discussed priors.

Once we have determined our weights, we can calculate the posterior probabilities, p_i, on each model by simply calculating

$$p_i = \frac{w_i ABF_i}{\sum_{i=1}^{2^n} w_i ABF_i}. \tag{2.9}$$

Before we can treat these p_i as probabilities, we need to check that they are well-calibrated—that is, models with posterior probability p will be the correct model that generated the data p proportion of the time. We investigate this in Section 2.6.

2.4 Generating simulated data

To test the frequentist properties of our method, we performed our analyses on summary statistics from millions of replicates of a SNP from a simulated GWAS.

The code required the following variables to be set upon initialization: the effect allele frequency (f), the prior correlation coefficient for analysis (ρ), the prior standard error for analysis (σ), the effect allele frequency (f), relative risk (RR), number of cases (n_{cases}), and number of controls ($n_{controls}$).

2.4.1 Outline of how simulated data were generated

The simulated datasets of effect size estimates and their standard errors were generated in R. Below is an outline of the process used to create them.

Step 1: Use the **rmvnorm** function to create the vector of potential real effects ($\boldsymbol{\beta}$) for all phenotypes at each SNP. To get relative risk (of the heterozygote compared to the homozygote for the major allele) at the ith phenotype RR_i, we have

$$RR_i = e^{\beta_i}.$$

Step 2: A model of the underlying true effects is chosen uniformly from the set of all possible models. This is represented as a binary string, similar to those shown in Table 2.1.

Step 3: Calculate the vector *gamma* ($\boldsymbol{\gamma}$) to be used when calculating the effect allele frequency in cases:

$$\boldsymbol{\gamma} = (1-f)^2 + 2f(1-f)\boldsymbol{RR} + \boldsymbol{RR}^2 f^2$$

Step 4: Calculate the vector allele frequency in controls, $f_{control0}, f_{control1}, f_{control2}$ for controls with 0, 1, and 2 effect alleles at the simulated SNP.

$$\boldsymbol{f_{control0}} = (1-f)^2$$

$$\boldsymbol{f_{control1}} = 2f(1-f)$$

$$\boldsymbol{f_{control2}} = f^2$$

Step 5: Calculate the vector of allele frequency in cases, $f_{case0}, f_{case1}, f_{case2}$ for cases with 0, 1, and 2 effect alleles at a given SNP.

$$\boldsymbol{f_{case0}} = \frac{(1-f)^2}{\boldsymbol{\gamma}}$$

$$\boldsymbol{f_{case1}} = \frac{2f(1-f)\boldsymbol{RR}}{\boldsymbol{\gamma}}$$

$$\boldsymbol{f_{case2}} = \frac{f^2\boldsymbol{RR}^2}{\boldsymbol{\gamma}}$$

Step 6: At each phenotype that is affected by the SNP in the true underlying model (chosen in step 1), we simulate allele counts for each "person" in our study, using the frequencies in cases and controls calculated in the previous two steps. This is done using the **sample** function in R.

Step 7: For phenotypes not affected by the SNP, allele counts are simulated in cases and controls using the frequencies in controls only.

Step 8: For each SNP at each phenotype, we perform a logistic regression on the data created in steps 6 and 7 and store the estimated effect sizes and standard errors.

This method works by simulating the data needed to perform a logistic regression and then performing it. Our simulation analyses were done in R, which gets p-values for the null hypothesis of $\hat{\beta}_G = 0$ from a z-test. If we wanted to simulate a less ideal scenario, such as shared controls, then we simply sample the controls (or a subset of them) once and use those data in each of the logistic regression models.

2.4.2 Simulating data from the normal distribution directly

In order to speed up the simulation procedure, we investigated generating the effect size estimates, $\hat{\beta}_i$ from the normal and multivariate normal distributions directly. The procedure was as follows:

Step 1: Define the characteristics of the study being simulated: number of cases (n_{cases}); number of controls ($n_{controls}$); true effect sizes (β), which can be set at some constant or drawn from a normal distribution centered on 0, with a standard deviation of σ_{true}; frequency of the effect allele (f). In the multivariate case, the effect sizes are drawn from a multivariate normal distribution centered on the vector $\mathbf{0}$, with a covariance matrix Σ, with diagonal elements σ^2_{true} and off-diagonal elements $\rho_{i,j}\sigma^2_{true}$, where ρ reflects the predefined, true underlying genetic correlation between phenotypes.

Step 2: We define n, the full sample size, and ϕ as

$$n = n_{cases} + n_{controls}$$

$$\phi = \frac{n_{cases}}{n}$$

Use the above to calculate the estimated standard error (SE) for the data using the formula

$$SE = \frac{1}{\sqrt{2nf(1-f)\phi(1-\phi)}}. \tag{2.10}$$

Step 3: From here, we can simulate observed effect sizes by sampling from a normal distribution (or a multivariate normal distribution), centered on β and with a standard error that is uniformly SE for all phenotypes. In the multivariate case, we can define correlations $\upsilon_{i,j} : i \neq j$ between phenotypes i and j and create a covariance matrix Σ where $\Sigma_{i,j} = \upsilon_{i,j}SE^2$ in the off-diagonal elements, and

$\Sigma_{i,i} = SE^2$. We generate the observed effect sizes by sampling from a multivariate normal distribution centered on the vector $\mathbf{0}$ and with a covariance matrix of $\Sigma + \Omega_{\hat{\beta}}$.

Step 4: To generate p-values, we perform a χ^2_1 test on the statistic $\left(\frac{\hat{\beta}}{SE}\right)^2$.

This method can also be adjusted to simulate situations with non-zero covariance between studies. In that instance, we need to calculate the expected correlation between studies with shared controls (or cases). Given

n_{0S}, the number of shared controls,

n_{1S}, the number of shared cases,

n_{0i}, the number of non-shared controls in study i,

n_{0j}, the number of non-shared controls in study j,

n_{1i}, the number of non-shared cases in study i, and

n_{1j}, the number of non-shared cases in study j,

then according to Zaykin *et. al.* [131] $\upsilon_{i,j}$, the correlation coefficient of the effect sizes of these studies under the null is

$$\upsilon_{i,j} = \frac{\frac{n_{0S}}{(n_{0j}+n_{0S})(n_{0i}+n_{0S})} + \frac{n_{1S}}{(n_{1S}+n_{1i})(n_{1S}+n_{1j})}}{\sqrt{\left(\frac{1}{n_{0i}+n_{0S}} + \frac{1}{n_{1S}+n_{1i}}\right)\left(\frac{1}{n_{0j}+n_{0S}} + \frac{1}{n_{1S}+n_{1j}}\right)}}. \tag{2.11}$$

Thus, the effect sizes are generated using a multivariate normal distribution, centered on the zero vector and with a covariance matrix that reflects both the underlying SNP effects and the covariance due to shared controls.

This is a very fast process, which can generate a million simulated data points in 0.0008 seconds, while the process outlined in Section 2.4 can take hours to generate the same number of data points and requires parallelization to implement on a large scale. It is therefore necessary to compare the two methods, to see if the process of generating the $\hat{\beta}_i$ directly creates the same distribution of test statistics and p-values as in the data generated from running the full logistic regression and performing a t-test.

Comparison of the two simulation methods

To start, we generated datasets of 100,000 replicants under both simulation schemes using the same parameters:

- Odds ratio, which took the values of 1, 1.1, 1.3, and 1.5.
- Study size, which took the values of 1000, 4000, and 14,000. in all studies, we assumed a 1:1 ratio of cases and controls.
- Minor allele frequency, which took the values of 0.5, 0.1, and 0.01, but was the same across all studies.

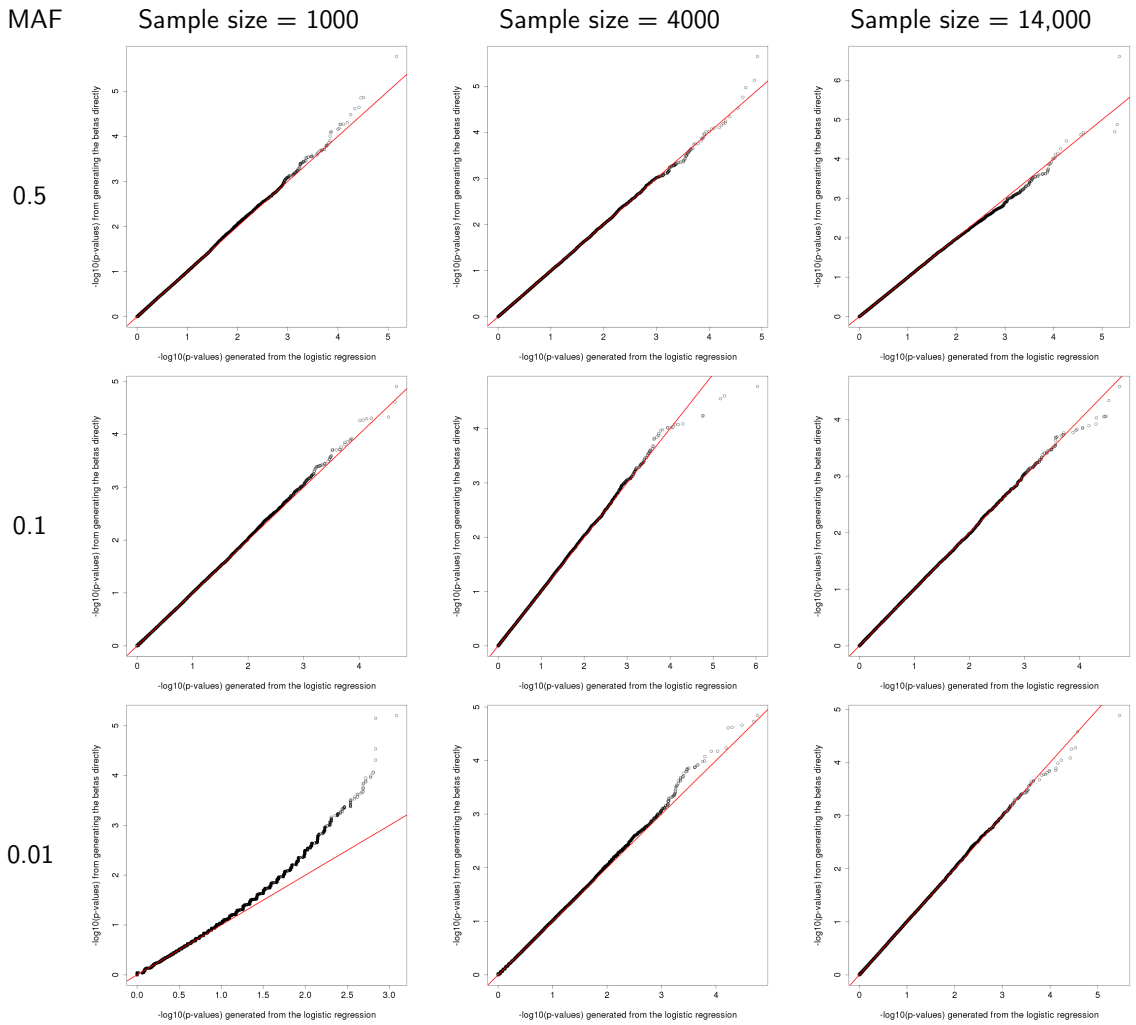
Figures 2.1 and 2.2 show the quantile-quantile (Q-Q) plots of the distribution of p-values for $\hat{\beta}$ and SE generated using the full logistic regression (x-axis) and for those generated from a normal distribution (y-axis). Ideally, we'd like to see these on the line $y = x$ (red line).

The most obvious feature of Figures 2.1 and 2.2 is that for all effect sizes, when the minor allele frequency is low (0.01) and the sample size is small (1000), the distribution of p-values generated from sampling from a normal distribution stochastically dominates the distribution generated from the logistic model. The next most striking thing is that for the largest effect sizes—with an odds ratios of 1.5 (Figure 2.2)—we see that one distribution frequently dominates the other stochastically and that the simulated p-values are rarely on the line $y = x$. In fact, for a study size of 1000 and with a minor allele frequency of 0.1, we see that the points are roughly below the $y = x$ line until about $x = 7$, when the points cross over and are from then on above. These patterns of deviation from the line $y = x$ could be seen when the true underlying OR was as low as 1.3.

What this suggests is that generating data directly from the normal distribution is probably not sufficient when the minor allele frequency is small or when effect sizes are large, but will work the rest of the time. However, these results are already slightly worrisome, as we would like to generate datasets where the effect sizes are drawn from a distribution. What this set of simulations tells us is that we should exercise caution when doing this, since the scheme that samples effect sizes directly does not produce results that are consistent with the simulation scheme that generates allele counts and performs the logistic regression.

We were able to obtain some improvements in the Q-Q plots when we changed

Figure 2.1: Q-Q plots on the distribution of p-values generated using the full logistic regression (x-axis) against the distribution generated from the normal distribution (y-axis), when the true OR = 1. The red line is the line $y = x$.

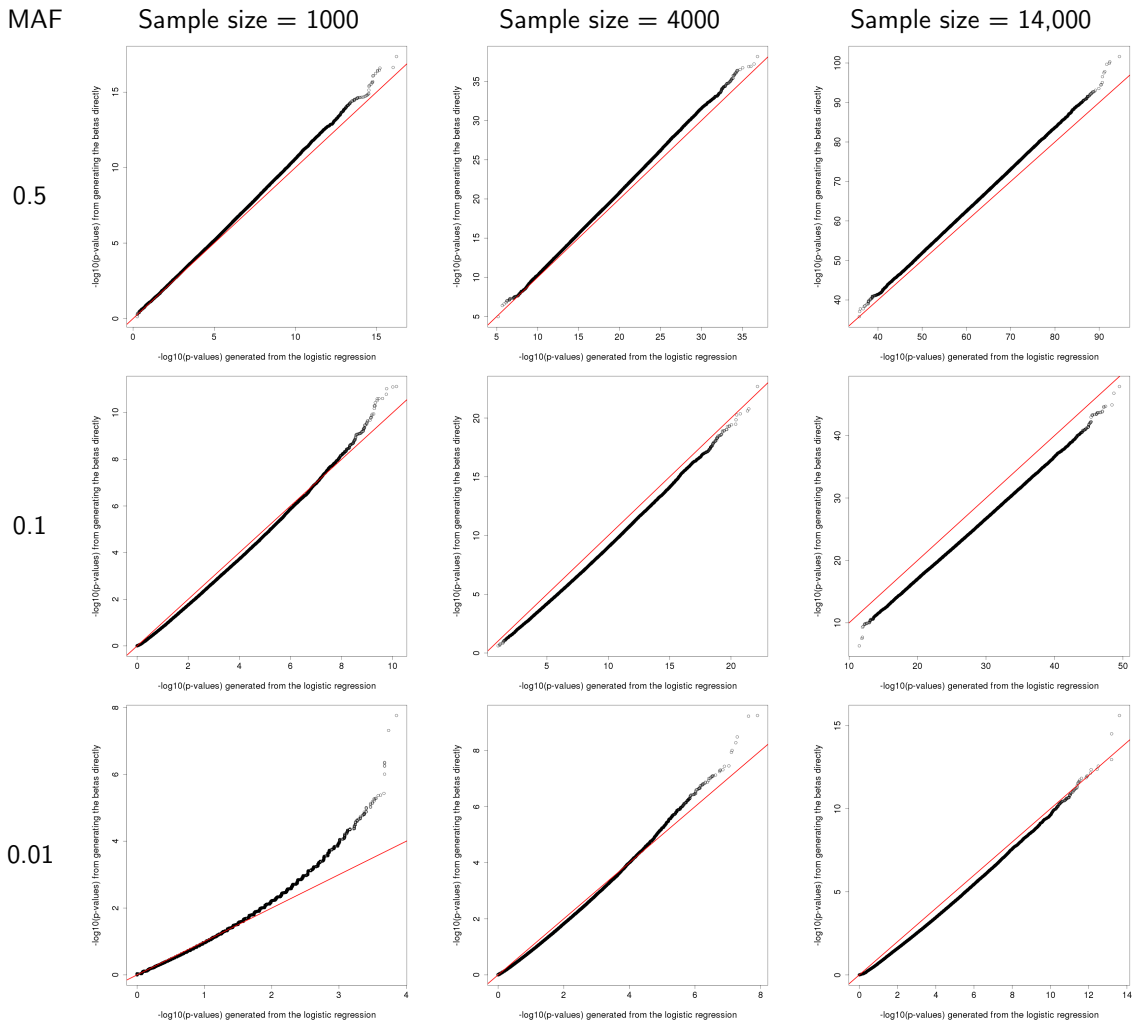


from a z-test in the logistic regression to a likelihood ratio test, comparing $\beta_G = \hat{\beta}_G$ to $\beta_G = 0$. This removed the upward curve from the line $y = x$ in most cases. However, the systematic deviations from the line $y = x$ remain.

Comparison of the multivariate case

We generated datasets for five independent studies under these two simulation schemes. The patterns observed in the multivariate data matched the univariate data very closely: the two methods gave roughly the same distribution of p-values when there was no true effect, but as effect sizes increased—even when only one of the five studies had a true effect—the distributions under the two simulation methods deviated from one another, as demonstrated in Figure 2.3. As we expected, these deviations became

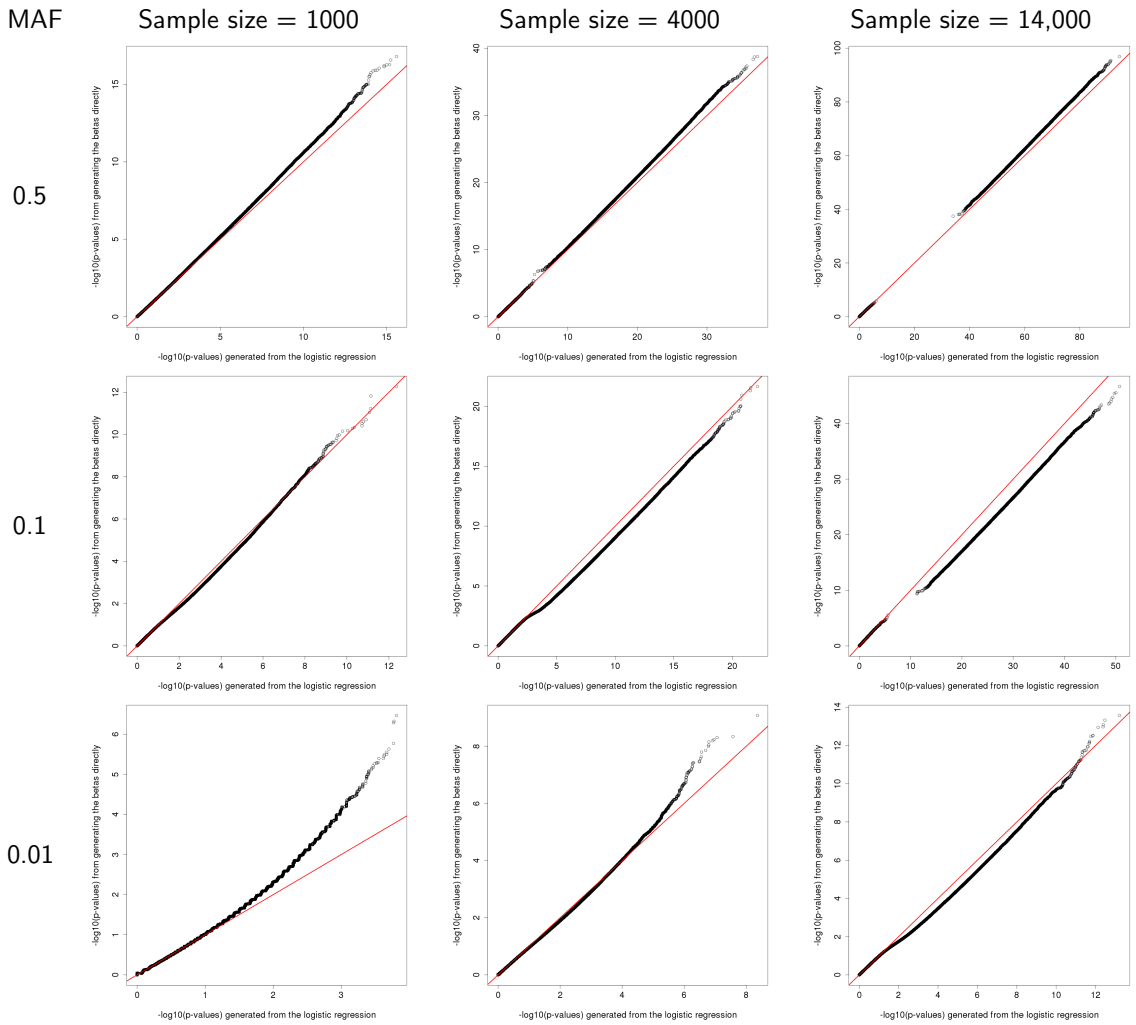
Figure 2.2: Q-Q plots on the distribution of p-values generated using the full logistic regression (x-axis) against the distribution generated from the normal distribution (y-axis), when the true OR = 1.5. The red line is the line $y = x$.



more pronounced when more studies showed true effects.

Once again, this tells us that we need to be careful about treating these two simulation methods as equal, especially if effect sizes are large (or drawn from a distribution that would allow for large effect sizes). However, when generating a null distribution for a large study with a large MAF, we probably can use the faster approach with no loss of accuracy in the distribution. This gain is not trivial, since generating null distributions allows us to understand the properties of Bayes factors and map them back to things like p-values, or simply compare the distribution of ABFs from the data against the expected distribution under the null, as we do in Chapters 4 (Section 4.3, Figure 4.3) and 5 (Section 5.3.1, Figure 5.2).

Figure 2.3: Q-Q plots on the distribution of p-values generated using the full logistic regression (x-axis) against the distribution generated from the normal distribution (y-axis) where RR = 1.5 for one phenotype and 1 for the rest. The red line is the line $y = x$.



2.4.3 Description of the simulations run

Using the code for simulating the full logistic regression described in Sections 2.4 and 2.4.1, we generated several batches of simulations. In all batches, the SNP had a minor allele frequency (MAF) of 0.5 and was associated with up to five phenotypes or studies, which we will call P1, P2,..., P5. For the parameters we did vary, such as effect sizes and true underlying models of association, we describe our processes for generating the data below. We do not intend for these simulation runs to be exhaustive over all the possible types of datasets we might encounter—for a start, this is impossible, as any single parameter used to generate these data can take an infinite number of values. Our goal in using these simulations is to learn about some

of the frequentist properties of the method.

Fixed effect sizes In these batches, we set a true underlying effect size for all phenotypes associated with the marker. The true underlying effect size for all unassociated phenotypes was 0. The effect was the natural logarithm of the relative risk of the heterozygote against the homozygote that did not have the effect allele. All effects were simulated to be additive. In these simulation runs, the effect sizes in truly associated SNPs were uniform. Once these were chosen, genotype counts for a given number of cases and controls were simulated and a logistic regression was performed to generate observed effect sizes and standard errors.

Random effect sizes In these simulations, the true underlying effect for each affected phenotype at each SNP was drawn randomly from a normal distribution centered on 0 with some standard deviation, σ_{gen}. Where we wanted the true effect sizes to be correlated between phenotypes, then true effect sizes were drawn from a multivariate normal distribution with covariance matrix S. Every diagonal element was σ^2_{gen} and for some given correlation coefficient, ρ_{gen}, every off-diagonal element was $\rho_{gen}\sigma^2_{gen}$. In almost all cases, ρ_{gen} and σ_{gen} were kept uniform across all phenotypes and pairs of studies. The exception was when we tested the effects of negative correlations. Then the correlations used were

$$\begin{bmatrix} 1 & -0.5 & 0.5 & -0.5 & 0.5 \\ -0.5 & 1 & -0.5 & 0.5 & -0.5 \\ 0.5 & -0.5 & 1 & -0.5 & 0.5 \\ -0.5 & 0.5 & -0.5 & 1 & -0.5 \\ 0.5 & -0.5 & 0.5 & -0.5 & 1 \end{bmatrix}.$$

As before, the true underlying effect size for all unassociated phenotypes was uniformly 0. The effects represented the natural logarithm of the relative risk of the heterozygote against the homozygote that did not have the effect allele and were additive. Once the true effects were chosen, genotype counts for a given number of cases and controls were simulated and a logistic regression was performed to generate observed effect sizes and standard errors, as above.

Models of association There were two ways in which we chose our true underlying models of association. One was to select from the $2^5 = 32$ possible models at random. Typically, we generated 32 million simulated data points when selecting the underlying models at random in order to ensure that each individual model would be chosen approximately 1 million times.

However, the simulated phenotypes are arbitrary—that is to say, for our purposes, there is no real difference between the model of association where only P1 and P2 are associated with a SNP and the model where P4, and P5 are associated. With this insight, we also performed simulations where there were only six possible models of association: the model where none of the phenotypes was associated, the model where only P1 was associated, the model where P1 and P2 were associated, and so on, adding P3, P4, and P5 to the list of associated phenotypes, in sequence. In general, when we used this method of selecting patterns of association, we simulated 6 million data points.

2.5 Power and model selection

2.5.1 Power

As discussed in Section 1.4.4 in Chapter 1, in frequentist statistics, power is the proportion of the time one rejects the null hypothesis when it is false. It is based on a value, α, which is the rate at which one rejects the null hypothesis when it is true—that is, it is proportion of false positives, or Type I error rate. When α is high (say, 0.2), we expect an increase in power because we are more likely to reject the null hypothesis in general than we would when α is set at a more stringent level. However, this increase in power comes at the expense of calling more false positives. Conversely, when α is low (say, 10^{-8}), we expect a loss of power in exchange for calling fewer false positives. We can apply this to the all-correlated Bayes factor approach. The Bayes factor is a test statistic, therefore it has a null distribution. When the test statistic is suitably extreme, we can reject the null hypothesis of no association with any of the phenotypes.

Because the subset-exhaustive approach gives 2^n Bayes factors, we need to choose some sort of summary statistic before discussing power. One such statistic

might be the maximum Bayes factor, while another might be the mean Bayes factor over all of the models. The mean Bayes factor also has a null distribution to which its values can be compared.

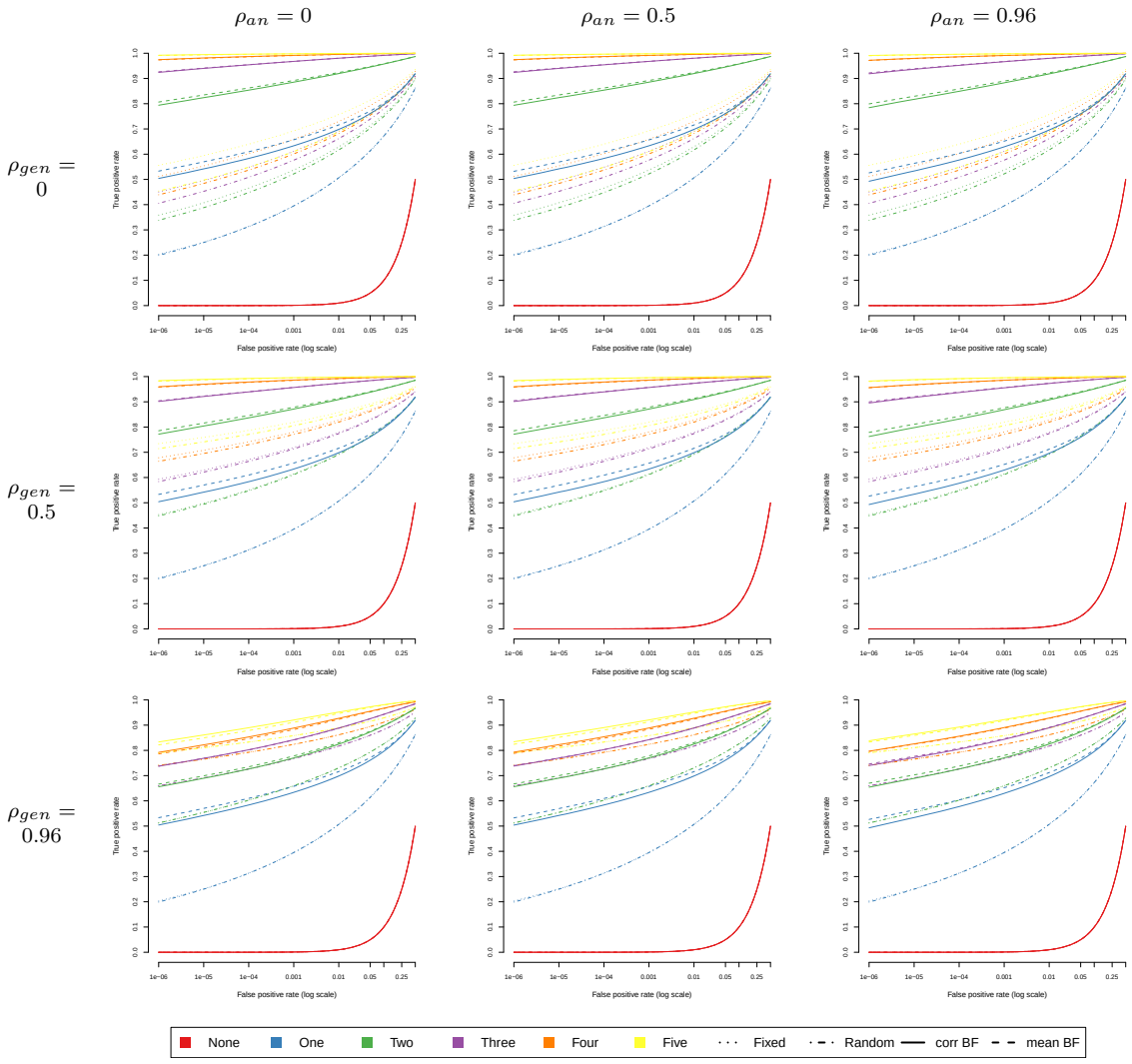
We created a dataset of 32 million simulated SNPs from a study of 7000 cases and 7000 controls, where models of association were randomly chosen from all possible models, and true effect sizes were drawn from a multivariate normal distribution, where each prior standard error was $\sigma_{gen} = 0.2$ and the correlation of true effect sizes between studies was $\rho_{gen} = 0.5$ (see Section 2.4.3 for explanation) and generated two more batches, changing only the value of r to either $\rho_{gen} = 0$ or $\rho_{gen} = 0.96$. For each SNP in the resulting datasets, we calculated both the all-correlated Bayes Factor and the mean Bayes factor, and performed both the fixed and random effects meta-analyses.

In each set, there were about 1 million data points that were generated from the null. We calculated both the mean and all-correlated Bayes factors for these points to estimate the null distributions for these test statistics. We split the data into subsets based on the number of true associations underlying each data point and generated power curves for each subset using both Bayes factor analyses as well as the fixed and random effects meta-analyses. We determined p-values for the latter two approaches by performing a Wald test (see Equation 1.3). This is done by comparing the squared z-scores of the meta-analyzed effect sizes and standard errors to a χ^2_1 distribution. Figure 2.4 shows the results for simulations where the true effect sizes were drawn from the prior, and Figure 2.5 shows the results when true effect sizes were set to be 1.1 uniformly.

We see that the ABF analyses generally have more power to detect effects than the traditional meta-analyses do. The exception is at low false positive rates and with either the odds ratio fixed at 1.1 (see Figure 2.5) or when the data were generated with $\rho_{gen} = 1$. In both of these scenarios, the underlying assumption of the fixed effects meta-analysis is correct—the true underlying effect is the same in all studies. Hence, the fixed effects meta-analysis has more power than the Bayes factor analysis to detect effects at low false positive rates.

We also note that the two Bayes factor approaches seem to have similar power to one another to detect effects, with the mean Bayes factor being slightly more pow-

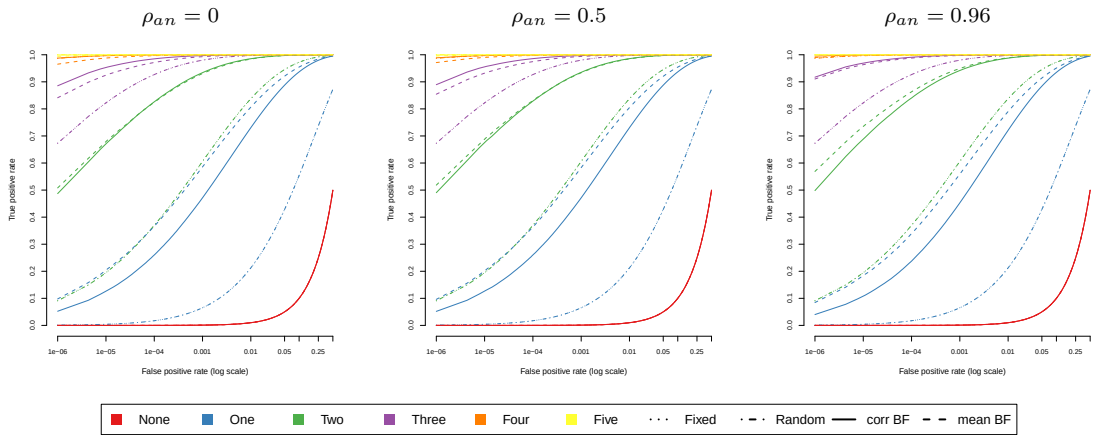
Figure 2.4: Power curves for data simulated under different values of ρ_{gen} with a flat σ_{gen} of 0.2. In all simulation runs, the data were generated assuming a minor allele frequency of 0.5 for all SNPs in a study of 7000 cases and 7000 controls, with effects in up to five different phenotypes. For each data run, we performed both a fixed and random effects meta-analysis ("Fixed" and "Random", respectively), and calculated both the all-correlated Bayes factor ("corr BF") as well as the mean of all the Bayes factors from the exhaustive approach ("mean BF"). We performed our Bayes factor analyses using different values of ρ (ρ_{an}) and a flat $\sigma = 0.2$ and compared the results to the distribution of these Bayes factors calculated using data where there was no effect in any of the phenotypes to obtain the analogue of p-values. All models in the Bayes factor calculations had a prior weight of 1. We plot the log of the false positive rate on the x-axis and the true positive rate on the y-axis (the red curve is the equivalent of the line $y = x$). For each curve, the type of line—solid, dotted, etc.—corresponds to the analysis and the color refers to the true number of phenotypes affected.



erful when there are fewer associated phenotypes, and the all-correlated Bayes factor being slightly more powerful when there are many associated phenotypes. This is unsurprising, since the prior on the all-correlated Bayes factor assumes all phenotypes are associated with the SNP and thus models the data well where all (or almost all) phenotypes are associated with the SNP.

If we set a significance threshold—say, $\alpha = 0.01$, we can compare the power

Figure 2.5: Power curves for data generated to have a true underlying odds ratio (OR) of 1.1. As before, these data were generated assuming a minor allele frequency of 0.5 for all SNPs in a study of 7000 cases and 7000 controls, with effects in up to five different phenotypes. For each data run, we performed both a fixed and random effects meta-analysis ("Fixed" and "Random", respectively), and calculated both the all-correlated Bayes factor ("corr BF") as well as the mean of all the Bayes factors from the exhaustive approach ("mean BF"). We ran our Bayes factor analyses using different values of ρ (ρ_{an}) and a flat $\sigma = 0.2$ and compared the results to the distribution of these Bayes factors calculated using data where there was no effect in any of the phenotypes to obtain the analogue of p-values. We plot the log of the false positive rate on the x-axis and the true positive rate on the y-axis (the red curve is the equivalent of the line $y = x$). For each curve, the type of line corresponds to the analysis and the color refers to the true number of phenotypes affected.



of our approach to that of traditional meta-analyses as well as Fisher's combined probability test. We show the results for analyses with varying values of ρ used in both the analysis and in generating the data in Figure 2.6. Analogous figures with data generated using varying values of σ to generate the data (Figure 2.A.2) and with a flat odds ratio of 1.1 (Figure 2.A.1) can be found in the Appendix.

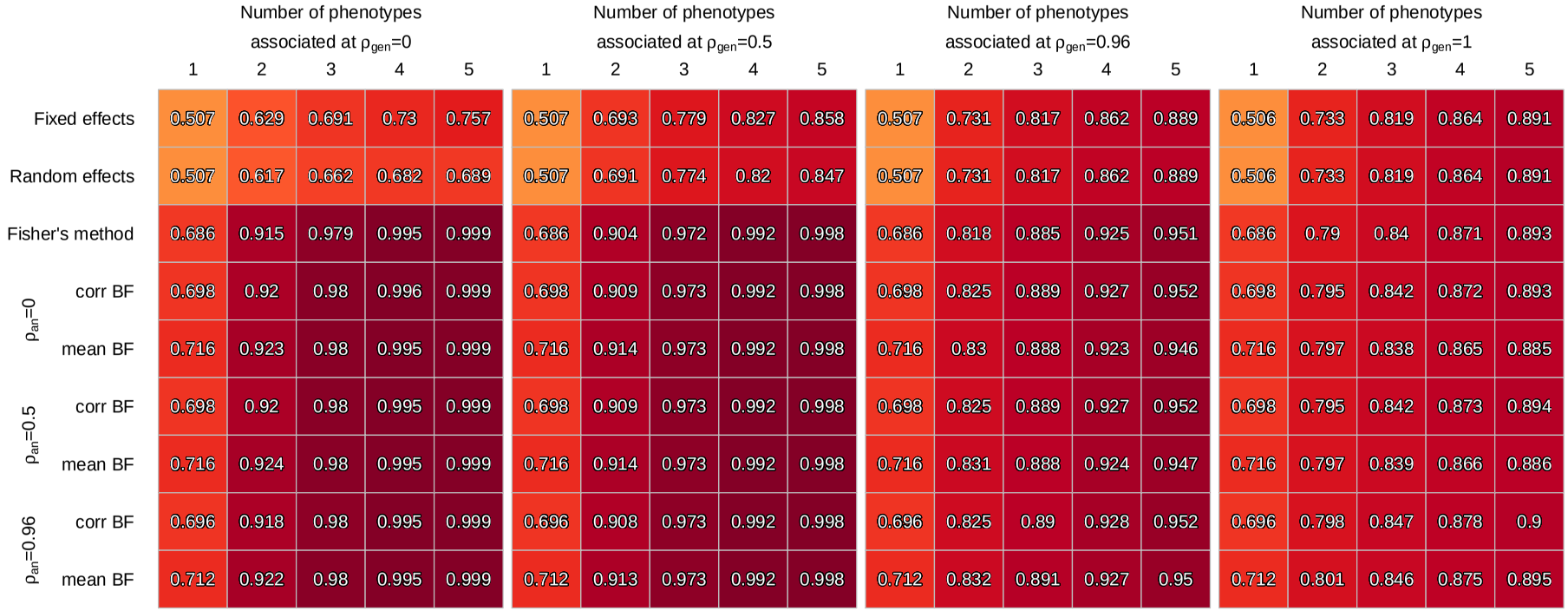
In Figure 2.6 we see that in all cases, the Bayes factor approaches are at least as powerful as the frequentist ones, especially when the number of true effects is small (1 or 2—the left two columns of all four panels). Of the three alternative approaches we tested, Fisher's method (fourth row) comes the closest in terms of power, often drawing level with the Bayes factor approaches when there are three or more associations to find. We also note that while we sometimes lose power when $\rho_{an} \neq \rho_{gen}$, the loss is not large—the results of the ABF methods using different ρ_{an} tend to differ by less than 0.005. This suggests that the Bayes factor approaches are fairly robust to the choice of prior ρ. If the prior is misspecified, then the data will overwhelm it and the analysis will be effective regardless.

Even when the underlying truth is the fixed effects model ($\rho_{gen} = 1$, the right-most panel, fifth column), we note that fixed effects is not more powerful than the

Bayes factor approaches at the 0.01 significance level. At even more stringent thresholds, the gains made from using a fixed-effects meta-anlysis are marginal.

One interesting feature of Figure 2.6 is that all methods show reduced power for the data set generated with $\rho_{gen} = 1$ (the right-most panel). We believe that this is a result of our simulation method, which simulates true effect sizes using a multivariate normal with a correlation of ρ_{gen} between studies. This means that for any simulated SNP, when there is an effect, all effect sizes are identical and the only difference between studies in observed effect sizes comes from sampling variation. However, with total study sizes of 14,000, it is unlikely that any of the observed effect sizes will be vastly more extreme than the true effect sizes, meaning that when the true effect size in one of the studies is close to 0, then the true effects in all studies are close to 0, and cannot be discovered. We expect this to be the case for some of the SNPs in our simulation. We note that fixed and random effects are the only methods that gain power in this dataset over the one where $\rho_{gen} = 0.96$ (third panel) when all five phenotypes are associated.

Figure 2.6: Heatmap and table of power for each of the different methods when $\alpha = 0.01$. We show the true positive rate at this level for data where 1-5 phenotypes were affected and where the effect sizes were generated using different values of the correlation coefficient of ρ_{gen}. All data were analyzed and generated using a prior σ of 0.2. The size of the study was 7000 cases and 7000 controls and in all cases the minor allele frequency was 0.5. The prior weight on all models was 1. A null dataset of 100,000,000 SNPs was created for which the all-correlated and mean Bayes factors (for each value of ρ_{an}) were calculated. This left us with a null distribution for each Bayes factor approach, allowing us to calculate p values. The random effects approach calculated τ by using the derSimonian and Laird formula [132].



2.5.2 Effect of model misspecification

To assess the effect of model misspecification, we look at how the ranking of models changes when we run our method on the same data, but with different priors. In the exhaustive Bayes factor approach, the model corresponding to the highest Bayes factor is the one that describes the data the best. Tables 2.2-2.5 show the accuracy, sensitivity, and specificity of this approach for analyses of data generated under the schemes described in Section 2.4.3 using various (flat) values of ρ. To some extent, these results represent a best-case scenario because we used $\sigma = 0.2$ in our analysis, which is the same parameter as was used to generate the data.

Here, we define the accuracy as the proportion of the time the model corresponding to the highest Bayes factor correctly called a phenotype as associated or not associated, for any phenotype at any replicated SNP. If one imagines a binary matrix of the true patterns of phenotypic association at each SNP, and another one of the models corresponding to the highest Bayes factor at each SNP, then the accuracy is simply the proportion of places where these two matrices agree. Sensitivity is the proportion of times that phenotypes associated with a SNP were correctly called as associated, and similarly, specificity is the proportion of times that the phenotypes that were not associated with the SNP were correctly called as such.

Table 2.2: The overall accuracy, sensitivity, and specificity resulting from selecting the model with the highest Bayes factor using different values of the prior ρ in the analysis and a flat $\sigma_{i,j} = 0.2$. The data were generated in five phenotypes, drawing the effect size from a normal distribution, using a correlation of $\rho_{gen} = 0$ for all pairs of phenotypes and $\sigma_{gen} = 0.2$ for all phenotypes.

Model	$\rho = 0$	$\rho = 0.5$	$\rho = 0.75$	$\rho = 0.96$	$\rho = 1$
Accuracy	0.884	0.884	0.880	0.827	0.752
Sensitivity	0.806	0.806	0.800	0.694	0.528
Specificity	0.962	0.962	0.959	0.959	0.975

Table 2.3: The overall accuracy, sensitivity, and specificity resulting from selecting the model with the highest Bayes factor using different values of the prior ρ in the analysis and a flat $\sigma_{i,j} = 0.2$. The data were generated in five phenotypes, drawing the effect size from a normal distribution, using a correlation of $\rho_{gen} = 0.5$ for all pairs of phenotypes and $\sigma_{gen} = 0.2$ for all phenotypes.

Model	$\rho = 0$	$\rho = 0.5$	$\rho = 0.75$	$\rho = 0.96$	$\rho = 1$
Accuracy	0.884	0.885	0.883	0.849	0.798
Sensitivity	0.806	0.807	0.803	0.732	0.623
Specificity	0.962	0.964	0.964	0.966	0.973

Table 2.4: The overall accuracy, sensitivity, and specificity resulting from selecting the model with the highest Bayes factor using different values of the prior ρ in the analysis and a flat $\sigma_{i,j} = 0.2$. The data were generated in five phenotypes, drawing the effect size from a normal distribution, using a correlation of $\rho_{gen} = 1$ for all pairs of phenotypes and $\sigma_{gen} = 0.2$ for all phenotypes.

Model	$\rho = 0$	$\rho = 0.5$	$\rho = 0.75$	$\rho = 0.96$	$\rho = 1$
Accuracy	0.884	0.889	0.893	0.902	0.902
Sensitivity	0.806	0.812	0.820	0.840	0.846
Specificity	0.962	0.965	0.965	0.965	0.959

Table 2.5: The overall accuracy, sensitivity, and specificity resulting from selecting the model with the highest Bayes factor using different (positive) values of the prior ρ in the analysis and a flat $\sigma_{i,j} = 0.2$. The data were generated in five phenotypes, drawing the effect size from a normal distribution, using a correlation of $\rho_{gen} = \pm 0.5$ for all pairs of phenotypes and $\sigma_{gen} = 0.2$ for all phenotypes.

Model	$\rho = 0$	$\rho = 0.5$	$\rho = 0.75$	$\rho = 0.96$	$\rho = 1$
Accuracy	0.884	0.884	0.881	0.832	0.751
Sensitivity	0.806	0.807	0.805	0.710	0.528
Specificity	0.962	0.961	0.957	0.955	0.974

We see from the tables that the analysis with the parameter $\rho = 0$ gives surprisingly consistent results—from the dataset generated using $\rho = 0$ for all pairs of correlations to the one where $\rho = \pm 0.5$. This is in part due to rounding—keeping a few extra digits shows the differences in the performance of analyses with this parameter. We also see that specificity is always greater than 0.95, which means that we do not call false positives more than 5% of the time. That said, there is quite a range in sensitivities, which drop to almost 0.5 when $\rho = 1$ and the data were generated using a different value. In fact, these tables indicate that performing the analysis with $\rho = 0.96$ will yield results that are at least as accurate as running it with $\rho = 1$, including when the data were generated by $\rho = 1$.

We also notice the surprising result that the accuracy of this approach to model selection does not appear to be harmed even when the sign of the correlation coefficient is incorrect. Table 2.5, whose data were generated with a correlation matrix on true effect sizes of

$$\begin{bmatrix} 1 & -0.5 & 0.5 & -0.5 & 0.5 \\ -0.5 & 1 & -0.5 & 0.5 & -0.5 \\ 0.5 & -0.5 & 1 & -0.5 & 0.5 \\ -0.5 & 0.5 & -0.5 & 1 & -0.5 \\ 0.5 & -0.5 & 0.5 & -0.5 & 1 \end{bmatrix}$$

is entirely consistent with the others, despite the fact that we would expect lower values across all analyses due to the true underlying negative correlations between some of the studies being assumed to be positively correlated in the analysis. It may be because if the sign of ρ is incorrect, then it is equally incorrect in all models of association and thus does not affect their final rankings.

Finally, we see that the main effect of choosing ρ wisely is on power to detect a true effect. This is somewhat comforting—if ρ is not chosen well, then the model corresponding to the highest Bayes factor will call extra false negatives, but it will not produce excessive false positives.

Another way a model may be misspecified is if there is unaccounted for sharing between cases and controls. As Figure 2.7 shows, our method is flexible enough to be able to account for sharing cases and controls across studies; however if this is not done—for instance, because researchers assume that the studies in the meta-analysis are independent and never test this assumption—there is inflation of the ABF compared to the distribution expected under the null, just as there is with both Fisher's method and fixed effects. The correlations between studies under the null can be calculated from Equation 2.11 or estimated empiricially using markers that show no evidence of association (for instance, a threshold of $p > 0.01$).

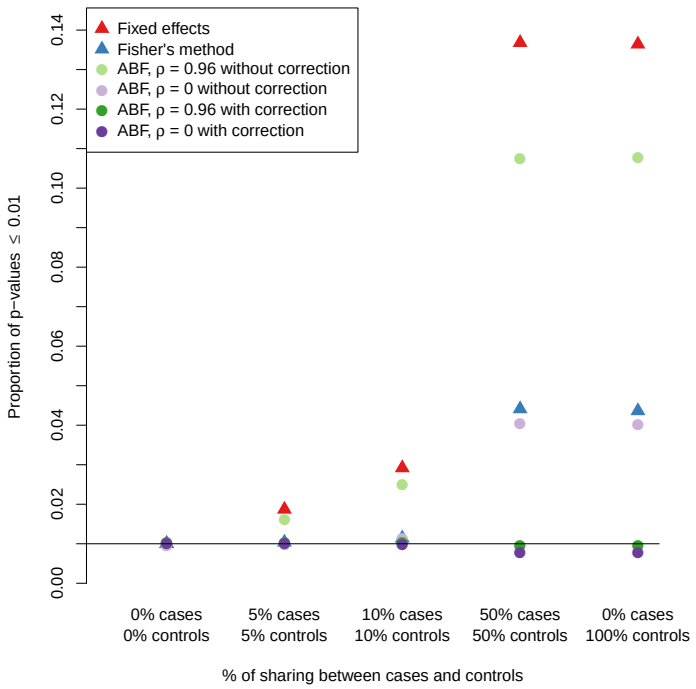


Figure 2.7: Plot showing the effect of of shared controls on standard meta-analysis methods (fixed effects and Fisher's method) as well as our method with and without accounting for sample sharing across studies at false positive rate $\alpha = 0.01$. Data were generated with no true underlying associations (all effect sizes were 0) assuming sample sizes of 1500 cases and 1500 controls across five studies, with the amount of shared cases and controls across all five studies in the proportions shown.

2.6 Calibration

To investigate the calibration of the posterior probabilities generated from the exhaustive Bayes factor approach, we used the simulated datasets described in Section 2.4.3. In each batch and for each replicate, we calculated the posterior probabilities using a flat prior on all the models. That is, $w_i = 1, \forall i$. Thus, for model i,

$$p_i = \frac{ABF_i}{\sum_{i=1}^{2^n} ABF_i}.$$

Each of the 32 p_i at each of the 32,000,000 replicates was sorted into 20 bins of equal size, partitioning the interval (0,1], described below.

$$\left\{ \{p_i : 0 < p_i \leq 0.05\}, \{p_i : 0.05 < p_i \leq 0.1\}, \ldots, \{p_i : 0.95 < p_i \leq 1\} \right\}.$$

For each bin, we counted how many of all of the p_i in the bin corresponded to the model that generated the data for the simulated SNP, and then plotted the

proportion of the "correct" p_i for each bin. This gave us the calibration curves that follow.

In Figure 2.8, we see the first set of curves resulting from data that were simulated from the prior directly, using parameters of $\rho_{gen} = 0.5$ and $\sigma_{gen} = 0.2$, and where the underlying models of association were randomly sampled from all possible models across the five traits. The curve corresponding to $\rho = 0.5$ is nearly perfectly aligned with the line $y = x$. This is what we want to see, especially when our prior matches how the data were generated. We also note that the model for $\rho = 0$ is very close to the line, rising slightly above it from about $p_i \geq 0.5$, indicating that the calculated probabilities are a little conservative compared to the true underlying ones.

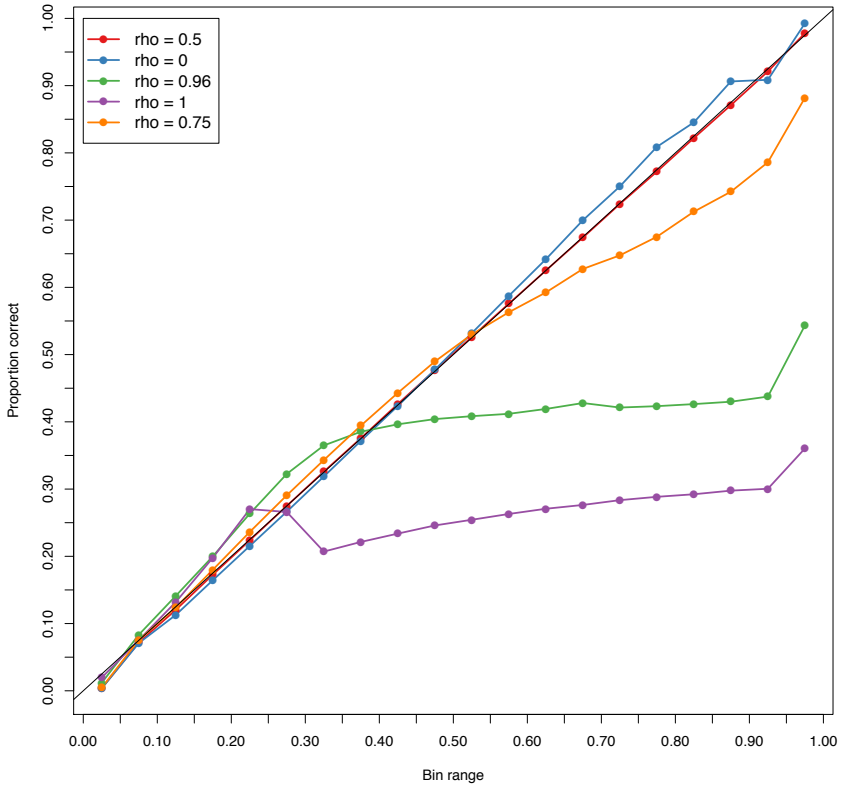
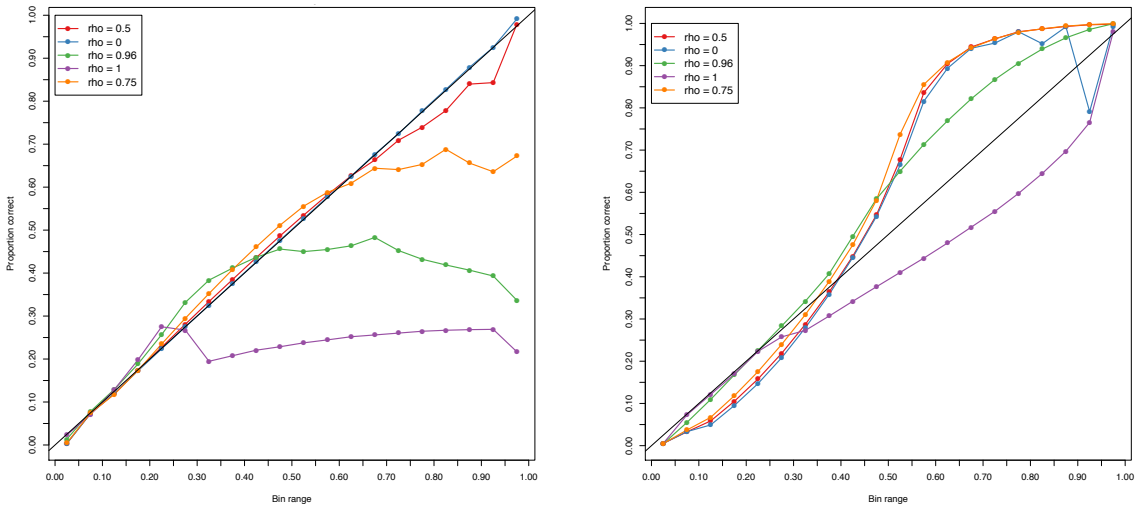


Figure 2.8: Calibration curves for the p_i calculated from the 32 million data points generated using randomly sampling to select the true underlying model of association. True effect sizes were drawn from a multivariate normal distribution centered on the zero vector, where the correlations between studies were $\rho_{gen} = 0.5$ and the standard deviation of effects were all $\sigma_{gen} = 0.2$. The x-axis denotes the bins (all of size 0.05) and y-axis shows the proportion of probabilities in each bin that corresponded to the model of association that generated the data for a simulated SNP. The colors of the curves correspond to the values of ρ that were used in the analysis. All data were analyzed using $\sigma = 0.2$.

Meanwhile, the curves for $\rho = 0.75$, $\rho = 0.96$, and $\rho = 1$ show a considerable deviation from the ideal, plateauing earlier with increasing difference between these values and the true value, $\rho = 0.5$, used to generate the simulated data. This suggests

that when choosing a value for ρ, it is best to choose the lowest plausible one.

In an effort to determine if it was true that it is better to underestimate ρ rather than to overestimate it, we regenerated these data with $\rho = 0$ and $\rho = 1$ as the true underlying parameter values.



(a) $\rho_{gen} = 0, \sigma_{gen} = 0.2$ **(b)** $\rho_{gen} = 0, \sigma_{gen} = 0.2$

Figure 2.9: Calibration curves for the p_i calculated from the 32 million data points generated by randomly selecting a true underlying model, and drawing the true effect sizes from a multivariate normal, with $\rho_{gen} = 0$ (2.9a) and $\rho_{gen} = 1$ (2.9b) as the true underlying correlation between true effect sizes. The standard deviations for each trait were uniformly $\sigma_{gen} = 0.2$. The colors of the curves correspond to the values of ρ that were used to analyze the data. All analysis was performed using $\sigma = 0.2$. The x-axis denotes the bins of size 0.05 into which the model posterior probabilities were sorted, and the y-axis shows the proportions of probabilities in each bin that corresponded to the true model that generated the data at a simulated SNP.

Figure 2.9a is consistent with our hypothesis that overestimating ρ leads to poor calibration of p_i and Figure 2.9b is also consistent in that beyond about $p_i > 0.35$, the estimates for ρ where $\rho \neq 1$ are all conservative. However, Figure 2.9b is a bit puzzling because when $\rho = 1$, the true value that was used to generate the data, we see that it peforms rather poorly. Previously, the curve for the value of ρ which was used to generate the data sat almost exactly on the line $y = x$.

To explain why it does not do so here, we may suggest that the sampling error among the five phenotypes may have reduced the correlations between their effect size estimates enough to reduce the accuracy of this value for the parameter ρ used in the analysis. We considered that it might be a problem with the code that was used to produce the figure or generate the data, though if we believe that the rest of our results are accurate, this explanation seems less likely, since the same scripts to

generate Figure 2.8 were used to generate Figure 2.9.

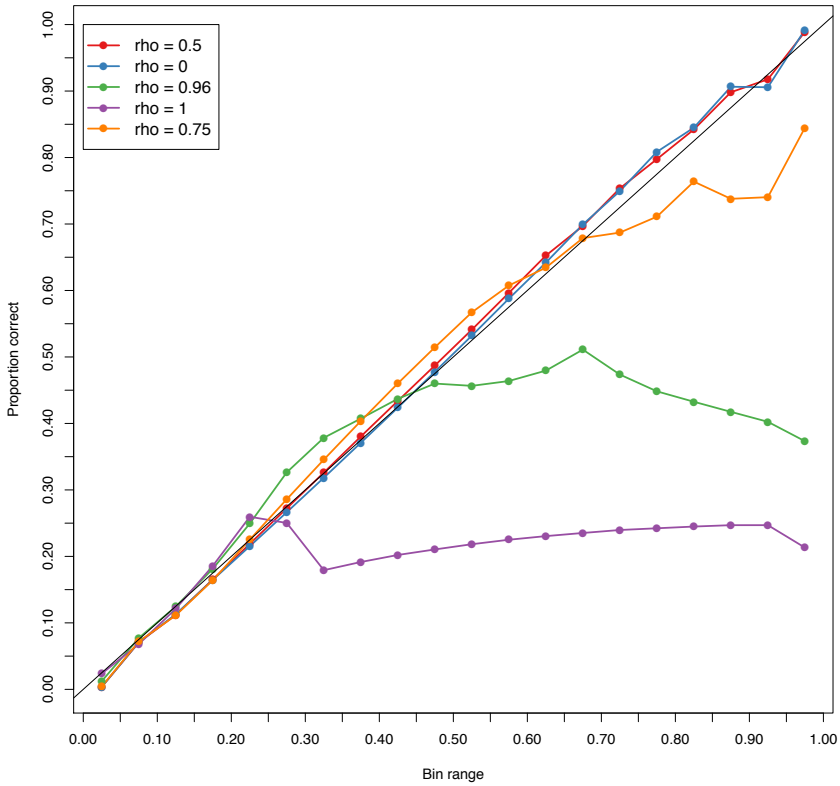


Figure 2.10: Calibration curve for the p_i calculated from the 32 million data points generated from randomly selecting the true underlying model and drawing true effect sizes from a multivariate normal where standard deviations are uniformly $\sigma_{gen} = 0.2$ and $\rho_{gen} = \pm 0.5$. The colors of the curves correspond to the values of ρ that were used in the analysis, while $\sigma = 0.2$ was uniformly used. The x-axis shows the bins of size 0.05 into which the model posterior probabilities were sorted. The y-axis shows the proportion of probabilities in each bin that corresponded to the model that was used to generate the data for the given simulated SNP.

We now move on to a set of simulated data where we investigated the effect of negative correlations on the calibration of p_i. We note that Figure 2.10 bears a striking resemblance to Figure 2.8 in that the curves for $\rho = 0$ and $\rho = 0.5$ stay close to the $y = x$ line, while the curves for the other values of ρ show poor calibrations. It is interesting that despite the direction of the effect being wrong in most instances, the posterior probabilities resulting from a flat prior of $\rho = 0.5$ are still well calibrated, if slightly conservative. This probably due to a loss of power to detect true effects at SNPs where sets of true effects are negatively correlated with one another.

Interestingly, while the curves for $\rho = 0.96$ and $\rho = 1$ in Figure 2.8 tended to curve up for the final bin of p_i, here the curves for $\rho = 0.96$ and $\rho = 1$ curve down, indicating that the final bins of p_i are depleted for associations with the true models that generated the data. This is despite the fact that these values are all

such that $0.95 < p_i \leq 1$. Not only are these very high, but it means that for a given simulated data point represented in this bin, there would be a Bayes factor of at least 19 for the likelihood of the model with the highest Bayes factor against the likelihood of the model with the next highest Bayes factor. We also note that the calibration for $\rho = 0.75$ is good until the $p_i \geq 0.65$, and then the probabilities become anticonservative. As before, $\rho = 0.96$ and $\rho = 1$ do very poorly early on ($p_i \leq 0.5$).

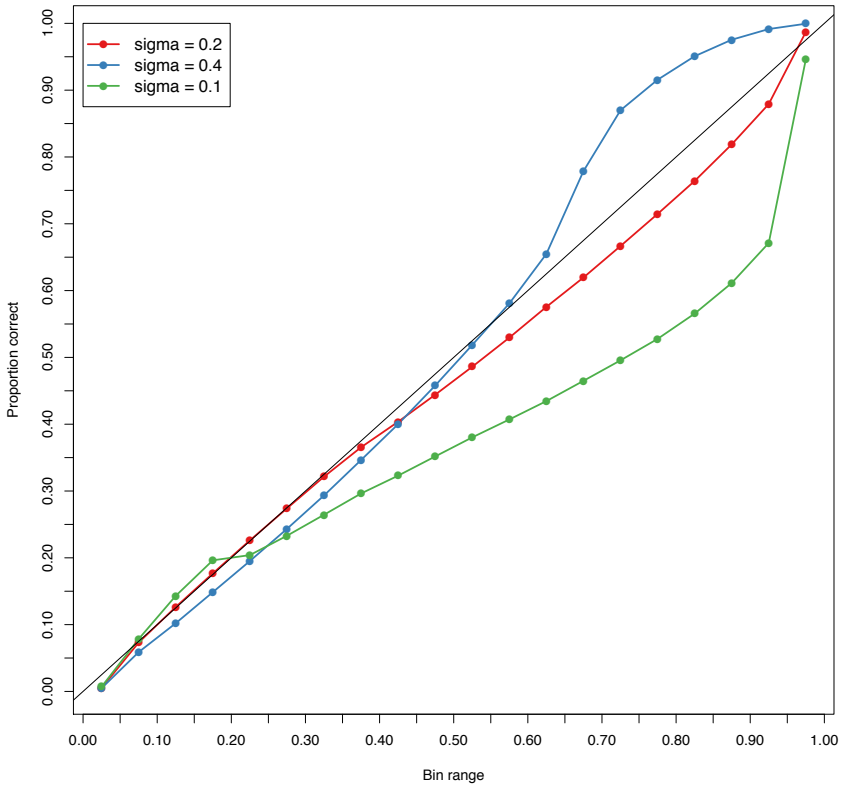


Figure 2.11: Calibration curve for the p_i calculated from the 32 million data points generated by selecting a random model of association and simulating true effect sizes from a multivariate normal using a flat $\rho_{gen} = 0.96$ between all the true effects and $\sigma_{gen} = 0.2$ for each individual study. This time, the color of the curves corresponds to the varying values of σ used in the analysis, while the correlation coefficients have been held at $\rho = 0.96$. As before the x-axis shows the bins (all of size 0.05) into which all posterior model probabilities were sorted, and the y-axis shows the proportion of probabilities in each bin that correspond to the model that generated the data for the given SNP.

Previously, we held sigma (σ) constant at $\sigma = 0.2$ for both generating the simulated data and for analyzing it. Here, we set $\rho = 0.96$ for both generating the data and for calculating the approximate Bayes factors and varied σ in our calculations. We see in Figure 2.11 that the true value of $\sigma = 0.2$ stays on the $y = x$ line for $p_i \leq 0.35$, but then starts to overestimate the true proportion of correct models for each bin. This overestimation is reminiscent of Figure 2.9b, where the curve for $\rho = 1$ was underneath the line $y = x$. It may be that when data are generated with $\rho = 0.96$

between all five phenotypes, that the sampling error breaks this down further so that our hypothesis for why the calculations with $\rho = 1$ performed so poorly in Figure 2.9b applies here as well.

We further note that if we overestimate σ, it appears that for $p_i \geq 0.55$, our estimate of the probability of the model being correct is conservative. If we underestimate σ, then we tend to overestimate the probability of a model's being correct. This makes a certain amount of sense, since the larger variance on the effect sizes in the alternative model means that there is probability density close to zero under the alternative model, and true effect sizes need to be more extreme in order to be detected. It also means that out in the tails, the alternative models with high prior variance will have higher densities than the null models (or alternative models with lower prior variances), resulting in more extreme Bayes factors.

Because generating data with a high correlation between the phenotypes leads to poor calibration when the correct value of ρ is used (as seen in Figure 2.9b), we reanalyzed the data where the models of association had been selected at random and the parameters for the multivariate normal distribution (centered on the zero vector) used to generate the true effect sizes were $\sigma_{gen} = 0.2$ for all studies and the true pairwise correlations were $\rho_{gen} = 0.5$. This time, in our analysis we held ρ at 0.5 and varied the values of σ. Figure 2.12 shows the results. Once again, we see that the analysis using the value that generated the data, $\sigma = 0.2$, shows good calibration. We also see that while $\sigma = 0.15$ and $\sigma = 0.3$ curve close to the line $y = x$, the former tends to be conservative at middling values for p_i ($0.2 < p_i < 0.7$) and anticonservative for higher values, while the latter tends to overestimate p_i in all but the highest bin. This figure suggests that, unlike for ρ, there is no way to choose σ so that you can be certain you are not overestimating p_i.

Figure 2.13 shows the calibrations of probabilities under different analyses on data that were generated so that when there was an effect, it was additive and with a uniform relative risk of 1.1 for the heterozygote, plus or minus some sampling error. We see that for $p_i \geq 0.25$, the value of p_is tends to be conservative in all analyses, and that for $p_i < 0.25$, the p_is tend to overestimate the probability of the corresponding model being the one that generated the data for the SNP in question. As before, if the p_is truly represented probabilities of the corresponding models' being the ones

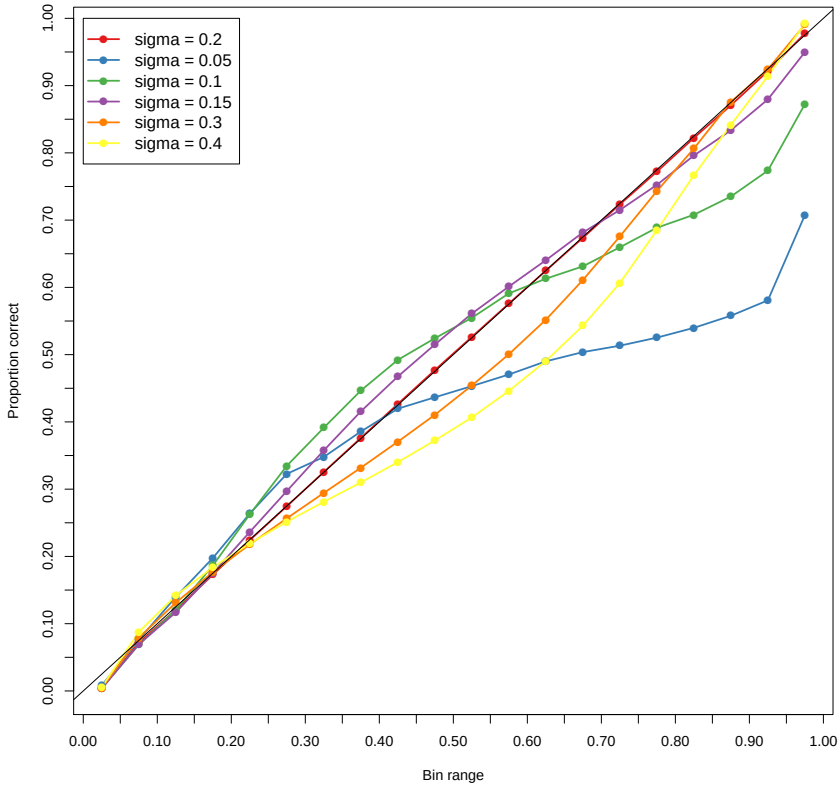


Figure 2.12: Calibration curves for the p_i calculated from the 32 million data points generated by randomly sampling true models of association from all possible models. True effect sizes were generated from a multivariate normal distribution centered on the zero vector, where the correlations between studies were $\rho_{gen} = 0.5$ and the standard deviation of effects were all $\sigma_{gen} = 0.2$. The colors of the curves correspond to the values of σ used in the analysis. All data were analyzed using $\rho = 0.5$. The x-axis denotes the bins (all of size 0.05) and y-axis shows the proportion of probabilities in each bin that corresponded to the model of association that generated the data for a simulated SNP.

that generated the data, we would expect to see the curves align perfectly with the line $y = x$ (black line). We observe that the model that is the closest to this line used $\rho = 1$ in the analysis, which is the the prior that most closely matches the way these data were generated.

As stated before, we would like these p_i to represent probabilities; however, it is not disastrous that they are conservative when $p \geq 0.25$ when priors are poor models for the data. In reality, we would not have a great deal of confidence in a model whose probability of being correct was less than 0.7 or 0.8, even if its probability were much greater than the probabilities corresponding to the other models. Finding out that under some circumstances, a model whose p_i is smaller than the true probability of that model being correct means we can potentially think of p_i as a lower bound for the true probabilities of the models underlying the data.

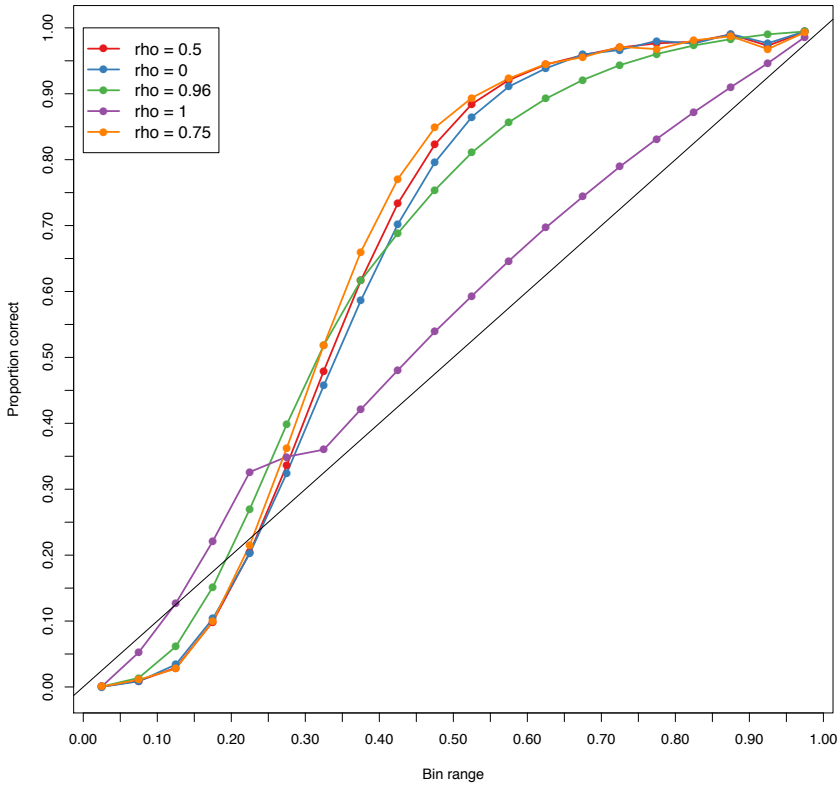
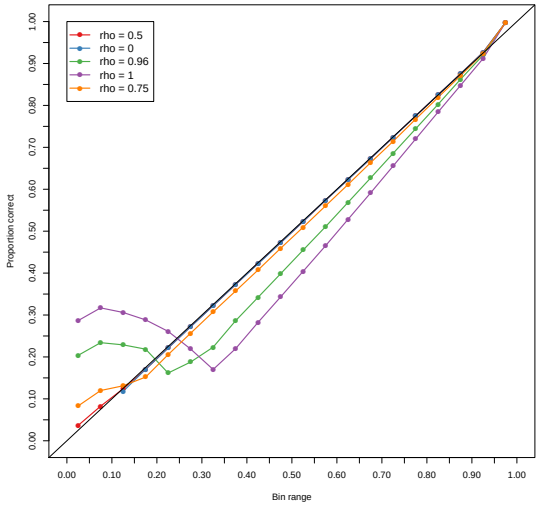


Figure 2.13: Calibration curves for the p_i calculated from 32 million simulated data points generated by selecting a true underlying model at random and assigning to all associated traits a relative risk of 1.1 for the heterozygote compared to the homozygote for the non-risk allele. All data were analyzed using $\sigma = 0.2$, and different values of ρ, which correspond to the colors of the curves. The x-axis denotes the bins (all of size 0.05) and y-axis shows the proportion of probabilities in each bin that corresponded to the model of association that generated the data for a simulated SNP.

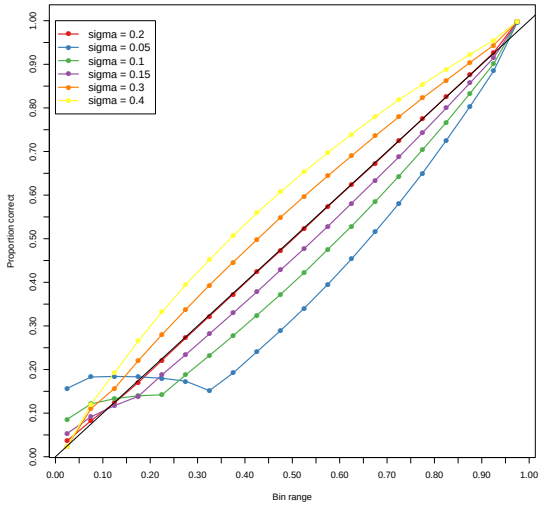
Overall, we find that our posterior probabilities are well calibrated when our priors reflect the parameters that were used to generate the data and that they are not well calibrated when they deviate from these. If the correlation between phenotypes is chosen to be low and, if anything, to underestimate the true correlation, then we can be reasonably certain that the true underlying probability of a model is at least p_i. However, if σ is poorly chosen—and especially if σ deviates from the true standard deviation on effect size by quite a lot, then the p_i will probably overestimate the true probability of the model if $0.7 < p_i < 0.95$. However, as shown in Section 2.5, poor calibration of posterior probabilities does not necessarily mean that we make poor choices in selecting the model of association.

2.6.1 Marginal posterior probabilities

Interestingly, the marginal probabilities of association for each of the phenotypes in dataset where true effects were drawn from the multivariate normal with $\rho_{gen} = 0.5$ for all pairs of effect sizes and $\sigma_{gen} = 0.2$ for each individual study of 7000 cases and 7000 controls (using the analyses that varied both ρ and σ) show a convergence to the true probability when the marginal probability is $p_i \geq 0.95$, as seen in Figure 2.14.



(a) Calibration curves where the parameters used for analysis were $\sigma = 0.2$, and the value of ρ corresponding to the color of the curve.



(b) Calibration curves where the parameters used for analysis were $\rho = 0.5$, and the value of σ corresponding to the color of the curve.

Figure 2.14: Calibration curves for the p_i calculated from 32 million data points that were generated by selecting a random model of association, and generating true effect sizes from a multivariate normal (centered on the zero vector), where the pairwise correlations between two true effects in different traits was $\rho_{gen} = 0.5$ and the standard deviation for each study was $\sigma_{gen} = 0.2$. These tables show the calibration of marginal probabilities of association for each individual trait, with different values of ρ used in the analyses and a uniform $\sigma = 0.2$ (2.14a), and for varying values of σ with $\rho = 0.5$ for all the analyses (2.14b). The x-axis denotes the bins of size 0.05 that each marginal probability was sorted into, while the y-axis shows that proportion of marginal probabilities in each bin that correspond to a phenotype that was truly associated with the simulated SNP.

What this tells us is that while the posterior probability on any individual model may not be accurate if the correct prior parameters are in doubt, that high marginal probabilities ($p_i \geq 0.95$) on the association of a given phenotype with the marker can be trusted. We also note that when $p_i > 0.35$, the calibration curves are fairly smooth and all curve towards 1 when $p_i > 0.95$. We see that in both Figures 2.14a and 2.14b, the distance of the curve from the line $y = x$ tracks with the parameter's distance from the the true underlying value. Since one would probably not believe that phenotype with a low marginal probability ($p_i \leq 0.35$) was associated with the

marker anyway, the poor calibration of low marginal posterior probabilities does not actually negatively affect the specificity—we come to the correct conclusion that the phenotype is probably not truly associated with the marker, but we may be wrong in the measure of uncertainty behind that conclusion.

Taken together, these plots show that the calibration of the posterior probabilities is dependent on the prior. To avoid estimating probabilities that are higher than the true underlying probabilities, it is best to choose the lowest plausible correlation coefficients between the studies. Additionally, it seems that any choice of σ that deviates too much from the true underlying value will lead to poor calibration of the posterior probabilities, at least for part of the interval $[0, 1]$. In particular, when σ is too small, then the posterior probabilities on the interval $[0.6, 1]$ are larger than the true probabilities.

Marginal posterior probabilities versus single-study PPA

Because it is possible to get a posterior probability of association (PPA) from the approximate Bayes factor for a single study (see Equation 1.17 in Section 1.5.3 of Chapter 1), we investigated whether or not meta-analyzing studies increases marginal posterior probabilities for truly associated traits versus the PPA. We used three datasets of 1 million simulated SNPs where the number of cases and the number of controls were both 3000, and where true effects were generated from a multivariate normal distribution centered on the zero vector. The standard deviation for each study was $\sigma_{gen} = 0.2$, and the pairwise correlations between true effects were $\rho_{gen} = 0.5$. We focussed our attention on four phenotypes (P1, P2, P3, and P4), where there were true effects in P1 and P2, and no true effects in P3 and P4.

We performed a meta-analysis of P1 and P2, P3 and P4, and P1 and P4 to test the difference between marginal probabilities from our meta-analysis against posterior probabilities of association, which we calculated assuming $\pi = 0.5$ as a prior on a true association (see Equation 1.16 of Section 1.5.3 of Chapter 1). While we would generally not have such a high prior on the probability of association, this is equivalent to the flat prior we have used in calculating the posterior probability distribution on models of association. When the meta-analysis is performed with $\rho = 0$, the marginal probabilities for each phenotype are equal to the PPAs, as one would expect, since

studies are assumed to be independent. We show the results for the comparison of marginal probabilities and the PPA in our analyses across different true values of ρ and ρ_{gen} in simulated data in Table 2.6.

Table 2.6: The mean and median (mean, median) differences between the marginal probabilities of the stated phenotypes, calculated using the subset-exhaustive ABF approach, and the posterior probability of association, calculated using Wakefield's approximate Bayes factor. The data were simulated for 1 million SNPs and four phenotypes, with only the first two (P1 and P2) showing true associations. Study sizes assumed 3000 cases and 3000 controls, a minor allele frequency of 0.5. True effects were drawn from a multivariate normal distribution where each study had a standard error $\sigma_{gen} = 0.2$ and pairwise correlations, ρ_{gen} were 0, 0.5, or 0.96. The analysis was performed using $\sigma = 0.2$ in all cases, and $\rho = \{0.5, 0.96\}$.

	Parameters	P1, P2	P3, P4	P1, P4
$\rho_{gen} = 0$	$\rho = 0.5$	-0.0047, 3.77×10^{-15}	0.0049, 0.0040	-0.0023, 0.0018
	$\rho = 0.96$	-0.1370, -1.22×10^{-8}	0.0401, 0.0353	-0.0318, -3.21×10^{-13}
$\rho_{gen} = 0.5$	$\rho = 0.5$	-0.0002, 9.19×10^{-8}	0.0049, 0.0040	-0.0023, 0.0018
	$\rho = 0.96$	-0.0732, -8.88×10^{-16}	0.0400, 0.0352	-0.0318, -2.10×10^{-13}
$\rho_{gen} = 0.96$	$\rho = 0.5$	0.0063, 0.0006	0.0049, 0.0040	-0.0023, 0.0018
	$\rho = 0.96$	0.0236, 0.0003	0.0401, 0.0353	-0.0318, -2.39×10^{-13}

When $\rho = 0.5$ we do not observe large differences between the marginal probabilities of association and the PPA. Individual simulated SNPs might show large differences between the two methods of calculating probabilities of association, but these were rare occurrences, and both the median and mean differences between the studies' marginal probabilities of association and their PPAs were less than 0.01, regardless of which two phenotypes were being meta-analyzed, and regardless of what value of ρ_{gen} was used to generate the data. Because the results of the two methods are so close, we would not expect calibration of probabilities to differ very much between them.

However, when $\rho = 0.96$, we observe what are sometimes very large differences between the PPA and the marginal posterior probabilities. Our prior encodes a belief almost perfect correlation between the effect sizes of the studies, which makes it more likely that the posterior probability weight will be on either the null model or the

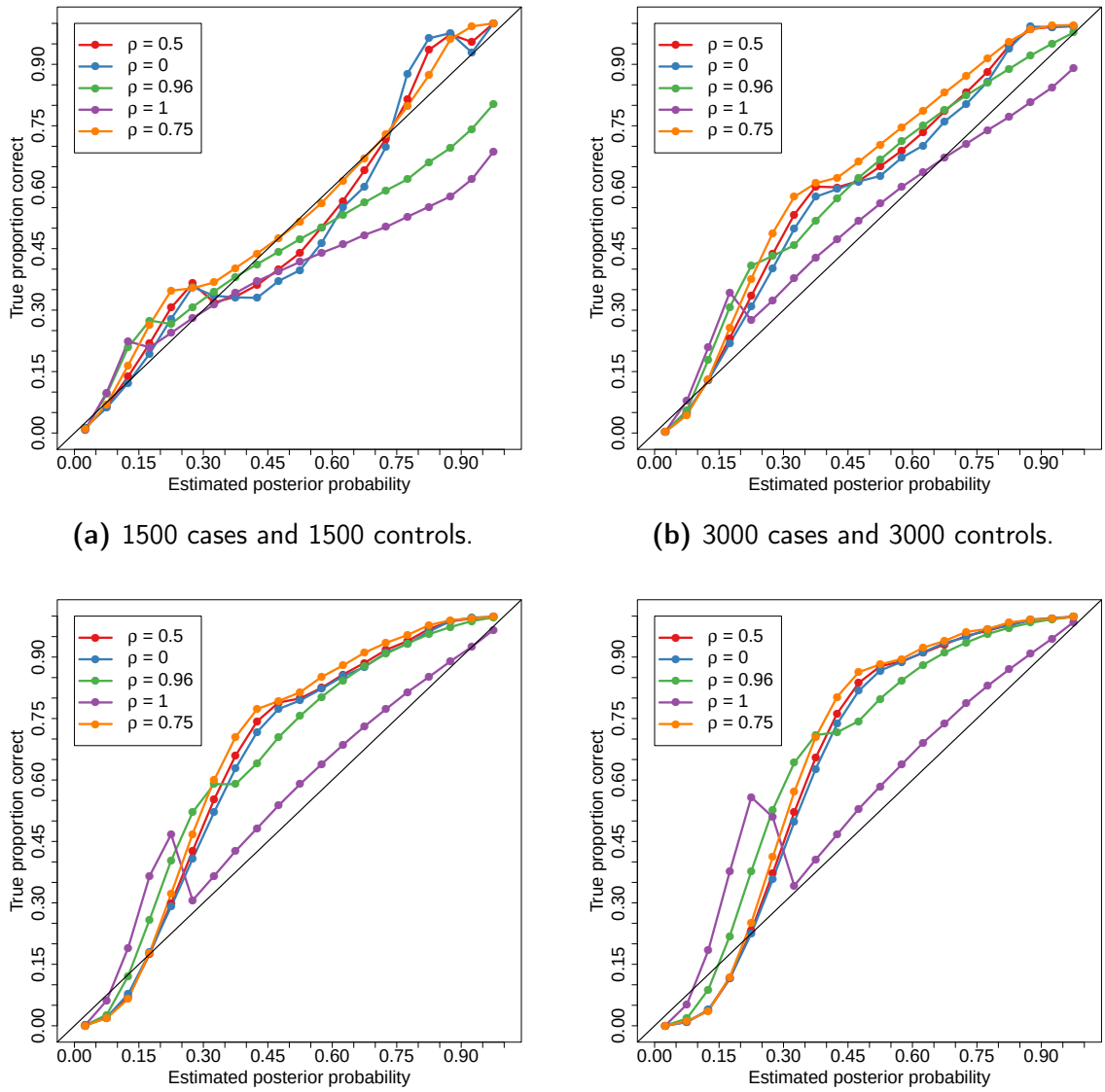
model where both studies are associated. As a result, posterior marginal probabilities of association are, on average, lower than the corresponding PPAs when at least one study is truly associated (except in the case when the effect sizes of two associated studies are strongly correlated), and higher when neither study is. This is because PPAs are independent of one another, so if one study shows an association, its PPA rises, but does not affect the PPAs of other studies. Thus, a false positive or false negative for the analysis of a single study does not increase or decrease the probability weight on models of association in another study.

The same is not true for the ABF analysis that assumes a strong correlation between effects. In the subset-exhaustive ABF approach, if the most likely model contains a false positive, models of association containing that study receive more posterior weight. This means that the model where every study is associated also receives some posterior weight, which increases the posterior marginal probability on the second study. Similarly, when the most likely model contains a false negative, the the posterior weight on all models that exclude that study increases, including the null model, which decreases the posterior probability of association on the second study.

2.6.2 Effect of power on calibration

Because Bayes factors naturally account for power, it would be expected that the effect of power on calibration would be negligible. However, it is also true that when power to detect true effects is reduced, we would expect that the model that generated the data would be selected as the most likely model less often, due to one or more of the true effects not being detected. In order to investigate this, we generated four datasets with identical properties, except for sample size. Each dataset was generated by selecting an underlying model—where none of the phenotypes was truly associated, where only Phenotype 1 was associated, where only P1 and P2 were associated, and so on up to the model where all phenotypes were associated—and simulating allele counts assuming that true effects were either 0 or had a relative risk of 1.1 for the heterozygote compared to the homozygote for the non-effect allele (effects were assumed to be additive), and recording the effect size estimates and standard errors from the logistic regression.

Studies were evenly split between cases and controls and contained 1500, 3000, 5000, or 7000 of each. Power to detect true effects should increase with sample sizes. We analyzed the data using different values of ρ and a flat σ of 0.2 and plot the calibration curves for posterior probabilities of various models in Figure 2.15.



(a) 1500 cases and 1500 controls.

(b) 3000 cases and 3000 controls.

(c) 5000 cases and 5000 controls.

(d) 7000 cases and 7000 controls.

Figure 2.15: Calibration curves for p_i calculated from simulated datasets where true effects were fixed so that the heterozygote had a relative risk of 1.1 compared to the homozygote for the non-risk allele. The effect was additive. Underlying true models of association were chosen so that there would be exactly 1 million simulated SNPs with no true effects, 1 million SNPs with exactly one true effect (always in the first phenotype), and on up through all five effects. Four datasets were created in this manner, differing only in the sample sizes used. Each dataset contained exactly 6 million simulated SNPs. The colors of the curves correspond to the values of ρ that were used in the analysis, while σ was always 0.2. The x-axis denotes the bins of size 0.05 that each marginal probability was sorted into, while the y-axis shows that proportion of probabilities in each bin that correspond to the true model that generated the data for a given simulated SNP.

We note that the calibration is not perfect for any analysis in any of the datasets. Based on the results of previous sections, we would expect the calibration for the analysis where $\rho = 1$ to be the best, and indeed, for all the studies but the one with 1500 cases and 1500 controls (Figure 2.15a), this seems to be the case. For the dataset with the smallest sample size, we see that the calibration is poor overall, and the posterior probabilities are anti-conservative once $p_i \geq 0.3$—that is, they overestimate the true probability of a given model being true model underlying the data. We attribute this to the fact that standard errors from the logistic regression will be relatively large, and so the observed effect sizes in traits that are truly associated with the marker will not be extreme enough to make any of the alternative hypotheses more likely than the null.

We also note that the posterior probabilities in the studies with larger samples sizes (and therefore more power) are not well calibrated in most of the analyses, either, but in this case, they tend to be anti-conservative once $p_i > 0.15$. This means that these posterior probabilities can be treated as a lower bound on the true probability of the associated model, which allows them to be useful in inference. We see that there is not a great difference between the curves in Figures 2.15c 2.15d, which have total study sizes of 10,000 and 14,000, respectively. We also note that and that Figure 2.15b (sample size of 6,000) shows curves that appear to be intermediate between the calibration curves resulting from the dataset where the sample size was only 3,000 and the datasets with the two larger sample sizes. Taken all together, this suggests is that calibration of posterior probabilities is affected by the power of the constituent studies.

2.7 Discussion

Analyzing these simulated data has given us some insight into where our method succeeds and where it fails, and when we can be confident in our answers. We can treat our normalized Bayes factors like probabilities (or at least, lower bounds on probabilities) if we believe that our priors are well chosen or that we have underestimated the correlation between true effect sizes and overestimated the prior variance on effect sizes in the alternative model.

Aside from ruining the calibration of posterior probabilities, the other drawback of misspecifying ρ in particular is that when applying the subset exhaustive Bayes factor approach, it is less likely that the model corresponding to the highest Bayes factor will detect true effects. Fortunately, it does not seem to increase the rate at which we detect false positives. This seems to hold, even when the true correlation between phenotypes is negative, but the analysis is done using a positive value for ρ. This suggests that the rankings of the models is fairly robust to the choice of prior ρ. However, our results also show that if individual studies have low power to detect true effects, the calibration may become anti-conservative.

In contrast, underestimating σ will lead to an increase in false positives, while overestimating σ will create a surplus of false negatives. This intuitively makes sense, since a small σ encodes a belief that true effect sizes will be small and so even very small deviations from an effect size of 0 will be treated as true associations. On the other hand, a large value of σ indicates a belief that only very large deviations from the null can be believed to be real, so the method becomes more conservative.

What will influence the rankings of subset-exhaustive models is the prior weight on each model. In Chapter 5, we investigate the potential of an approach to learn these weights empirically. Another option is to use liability models (see Chapter 6) to try to choose these weights, but an implementation that can return an answer quickly is needed. Furthermore, the results may be too sensitive to our initial Bayes factor analysis to be useful. Choosing weights remains an open problem and different approaches will be appropriate for different analyses.

In any case, we can always turn to the marginal posterior probabilities on each of the phenotypes and be confident that when these are are high ($p_i > 0.95$), there is probably a true association between the individual phenotype and the marker. We are also unlikely to call false positives when using the marginal probabilities, though they are not well calibrated when they are low ($p_i < 0.35$) and the priors are misspecified. This means that we risk being wrong about the level of uncertainty about a phenotype's association: p_i may be equal to 0 in our calculations, but the true value might be closer to 0.3, and similarly, we may believe that $p_i = 0.3$ when it is actually closer to 0.2. However, this difference will not ultimately change the final conclusion.

One caveat to our analysis of simulated data is that we have often performed our analyses using at least one parameter (σ or ρ) that was also used to generate the data. Additionally, in our assessment of the calibration of posterior probabilities, we have implicitly assumed a flat prior on each possible model of association, which exactly matches how each true underlying model of association was chosen—that is, the true models were chosen at random and with equal probability. Our justification is that this helps us to understand the effect of misspecifying a single parameter, however it may be that the effects of misspecifying multiple parameters are not additive and result in worse (or better, if the effects of misspecificiation somehow cancel each other out) power, calibration, or sensitivity/specificity than expected from the individual analyses.

We have not included missing data in any of our simulations. Missing data is a frequent occurrence in real data and is part of our datasets in Chapters 3 and 4. We use a naïve approach to handling missing data: we simply remove the missing study from the dataset at a given SNP and its corresponding row and column in the prior correlation and cryptic covariance matrices, and proceed with the analysis without it.

There is no limit to the number of further simulations we could do. In addition to testing the effects of misspecifying both ρ and σ simultaneously, we have not investigated what happens when meta-analyzing datasets of varying levels of power—for instance, if we had several studies with large sample sizes combined with some that had small ones. We have also not looked at what might happen when our assumption that the summary statistics represent solely the effect of the effect allele on the trait is violated. For instance, in the presence of covariates that are unaccounted for by the model used to create the effect size estimates—can these be captured by the $\mathbf{\Omega}_{\hat{\boldsymbol{\beta}}}$ matrix? Or does this matrix only capture the effect of shared samples, as we expect? To understand this, we would need to create raw data that contains both a true effect size and a covariate measurement and then test how $\mathbf{\Omega}_{\hat{\boldsymbol{\beta}}}$ changes when the covariate is accounted for by the regression model and when it is not, as well as the downstream effect on inference.

Finally, our simulations have only looked at true effects in independent markers, since our method tests each marker individually. Our approach alone cannot detect the situation when two typed variants associated with different traits colocalize to

the same untyped causal variant; further data and approaches would be needed to identify this situation. Furthermore, in real GWAS data, there is an underlying LD structure among the markers that can introduce the appearance of cross-trait associations where there are none. This begs the question of what sort of results our method produces under these circumstances. For instance, are there any obvious and predictable patterns in the posterior distribution on models of association or in marginal posterior probabilities across SNPs in the LD region? Is there any way to mitigate the problem by using these potential patterns or by using the single-study ABFs? A new simulation framework—one that can simulate LD across regions of the genome—would need to be developed before we could begin to the answer these questions.

Appendix

2.A Supplementary heatmap of power

Figure 2.A.1: Heatmap and table of power for each of the different methods when $\alpha = 0.01$. We show the true positive rate at this level for data where 1-5 phenotypes were affected and where the odds ratio of the effect was 1.1. The size of the study was 7000 cases and 7000 controls and in all cases the minor allele frequency was 0.5. The prior weight on all models was 1. A null dataset of 100,000,000 SNPs was created for which the all-correlated and mean Bayes factors (for each value of ρ_{an}) were calculated. This left us with a null distribution for each Bayes factor approach, allowing us to calculate p values.

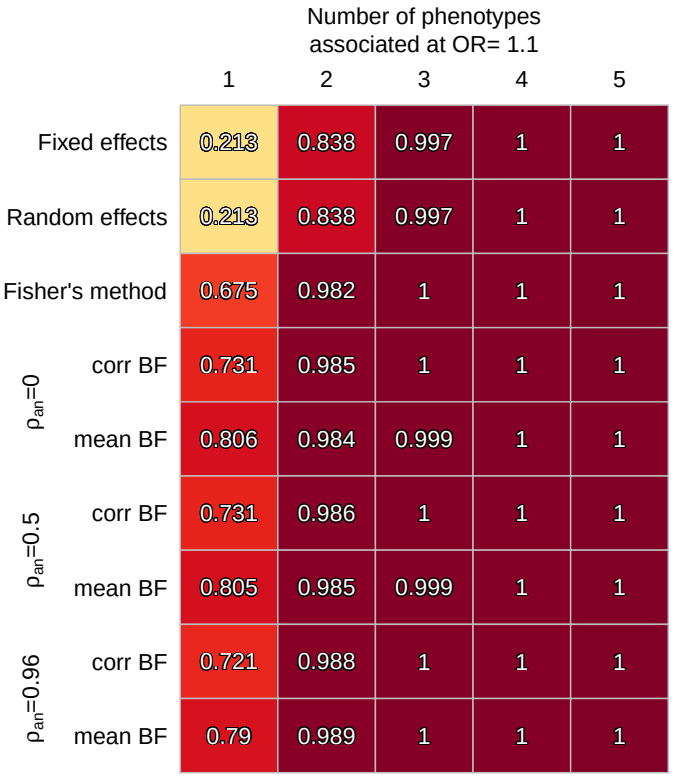
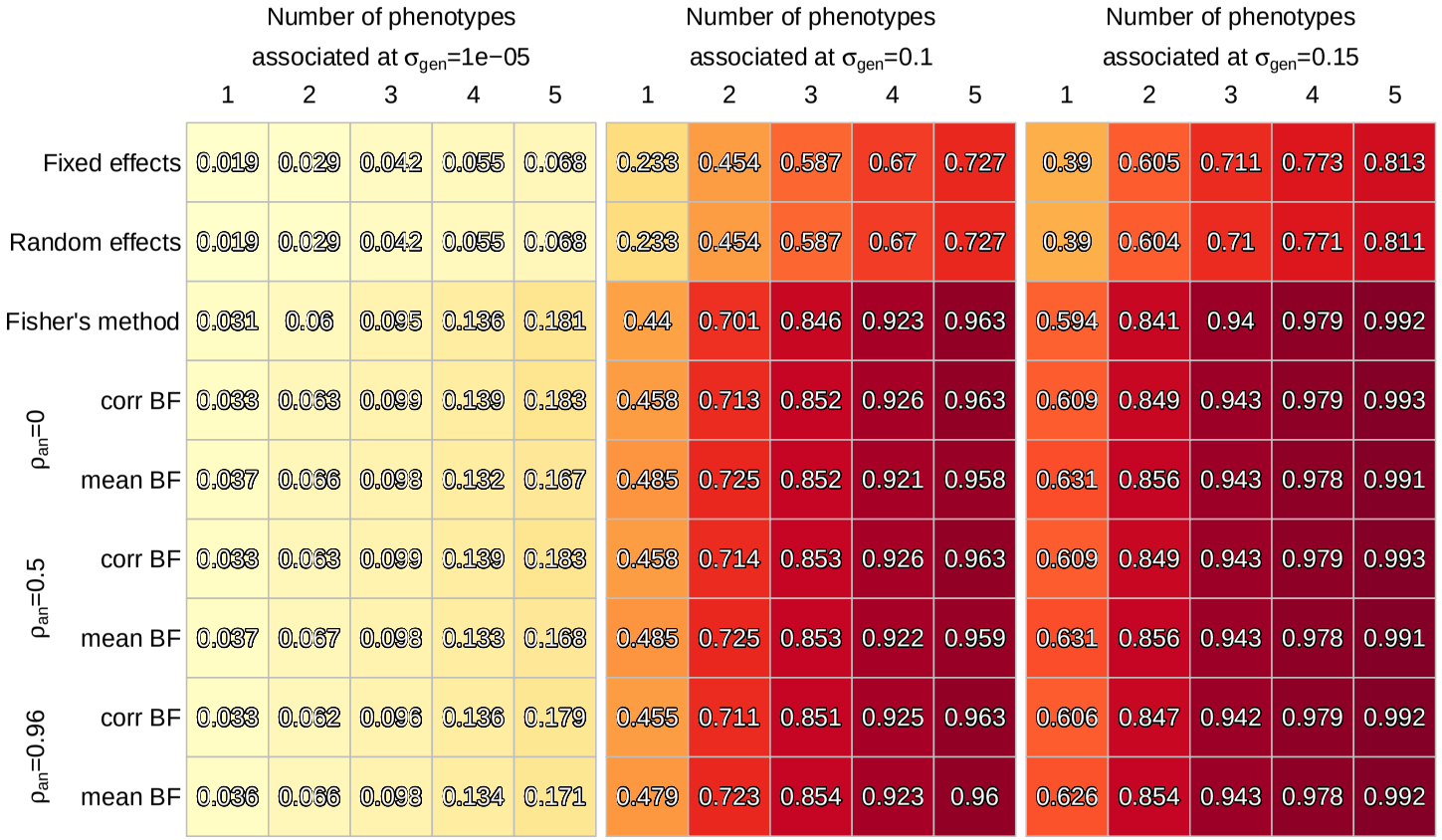


Figure 2.A.2: Heatmap and table of power for each of the different methods when $\alpha = 0.01$. We show the true positive rate at this level for data where 1-5 phenotypes were affected and where the effect sizes were generated using different values of the prior σ_{gen}. All data were generated using a prior $\rho_{gen} = 0.5$ and analyzed using a prior $\sigma = 0.2$. The size of the study was 7000 cases and 7000 controls and in all cases the minor allele frequency was 0.5. The prior weight on all models was 1. A null dataset of 100,000,000 SNPs was created for which the all-correlated and mean Bayes factors (for each value of ρ_{an}) were calculated. This left us with a null distribution for each Bayes factor approach, allowing us to calculate p values.



2.B Proof that ABFs from independent effects model is equal to the product of Wakefield ABFs

Proof. We begin by observing that the product of two approximate univariate Bayes factors, ABF_1 and ABF_2 is (from equation 1.15)

$$ABF_1 ABF_2 = \sqrt{\frac{SE_1^2 SE_2^2}{(SE_1^2+\sigma_1^2)(SE_2^2+\sigma_2^2)}} \exp\left\{\frac{1}{2}\left[\left(\frac{\hat{\beta}_1^2}{SE_1^2}+\frac{\hat{\beta}_2^2}{SE_2^2}\right) - \left(\frac{\hat{\beta}_1^2}{SE_1^2+\sigma_1^2}+\frac{\hat{\beta}_2^2}{SE_2^2+\sigma_2^2}\right)\right]\right\}.$$

From here, it is easy to see that this generalizes to

$$\prod_{i=1}^{n} ABF_i = \sqrt{\frac{\prod_{i=1}^{n} SE_i^2}{\prod_{i=1}^{n}(SE_i^2+\sigma_i^2)}} \exp\left\{\frac{1}{2}\left(\sum_{i=1}^{n}\frac{\hat{\beta}_i^2}{SE_i^2} - \sum_{i=1}^{n}\frac{\hat{\beta}_i^2}{SE_i^2+\sigma_i^2}\right)\right\} \tag{2.12}$$

In the multivariate case, $\mathbf{\Omega}$ and $\mathbf{\Sigma}$ have the forms

$$\mathbf{\Omega} = \begin{bmatrix} SE_1^2 & & & 0 \\ & SE_2^2 & & \\ & & \ddots & \\ 0 & & & SE_n^2 \end{bmatrix},$$

and

$$\mathbf{\Sigma} = \begin{bmatrix} \sigma_1^2 & & & 0 \\ & \sigma_2^2 & & \\ & & \ddots & \\ 0 & & & \sigma_n^2 \end{bmatrix}.$$

This means that $|\mathbf{\Omega}| = \prod_{i=1}^{n} SE_i^2$ and $|\mathbf{\Omega}+\mathbf{\Sigma}| = \prod_{i=1}^{n}(SE_i^2+\sigma_i^2)$. Furthermore, $\mathbf{\Omega}^{-1}$ and $(\mathbf{\Omega}+\mathbf{\Sigma})^{-1}$ have the forms

$$\boldsymbol{\Omega}^{-1} = \begin{bmatrix} \frac{1}{SE_1^2} & & & 0 \\ & \frac{1}{SE_2^2} & & \\ & & \ddots & \\ 0 & & & \frac{1}{SE_n^2} \end{bmatrix}.$$

and

$$(\boldsymbol{\Omega} + \boldsymbol{\Sigma})^{-1} = \begin{bmatrix} \frac{1}{SE_1^2 + \sigma_1^2} & & & 0 \\ & \frac{1}{SE_2^2 + \sigma_2^2} & & \\ & & \ddots & \\ 0 & & & \frac{1}{SE_n^2 + \sigma_n^2} \end{bmatrix}.$$

The terms $\hat{\boldsymbol{\beta}}^{\intercal}\boldsymbol{\Omega}^{-1}\hat{\boldsymbol{\beta}}$ and $\hat{\boldsymbol{\beta}}^{\intercal}\left(\boldsymbol{\Omega} + \boldsymbol{\Sigma}\right)^{-1}\hat{\boldsymbol{\beta}}$ have the forms

$$\hat{\boldsymbol{\beta}}^{\intercal}\boldsymbol{\Omega}^{-1}\hat{\boldsymbol{\beta}} = \frac{\hat{\beta}_1^2}{SE_1^2} + \frac{\hat{\beta}_2^2}{SE_2^2} + \ldots + \frac{\hat{\beta}_n^2}{SE_n^2} = \sum_{i=1}^{n} \frac{\hat{\beta}_i^2}{SE_i^2},$$

and

$$\hat{\boldsymbol{\beta}}^{\intercal}\left(\boldsymbol{\Omega} + \boldsymbol{\Sigma}\right)^{-1}\hat{\boldsymbol{\beta}} = \frac{\hat{\beta}_1^2}{SE_1^2 + \sigma_1^2} + \frac{\hat{\beta}_2^2}{SE_2^2 + \sigma_2^2} + \ldots + \frac{\hat{\beta}_n^2}{SE_n^2 + \sigma_n^2} = \sum_{i=1}^{n} \frac{\hat{\beta}_i^2}{SE_i^2 + \sigma_i^2}.$$

Applying these to equation 2.6, we have

$$ABF = \sqrt{\frac{\prod_{i=1}^{n} SE_i^2}{\prod_{i=1}^{n}\left(SE_i^2 + \sigma_i^2\right)}} \exp\left\{\frac{1}{2}\left(\sum_{i=1}^{n} \frac{\hat{\beta}_i^2}{SE_i^2} - \sum_{i=1}^{n} \frac{\hat{\beta}_i^2}{SE_i^2 + \sigma_i^2}\right)\right\},$$

which is identical to the product of n univariate ABFs as defined in equation 2.12. □

Chapter 3

Application to 107 Autoimmune Associated SNPs

3.1 Introduction

In Chapter 1, we discussed several methods for meta-analyzing genome-wide association studies (see Sections 1.6 and 1.7). In Chapter 2, we introduced our own approach and investigated its performance relative to standard meta-analyses, discussed in Section 1.6. In this chapter, we use a small dataset from "Pervasive Sharing of Genetic Effects in Autoimmune Disease" by Cotsapas *et al.* [52] to compare our method to four others: the cross-phenotype meta-analysis (*CPMA*) test [52], ASSET [111], the binary effects model [112], and CPBayes [116]. We assess the speed of each method on this dataset of 107 SNPs across up to seven autoimmune diseases and compare the findings of each method to the others, as well as to the literature. In doing so, we also illustrate some of the difficulties inherent to using real data.

In "Pervasive Sharing of Genetic Effects in Autoimmune Disease" by Cotsapas *et al.* [52], the authors introduced their method: the cross phenotype meta-analysis (*CPMA*) test statistic. We provided an overview of this method in Chapter 1 1.7.1, along with the other methods discussed in this chapter (Sections 1.7.4, 1.7.3 and 1.7.11). They applied the *CPMA* to GWAS summary data for the following diseases: celiac disease (CeD), Crohn's disease (CD), multiple sclerosis (MS), psoriasis (PS), rheumatoid arthritis (RA), systemic lupus erythematosus (SLE), and type 1 diabetes (T1D).

Using only the p-values for the association of 107 single nucleotide polymorphisms (SNPs) with each of these diseases, the authors tested each SNP for multiple associations. They determined that 47* of the SNPs they tested showed evidence of cross-phenotype associations, with nine of them showing evidence of effects in opposite directions. In particular, their analysis highlighted SNPs in the *IL23R* gene, which is known for its associations with multiple autoimmune diseases [59, 60, 61, 62]. Also represented among the paper's highest ranked SNPs are SNPs in *PTPN2*, which is another gene we would expect to see, based on the literature [133, 134, 135, 136, 137], along with *CTLA4* [138, 135, 139, 137] and *CLEC16A* [138, 135, 140, 141].

To study the biological implications of their results, the authors performed hierarchical clustering of the SNPs based on their patterns of association with these diseases and were able to divide the 47 SNPs that showed cross-phenotype associations into four clusters, three of which corresponded to sets of proteins that are known to interact with one another. Of these three sets of protein interaction networks, two appeared to be preferentially expressed in immune cell subtypes.

Though the *CPMA* statistic is fast and easy to calculate, it does not say anything about pattern of associations in SNPs that are found to have cross-phenotype associations—the paper's authors had to devise a strategy based on binning p-values in order to perform their follow-up analysis. Additionally, in using only p-values as the input, the method throws away other widely-reported summary statistics, such as study sizes, effect sizes, and standard errors, all of which affect p-values and statistical power.

3.2 The data

The data we use in this chapter come from Dataset S1 in the supplement of the Cotsapas *et al.* paper [52]. This table gives the following information for each SNP: rsid, chromosome number, base pair position (according to NCBI Build 36, also known as hg18), the major allele, the minor allele, z scores and p-values for each of the seven studies, the disease the study was first ascertained in, the *CPMA* p-value, and the gene or region associated with the SNP. The z scores and p-values come from the

*On 15 July 2015, the co-author Benjamin Voigt submitted a correction of the reported CPMA p-values that reduced this number to 45. However, the subsequent analyses were not updated.

original papers cited by Cotsapas *et al.*, though the authors state that some z scores were calculated from a χ^2_1 statistic based on the p-value, due to their being unavailable in the original papers. We recreate Table 1 from the Cotsapas *et al.* paper in our own Table 3.1.

Disease	**Cases**	**Controls**	**Ref.**
Celiac disease	3,796	8,154	[135]
Crohn's disease	3,230	4,829	[142]
Multiple sclerosis	2,624	7,220	[143]
Psoriasis	1,359	1,400	[144]
Rheumatoid arthritis	5,539	20,169	[139]
Systemic lupus erythematosus	1,963	4,329	[145]
Type 1 diabetes	7,514	9,045	[138]

Table 3.1: Summary the number of cases and controls, along with the study reference for each disease used the meta-analysis performed by Cotsapas *et al.* The subjects in all studies were people of European descent [52].

Here, we encounter the first potential problem with these data—p-values and χ^2_1 statistics do not contain information on the direction of the effect. This does not matter for the *CPMA* calculation, but it matters for our method, as well as others. The authors state, "Sign was assigned with reference to the minor allele declared in the psoriasis GWAS (chosen arbitrarily)," but make no mention of whether or not the alleles in each study were reported on the same strand.

The lack of strand data makes it impossible to verify the results for SNPs with complementary alleles; that is, alleles that are also nucleotide bases that pair with each other in the DNA molecule. In this dataset, one such SNP is rs11755527 on chromosome 6, which has alleles C and G. If one study reports the effect for the G allele, and another reports the effect for the C allele, then absent any further information about the strand the alleles are reported on, we cannot tell if the data are truly being reported for the two different alleles, or if they are being reported for the same allele on different strands. In the former case, we would choose one of the alleles as the effect allele, and the effect size estimate for the study reporting on the non-effect allele would be multiplied by -1 to align it to the other study. In the latter

case, no such conversion would be needed.

If allele frequencies are reported in each study,[†] then it may be possible to ensure that the effects in each study are reported for the same allele. However, according to 1000 Genomes, rs11755527 has a minor allele frequency of 0.472 in Europeans (G allele, forward strand). This means that due to sampling variation, it may be that the frequency of the "minor allele" in the samples used in each individual GWAS is actually above 0.5. Alternatively, it maybe that some of the studies were performed on a subpopulation where the G allele is actually the major allele. Indeed, two of the five European subpopulations in the 1000 Genomes Phase 3 data show a frequency above 0.5 for this allele. As such, the strand information from each study is vital in order to ensure that the effects are being reported for the same allele of this SNP in each of the studies.

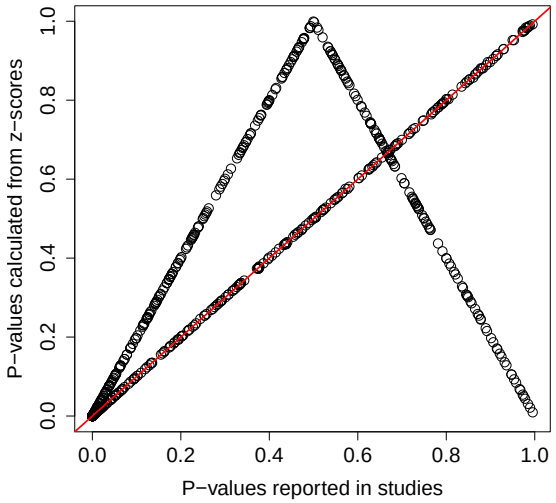


Figure 3.1: Comparison of the reported p-values (x-axis) and the p-values calculated from the reported z scores using a χ^2_1 test (y-axis). The red line is the line $y = x$. The data used to make this plot all come from Dataset S1 of "Pervasive Sharing of Genetic Effects in Autoimmune Disease" by reported by Cotsapas *et al.* [52]

In Figure 3.1, we plot the reported p-values against the p-values calculated from a performing a χ^2_1 test on the square of the z scores. We see that 63.7% of the data lie on the line $y = x$, showing concordance between the reported p-values and the reported z scores. However, the remaining 36.3% show what are at times very large (difference ≥ 0.5) discrepancies. Closer examination of the data shows that these

[†]The authors provide no allele frequency information anywhere in the paper or supplementary material.

discrepancies arise in the CeD, MS, SLE, and T1D datasets, but never in the CD, PS, and RA datasets, suggesting that there was some systematic difference in the way the data were gathered between the first group of studies and the second. We speculate that the second group of studies were the ones for which the z scores were calculated based on the p-values directly, and that at least some of the the z scores and p-values for the first group were collected separately.

Within the studies, the discordant summary statistics do not seem to have an obvious explanation—for instance, the z scores for these data points range from -2.598 to 2.304, with a median of 0.198. While this misses out the extreme values of the whole set of z scores (which range from -15.278 to 20.400, with a median of 0.351), there does not seem to be any reason that this subset of results should show a discrepancy between z scores and p-values. Additionally, the total number of SNPs that were deemed to have discrepant summary statistics—defined as $|\text{reported } p\text{-value} - p\text{-value from } z\text{-score}| \geq 0.1$—is 100. This suggests that these are not due to some feature of the SNPs involved, such as minor allele frequency or genomic location, since nearly the whole dataset is represented among this set of SNPs.

There is an additional issue only briefly mentioned by the study's authors: that of control sharing. In Table 1 of the paper, summarizing the number of cases and controls in each of the studies contributing to the analysis, they note that "controls overlap in some cases due the use of common shared sample genotypes." This necessarily introduces a correlation of summary statistics between studies that share controls. When this occurs, a small p-value in one study increases the chances of a small p-value in the other studies, even if there is no true association between the SNP and the other phenotypes [131]. Examination of the papers cited by Cotsapas *et al.* suggests that every study except for the one of PS [144] made use of samples from a cohort that was also represented in at least one other study [135, 143, 139, 142, 142, 145]. In the analysis of their summary data, the authors did not account for these correlations. If they had—for instance, by reducing their p-value threshold from $p \leq 0.01$ to $p < 0.001$—they would have reduced the number of SNPs that displayed cross-phenotype associations by about 25%.

It is difficult to ascertain the exact number of controls shared between studies. If the dataset were larger and contained many SNPs we believed to have no true

associations, we might be able to estimate the correlations between effect sizes under the null empirically. However, on a dataset of 107 SNPs, we are unable to do this with the desired accuracy. We therefore choose to run all of the analyses without accounting for shared controls and effectively raise any thresholds to determine associations or reject null hypotheses.

The final concern is that the authors admit, "Where data for the reported SNP itself were not available in our GWA studies ... we chose a proxy in high linkage disequilibrium to the reported marker ($r^2 > 0.9$ in HapMap/CEU)." However, nowhere do they list which SNPs or studies for which proxies were used. Despite the high correlation between the proxy SNPs and the reported ones, this still introduces the possibility that the evidence of association for a given study at a given SNP may be inflated or deflated compared to the true, unreported effect at the SNP.

However, despite these shortcomings with their data, the conclusions they have been able to draw from their analysis appear to be valid. We took the opportunity to reanalyze these data using our own method, as well as three others. These analyses allowed us to investigate the results of our method in their own right, and compared to the results of other methods.

3.2.1 Processing steps

Despite the open questions about the data, we proceeded under the assumption that Cotsapas *et al.* aligned their data correctly so that the relative directions of effect are correct. The authors reported the effect of the minor allele in their data. We compared the minor alleles reported in the paper to the minor allele at that SNP in the European populations of the 1000 Genomes Phase 3 data, as every study in this meta-analysis used subjects who were of European descent. None of the SNPs with complementary alleles reported a minor allele that differed with the one in 1000 Genomes European populations—though as noted above it would be hard to determine this with rs11755527, whose minor allele frequency is very close to 0.5.

In the end, only one SNP, rs3184504, reported a minor allele that differed from the one in 1000 Genomes. Since the alleles are C/T in both the paper's data and in 1000 Genomes, there was no confusion about which allele was the effect allele. In 1000 Genomes, T is the minor allele across European populations, with a frequency of

0.464, however we note that for the GBR and TSI subpopulations, this allele occurs at frequencies of 0.5 and 0.528 respectively. It is therefore not impossible that in one or more of the studies, C was indeed the minor allele, hence why the study's authors reported it as such in their table. In our data, we changed the minor allele to T at this SNP, and multiplied the reported z scores by -1.

Next, we calculated the estimated standard errors of the effect sizes ($\hat{\beta}$s) for each study at each SNP and converted the z scores into effect size estimates. For this, we used information about the numbers of cases and controls (provided in Table 3.1) and the allele frequency information. If the ith SNP has minor allele frequency f_i, the jth study has a total sample size (cases and controls) of n_j, and ϕ_j is the proportion of cases in the sample, then

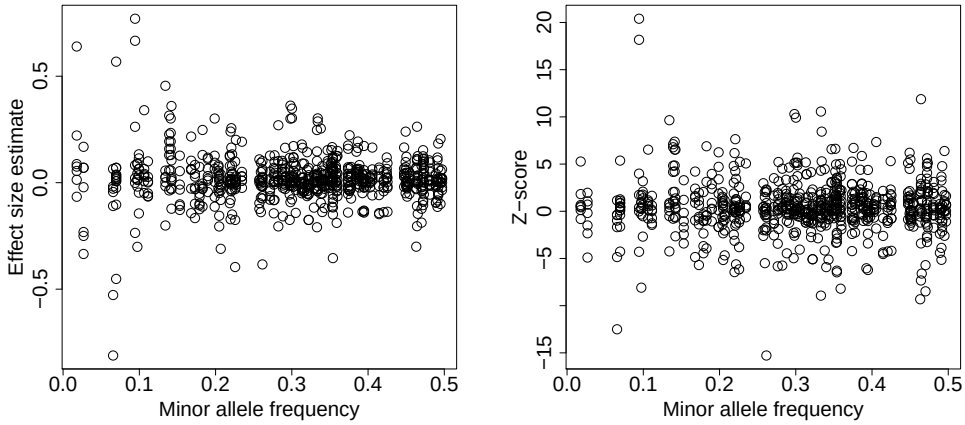
$$SE_{i,j} = \frac{1}{\sqrt{2n_j f_i(1-f_i)\phi_j(1-\phi_j)}}, \tag{3.1}$$

and for $z_{i,j}$ the z score of the jth study at the ith SNP, $\hat{\beta}_{i,j}$ is simply

$$\hat{\beta}_{i,j} = z_{i,j} SE_{i,j}. \tag{3.2}$$

We observe in Figure 3.2a that the largest effect sizes occur at low minor allele frequencies (MAF ≤ 0.1) and that in general, effect sizes tend to shrink as MAF increases. Figure 3.2b, which shows z scores, indicates that these large effect sizes tend to be paired with larger standard errors. While there are three outliers that appear when MAF ≤ 0.1, there is also one that occurs with at a MAF of 0.261.

We used these data to rerun the analysis by Cotsapas *et al.*, as well as our approximate Bayes factor (ABF) analysis. For further comparisions to other methods, we also ran ASSET [121], Han and Eskin's binary effects model and random effects meta analysis [102, 112, 54], and CPBayes [116].



(a) Estimates of effect sizes (y-axis) plotted against minor allele frequencies (x-axis).

(b) The z scores for the Cotsapas *et al.* data (y-axis) plotted against minor allele frequencies (x-axis).

Figure 3.2: The effect size estimates we calculated for the Cotsapas *et al.* data from allele frequencies reported in 1000 Genomes and z-scores (y-axis) plotted against minor allele frequencies (x-axis).

3.3 Analysis

Because the dataset is so small, we were able to use our subset-exhaustive approach on every SNP. The maximum number of subset models is $2^7 = 128$, for which approximate Bayes factors at each SNP can be computed fairly quickly (see Table 3.5).

We began by investigating how changing the prior values of ρ and σ changed the rankings of SNPs by mean ABF across all models of association. Note that ρ was kept constant among all possible pairs of correlations, and σ was similarly kept constant as the prior across all studies in the meta-analysis. Possible values were $\rho = \{0, 0.5, 0.96, \pm 0.5\}$, while $\sigma = \{0.2, 0.4\}$. Because of the short-comings in the data, we used the results from the analysis with $\rho = 0$ and $\sigma = 0.2$ to select SNPs for follow-up analyses, since the prior is agnostic to the direction of effect, meaning that its results should be valid even if some of our effects are reported for the wrong allele, though we would expect to lose power from not taking advantage of the probable correlations in true effect sizes among the studies.

We included the other analyses because we thought it would be informative to see how results change depending on the prior. In particular, we were interested in the effect of including negative prior correlations between effect sizes. For the analyses that used priors of $\rho = \pm 0.5$, the analysis created a new prior correlation matrix for each SNP. In the matrix, the signs of the off-diagonal elements, $\rho_{i,j}$, were negative if $\hat{\beta}_i$ and $\hat{\beta}_j$ for the SNP had different signs, and were positive if they had the same sign.

We also investigated the effect of using different priors on models of association. These were briefly discussed in Chapter 2, Section 2.3 and are as follows:

Flat prior: This assumes that every model of association is just as likely as every other model of association—that is, all weights, w_i in Equation 2.8 have a weight of 1, and thus a prior probability p of

$$p = \frac{1}{2^n}. \tag{3.3}$$

Combinatorial prior: When calculating the prior weight on models with m out of a possible n associated phenotypes under the flat prior, one sees that models with $\frac{n}{2}$ phenotypes ($\frac{n}{2} \pm 0.5$ if n is odd) have the most prior weight, due to the fact that there are more of these models. The combinatorial prior forces the prior weights on the number of associations to be the same. That is, for a model with a total of m associations, the prior probability p on the model is

$$p = \frac{m!\,(n-m)!}{(n+1)!}. \tag{3.4}$$

Binomial prior: Both the flat prior and the combinatorial priors assume that a model with no associations is as likely as a model where all n phenotypes are associated. In reality, we might assume that each of the studies has some probability $q_i : i \in \{1, \ldots, n\}$ of having a true association with a SNP. Alternatively, these q_i can be thought of as the proportion of SNPs that the ith study is expected to be truly associated with. This is similar to π, the prior on the expected proportion of true effects that Stephens and Balding [93] discuss for the calculation of the posterior probability of association (see Section 1.5.3 in Chapter 1). For a given model with $K \subseteq \{1, \ldots, n\}$, the subset of studies associated with the SNP and $L = \{1, \ldots, n\} \setminus K$, the subset of studies that are not, then the prior probability on the model is

$$p = \prod_{i \in K} q_i \prod_{j \in L} (1 - q_j)\,. \tag{3.5}$$

The flat prior can be recreated in this prior by setting all $q_i = 0.5$.

In addition to the analysis using our own method, we also recalculated the CPMA test statistic and replicated the p-values from the paper by Cotsapas *et al.* (See Section 1.7.1 for details). For ASSET, CPBayes, and Han and Eskin's binary effects model, we assumed no cryptic correlation between the studies, as we did in our own analyses, even though we know this not to be the case. The binary effects model also requires a prior σ the standard deviation of the normal distribution from which true effect sizes are drawn. Since we ran each method using default parameters, we did not change the default value of σ from 0.2. ASSET requries the number of cases and controls for each study, which we provided using the data in Table 3.1. The dataset was small enough that exact m values for the binary effects model could be calculated. The inputs for ASSET, CPBayes, and the binary effects model were the effect size estimates and standard errors that we calculated for the data, as described in Section 3.2.1.

3.3.1 Simulations based on the data

To better understand the accuracy and power of our approach when applied to these data and to compare it to the *CPMA* statistic, we tested the performance of the two methods using simulated data. For each simulation run, we generated data for 8 million SNPs across seven studies. All seven studies had a composition of 3,500 cases and 8,000 controls—similar to the average number of cases and controls in our dataset. The SNPs all had a minor allele frequency of 0.5 and could be subdivided into eight groups of exactly 1 million SNPs each. These groups were identified by the true underlying models of association that generated the data. One of them was the set of SNPs with no true underlying associations with any of the studies, another comprised SNPs that were associated with exactly one study (the same study in all SNPs), and so on up to the set containing SNPs where all seven studies were associated.

True effect sizes were either set to have an odds ratio of 1.05 or 1.1 or were generated from a multivariate normal distribution (for the subsets of SNPs with multiple associations) with study standard deviations of either 0.2 or 0.4 and correlations between the studies of 0, 0.5 and 0.96. In either case, the same parameter value was applied to all studies. Using the true effect sizes, we sampled alleles in cases and controls and then performed a logistic regression to generate the summary data. As in Chapter 2, we did not aim to explore all possible types of data that we might

encounter. For instance, we never varied the minor allele frequency, despite the fact that SNPs all along the allele frequency spectrum appear in our data, nor did we attempt to simulate missing data, even though the dataset contained such SNPs.

Power comparison of the ABF and *CPMA* approaches

To calculate p-values for our all-correlated ABFs, we used the distribution quadratic form (see 2.2.2 for details), which gives similar p-values when compared to the null distribution (i.e. the set of SNPs for which there are no true underlying associations). Then, in each simulation run, for different significance thresholds, we plotted the proportion of simulations whose p-values were below the given threshold, generating the receiver operating characteristic curves (Figure 3.1). On this type of plot, a method that performs perfectly would be on the line $y = x$ when the null model was true (i.e. when there was no association in any study) and would have y values of 1.0 from $x > 0$ in all other scenarios.

We found that when treating the single Bayes factor as a test statistic, we had more power to detect SNPs with multiple effects than the *CPMA* statistic. Changing the values of ρ and σ in our analysis did not change the shape of the curves very much overall regardless of true underlying effect size (fixed with a given risk ratio, or generated randomly from a multivariate normal distribution with given values of ρ and σ). Figure 3.1 provides two representative examples of how the ABF and *CPMA* approaches performed on the data we simulated.

We note that the all-correlated Bayes factor test statistic has more power than the *CPMA* test statistic for every simulation run in which we tested them. As we see in Figure 3.1b, where the risk ratio is 1.05, the all-correlated ABF on a dataset with m truly associated phenotypes tends to perform almost as well as the *CPMA* test on the dataset with $m + 1$ truly associated phenotypes. However, neither one performs particularly well when the relative risks are low (≈ 1.05) and when the number of affected phenotypes is one or two. With six or seven affected phenotypes, both methods of analysis show high power to detect effects, even at low false positive rates ($\alpha \leq 0.05$). This is expected, since so many affected phenotypes mean that these tests are more likely to find multiple associations, even if they underestimate the number of associations.

(a) ABF analysis with $\rho = 0$, $\sigma = 0.2$, risk ratio = 1.05.

(b) ABF analysis with $\rho = 0$, $\sigma = 0.2$, generated with $\rho = 0$ and $\sigma = 0.2$.

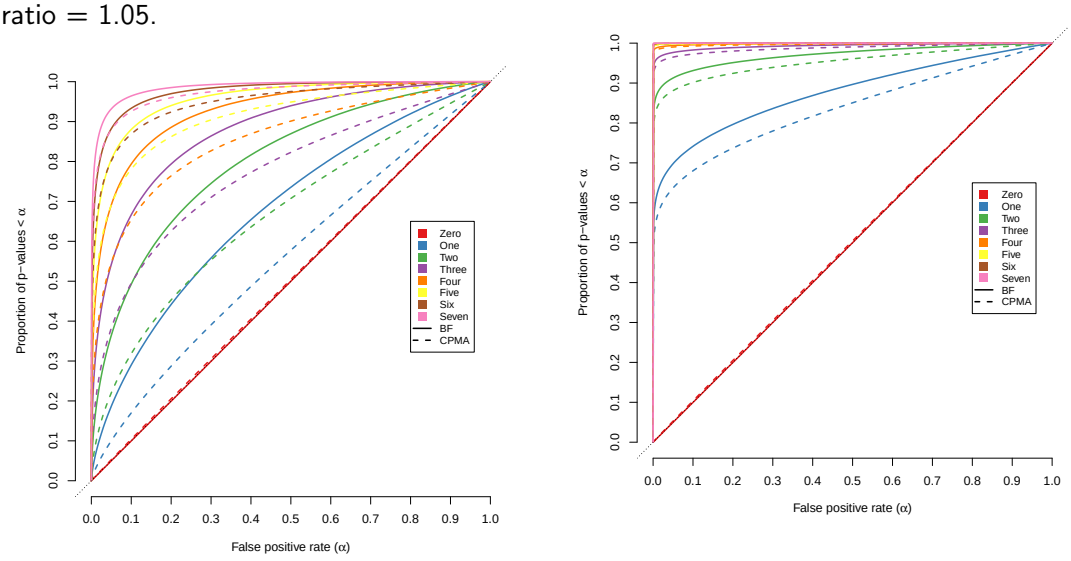


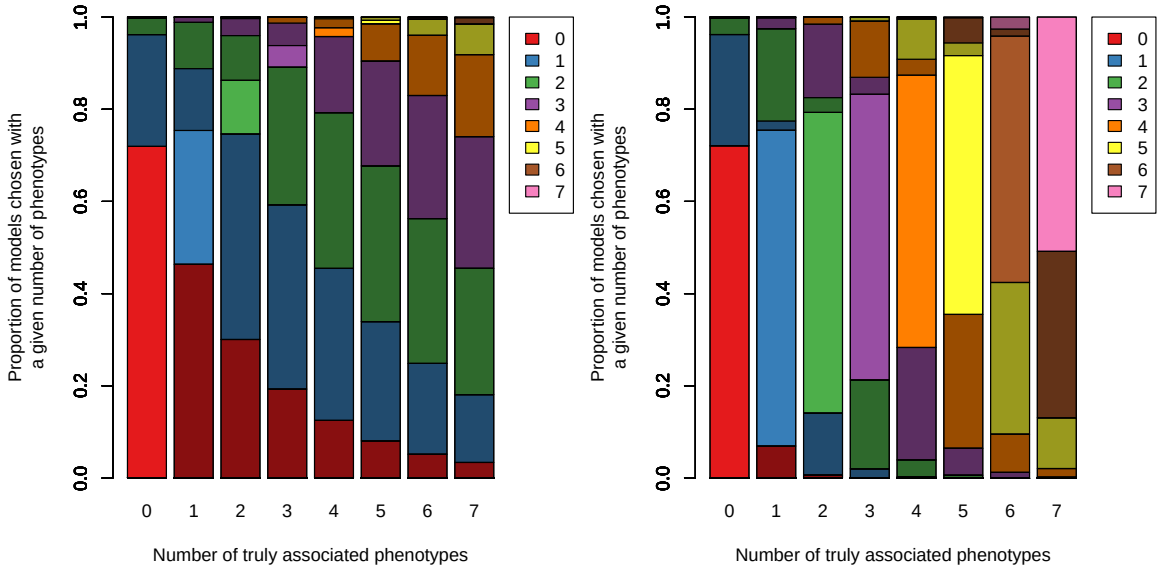
Figure 3.1: Receiver Operation Characteristic curves for the all-correlated Bayes factor analysis (solid lines) and the *CPMA* analysis (dotted lines) on simulated data, subdivided by the number of truly affected phenotypes, indicated by colors. In both panels, the x-axis is the false positive rate, α, and the y-axis is the proportion of test statistics whose p-values are $\leq \alpha$. Different risk ratios were used to generate the data and different prior values for ρ and σ were used to perform the Bayes factor analysis. The data were simulated with the parameters indicated, assuming studies with 3500 cases and 8000 controls (similar to the sizes used in the original studies from the Cotsapas *el al.* data. For each number of associated traits, exactly one million markers were simulated. Models of association were the null model, the model where there was an association in the first study only, the model where there were associations in the first and second studies, and so on up to the model where there were associations in all studies.

ABF model selection in simulated data

We also performed the subset-exhaustive analysis on the simulated data to see how often the model with the highest Bayes factor (for given values of ρ and σ) corresponded to the correct model of associated phenotypes. We used the flat prior on the models of association. This puts the most prior weight on models where the SNP affects three or four phenotypes, since there are 70 such models, while there is only one model where the SNP affects no phenotypes and only one where it affects all phenotypes. There are a total of 128 models being compared for each simulated SNP. A model chosen at random has a 0.78% chance of being the true model underlying the data.

Since we knew exactly which phenotypes were affected in the simulated data, we could see how often this approach resulted in the correct model receiving the most posterior probability density. As Figure 3.2a shows, for low risk ratios (again, ≈ 1.05), our method struggles to detect the correct model of association, unless the null model

is the correct one. We note that the method tends to be conservative about the number of phenotypes affected, though when it selects a best model with the correct number of affected phenotypes, it tends to select the true underlying model. However, as risk ratios increase (Figure 3.2b), the signal becomes stronger, and so the accuracy of this approach improves dramatically. We note that in either case, the models selected by this approach tend to err on the side of fewer associations than the true underlying model, rather than more.



(a) For $\rho = 0.5$, $\sigma = 0.2$, risk ratio = 1.05. **(b)** For $\rho = 0$, $\sigma = 0.2$, risk ratio = 1.1.

Figure 3.2: The bar plots show the frequency with which the subset-exhaustive approach returns the highest Bayes factor for a model with a given number of affected phenotypes (coded by the bar color) when the underlying data was simulated with the number of affected phenotypes indicated by the clusters' x-axis labels (assuming 3500 cases and 8000 controls in each study). Grey overlays are on models where the number of affected phenotypes in the chosen model did not match the number of affected phenotypes in the underlying simulated data, which gives an indication of the frequency with which the model with the highest Bayes factor was identical to the true underlying model that generated the simulated data.

Figure 3.2 does not show the distribution of approximate Bayes factors over models. It may be that the highest ABF is not much higher than the next highest ABF—or even the third or fourth highest ABFs, one of which may be the true underlying model. Combined with the tendency of the model selection to underestimate the number of true associations, it is likely that model determined to be the most likely by our approach is a subset of the true set of associated models. This is particularly true in the case where true underlying risk ratios are about 1.05, which is more similar to the ratios observed in GWAS studies than 1.1. In this, we see that our approach loses power to detect effects as the true underlying effect becomes smaller, as is the

case with all methods. By meta-analyzing data across studies, we hope to be able to detect some true effects, despite this loss of power, and despite the fact that some of the true underlying effects will inevitably remain undiscovered.

3.4 Results

Here we present the results of our comparisons of analyses using different values for the prior parameters, ρ and σ, as well as for the three priors on models of association we discussed in the previous section.

3.4.1 Effect of prior ρ and σ

We wanted to understand the effect of using different priors on our results. One possible effect is on how SNPs are ranked when we take the mean ABF over all possible models of association. Figure 3.1 represents ranks in color—the shade of green depends on rank, with darker shades corresponding to higher ranks and paler shades to lower ones. In general, there was good concordance in ranks across all studies, especially for the highest and lowest ranked SNPs.

We calculated the correlation of ranks among the analyses using Kendall's τ. The lowest resulting correlation was $\tau = 0.868$ between the analysis where $\rho = 0$ and $\sigma = 0.4$ and the one where $\rho = 0.96$ and $\sigma = 0.2$. This is perhaps expected, as one prior assumes no correlation among true effect sizes, which are expected to be large, while the other assumes high correlations among true effect sizes which are expected to be smaller—these beliefs are, in some sense, opposites. The highest rank correlation was $\tau = 0.995$ between the analyses performed with $\rho = 0.5$ and $\sigma = 0.4$ and the one where $\rho = \pm 0.5$ and $\sigma = 0.4$. For six of the SNPs in the study, these priors are identical, since all the effect size estimates at those SNPs all had the same sign; however that alone would not explain the high rank correlation. It may be that for purposes of ranking SNPs, the sign of the correlation coefficient does not matter as much as its modulus.

We also looked across our analyses to determine which set of priors produced the highest approximate Bayes factors. Higher ABFs suggest that the alternative model explains the data better, compared to the null, which has the same probability

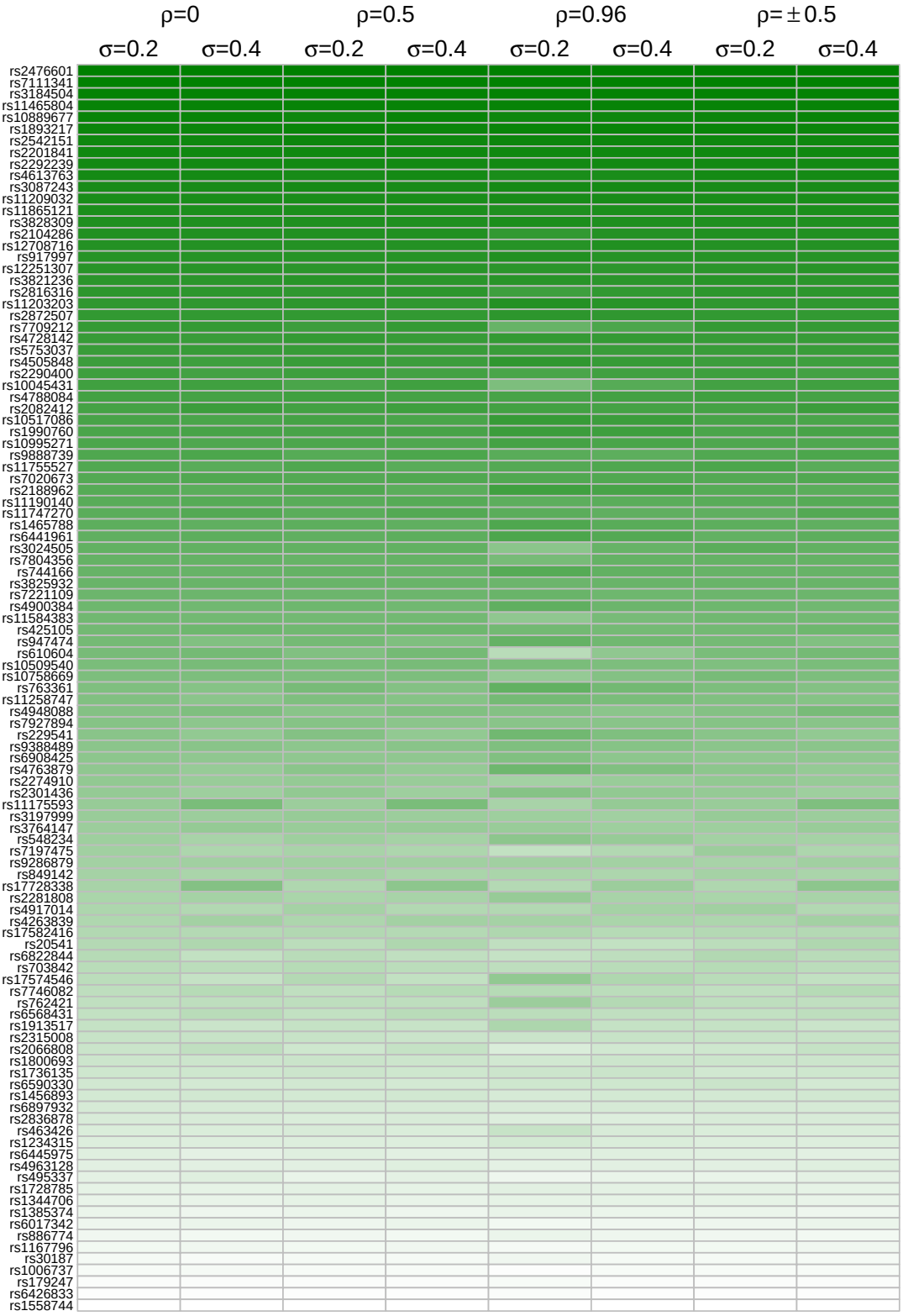


Figure 3.1: Comparison of how the SNPs in the meta-analysis were ranked in each of the analyses (parameters used are at the top of the plot). Darker greens indicate higher ranks (indicating higher ABFs), while paler colors correspond to lower ranks. The table is ordered by the ranks in the first column, where the analysis was conducted with $\rho = 0$ and $\sigma = 0.2$.

density across all analyses at a given SNP. For 64 of the 107 SNPs (almost 60%), the highest mean ABFs were observed in the analysis with priors $\rho = \pm 0.5$ and $\sigma = 0.2$. A further third of the SNPs (37) had the highest mean ABFs with $\rho = 0.96$ and $\sigma = 0.2$. When looking for the maximum ABFs at each SNP across all subsets of association in all models, these two analyses produced the highest ABFs for 69 and 29 SNPs, respectively.

It is not especially surprising that that the model using $\rho = \pm 0.5$ would yield higher ABFs than those from analyses that used the same flat prior correlation among the studies at all SNPs. In this analysis, the prior correlation on true effect sizes was created using information on direction of effect at each SNP—that is, it describes the data better than the other analyses, because the data used in the analyses was also used to create the prior. When the two analyses that made use of this prior are removed from consideration, the analyses producing the highest mean ABFs and the highest ABFs for each SNP across all analyses and subset models tend to have a slightly more even distribution among the remaining priors. We summarize these in Tables 3.1 and 3.2. We note that the highest means and overall ABFs tend to correspond to prior $\sigma = 0.2$ rather than $\sigma = 0.4$, suggesting that the majority of effect sizes are not large.

$\rho \backslash \sigma$	0.2	0.4
0	22	3
0.5	39	3
0.96	39	1

Table 3.1: Distribution of the number of SNPs whose highest mean ABF across all analyses corresponded to each value of ρ and σ when $\rho = \pm 0.5$ is removed from consideration.

$\rho \backslash \sigma$	0.2	0.4
0	49	7
0.5	14	2
0.96	33	2

Table 3.2: Distribution of the number of SNPs whose highest ABFs across all analyses and subset models corresponded to each value of ρ and σ when $\rho = \pm 0.5$ is removed from consideration.

Finally, different priors had moderate effects on the rankings of models of association at a given SNP. At 39 SNPs, there was no disagreement among the methods about the most likely true underlying model of association. At 42 SNPs, the highest ABFs across all analyses were associated with models that agreed on the association

status for all but one of the studies. At 20 SNPs, the analyses suggested three possible models of association, depending on the prior parameters; and only 6 SNPs had four different models associated with the highest ABF, depending on the prior.

3.4.2 Effect of priors on models of association

We also conducted an investigation into the effect of the different priors on models of association described in Section 3.3. Figure 3.2 shows prior and posterior probabilities on the number of associations over the whole dataset. Six SNPs were missing summary data for one of the studies and one SNP was missing data for three of the studies. This is why the distibution of prior probabilities under the flat prior (Figure 3.2a) is not perfectly symmetrical, and also why the combinatorial prior (Figure 3.2b) shows slightly smaller probabilities for models with 5-7 associations. At models with full data, the weights on number of associations do form a symmetric distribution, and are equal under the combinatorial prior.

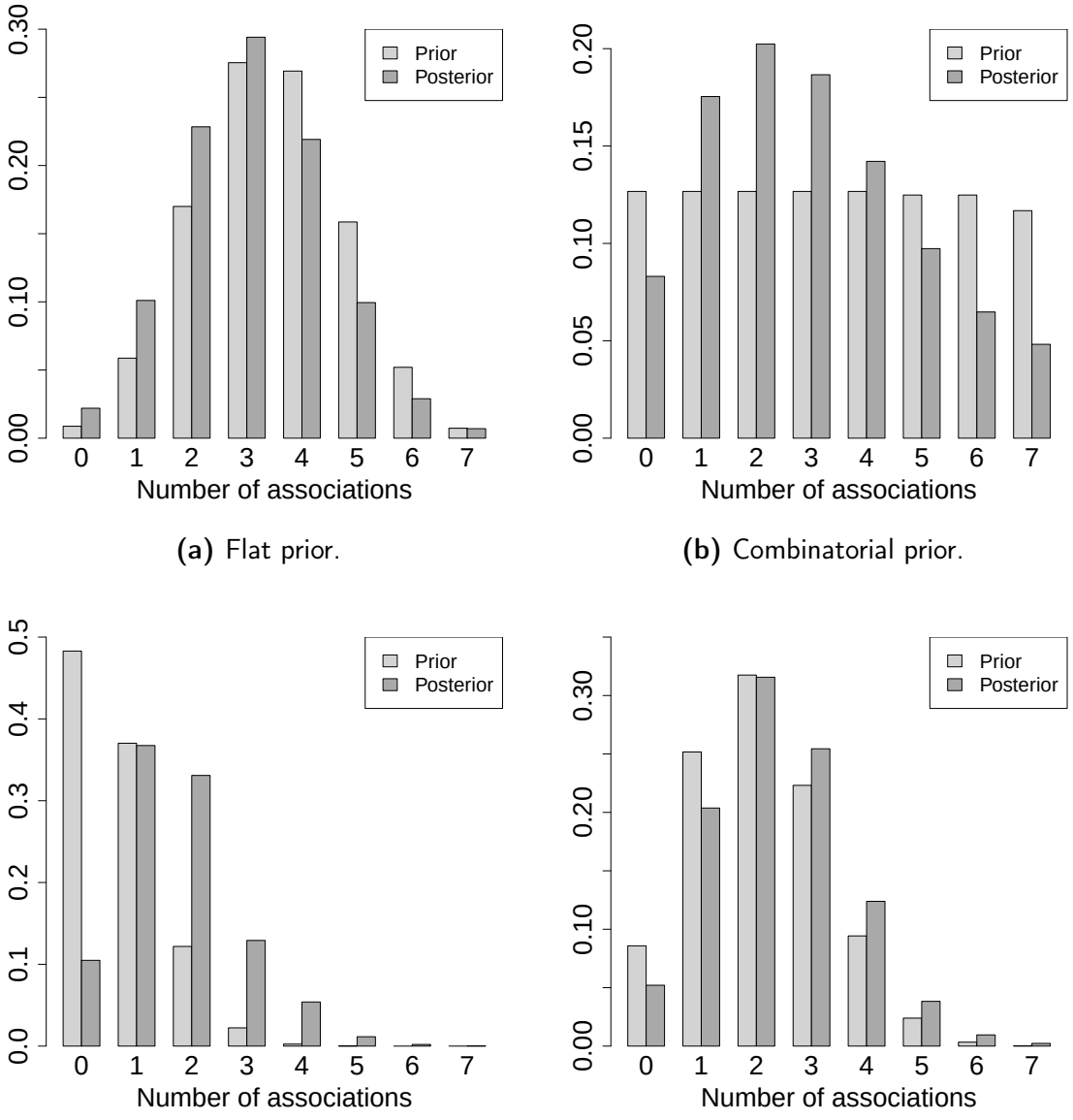
The posterior probabilities were calculated by multiplying the ABFs calculated with $\rho = 0$ and $\sigma = 0.2$ by the prior weights and then normalizing at each SNP, and then dividing by 107 (the number of SNPs in the study) to get the distribution of weights over the whole dataset on the same scale as the prior weights. We note that all posterior distributions display some degree of positive skewness, suggesting that most of the SNPs in our data show evidence of true associations with three or fewer studies. We also note that the degree of skewness increases based on the skewness of the prior. The posterior calculated for the flat prior is the most symmetric of the four distributions.

We looked at the effect of these priors on model weights of the approximate Bayes factors. Though visual inspection of the plots in Figure 3.2 suggests that the binomial prior with $p_i = 0.3$ might be the one that fits the data the best, it was actually the flat prior that yielded the highest overall approximate Bayes factors. However, this does not necessarily mean that we should not consider using these other priors. Despite the fact that the prior with $\rho = 0$ and $\sigma = 0.2$ yields smaller ABFs than the one with $\rho = \pm 0.5$ and $\sigma = 0.2$, we chose the former prior to conduct our follow-up analyses because it is agnostic to the direction of a SNP's effect on disease risk. The priors where $\rho \neq 0$ assume that the directions and sizes of effects are accurately reported in

our data, which we cannot guarantee.

However, all of that being said, there is reason to prefer the flat prior in our further analyses. The SNPs chosen for inclusion in this dataset were chosen because of their association with at least one immune-mediated disease—though not necessarily one of the seven in the dataset. Thus, there is no reason to make the usual assumption in GWAS that the majority of SNPs in the dataset will not be associated with any phenotype. Additionally, the known shared genetic relationships among autoimmune diseases make it plausible that all phenotypes might be associated with at least a few markers in the dataset. Hence, the binomial priors do not seem applicable here.

Furthermore, while the combinatorial prior is symmetric, and thus encodes a belief that the null model of association is just as likely as the model where every phenotype is associated with the SNP, it also encodes the belief that both models are n (the number of phenotypes) times as likely as a model where exactly one disease is associated, or a model where all but one disease is associated. Similarly, at a SNP where we have data for all seven studies, the null model (or the model of total association) is 35 times as likely as a model with just three or four associations. There is no real justification for believing any of that to be the case in our data. Furthermore, it seems odd to enforce a belief about the number of associations, rather than the specific traits that are associated. It is hard to imagine a circumstance where one believes *a priori* that there is a particular probability that m traits are associated with a marker, but is agnostic about which specific traits those are. Hence, we choose the flat prior for our further analyses.



(a) Flat prior.

(b) Combinatorial prior.

(c) Binomial prior with $p_i = 0.1$.

(d) Binomial prior with $p_i = 0.3$.

Figure 3.2: The bar plots show distribution of prior and posterior probabilties for different numbers of true associations over the whole data set, under different priors on models of association. Posterior probabilities were calculated using approximate Bayes factors with prior parameters $\rho = 0$ and $\sigma = 0.2$.

3.4.3 Application of the ABF approach to the data

We summarize the results of the analysis with $\rho = 0$ and $\sigma = 0.2$ for each SNP in Figure 3.3. As described above, we used a flat prior on the models of association. The SNPs are ordered according to their mean ABF (the $\log_{10}$ values of these means are reported) over all possible models of association at the SNP. We note that if we require that the mean ABF across all models of association at a given SNP is as high as 6—a stringent threshold, which might be necessary, given some of the uncertainties

about the data—then 44 SNPs are highlighted by our analysis.

We see immediately that the number of assocations in the model corresponding to the highest ABF does not necessarily correspond to higher mean ABFs. We also note that while all studies with high mean ABFs have at least one strong association, with a marginal probability that rounds to 1.000, the evidence for further associations may not have such high marginal probabilities. In fact, rs12251307 has a $\log_{10}$(mean ABF) of 11.830, but the most likely model shows a single association with T1D.

As Figure 3.2 indicated, most of the SNPs in our dataset show associations with three studies or fewer. Notable exceptions are rs2476601, rs3184504, rs1893217, and rs2542151, which are all in among the top 10 mean ABFs listed in Table 3.3.

The 10 SNPs with the highest ABFs

The SNP with the largest mean ABF was rs2476601, which is a well-known a missense variant in *PTPN22*. It has been associated with Crohn's disease [142], rheumatoid arthritis [139], systemic lupus erythematosus [141], and type 1 diabetes [134], which is the model of association found in Figure 3.3. Similarly, the SNP rs7111341 has been previously associated with type 1 diabetes [142], but not with any other autoimmune diseases—which matches the pattern of association our analysis found for it (though we did find some weak evidence for an association in celiac disease, as well). SNPs that appear to be associated solely with type 1 diabetes, both in our results and in the literature, are rs12251307 and rs5753037 [146, 138].

The SNP rs3184504 is the only SNP that our analysis found to be associated with all studies. It is a missense variant of *SH2B3* and is associated with a diverse set of traits, from rheumatoid arthritis [139] and type 1 diabetes [138] to coronary artery disease [147] and colorectal cancer [148]. On the opposite end of the spectrum, rs11465804 only seems to have one association in the literature, with Crohn's disease [142]. Our analysis put a posterior probability of 0.9997 on an association between rs11465804 and psoriasis. This finding may be due to the SNP being in high LD ($R^2 = 0.9034$ in the 1000 Genomes European populations) with a rs11209026, which has been associated with psoriasis [62].

Further upstream, though still in the gene *IL23R*, rs10889677—a known in-

SNP	$\log_{10}$ (Mean ABF)	Gene/Region	Mean Number of Associations	GWAS Catalog
rs2476601	159.33	*PTPN22*	5.00	6 (2)
rs7111341	48.01	*INS*, *TH*, *ASCL2*	2.49	2 (1)
rs3184504	43.35	*SH2B3*	6.34	18 (16*)
rs11465804	32.56	*IL23R*	3.58	1 (0)
rs10889677	25.18	*IL23R*	3.42	2 (1)
rs1893217	24.70	*PTPN2*	5.21	5 (1)
rs2542151	24.44	*PTPN2*	5.23	2 (0)
rs2201841	23.38	*IL23R*	3.24	3 (1)
rs2292239	21.99	*ERBB3*	2.97	2 (1)
rs4613763	20.00	*PTGER4*	3.04	2 (0)
rs1344706	-0.88	*ZNF804A*	2.13	1 (1)
rs1385374	-1.02	*SLC15A4*	2.05	1 (0)
rs6017342	-1.10	*HNF4A*	1.25	2 (2*)
rs886774	-1.19	*LAMB1*	1.72	1 (1)
rs1167796	-1.21	7q11.23	1.19	1 (0)
rs30187	-1.36	*ERAP1*	1.48	1 (1)
rs1006737	-1.38	*CACNA1C*	1.42	2 (2)
rs179247	-1.42	*TSHR*	1.37	NA
rs6426833	-1.44	1p36.1	1.35	1 (1)
rs1558744	-1.59	12q15	1.09	1 (0)

Table 3.3: Table showing the top and bottom 10 SNPs (using prior $\sigma = 0.2$ and $\rho = 0$), ranked by mean ABF across all possible models of association. Additionally, we show the the consensus candidate gene reported by Cotsapas *et al.*, the GWAS Catalog [25], and the Ensembl genome browser; the weighted mean number of phenotypic associations, (weights are calculated from the posterior probability distribution, which is calculated from the ABFs on all models); and the number of traits associated with the SNP in the GWAS catalog, Immunobase, and Ensembl, with the number of new traits that are not in our study in parentheses. Possible new diseases included inflammatory bowel disease (IBD), which comprises Crohn's disease and ulcerative colitis. Unless the marker was associated with both diseases individually, we treated this as a new trait, but marked entries that included it with (*). There was no information available in the GWAS Catalog for the rs179247.

flammatory bowel disease locus [149, 60]—appears in our data to be associated with psoriasis (marginal probability ≈ 1.000) and multiple sclerosis (marginal probability 0.818), as well Crohn's disease (marginal probability rounds to 1). The psoriasis as-

sociation is particularly hard to explain, since the model of association at this SNP depicted in Figure 3.3 is 45,000 times likelier than any model without a psoriasis association. *IL23R* does have other psoriasis associations, as previously discussed. It may be that the apparent association at this SNP is tagging a true association nearby.

The SNPs rs1893217 and rs2542151 display similar patterns of assocation and are in high LD with one another ($R^2 = 0.9673$ in European populations). The former is associated with more diseases in the literature: celiac disease [135], Crohn's disease [59], type 1 diabetes [138], and ulcerative colitis [137]. Additionally, there is some evidence that it may be associated with rheumatoid arthritis, which is often comorbid with celiac disease [150]. The SNP rs2542151, on the other hand, only has associations with Crohn's disease [142] and type 1 diabetes [134]. The high LD between the two SNPs and their associations in the literature explains their similar patterns of associations in Figure 3.3; however we could not find anything to corroborate the putative association with psoriasis in our analysis. With a marginal probability of association of 0.700 at rs1893217 and 0.688 at rs2542151, it is entirely possible that these associations are spurious.

The third representative of *IL23R* in our table is rs2201841, which has published associations with Crohn's disease [149], ulcerative colitis [151], and psoriasis [144]. Our analysis replicates the associations with Crohn's disease and psoriasis (the respective marginal probabilties of both traits round to 1), and suggests a possible association with multiple sclerosis, which has a marginal probability of 0.659. Similarly, at rs2292239, there is a very strong association with type 1 diabetes (which replicates published findings [134, 138]), and two weaker associations with psoriasis and celiac disease (marginal probabilties of 0.547 and 0.530). The only other trait this SNP appears to be associated with in the literature is alopecia areata [152], which does not explain the extra associations suggested by our analysis. These SNPs stand in contrast to rs4613763, which, according to the literature, is associated with multiple sclerosis [140] and Crohn's disease [142], which are precisely the two diseases highlighted at this SNP in our analysis. The marginal probabilities on the other diseases were at most 0.315 (in psoriasis). This is in stark contrast to the high marginal probabilities on multiple sclerosis and Crohn's disease (0.999 and 1.000, respectively).

The 10 SNPs with the lowest ABFs

Our 10 lowest ranked SNPs all had mean ABFs that were less than 1 (or, on a $\log_{10}$ scale, less than 0), suggesting that the evidence for disease associations at these SNPs is generally not significantly better than the evidence for the null model. The two SNPs for which the highest ABF corresponded to non-null models—rs1344706 and rs6017342—had ABFs of 1.92 and 1.53, respectively, which are generally not considered high enough to accept the alternative model. Despite the lack of evidence in our data, rs1344706 has a possible association with schizophrenia [153]. Another schizophrenia locus on our list is rs1006737 [154], which also shows possible associations with bipolar disorder [155]. Rs1385374 and rs1167796 have associations with systemic lupus erythematosus [156, 157], while rs6017342, rs886774, rs6426833, and rs1558744 are all associated with ulcerative colitis [158, 159, 160, 151]. The SNP rs30187 is a coding variant of *ERAP1* and is a well-known ankylosing spondylitis locus [61], which also showed a possible association with psoriasis [62].

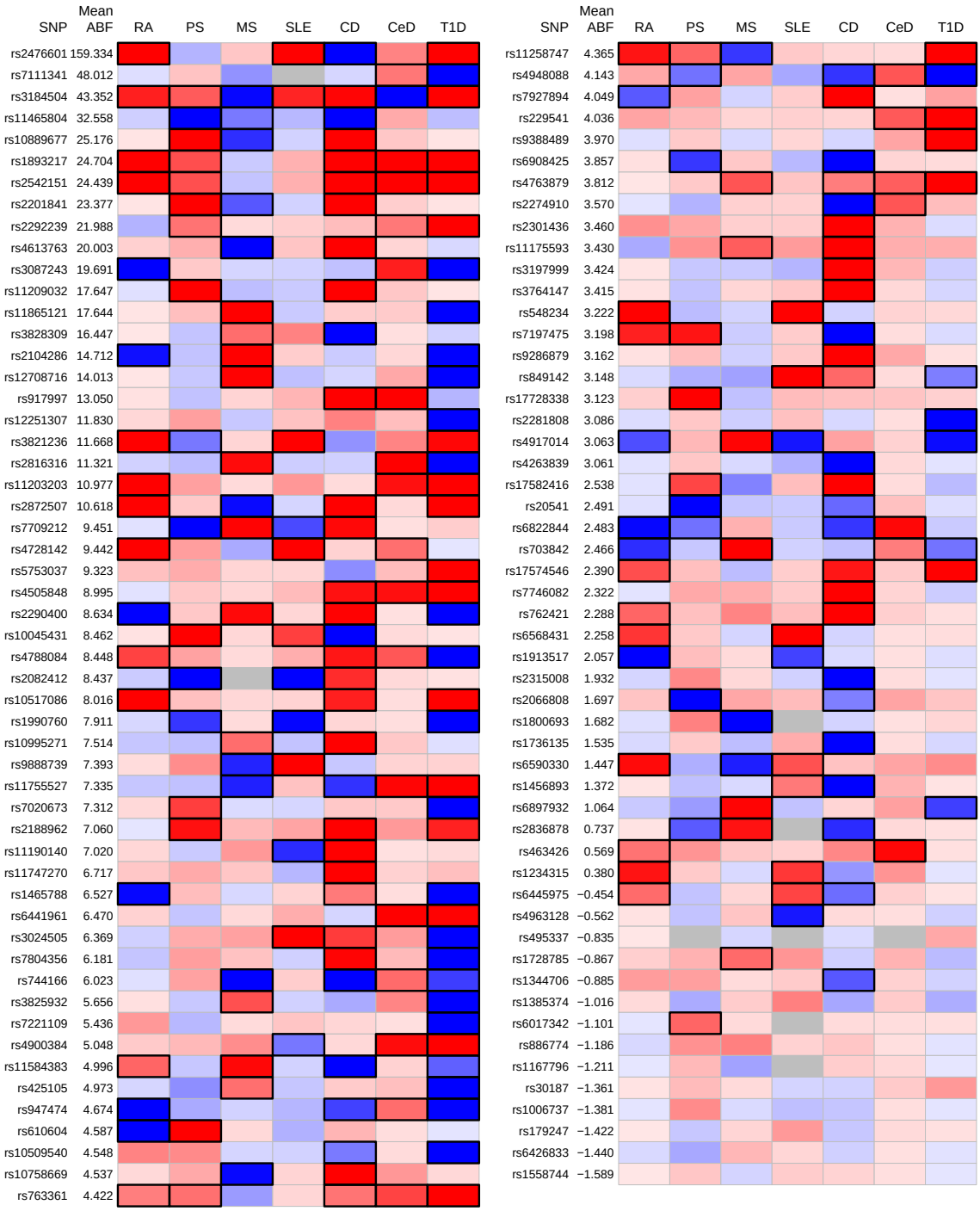


Figure 3.3: Results of our analysis for each SNP. Rectangles outlined in black correspond to diseases that were significantly associated with the SNP according to the model with the highest ABF. The color of the rectangle is based on the direction of the effect (the sign of the beta value) with blue indicating a protective effect and red indicating an increased risk of disease. The intensity of the color corresponds to the amount of marginal posterior probability weight on the association of the disease with the SNP—darker colors represent increased posterior probability compared to paler colors. Grey boxes indicate diseases for which the data necessary to perform the analysis was missing. The $\log_{10}$ (Mean ABF) values are reported in the "Mean ABF" column.

As one might expect from our top and bottom 10 SNPs, the top 10 had more associations in our data—Welch's T-test for the distribution of mean associations for the SNPs with the 10 highest mean ABFs compared to the ones with the 10 lowest returned a p-value of 0.0001, suggesting that the respective sample means of 4.05 and 1.51 are significantly different from one another. What is interesting is that searching through databases and the literature showed that the bottom 10 SNPs in terms of mean ABF did not necessarily lack associations with any trait—only rs179247 had no known associations. The rest were either associated with exactly one disease in our dataset that we could not replicate, or were associated with traits for which we had no data.

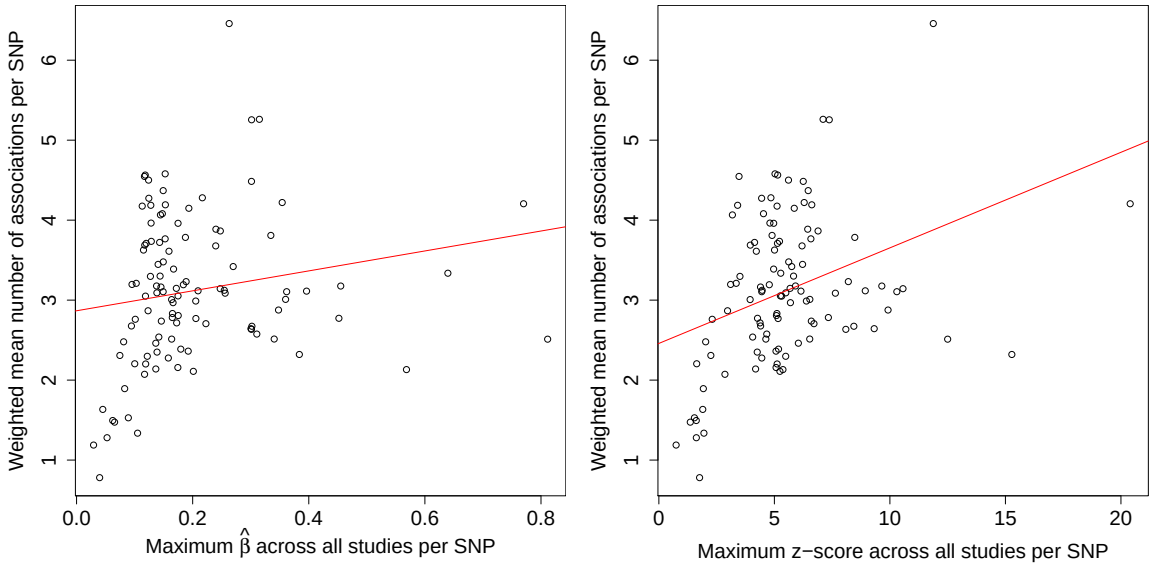
Since a high mean ABF can be the result of a very strong association in a single study or of multiple associations, it is worth asking the question: what is the effect of a large $\hat{\beta}$ in one study on the number of associations overall? To answer this, we collected the absolute values of most extreme effect sizes and z scores (positive or negative) for each study per SNP and compared them to the mean number of associations at each SNP. The latter is simply the mean of the number of possible associations at a SNP, weighted by the marginal posterior probability on the number of associations, based on the distribution of normalized ABFs at that SNP. That is, for n associations ($n \in \{0, \ldots, 7\}$) with posterior probability p_n, the estimated number of associations at the ith SNP, A_i, is simply

$$A_i = \sum_{n=0}^{7} np_n \tag{3.6}$$

These values tell us, at each SNP, how many associations we expect based on the posterior probability distribution on models of association.

The correlation between the mean number of effects and the maximum $\hat{\beta}$s was 0.336, and between the mean number of effects and maximum z score, the correlation was 0.494. In a linear regression, shown in Figure 3.4, the regression coefficient for the number of associations regressed onto maximum $\hat{\beta}s$ was 2.3651 with a standard error of 0.6475. This means that if the absolute value of $\hat{\beta}$ at a SNP increases by 0.1, the mean number of associations at that SNP is expected to increase by 0.237. For the regression of mean number of effects on z scores, the regression coefficient was 0.1589 with a standard error of 0.0288. This means that for every gain of 1.5 in the

absolute value of the z score, the mean number of associations increases by 0.238.

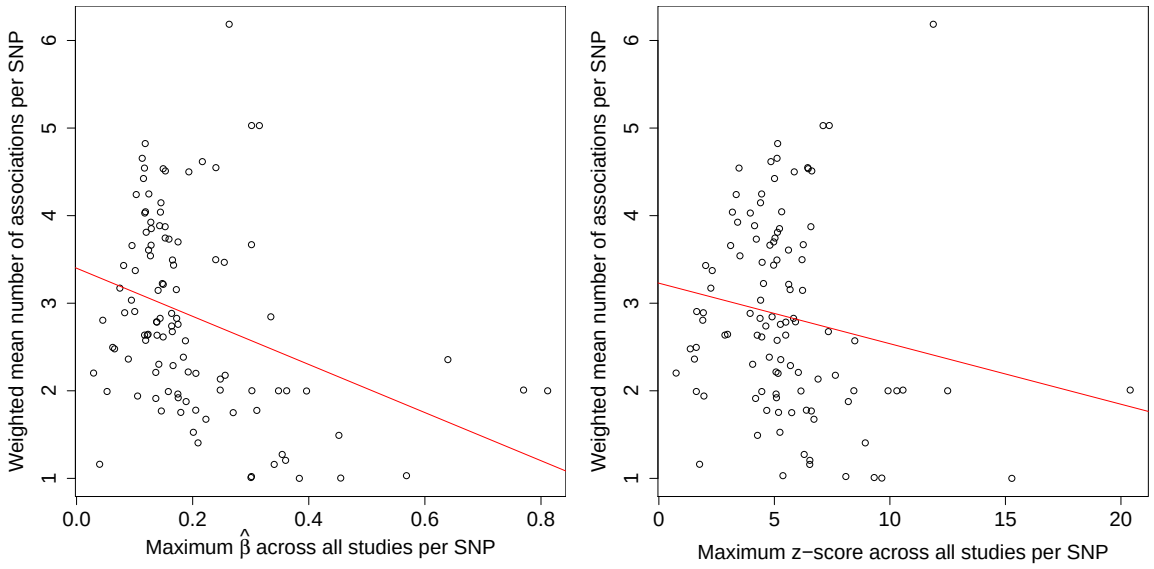


(a) The weighted mean number of associations per SNP as a function of maximum $\hat{\beta}$.

(b) The weighted mean number of associations per SNP as a function of maximum z score.

Figure 3.4: Plots of the weighted mean number of associations at each SNP according to the ABF analysis that used priors $\rho = 0.5$ and $\sigma = 0.2$ (y-axis) as a function of either maximum $\hat{\beta}$ or maximum z score (x-axis). The red lines shows the least-squares regression line through the data.

Initially, it seems that the answer to our question is that a large effect size—either in terms of the estimated effect size, $\hat{\beta}$ or in terms of the z score—in one study increases the average number of associated phenotypes we would expect to see at that SNP. However, the distribution of posterior probabilties on models is highly dependent on the prior we choose. For instance, in our dataset, if we choose a prior $\rho = 0.96$ and $\sigma = 0.2$, then Figure 3.5 shows that larger effect sizes lower the mean number of associations.



(a) The weighted mean number of associations per SNP as a function of maximum $\hat{\beta}$.

(b) The weighted mean number of associations per SNP as a function of maximum z score.

Figure 3.5: Plots of the weighted mean number of associations at each SNP according to the ABF analysis that used priors $\rho = 0.96$ and $\sigma = 0.2$ (y-axis) as a function of either maximum $\hat{\beta}$ or maximum z score (x-axis). The red lines shows the least-squares regression line through the data.

3.4.4 Principal Component Analysis

Often when working with large datasets, researchers will use Principal Component Analysis (PCA) to account for underlying population structure and as a way of learning about the relatedness of different subpopulations to one another (see Section 1.4.3 in Chapter 1).

We applied this to investigate how the autoimmune diseases in our dataset relate to one another. In this analysis, we treated each disease as if it were an individual. From the results of the subset-exhaustive approach we created a matrix of marginal posterior probabilities for each SNP. The cell corresponding to a given row (SNP) and a given column (diseases) shows the probability of association of the disease at that SNP.

We performed a singular value decomposition on this matrix and stored the loadings. At each SNP, there is a probability distribution over all possible patterns of associated diseases with the SNP. We wanted to incorporate the uncertainty of the model selection into our analysis, since the probability distributions on these models are what determine the relationships between the diseases according to the PCA. To this end, we generated new matrices by taking random samples from the posterior

distributions at each SNP, projecting the resulting 107×7 binary matrices through the stored loadings, and coloring the seven newly projected points according to the corresponding disease. We performed 1000 replications of this analysis. The result of this sampling scheme are clusters of diseases, as opposed to single points. The more dispersed clusters indicate more uncertainty about the correct models of association in the SNPs used in the analysis, and consequently, the relationships between the diseases. The results for the first two PCs, which account for 29.7% and 18.0% of the total variance, can be seen in Figure 3.6.

Our PCA plot indicates that Crohn's disease and type 1 diabetes are more different than the rest of the autoimmune diseases, though they are aligned along along PC1. We also see that rheumatoid arthritis, celiac disease, systemic lupus erythematosus, and multiple sclerosis, form one large cluster, with part of the psoriasis cluster at the bottom. We only had about 100 SNPs in our analysis, which begs the question of what this plot would look like when created from a larger dataset—for example, would we see better separation of the four or five diseases that cluster together in this analysis? We will attempt this with the Immunobase data in Chapter 5.

Few studies look at more than two of these diseases together, however a similar, gene-based approach has been published [161], and their results, shown in Figure 3.7a, show similar differentiation of Crohn's disease and other autoimmune diseases. However, they disagree with our analysis in that they found multiple sclerosis, rheumatoid arthritis, and psoriasis to cluster close together, while celiac disease and systemic lupus erythematosus lay further out on their PC1 axis. Type 1 diabetes lay between the clusters containing multiple sclerosis and celiac disease. When the genes from the HLA region were removed from the analysis, as in Figure 3.7b, we see that the differentiation of Crohn's disease from the other autoimmune diseases remains, but that celiac disease and systemic lupus erythematosus cluster with the other autoimmune diseases, save for one of the multiple sclerosis cohorts and vitiligo, which is not in our dataset [161].

A review of autoimmune diseases [47] looked at the comorbidities of immune-mediated disease with one another, as well as the types of diseases that tend to occur within families, given a proband with a particular illness. They found that Crohn's

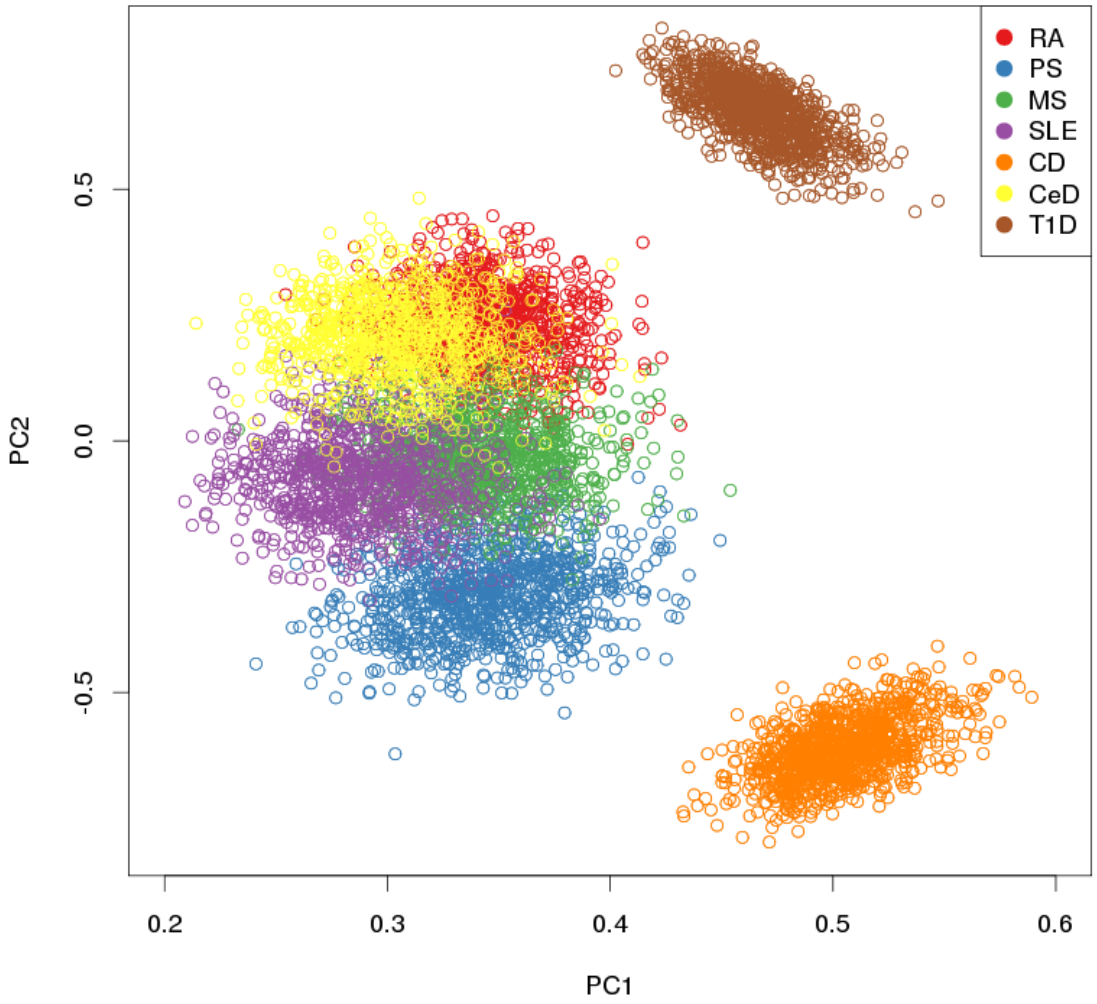
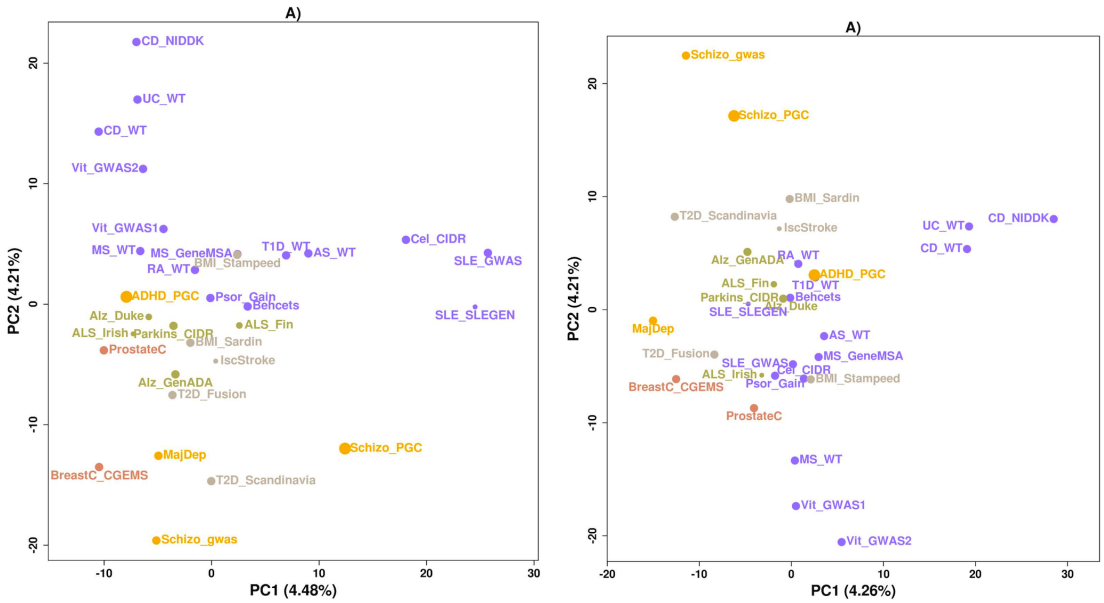


Figure 3.6: Results of the PCA analysis. We performed our ABF analysis on the 107 SNPs from the Cotsapas *et al.* paper [52] and obtained probability distributions over all models of association at each SNP, which we used to calculate marginal probabilities of association for each disease at each SNP. We performed a singular value decomposition using the matrix of marginal probabilities and stored the loadings. We performed 1000 resamplings of models of association according to the probability distribution at each SNP to create new binary matrices of patterns of association of the seven diseases over all SNPs in the dataset and projected these through the stored loadings. The resulting clusters give some indication of the relationship between the diseases in our dataset, with some measure of uncertainty based on the probability distribution over models of association.

disease and psoriasis showed evidence of aggregating in families together, while type 1 diabetes, celiac disease, and rheumatoid arthritis formed another cluster of diseases that tend to afflict members of the same family. Family clusters could not be determined for celiac disease, systemic lupus erythematosus or psoriasis, but multiple sclerosis showed increased occurrence in patients with psoriasis or inflammatory bowel disease. Our PCA analysis does not suggest much of a relationship between multiple



(a) Figure 4A of Chang *et. al.*'s paper. This analysis included genes from the HLA region. **(b)** Figure 6A of Chang *et. al.*'s paper. This analysis excluded genes from the HLA region.

Figure 3.7: Results of a gene-based PCA analyses of various diseases, including autoimmune diseases (in purple), performed by Chang *et. al.* [161]

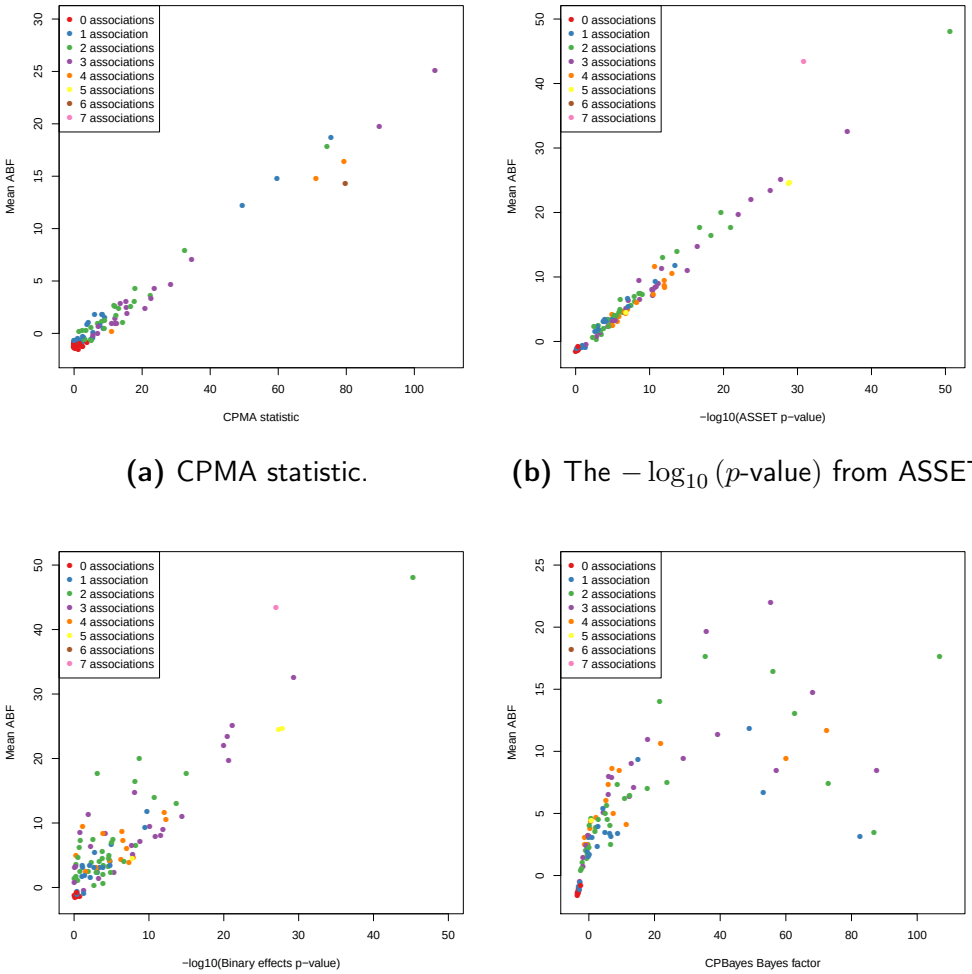
sclerosis and Crohn's disease, though multiple sclerosis does cluster next to psoriasis on our plot. Finally, according to the Zhernakova *et al.* study, rheumatoid arthritis co-occurs with type 1 diabetes, systemic lupus erythematosus, and multiple sclerosis, which are relationships that are suggested by Figure 3.6, though the cluster representing celiac disease, which lies at the intersection of the rheumatoid arthritis, multiple sclerosis, and systemic lupus erythematosus clusters remains uncorroborated.

3.5 Comparison with other methods

Now that we have seen some of the results of applying our method to the data, we look at the conclusions we might draw from applying other methods to this dataset. The full results for ASSET, the binary effects model, and CPBayes are in Section 3.A of the Appendix. The full results for the *CPMA* can be found in Cotsapas *et al.* (2011). Here, we discuss general trends and patterns among the results.

3.5.1 Correlations between test statistics

Figure 3.1 shows the mean ABF (from the analysis where $\rho = 0$ and $\sigma = 0.2$) plotted against the test statistics of the various other methods. The dots are colored by the



(a) CPMA statistic.

(b) The $-\log_{10}(p\text{-value})$ from ASSET

(c) $-\log_{10}(p\text{-values})$ from the binary effects model statistic.

(d) CPBayes Bayes factors.

Figure 3.1: Plots of the $\log_{10}(\text{mean ABFs})$ (y-axes) against the test statistics of each of the other methods (x-axes). The colors of the dots correspond to the number of phenotypic associations in the model associated with the highest approximate Bayes factor in the subset-exhaustive approach. In each analysis, there was an outlying value at rs2476601, which we removed from the plots, except in the comparison to ASSET, where the extreme value did not cause readability issues.

number of associations in the model corresponding to the highest ABF in our analysis. This is to see if there is any pattern to the types of SNPs where the two methods are especially concordant or especially divergent. We can see that the ABFs we calculated are highly correlated with the test statistics from ASSET and the CPMA and have a lower correlation with the binary effects model test statistic. Surprisingly, there does not appear to be a high correlation between CPBayes' Bayes factors and the ABFs, and we note that CPBayes' Bayes factors show a larger range of values than the ABFs do.

One challenge presented by looking across all these methods is to define what a significant SNP is—is it a SNP that shows multiple associations? A SNP with a low p-value? Even if we accept any SNP for which we reject the null hypothesis of no association, we are still left with the problem of determining the threshold (p-value or Bayes factor) for each method. This needs to take into account features of the method itself, as well as the complexities of our data—unaccounted for shared controls might warrant a more stringent threshold than the one used for GWAS, but a dataset of 107 SNPs might suggest that a lower threshold should be employed, since the number of tests being done is much smaller than in a GWAS. We sidestep this issue by comparing the rankings of SNPs in each method, and calculating Kendall's τ between the results from each method, which are reported in Table 3.1 (and visualized these in Figure 3.B.1 in Section 3.B).

Method	ABF	CPMA	ASSET	Binary effects	CPBayes
ABF	1	0.441	0.906	0.612	0.792
CPMA	0.441	1	0.470	0.348	0.338
ASSET	0.906	0.470	1	0.587	0.719
Binary effects	0.612	0.348	0.587	1	0.513
CPBayes	0.792	0.338	0.719	0.513	1

Table 3.1: Correlation matrix (correlation coefficients are Kendall's τ) calculated between between the results of each pair of methods.

We see that ASSET's rankings are the most like our own, followed by CPBayes, then the binary effects model, and finally the CPMA. We also note that the highest correlations for ASSET, binary effects, and CPBayes all occur when paired with the ABF approach. Unsurprisingly, the lowest correlations for these methods as well as our own occur in pairings that involve CPMA method. This is because the *CPMA* is a test of further associations; it requires that the study for which an association with a given SNP was first established be removed from the meta-analysis. It is unique among the methods for having this requirement. As a result, SNPs like rs7111341, rs2292239, and rs3828309—which all have $\log_{10}(\text{mean ABF}) \geq 16.447$ and are ranked in the top quarter of all SNPs by the non-CPMA methods—are ranked in the second half of all SNPs using the *CPMA* statistic.

At these SNPs, the other methods calculate high marginal probabilities of association with the ascertainment disease and low or moderate marginal probabilities

on the further associations. Because these data are excluded from the *CPMA*, it is understandable that it would rank these SNPs lower than the other analyses do. Additionally, the one method that selected a model that included extra associations at these SNPs was the ABF approach. These extra associations are probably spurious, as they are not corroborated in Immunobase or the GWAS catalog. This highlights a potential danger of our method: it produces the model corresponding to the highest ABF, "the most likely model", however when marginal probabilities of association are not high (≥ 0.90), it is necessary to exercise caution about declaring associations.

We now consider the patterns of association that the non-CPMA methods find for each SNP. In order to examine these results, we consider the top 40 markers from each analysis. This is because we want to examine results for which we can be reasonably certain we would not accept (or fail to reject) the null hypothesis. Since both the ABF and CPMA approaches found over 40 "significant" SNPs in this dataset, it seems plausible that the other methods would find a similar number of associations.

Table 3.2 summarizes the thresholds corresponding to the top 40 SNPs for each analysis. Any one of these is theoretically defensible, given the features of our data. Admittedly, choosing the top 40 markers in each analysis—as opposed to any other number—is still somewhat arbitrary; however it avoids problems such as determining a p-value for CPBayes' Bayes factors or justifying whether the p-values for the CPMA, binary effects, and ASSET methods should all be subject to the same thresholds or a different one for each, and in any case, trying to determine exactly what those thresholds should be.

ABF	CPMA	ASSET	Binary effects	CPBayes
mean ABF > 6.5	$p < 4.1 \times 10^{-3}$	$p < 1.1 \times 10^{-8}$	$p < 1.0 \times 10^{-5}$	BF > 9.0

Table 3.2: Effective thresholds for each analysis method for the SNPs with the 40 highest test statistics.

3.5.2 Models of association in the non-CPMA methods

Table 3.3 summarizes how many "significant" SNPs overlap in each pair of analysis methods when using this threshold (lower triangle). It also states for how many of these shared SNPs the two analysis methods predicted the same pattern of association (upper triangle). The patterns of association determined by the binary effects model

are simply the studies that achieve an m score of at least 0.9 at each SNP. We used both this threshold and a lower one of $m \geq 0.5$ to calculate patterns of association for this method.

Method	ABF	ASSET	Binary effects ($m \geq 0.9$)	Binary effects ($m \geq 0.5$)	CPBayes
ABF	×	10	10	9	14
ASSET	38	×	14	14	13
Binary effects ($m \geq 0.9$)	31	31	×	28	18
Binary effects ($m \geq 0.5$)	31	31	40	×	14
CPBayes	34	33	27	27	×

Table 3.3: Table whose lower triangle contains the number of markers whose test statistic ranks were between 1 and 40 in both methods for each pair of methods. The upper triangle shows the number markers at which both methods selected the same subset model of association as the most likely model, among the markers that were ranked between 1 and 40 by both methods.

Despite giving the highest correlations when comparing test statistic rankings among the non-CPMA methods, the ABF approach shows considerably less concordance with the other methods when we consider the patterns of associations at the top 40 SNPs for each analysis. Unsurprisingly, the two binary effects methods agree about the pattern of association the most frequently (70% of the time), though CPBayes also shows a high degree of agreement with the binary effects models. While the total number of "significant" SNPs shared between the two methods is low (only 27), CPBayes estimates the same pattern of association at 67% and 52% of these SNPs when $m \geq 0.9$ and 0.5 in the binary effects model, respectively. None of the other pairwise comparisons agree as frequently as 50% of the time, though they do all have more significant SNPs in common.

We investigated further the differences in the patterns of association determined by each method and summarize our findings in Table 3.4. For a given pair of methods, again we consider the set of SNPs found to be significant by both methods. Each entry shows the number of associations across all SNPs in the set which were found in Method 1 (rows), but not in Method 2 (columns).

We see a divide—the ABF approach and ASSET tend to call far more associations that are not found by binary effects and CPBayes than vice versa. The binary effects model where $m \geq 0.9$ is the method that is the most conservative about calling

		Method 2				
		ABF	ASSET	Binary effects ($m \geq 0.9$)	Binary effects ($m \geq 0.5$)	CPBayes
Method 1	ABF	×	22	31	29	24
	ASSET	21	×	33	28	32
	Binary effects ($m \geq 0.9$)	0	3	×	0	4
	Binary effects ($m \geq 0.5$)	5	5	19	×	10
	CPBayes	0	11	13	12	×

Table 3.4: Table showing the number of disparate calls between results from each method at markers that were ranked between 1 and 40 by both methods. Each cell shows the number of associations in the most likely models according to Method 1 (rows) that were not found in the most likely models from Method 2 (columns).

associations. Since the m score is the marginal probability of association between a study and the SNP, and since none of the other methods require equally stringent marginal probabilities of association, it is to be expected that this model would call fewer associations compared to the others.

However, this conservatism can cause true associations that are detectable by other methods to be overlooked. As an example, we consider the SNP rs3184504, where both the ABF and CPBayes analyses calculate true associations in all seven diseases, though both with a wide range of marginal probabilities ($[0.575 - 1]$ for the ABF analysis and $[0.694 - 1]$ for CPBayes). The binary effects approach only finds an effect in T1D, despite the fact that the literature states that there should be at least one more with rheumatoid arthritis. Additionally, ASSET—despite calling so many associations overall—found only two associations at this SNP: one in SLE and the other in T1D. We note that for the subset of traits with negative effect sizes, ASSET failed to calculate a p-value, which explains why it found no effects in multiple sclerosis or celiac disease (both diseases had $\hat{\beta} < 0$), but it does not explain why it found an effect in SLE but not in Crohn's disease. The effect size for SLE was smaller at this SNP than CD (0.104 for SLE vs 0.110 for CD) and the standard error was larger (0.039 for SLE vs 0.032 for CD). It is possible that the model of association among all diseases and rs3184504 is also not correct; however, given the literature, it seems that ASSET and the binary effects approaches have both somehow overlooked at least one true association with rheumatoid arthritis that the ABF and CPBayes approaches

were able to find.

However, this is not to say that more associations called necessarily means these associations are true. While we have discussed the likely spurious associations found by the ABF approach at rs7111341, rs2292239, and rs3828309, ASSET also reports some dubious findings. For instance, rs12251307 appears to be a type 1 diabetes locus [146, 138] without any further associations. This is the consensus among the ABF, CPBayes, and binary effects approaches. However, ASSET finds the optimal model to be the one with associations at type 1 diabetes, psoriasis, Crohn's disease, and celiac disease. Just as the ABF approach necessitates looking at marginal probabilities of association to ensure valid results, it is equally important to pay attention to the p-values associated with ASSET's "+" and "−" subsets. At this SNP, the "−" subset has a subset p-value of 2.65×10^{-15} and contains only type 1 diabetes; the others three traits are in the "+" group, which has a p-value of 0.387, suggesting that we should ignore this subset.

However, within subsets, there is no way of determining marginal probabilities using ASSET. For instance, at rs10995271, both ASSET and the ABF approach agree that there is evidence of association with Crohn's disease and multiple sclerosis. The other three methods do not find evidence of association with multiple sclerosis at this marker. The ABF approach puts a marginal probability of association with MS at 0.459. Published studies also indicate no association between this marker and MS. This demonstrates the advantage of calculating the probability distribution on all models of association, as the marginal probability of association says the balance of probability is against the association of the SNP with MS.

Furthermore, two of these methods—ASSET and binary effects—have difficulty handling data with large effect sizes, as at rs2476601. This SNP has two very large $\hat{\beta}$s—for rheumatoid arthritis and type 1 diabetes—which are 0.666 and 0.770, respectively, with standard errors of 0.037 and 0.038. These results are so extreme that ASSET cannot perform its calculation using these data—it simply prints out an error.

The difficulty that the binary effects model has manifests more subtly. There are potentially two more signals at this SNP in systemic lupus erythematosus and Crohn's disease (their respective effect sizes are 0.263 and -0.236 and standard errors are 0.066 abd 0.055). Both the ABF approach and CPBayes put high posterior probabilites on

the marginal associations of these diseases with this SNP. CPBayes calculates that these probabilities are 1 for RA, T1D, SLE, and CD, while the ABF analysis says the marginal probabilities are 1 for RA and T1D, 0.998 for SLE, and 0.996 for CD. Both methods select as their most likely models the model of association that includes all four of these diseases.

The binary effects method calculates an appropriately low p-value for the SNP, $p = 3.32 \times 10^{-155}$, but its m values (marginal probabilities of association) are 1 for RA and T1D and 0 for every other study, including SLE and CD. While it is possible that not every analysis method would agree that the most likely model of association includes SLE and CD, it is odd that this one puts the probability of association with these two diseases at 0. We have reason to believe that the associations found by the ABF approach and CPBayes are true based on the literature.

ASSET also displays the bizarre behavior of having fewer associations in the SNPs with the lowest (most extreme) p-values, while having many associations in SNPs with high p-values, for which we do not reject the null hypothesis of no association. The ABF, CPBayes, and binary effects approaches also ascribe non-null models to SNPs whose test statistic indicates that we should accept (or not reject) the null hypothesis. The two Bayes factor approaches have non-null models associated with SNPs where the $\log_{10}$(Bayes factor) or $\log_{10}$(mean Bayes factor) is negative. The binary effects model calculates high marginal probabilities of association for diseases with SNPs whose p-values are greater than 0.5. All of this is undesirable behavior—ideally SNPs for which we consider the null model to the be most likely model of association should not have any marginal associations calculated for them.

3.5.3 Method runtimes

A final point of comparison is how long each method takes to perform its calculation. We repeated each calculation for the full dataset 100 times in order to get a distribution of calculation times. These runtimes included things like storage of calculation outputs in variables for downstream analysis, as well as the calculations themselves. The results (range of runtimes as well as the median runtime—all reported in seconds) are reported in Table 3.5.

It comes as no surprise that the two fastest methods are the CPMA statistic and

Method	Range	Median
CPMA	$0.054 - 0.333$	0.055
ASSET	$360.464 - 497.891$	378.897
Binary effects	$82.761 - 86.582$	84.543
CPBayes	$453.717 - 580.392$	454.984
All-correlated ABF	$0.067 - 0.947$	0.073
Exhaustive ABF	$4.816 - 7.073$	5.801

Table 3.5: Range and median run times (in seconds) for 100 runs of the data from the Cotsapas *et al.* paper [52]. The timing included assignment of variables relevant to the analysis—such as priors for the ABF methods—as well as assigning results to matrices and vectors to allow further investigation. All analyses were conducted in R, except for binary effects, which was performed using the METASOFT package, written in Java [112].

the all-correlated ABF. Both methods are the calculation of a test statistic to reject (or, in the case of the ABF, accept) the null hypothesis or model, respectively, and neither approach attempts to answer the question of which model of association is the likeliest. Of the methods that do attempt to answer this question, the exhaustive ABF approach beats the others by being between 14.6-78.4 times faster. CPBayes is the slowest of the methods because it uses Gibbs sampling to calculate its Bayes factors and to calculate the marginal probabilities of association. Binary effects also uses a sampling method if the number of studies in the meta-analysis is large; however, with only seven studies in our dataset, it was possible to calculate the exact m values. If this were not the case, we would have had to use its sampling method to calculate these values, which we expect would increase the run time. ASSET calculates a test score for a number of (though not necessarily all) subset models, which might explain its runtime. The relative speed of the exhaustive ABF approach compared to these other methods highlights the advantage of the simplicity of the approximate Bayes factor—which is very easy to calculate—over a real one, or a more complicated test statistic.

3.6 Discussion

Despite the flaws in the dataset, we have demonstrated that our method is able to find true associations between the studies in the meta-analysis performed by Cotsapas *et al.* and we note that higher mean ABFs (where the mean is taken over all the possible

subset models) generally seem to correspond to SNPs with more true associations, though this can be confounded by markers that have very large effect sizes in a single study. We also note that because of the way this dataset was constructed—all SNPs had been found to be associated with at least one immune-mediated disease, even if it was not one of the diseases in the meta-analysis—it was almost inevitable that our follow-up analyses would find associations with something.

Our prior assumes no correlations between the effect sizes. While this meant that our approach was agnostic to the direction of effect, and therefore did not lose power when studies had opposite directions of effects, we probably did lose power to detect effects overall. Because autoimmune diseases are known to share genetic risk factors [52, 158, 162], it is reasonable to use a prior that encodes a belief in correlated effect sizes between the studies in the meta-analysis. The fact that we did not means that we potentially missed some true associations as a result.

Based on the further simulation work we have done, we expect the subset-exhaustive ABF approach to pick out the true underlying model only when true effect sizes are relatively large (Odds ratio ≥ 1.1). When the model selected by this approach is incorrect, it tends to be too conservative, selecting fewer associations than the the underlying model that generated the data. However, this does not stop our method from selecting models that include associations that are probably spurious, as we saw in the comparison to other methods. On the other hand, the most conservative method, the binary effects approach, failed to call associations that were probably true. This is the conflict between sensitivity (the ability to call true positives) and specificity (the ability not to call false positives). An ideal method is both highly sensitive and highly specific; however, since no method is ideal, researchers have to optimize the relative costs of calling false positives and the costs of calling false negatives.

Our results allowed us to investigate the question of whether a large effect in one trait predicts further effects in others. Our findings indicate that this could be true, though the most likely pattern of association at a SNP is determined by the priors chosen for the calculation. While there appeared to be some positive correlation between the largest effect size at a given SNP and likely the number of associations it had under the independent effects model ($\rho = 0, \sigma = 0.2$), this disappeared under

a fixed effects model ($\rho = 0.96, \sigma = 0.2$). If we want to answer this question more definitively, we will need to be certain that our priors are well chosen.

Another follow-up analysis was principal component analysis on the posterior probabilities of association on each disease at each SNP. This suggested possible relationships between the diseases in the meta-analysis. The results of this were ultimately inconclusive, as four of the five studies clustered very closely together. Furthermore, we had difficulty validating any of our findings in other studies. Ideally, a future analysis of this kind would include both ulcerative colitis and Crohn's disease as a positive control—these two diseases are known to share a large number of predisposing genetic loci, so we would have more confidence that our analysis was correct if we saw these two diseases cluster together. Ideally there would also be a non-autoimmune phenotype in the meta-analysis to act as the negative control, which we would expect to see cluster far away from the others.

We note that our analyses have taken place over a small set of markers that are scattered throughout the genome. As such, we do not necessarily expect the SNPs highlighted by our analyses to be causal for the diseases for which they show an association; rather, we see this as an analysis that has highlighted genomic regions that show cross-trait associations.

Due to the speed of the calculation of both the ABF and CPMA methods, it was feasible to compare the power of these two approaches across a wide range of false positive rates. An obvious next step would be to apply the other methods to our simulated datasets and perform further comparisons of power, as well as sensitivity and specificity analysis for model selection. It would also be interesting to test the calibration of the marginal probabilities of associations that we obtain from CPBayes and the binary effects approaches, and to test their power to detect effects under various circumstances, like we did for the ABF in Chapter 2.

An advantage that the ABF, binary effects, and CPBayes approaches all have over the CPMA and ASSET methods is the ability to provide marginal probabilities of association between each study and each SNP. This provides extra information that can be used to determine whether a putative association is real. In general, a high marginal probability suggests a true association while a lower one ($p_{marginal} \approx 0.5$ or 0.6) suggests that outside evidence—for instance from other studies or from functional

analysis—might be needed to corroborate the association before we believe that it is true.

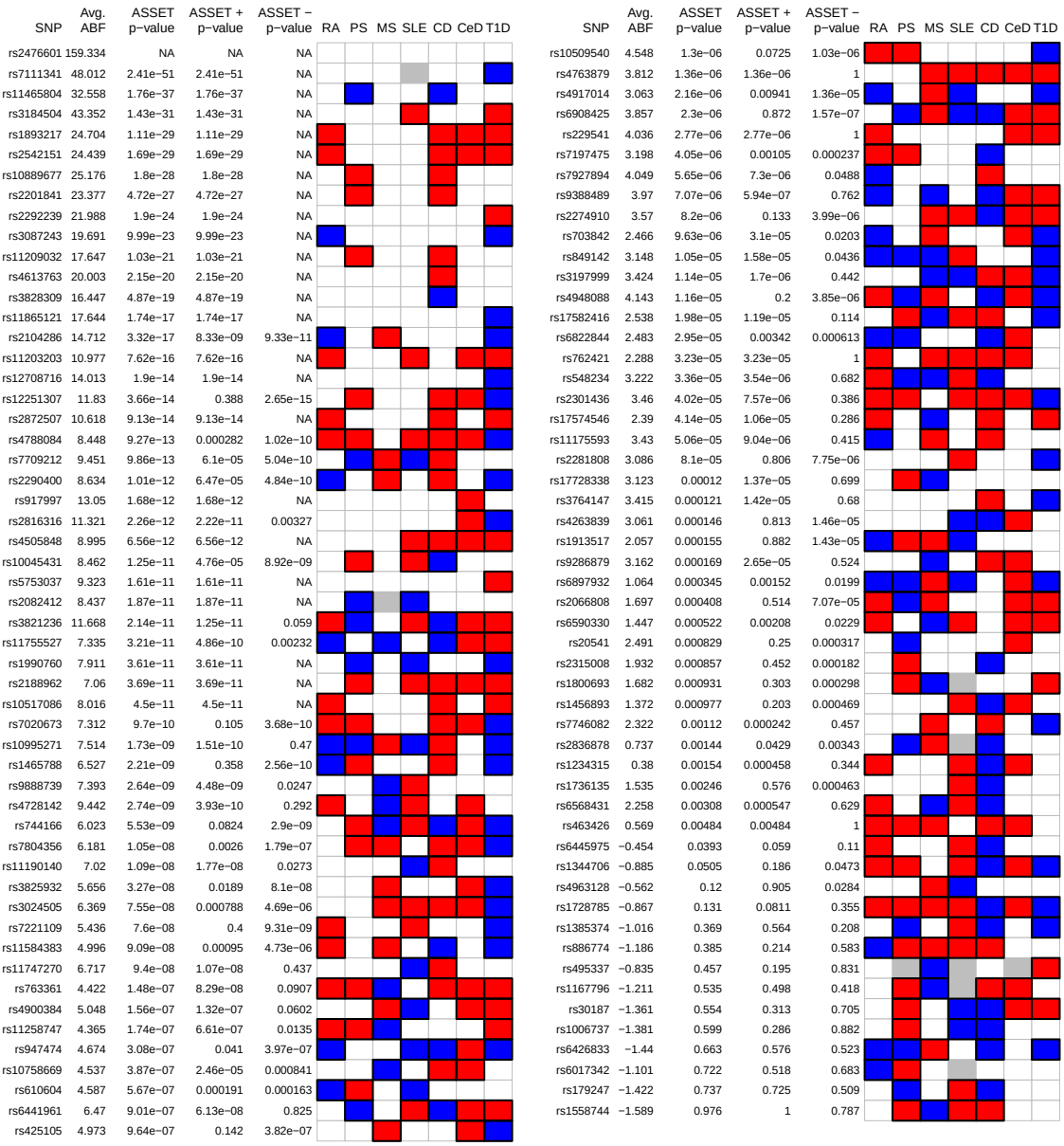
Finally, while ASSET, binary effects, and CPBayes may be able to account for covariance in effects due to shared samples or population structure, our method is unique in its ability to incorporate prior information on correlation between effects sizes due to true shared genetic effects. Additionally, our method allows us to place priors on particular patterns of association. While it might be theoretically possible to apply these priors to some of these other methods, none of the current implementations allow for it without sacrificing the ability to calculate marginal probabilities.

Of the methods we have investigated, including our own, there is no single approach that is universally better than all the others. In any genetic study, researchers have to balance the desire to discover novel associations with the desire to avoid calling false associations. Similarly, the flexibility of a Bayesian approach such as ours may be attractive to those who have concrete ideas about the priors that will be appropriate for their data, while being off-putting to researchers who do not and who may be concerned about loss of power due to inappropriately chosen priors. As ever, it is up researchers to decide what analysis is most appropriate for their data and their researh objectives.

Appendix

3.A Plots of subset associations at each SNP using different methods

3.A.1 ASSET



SNP	Avg. ABF	ASSET p−value	ASSET + p−value	ASSET − p−value
rs2476601	159.334	NA	NA	NA
rs7111341	48.012	2.41e−51	2.41e−51	NA
rs11465804	32.558	1.76e−37	1.76e−37	NA
rs3184504	43.352	1.43e−31	1.43e−31	NA
rs1893217	24.704	1.11e−29	1.11e−29	NA
rs2542151	24.439	1.69e−29	1.69e−29	NA
rs10889677	25.176	1.8e−28	1.8e−28	NA
rs2201841	23.377	4.72e−27	4.72e−27	NA
rs2292239	21.988	1.9e−24	1.9e−24	NA
rs3087243	19.691	9.99e−23	9.99e−23	NA
rs11209032	17.647	1.03e−21	1.03e−21	NA
rs4613763	20.003	2.15e−20	2.15e−20	NA
rs3828309	16.447	4.87e−19	4.87e−19	NA
rs11865121	17.644	1.74e−17	1.74e−17	NA
rs2104286	14.712	3.32e−17	8.33e−09	9.33e−11
rs11203203	10.977	7.62e−16	7.62e−16	NA
rs12708716	14.013	1.9e−14	1.9e−14	NA
rs12251307	11.83	3.66e−14	0.388	2.65e−15
rs2872507	10.618	9.13e−14	9.13e−14	NA
rs4788084	8.448	9.27e−13	0.000282	1.02e−10
rs7709212	9.451	9.86e−13	6.1e−05	5.04e−10
rs2290400	8.634	1.01e−12	6.47e−05	4.84e−10
rs917997	13.05	1.68e−12	1.68e−12	NA
rs2816316	11.321	2.26e−12	2.22e−11	0.00327
rs4505848	8.995	6.56e−12	6.56e−12	NA
rs10045431	8.462	1.25e−11	4.76e−05	8.92e−09
rs5753037	9.323	1.61e−11	1.61e−11	NA
rs2082412	8.437	1.87e−11	1.87e−11	NA
rs3821236	11.668	2.14e−11	1.25e−11	0.059
rs11755527	7.335	3.21e−11	4.86e−10	0.00232
rs1990760	7.911	3.61e−11	3.61e−11	NA
rs2188962	7.06	3.69e−11	3.69e−11	NA
rs10517086	8.016	4.5e−11	4.5e−11	NA
rs7020673	7.312	9.7e−10	0.105	3.68e−10
rs10995271	7.514	1.73e−09	1.51e−10	0.47
rs1465788	6.527	2.21e−09	0.358	2.56e−10
rs9888739	7.393	2.64e−09	4.48e−09	0.0247
rs4728142	9.442	2.74e−09	3.93e−10	0.292
rs744166	6.023	5.53e−09	0.0824	2.9e−09
rs7804356	6.181	1.05e−08	0.0026	1.79e−07
rs11190140	7.02	1.09e−08	1.77e−08	0.0273
rs3825932	5.656	3.27e−08	0.0189	8.1e−08
rs3024505	6.369	7.55e−08	0.000788	4.69e−06
rs7221109	5.436	7.6e−08	0.4	9.31e−09
rs11584383	4.996	9.09e−08	0.00095	4.73e−06
rs11747270	6.717	9.4e−08	1.07e−08	0.437
rs763361	4.422	1.48e−07	8.29e−08	0.0907
rs4900384	5.048	1.56e−07	1.32e−07	0.0602
rs11258747	4.365	1.74e−07	6.61e−07	0.0135
rs947474	4.674	3.08e−07	0.041	3.97e−07
rs10758669	4.537	3.87e−07	2.46e−05	0.000841
rs610604	4.587	5.67e−07	0.000191	0.000163
rs6441961	6.47	9.01e−07	6.13e−08	0.825
rs425105	4.973	9.64e−07	0.142	3.82e−07

SNP	Avg. ABF	ASSET p−value	ASSET + p−value	ASSET − p−value
rs10509540	4.548	1.3e−06	0.0725	1.03e−06
rs4763879	3.812	1.36e−06	1.36e−06	1
rs4917014	3.063	2.16e−06	0.00941	1.36e−05
rs6908425	3.857	2.3e−06	0.872	1.57e−07
rs229541	4.036	2.77e−06	2.77e−06	1
rs7197475	3.198	4.05e−06	0.00105	0.000237
rs7927894	4.049	5.65e−06	7.3e−06	0.0488
rs9388489	3.97	7.07e−06	5.94e−07	0.762
rs2274910	3.57	8.2e−06	0.133	3.99e−06
rs703842	2.466	9.63e−06	3.1e−05	0.0203
rs849142	3.148	1.05e−05	1.58e−05	0.0436
rs3197999	3.424	1.14e−05	1.7e−06	0.442
rs4948088	4.143	1.16e−05	0.2	3.85e−06
rs17582416	2.538	1.98e−05	1.19e−05	0.114
rs6822844	2.483	2.95e−05	0.00342	0.000613
rs762421	2.288	3.23e−05	3.23e−05	1
rs548234	3.222	3.36e−05	3.54e−06	0.682
rs2301436	3.46	4.02e−05	7.57e−06	0.386
rs17574546	2.39	4.14e−05	1.06e−05	0.286
rs11175593	3.43	5.06e−05	9.04e−06	0.415
rs2281808	3.086	8.1e−05	0.806	7.75e−06
rs17728338	3.123	0.00012	1.37e−05	0.699
rs3764147	3.415	0.000121	1.42e−05	0.68
rs4263839	3.061	0.000146	0.813	1.46e−05
rs1913517	2.057	0.000155	0.882	1.43e−05
rs9286879	3.162	0.000169	2.65e−05	0.524
rs6897932	1.064	0.000345	0.00152	0.0199
rs2066808	1.697	0.000408	0.514	7.07e−05
rs6590330	1.447	0.000522	0.00208	0.0229
rs20541	2.491	0.000829	0.25	0.000317
rs2315008	1.932	0.000857	0.452	0.000182
rs1800693	1.682	0.000931	0.303	0.000298
rs1456893	1.372	0.000977	0.203	0.000469
rs7746082	2.322	0.00112	0.000242	0.457
rs2836878	0.737	0.00144	0.0429	0.00343
rs1234315	0.38	0.00154	0.000458	0.344
rs1736135	1.535	0.00246	0.576	0.000463
rs6568431	2.258	0.00308	0.000547	0.629
rs463426	0.569	0.00484	0.00484	1
rs6445975	−0.454	0.0393	0.059	0.11
rs1344706	−0.885	0.0505	0.186	0.0473
rs4963128	−0.562	0.12	0.905	0.0284
rs1728785	−0.867	0.131	0.0811	0.355
rs1385374	−1.016	0.369	0.564	0.208
rs886774	−1.186	0.385	0.214	0.583
rs495337	−0.835	0.457	0.195	0.831
rs1167796	−1.211	0.535	0.498	0.418
rs30187	−1.361	0.554	0.313	0.705
rs1006737	−1.381	0.599	0.286	0.882
rs6426833	−1.44	0.663	0.576	0.523
rs6017342	−1.101	0.722	0.518	0.683
rs179247	−1.422	0.737	0.725	0.509
rs1558744	−1.589	0.976	1	0.787

Figure 3.A.1: The table shows the SNP rsid, the $\log_{10}$ (mean approximate Bayes factor) (our method), and the p-values for both sides of the test. The colored boxes outlined in black show for each SNP which subset of diseases were found by ASSET to be associated. Red boxes indicate the diseases whose estimated effect sizes were positive, and blue boxes indicate diseases whose estimated effect sizes were negative. The SNPs are listed in the order of increasing p-value (from the meta-analyzed p-values calculated by the method for both the $+$ and $-$ subsets).

3.A.2 Binary effects

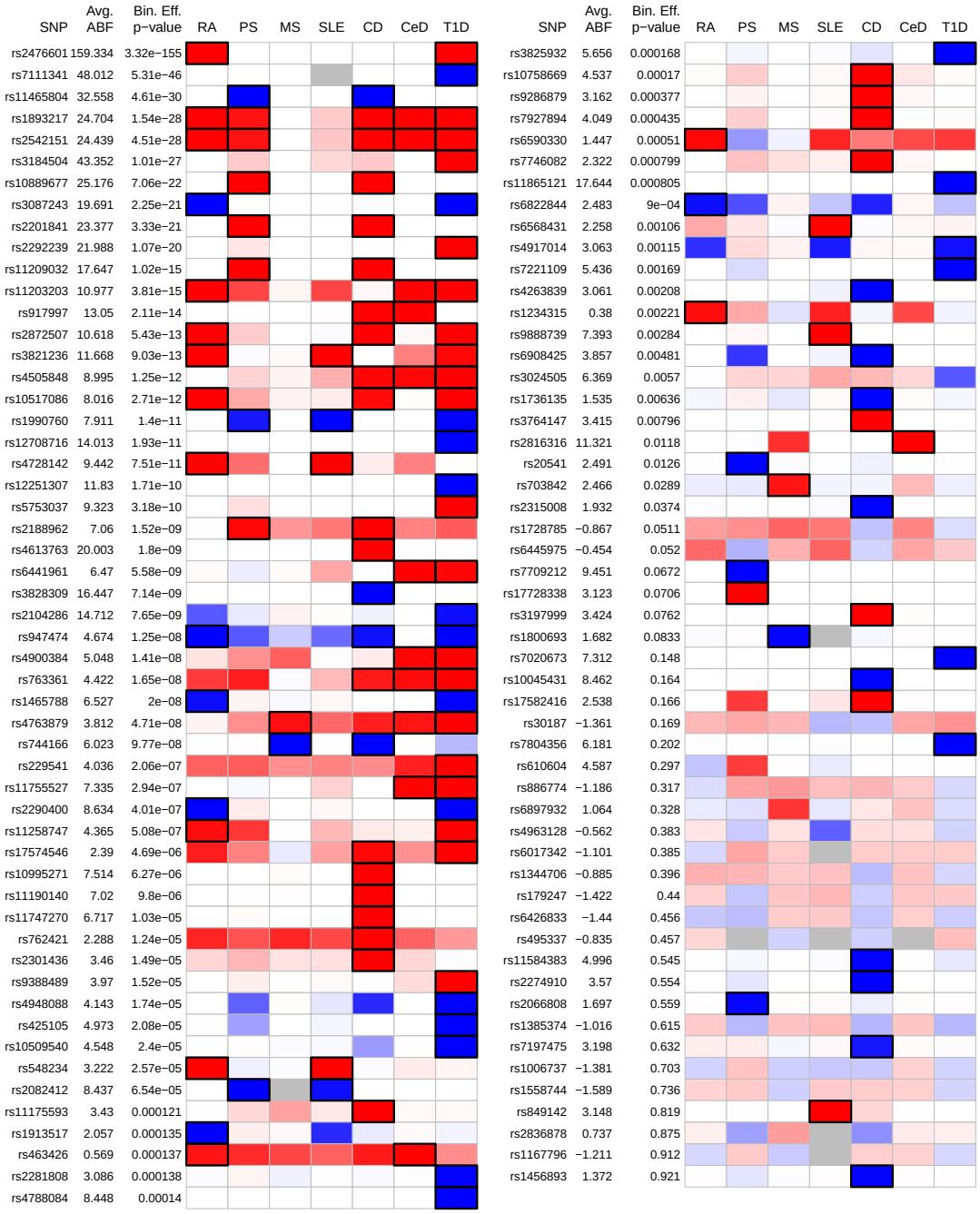


Figure 3.A.2: The table shows the SNP rsid, the $\log_{10}$(mean approximate Bayes factor) (our method), and the p-values for binary effects analysis. The colored boxes outlined in black show for each SNP which subset of diseases were found by this method to be associated. Red boxes indicate the diseases whose estimated effect sizes were positive, and blue boxes indicate diseases whose estimated effect sizes were negative. The intensity of the color indicates an m-value (marginal value of association), with darker colors showing that this value is closer to 1 than paler colors. The SNPs are listed in the order of increasing p-value.

3.A.3 CPBayes

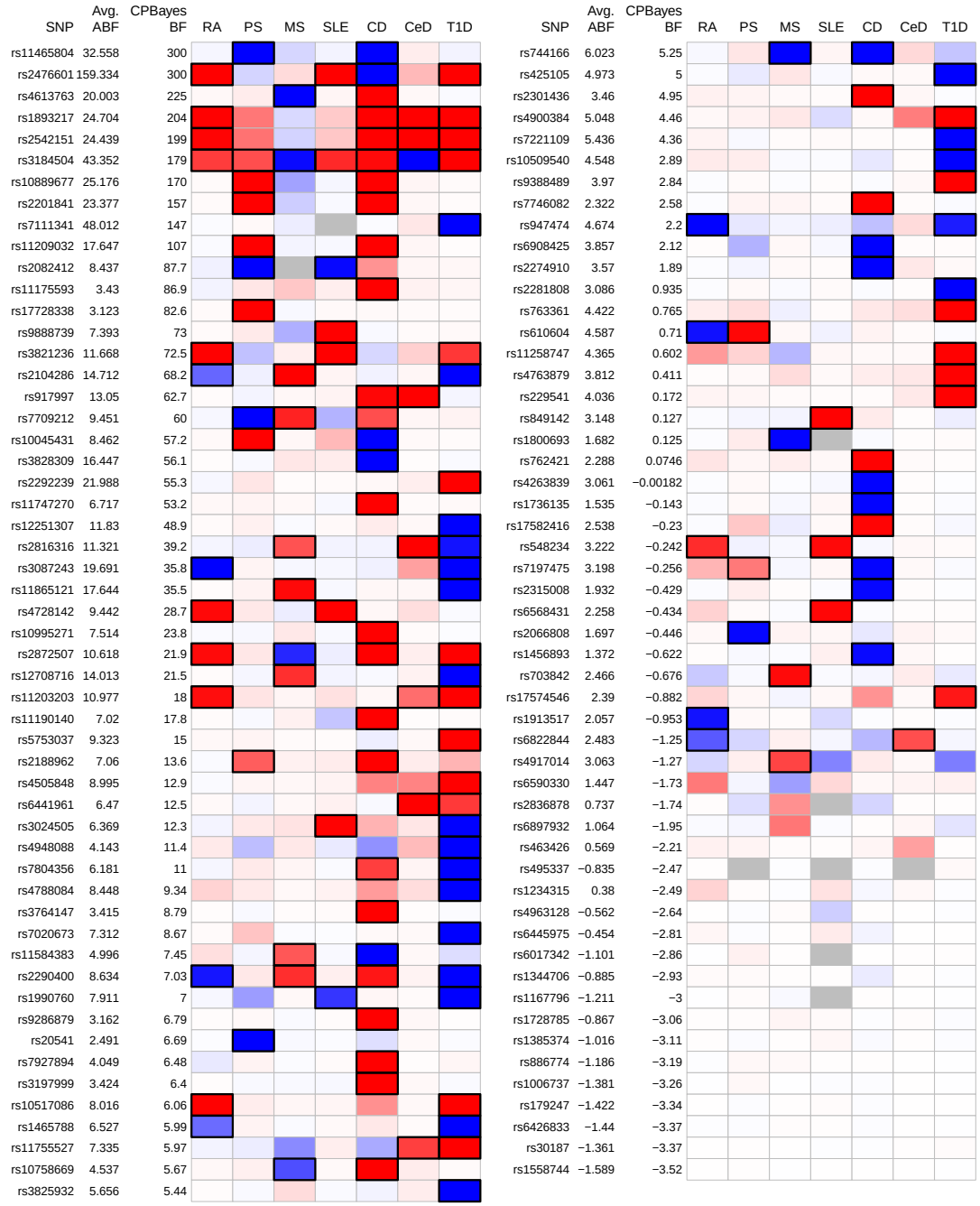


Figure 3.A.3: The table shows the SNP rsid, the $\log_{10}$(mean approximate Bayes factor) (our method), and the $\log_{10}$(Bayes factor) calculated by CPBayes. The colored boxes outlined in black show for each SNP which subset of diseases were found by this method to be associated. Red boxes indicate the diseases whose estimated effect sizes were positive, and blue boxes indicate diseases whose estimated effect sizes were negative. The intensity of the color indicates an marginal probability of association, with darker colors showing that this value is closer to 1 than paler colors. The SNPs are listed in the order of increasing p-value.

3.B Visualization of rankings of different methods

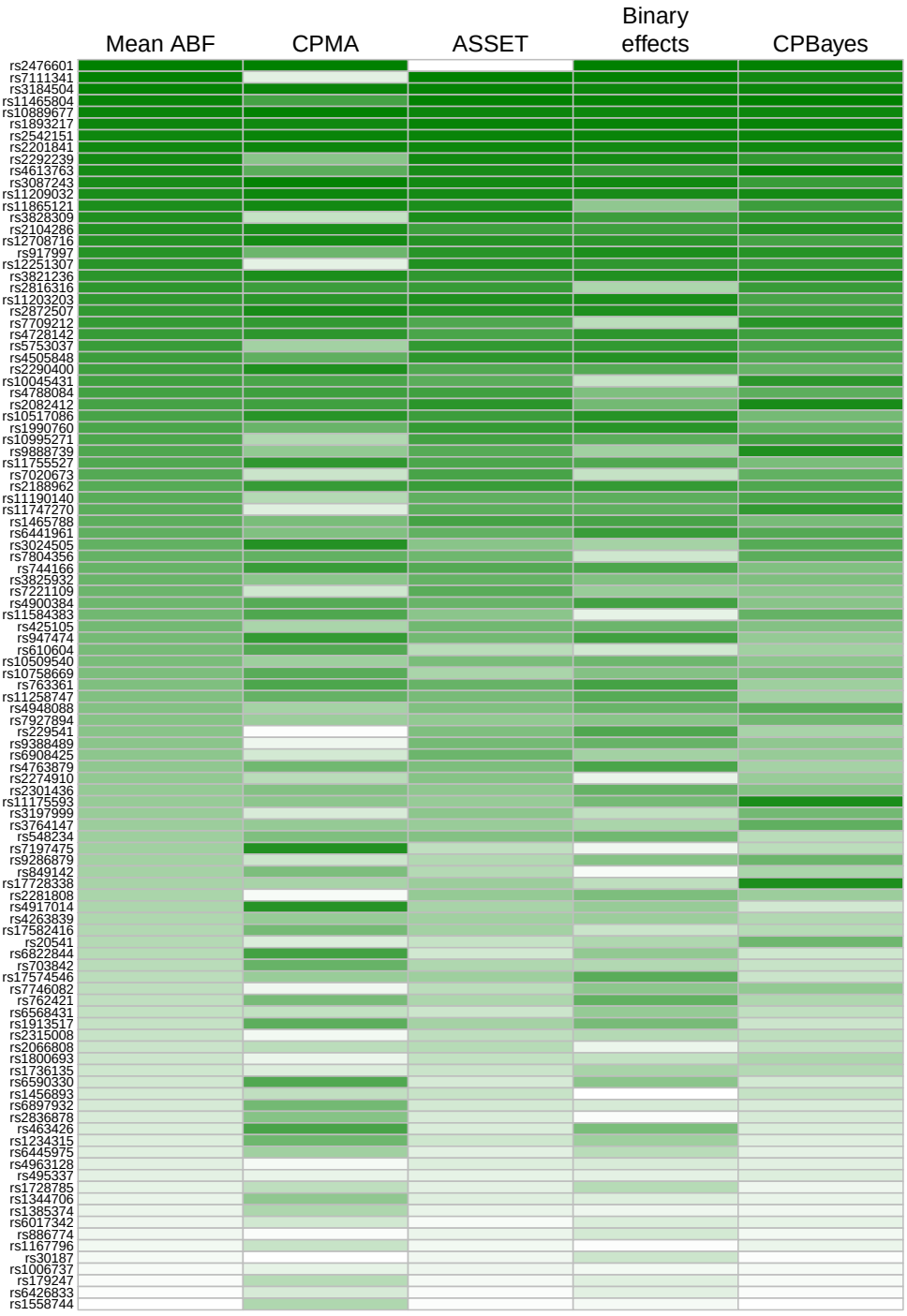


Figure 3.B.1: Comparison of how the SNPs were ranked in each of the analyses (parameters used are at the top of the plot. Darker greens indicate higher ranks, while paler colors correspond to lower ranks. The table is ordered by the ranks in the first column, where the ABF analysis was conducted with $\rho = 0$ and $\sigma = 0.2$.

Chapter 4

Analysis Across the Wellcome Trust Case Control Consortium 2

4.1 Introduction

In the previous chapter, we applied our method as well as others to a small dataset. In doing so, we highlighted several problems with that dataset: our uncertainty in the effect allele for a given marker across all studies, the shared controls that could not be satisfactorily accounted for, and the discrepancy between the reported z-scores and the p-values. In this chapter, we apply our method to the Wellcome Trust Case Control Consortium 2 (WTCCC2), a larger dataset in terms of both the number of constituent studies and the number of SNPs.

As a follow-up to the WTCCC1 (see Section 1.1 in Chapter 1), The Wellcome Trust Case Control Consortium 2 was created to perform 15 genome-wide association studies, of which all but one have been published. Seven of these studies took advantage of a pool of controls—as before, these samples came from the 1958 British Birth Cohort and from the UK Blood Services—though this time with larger study sizes than those in WTCCC1. The samples were genotyped on one of three chips: the Affymetrix 6.0 SNP array (cases and controls) or the Illumina 660W-Quad array (cases) and the Illumina Human 1.2M-Duo array (controls). Additionally, some studies were imputed to increase the number of SNPs in the GWAS. Table 4.1 summarizes the study names, chip and imputation information, and references. Table 4.2 goes into further detail about the samples used in each study.

Table 4.1: Summary of the data available from the Wellcome Trust Case Control Consortium Study 2. We list the phenotypes, the short-hand we use for them ("Abbr."), the genotyping chips (cases/controls, if different chips were used for each), the reference panel to which they were imputed (if applicable), and the reference number of the published paper.

Trait	**Abbr.**	**Genotyping chip (case/control)**	**Reference panel**	**Reference**
Ulcerative colitis	UC	Affymetrix 6.0 SNP array	Not imputed	[160]
Psoriasis	PS	Illumina 660W-Quad array/Illumina Human 1.2M-Duo array	Not imputed	[62]
Parkinson's disease	PD	Illumina 660W-Quad array/Illumina Human 1.2M-Duo array	Not imputed	[163]
Glycemic response to metformin in type 2 diabetes	PR	Affymetrix 6.0 SNP array	Not imputed	[164]
Ankylosing spondylitis	AS	Illumina 660W-Quad array/Illumina Human 1.2M-Duo array	HapMap2 CEU, HapMap3 CEU	[61]

(Continued on next page)

(continued from previous page)

Trait	Abbr.	Genotyping chip (case/control)	Reference panel	Reference
Multiple sclerosis	MS	Illumina 660W-Quad array/Illumina Human 1.2M-Duo array	Not imputed	[140]
Ischemic stroke	IS	Illumina 660W-Quad array/Illumina Human 1.2M-Duo array	Not imputed	[125]
Barrett's esophagus	BO	Illumina 660W-Quad array/Illumina Human 1.2M-Duo array	Not imputed	[165]
Schizophrenia, schizoaffective disorder, and schizophreniform disorder	SP	Affymetrix 6.0 SNP array	Not imputed	[166]
Visceral leishmaniasis	VL	Illumina 660W-Quad array	Not imputed	[167]

(Continued on next page)

(continued from previous page)

Trait	Abbr.	Genotyping chip (case/control)	Reference panel	Reference
Intra-ocular pressure (glaucoma endophenotype)	GL	Illumina 660W-Quad array	1000 Genomes Pilot CEU haplotypes	[99]
Psychosis	PE	Affymetrix 6.0 SNP array	Not imputed	[168]
Reading and mathematical performance at age 12	RM	Affymetrix 6.0 SNP array	HapMap2 CEU, HapMap3 CEU	[169]
Bacteremia susceptibility	BS	Affymetrix 6.0 SNP array	1000 Genomes Phase 1	[170]
Pre-eclampsia	PA	Affymetrix 6.0 SNP array	1000 Genomes Phase 3	Not published

Across all WTCCC2 studies, some, such as the multiple sclerosis study, identified many loci and some, such as the visceral leishmaniasis studies, identified very few or no loci. This may be due to differing sample sizes, which affect power. The MS study was very large, with 27,148 individuals, while the VL study had a total sample size of 4,048—a little over a seventh as large. Diseases with heterogeneous etiologies, such as stroke or schizophrenia may also suffer from reduced power due to underlying genetic causes differing between undetected subtypes. Finally some traits, such as reading and mathematical abilities or psychosis, may simply not be influenced by any loci of large effect, so even a study of moderate size may be underpowered to detect true effects. Meanwhile, others—especially the studies of quantitative traits, such as intraocular pressure—may have loci whose effects are large enough to be discovered even with modest sample sizes.

Though there are now many databases of GWAS summary statistics available—for instance anthropometric traits from the GIANT consortium [171] or psychiatric traits from the Psychiatric GWAS Consortium [172], we chose the WTCCC2 data because the analysis had been performed at the Wellcome Trust Centre for Human Genetics, which facilitated the collection of the data and ensured that we had the summary statistics we needed; summary statistic information from other sources does not always include effect size estimates and standard errors.

Table 4.2: Description of each study, with the total numbers of cases and controls, number of controls from the WTCCC2 pool of shared controls, the population of the samples, and any miscellaneous comments. Separate cohorts are appended. For instance, the visceral leishmaniasis cohort from India is labelled VL India. For studies with multiple cohorts, we give the total numbers across all studies. These rows are italicized. For studies investigating a quantitative trait, the total number of participants is listed in the cases column.

			WTCCC2		
Study	**Cases**	**Controls**	**Controls**	**Population**	**Notes**
UC	2361	5417	5417	European	
PS	2178	5175	5175	European	
PD	1705	5175	5175	European	
PR	1024			European	
AS	1787	4800	4800	European	
MS UK	1854	5175	5175	European	
MS nonUK	7918	12201	0	European	
MS	*9772*	*17376*	*5175*	*European*	
IS TOAST 1	844	5972	5175	European	Large vessel stroke cohort
IS TOAST 2	580	5972	5175	European	Small vessel stroke cohort
IS TOAST 3	790	5972	5175	European	Cardioembolic stroke cohort
IS	*3548*	*5972*	*5175*	*European*	*Includes samples not in TOAST cohorts.*

(Continued on next page)

(continued from previous page)

			WTCCC2		
Study	**Cases**	**Controls**	**Controls**	**Population**	**Notes**
BO	1852	5172	5172	European	
SP	1606	1794	0	European	
VL India	989	1089	0	Indian	
VL Brazil	357	1613	0	Brazilian	Cases and controls drawn from 308 families.
VL	*1346*	*2702*	*0*	*Indian and Brazilian*	
GL	2302			European	
PE	1239	3596	0	European	857 of the controls are unaffected relatives of cases.
RM	2794			European	
BS pneumococcal	429	2677	0	Kenyan	
BS overall	1536	2677	0	Kenyan	
PA	1924	2451	0	Colombian	

The diversity of data in the WTCCC2 provides numerous challenges to meta-analysis: these data come from different ethnic groups, two different genotyping protocols were used, some traits are binary and some are quantitative, and some traits (for instance, autoimmune diseases) are known to share genetic risk factors while others are not known or expected to have any etiological relationship. We note that autoimmune diseases studied in European populations were well represented in both the WTCCC1 and the WTCCC2. Diseases of this class are known to have complex genetic architectures and to share genetic variants [52, 53]. These qualities make them prime candidates for both genome-wide association analyses and for subsequent meta-analysis in the hopes of discovering signals that have been previously overlooked due to their failure to achieve genome-wide significance in a single study [118] (see Section 1.4.4 in Chapter 1 for a discussion of this concept).

One of our aims in performing a meta-analysis on these data is to demonstrate that our method is applicable to genome-wide data and across a diverse group of studies. We also investigate whether or not our method can leverage the relationships between true genetic effects associated with these diseases to increase power to discover signals of association that were missed in the original studies. Furthermore, we are interested in determining whether this leveraging can also uncover signals in a study like pre-eclampsia, where no genome-wide significant associations were found. Finally, we hope to discover novel loci common to multiple phenotypes. Such loci provide clues about the shared etiology of these complex phenotypes. Our method potentially provides a principled way of using the observation that some phenotypes—say, autoimmune diseases—may share genetic risk factors, and thus if one trait is associated at a SNP, we may be willing to accept that another, related trait is also associated at that same SNP, even if the evidence of association for the second trait does not achieve the same level of statistical significance as the first. While some of our findings may be supported by subsequent studies of these traits (when such studies have been conducted), this analysis is necessarily an exploration of the feasibility and utility of our approach.

4.2 Data and preprocessing

4.2.1 GWAS data

As noted earlier, because the WTCCC2 GWAS analyses were performed in the Wellcome Trust Centre for Human Genetics, it was relatively easy to obtain the required summary statistics. Chris Spencer created standardized comma-separated value (.csv) files containing for each SNP its chromosome and base pair position (on NCBI human genome build 36), rsid, alleles (coded "allele A" and "allele B"), frequency of the B allele, the estimated effect size of the SNP on the trait, the standard error of the estimated effect size, and the p-value. Associated with each file was a list of SNP inclusions or exclusions from the original studies. These were applied to the .csv files to ensure that only the SNPs that passed quality control in the original studies were included in the meta-analysis.

The first step was to check the internal consistency between the effect size estimates and their standard errors and the reported p-values. We saw in Figure 3.1 of Section 3.2 in Chapter 3, the discrepencies between the reported z scores and the reported p-values in about one third of the data. In Figure 4.1, we see that the p-values calculated from the effect size estimates and standard errors are positively correlated ($\rho \approx 0.99997$) with those reported in the summary statistic tables and that while these values do not always match perfectly, the results do appear to be consistent with one another. This suggests that there have not been major errors in the tables of summary statistics for each study and that the effect size estimates and p-values are generally being reported correctly for each cohort across all SNPs.

Because the data are so diverse and specifically contain both quantitative and binary traits, we also checked the distributions of effect size estimates ($\hat{\beta}$s) across all studies. We show the distribution of the $\hat{\beta}$ for each study in Figure 4.2, coloring the curves based on whether the trait was quantitative (blue) or binary (red). We note that these distributions all appear to be normal about a mean of 0, in line with the assumptions of the approximate Bayes factor (see 1.5.2 in Chapter 1 and 2.2.1 in Chapter 2). We normalize the raw $\hat{\beta}$s to have an area under the curve of 1 to show these distributions as probability density functions.

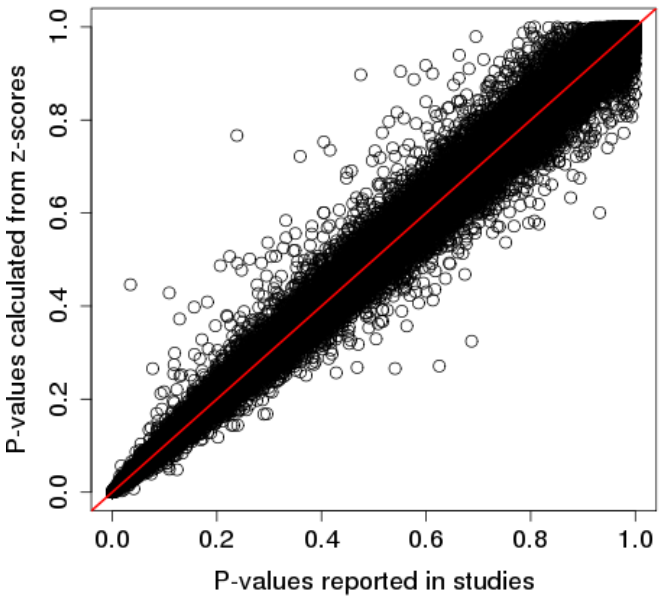


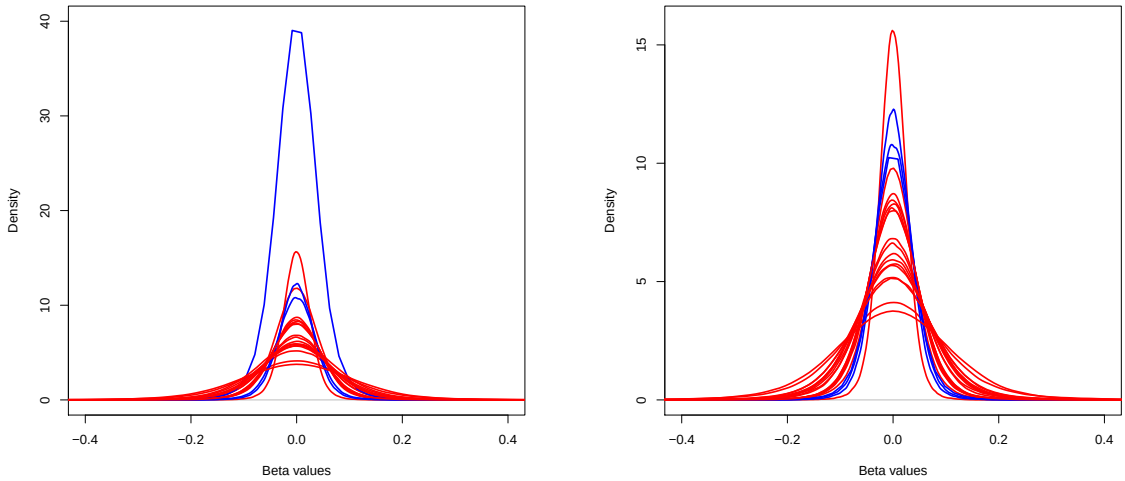
Figure 4.1: Comparison of the p-values calculated from z scores (y-axis) and the reported p-values (x-axis) in the WTCCC2 data. The red line is the line $y = x$.

4.2.2 Strand alignment

To avoid one of the ambiguities that plagued the Cotsapas dataset in the previous chapter, we created a pipeline to align the SNPs in each study to the forward strand, thus ensuring that the A and B alleles of each SNP were the same across all studies. We describe the process below.

We first compared the rsids and positions of each SNP to files that map probes from the SNP chips used to genotype samples in the studies to the human reference sequence, available online at `http://www.well.ox.ac.uk/~wrayner/strand` [173]. These files also include strand information. For each study, there were a small number of anomolous SNPs whose positions were not found in the file for their genotyping chip and were thus removed.

The next step was to align all SNPs to the forward (+) strand. This meant that for SNPs aligned to the backward (-) strand, we swapped Allele A and Allele B around, subtracted the stated allele frequency from 1 to reflect the frequency of the new B allele, and changed the sign of the estimated effect size. SNPs that were already aligned to the forward strand were left unchanged. This step prevented the removal of SNPs with complementary alleles (see Section 3.2 of Chapter 3 for a definition of



(a) Distribution of the raw $\hat{\beta}$s for each WTCCC2 study.

(b) Distribution of the $\hat{\beta}$s for each WTCCC2 study normalized to have an area of 1 under each curve.

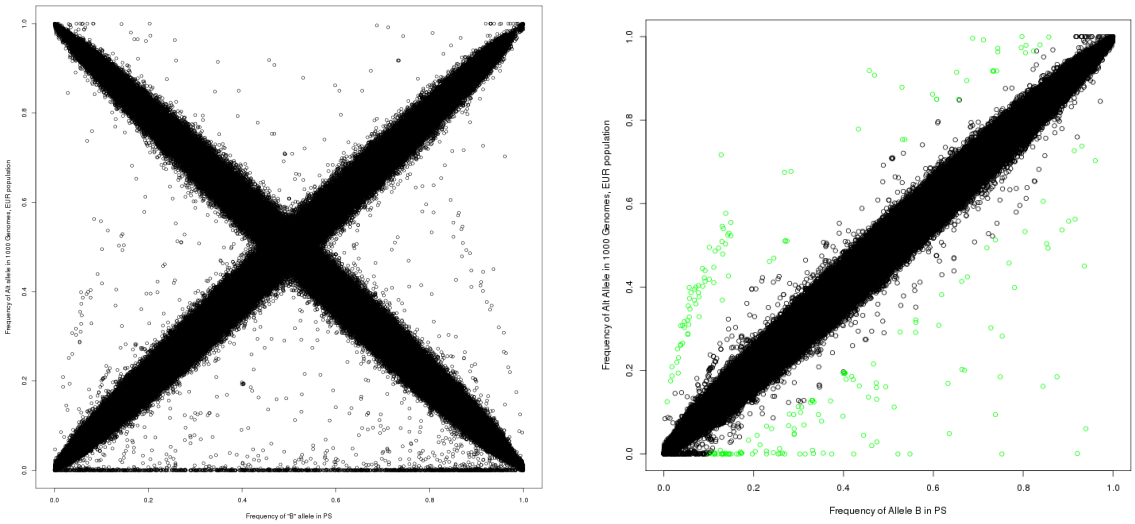
Figure 4.2: Distribution of the $\hat{\beta}$s for all WTCCC2 studies. Red curves denote binary traits, while blue curves denote quantitative ones.

these) from the dataset.

We then compared each file to the 1000 Genomes Phase 3 data to ensure that our strand alignment was correct. The SNPs in 1000 Genomes data are all aligned to the forward strand, meaning that the "ref" allele was expected to correspond to our Allele A and the "alt" was expected to correspond to our Allele B. SNPs whose alleles did not match those in 1000 Genomes were removed from the data. Finally, for the SNPs in our datasets, we calculated Wright's F_{ST} between our study samples and the corresponding 1000 Genomes population using both our and the 1000 Genomes allele frequency data. We removed any SNPs where $F_{ST} \geq 0.1$. This step was taken to remove SNPs where we were uncertain that the strand alignment had worked correctly. Allele frequency discrepencies suggest that there is a problem—perhaps the SNP has more than two alleles, or perhaps the SNP has been misidentified on the genotyping chip used in the WTCCC2 study. On the other hand, we risk removing SNPs whose effect sizes are so large that it alters the allele frequency in the sample, due to overrepresentation of cases in a study versus the general population.

As a final check, we plotted the allele frequencies of the SNPs listed in the original .csv files for each study, before any alignment work was done, to the alternate allele frequency in the corresponding 1000 Genomes population. We show the before

and after images for the psoriasis study (PS) below and report the number of SNPs at the start of the pipleine and at the end for each cohort in Table 4.1.



(a) Allele frequencies of the alt allele from the 1000 Genomes EUR population (y-axis) plotted against the allele frequencies of the B allele in the psoriasis study (x-axis), before alignment and filtering.

(b) Allele frequencies of the alt allele from the 1000 Genomes EUR population (y-axis) plotted against the allele frequencies of the B allele (x-axis) in the psoriasis study, after alignment and filtering. Green circles represent the SNPs that were filtered out due to $F_{ST} \geq 0.1$.

Figure 4.3: Comparison of allele frequencies in the psoriasis study (x-axes) and in 1000 Genomes EUR population (y-axes) before alignment and filtering (left) and after (right).

Table 4.1: Table showing for each cohort the number of SNPs at in the .csv files ("SNPs at start"), the number of SNPs after the strand alignment pipeline ("SNPs at end"), and the percent of the SNPs at the start that are left at the end of the pipeline. The IS TOAST cohorts had the same SNPs and allele frequencies, so they are listed as one cohort instead of three.

Cohort	SNPs at start	SNPs at end	% left
AS	2,078,665	2,059,149	99.06
BO	521,744	516,922	99.08
BS overall	12,785,625	847,650	6.63
BS pneumo.	12,781,759	849,543	6.65
GL	7,642,395	4,910,776	64.26
IS TOAST	495,896	492,156	99.25
MS UK	465,434	462,014	99.27

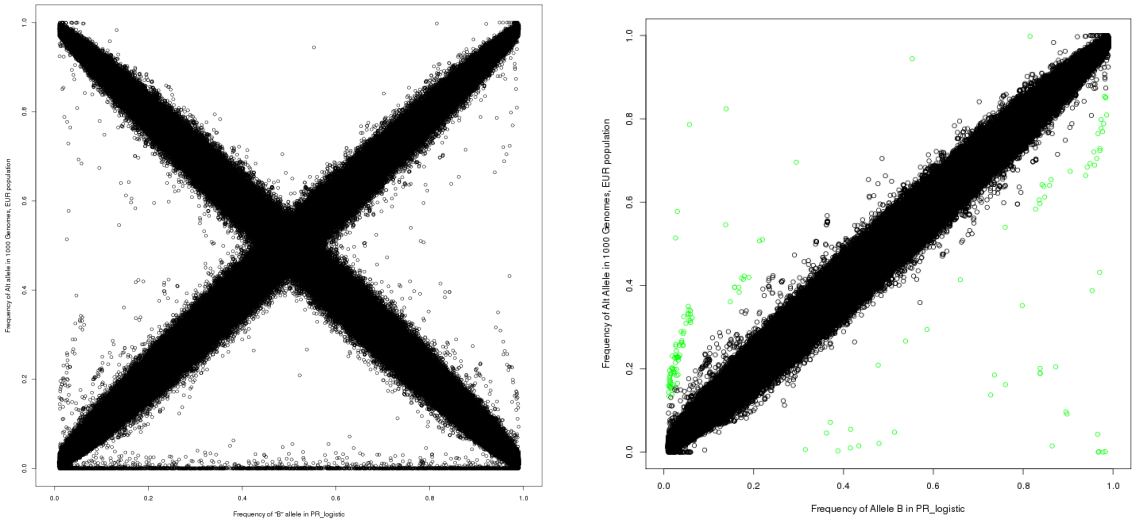
(Continued on next page)

(continued from previous page)

Cohort	SNPs at start	SNPs at end	% left
MS nonUK	465,434	462,016	99.27
PA	11,517,889	11,448,177	99.39
PE	695,193	689,216	99.14
PR	705,124	73,775	10.46
PS	523,067	518,225	99.07
RM reading	1,588,650	1,575,073	99.15
RM maths	1,588,284	1,574,713	99.15
SP	696,951	687,968	98.71
UC	757,931	747,756	98.66
VL Brazil	574,409	542,535	94.45
VL India	526,731	521,853	99.07

Inspection of these plots satisfied us that our pipeline had worked correctly: we knew that for every cohort, the alleles matched those in 1000 Genomes and in those the files that mapped SNP probes to the human reference genome. The plots showed us that the frequencies of Allele B in our studies were consistent with those of the alt alleles in the 1000 Genomes populations that corresponded to the populations from which our samples were drawn. The only trait for which this was not the case was PR (see Figure 4.4). When this dataset when through the alignment pipeline, 89.5% of the SNPs in the initial .csv file were removed, and among the remaining SNPs were some whose allele frequencies suggested that they had not been aligned correctly. Consequently, we excluded this study from our analyses. This left us with 20 individual cohorts across 14 studies.

We also note that there is an anomaly with some of the imputed cohorts—namely, GL and the two BS studies. Compared to the other studies, which retain over 90% of the SNPs from the initial data files, GL only retains 64.26% of its SNPs, and



(a) Allele frequencies of the alt allele from the 1000 Genomes EUR population plotted against the allele frequencies of the B allele in the psoriasis study, before alignment and filtering.

(b) Allele frequencies of the alt allele from the 1000 Genomes EUR population plotted against the allele frequencies of the B allele in the psoriasis study, after alignment and filtering. Green circles represent the SNPs that were filtered out due to $F_{ST} \geq 0.1$.

Figure 4.4: Comparison of allele frequencies in the study of metformin response and in 1000 Genomes EUR population before alignment and filtering (left) and after (right).

the BS cohorts only keep about 6.6% each, which is even worse than the proportion of SNPs removed from the PR analysis. It is only because the BS cohorts had over 12 million SNPs to start with that an acceptable number are left after the removal of over 93% of them. To determine whether any chromosome was overrepresented among the SNPs that were filtered out during the alignment process, we plot the proportion of SNPs on each chromosome (sequentially from left to right) in each cohort, before and after alignment in Figure 4.5.

We observe that some chromosomes were indeed overrepresented in the SNPs that were filtered out of the BS and GL cohorts. While the differences in the proportions of SNPs before and after alignment on each are not large in absolute terms—they are all less than 0.01—even a small deviation in absolute terms can represent a large proportional change if proportion of SNPs on a given chromosome was small to begin with. The most extreme example of this is the proportion of SNPs on chromosome 19 in the BS cohorts. This proportion shrinks nearly 25% after alignment. However, in absolute terms, the proportion of SNPs in the dataset before and after alignment that are on this chromosome drops from 0.018 to 0.013. While we would prefer the distri-

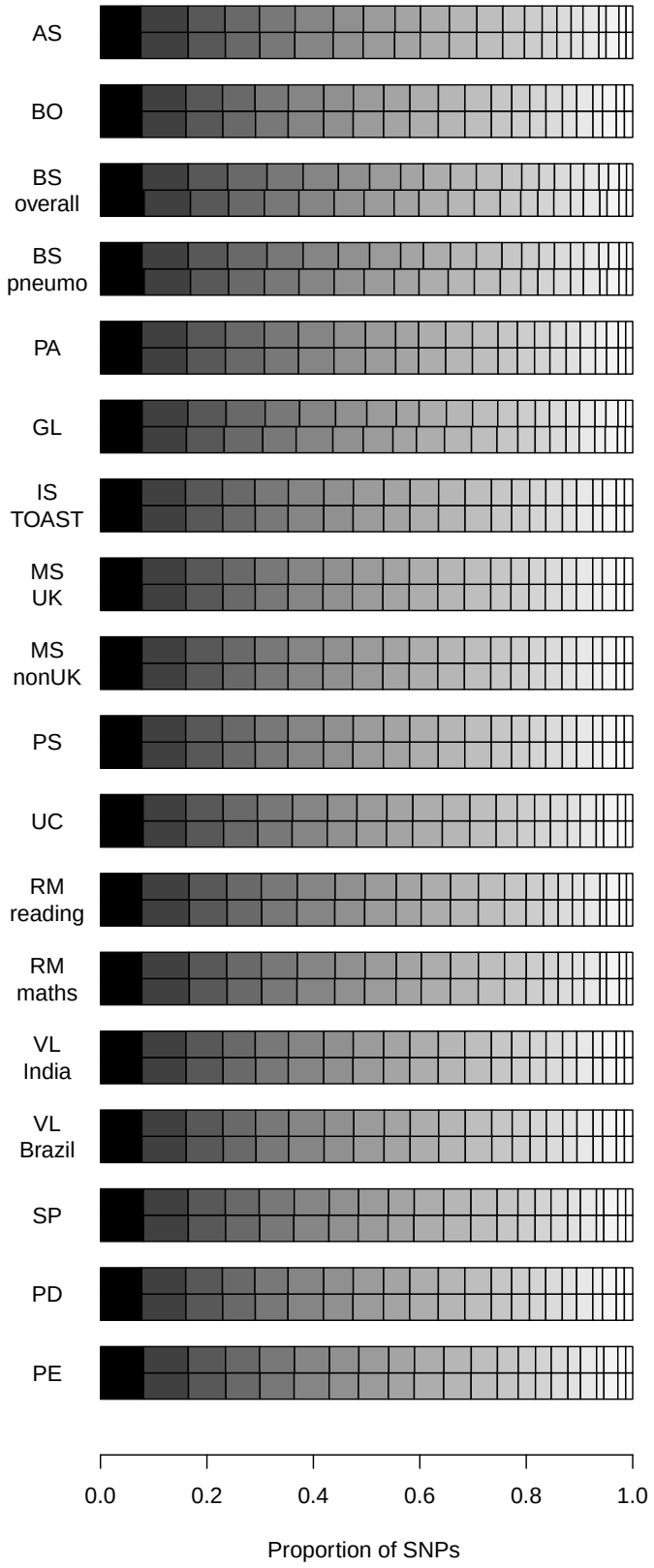


Figure 4.5: Comparison of the proportion of SNPs on each chromosome, before and after allele alignment and SNP filtering, for each WTCCC2 cohort. The bars show proportion of SNPs on each chromosome (1 is colored black and at the far left of each bar; 22 is colored white and at the far right of each bar) for each cohort, before alignment (top bar) and after (bottom bar).

bution of SNPs on each chromosome to be more consistent before and after alignment, as they are in the other cohorts, we have no choice but to accept this consequence of our alignment pipeline and be aware that this may affect our results.

After applying a final filter to remove any SNPs whose reported standard errors were less than or equal to zero, we were left with a set of 2,058,911 SNPs that had effect size estimates and associated standard errors in at least two of the cohorts. This was the dataset used in our subsequent analyses.

4.3 Analysis

Once we had the set of aligned SNPs, we could turn to choosing the parameters for our analyses, as discussed in Sections 2.2 and 2.2.1 of Chapter 2. In particular, as we can see in Table 4.2, nine of the WTCCC2 studies used the shared controls, the reading and mathematics traits were measured on the same set of individuals, and the BS pneumococcal cohort was a subset of the BS overall group. This sharing of samples across cohorts introduces cryptic correlation in the effect size estimates. In order to account for this, we needed to estimate a covariance matrix under the null. There are several possible approaches we could take—estimating the correlations from the data directly, using LD score regression [84, 85] to partition the correlation of summary statistics into those due to shared genetic effects and those due to confounding (see 1.4.3 in Chapter 1), or we could calculate the correlation based on the number of shared cases and controls using the formula from Zaykin *et al.* (see 2.4.2) [131].

We applied five different methods of estimation to our data and report the results in Table 4.1. The first three columns of results show the correlations calculated directly from the effect size estimates of the intersection of SNPs typed in both cohorts. In one instance, we excluded all of the SNPs in the HLA region as well as those which showed a genome-wide significant association in at least one of the cohorts. In another, we thinned each pair using Inthinnerator [174] to ensure that SNPs were all at least 0.25 centiMorgans apart and thus roughly independent of one another. For each pair of cohorts, this left a set of between 9,000 and 11,000 SNPs. We also calculated the pairwise correlations using all SNPs shared between each pair of cohorts, largely for the sake of comparison.

The fourth column of results shows the correlation calculated by LD score regression—this value is the intercept of the genetic covariance regression, which provides an estimate of the test statistic inflation due to confounding factors, such as shared samples [85, 84]. LD score regression requires reference panels to run; we did not have appropriate reference panels for the non-European cohorts. Additionally, due to low numbers (<200,000) of SNPs shared between studies—in particular between the UC cohort and the BO, IS, MS, and PS cohorts, we must treat the correlations calculated using this method with caution. Furthermore, for reasons that remain unclear, LD score regression could not calculate values for 17 pairs of cohorts (11 of these involved at least one of the BS cohorts or the GL cohort), so these are marked as "NA".

The final column shows the results obtained from using the analytic formula from Zaykin *et al.* This formula is based entirely on the number of cases and controls in each study and number of shared cases and controls between studies. Unlike the other methods, it does not use the $\hat{\beta}$s to estimate the correlations. Because this formula explicitly applies to case/control studies, we did not use it to estimate the cryptic correlation for cohort pairs where one or both of the traits was quantitative.

To be included in Table 4.1, we used the standard errors of the correlation estimates from each respective calculation, except for the analytic one, to determine whether or not the estimate was significantly different from zero. We report the correlations for pairs where at least one of the methods returned an estimate of at least 0.05 and where that estimate was statistically significant at the Bonferroni-corrected 0.01 level (5.26×10^{-5}). In practice, because of the large number of SNPs used in each calculation, all of the estimates above 0.05 were statistically significant at a much more stringent threshold than this ($p \ll 10^{-10}$). For the pairs of cohorts where the LD score regression calculation was able to run, we also calculated the genetic correlation. None of these correlations were statistically significant. In fact, the only correlation that achieved even nominal significance ($p \approx 0.019$) was between the schizophrenia (SP) and psychosis endophenotype (PE) cohorts, with $\rho_{gen} = -0.9163$, however this does not survive correction for 190 possible pairs of correlations.

Table 4.1: Comparison of the 39 non-zero cryptic correlations calculated between each of the studies using different methods: using the set of SNPs shared between the two phenotypes with genome-wide significant SNPs and those in the HLA region removed ("No signal"), the set of SNPs shared between the two phenotypes, thinned so that all SNPs are at least 0.25 centiMorgans apart ("Thin"), the set of all SNPs shared between the two phenotypes ("All"), the intercept of the genetic covariance, calculated by LD Score regression [85, 84] ("LDSC"), and calculated from the study sizes and the number of shared cases and controls in each study ("Calc")—see Section 2.4.2 of Chapter 2 for how this is calculated.

Cohort 1	Cohort 2	No signal	Thin	All	LDSC	Calc
AS	BO	0.192	0.162	0.177	0.196	0.258
AS	IS TOAST 1	0.097	0.0946	0.0932	0.0939	0.164
AS	IS TOAST 2	0.0995	0.0892	0.0868	0.0998	0.139
AS	IS TOAST 3	0.0885	0.0817	0.0804	0.096	0.160
AS	MS UK	0.184	0.143	0.147	0.199	0.258
AS	PD	0.198	0.185	0.169	0.206	0.250
AS	PS	0.198	0.147	0.152	0.206	0.273
AS	UC	0.213	0.203	0.209	0.225	0.270
BO	IS TOAST 1	0.122	0.134	0.122	0.127	0.168
BO	IS TOAST 2	0.133	0.162	0.132	0.135	0.142
BO	IS TOAST 3	0.113	0.121	0.113	0.126	0.163
BO	MS UK	0.229	0.199	0.226	0.241	0.264
BO	PD	0.237	0.227	0.237	0.236	0.256
BO	PS	0.284	0.277	0.273	0.293	0.279
BO	UC	0.244	0.251	0.244	0.268	0.276
BS overall	BS pneumo.	0.619	0.623	0.619	NA	0.616
IS TOAST 1	IS TOAST 2	0.121	0.133	0.120	0.127	0.105
IS TOAST 1	IS TOAST 3	0.169	0.175	0.168	0.164	0.120
IS TOAST 1	MS UK	0.115	0.098	0.111	0.118	0.168
IS TOAST 1	PD	0.111	0.118	0.111	0.110	0.163
IS TOAST 1	PS	0.129	0.119	0.124	0.127	0.178
IS TOAST 1	UC	0.114	0.112	0.115	NA	0.176
IS TOAST 2	IS TOAST 3	0.117	0.137	0.118	0.114	0.102

(Continued on next page)

(continued from previous page)

Cohort 1	Cohort 2	No signal	Thin	All	LDSC	Calc
IS TOAST 2	MS UK	0.124	0.127	0.121	0.127	0.142
IS TOAST 2	PD	0.128	0.117	0.127	0.128	0.138
IS TOAST 2	PS	0.134	0.136	0.127	0.139	0.151
IS TOAST 2	UC	0.124	0.161	0.123	0.128	0.149
IS TOAST 3	MS UK	0.112	0.0866	0.110	0.116	0.163
IS TOAST 3	PD	0.112	0.102	0.112	0.104	0.158
IS TOAST 3	PS	0.113	0.108	0.107	0.122	0.173
IS TOAST 3	UC	0.109	0.131	0.109	NA	0.171
MS UK	MS nonUK	0.0529	0.132	0.0927	0.0191	0
MS UK	PD	0.243	0.245	0.242	0.252	0.256
MS UK	PS	0.243	0.218	0.222	0.254	0.280
MS UK	UC	0.233	0.224	0.231	0.244	0.277
PD	PS	0.249	0.269	0.237	0.253	0.271
PD	UC	0.240	0.251	0.239	0.246	0.267
PS	UC	0.257	0.247	0.240	0.271	0.293
RM reading	RM maths	0.535	0.535	0.535	0.543	NA

We note that the estimates from LD score regression, when they could be calculated, are similar to those calculated by taking correlations of the effect size estimates directly. Furthermore, the results are consistent with expectations: the studies that used WTCCC2 shared controls are the ones between which nonzero cryptic correlations are calculated. Additionally, the methods find correlations between the two RM cohorts—which measured two different traits on the same set of samples—and the BS cohorts, where the pneumococcal samples are a subset of the overall samples. In these last two cases, the estimated correlations are very high, presumably due to the complete sharing of samples between cohorts—we note that for the BS cohorts, the

correlations calculated from the data directly using non-GWAS SNPs and all SNPs are the same as the correlation calculated from Zaykin's formula, which gives the expected correlation due to shared samples. This suggests that for the BS cohorts, the high correlations (≈ 0.62) are due solely to shared controls. While this formula cannot be applied to quantitative traits, like reading and mathematics test scores, we see that the correlations between these two datasets are also high (≥ 0.5), and posit that this is also due to these traits sharing all the same samples; however, we cannot rule out the possibility that these correlations may be inflated due to covariates that have not been properly accounted for (for example, some environmental exposure), or because of variants with true but small effects that are shared by the two traits.

We note other discrepancies in that methods that calculated the correlations from the effect size estimates directly estimate cryptic correlations between the MS UK data and the MS nonUK data, while LD score regression and the analytic formula do not. Because samples were not shared between the two MS cohorts, we would not expect there to be any cryptic correlation between them, unless unmeasured confounders—for example, environmental exposures—somehow induced such correlations. We observe that the method that used the set of SNPs that excluded genome-wide significant associations and the HLA region calculated the lowest correlation between effect size estimates, and posit that for all of the methods that calculate the correlations from the data directly, the apparent correlation between the two MS cohorts is due true effects of small size rather than cryptic relatedness between them.

We also see that the correlations estimated directly from the data are usually lower than those calculated by Zaykin's formula. Zaykin's formula applies to null SNPs, but we cannot guarantee that all of the SNPs used to estimate the correlations from the data directly are null. It may be that each dataset contains variants of small, true effects for one of the studies, which reduces the estimated correlation. It may also be that unaccounted for covariates are introducing noise into the effect size estimates and reducing the estimated correlations between the studies. We discuss how we might test for this in the discussions of both this chapter (Section 4.5) and of Chapter 2 (Section 2.7).

For the analyses presented here, we used the cryptic correlation matrix calculated from the effect size estimates that excluded SNPs that had genome-wide signifi-

cant associations or were in the HLA region. This is because we were able to calculate a complete set of correlations between all pairs of cohorts and its results were close to those of LD score regression, which we believe is the most accurate method when applied to the data from European populations. While we suspect that the nonzero correlation between the MS cohorts is due to real effects of small magnitude, we do not believe that the potential overcorrection will cause many true signals to be missed. This is because we estimate a small correlation ($\approx 5.3\%$) and it is being applied to the study with the largest sample size. We assume that the power of the MS studies (especially the MS nonUK cohort) will be enough to overcome a potentially small overcorrection for confounding.

Because of the diversity of diseases represented in the WTCCC2, choosing prior values for the correlation between diseases is particularly challenging. Additionally, the number of studies made using the subset-exhaustive approach impractical for the initial scan. We instead chose to perform several all-correlated scans using different prior matrices and then take the average over all the results to determine which SNPs to take forward for further analysis. We summarize the prior matrices used in Table 4.2 and visualize them as heatmaps, which we show in Figures 4.1 and 4.2.

Table 4.2: Description of all the prior correlation matrices tested in the initial meta-analysis, along with a reference name

Matrix name	Description
Correlated 1	Independent effects: the pairwise correlations between all cohorts were uniformly set to 0.
Correlated 2	Fixed effects: the pairwise correlations between all cohorts were uniformly set to 1.
Correlated 3	Highly correlated effects: the pairwise correlations between all cohorts were uniformly set to 0.96.
Correlated 4	Moderately correlated effects: the pairwise correlations between all cohorts were uniformly set to 0.5.

(Continued on next page)

(continued from previous page)

Matrix name	Description
Correlated 5	Weakly correlated effects within studies: the pairwise correlations between cohorts from different studies were uniformly set to 0. Within phenotypes or cohorts that were part of the same study—i.e. the BS, IS, MS, and RM cohorts—the correlations were set to 0.1. So, for example, the pairwise correlation between BS overall and MS UK was 0, but the correlation between BS overall and BS pneumococcus was 0.1 and similarly, the correlation between MS UK and MS nonUK was 0.1.
Correlated 6	Moderately correlated effects within studies: the pairwise correlations between cohorts from different studies were uniformly set to 0. Within phenotypes or cohorts that were part of the same study—i.e. the BS, IS, MS, and RM cohorts—the correlations were set to 0.5. This is the same as Correlated 5, but with a higher correlation coefficient between cohorts within a study.
Correlated 7	Highly correlated effects within studies: the pairwise correlations between cohorts from different studies were uniformly set to 0. Within phenotypes or cohorts that were part of the same study—i.e. the BS, IS, MS, and RM cohorts—the correlations were set to 0.96. This is the same as Correlated 5 and 6, but with a higher correlation coefficient between cohorts within a study.

(Continued on next page)

(continued from previous page)

Matrix name	Description
Correlated 8	Highly correlated effects within studies, weakly correlated effects within disease classes: the pairwise correlations between cohorts from different studies were usually set to 0. The exceptions were within studies of diseases in the same class. That is, between all pairs of autoimmune diseases (AS, PS, MS cohorts, and UC), all pairs of infectious diseases (BS cohorts and VL cohorts), and the two psychiatric phenotypes (SP and PE), there was a correlation coefficient of 0.1. Within cohorts of the same study—i.e. the BS, IS, MS, and RM cohorts—the correlations were set to 0.96.
Correlated 9	Highly correlated effects within studies, moderately correlated effects within disease classes: the pairwise correlations between cohorts from different studies were usually set to 0. The exceptions were within studies of diseases in the same class. That is, between all pairs of autoimmune diseases (AS, PS, MS cohorts, and UC), all pairs of infectious diseases (BS cohorts and VL cohorts), and the two psychiatric phenotypes (SP and PE), there was a correlation coefficient of 0.5. Within cohorts of the same study—i.e. the BS, IS, MS, and RM cohorts—the correlations were set to 0.96. This is the same as Correlated 8, but with the correlations between diseases in the same class (autoimmune, infections, psychiatric) raised to 0.5.

(Continued on next page)

(continued from previous page)

Matrix name	Description
Correlated 10	Moderately correlated effects within studies, weakly correlated effects within disease classes: the pairwise correlations between cohorts from different studies were usually set to 0. The exceptions were within studies of diseases in the same class. That is, between all pairs of autoimmune diseases (AS, PS, MS cohorts, and UC), all pairs of infectious diseases (BS cohorts and VL cohorts), and the two psychiatric phenotypes (SP and PE), there was a correlation coefficient of 0.1. Within cohorts of the same study—i.e. the BS, IS, MS, and RM cohorts—the correlations were set to 0.5. This is the same as Correlated 8, but with the correlations within cohorts in the same study lowered to 0.5.
Correlated 11	Moderately correlated effects within studies, moderately correlated effects within disease classes: the pairwise correlations between cohorts from different studies were usually set to 0. The exceptions were within studies of diseases in the same class. That is, between all pairs of autoimmune diseases (AS, PS, MS cohorts, and UC), all pairs of infectious diseases (BS cohorts and VL cohorts), and the two psychiatric phenotypes (SP and PE), there was a correlation coefficient of 0.5. Within cohorts of the same study—i.e. the BS, IS, MS, and RM cohorts—the correlations were also set to 0.5. This is the same as Correlated 9 and 10, but using 0.5 as the correlation coefficient between disease classes and between study cohorts.

(Continued on next page)

(continued from previous page)

Matrix name	Description
Correlated 12	Highly correlated effects within studies, highly correlated effects within disease classes: the pairwise correlations between cohorts from different studies were usually set to 0. The exceptions were within studies of diseases in the same class. That is, between all pairs of autoimmune diseases (AS, PS, MS cohorts, and UC), all pairs of infectious diseases (BS cohorts and VL cohorts), and the two psychiatric phenotypes (SP and PE), there was a correlation coefficient of 0.96. Within cohorts of the same study—i.e. the BS, IS, MS, and RM cohorts—the correlations were also set to 0.96. This is the same as Correlated 11, but raising the correlation coefficient used to 0.96.

Figure 4.1: Visualizations of the eighth correlation matrix, described in Table 4.2.

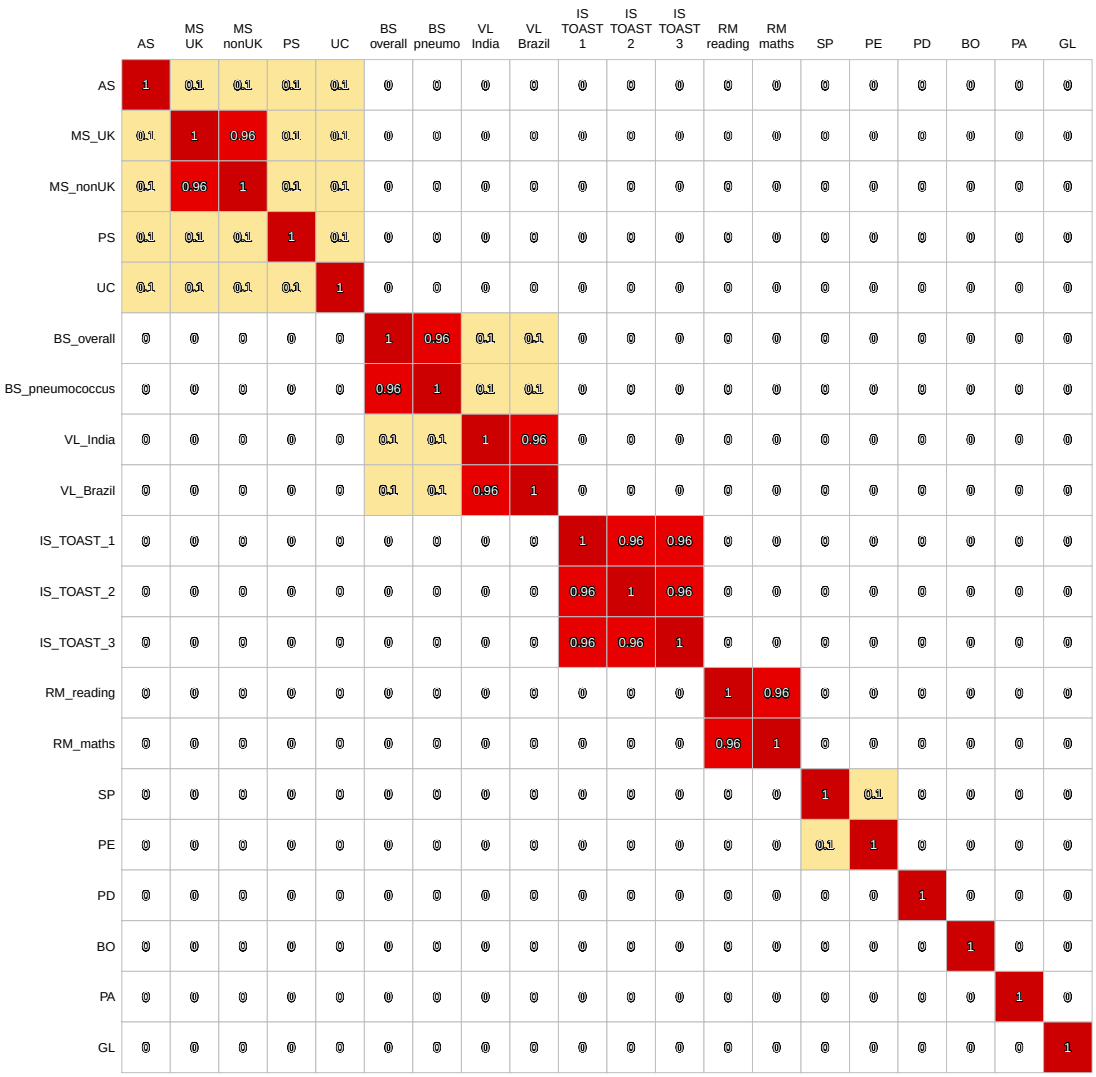
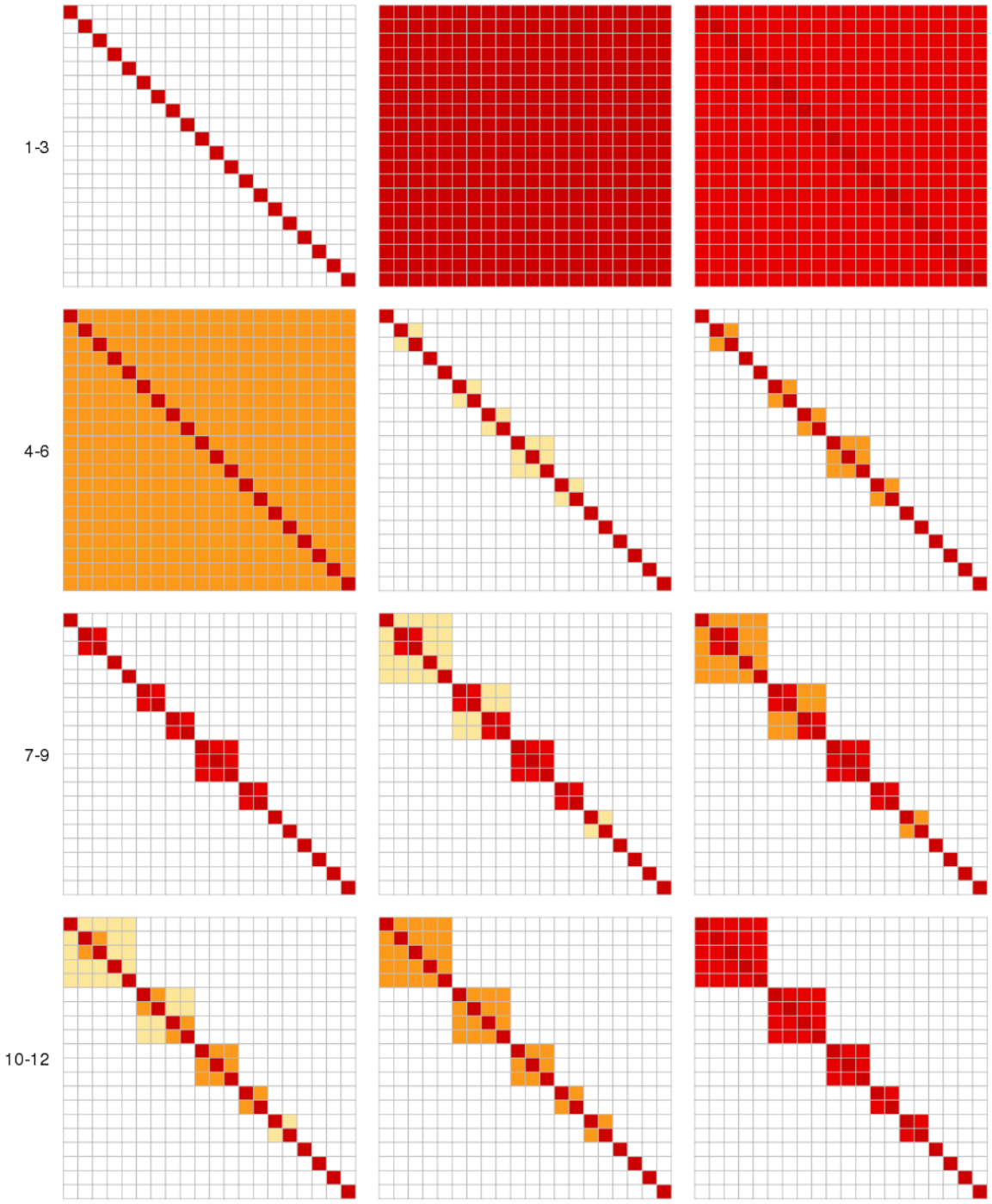


Figure 4.2: Visualizations of the prior correlation matrices described in Table 4.2. The darkest red boxes reflect correlation coefficients of 1, boxes colored in the next darkest red correspond to correlations of 0.96, yellow-orange reflects correlations of 0.5, light yellow is for correlations of 0.1, and white means a correlation coefficient of 0.



In addition to performing the analysis across multiple patterns of genetic correlation among the studies, we also varied the prior on true effect sizes (σ, see Section 2.2.1 of Chapter 2 and 1.5.2 of Chapter 1), using 0.1, 0.2, and 0.4 as possible values, which encode a belief in small, medium, and large effect sizes, respectively. We translate these into relative risk ratios for the risk allele for the top 5% and top 1% of effect size estimates in Table 4.3, as discussed in 2.2.3 in Chapter 2. This created a total of 36 all-correlated ABFs calculated for each SNP. It is over these models that we took the mean all-correlated ABF used in our results.

σ	5%	1%
0.1	1.217	1.294
0.2	1.480	1.674
0.4	2.190	2.802

Table 4.3: Expected minimum relative risk ratios of the risk allele for either the top 5% or 1% of effect sizes under different prior values of σ. The relative risk ratios for the corresponding protective alleles are simply the inverse of these values.

In a standard GWAS, it is common to check the distribution of the test statistic using a quantile-quantile (QQ) plot. Ideally, the test statistic (or its associated p-values) will follow the null distribution—this appears as the points of the plot following the line $y = x$, deviating from it only at extreme values. This deviation signifies the true associations found by the GWAS. A test statistic that deviates from the null distribution at the majority of data points—that is, the points are offset from the line $y = x$—suggests problems with the data. This could be the result of inaccurate genotyping or phenotyping, or it could be a problem with the analysis—perhaps the researchers failed to account adequately for population structure or other confounders. Because we are conducting our meta-analysis on a genome-wide set of SNPs (as opposed to a set of SNPs that have been selected based on results suggesting they are associated with a trait), we make the same assumption as a standard GWAS about the distribution of our approximate Bayes factors, and expect the majority of them to follow the null distribution. However, in order to make this comparison, we first need to generate a null distribution to which to compare our data.

To that end, we simulated data under the null using our cryptic correlation matrices and the standard errors for each SNP in each study. For each SNP that was not in the HLA region, we used the **rmvnorm** function from the **mvtnorm** package

in R to generate ten sets of effect size estimates under the null for the n cohorts for which we had data at the SNP. The covariance matrix to generate the estimates was created using the cryptic correlation matrix and the standard errors of the n cohorts. We calculated all-correlated ABFs for each of these simulated SNPs using every combination of prior matrix and σ value, and compared the distribution of the ABFs from this null dataset to the corresponding ABFs in the real data calculated using the same priors.

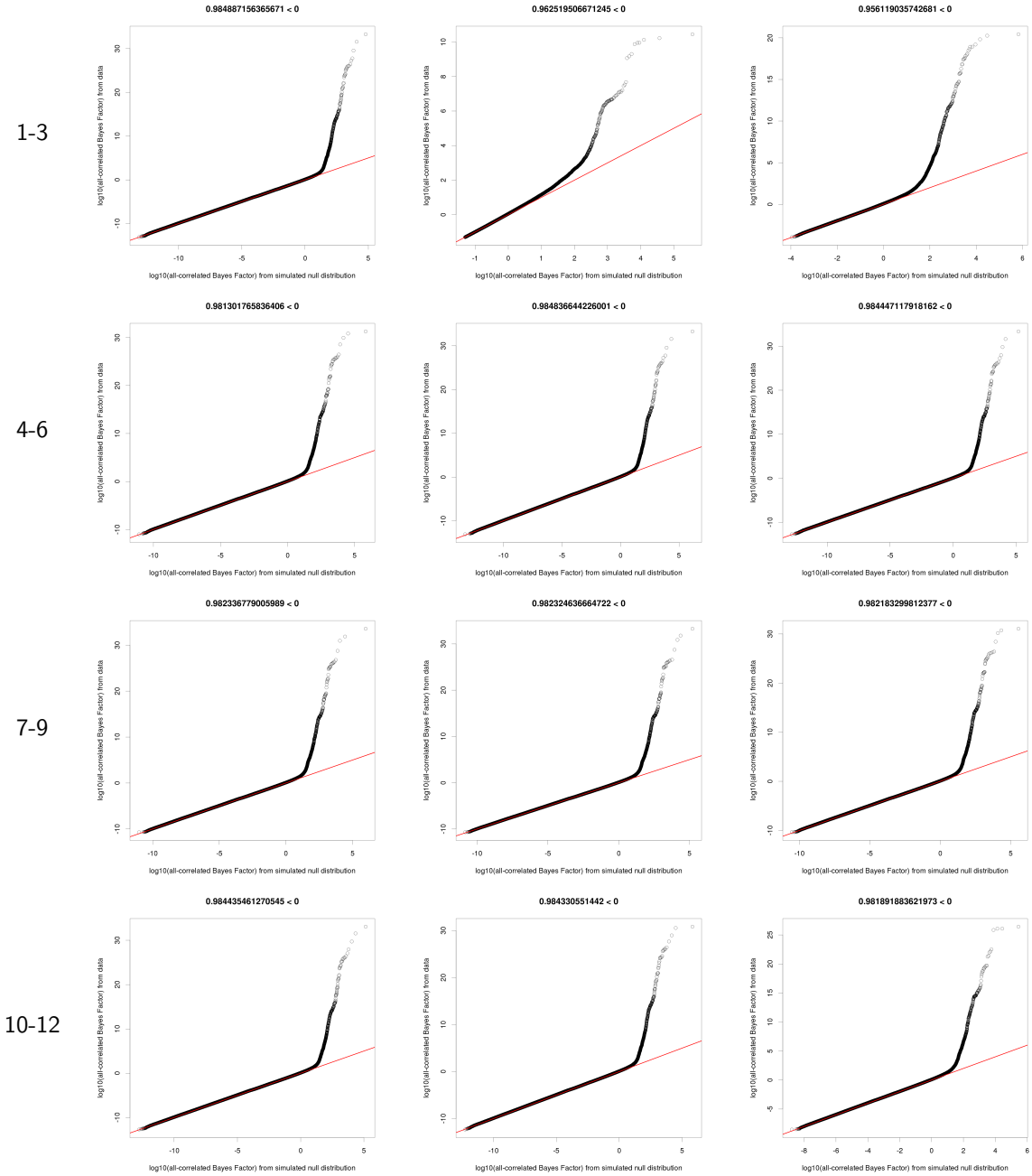
We present the Q-Q plots for approximate Bayes factors calculated under each of the 12 prior correlation matrices in Figure 4.3. The Q-Q plots shown here used the cryptic correlation matrix calculated from effect size estimates from the summary statistics, excluding genome-wide significant SNPs and those in the HLA region. These calculations used a prior standard error of $\sigma = 0.2$. The results for $\sigma = 0.4$ and $\sigma = 0.1$ were similar (see Figures 4.B.1 and 4.B.2 in the Appendix), though in the former case, a larger proportion of SNPs tended to have $\log_{10}(\text{ABFs}) < 0$ than in the $\sigma = 0.2$ case (usually the proportion was upwards of 0.98). In the latter case, a smaller proportion of SNPs tended to have $\log_{10}(\text{ABFs}) < 0$, though the proportion was always greater than 0.91. This variation in the proportion of ABFs that show evidence in favor of the null is not unexpected. A smaller standard error corresponds to a belief that true effects will be smaller, leading to more effects being "found" and consequently more positive $\log_{10}(\text{ABFs}) < 0$. Analogously, a larger standard error means fewer effects will be found and fewer positive $\log_{10}(\text{ABFs}) < 0$.

These plots show us that the data follow the expected null distribution most of the time, suggesting that the observed deviation from the null is the result of interesting true effects. When choosing markers for follow-up, we required that

$$\log_{10}(\text{mean all-correlated ABF}) \geq 4.$$

These plots also show us that under some of these prior correlation matrices positive values of $\log_{10}$(all-correlated ABFs) are still consistent with results from the null distribution; however, by the time these values are as high as 4, they deviate from the null significantly, making this a reasonable cut-off, especially as a stringent cut-off is needed to account for the increased number of postive $\log_{10}$(all-correlated ABFs) when calculated using a prior of $\sigma = 0.1$.

Figure 4.3: Q-Q plots for the distribution of $\log_{10}(\text{approximate Bayes factors of the null data})$ data (x-axes) vs. $\log_{10}(\text{approximate Bayes factors of the real data})$ (y-axes) calculated using each of the 12 different correlation matrices and the cryptic correlation matrix calculated from effect size estimates excluding genome-wide significant SNPs and the HLA region. The red line is the line $y = x$. The number at the top of each plot is the proportion of $\log_{10}(\text{approximate Bayes factors})$ that were less than zero. In all cases shown here, this number was at least 0.95.



4.3.1 Autoimmune analysis

We have previously highlighted the representation of autoimmune diseases in this dataset. Because of the known shared genetic risk factors among members of this class of diseases, they form a particularly appealing set of traits to which to apply our method. Additionally, the study subjects in the MS, AS, and PS cohorts were all genotyped on the Illumina chips, which meant that there are 435,249 markers common to all four datasets. LD score regression was also able to calculate the cryptic relatedness between each pair of studies, which gave us the opportunity to use our preferred cryptic covariance matrix in this analysis. In reality, it did not differ greatly from the matrix used in the whole WTCCC2 analysis, except that the matrix calculated using LD score regression found no cryptic relatedness between the two MS cohorts.

The smaller number of cohorts allowed us to perform the subset exhaustive ABF approach over all combinations of priors. This permitted us to take the mean ABF over all possible subsets of models using all possible priors. Unlike the analysis of all the cohorts, we opted for correlation matrices that assumed the same pairwise correlations, ρ, between all of the cohorts. We performed this analysis using $\rho = \{0, 0.1, 0.5, 0.96\}$ and $\sigma = \{0.1, 0.2, 0.4\}$, as before. In all prior matrices, we assumed a strong correlation ($\rho = 0.96$) between the two multiple sclerosis cohorts and treated them as one study.

This analysis of a subset of the traits was performed to help assess whether there is any added benefit to meta-analyzing traits that are not expected *a priori* to have any common genetic causes. It also gives us insight into the potential advantages of having data for every trait at every marker versus a dataset that is missing information for some traits at some markers.

4.4 Results

Here we present the results of the analyses we have performed. As previously stated, these are necessarily exploratory in nature and must be viewed in the context of assessing the performance of our method rather than searching for novel results, as we have no replication cohorts, and detailed functional follow-up to validate potentially

novel results is beyond the scope of this project. However, some results may be validated through a search of the literature to see if other studies of these traits have found similar results.

For both the autoimmune and full analyses, we identified a set of top independent SNPs by using the 1000 Genomes Phase 3 recombination map. Because this map is on NCBI human genome build 37, we lifted the map coordinates over to build 36 and interpolated the genetic distance for any loci that were not successfully lifted over. To choose the SNPs for follow-up, we began with a list of all SNPs whose mean all-correlated ABFs were at least 4. The SNP with the highest mean ABF was selected and any SNPs within 0.25 centiMorgans of it were removed from consideration. Then the SNP with the next highest ABF was selected, the SNPs within 0.25cM of it were removed from the list of candidates, and so on until there were no SNPs left. For both analyses, this process left us with a list of approximately independent SNPs which showed evidence of cross-trait associations.

4.4.1 Full WTCCC2 meta-analysis

We performed our all-correlated analysis as described in Section 4.3 and show the results in Figure 4.1. We also show a summary of the results of the subset-exhaustive analysis for the SNPs associated with the highest ABF in a given region in Figure 4.2. We discuss some of these top SNPs in detail below, highlighting those that show the potential utility and short-comings of our approach, as well as those whose results show an interesting combination of associated traits.

The ABF approach allows definitive statements to be made about association and non-association on chromosome 1

The SNP rs3890745 on 1p36.32, is a good example of the utility of our approach. Our analysis gives high posterior probabilites on this SNP's association with both MS cohorts and UC. This SNP is in high LD ($R^2 = 0.9954$ in EUR populations) with rs4648356, which was a known MS SNP replicated in the original WTCCC2 MS GWAS. However, there is no data at rs4648356 for UC, which is probably why the analysis flags a variant in high LD with it as a probable locus for both diseases. Additionally, rs3890745 has been previously associated with rheumatoid arthritis,

another autoimmune disease [175, 139, 176], which makes the further association with ulcerative colitis more plausible. However, while there is reason to believe that this locus is associated with two of the four autoimmune diseases in our dataset, we note that there is very little marginal posterior probability on AS or PS, suggesting that they are explicitly not associated with this locus.

A potential link between autoimmune diseases and pre-eclampsia on chromosome 5

Our analysis also picks out the coding variant rs30187 in the gene *ERAP1* (formerly *ALAP*) on 5q15 as being associated with both AS and PS. This SNP was previously known as an ankylosing spondylitis locus and was flagged as a possible psoriasis SNP in the WTCCC2 GWAS. Ultimately, the authors of that study chose rs27524 ($R^2 = 0.7583$ in EUR populations) as the likelier locus since the model of rs27524 interaction with the HLA Cw*0602 allele fit the data better than a model involving rs30187, however the latter could not be ruled out entirely.

The reason why rs30187 was highlighted by our analysis is because there is also some evidence for association with pre-eclampsia, with a posterior probability of association of 0.728 for this trait when calculating the subset-exhaustive ABFs using the prior parameters that maximized the ABF over all prior parameters and models of association (these parameters were the same for both SNPs). The posterior probability for pre-eclampsia on rs27524 was 0.205, suggesting that it is not associated with this locus. The LD between the two SNPs in the 1000 Genomes AMR populations is lower than in the EUR cohorts, with $R^2 = 0.4479$.

From the literature, there is further reason to believe that the association of rs30187 with pre-eclampsia could be real. This locus was significantly associated with hypertension in a Japanese cohort [177]. There was also some evidence of the association of *ERAP1* with pre-eclampsia in a study of Australian and Norwegian cohorts, though there was stronger evidence for *ERAP2*, which lies about 71kb downstream of *ERAP1*. Furthermore, the study's authors note that "The *ERAP1* and *ERAP2* genes encode enzymes that are reported to play a role in blood pressure regulation and essential hypertension in addition to innate immune and inflammatory responses," which is also relevant to the associations with AS and PS [178].

An overlooked association is found by the ABF method, along with a possible VL association on chromosome 10

The case of rs793108 on chromosome 10 highlights the potential of our method to find true associations that were overlooked by traditional GWASs. At this SNP, the ABF approach puts posterior probabilities of 0.999 and 0.972 on an association between this SNP and the UK and nonUK multiple sclerosis cohorts, respectively. While this SNP was not a genome-wide significant locus in the original WTCCC2 study [140], it was genome-wide significant in a follow-up study using the Immunochip [179], and is also a rheumatoid arthritis locus [34].

Our analysis of rs793108 shows another high posterior probability (≈ 0.982) of association with the Indian visceral leishmaniasis cohort. In fact, this study has the most evidence of association when looking at the Wakefield ABFs of all studies at this SNP (50.57 for VL India, as opposed to 42.84 for MS UK and 26.32 for MS nonUK). There is nothing in the literature that validates this finding, however there have been few studies of the genetic risk of visceral leishmaniasis, and the ones that have been performed have lacked sufficient power to find associations with modest effect sizes [180, 167]. The closest gene is *ZNF438*, which is assumed to be associated with this marker. It shows further associations with height in East Asians [181], blood lipid response to fenofibrate [182], and congenital heart defects [183], none of which provide any further information about our putative VL association.

A probable spurious result on chromosome 11

The SNP rs16920014 on chromosome 11, highlighted by our analysis is notable for the lack of SNPs in LD around it showing high ABFs, and also for the lack of information about it in the literature. This SNP has a low minor allele frequency—in our data it ranges from 0.033 in the GL cohort to 0.057 in the bacteremia cohorts. Additionally, the SNPs surrounding it have higher MAFs; in the GL cohort, rs16920014 is clearly an outlier, as it is surrounded by SNPs with MAFs around 0.42. Furthermore, the association at this marker is driven by a very large estimated effect in UC ($\hat{\beta} = 0.5513$), which had a very small p-value ($p = 2.8 \times 10^{-18}$). As this result was not reported in the UC GWAS [160], we assume that it did not replicate and that this apparent association is the result of an excess of cases or controls having the minor allele (MAF

was 0.041 in the UC cohort) due to sampling variation.

Non-autoimmune traits associated with the chromosome 17 inversion

There is a well-known 970 kb inversion on 17q21.31 [184, 185]. The inversion spans the genomic region containing the SNPs rs393152, rs17769552, rs1981997, and rs2668692, which have been highlighted by our analysis. This inversion is known as the *MAPT* inversion and is associated with two main haplotypes, H1 and H2, with H1 occurring in all populations and matching the human reference genome, while H2 contains the inversion and is more commmon in Europeans—occurring in about 20% of chromosomes in this population—than in other continental populations [184].

Despite our filter for genetic distance, these SNPs are all in high LD with one another (the lowest pairwise R^2 was 0.925 between rs17769552 and rs2668692 in EUR populations). This region is known to be associated with Parkinson's disease, particularly in variants on the H1 haplotype [186, 187, 188]. Indeed, in our data, we see that all four of the highlighted SNPs show associations with PD. The SNP rs393152 was cited in the PD GWAS as a previously known PD locus in the *MAPT* gene [189] and has also been associated with corticobasal degeneration [190].

The region also has known Alzheimer's disease loci. The Alzheimer's SNP rs7207400 has $R^2 \geq 0.729$ with our 17q21.31 SNPs in EUR populations. This is interesting because there is some evidence that glaucoma is comorbid with Parkinson's disease and Alzheimer's disease [191, 192], though the association with PD was disputed by an eight year longitudinal study [193]. It may be that the associations of rs393152, rs17769552, and rs1981997 with GL are mediated through one or both of Parkinson's disease and Alzheimer's disease.

Additionally, all four SNPs show some degree of association with pre-eclampsia, while rs393152, rs1981997, and rs2668692 show strong associations with BO. The SNP rs1981997 is associated with idiopathic pulmonary fibrosis—scarring of the lung tissue—however because of the high LD and large number of genes around the SNP, the authors of the study declined to state which gene was associated with the disease [194]. This disease does not have a clear analogy to any of the traits in the WTCCC2.

Other traits associated with this region and with SNPs in high LD with our variants included bone mineral density ($R^2 \geq 0.941$ for our variants and the bone

mineral density variant rs1864325), intra-cranial volume in infants ($R^2 \geq 0.906$ for our variants and rs9303525), male pattern baldness ($R^2 \geq 0.951$ for our variants and rs12373124), neuroticism ($R^2 \geq 0.946$ for our variants and rs111433752), progressive supranuclear palsy—another neurodegenerative movement disorder—($R^2 \geq 0.936$ for our variants and rs8070723), and—perhaps unsurprisingly, given the association of the region with corticobasal degeneration—subcortical brain region volumes ($R^2 \geq 0.951$ for our variants and rs8072451 and $R^2 \geq 0.946$ for rs17689882).

All this is to say that this is clearly a complex and highly pleiotropic region. LD is high across large sections of it—no doubt due in part to the fact that H2 haplotypes do not recombine among themselves or with H1 haplotypes [195]. Because the region spans multiple genes, it makes it very hard to choose causal variants associated with a phenotype and to map these back to the correct gene. According to the UCSC Genome Browser [196], the nine genes contained in the genomic region containing our four SNPs are *LINC02210*, *CRHR1*, *MAPT-AS1*, *SPPL2C*, *MAPT*, *MAPT-IT1*, *STH*, *KANSL1*, and *KANSL1-AS1*.

Another overlooked association, which is also a potential stroke locus is found on chromosome 19

The SNP rs8106664, an intronic variant of the gene *SLC44A2* on 19p13.2 once again highlights the potential of this method to find associated loci that were overlooked in GWAS. This SNP is in high LD ($R^2 = 0.935$) with the coding variant rs2288904, which was found to be associated with MS in a 2013 study [179]. Both SNPs show some evidence of association in the WTCCC2 GWAS—after combining the evidence from both the UK and nonUK cohorts, the p-values were 1.787×10^{-4} for rs8106664 and 3.240×10^{-4} for rs2288904. Additionally, the marginal associations for PS and PD were stronger at rs8106664 than at rs2288904, with $p = 7.292 \times 10^{-4}$ for PS at rs8106664 as opposed to $p = 4.172 \times 10^{-3}$ at rs2288904 and similarly, $p = 8.575 \times 10^{-3}$ for PD at rs8106664, while $p = 1.053 \times 10^{-2}$ at rs2288904. Once again, the priors that maximized the ABF over all subsets and prior parameters (matrix 12 and $\sigma = 0.1$) were the same for both SNPs.

Additionally, rs2288904 was found to be associated with venous thromboembolism [197], which would explain the evidence for association with the IS cohorts,

especially large and small vessel stroke. Once again, the marginal evidence in the WTCCC2 data slightly favored rs8106664 over rs2288904 in each of the traits. Because of this our analyses show higher posterior probabilities across all associated traits, including the more marginally associated IS cohorts.

There is less in the literature to explain the associations with psoriasis and Parkinson's disease. There are numerous loci associated with PS in 19p13.2—the closest one to *SLC44A2* is in *QTRT1*, about 90kb away. However, none of these loci show strong LD with rs8106664 or rs2288904 ($R^2 < 0.35$). It may be that this signal is confounded by the SNPs in *QTRT1*, or it may be that there is some undetected PS locus in *SLC44A2*. There do not seem to be any known associations at this locus or nearby for Parkinson's disease.

An association between autoimmune disease and reading and mathematical abilities in the Down's syndrome critical region of chromosome 21

The final SNP we highlight is rs2094871 on 21q22.2. This SNP was previously associated with UC [198], though it did not meet the threshold for genome-wide significance in the WTCCC2 study ($p = 1.6 \times 10^{-6}$). 21q22 was associated with AS in the WTCCC2, but at the SNP rs378108, which has $R^2 = 0.5248$ with rs2094871 in EUR populations. Addtionally, rs2242944 (which is in LD with rs2094871—$R^2 = 0.881$ in EUR populations) was associated with AS disease in a different study [199]. Unsurprisingly, rs2094871 is in moderate LD ($R^2 = 0.641$) with other SNPs associated with UC, Crohn's disease, and inflammatory bowel disease in general. The SNP rs2836878 has an added association with C-reactive protein levels, which are a marker of inflammation. We plot the individual results of these SNPs in this region in Figure 4.3.

In addition to the strong association we see with both of these autoimmune diseases, there is also a strong signal in both of the RM phenotypes. The posterior probabilities for reading and maths were 0.987 and 0.998, respectively. However, unlike the autoimmune phenotypes, there are no other SNPs in LD with rs2094871 that relate to phenotypes relevant to educational outcomes, intelligence, or brain function. The only possible exception is rs75884327 ($R^2 = 0.457$ in EUR populations), which showed some evidence (though not at a genome-wide significant level) with a

daytime napping phenotype in adults [200]. However, chromosome 21 is famous for its association with Down's syndrome, which is caused by a trisomy, and results in intellectual disability as well as physical phenotypes. It is believed that 21q22.2 is part of the Down's syndrome critical region—that is, if there is only a partial trisomy of chromosome 21, then 21q22.2 and 21q22.3 must be triplicated in order for Down's syndrome to result [201, 202].

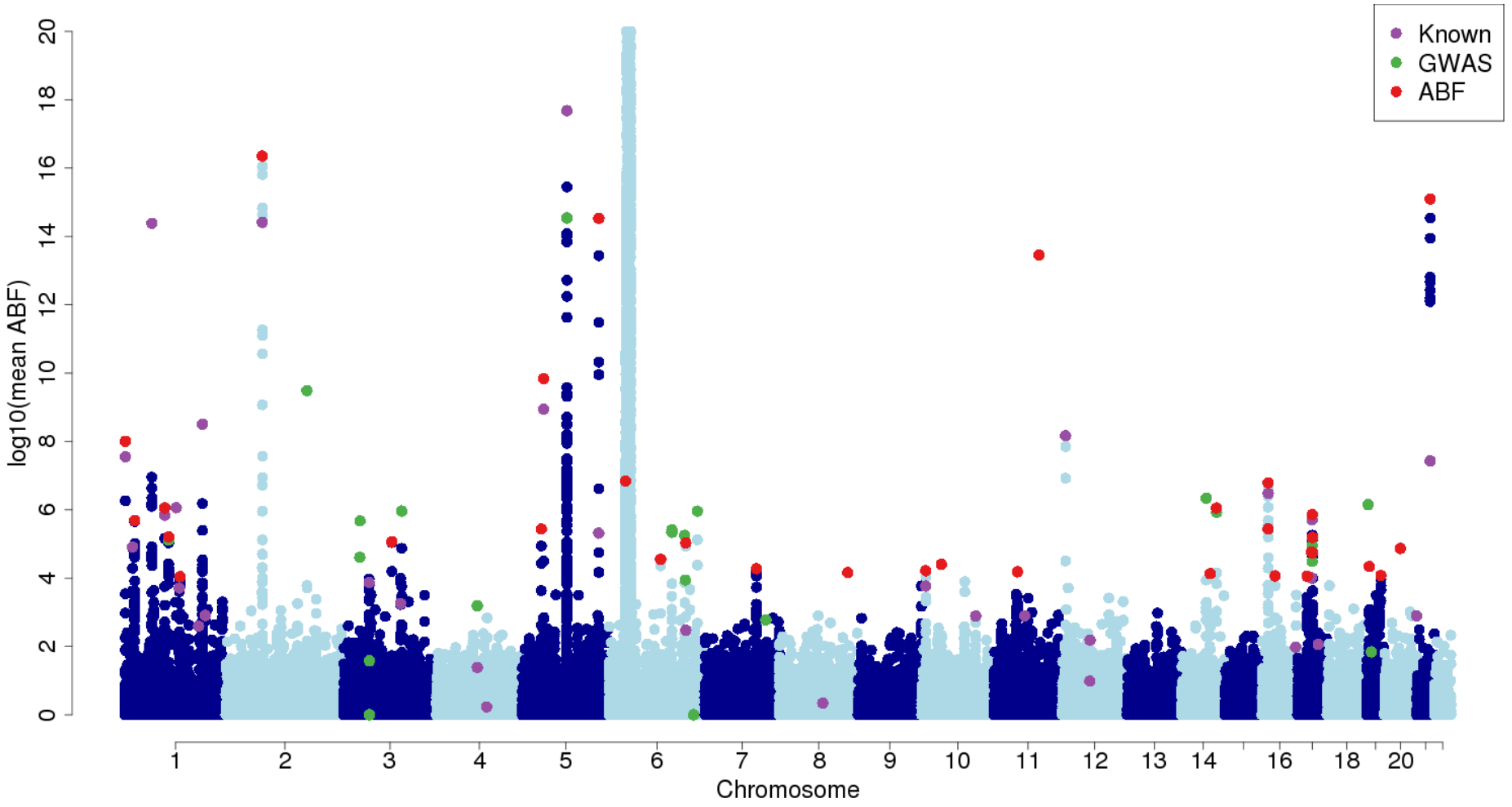


Figure 4.1: Manhattan plot of the $\log_{10}$(mean ABF) for all WTCCC2 SNPs across 14 of the 15 studies—the data from study on metformin response could not be aligned properly and was thus dropped from the analysis. Each study contained a list of loci that were previously known to be associated with at least one phenotype. These are highlighted in purple ("Known"). SNPs that achieved genome-wide significance and were novel in the original GWAS are colored green ("GWAS"), and novel SNPs found by our method are colored red ("ABF").

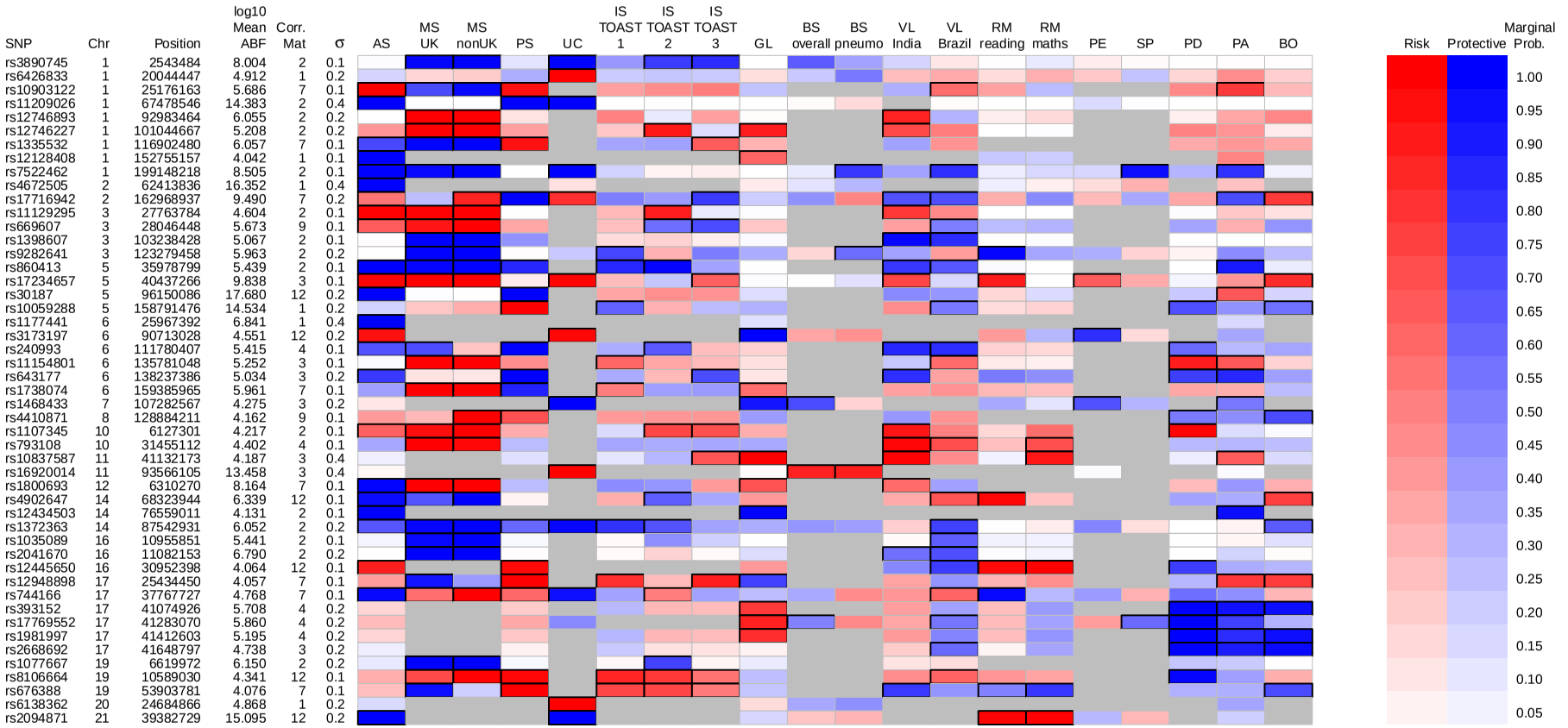


Figure 4.2: Table showing the SNPs that were taken forward for further analysis of their cross-trait associations. In the table on the left, we give the rsid, chromosome, genomic position, the $\log_{10}$ (mean all-correlated ABF), the correlation matrix and value of σ associated with the highest ABF across all possible subsets (and all possible prior correlation matrices and values for σ). In the chart on the right, we show the pattern of association that corresponded to the highest Bayes factor using the subset-exhaustive approach and the prior parameters in the table at the left. Black boxes are drawn around the traits associated with the SNP in the model with the highest ABF, and the color indicates whether the effect allele was a risk (red) or protective allele (blue). The intensity of the color is based on the marginal probability of association at the SNP, with the scale at the far right.

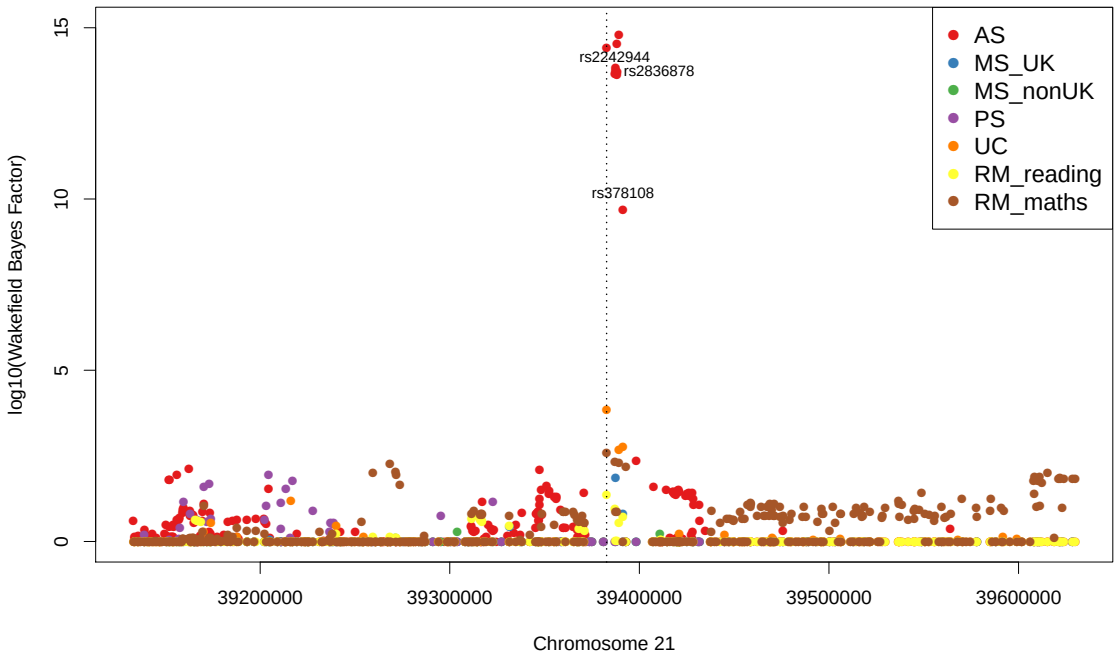


Figure 4.3: Hit plot of the region around rs2094871. The $\log_{10}$ (Wakefield ABF) for each autoimmune disease are plotted on the y-axis and the position on chromosome 21 plotted on the x axis. The dotted line shows the index SNP. We highlight three other SNPs (labelled) discussed in relation to their associations with autoimmune diseases and the reading and mathematics traits in this region.

4.4.2 Autoimmune meta-analysis

The results of the full WTCCC2 suggest that there are disadvantages to implementing this method on datasets with missing data. We also see that the three autoimmune diseases from this analysis are well represented in the results of the full analysis. We now compare the autoimmune analysis results to the ones obained from the full analysis.

Once again, we present our results for the four autoimmune disease cohorts as a Manhattan plot. The SNPs with the highest ABF in a region (within 0.25 cM of the lead SNP, as before) are colored based on the pattern of association in the model with the highest ABF across all priors. The "fixed", "correlated", and "independent" models were the prior matrices that assumed each study was associated with correlations between true effects in each of the cohorts of $\rho = 0.96$ in the fixed effects model, $\rho = \{0.1, 0.5\}$ in the correlated effects model, and $\rho = 0$ in the independent effects model.

We note that 48 of the 57 associations contain at least one of the MS cohorts.

This is in contrast to 31 associations for AS and 27 associations for PS. It is likely that this is due to the power of the multiple sclerosis GWAS. We also note that there is much overlap between the regions found by this analysis and the regions found in the analysis using all cohorts.

The autoimmune meta-analysis highlights a region of association on chromosome 16

One key difference between the autoimmune analysis and the whole data analysis is the peak at the start of chromosome 16. It is much higher when looking only at autoimmune diseases. This region is associated with the *CLEC16A* gene and was identified as showing "strong evidence" of association in the multiple sclerosis study [140]. The lead SNP in that study, rs7200786, is correlated with both rs1035089 ($R^2 = 0.471$ in European populations) and rs12103174 ($R^2 = 0.683$ in European populations) highlighted by our analysis and lies between them. We note that the $\log_{10}$(mean ABF) for rs12103174 is 10.77, and for rs7200786, it was 10.68, suggesting that the amount of evidence for assocation at both SNPs is about equal. It may well be that rs7200786 is true lead signal in our data as well, but due to stochastic noise from the other studies in the meta-analysis, it just misses out on being the highest ranked SNP in the region.

Associations between AS and PS that do not appear in the whole WTCCC2 analysis

We note that rs30187 appears as an ankylosing spondylitis and psoriasis SNP in this analysis, as it did in the full WTCCC2 analysis. However, unique to the autoimmune disease results is the SNP rs10865331 on chromosome 2, which shows a very high posterior probability of association with both AS and PS as well. While this was a known AS locus before the WTCCC2 study [199], a subsequent study using the Immunochip has shown it to be associated with PS as well [203], and additionally, it is also believed to be a Crohn's disease/IBD locus [158]. Similarly, rs2546890, which was found by the WTCCC2 GWAS for MS [140] appears to to have a further association with psoriasis in our analysis, which is confirmed by another psoriasis GWAS [204].

SNPs on chromosome 21 produce a probable spurious result and highlight the problem of multiple causal variants tagged by a single marker

One issue that is highlighted by comparing this analysis to the full dataset is when a single marker tags multiple causal variants. As an example, we consider association tagged by rs2836878 on chromosome 21. This is a known inflammatory bowel disease locus [159, 158], however despite ulcerative colitis not being part of this analysis, we still find an association with AS and MS at this SNP. This marker is not even nominally associated in the MS nonUK cohort; the Wakefield ABF using a prior $\sigma = 0.1$ is 0.904, suggesting that even assuming small effect sizes, the null model describes the data at this SNP for this cohort better. Thus, we are highly skeptical of this putative association with MS, especially as this region has no known associations with that disease. However, this SNP is in LD ($R^2 = 0.606$ in Europeans) with rs2242944, which we discussed in relation to our findings at rs2094871 (which is also in LD with rs2242944, with $R^2 = 0.641$). Neither rs2242944 nor rs2094871 are in this dataset, suggesting that perhaps this result is tagging the true AS association through LD with one of these SNPs. We plot the Wakefield ABFs for each trait at the SNPs in the 500 kb region around rs2836878 in Figure 4.5. We note the dearth of SNPs with information about UC in the area immediately surrounding rs2836878.

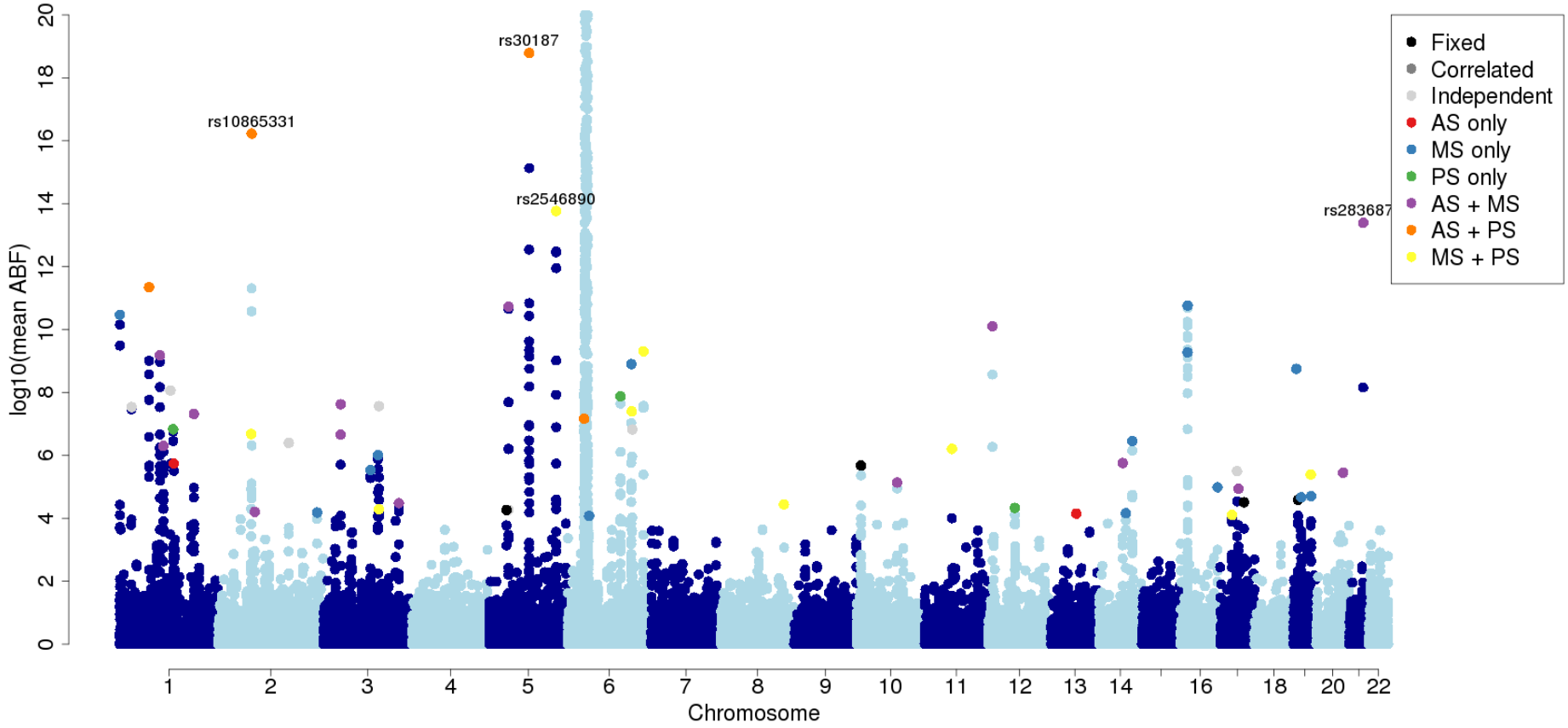


Figure 4.4: Manhattan plot of the $\log_{10}$ (mean ABF) for the SNPs that contained data for the AS, PS, and two MS cohorts. The means were taken across all subset models analyzed using 12 possible priors. Here we color the SNPs with the highest mean ABF in a given region based on the pattern of association at the model with the highest ABF over all subsets and all prior parameters.

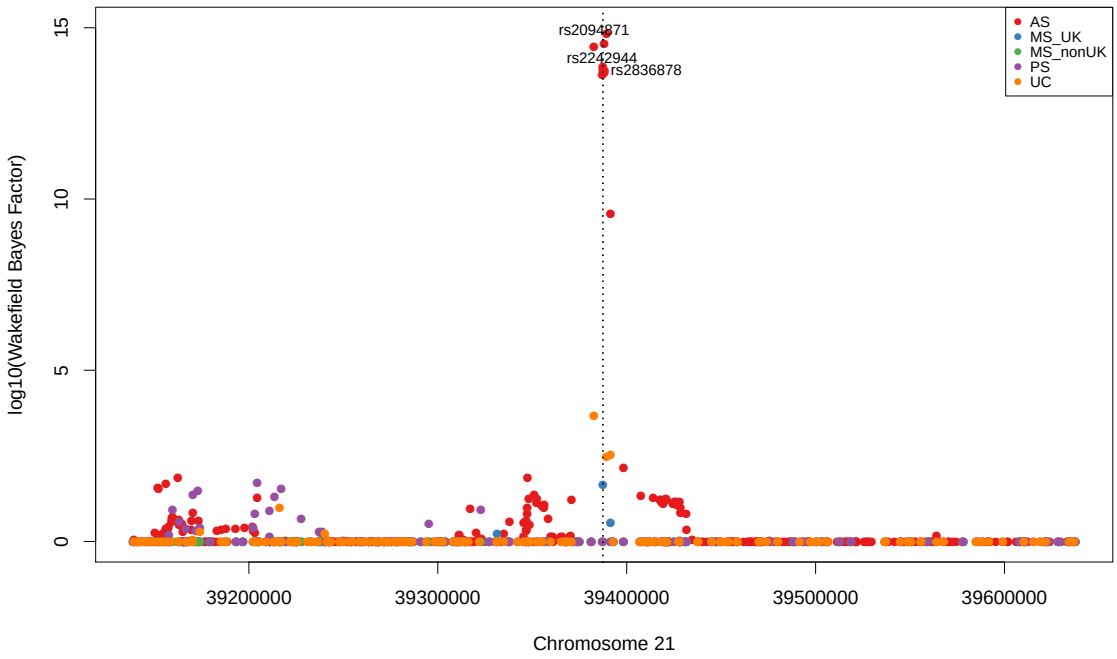


Figure 4.5: Hit plot of the region around rs2836878. The $\log_{10}$ (Wakefield ABF) for each autoimmune disease are plotted on the y-axis and the position on chromosome 21 plotted on the x axis. The dotted line shows the index SNP. We highlight two other SNPs discussed in relation to their associations with autoimmune diseases in this region.

4.4.3 Gene enrichment in WTCCC2 non-HLA results

Because one of the goals of studying the genetic architecture of disease is to gain insights into disease etiology, we were interested in learning if there was any biological commonality among the associations we found in our whole WTCCC2 analysis. For instance, there might be a biological pathway the genes associated with our results are involved in. Alternatively, we may find that these genes are often expressed in a particular tissue. Knowing this can give our findings a biological context. However, we also must note that we are not entirely certain about the validity of some of our findings and the number of significant loci returned by our analysis is small, which makes it more challenging to determine that the set is enriched for anything specific.

We mapped the SNPs highlighted in Figure 4.2 to genes using SNPsnap [205]. The gene assigned to each SNP was the gene it was deemed to lie within or, if there was no such gene, then the gene whose start site was the closest to the SNP. We then analyzed this list through the Reactome's online tool [206] for enrichment in any biological pathways. Using a false discovery rate of 0.05, the pathways that appeared

to be significant in this analysis (in order of decreasing overrepresentation compared to what would be expected genome-wide) were interleukin-10 signalling, signalling by interleukins, cytokine signaling in the immune system, interleukin-4 and 13 signaling, transcriptional regulation of pluripotent stem cells, TNFR1-induced proapoptotic signaling, and RUNX1 and FOXP3 control and development of regulatory T lymphocytes.

These suggest that the pathways affected by the markers highlighted by our analysis are largely related to immune response, particularly through interleukins, which are pro-inflammatory cytokines [206], as is tumor necrosis factor (TNF). In fact, anti-TNF therapies that target the extracellular domain's tumor necrosis factor receptor 1 (TNFR1) to prevent it from interacting with TNF have been successfully employed in the treatment of rheumatoid arthritis and inflammatory bowel disease [207].

FORGE analysis [208] using 10,000 matched SNPs and data from Roadmap Epigenomics [209] shows that the set of markers highlighted by our analysis is enriched for overlap with DNAseI sensitive regions in certain tissues, suggesting tissue-specific functional relevance. Of the 26 cell types whose enrichment p-values were were below 0.01 (Bonferroni-corrected for the number of tissues tested), 18 were expressed in the blood, four were expressed in the fetal thymus, three in the skin, and one in the breast. The blood cells represented in this enrichment analysis were positive for the antigens CD3, CD4, CD8, CD14, CD19, CD34, and CD56. The first three and CD56 are antigens appearing on T cells, while CD14 appears on macrophages and granulocytes. CD19 appears in B cells, and CD34 is a hematopoietic progenitor cell [210, 211, 212].

Taken all together, these results suggest that the regions highlighted by our analyses affect immune response, particularly through the action of leukoctyes. Given that all but five of the 49 SNPs highlighted in Figure 4.2 are associated with at least one autoimmune disease, these results are not particularly surprising. Unfortunately, they provide little insight into the genetic and physiological bases of the other traits in our meta-analysis, since all of them except for the reading and mathematics traits are medical conditions of one sort or another and therefore might all be expected to involve the immune system to some extent.

4.5 Discussion

Our analysis of the WTCCC2 studies has demonstrated this meta-analysis approach does indeed scale to genome-wide data on up to 20 cohorts. Furthermore, it is possible for this analysis to find true signals that were overlooked in the original GWAS. By testing over a range of priors, we can find the one that maximizes the approximate Bayes factor associated with our data, which helps ensure that we do not miss associations due to poor choice of prior. We also have the option of using the average over the results of the different priors to reduce some of the uncertainty about exactly which prior is the most appropriate. The intuition is that the signal from priors that fit the data well will overwhelm the lack of signal from priors that do not.

Our results showed enrichment in pathways and cell expression relevant to inflammation and immune response. Furthermore, the loci implicated in our analysis correspond to DNAseI peaks in blood cells containing clusters of differentiation antigens that mark leukocytes, which are critical to immune response. Since autoimmune diseases are more closely related to one another than any other set of five cohorts in our meta-analysis, it is to be expected that loci related to them would dominate our results, especially since the power of our method is derived from the shared genetic bases of different traits. This led to an enrichment of tissues and biological pathways relevant to the etiologies of these diseases.

Additionally, we believe that this type of meta-analysis does have advantages over single study analyses. In Figure 4.1, we compare the mean ABFs over all 12 correlation matrices and all three values of σ at each SNP to the mean Wakefield ABFs across all values of σ for each individual study. We see that the means of the meta-analyzed ABFs are rarely lower than the means of the study-specific ABFs (above the red $y = x$ line), and when they are, the difference between the two is not great—especially not compared to some of the differences between the meta-analyzed ABFs and the study-specific ones when the meta-analyzed ABFs are higher. This suggests that the meta-analysis approach will usually have at least as much power to detect effects as the single study approach, probably due to the fact that its calculations include information from other, potentially related traits.

However, the analysis has limitations. For instance, while we have shown that we can estimate cryptic correlations between studies from the effect size estimates

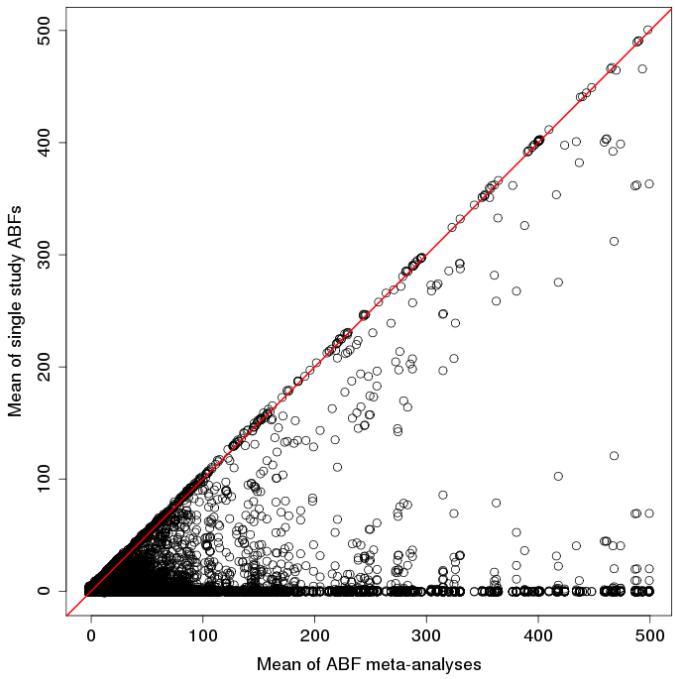


Figure 4.1: Plot of the $\log_{10}$(mean all-correlated ABFs) calculated across all possible prior parameters (x-axis) against the $\log_{10}$(mean Wakefield ABFs) calculated across all possible values of σ. The line $y = x$ is colored red.

alone, we do not fully understand what these correlations are capturing. Our results do not differ greatly from those found using LD score regression, but tend to be lower than those calculated from an analytic formula that calculates correlations due to shared samples. As discussed in Section 2.7 in Chapter 2, we would need to perform more simulations to truly understand what can be captured by the cryptic covariance matrix, $\boldsymbol{\Omega}_{\hat{\boldsymbol{\beta}}}$—whether it is only the effect of shared controls, or whether it can account for other confounders, as well.

Furthermore, it is impossible to tell if an apparent cross-trait association is due to the same underlying causal variant for both (or all) associated cohorts, or if it is the result of multiple causal variants that are tagged by the same typed variant. A third possibility is that the associated variant is correlated with multiple other typed variants whose marginal results are stronger in a single study, but the aggregate evidence favors the SNP that has more moderate marginal results in more of the studies, as we see when comparing the results for the full WTCCC2 analysis at rs10059288 and rs2546890 in Figure 4.2. We highlighted this problem while investigating a possible association of rs2836878 with MS and AS in the autoimmune disease analysis.

Such convoluted signals have less to say about the underlying biology causing the

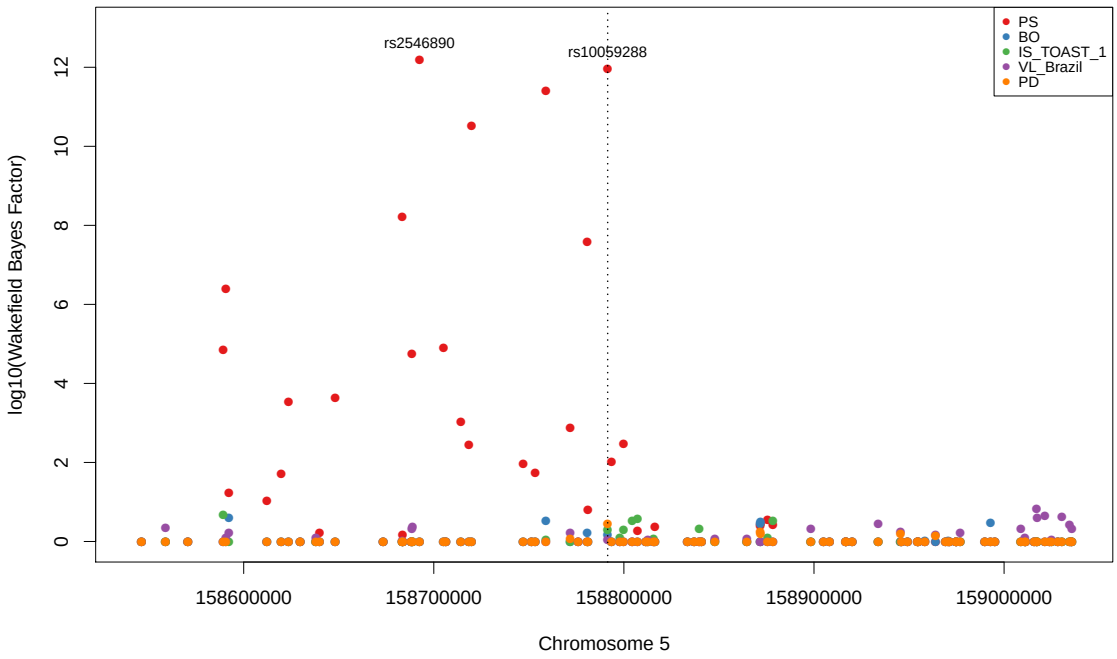


Figure 4.2: Hit plot of the region around rs10059288. The $\log_{10}$ (Wakefield ABF) for each autoimmune disease are plotted on the y-axis and the SNP's position on chromosome 5 is plotted on the x axis. The dotted line shows the index SNP. The locus rs2546890 is marked. It shows a higher marginal effect in PS than the index SNP, but lower marginal effects in all the other traits.

apparent cross-trait association, since the causal variants may be part of completely different pathways. While it may be interesting from an evolutionary or population genetics standpoint to investigate why certain haplotypes maintain a correlation between traits, for those who wish to learn about the genetic etiology of different traits, the tagging of multiple causal variants at a single typed locus reduces power to identify variants with cross-trait associations, and our method appears to be vulnerable to it. It is vital to perform follow-up analysis to understand the associations it finds. There are also methods available that specifically aim to determine whether two association signals in a locus share a causal variant [42, 44], which we could try implementing as part of the data quality control process, ahead of analysis, using GWAS p-values or single-study ABFs to find the lead SNP for each disease in a given region. Alternatively, these methods could be applied after the meta-analysis to filter results.

Follow-up analysis is important to gain understanding of results generated by our method; however, the reliance on other published studies to validate our results is a hindrance to understanding potentially novel results. For instance, the result at rs2094871 on chromosome 21 can be verified for the two autoimmune traits, but we

cannot find any way of replicating the results for reading and mathematical ability, in part because they are unusual traits for study. We assumed that IQ and educational attainment were related traits and tried to follow up our results in two GWASs of educational attainment [213, 214] as well as a meta-analysis of intelligence tests that either correlated highly with scores for g—the generalized intelligence factor—or tested it directly [215]. These studies were all large—the two studies of educational attainment had more than 100,000 samples each, while the intelligence study meta-analyzed results for 78,308 samples across 13 cohorts. The summary statistics for these studies are all available on the Social Science Genetic Association Consortium's website, at `https://www.thessgac.org/data`.

The locus on chromosome 21 did not replicate in any of these studies. However, nothing replicated between the WTCCC2 reading and mathematics traits and these three datasets. The only genome-wide significant result in the WTCCC2 study was an association for reading scores with a region on chromosome 19. This region is not associated with educational attainment or intelligence in any of these three studies. Additionally, none of the genome-wide associated regions in the studies were associated with the WTCCC2 traits, even at a p-value threshold of $p \leq 0.0001$ for the WTCCC2 studies. The reverse was also true. SNPs significant at at $p \leq 0.0001$ in the WTCCC2 studies corresponded to regions where $p \geq 10^{-6}$ in these studies.* What any of this means for our own analysis is unclear, though it might suggest that reading and mathematical ability at the age of 12 does not have the same genetic basis as adulthood intelligence or educational attainment. However, if this is true, then it begs the question of what traits they are related to.

The case of rs16920014 is an example of how our approach is sensitive to strong (and in this case, spurious) results in a single study. We see this among the SNPs that are near the $y = x$ line in Figure 4.1—the high results in a single study leads to a high mean all-correlated ABF as well. Another example of this phenomenon is the SNP rs6426833 on 1q36.13, which shows a strong marginal association with ulcerative colitis, at $p = 6.97 \times 10^{-11}$ (see Figure 4.3). The $\log_{10}$(Wakefield ABF) $= 7.850$ for the association of UC at this SNP against the null. Without the data from UC, under the priors that maximized the ABF for the full dataset at that SNP, the

*The higher p-value threshold is because we expect the non-WTCCC2 studies to be better powered than the WTCCC2 studies, due to their larger sample sizes.

$\log_{10}$(all-correlated ABF) $= -3.91218$. Moreover, further investigation showed that the large estimated effect in UC did not replicate in the original study and this locus was not included among the study's findings.

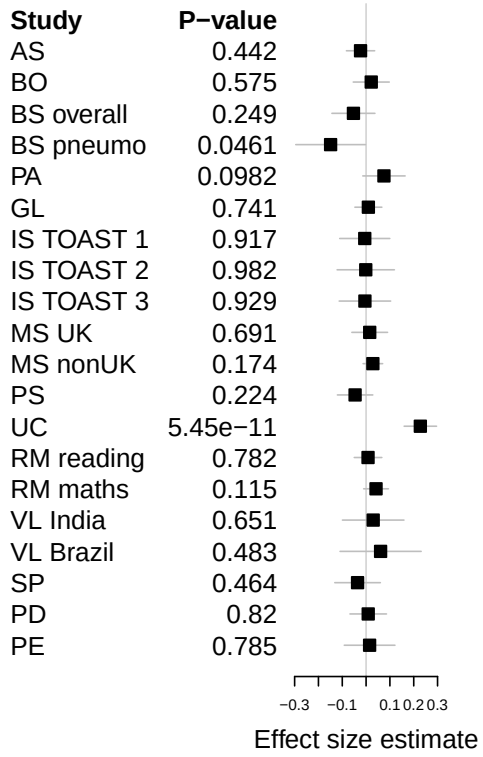


Figure 4.3: Forest plot of the effect size estimates and 95% confidence intervals for all WTCCC2 cohorts at the SNP rs6426833.

We also see from these analyses how we lose power when we are missing data at a given variant for one or more of the studies. For instance, the SNP rs3890745 was in high LD with a multiple sclerosis locus tagged by rs4648356. However, despite rs4648356 having the marginally lower p-value for the two MS cohorts (3.086×10^{-14} vs 5.707×10^{-14}) and because there is no data for UC at rs4648356, our analysis finds rs3890745 as the top SNP in the region. Taking the data from UC out would put the highest all-correlated ABF for rs3890745 0.7 orders of magnitude below that of rs4648356. Thus, if there is a true genetic signal for UC at rs4648356 that is similar to or stronger than the one at rs3890745, then it seems likely that the true cross-trait association for this region should be at rs4648356, and that it has been penalized for missing data that is available at a correlated SNP. This issue is rife in our data, due to some studies being imputed and others not, and because of differences in the chips

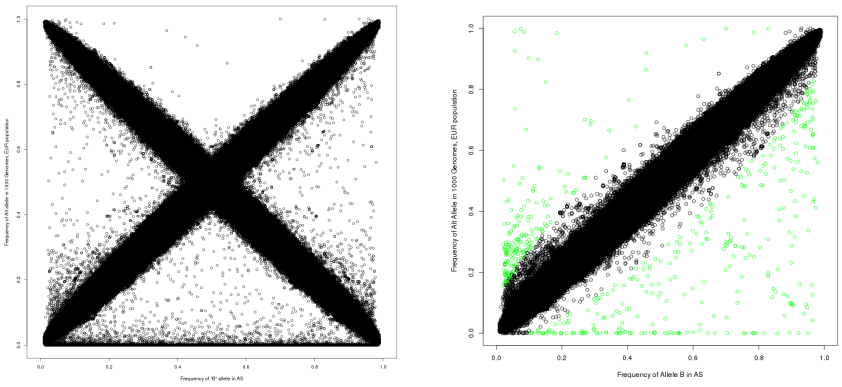
used for genotyping. One potential way around this is to impute missing summary statistics, as was done for the Immunobase data in the next chapter. Another potential solution is to implement one of the colocalization methods [42, 43, 44], to determine if two associations share a causal variant.

We must also note that we have not taken full advantage of the flexibility of our approach. There is no reason to assume that σ would be constant across all cohorts at all places in the genome. Similarly, the prior correlation matrix that models the data best may also change depending on genomic context. While we have attempted to solve this by performing the analysis using different prior matrices, we have made no attempt to vary σ across studies. For instance, in the example of rs6426833 above, if we change the prior values of σ so that $\sigma = 0.4$ for UC and 0.1 for the other cohorts, the maximum $\log_{10}$(all-correlated ABF) = 6.360, up from 5.983. In real terms, this raises the all-correlated ABF by about 2.385, which is not very much, but it demonstrates that we can improve the way alternative models describe the data by letting σ vary across individual cohorts within a SNP, as well as allowing it to vary across SNPs. Another avenue of further investigation is another round of simulations to see whether this increases our power to detect effects true effects and select the correct underlying model of association at a given SNP, and if so, how much this changes from the inference we do already.

While it is clear that our method can scale to genome-wide data and can find valid results, our results may also show extra associations at SNPs that we cannot verify. It is unclear how to proceed in these instances. Additionally, we need to be concerned that the multiple associations we find are the result of a single SNP tagging multiple causal variants; this problem is inherited from GWAS, where typed SNPs may or may not be causal and, in the latter case, may tag more than one causal variant. More worrisome is the tendency of the method to calculate higher ABFs for SNPs that have more data. This can potentially be overcome with summary statistic imputation, but it is nevertheless a short-coming of the method. Future work should include investigating a way to remove this in-built penalty for missing data.

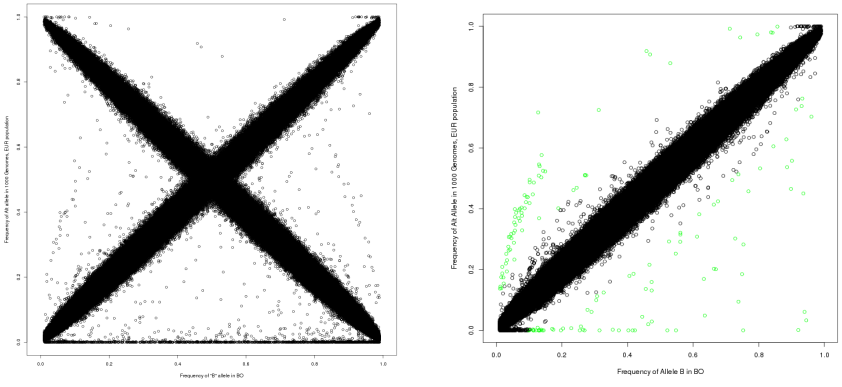
Appendix

4.A Allele frequency plots before and after strand alignment and filtering



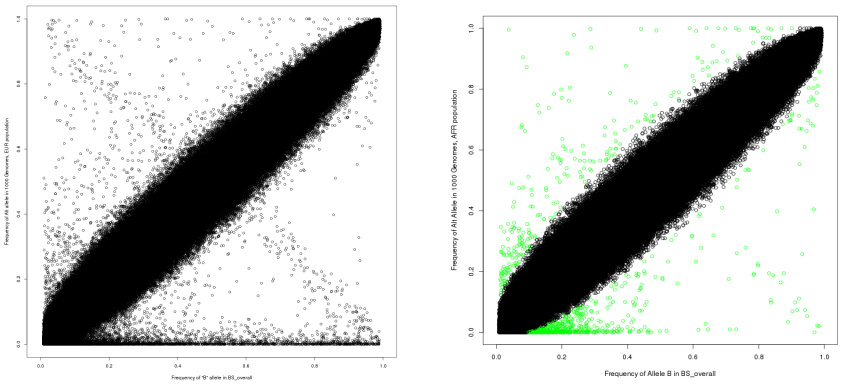
(a) (b)

Figure 4.A.1: Comparison of allele frequencies in the ankylosing spondylitis study and in 1000 Genomes EUR population before alignment and filtering (left) and after (right).



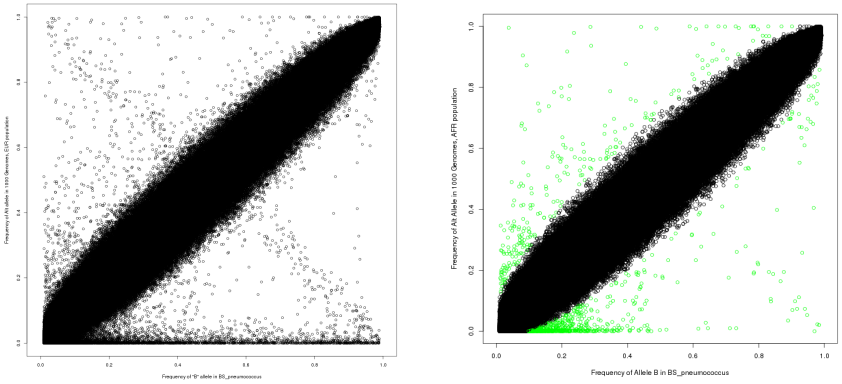
(a) (b)

Figure 4.A.2: Comparison of allele frequencies in the Barrett's esophagus study and in 1000 Genomes EUR population before alignment and filtering (left) and after (right).



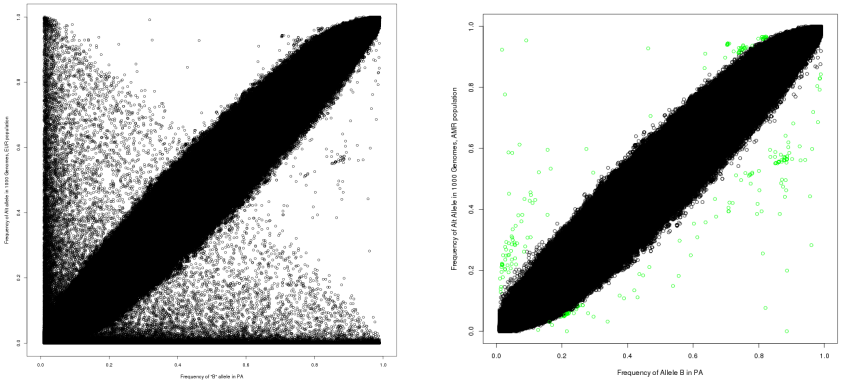
(a) (b)

Figure 4.A.3: Comparison of allele frequencies in the study of overall bacteremia susceptibility and in 1000 Genomes AFR population before alignment and filtering (left) and after (right).



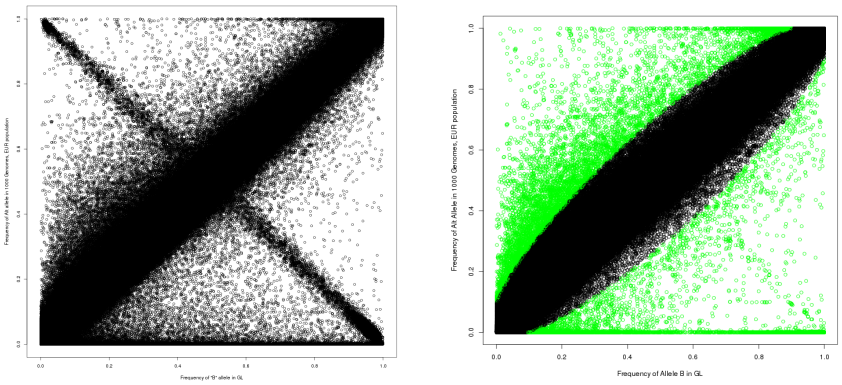
(a) (b)

Figure 4.A.4: Comparison of allele frequencies in the study of pneumococcal bacteremia susceptibility and in 1000 Genomes AFR population before alignment and filtering (left) and after (right).



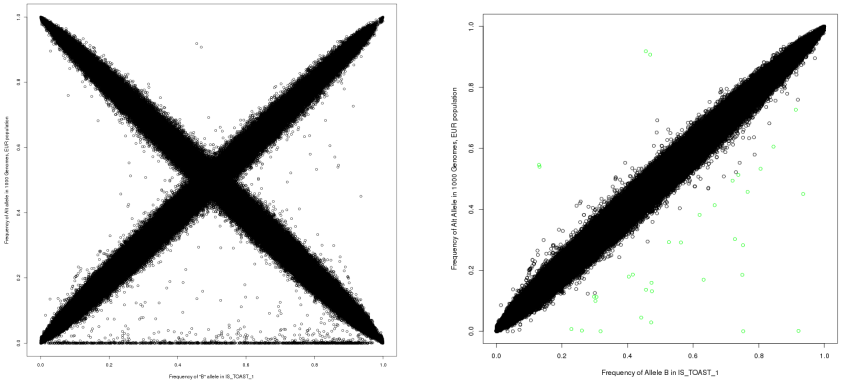
(a) (b)

Figure 4.A.5: Comparison of allele frequencies in the pre-eclampsia study and in 1000 Genomes AMR population before alignment and filtering (left) and after (right).



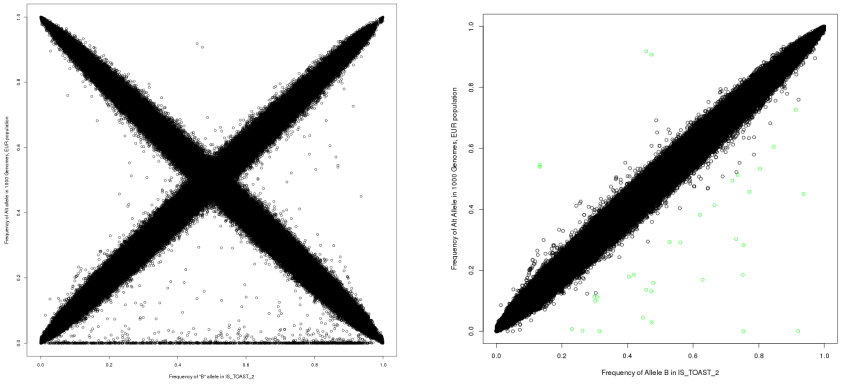
(a) (b)

Figure 4.A.6: Comparison of allele frequencies in the glaucoma study and in 1000 Genomes EUR population before alignment and filtering (left) and after (right).



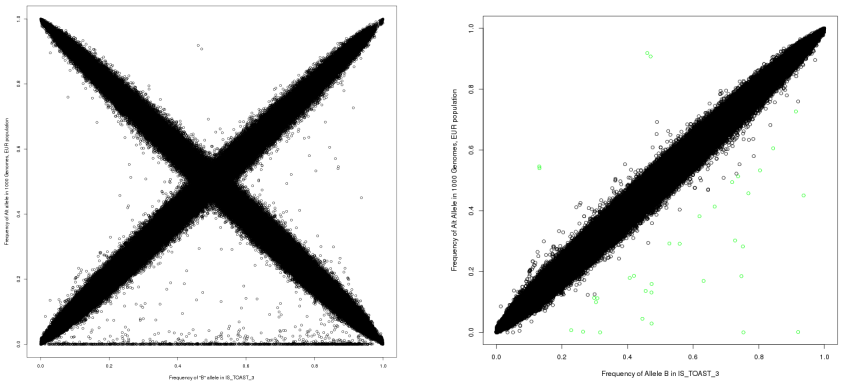
(a) (b)

Figure 4.A.7: Comparison of allele frequencies in the ischemic stroke (TOAST 1) study and in 1000 Genomes EUR population before alignment and filtering (left) and after (right).



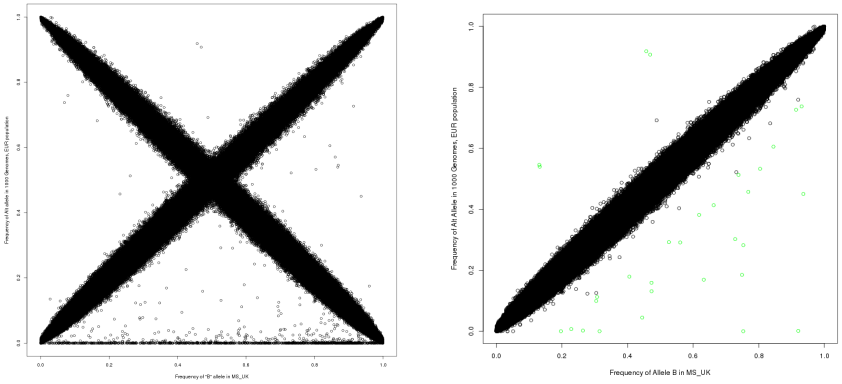
(a) (b)

Figure 4.A.8: Comparison of allele frequencies in the ischemic stroke (TOAST 2) study and in 1000 Genomes EUR population before alignment and filtering (left) and after (right).



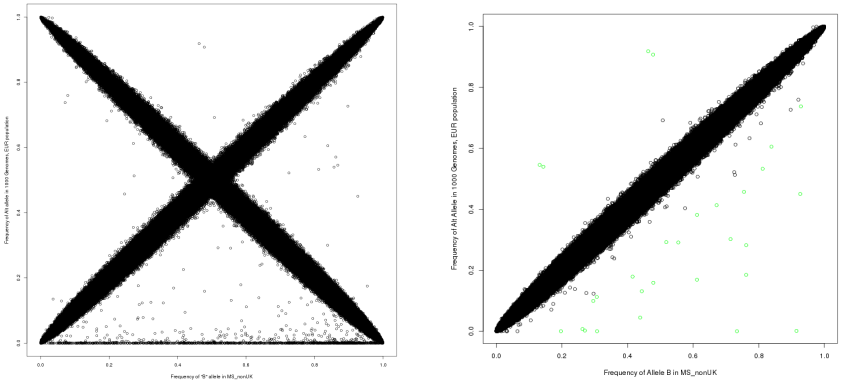
(a) (b)

Figure 4.A.9: Comparison of allele frequencies in the ischemic stroke (TOAST 3) study and in 1000 Genomes EUR population before alignment and filtering (left) and after (right).



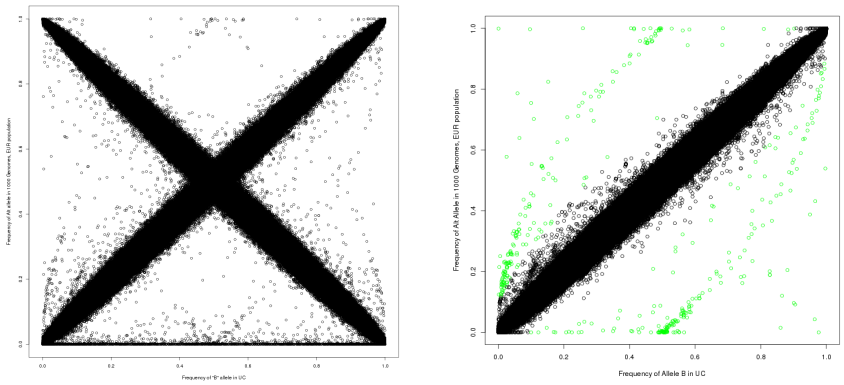
(a) (b)

Figure 4.A.10: Comparison of allele frequencies in the multiple sclerosis (UK cohort) study and in 1000 Genomes EUR population before alignment and filtering (left) and after (right).



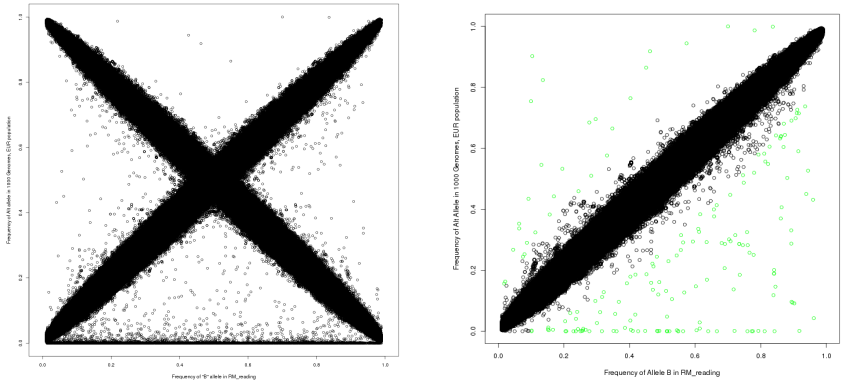
(a) (b)

Figure 4.A.11: Comparison of allele frequencies in the multiple sclerosis (nonUK cohorts) study and in 1000 Genomes EUR population before alignment and filtering (left) and after (right).



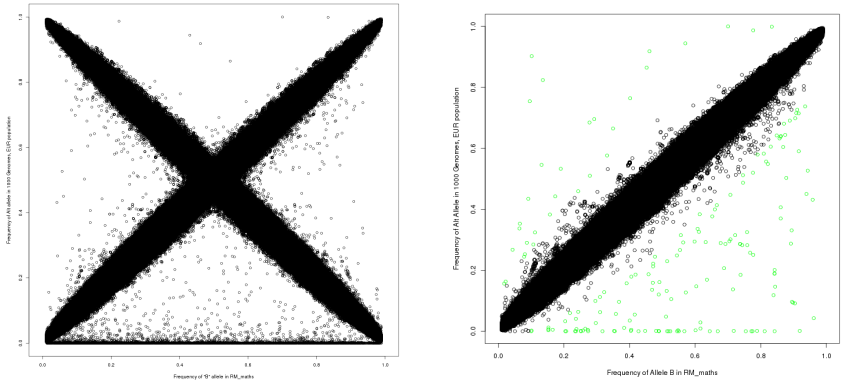
(a) (b)

Figure 4.A.12: Comparison of allele frequencies in the ulcerative colitis study and in 1000 Genomes EUR population before alignment and filtering (left) and after (right).



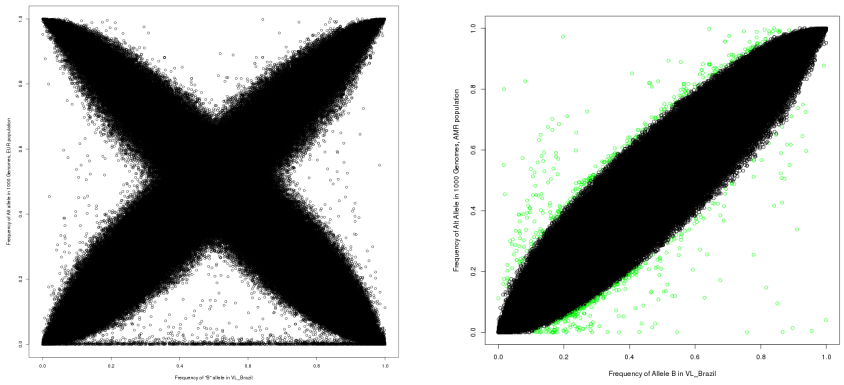
(a) (b)

Figure 4.A.13: Comparison of allele frequencies in the study of reading ability at age 12 and in 1000 Genomes EUR population before alignment and filtering (left) and after (right).



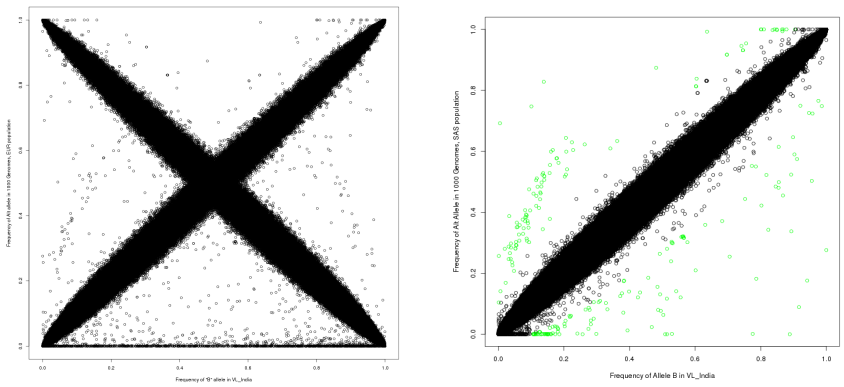
(a) (b)

Figure 4.A.14: Comparison of allele frequencies in the study of mathematical ability at age 12 and in 1000 Genomes EUR population before alignment and filtering (left) and after (right).



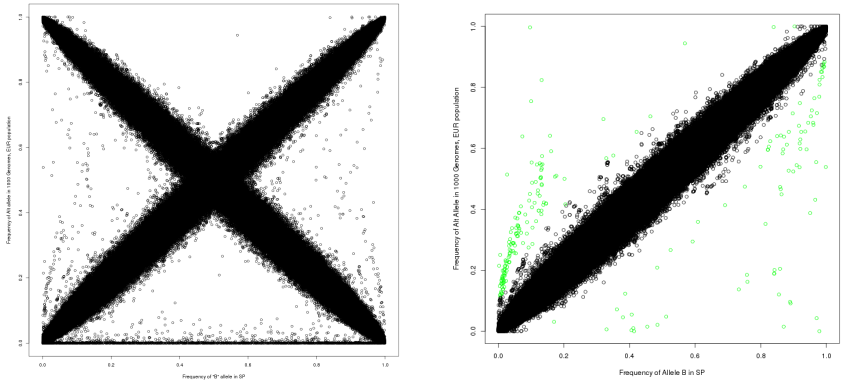
(a) (b)

Figure 4.A.15: Comparison of allele frequencies in the visceral leishmaniasis (Brazil) study and in 1000 Genomes AMR population before alignment and filtering (left) and after (right).



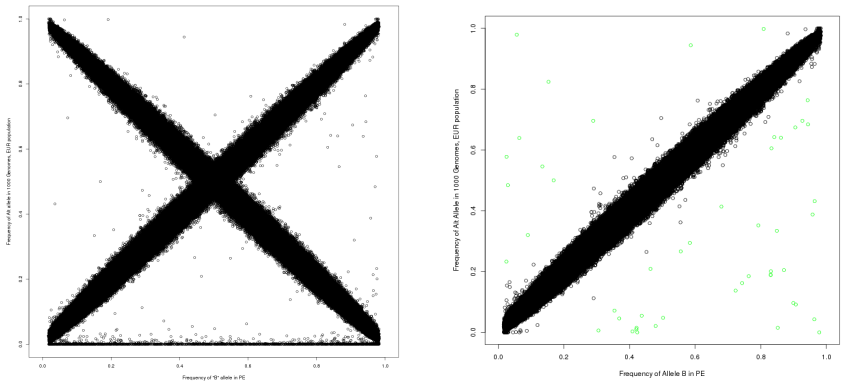
(a) (b)

Figure 4.A.16: Comparison of allele frequencies in the visceral leishmaniasis (India) study and in 1000 Genomes SAS population before alignment and filtering (left) and after (right).



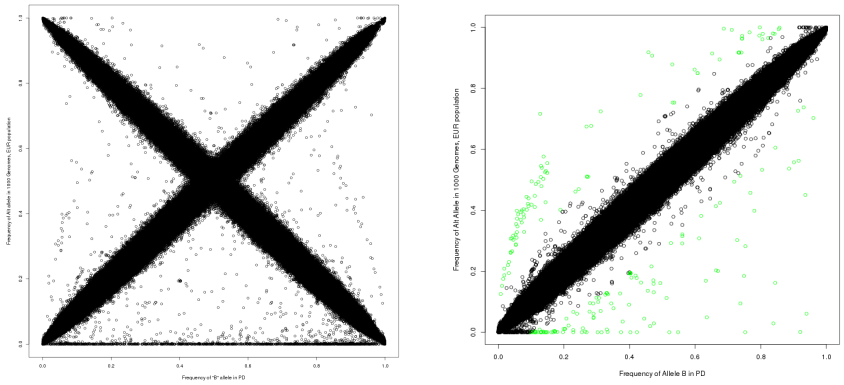
(a) (b)

Figure 4.A.17: Comparison of allele frequencies in the schizophrenia study and in 1000 Genomes EUR population before alignment and filtering (left) and after (right).



(a) (b)

Figure 4.A.18: Comparison of allele frequencies in the study of psychosis endophenotypes and in 1000 Genomes EUR population before alignment and filtering (left) and after (right).



(a) (b)

Figure 4.A.19: Comparison of allele frequencies in the Parkinson's disease study and in 1000 Genomes EUR population before alignment and filtering (left) and after (right).

4.B Q-Q plots for ABF calculations using different priors on the standard errors

Figure 4.B.1: Q-Q plots for the distribution of $\log_{10}$(approximate Bayes factors of the null data) data vs. $\log_{10}$(approximate Bayes factors of the real data) calculated using each of the 12 different correlation matrices, a prior $\sigma = 0.4$, and the cryptic correlation matrix calculated from effect size estimates excluding genome-wide significant SNPs and the HLA region. The red line is the line $y = x$. The number at the top of each plot is the proportion of $\log_{10}$(approximate Bayes factors) that were less than zero. In all cases, this number was at least 0.98.

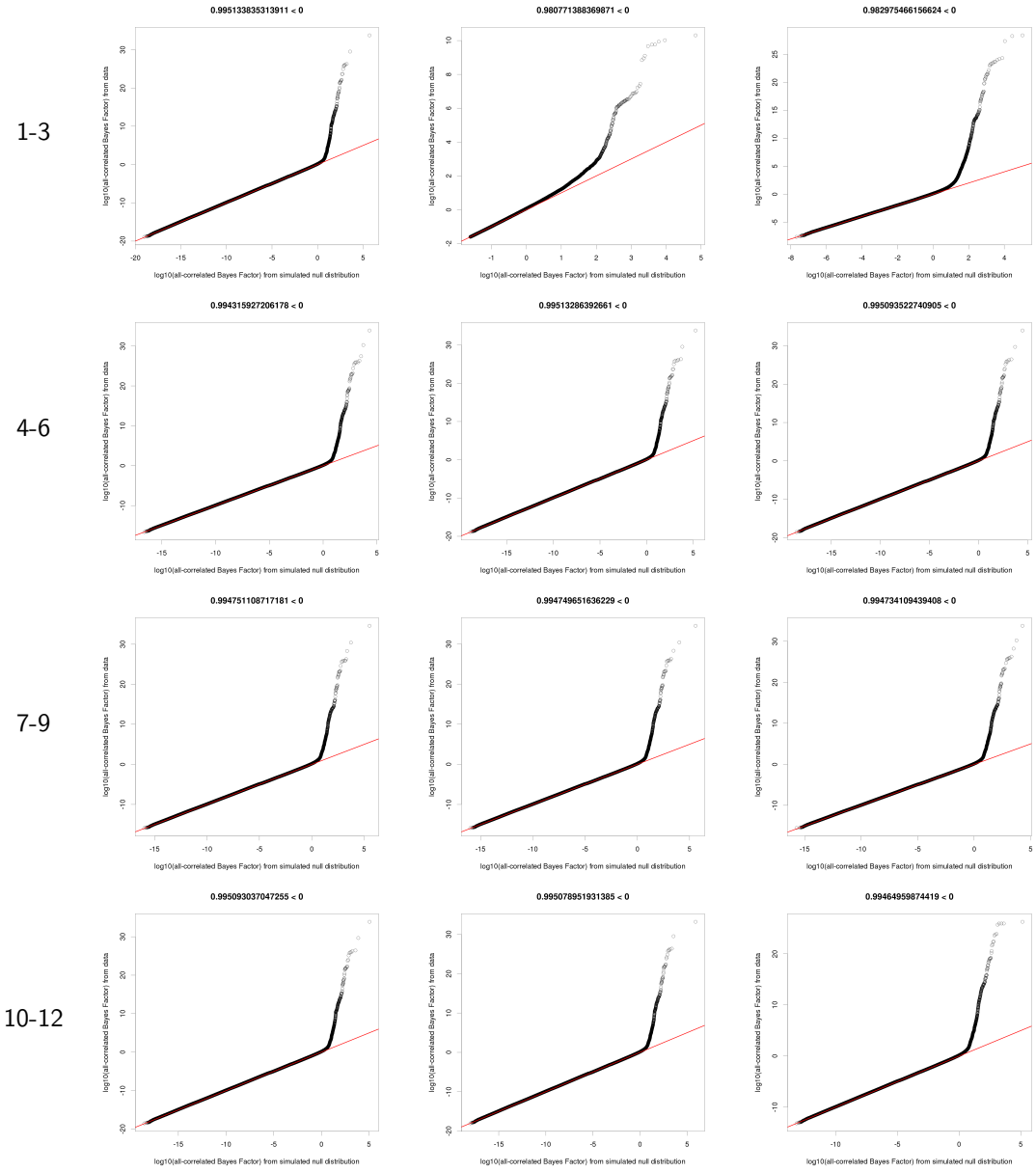
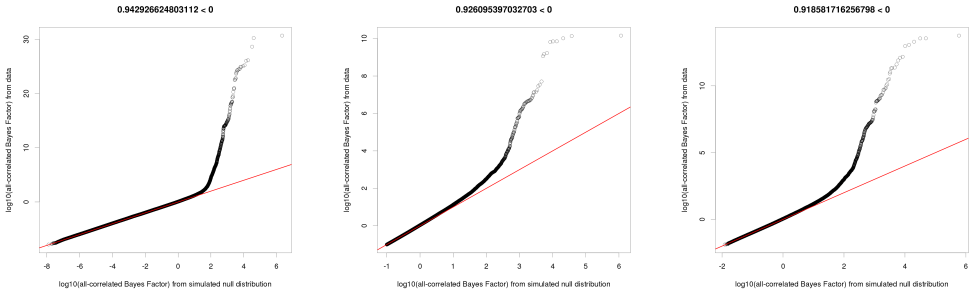
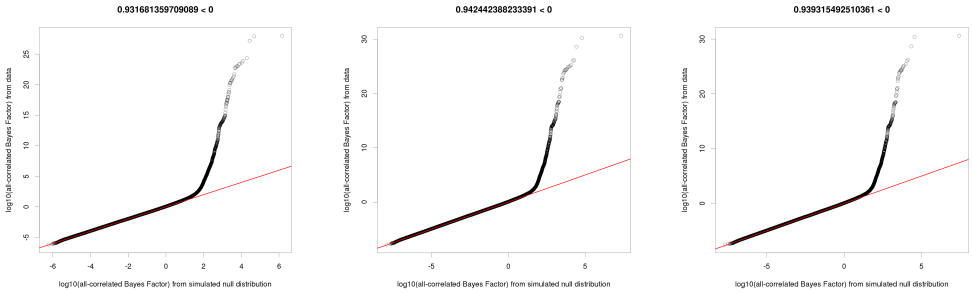


Figure 4.B.2: Q-Q plots for the distribution of $\log_{10}$(approximate Bayes factors of the null data) data vs. $\log_{10}$(approximate Bayes factors of the real data) calculated using each of the 12 different correlation matrices, a prior $\sigma = 0.1$, and the cryptic correlation matrix calculated from effect size estimates excluding genome-wide significant SNPs and the HLA region. The red line is the line $y = x$. The number at the top of each plot is the proportion of $\log_{10}$(approximate Bayes factors) that were less than zero. In all cases, this number was at least 0.91.

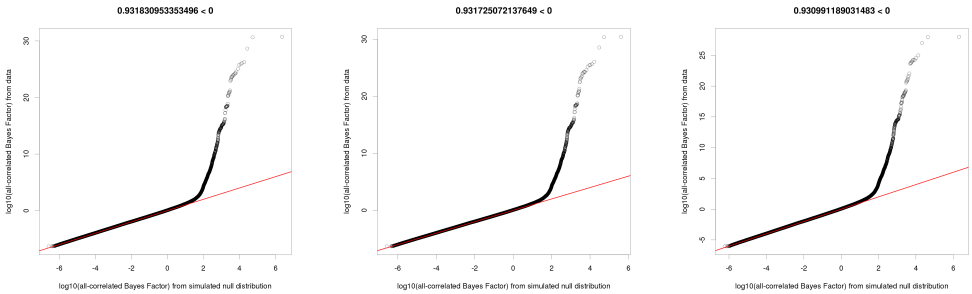
1-3



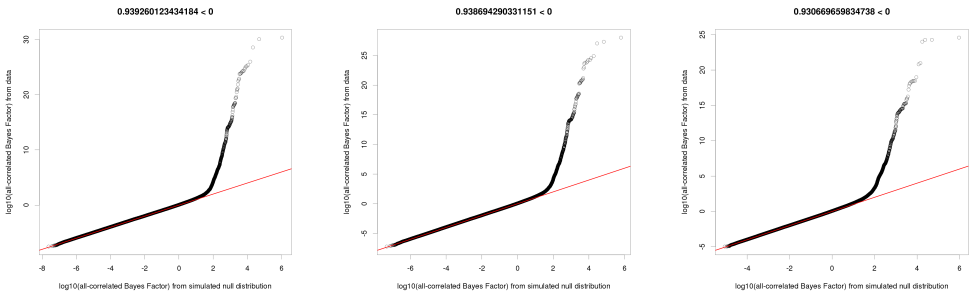
4-6



7-9



10-12



Chapter 5

Application to Immunobase

5.1 Introduction

In Chapter 3, we investigated the relationship between seven autoimmune diseases in a small dataset of only 107 SNPs. In Chapter 4, we applied our method genome-wide, but due to the number of cohorts (20), we could not apply the subset-exhaustive approach genome-wide, and instead used a staged approach to choose a set of SNPs whose patterns of association we wished to investigate. In doing so, we discovered how missing data affects the results of our meta-analysis (see Section 4.5 of Chapter 4). In this chapter, we use data from Immunobase [216], which has been imputed so that we have full summary data at each SNP for each study.

We also investigate an empirical prior distribution, estimated from applying the subset-exhaustive approach genome-wide. Because determining prior distributions on the models of association, as well as the ramifications of said distributions, could easily be a thesis in and of itself, we limit our scope to one particular type of empirical prior and compare the analysis results between this prior and the flat prior. We are interested in understanding what, if anything, our method can tell us about the relationships between diseases and the effect of this empirical prior on inference, as well as the relationship between effect sizes and the number of associations, which we discussed in Chapter 1, Section 1.2, as well as Chapter 3, Section 3.4.3.

5.1.1 Immunochip and Immunobase

Despite the explosion in genome-wide association studies discussed in Chapter 1, few variants were discovered that had clear functional relevance to disease risk. This created a demand for custom arrays that allowed for dense genotyping in regions shown to be associated with a trait in order to map the causal loci. One such array is the Immunochip, designed by the Immunochip Consortium, which was created to aid the fine-mapping of regions associated with immune-mediated diseases [19, 53]. The chip genotypes 195,806 SNPs and 718 insertion-deletions, providing dense genotyping of 186 regions of the genome. These regions were chosen based on the GWAS associations with Immunoglobulin A deficiency [217, 53] and with the following diseases:

- autoimmune thyroid disease
- ankylosing spondylitis
- Crohn's disease
- celiac disease
- multiple sclerosis
- primary biliary cirrhosis
- psoriasis
- rheumatoid arthritis
- systemic lupus erythematosus
- type 1 diabetes
- ulcerative colitis.

A recent study by Ellinghaus *et al.* [48] used genotyping data from the Immunochip to compare the genetic bases of five diseases: Crohn's disease, ulcerative colitis (often denoted collectively as IBD by the study's authors), ankylosing spondylitis, primary sclerosing cholangitis, and psoriasis. In addition to finding 27 novel disease loci and three new loci shared across diseases, they observed a high rate of comorbidity between IBD and primary sclerosing cholangitis, and found evidence that people with both diseases are genetically distinct from those who have only one. Additionally, they observed that the symptoms of both disorders were slightly different in patients with the comorbidity than in patients without. Such results highlight the potential of this chip to increase the understanding of the genetic architecture of autoimmune diseases and which loci influence the risk of multiple diseases.

Immunobase is a curated database of information on the loci, summary statistics, and papers related to the immune-mediated diseases whose known genomic associations were used in the creation of the Immunochip. It also includes information about juvenile idiopathic arthritis (JIA), and several other diseases known or suspected to have an immune-related etiological component: alopecia areata, inflammatory bowel disease (in addition to CD and UC), immunoglobulin E and allergic sensitization, narcolepsy, primary sclerosing cholangitis, Sjögren's syndrome, systemic scleroderma, and vitiligo [216]. It can be found online at `https://www.immunobase.org`.

With data downloaded from Immunobase, we generated an empirical prior on models of association and examined its effect on the results of our analyses. Due to the nature of the data, there were a great many associations between the diseases and SNPs in our meta-analysis; as a result, we neglect discussing specific genomic regions in favor of trying to learn about overall relationships between diseases. We also investigate further the question of whether a large effect size in one study correlates with further associations in another study, as we did in Chapter 3.

5.2 The data

The summary statistics for the Immunochip studies were downloaded from Immunobase [216]. We summarize the studies in Table 5.1 and note that while the studies of inflammatory bowel disease (comprising Crohn's disease and ulcerative colitis) looked across continental populations to replicate its findings, our data come from the discovery analysis, which was performed on people of European descent. Similarly, the ankylosing spondylitis and celiac disease studies contained non-European samples (noted in the table).

All studies were successful in finding new disease-associated loci, though often GWAS data was used in replication or to meta-analyze the Immunochip data [179, 219, 137, 203]. Except for the celiac disease study, all of the studies discussed which of the markers highlighted in their analyses overlapped with loci in other autoimmune diseases. For instance, in the type 1 diabetes study, the authors took the set of T1D associated markers and determined whether they were enriched for low p-values in studies of other autoimmune diseases. This analysis suggested genetic sharing between

Disease	Abbr.	Cases	Controls	Ref.
Crohn's disease*	CD	14,594	26,715	[137]
Ulcerative colitis*	UC	10,679	26,715	[137]
Ankylosing spondylitis**	AS	10,619	15,145	[162]
Multiple sclerosis	MS	14,498	24,091	[179]
Celiac disease†	CeD	12,041	12,228	[217]
Juvenile idiopathic arthritis	JIA	772	8,530	[218]
Rheumatoid arthritis	RA	11,475	15,870	[219]
Type 1 diabetes‡	T1D	12,010	12,331	[220]
Psoriasis††	PS	207	4,822	[203]
Primary biliary cirrhosis	PBC	2,861	8,514	[221]

* These diseases were part of the same study on inflammatory bowel disease (IBD) and use the same controls.

** 1,550 cases and 1,567 controls were people of East Asian descent.

† 229 cases and 391 controls were people of Indian (Punjabi) descent.

‡ These summary statistics were from a meta-analysis of a case/control analysis in unrelated samples, as well as family-based data.

†† These data are listed as coming from the UK cohort only, but the standard errors listed in the file of summary statistics are more consistent with a larger study that is more balanced between cases and controls.

Table 5.1: Summary of the immunochip studies from Immunobase. We give the disease name, its abbreviation ("Abbr."), the number of cases and controls used in the sample that generated the summary statistics (some studies had replication cohorts as well, whose summary statistics were not included in our data), and a reference to the published study ("Ref.").

type 1 diabetes and juvenile idiopathic arthritis, alopecia areata, primary sclerosing cholangitis, rheumatoid arthritis, and primary biliary cirrhosis, as well as possible enrichment with celiac disease and autoimmune thyroid disease [220]. The psoriasis study noted an epidemilogical relationship between psoriasis and celiac disease and reported 10 loci that overlapped between the two, as well as 10 other loci shared between celiac disease and Crohn's disease [221].

The markers listed in the summary files were aligned to 1000 Genomes by first matching on rsids, and—where this was not possible—matching on chromosome, base pair position, and minor allele, provided that the minor allele frequency (MAF) was ≤ 0.4. Triallelic variants and SNPs with complementary alleles and MAF > 0.4 were removed from the dataset. SNPs were aligned based on alleles in SNPs with non-complementary alleles and based on MAF in SNPs with complementary alleles.

In order to get full data at each marker, and to increase the number of markers at which we had data, summary statistic imputation was performed. This method takes advantage of the linkage disequilibrium between markers on reference panels to calculate the expected correlation between summary statistics based on summary

data at typed SNPs [222, 223]. This avoids performing imputation of genotypes in order to then perform association analysis, which is not only an extra step, but also requires individuals' genotypes, which are less easily shared due to privacy concerns, and which also increase the amount of storage space needed for the data by several orders of magnitude.

Summary statistic imputation was carried out in each study individually using DIST v.1.0.0 [222] with default parameters. This algorithm uses a reference panel to calculate correlations between the summary statistics at typed and untyped variants. This correlation matrix, along with the information for typed variants within an "extended window", is used in the conditional mean formula for multivariate normal distributions in order to estimate the effect sizes of untyped variants in a subset of the extended window. The reference data came from the 1000 Genomes EUR samples. Variants in the final dataset had an info score of at least 0.8 in a minimum of five of the studies [224]. Any SNPs with residual missing data were removed from our analysis. The final data set contained summary statistics for 774,202 markers across all 10 diseases. The data alignment and imputation were all performed by Luke Jostins.

Figure 5.1 shows the frequency of imputation info scores across all studies in this dataset, in bins of size 0.05. An info score of 1 indicates that the SNP was either genotyped directly or could be imputed from typed SNPs with perfect accuracy. We note that the AS and T1D studies seem to have the lowest rates of info scores in the 0.8-1 range, suggesting that imputation was less successful in these two studies than in the others.

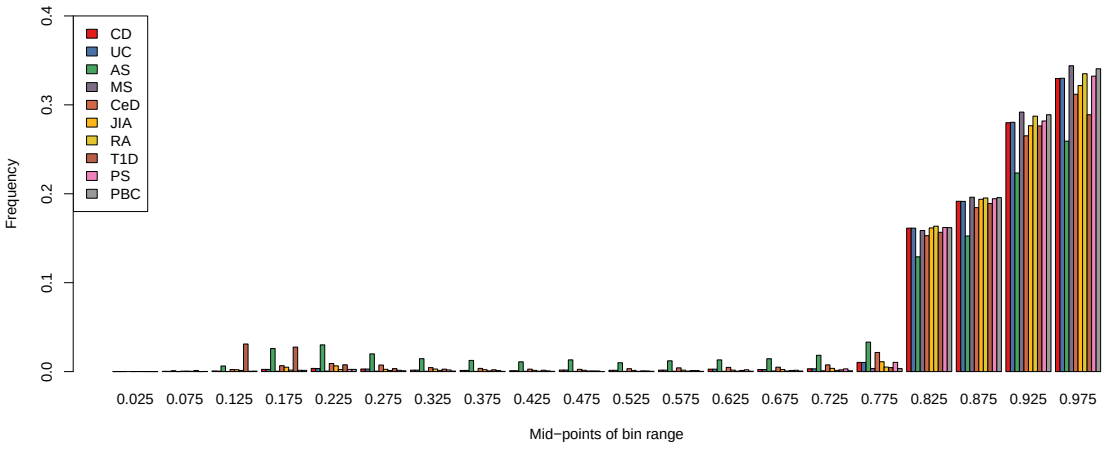


Figure 5.1: Histogram of imputation info scores across all studies, with high scores indicating more confidence in the imputation.

All studies showed negative and statistically significant ($p \ll 0.0001$) correlations between info scores and standard errors (the correlation coefficients were on the interval $[-0.3121, -0.0438]$), suggesting that lower info scores lead to higher imputed standard errors. The correlation between info scores and effect size estimates were generally closer to zero—the most extreme estimate was for the MS study, which had a correlation coefficient of -0.0524. The rest all had a modulus of < 0.015 and were not consistently statistically significant. Thus, the effect of uncertainty around imputation seems to be on the estimated standard error, which is intuitive; the less confident we are in the imputation, the more uncertainty there should be around the effect size estimate. Furthermore, both standard errors and imputation info scores are dependent on the minor allele frequency of the marker, which would induce a correlation between the two measures. Minor allele frequency has less effect on the point estimate, hence these have lower correlations with the info score, and are not always statistically significant.

5.3 Analysis

Many studies in this data set used shared controls—particularly from the WTCCC2 pool. Therefore, it was once again necessary to adjust for cryptic relatedness. We would have liked to populate the cells of the correlation matrix on the null model with the estimates of genetic covariance due to confounding calculated by LD score regression (see Section 1.4.3 in Chapter 1 and Section 4.3 in Chapter 4), but we could not because fewer than 200,000 SNPs in our data matched variants in their reference panel, meaning that the calculation was unlikely to be accurate and led to negative heritability estimates in some of the traits. Furthermore, the software was unable to perform 13 of the 45 calculations, citing invalid values in its internal processes. The empirical calculation of the correlation due to control sharing between UC and CD (using Zaykin's formula, Equation 2.11 in Section 2.4.2 of Chapter 2) is 0.318. LD score regression, however, estimates this value as 0.7226. Meanwhile, the empirical estimation of the correlation directly from the data (removing HLA and genome-wide significant SNPs) is 0.5065.

The discrepencies among these methods are troublesome. However, we chose to

follow the methodology of the WTCCC2 analysis in Chapter 4 (Section 4.3) and used the matrix generated from estimating the correlations from the effect size estimates directly. As we see in the example above, its values form a potential compromise between the high correlations calculated by LD score regression and the low ones estimated using Zaykin's formula. Furthermore, as we do not know (and have no way of estimating accurately) exactly how many samples were shared across studies, Zaykin's formula proves to be as impractical as LD score regression.

Figure 5.1: Heatmap of the cryptic correlations calculated between all pairs of diseases. We required that the be at least 0.05 and significant at a Bonferroni corrected level of 0.01—however given the low standard errors associated with the correlation coefficients we estimated, the requirement that the coefficients be at least 0.05 was the more stringent criterion.

We used the cryptic correlation matrix when performing the subset-exhaustive approach on the Immunobase data using priors $\rho = \{0, 0.5, 0.96\}$, $\sigma = \{0.1, 0.2, 0.4\}$.

We performed an initial analysis using a flat prior over all possible models of association, and then used the results to calculate prior probabilties for each model of association (see Section 5.3.1 below). We performed the analysis again using these new priors.

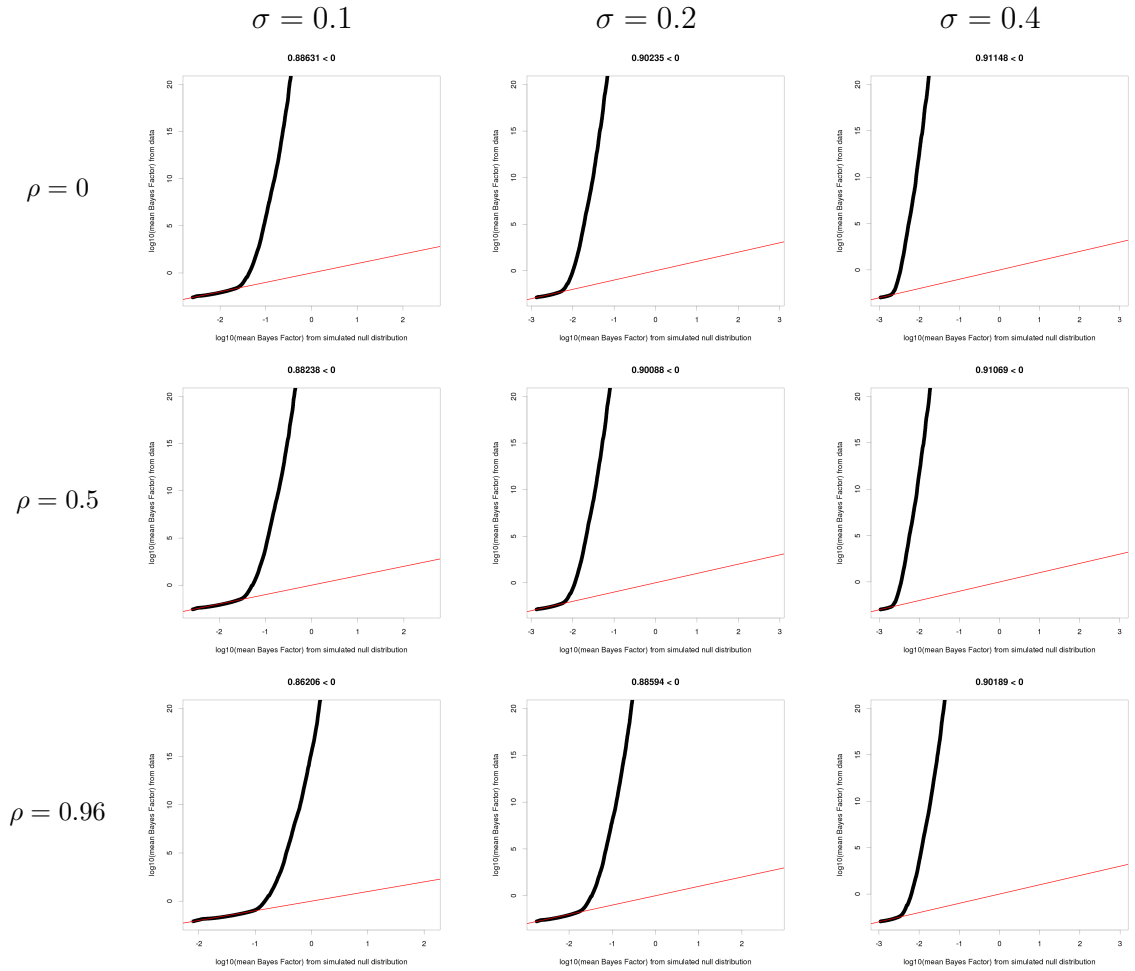
5.3.1 An empirical prior over all models of association, calculated genome-wide

There are infinite ways to calculate an empirical prior on models of association; we calculated and explored the effects on results of just one of them. Our analysis always includes a prior on models of association, even when we do not explicitly acknowledge one. The prior that is naturally assumed by our subset-exhaustive approach is a flat prior, where every model of association is as likely as every other model. This means that once we have calculated the ABFs over every model at a given SNP, we do no further reweighting of the results. We used such a prior in our initial analyses; we calculated the subset-exhaustive ABFs using varying values of ρ and σ and that was all.

As in Chapter 4, we compared the mean ABFs calculated from the real data against a null distribution generated from the cryptic correlation matrix and the standard errors of each disease at each SNP. For this analysis, we excluded the extended MHC region on chromosome 6 [225]. We present the quantile-quantile plots in Figure 5.2 and note that the mean ABFs calculated from the data very quickly deviate from the null distribution. This suggests an inflation of our test statistic, which could indicate that we have not properly accounted for the cryptic covariance between the studies. It is also possible that because the genomic regions targeted by the Immunochip were selected based on previously known associations in nine out the ten studies in our meta-analysis, the standard assumption that associations with any of our traits will be rare do not apply. We also note that while the number of ABFs that are smaller than 0 increases with the prior σ, it appears that this also corresponds to deviating from the null distribution at lower quantiles than analyses with smaller values of σ.

To calculate our empirical prior, we used the results of our initial analysis, which assumed this flat prior over all models of association. For each SNP, we calculated

Figure 5.2: Q-Q plots for the distribution of $\log_{10}$(mean ABFs of the null data) data (x-axis) vs. $\log_{10}$(mean ABFs of the real data) (y-axis) calculated using each combination of ρ and σ and the cryptic correlation matrix calculated from effect size estimates excluding genome-wide significant SNPs and the HLA region. The red line is the line $y = x$. The number at the top of each plot is the proportion of $\log_{10}$(mean ABFs) that were less than zero.



$2^{10} = 1024$ approximate Bayes factors. We normalized the ABFs to get a posterior distribution of probabilities on each model, and then summed the posterior probability on each model over all SNPs and divided by the number of SNPs to get the genome-wide prior probability for each model. That is, if we let i index the SNPs ($i \in \{1, 2, \ldots, 774202\}$) and j index the models of association ($j \in \{1, 2, \ldots, 1024\}$). Let $\text{ABF}_{i,j}$ be the approximate Bayes factor of the the jth model at the ith SNP in the intial analysis. Then initial posterior probability on the jth model at the ith SNP, $p_{i,j}$ is

$$p_{i,j} = \frac{\text{ABF}_{i,j}}{\sum_j \text{ABF}_{i,j}}, \tag{5.1}$$

and we calculate the prior probability of the jth model as simply

$$p_j = \frac{\sum_i p_{i,j}}{774202}. \tag{5.2}$$

Figure 5.3 shows the distribution of prior probabilties for the 20 models with the most prior density over all possible values of ρ and σ. We note that the null model (furthest left) has the most prior weight in all but one of the analyses*, though the range is considerable between analyses that assume small true effects ($\sigma = 0.1$) and large effects ($\sigma = 0.4$). We also note that contrary to the standard GWAS assumption that true genomic associations with the trait are rare, which would correspond to large amount of prior density on the null model [93], we see that the empirical prior calculated using the whole genome puts only a modest amount of weight on the null model—between 0.006 and 0.222, depending on the analysis parameters. Given that the Immunochip was designed to fine-map regions associated with various immune-mediated diseases—nine of which are part of our meta-analysis—this is perhaps not too surprising.

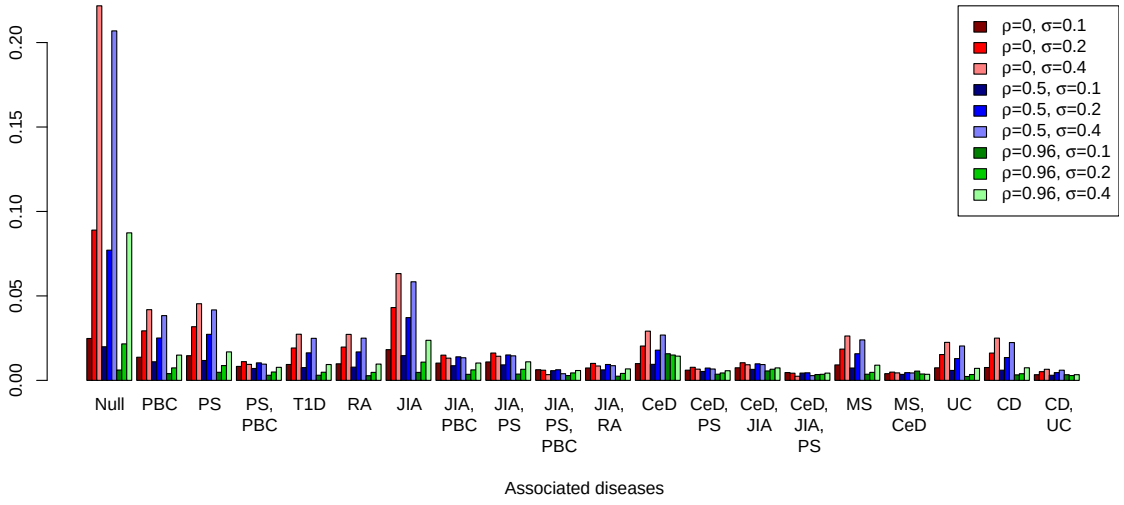


Figure 5.3: Distribution of prior probabilties over all Immunobase SNPs on each combination of ρ and σ used in the analysis. Bars are colored by the parameters used in each analysis. We show the 20 models with the most prior density.

We also note that the non-null models that show increased prior probabilities compared to the flat prior are all for models associated with a maximum of three diseases. This suggests that despite the known genetic risk shared among autoimmune diseases, that these risks tend to be shared between subsets of autoimmune diseases,

*The analysis with $\rho = 0.96$ and $\sigma = 0.1$ is the exception—the highest prior weight in that analysis was on the model where only CeD was associated.

rather than the class of diseases as a whole. We note among our priors the prevalence of PBC, PS, JIA and CeD in the models of multiple associations, while T1D only ever appears on its own. None of these highly weighted models ever include AS—even on its own—though this prior does increase the weight on the AS-only model compared to the flat prior in all but one of the analyses (the exception was where $\rho = 0.96$, $\sigma = 0.1$). This may be a result of the relatively lower success of imputing AS summary statistics (see Section 5.2).

Taken together, this suggests that in most analyses, we should expect that the single most likely model is the null model, and that when a true effect exists, it is most likely to be associated with exactly one disease. When a SNP is associated with multiple diseases, we expect the associations to be some combination of {PS, PBC, JIA, CeD}, of {MS, CeD}, or of {CD,UC}. We expect associations with AS comparatively rarely.

It is notable, if unsurprising, that the analysis where correlations between studies is expected to be high ($\rho = 0.96$) and the true effect sizes are expected to be small ($\sigma = 0.1$) was the only one for which the model with the most prior weight was not the null model. For this set of parameters, the model associated with only celiac disease was the one with the highest prior weight (0.0158). Table 5.1 shows the number of models whose prior weights increase under this empirical prior compared to the flat prior. We see that the number of models with increased prior weight goes up as ρ increases and goes down as σ increases.

Analysis	**Number of models**
$\rho = 0, \sigma = 0.1$	262
$\rho = 0, \sigma = 0.2$	168
$\rho = 0, \sigma = 0.4$	114
$\rho = 0.5, \sigma = 0.1$	291
$\rho = 0.5, \sigma = 0.2$	191
$\rho = 0.5, \sigma = 0.4$	125
$\rho = 0.96, \sigma = 0.1$	476
$\rho = 0.96, \sigma = 0.2$	357
$\rho = 0.96, \sigma = 0.4$	233

Table 5.1: Table of the number of models whose prior weight, p, was such that $p > \frac{1}{2^{10}}$ for each set of ρ and σ used for analysis.

The results in Figure 5.3 and Table 5.1 are intuitive—when correlations between true effect sizes are assumed to be high, then finding an effect in one study increases the likelihood of finding further effects. Similarly, if true effect sizes are assumed to be small, then we expect that more studies will show true effects, thus more models will have increased prior probability, and there will be less prior weight on the null model than in analyses where true effect sizes are expected to be larger.

5.4 Results

Here we present the results of our analyses, and try to interpret their meaning regarding the relationships among the Immunochip diseases.

5.4.1 Change in posterior probabilities of models of association

Having defined the a prior based on the posterior probabilities of association on each model from the intial analysis in Section 5.3.1, we now examine how these priors affect the posterior probabilities on different models of association when they are applied to the analysis. Figure 5.1 shows the mean posterior weights on the models with the 20 highest mean changes (positive or negative) for each model. These were calculated by subtracting the mean posterior weights on each model from the analysis using the flat prior from the mean posterior weights on each model from the analysis using the the empirical prior.

Unsurprisingly, given the prior weights used, the largest changes in posterior weight occurred for the null model. Once again, in all but one of the analyses (where $\rho = 0.96$ and $\sigma = 0.1$), this was the largest change in either direction. Also expected are the increased posterior probabilties on models with just one or two associations. We also note how rapidly the modulus of the change in posterior probability decreases among these models, which is similar to the rapid decline in prior weights seen in Figure 5.3.

However, some of these changes are very surprising. We note, for instance, that despite some of the models having increased prior weights in all analyses, some of the posterior changes are negative (or close to 0), particularly when $\rho = \{0, 0.5\}$ and

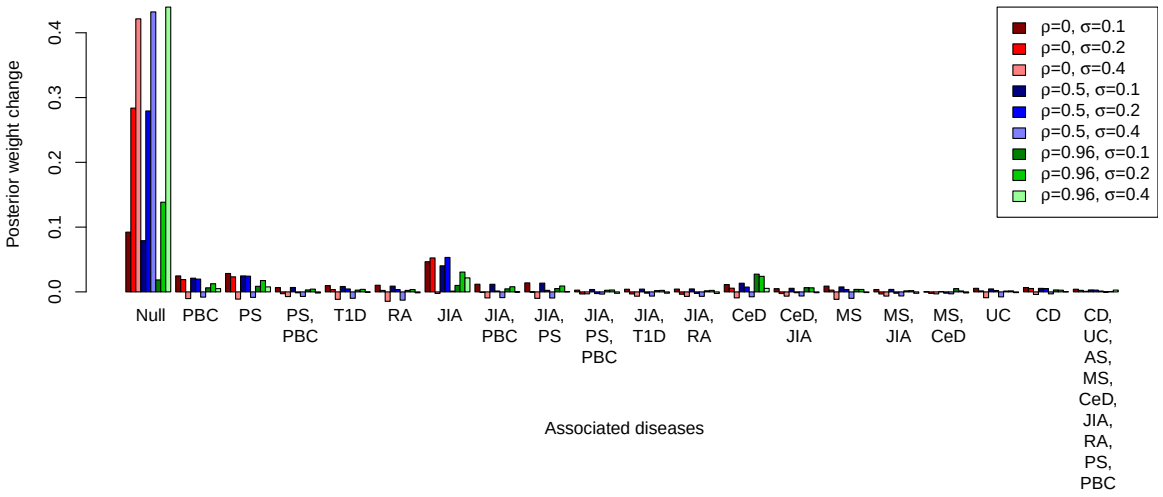


Figure 5.1: Difference in mean posterior weights per analysis between the analysis that used the genome-wide empirical prior and the analysis that used a flat prior. Bars are color-coded by the analysis parameters.

$\sigma = 0.4$. We also note that, while there is a good deal of overlap in the models shown in Figures 5.3 and 5.1, there are some noticeable differences in which models show the greatest change in posterior probability and which had the most prior weights. For instance, the models {CeD, PS}, {CeD, JIA, PS}, and {CD, UC} are in the top 20 for increased prior weight, but are not for posterior weight change, and are replaced by {JIA, T1D}, {MS, JIA}, and {CD, UC, AS, MS, CeD, JIA, RA, PS, PBC}. These differences are undoubtedly due the combined effect of reweighting every model of association at every SNP across the genome.

5.4.2 Principal Component Analysis

In Chapter 3, Section 3.4.4, we performed a principal component analaysis (PCA) using the marginal probabilities on all 107 SNPs in the data set. We would like to be able to interpret disease clusters that are close together on a plot of the first two principal components to correspond to similar genetic profiles; however, we could not verify our PCA results for the data in Chapter 3. In particular, we had no controls in the form of diseases we would expect to see cluster together and diseases we would expect to see that are highly differentiated from the others. With the Immunobase data, we have a positive control in the form of the two disorders that make up inflammatory bowel disease: Crohn's disease and ulcerative colitis.

As in Section 3.4.4, we conducted our principal component analysis by per-

forming a singular value decomposition on the matrix of marginal probabilities of association between each disease at each SNP. This gave us a matrix of loadings, which we stored. As before, we sampled 1000 models from the posterior distribution at each SNP, and used the pre-calculated loadings to project them onto the first two PCs, allowing us to visualize the uncertainty around the relationships between and among diseases, due to our uncertainty about the correct model of association at each SNP.

The Immunobase data presents the challenge of determining which SNPs to use to calculate the principal components. We face three main decisions: how do we ensure the independence of our SNPs? Should SNPs from the extended HLA region be included? Should SNPs be chosen randomly, or should we preferentially choose SNPs that show associations in our ABF analysis? In addition to these, we are interested in what, if any, effect different values of ρ and σ have, as well as the the effect of using the two different priors on models of association.

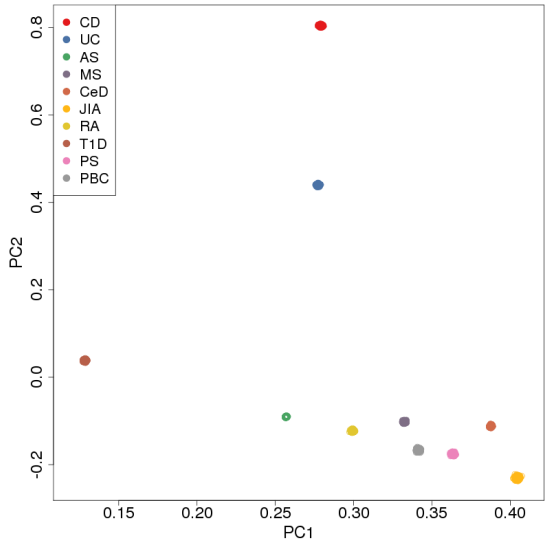
Figure 5.2 is a typical example of the differences we observed when the only variable between PCA analyses is the number of SNPs included in the singular value decomposition. We observe that as the parameters for inclusion become more stringent (i.e. when the SNPs are required to be further apart in both genetic and physical distance), fewer SNPs are included in the analysis, and the disease clusters become larger and closer together. Additionally, removing the extended HLA region from the analysis (about 5.8% of all SNPs in the dataset) made a negligible difference to the sizes and positions of the clusters, so we left the SNPs from that region in. When all SNPs are included, we see that we are able to plot perfectly each disease in two dimentional space, and there is no overlap between clusters. Essentially, when using all the association data, each disease can be perfectly characterized.

We note the similarity between the Figures 5.2b and 5.2c, despite the former using nearly twice the number of SNPs in its analysis as the latter. This was consistent across all analyses, resulting in a decision to focus on generating results using the parameters in Figure 5.2c, where all SNPs are at least 1 centiMorgan and 100 kilobases apart. This ensures a dataset of SNPs that are reasonably independent, but without removing so many SNPs that the analysis cannot differentiate among diseases.

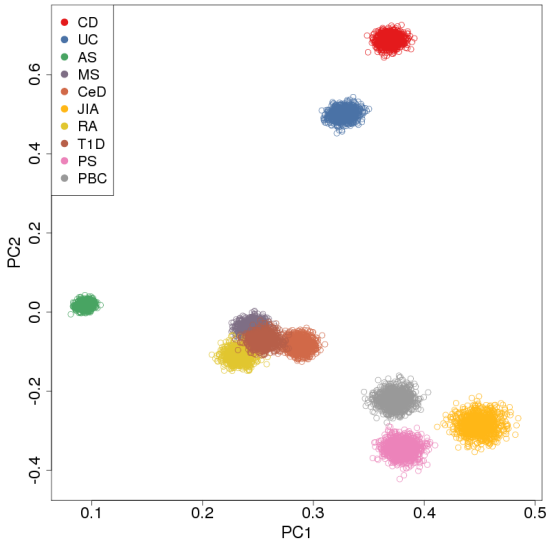
Our next decision concerned whether to choose SNPs based on evidence of asso-

Figure 5.2: PCA cluster plots generated by performing singular value decomposition with differing numbers of SNPs. Otherwise, all sets shown were identical, using the results from the analysis $\rho = 0$ and $\sigma = 0.2$ with the empirical prior on models of association. SNPs were chosen based on their mean ABFs across all models and all priors. The selection process chose the SNP with the highest mean ABF and, when SNPs were removed, removed those according to the parameters indicated next to the plot.

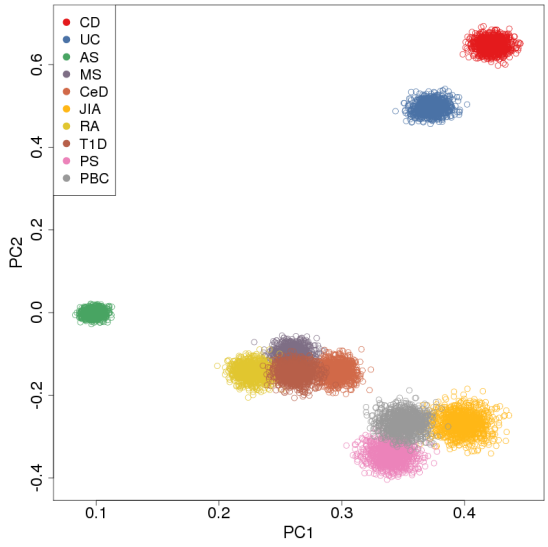
(a) All 774,202 SNPs in the dataset used.



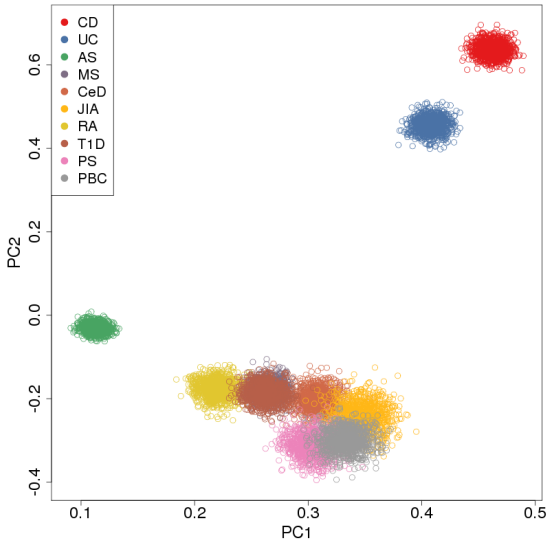
(b) All 5,067 SNPs are at least 0.5 cM and 50 kb apart.



(c) All 2,688 SNPs are at least 1 cM and 100 kb apart.



(d) All 1,357 SNPs are at least 2 cM and 200 kb apart.

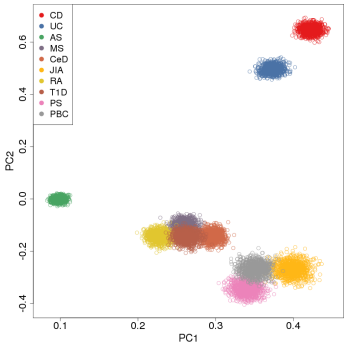


ciation or to choose them randomly. We examine the effects of choosing SNPs based on level of association in Figure 5.3. In Figure 5.3a, SNPs have been chosen using the algorithm that prioritizes SNPs with large amounts of evidence of association. Figure 5.3b shows results of our PCA analysis for SNPs that have been chosen randomly, and Figure 5.3c shows the PCA analysis performed on the simulated null data used

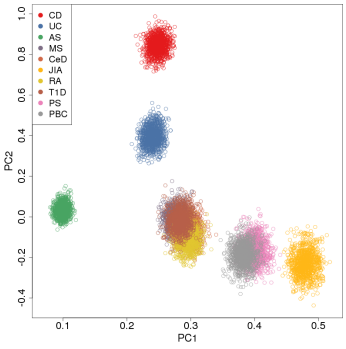
to create the Q-Q plots in Figure 5.2. We observe that there is structure in even the null data and that it separates JIA from the other studies. This may be the result of the large standard errors associated with JIA compared to the other studies. Indeed, it appears that clusters in the PCA of null data seem to correlate with the study sizes listed in Table 5.1, with studies of similar size lying closer to one another.

Figure 5.3: PCA cluster plots generated by performing the singular value decomposition with different criteria for SNP inclusion for 5.3a and 5.3b, while the data used in 5.3c were selected randomly from among the set of SNPs with effect sizes simulated under the null. All data were analyzed using $\rho = 0$ and $\sigma = 0.2$ with the empirical prior on models of association. All SNPs in the analysis were at least 1 cM and 100 kb apart.

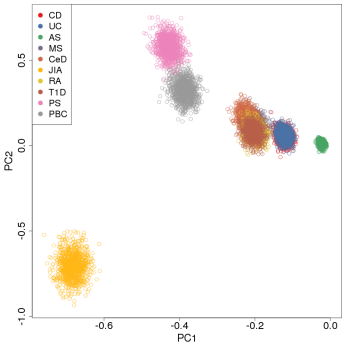
(a) SNPs chosen based on evidence of association.



(b) SNPs chosen randomly.



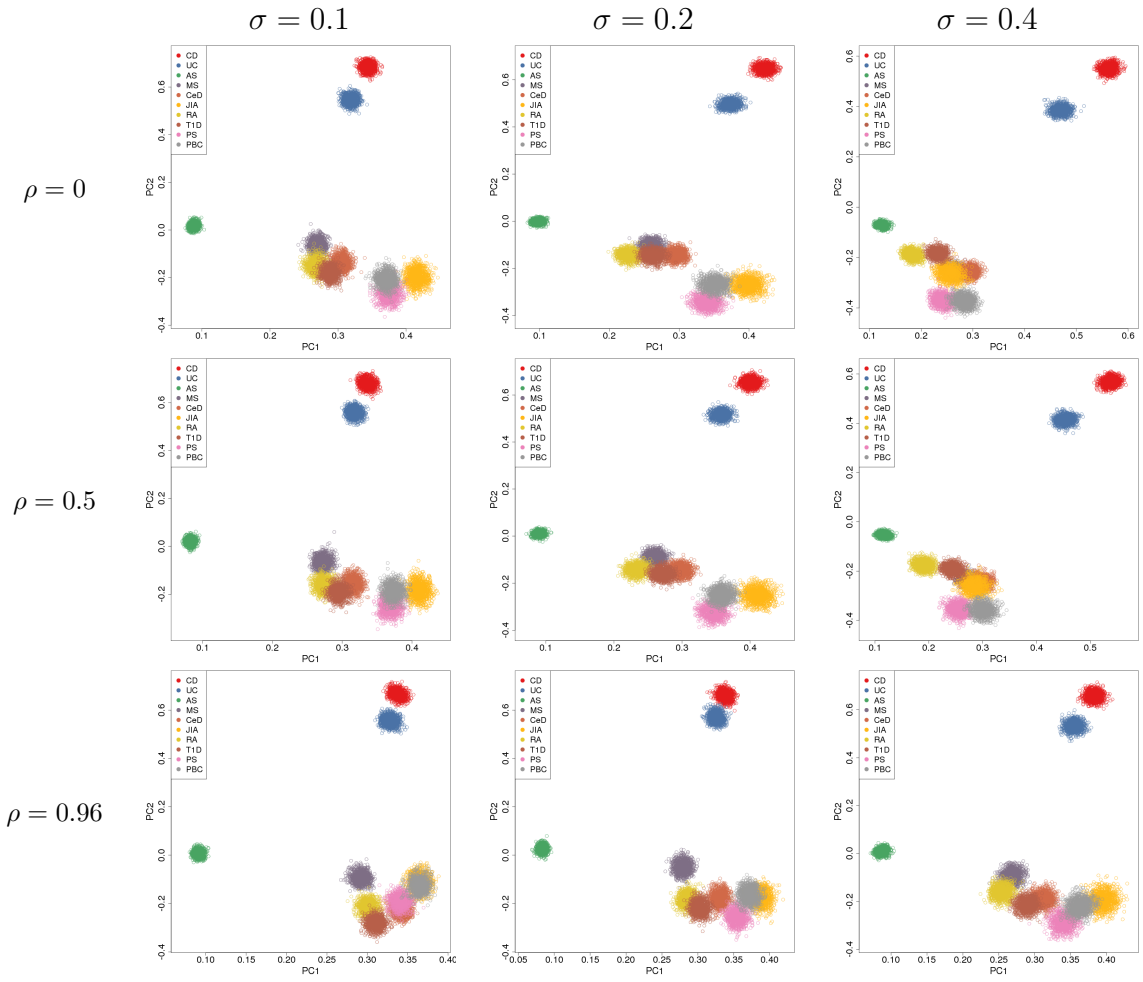
(c) SNPs from simulated data with no true effects.



We also note that Figure 5.3b seems to be an intermediate step between Figures 5.3a and 5.3c. The clusters of 5.3b are a little more spread out than those of 5.3c, but neither one shows the separation of diseases that 5.3a does. We speculate that the structure observed in the null data is the result of the way standard errors correlate between studies, as well as the cryptic correlation matrix we used to generate the data and note that there does not seem to be any gain from choosing SNPs randomly from the real data, and indeed, there is a good argument for not doing so: by preferentially selecting SNPs that show evidence of association in our analyses (i.e. based on the mean ABF across all models and for all values of ρ and σ), we ensure that our PCA is conducted based on the markers that are the most relevant. However, this method could introduce a bias if SNPs with high ABFs are systematically more likely to be correlated with a particular subset of diseases, while SNPs that are informative about other relationships between diseases show lower ABFs, and are potentially removed from the set included in the PCA.

Regardless of the exact parameters used to perform the singular value decompo-

Figure 5.4: Plot of all PCA using the results of the various ABF analysis with the genome-wide prior on models of associations.



sition and thus the PCA, we find that, except in Figure 5.2a, the results are consistent in separating AS from all the other diseases, and in separating UC and CD from the other diseases, with each as the other's closest cluster. Additionally UC is always between CD and the other clusters. We also see that CeD, MS, RA, and T1D often cluster near one another, while PBC, PS, and JIA often form another group. However, in Figure 5.4, which shows the PCA plots for all values of ρ and σ under the genome-wide empirical prior, we see that when ρ or σ is large (equal to 0.96 or 0.4, respectively), these two sets of diseases tend to move closer together, with CeD forming a bridge between the two. We note that, like Chang *et al.* [161], we find that MS, RA, and T1D cluster (relatively) close together, though unlike them, we do not find that PS clusters closely with CeD and MS.

We recall from Section 5.3.1 that the prior put more weight on models of association containing combinations of {PS, PBC, JIA}, as well as models containing

{CeD, JIA, and PS}. These relationships appear to be reflected in the plots in Figure 5.4. However, we note that the PC analysis suggests that CeD is more closely related to MS, RA, and T1D than it is to PS, PBC, and JIA, contrary to the prior and to the change in posterior model weights discussed in Section 5.4.1. We note that our findings are more or less consistent with—though not necessarily confirmed by—those found by Zhernakova *et al.* [47] and Parkes *et al.* [53]

Additionally, the results of the PCA calculated using the analyses with the flat prior on models of association showed similar results, which reassures us that these putative relationships are not due solely to our choice of prior. The chief difference between PCAs performed using results from the flat prior and the genome-wide empirical prior were that the {CeD, MS, RA, T1D} cluster and {JIA, PBC, PS} cluster were often plotted closer to one another under the flat prior, and did not have such extreme clustering or movement towards the AS cluster when $\sigma = 0.4$ and $\rho = 0.96$. These plots can be found in Figure 5.A.2 in Section 5.A of the Appendix.

5.4.3 The relationship between large estimated effect sizes and the number of associations

Once again, we turn to the question of whether a large effect size in one study at a given SNP predicts more associations at that same SNP. Immunobase is an interesting dataset with which to explore this question because the diseases in the analysis are known to share genetic variants, and because the regions of the genome interrogated by the Immunochip were chosen due to their associations with these diseases. In essence, these circumstances are so favorable to the hypothesis that strong effects in a single study correlate with more associations overall, that if these data are not consistent with it then it is hard to imagine that there are any circumstances under which this hypothesis would be true.

We analyze this question in each of the results of the exhaustive ABF analyses we have performed. For each analysis, we calculated the posterior probabilties of association across all possible models at each SNP, using Equation 3.6 from Section 3.4.3 of Chapter 3, though this time with $n \in \{1, \ldots, 10\}$.

We next determine the maximum effect size at each SNP. Because we are interested in the relationship between extreme effect sizes and expected number of asso-

ciations (as opposed to effect sizes that are explicitly greater than zero), we examine the absolute values of the effect size estimates in our data. Table 5.1 shows that the most extreme effects are not distributed uniformly over all the studies. Additionally, we recall from Table 5.1 in Section 5.2 that study sizes vary wildly among our data, and we may end up choosing extreme effect size estimates which are actually not significantly different from 0.

The logical conclusion is to examine the relationship between z scores and the estimated number of associations. Furthermore, the larger standard errors associated with poorly imputed summary statistics provide a natural penalty, making them less likely to be included as the largest z score at a SNP. We also see in Table 5.1 that, unlike maximum effect size estimates, maximum z scores are more evenly distibuted across all diseases, ensuring that the results of this analysis are not dominated by the effects of a single study.

Disease	**Prop. ($\hat{\beta}$)**	**Prop. (z)**
JIA	0.3909	0.0898
PBC	0.1885	0.0847
PS	0.1851	0.1103
CeD	0.0530	0.0920
T1D	0.0451	0.0977
RA	0.0423	0.0831
CD	0.0387	0.1288
MS	0.0307	0.1035
UC	0.0256	0.1103
AS	3.427×10^{-5}	0.0998

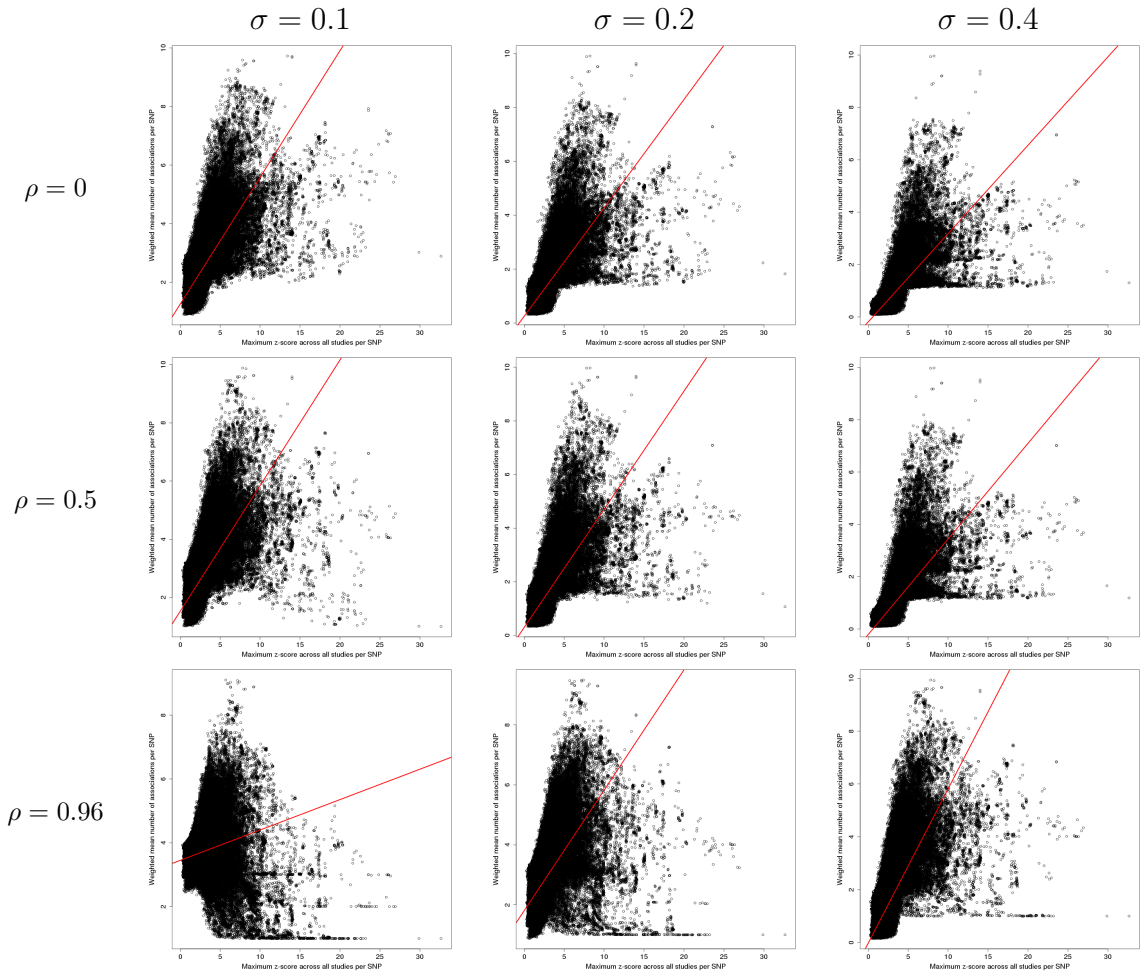
Table 5.1: Proportion of the time that the effect size estimate—"Prop. ($\hat{\beta}$)"—for each study was the most extreme of all studies, and the proportion of times that the z score—"Prop. (z)"—for each study was the most extreme of all studies. Results are listed in descending order by the proportion of $\hat{\beta}$ that were the largest for each study.

For each ABF analysis (using the genome-wide empirical prior), we plot the estimated number of associations against the maximum z scores, and draw the least squares regression line for the data. We note that in all but one of the studies (where $\rho = 0.96$ and $\sigma = 0.1$), the slope of the regression line lies on the interval of $[0.337, 0.0.584]$, indicating that in most analyses, between 1.71 and 2.96 unit changes in z score are required to increase the expected number of associations by 1. For each analysis, we report the estimates of the slope and their associated standard

errors in Table 5.2. The analogous figure and table for analyses that used the flat prior can be found in Section 5.B of the Appendix. In all cases, the p-values were very small ($p \ll 0.00001$).

Our question is a difficult one to answer because of the differences in power between the studies, as well as the fact that our method's power to detect true effects depends on the prior. Any effect that can be detected when $\sigma = 0.4$ is also large enough to be detected when $\sigma = 0.1$, so using $\sigma = 0.1$ results in a prior that is always at least as powerful as $\sigma = 0.4$, though it may call more false positives. As a result, when $\sigma = 0.1$, we call more associations in general, shown by the the fact that the regression intercepts in Figure 5.5 are negatively correlated with σ. Additionally, unlike the data in Chapter 3, we have probably counted some of the mean associations multiple times due to the LD structure between SNPs in the Immunobase data.

Figure 5.5: For each analysis, the estimated number of effects (y-axis) plotted as a function of the maximum z score (x-axis), per SNP. The red line is the least squares regression line calculated from the data. All analyses used the genome-wide empirical prior.



Analysis	Slope	SE
$\rho = 0, \sigma = 0.1$	0.4325	0.0005
$\rho = 0, \sigma = 0.2$	0.4023	0.0004
$\rho = 0, \sigma = 0.4$	0.3373	0.0003
$\rho = 0.5, \sigma = 0.1$	0.4320	0.0005
$\rho = 0.5, \sigma = 0.2$	0.4406	0.0005
$\rho = 0.5, \sigma = 0.4$	0.3655	0.0003
$\rho = 0.96, \sigma = 0.1$	0.0953	0.0004
$\rho = 0.96, \sigma = 0.2$	0.4018	0.0006
$\rho = 0.96, \sigma = 0.4$	0.5839	0.0006

Table 5.2: Table of the regression slopes and their standard errors each set of ρ and σ used for analysis (using the genome-wide empirical prior).

Taken together, we see that there is evidence in our analysis that having an extreme effect in one of the studies at a given SNP predicts more associations at that SNP. The analysis where $\rho = 0.96$ and $\sigma = 0.1$ is a special case that merits some attention. This prior assumes that true effects are highly correlated among all diseases and that these effects will be small. A dataset for which these conditions hold would not be an ideal one in which which to test our hypothesis that extreme effects correlate with more associations, because these conditions explicitly assume that effects are not extreme in any study. This makes it harder for sets of studies with heterogeneous sizes and directions of effect to be correctly identified as being associated. This may be the reason that the regression slopes for the analyses where $\rho = 0.96$ tend to be lower than the slopes where ρ takes a lower value. Similarly, within the set of slopes for a given value of ρ, the slope for the analysis where $\sigma = 0.1$ is the lowest. The low regression slope estimated using this particular set of priors seems to be the result of the interaction of these two trends.

5.5 Discussion

In this chapter, we have applied our method to a third and final dataset of immune-mediated diseases that have been typed on the Immunochip. Summary statistic imputation has given both a larger dataset in terms of the number of markers, and complete data at each marker. Additionally, the fact that there are 10 studies in this meta-analysis makes performing the subset-exhaustive approach tractable.

However despite these improvements over previous datasets, we struggled to verify our estimates of the cryptic correlation between the effect sizes of each disease. While we could still estimate these correlations from null SNPs that were not in the HLA region, the few estimates we were able to obtain using LD score regression differed so wildly from our empirical estimates, as well as from those calculated using Zaykin's analytic formula, that it was impossible for there to be any sort of consensus in the correlation coefficients across all methods.

To add to our concern, our quantile-quantile plots show that the distribution of mean ABFs calculated from the data very quickly deviates from the null distribution; the speed with which it deviates appears to be positively correlated with the prior value of σ. While it may be that this particular dataset contains many true signals across our studies, we cannot guarantee that our test statistic is not inflated, which may affect our other results—for instance by causing us to choose SNPs with spuriously high ABFs in our principal component analysis over those with true signals, or by causing us to overestimate the expected number of associated diseases in our regression analysis.

The dataset itself had a few oddities, as well. Two of the studies included people of non-European descent, despite the majority of the cohorts being sampled from Europeans. One of the disadvantages of working only with summary data is that it is impossible to tell what effect—if any—this has on the effect size estimates and standard errors that we use in our analysis. Theoretically, if the differences in genetic background are properly accounted for in the statistical model, then there should be no problem. Alternatively, it may be that the association testing was performed in each cohort separately and the meta-analyzed results are what were reported. However, we cannot know from the data alone what was done and have to take it on faith that the estimates we receive are free of biases related to population structure or other confounders. Additionally, we noticed a discrepancy between the stated sample size for psoriasis and the reported standard errors; we could not verify exactly where our psoriasis data came from, but it appeared to be from a larger study than the one claimed by Immunobase.

Nevertheless, we used these data to investigate the feasibility of applying our method to learn about the relationships between traits in the meta-analysis. We

estimated a prior on models of association using the results of an analysis with a flat prior. We should note that by doing this, we have technically used the data twice and our prior is no longer a prior under the traditional definition of the term. However, empirical Bayesian methods are common and have been applied to genetics before [226, 227, 228]. While we cannot say for certain whether the use of this empirical prior leads to more accurate results over what we find when using a flat prior—we would need to carry out more work on simulated datasets to establish this—it does seem that the empirical prior boosts signal. For instance, when comparing the results of the PCA with the empirical prior against those without, it was often easier to differentiate between disease clusters (by visual inspection) in the analyses that used the empirical prior than in those that did not. Additionally, in our analyses of the relationship between extreme effect sizes and estimated number of associations, the slopes of the regression lines were all higher when the empirical prior had been used.

As for the results themselves, we note that our PCA separates Crohn's disease and ulcerative colitis from the other studies, which is what we were hoping to see, as these two diseases are subcategories of inflammatory bowel disease. While we might have expected to see them cluster more closely than any two other diseases in our dataset, this did not occur. It may be that the first two principal components were especially efficient at separating CD and UC, which are mutually exclusive diseases (people afflicted with one disease will not have the other), while being less efficient at differentiating between diseases that often co-occur, like multiple sclerosis, rheumatoid arthritis, and type 1 diabetes [47]. It also might be that, as two of the most well-powered studies in the dataset, there were more associations with these two studies that the PCA could use to differentiate them from the others.

Among the other studies, the relationships we find do not appear to be wildly inconsistent with the published literature, but there does not appear to be any resource that can strongly confirm our findings either. The separation of ankylosing spondylitis from all the other studies is confusing from a biological standpoint—one might expect the arthritic diseases (AS, RA, and JIA) to cluster together. However, given that AS was the only disease that did not receive a significant amount of extra weight under our empirical prior—recall that none of the 20 models receiving the highest amount of prior weight, depicted in Figure 5.3, included AS—we suspect that this is an artifact

of the data and that AS has fewer associations compared to the other studies overall, which causes it to be differentiated from the others in the PCA.

Finally, there is some evidence in this dataset that an extreme effect in one study correlates to an increased number of associations overall. Of our nine analyses using various combinations of ρ and σ, only one showed a low estimated effect ($\hat{\beta} < 0.1$) on the maximum z score across all studies at a SNP and the number of disease associations with that SNP. This analysis assumes a prior where true effects are both small ($\sigma = 0.1$) and highly correlated across studies ($\rho = 0.96$) which—if true for the data—are conditions that would not be conducive to establishing the relationship that we were testing.

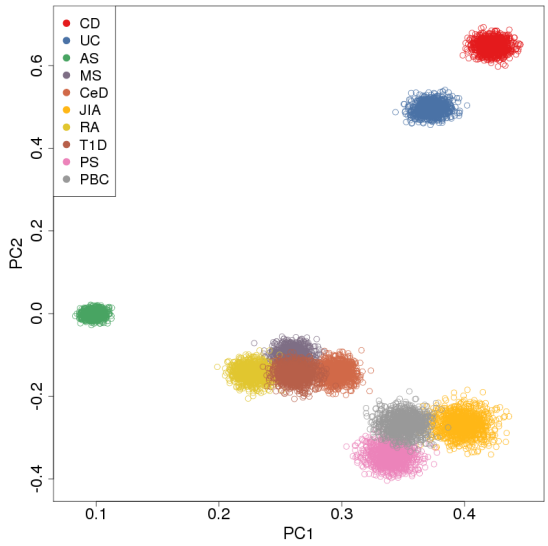
It is also worth noting that the analysis where $\rho = 0.96$ and $\sigma = 0.1$ is an outlier in other respects. When calculating the empirical prior on models of association, this combination of ρ and σ increased the prior weight on the most models, but none of these increases were large in absolute terms (the maximum was around 0.016). It was also the only set of parameters for which the model with the highest amount of prior weight (when calculating the empirical prior) was not the null model. The interpretation of this is difficult; for us, it suggested that aberrant results under this set of prior parameters might be expected. Further work should be carried out—either in the form of simulations, or through application to real datasets—to establish under what circumstances this prior yields results unlike those returned from analyses using the other sets of priors we have used, and whether any other combinations of priors can be found that give similar results to it. This would help us interpret the results from this prior and understand the circumstances under which it is best to use it.

Appendix

5.A More PCA comparisons

Figure 5.A.1: PCA cluster plots generated by performing the singular value decomposition with SNPs chosen preferentially if they showed high evidence of association in our analysis (Figure 5.A.1a) or randomly. Otherwise, both sets shown were identical, using the results from the analysis $\rho = 0$ and $\sigma = 0.2$ with the empirical prior on models of association. All SNPs included in either analysis were at least 1 cM and 100 kb apart.

(a) SNPs chosen based on evidence of association.



(b) SNPs chosen randomly.

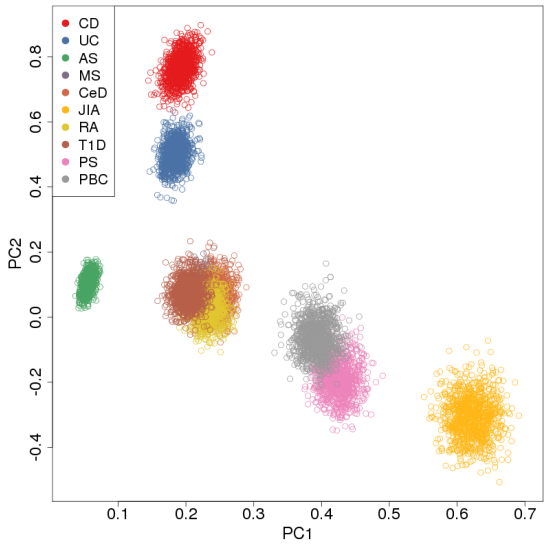
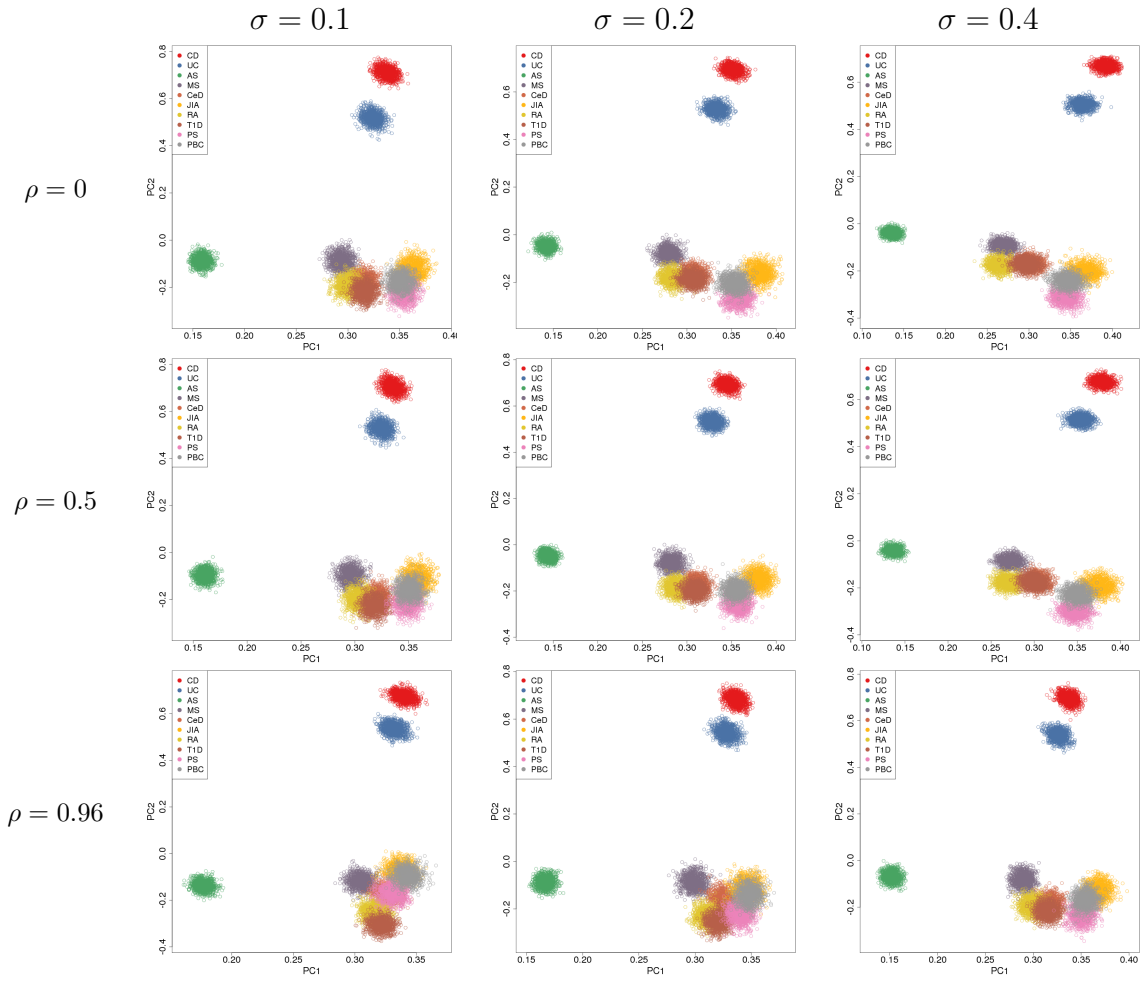


Figure 5.A.2: PCA plots from the exhaustive analyses performed using the flat prior on models of association.

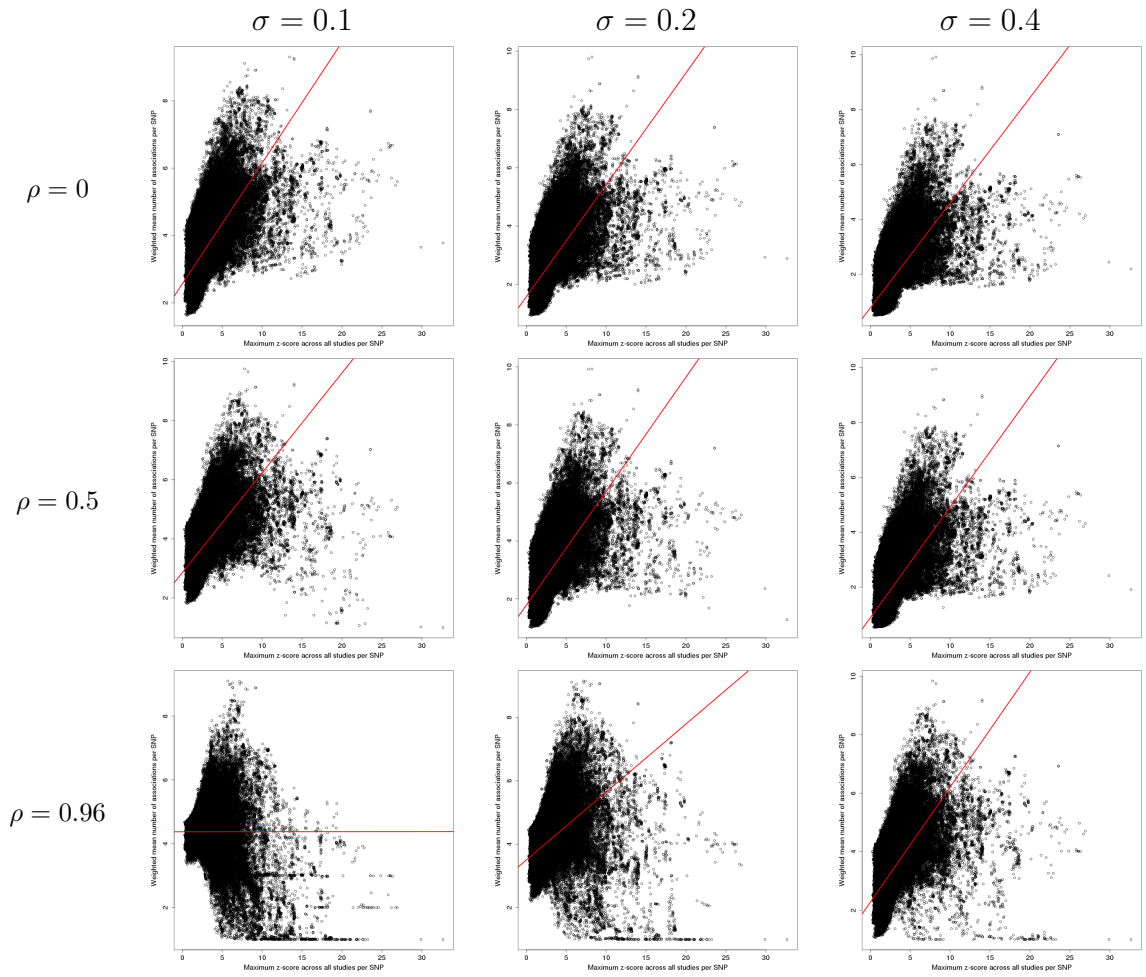


5.B Large estimated effects and analyses with the flat prior

Analysis	Slope	SE of slope
$\rho = 0, \sigma = 0.1$	0.3590	0.0006
$\rho = 0, \sigma = 0.2$	0.3844	0.0006
$\rho = 0, \sigma = 0.4$	0.3833	0.0005
$\rho = 0.5, \sigma = 0.1$	0.3381	0.0006
$\rho = 0.5, \sigma = 0.2$	0.3935	0.0006
$\rho = 0.5, \sigma = 0.4$	0.4041	0.0006
$\rho = 0.96, \sigma = 0.1$	0.0001	0.0003
$\rho = 0.96, \sigma = 0.2$	0.2145	0.0005
$\rho = 0.96, \sigma = 0.4$	0.3939	0.0007

Table 5.B.1: Table of the regression slopes and their standard errors each set of ρ and σ used for analysis, using the flat prior on models of association.

Figure 5.B.1: For each analysis, the estimated number of effects (y-axis) plotted as a function of the maximum z score (x-axis), per SNP. The red line is the least squares regression line calculated from the data. All analyses used the flat prior.



Chapter 6

Discussion

6.1 Introduction

The proliferation of genome-wide association studies has led to a concomitant rise in consortia established to create large samples in which to investigate traits of interest. Examples of these include the Wellcome Trust Case Control Consoritia 1 and 2 [26, 229]— which examined a number of diseases and traits and made use of shared pools of controls—the Cohorts for Heart and Aging Research in Genomic Epidemiology (CHARGE) Consortium [230], and the The Genetic Investigation of Anthropometric Traits (GIANT) Consortium [171]. More recent years have seen the rise of biobanks—such as the China Kadoorie Biobank [73] and UK Biobank [72], which gather not only a large number of samples, but for each sample, phenotypic data for a a wide variety of traits. In addition to standard genome-wide association studies, these datasets invite analysis of the shared genetic bases between traits.

Unlike genome-wide association studies, there is currently no standard set of methods to determine cross-trait associations, and many have been proposed. We reviewed some of these methods in Chapter 1, Section 1.7, and have proposed our own method, in Section 2.2 of Chapter 2. We used the rest of this thesis to investigate the properties of our method in both simulated and real data. In this chapter, we summarize what we have learned so far and discuss the implications, as well as possible directions of future research.

6.2 Our method

6.2.1 ABF results are comparable to those of other methods

Our approximate Bayes factor approach, like many others (see Table 1.1 in Section 1.7 of Chapter 1), uses summary statistics as its input. Summary statistics are often publicly available and are more easily shared across research groups. Additionally, because the calculation involved in our method is simply the ratio of multivariate normal densities, it can be performed quickly using just about any programming language or statistical package. This is in contrast to some of the competing methods we examined in Chapter 3, Section 3.5, which took longer to run and required downloading of particular software or R packages.

Across all methods, there was variation in the markers highlighted, as well as in the patterns of association chosen by each method as the most likely; however our results did not deviate significantly from those of the others in either respect. Our subset exhaustive approach also allows us to calculate a posterior distribution on patterns of association, which we can use to explore the relationships between traits. While other methods also provide estimates of the marginal probabilities of associations of each trait with each marker, none of the ones we have investigated provide a full distribution across all patterns of associations.

6.2.2 The ABF approach is scalable to genome-wide data that share controls across studies

Through our applications to the Wellcome Trust Case Control Consortium 2 data (see Chapter 4) and to the Immunobase data (see Chapter 5), we have also demonstrated that our method scales to genome-wide datasets with 10-20 cohorts, though the subset exhaustive approach was only feasible for the Immunobase data (comprising 10 cohorts). Our WTCCC2 analysis also demonstrated the method's ability to account for covariance in effect size estimates due to things like shared controls. Because so much potential data will come from consortia and biobanks, where samples may be shared across studies of different traits, meta-analysis methods will need to be able to account for this. Table 4.1 of Chapter 4, Section 4.3, suggests that it is possible

to estimate correlation coefficients from the data directly and that the results from this rather naïve approach are in line with those of a more sophisticated method like LD score regression [85]. However, our difficulties in getting LD score regression to produce estimates for some pairs of WTCCC2 cohorts, as well as the low overlap in markers in the LD score reference panels and the Immunobase data (see Chapter 5, Section 5.3), meant that we could not corroborate the similarity of the results in another dataset.

6.2.3 Weaknesses of the ABF approach

However, like all methods, ours also has its limitations. We noticed in our WTCCC2 analysis that our method tends to prioritize markers for which there is more information available over markers with missing data, even if the marginal evidence of association is slightly stronger at the marker with the missing data (see Section 4.5 of Chapter 4). We can try to overcome this problem by only meta-analyzing studies for which we have information at a large number of markers—as we did in our analysis of four WTCCC2 autoimmune disease cohorts across three diseases (Sections 4.3.1 and 4.4.2 of Chapter 4)—but this cuts traits out of the analysis that we might want to include. For instance, we excluded ulcertaive colitis from our autoimmune analysis because that cohort was genotyped on a different chip from the other four (see Table 4.1 in Section 4.1 of Chapter 4), and consequently had few variants in common with them.

Summary statistic imputation provides a possible solution to this problem. However in Section 5.2 of Chapter 5, we saw that imputed summary statistics have more uncertainty around their effect size estimates. Variants with low imputation info scores have higher standard errors and, in some cohorts, appear to bias effect size estimates as well, though the effect of the latter is small. We also note that low info scores were particularly prevalent in two of the cohorts—type 1 diabetes and ankylosing spondylitis. When examining our results in Section 5.5 of Chapter 5, we noted that ankylosing spondylitis appeared to have fewer overall associations in Immunobase than the other diseases. One reason for this could be that the summary statistics were imputed with less confidence in this dataset than in the others, making it more difficult for effect size estimates to be assessed as significantly different from

zero.

Furthermore, while summary statistics are more conveniently obtained than raw genetic and phenotypic data, this can also introduce confusion into exactly what data are being summarized. In the dataset of 107 SNPs used in the analyses in Chapter 3 (see Section 3.2), we noticed that reported effect size estimates and standard errors did not always match the corresponding p-values. We also could not verify that the alleles were consistently aligned across all studies, nor did we know how many controls were shared across studies in order to estimate cryptic relatedness due to sample sharing. While this can be estimated using LD score regression or from the effect size estimates directly, on a small dataset such as that one, these methods are not feasible. Furthermore, while the authors who collected the data [52] did mention that some of the data came from proxies for the reported SNPs, they did not say which datapoints these were, illustrating another potential problem with summary data: one cannot tell which markers have been genotyped directly and which ones have been imputed (either genotypes or summary statistics).

The Immunobase data in Chapter 5 (see Section 5.2) had its own set of challenges, as well. Though we did have information about which studies contained imputed summary statistics, and we were confident in the data alignment pipeline, we noticed that there was a discrepency in the psoriasis data—the standard errors reported in our summary data were much lower than we would have expected, given the reported sample size of the UK cohort. Additionally, two of the studies used samples from both European and non-European populations, and a further study contained data from a meta-analysis of two studies, one of which used related individuals, while the other contained only unrelated individuals (see Table 5.1 for details). When given only summary data, it is hard to be certain exactly which potential confounders—such as population structure—have been accounted for and whether this has been done consistently across all studies. All methods will give spurious results when applied to data that has been poorly quality controlled; it is therefore important to have a full understanding of the provenance of the summary statistics one is working with.

6.2.4 Challenges in validating results

We know from our simulation work in Chapter 2 (Section 2.5.2) and from our results in Chapter 3 Sections 3.4 and 3.5 that our method is not immune to calling spurious associations. The ability to search systematically for shared genetic variants across a large number of traits and cohorts is relatively new, and thus unexpected novel results can be difficult to validate. In Chapter 4, Section 4.4, we highlighted loci like rs793108 on chromosome 10, which did not show a statistically significant association with multiple sclerosis when standard GWAS techniques were applied to the data, but did when our method was applied. Without confirmation of this association from the Immunochip study of MS [179], we would have been uncertain whether this association was real or not.

Absent information from other studies, our other options are functional follow-up or to attempt to replicate the findings in other datasets. Functional follow-up requires information about the pathways affected by the gene associated with the marker. When the associated gene is unclear or does not exist, this quickly becomes a dead-end for the researcher. Even when the genetic variant can be mapped to a gene, biological plausibility can only be established if there is sufficient information on the metabolic pathways affected by the gene, the cell types in which the gene is expressed, and if these pathways and cell types are known (or suspected) to affect the traits associated with the variant. Often, we do not have such complete information, leaving this sort of analysis to be used rarely. Replication in other studies may be more feasible, but presents its own challenges. When the research of the genetic architecture of a particular trait or set of traits is dominated by one research group or consortium, then finding completely independent datasets in which to replicate one's findings is difficult, as we saw for the potential visceral leishmaniasis locus and the possible association between autoimmune diseases and reading and mathematical abilities in Chapter 4, Section 4.4.1.

Additionally, we saw that a putative cross-trait association may be the result of a single SNP tagging different causal loci—see Figures 4.5 and 4.2 in Sections 4.4.2 and 4.5 (respectively) of Chapter 4. This problem is inherited from GWAS, where the causal variant may not have been genotyped, but is in linkage disequilibrium with a variant that has been, leading to a statistically significant association between

the typed variant and the trait under investigation. As discussed above, our method favors SNPs with more information over those with missing information. It also favors SNPs that show marginal evidence of association in more cohorts over SNPs that show stronger evidence of association in a single cohort. These two apsects of our method make it particularly vulnerable to calling spurious cross-trait associations that arise from LD between tag SNPs and multiple causal variants. One potential solution is to use a method like COLOC [42], eCAVIAR [43], or *enloc* [44] to try to determine whether the markers with the highest individual associations with each trait are likely to colocalize to the same causal variant.

6.3 Future directions

Our investigation of this method and results from its applications invite several questions, some of which we attempt to answer, some of which will be left open.

6.3.1 The subset-exhaustive approach for large meta-analyses

One potential strength of our approach is in the ability to calculate posterior probabilties on all possible patterns of association. Summing over the probabilities of these models can give us marginal probabilities of association for each trait, which allows us to say which ones we believe are associated with the SNP, which ones we do not believe are associated, and how certain we are in either case. We note that in Section 4.4.1 of Chapter 4, our analysis highlighted rs3890745 as being associated with both multiple sclerosis and ulcerative colitis, and predicted explicitly that this locus was not associated with ankylosing spondylitis or psoriasis. Being able to understand which traits are not associated with a genetic marker, as well as those that are, can improve our understanding of the relationships between traits over traditional analyses which focus only on whether or not there is evidence for shared associations.

However, we saw in the analysis of the WTCCC2 data that there is a practical limit to how many studies can be part of the meta-analysis and still allow for the subset-exhaustive approach to be feasible. Moreover, as more data becomes available, we may be interested in looking across hundreds of cohorts/traits. When the number of studies becomes this large, the calculation of approximate Bayes factors for each

possible model of association may become intractable even at one SNP, never mind genome-wide. The memory requirements may become too great, or the calculation may take too long to be feasible.

One possible solution to this problem is the stochastic shotgun search (SSS) [231]. This iterative approach can quickly identify the models associated with the highest Bayes factors and explore related models, meaning that the subset of models explored is likely to contain the one that best fits the data. It has been successfully implemented in FINEMAP [22] to discover causal variants in a region without having to search over every possible combination of variants.

Briefly, this approach works by taking an initial model, γ_0 with n associated traits, and then calculating the neighborhood, $\text{nbd}(\gamma_0) = \{\gamma^+, \gamma^0, \gamma^-\}$ around it. If we write γ_0 as a binary string as in Table 2.1 of Chapter 2, Section 2.2.4, then γ^+ is the set of strings where one of the 0 entries in γ_0 is changed to a 1. Similarly, γ^- contains all the strings where one of the 1 entries in γ_0 is changed to a 0. Finally, γ^0 is the set of all strings that change exactly one element that was 1 in γ_0 to 0 and one 0 element of γ_0 and change it to a 1.

Within each of γ^+, γ^0, and γ^-, the approximate Bayes factor for each model is calculated and stored. These are then normalized within each subset (see Equation 2.9) and then one model is chosen as a representative with a probability equal to its normalized Bayes factor. The scores for the each representative are then normalized again, and one of them is chosen as γ_1 with probability equal to approximate Bayes factor normalized within this set. A new neighborhood is calculated as above and the alogrithm repeats for some predefined number of runs or until some other stopping criteria is reached.

Because the approximate Bayes factors calculated for each model are stored—usually in a hash table—we do not need to recalculate any models appearing in $\text{nbd}(\gamma_i)$ that have appeared in $\text{nbd}(\gamma_{j<i})$. Further gains in efficiency can be made by parallelizing the calculations of γ^+, γ^0, and γ^-, the approximate Bayes factors within, and the and selection of a representative from each.

In the datasets we have examined, the distributions of ABFs in the subset-exhaustive approach have been unimodal and the models associated with the highest ABFs tend to differ from each other only by one extra (or one fewer) association. For

example, if the highest ABF is associated with five traits, $\{T1, T2, \ldots, T5\}$, then the model associated with the next highest ABF will either be associated with four of the five traits, or with all five traits plus an extra one ($T6$), while the next highest ABF will be associated with be some other combination of $\{T1, T2, \ldots, T5, T6\}$. Conversely, the models associated with the lowest ABFs show similar consistency. We have never observed a situation where similar models of association do not have similar ABF rankings.

While the SSS shows promise, there are some questions about its implementation that are unresolved. The first is, what is an appropriate stopping criterion? Some possibilities are setting a hard limit on the number of iterations performed, stopping once the ABFs for a certain proportion of all possible models have been calculated, or stopping once the algorithm has performed a certain number of iterations without calculating any new ABFs. It is unclear which of these would be best, or if different circumstances would call for different stopping criteria.

We would also require some investigation into how well the posterior probability distrubtion could be estimated from the subset of models calculated using the SSS. It may be that the probabilities of models whose ABFs are not calculated by the SSS are sufficiently close to zero that they can be ignored, or it may be that a significant portion of the posterior distribution is ignored, and thus, posterior probabilities cannot be estimated from this approach with any accuracy.

6.3.2 Choosing priors

One of the most challenging aspects of our analysis has been choosing appropriate prior values. In Chapter 2, Sections 2.5.2 and 2.6, we saw how poor choices of ρ and σ could impact both inference and the posterior distribution on models of association. In chapter 5, we noted that analysis under a particular set of priors parameters ($\rho = 0.96$ and $\sigma = 0.1$) seemed to yield different results about the relationships between diseases than the analyses using other parameters. Finding principled ways of selecting values of ρ, σ, as well as selecting the weights on models of association that approximate the underlying reality well, could not only improve our method's power to detect effects, but could also improve our inference of which traits are associated with which loci.

In all our analyses thus far, we have hedged by calculating ABFs over a range

of values and taking means, looking at the results that yield the maximum ABF over all prior values, or looking for patterns that are consistent across all combinations of prior values. Despite calculating mean ABFs, we have not taken the logical next step of investigating posterior probability distributions derived from them directly. It would be interesting to see how well they approximate the true models underlying the data.

Furthermore, as we discussed in Chapter 4, Section 4.5, our analyses tend to assume that σ and ρ do not vary throughout the genome or across traits within a single marker. These assumption keep the number of calculations manageable, but probably do not reflect the underlying biological reality very well—for instance, we might expect that true effect sizes for immune-mediated diseases will be large in the HLA region on chromosome 6, but smaller elsewhere in the genome.

Using empirical priors, which are calculated from the data, is one option for systematically choosing priors that vary both by trait and across the genome. Additionally, such priors would be expected to model the underlying reality better than more arbitrarily chosen priors. In Chapter 5, we used one type of empirical prior (Section 5.3.1) on models of association. However, the genome-wide prior we used necessarily assumes that the probability distribution is the same across the genome, when—as with the values of ρ and σ discussed above—this is probably not the case. It may be possible to estimate a more accurate prior by using a window approach across each chromosome.

It might also be possible to determine the prior probability distribution by using liability models. Briefly, liability models assume that there is an underlying phenotype liability that varies throughout the population. People whose liabilities cross some threshold amount will have the phenotype, while those whose liabilities do not cross the threshold will not [232]. So, for example, in our analyses, we might assume that the underlying liability of each disease is normally distributed throughout the population, and that these liabilities—or a subset of them—may have some sort of correlation with one another. It may be possible to determine these liability thresholds, estimate their underlying structure, and use this information to calculate weights on models, which can then be used as priors in the ABF analysis.

6.3.3 The effect of large effect sizes on the number of associations

If we accept the omnigenic model [46], discussed in Section 1.2 of Chapter 1, then pleiotropy (both vertical and horizontal) is pervasive around the genome and what we currently think of as pleiotropic loci are simply those with detectably large effects on multiple traits—that is, those that are part of the "core set" of genes. The finding that most genomic regions transcribed in tissues relevant to the trait of interest seem to contribute to the heritability of that trait—even those that are not in the set of core genes—suggests that a natural next step in our analysis would be the integration of expression data. We could achieve this, for example, by lowering the prior σ on true effect sizes in regions that are transcribed in biologically-relevant tissues, or by increasing the correlation coefficients between traits whose set of biologically-relevant tissues overlap.

This hypothesis also begs the question of whether large effects in multiple traits are more or less likely for a given SNP, if it is already known to have a large effect on another trait. Our own analyses in Chapter 3, Section 3.4.3 (Figures 3.4 and 3.5) and Chapter 5, Section 5.4.3 (Figure 5.5) suggest that this may indeed be the case—though the strength of evidence changes depending on the priors used to calculate our ABF results, which are then used to calculate the number of expected associations at each marker. Moreover, both datasets in which we performed this analysis comprised only autoimmune diseases, which are known to have shared genetic bases. Furthermore, given our method's inability to disentangle true association at a marker with those induced by LD with other markers (discussed in Section 6.2.3 of this chapter), it may be that the mean numbers of associations we calculated are inflated, and that this systematically affects markers with a single large effect size more than those without.

It may also be speculated that many of these shared associations are due to effects on the same metabolic pathways that affect the risk of multiple immune-mediated diseases. If this is true, then those who reject the idea that vertical pleiotropy is true pleiotropy would argue that that these loci are not truly associated with multiple traits—that instead, these loci are associated with endophenotypes that underlie multiple diseases. It would be interesting to see if these results could be replicated in sets of traits where such sharing of genetic loci was neither previously observed

nor expected. This would suggest that horizontal pleiotropy—which no one denies is pleiotropy—can also be expected at loci with large effects on at least one trait.

6.4 Closing remarks

Despite the work that has been done in establishing the properties of this method and applying it to data, there is always more that can be done: more simulations to learn how results change under different circumstances, investigations into how to choose prior parameters, and increasing numbers of datasets to which the method can be applied. Our approach is one of many that attempt to establish associations across multiple genome-wide association studies from summary statistics. While we do not (and cannot) argue that our method is better than any other method, we hope we have established some of the strengths and weakness of our method and shown that it has the flexibility to be deployed in multiple settings to answer questions about the genetic relationships between traits.

Works Cited

[1] R. Punnett, Mendelism. New York, New York: Whilshire Book Co., american, 1st ed., 1909.

[2] N. W. Gillham, "The Battle Between the Biometricians and the Mendelians: How Sir Francis Galton's Work Caused his Disciples to Reach Conflicting Conclusions About the Hereditary Mechanism," Science & Education, vol. 24, pp. 61–75, Jan. 2015.

[3] R. A. Fisher, "The Correlation between Relatives on the Supposition of Mendelian Inheritance.," Transactions of the Royal Society of Edinburgh, vol. 52, no. 02, pp. 399–433, 1918.

[4] A. H. Sturtevant, "The linear arrangement of six sex-linked factors in Drosophila, as shown by their mode of association," Journal of Experimental Zoology, vol. 14, pp. 43–59, Jan. 1913.

[5] A. D. Hershey and M. Chase, "Independent functions of viral protein and nucleic acid in growth of bacteriophage," The Journal of General Physiology, vol. 36, pp. 39–56, May 1952.

[6] J. D. Watson and F. H. Crick, "Molecular structure of nucleic acids; a structure for deoxyribose nucleic acid," Nature, vol. 171, pp. 737–738, Apr. 1953.

[7] J. C. Venter, M. D. Adams, E. W. Myers, P. W. Li, R. J. Mural, G. G. Sutton, H. O. Smith, M. Yandell, C. A. Evans, R. A. Holt, J. D. Gocayne, P. Amanatides, R. M. Ballew, D. H. Huson, J. R. Wortman, Q. Zhang, C. D. Kodira, X. H. Zheng, L. Chen, M. Skupski, G. Subramanian, P. D. Thomas, J. Zhang, G. L. Gabor Miklos, C. Nelson, S. Broder, A. G. Clark, J. Nadeau, V. A. McKusick, N. Zinder, A. J. Levine, R. J. Roberts, M. Simon, C. Slayman, M. Hunkapiller, R. Bolanos, A. Delcher, I. Dew, D. Fasulo, M. Flanigan, L. Florea, A. Halpern, S. Hannenhalli, S. Kravitz, S. Levy, C. Mobarry, K. Reinert, K. Remington, J. Abu-Threideh, E. Beasley, K. Biddick, V. Bonazzi, R. Brandon, M. Cargill, I. Chandramouliswaran, R. Charlab, K. Chaturvedi, Z. Deng, V. D. Francesco, P. Dunn, K. Eilbeck, C. Evangelista, A. E. Gabrielian, W. Gan, W. Ge, F. Gong, Z. Gu, P. Guan, T. J. Heiman, M. E. Higgins, R.-R. Ji, Z. Ke, K. A. Ketchum, Z. Lai, Y. Lei, Z. Li, J. Li, Y. Liang, X. Lin, F. Lu, G. V. Merkulov, N. Milshina, H. M. Moore, A. K. Naik, V. A. Narayan, B. Neelam, D. Nusskern, D. B. Rusch, S. Salzberg, W. Shao, B. Shue, J. Sun, Z. Y. Wang, A. Wang, X. Wang, J. Wang, M.-H. Wei, R. Wides, C. Xiao, C. Yan, A. Yao, J. Ye, M. Zhan, W. Zhang, H. Zhang, Q. Zhao, L. Zheng, F. Zhong, W. Zhong, S. C. Zhu, S. Zhao, D. Gilbert, S. Baumhueter, G. Spier, C. Carter, A. Cravchik, T. Woodage, F. Ali, H. An, A. Awe, D. Baldwin, H. Baden, M. Barnstead, I. Barrow, K. Beeson, D. Busam, A. Carver, A. Center, M. L. Cheng, L. Curry, S. Danaher, L. Davenport, R. Desilets, S. Dietz, K. Dodson, L. Doup, S. Ferriera, N. Garg, A. Gluecksmann, B. Hart, J. Haynes, C. Haynes, C. Heiner, S. Hladun,

D. Hostin, J. Houck, T. Howland, C. Ibegwam, J. Johnson, F. Kalush, L. Kline, S. Koduru, A. Love, F. Mann, D. May, S. McCawley, T. McIntosh, I. McMullen, M. Moy, L. Moy, B. Murphy, K. Nelson, C. Pfannkoch, E. Pratts, V. Puri, H. Qureshi, M. Reardon, R. Rodriguez, Y.-H. Rogers, D. Romblad, B. Ruhfel, R. Scott, C. Sitter, M. Smallwood, E. Stewart, R. Strong, E. Suh, R. Thomas, N. N. Tint, S. Tse, C. Vech, G. Wang, J. Wetter, S. Williams, M. Williams, S. Windsor, E. Winn-Deen, K. Wolfe, J. Zaveri, K. Zaveri, J. F. Abril, R. Guigó, M. J. Campbell, K. V. Sjolander, B. Karlak, A. Kejariwal, H. Mi, B. Lazareva, T. Hatton, A. Narechania, K. Diemer, A. Muruganujan, N. Guo, S. Sato, V. Bafna, S. Istrail, R. Lippert, R. Schwartz, B. Walenz, S. Yooseph, D. Allen, A. Basu, J. Baxendale, L. Blick, M. Caminha, J. Carnes-Stine, P. Caulk, Y.-H. Chiang, M. Coyne, C. Dahlke, A. D. Mays, M. Dombroski, M. Donnelly, D. Ely, S. Esparham, C. Fosler, H. Gire, S. Glanowski, K. Glasser, A. Glodek, M. Gorokhov, K. Graham, B. Gropman, M. Harris, J. Heil, S. Henderson, J. Hoover, D. Jennings, C. Jordan, J. Jordan, J. Kasha, L. Kagan, C. Kraft, A. Levitsky, M. Lewis, X. Liu, J. Lopez, D. Ma, W. Majoros, J. McDaniel, S. Murphy, M. Newman, T. Nguyen, N. Nguyen, M. Nodell, S. Pan, J. Peck, M. Peterson, W. Rowe, R. Sanders, J. Scott, M. Simpson, T. Smith, A. Sprague, T. Stockwell, R. Turner, E. Venter, M. Wang, M. Wen, D. Wu, M. Wu, A. Xia, A. Zandieh, and X. Zhu, "The Sequence of the Human Genome," Science, vol. 291, pp. 1304–1351, Feb. 2001.

[8] The International HapMap Consortium, "The International HapMap Project," Nature, vol. 426, pp. 789–796, Dec. 2003.

[9] The 1000 Genomes Project Consortium, "A map of human genome variation from population-scale sequencing," Nature, vol. 467, pp. 1061–1073, Oct. 2010.

[10] M. MacDonald, C. M. Ambrose, M. P. Duyao, R. H. Myers, C. Lin, L. Srinidhi, G. Barnes, S. A. Taylor, M. James, N. Groot, H. MacFarlane, B. Jenkins, M. A. Anderson, N. S. Wexler, and J. F. Gusella, "A novel gene containing a trinucleotide repeat that is expanded and unstable on Huntington's disease chromosomes," Cell, vol. 72, pp. 971–983, Mar. 1993.

[11] Y. Miki, J. Swensen, D. Shattuck-Eidens, P. A. Futreal, K. Harshman, S. Tavtigian, Q. Liu, C. Cochran, L. M. Bennett, and W. Ding, "A strong candidate for the breast and ovarian cancer susceptibility gene BRCA1," Science (New York, N.Y.), vol. 266, pp. 66–71, Oct. 1994.

[12] R. Wooster, G. Bignell, J. Lancaster, S. Swift, S. Seal, J. Mangion, N. Collins, S. Gregory, C. Gumbs, G. Micklem, R. Barfoot, R. Hamoudi, S. Patel, C. Rices, P. Biggs, Y. Hashim, A. Smith, F. Connor, A. Arason, J. Gudmundsson, D. Ficenec, D. Kelsell, P. Tonin, D. Timothy Bishop, N. K. Spurr, B. A. J. Ponder, R. Eeles, J. Peto, P. Devilee, C. Cornelisse, H. Lynch, S. Narod, G. Lenoir, V. Egilsson, R. Bjork Barkadottir, D. F. Easton, D. R. Bentley, P. A. Futreal, A. Ashworth, and M. R. Stratton, "Identification of the breast cancer susceptibility gene BRCA2," Nature, vol. 378, pp. 789–792, Dec. 1995.

[13] H. K. Tabor, N. J. Risch, and R. M. Myers, "Candidate-gene approaches for studying complex genetic traits: practical considerations," Nature Reviews Genetics, vol. 3, pp. 391–397, May 2002.

[14] R. J. Klein, C. Zeiss, E. Y. Chew, J.-Y. Tsai, R. S. Sackler, C. Haynes, A. K. Henning, J. P. SanGiovanni, S. M. Mane, S. T. Mayne, M. B. Bracken, F. L. Ferris, J. Ott, C. Barnstable, and J. Hoh, "Complement Factor H Polymorphism

in Age-Related Macular Degeneration," Science, vol. 308, pp. 385–389, Apr. 2005.
[15] K. G. Ardlie, L. Kruglyak, and M. Seielstad, "Patterns of linkage disequilibrium in the human genome," Nature Reviews Genetics, vol. 3, pp. 299–309, Apr. 2002.
[16] B. Devlin and N. Risch, "A Comparison of Linkage Disequilibrium Measures for Fine-Scale Mapping," Genomics, vol. 29, pp. 311–322, Sept. 1995.
[17] C. S. Carlson, M. A. Eberle, M. J. Rieder, Q. Yi, L. Kruglyak, and D. A. Nickerson, "Selecting a Maximally Informative Set of Single-Nucleotide Polymorphisms for Association Analyses Using Linkage Disequilibrium," The American Journal of Human Genetics, vol. 74, pp. 106–120, Jan. 2004.
[18] National Human Genome Research Institute, "The Cost of Sequencing a Human Genome - National Human Genome Research Institute (NHGRI)," July 2017.
[19] A. Cortes and M. A. Brown, "Promise and pitfalls of the Immunochip," Arthritis Research & Therapy, vol. 13, no. 1, p. 101, 2010.
[20] G. Galarneau, C. D. Palmer, V. G. Sankaran, S. H. Orkin, J. N. Hirschhorn, and G. Lettre, "Fine-mapping at three loci known to affect fetal hemoglobin levels explains additional genetic variation," Nature Genetics, vol. 42, pp. 1049–1051, Dec. 2010.
[21] K. K.-H. Farh, A. Marson, J. Zhu, M. Kleinewietfeld, W. J. Housley, S. Beik, N. Shoresh, H. Whitton, R. J. H. Ryan, A. A. Shishkin, M. Hatan, M. J. Carrasco-Alfonso, D. Mayer, C. J. Luckey, N. A. Patsopoulos, P. L. De Jager, V. K. Kuchroo, C. B. Epstein, M. J. Daly, D. A. Hafler, and B. E. Bernstein, "Genetic and epigenetic fine mapping of causal autoimmune disease variants," Nature, vol. 518, pp. 337–343, Oct. 2014.
[22] C. Benner, C. C. Spencer, A. S. Havulinna, V. Salomaa, S. Ripatti, and M. Pirinen, "FINEMAP: efficient variable selection using summary data from genome-wide association studies," Bioinformatics, vol. 32, pp. 1493–1501, May 2016.
[23] H. Huang, M. Fang, L. Jostins, M. Umićević Mirkov, G. Boucher, C. A. Anderson, V. Andersen, I. Cleynen, A. Cortes, F. Crins, M. D'Amato, V. Deffontaine, J. Dmitrieva, E. Docampo, M. Elansary, K. K.-H. Farh, A. Franke, A.-S. Gori, P. Goyette, J. Halfvarson, T. Haritunians, J. Knight, I. C. Lawrance, C. W. Lees, E. Louis, R. Mariman, T. Meuwissen, M. Mni, Y. Momozawa, M. Parkes, S. L. Spain, E. Théâtre, G. Trynka, J. Satsangi, S. van Sommeren, S. Vermeire, R. J. Xavier, R. K. Weersma, R. H. Duerr, C. G. Mathew, J. D. Rioux, D. P. B. McGovern, J. H. Cho, M. Georges, M. J. Daly, and J. C. Barrett, "Fine-mapping inflammatory bowel disease loci to single-variant resolution," Nature, vol. 547, pp. 173–178, June 2017.
[24] J. Liu, X. Wan, C. Wang, C. Yang, X. Zhou, and C. Yang, "LLR: A latent low-rank approach to colocalizing genetic risk variants in multiple GWAS," Bioinformatics, Aug. 2017.
[25] T. Burdett, P. Hall, E. Hastings, L. Hindorff, H. Junkins, A. Klemm, J. MacArthur, T. Manolio, J. Morales, H. Parkinson, and D. Welter, "The NHGRI-EBI Catalog of published genome-wide association studies.," 2016.
[26] The Wellcome Trust Case Control Consortium, "Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls," Nature, vol. 447, pp. 661–678, June 2007.
[27] B. Maher, "Personal genomes: The case of the missing heritability," Nature,

vol. 456, pp. 18–21, Nov. 2008.

[28] A. L. Price, C. C. A. Spencer, and P. Donnelly, "Progress and promise in understanding the genetic basis of common diseases," Proceedings of the Royal Society B: Biological Sciences, vol. 282, p. 20151684, Dec. 2015.

[29] J. Yang, B. Benyamin, B. P. McEvoy, S. Gordon, A. K. Henders, D. R. Nyholt, P. A. Madden, A. C. Heath, N. G. Martin, G. W. Montgomery, M. E. Goddard, and P. M. Visscher, "Common SNPs explain a large proportion of the heritability for human height," Nature Genetics, vol. 42, pp. 565–569, July 2010.

[30] T. A. Manolio, F. S. Collins, N. J. Cox, D. B. Goldstein, L. A. Hindorff, D. J. Hunter, M. I. McCarthy, E. M. Ramos, L. R. Cardon, A. Chakravarti, J. H. Cho, A. E. Guttmacher, A. Kong, L. Kruglyak, E. Mardis, C. N. Rotimi, M. Slatkin, D. Valle, A. S. Whittemore, M. Boehnke, A. G. Clark, E. E. Eichler, G. Gibson, J. L. Haines, T. F. C. Mackay, S. A. McCarroll, and P. M. Visscher, "Finding the missing heritability of complex diseases," Nature, vol. 461, pp. 747–753, Oct. 2009.

[31] O. Zuk, S. F. Schaffner, K. Samocha, R. Do, E. Hechter, S. Kathiresan, M. J. Daly, B. M. Neale, S. R. Sunyaev, and E. S. Lander, "Searching for missing heritability: Designing rare variant association studies," Proceedings of the National Academy of Sciences, vol. 111, pp. E455–E464, Jan. 2014.

[32] T. M. Frayling, N. J. Timpson, M. N. Weedon, E. Zeggini, R. M. Freathy, C. M. Lindgren, J. R. B. Perry, K. S. Elliott, H. Lango, N. W. Rayner, B. Shields, L. W. Harries, J. C. Barrett, S. Ellard, C. J. Groves, B. Knight, A.-M. Patch, A. R. Ness, S. Ebrahim, D. A. Lawlor, S. M. Ring, Y. Ben-Shlomo, M.-R. Jarvelin, U. Sovio, A. J. Bennett, D. Melzer, L. Ferrucci, R. J. F. Loos, I. Barroso, N. J. Wareham, F. Karpe, K. R. Owen, L. R. Cardon, M. Walker, G. A. Hitman, C. N. A. Palmer, A. S. F. Doney, A. D. Morris, G. D. Smith, The Wellcome Trust Case Control Consortium, A. T. Hattersley, and M. I. McCarthy, "A Common Variant in the FTO Gene Is Associated with Body Mass Index and Predisposes to Childhood and Adult Obesity," Science, vol. 316, pp. 889–894, May 2007.

[33] T. M. Frayling, "Genome–wide association studies provide new insights into type 2 diabetes aetiology," Nature Reviews Genetics, vol. 8, pp. 657–662, Sept. 2007.

[34] Y. Okada, D. Wu, G. Trynka, T. Raj, C. Terao, K. Ikari, Y. Kochi, K. Ohmura, A. Suzuki, S. Yoshida, R. R. Graham, A. Manoharan, W. Ortmann, T. Bhangale, J. C. Denny, R. J. Carroll, A. E. Eyler, J. D. Greenberg, J. M. Kremer, D. A. Pappas, L. Jiang, J. Yin, L. Ye, D.-F. Su, J. Yang, G. Xie, E. Keystone, H.-J. Westra, T. Esko, A. Metspalu, X. Zhou, N. Gupta, D. Mirel, E. A. Stahl, D. Diogo, J. Cui, K. Liao, M. H. Guo, K. Myouzen, T. Kawaguchi, M. J. H. Coenen, P. L. C. M. van Riel, M. A. F. J. van de Laar, H.-J. Guchelaar, T. W. J. Huizinga, P. Dieudé, X. Mariette, S. Louis Bridges Jr, A. Zhernakova, R. E. M. Toes, P. P. Tak, C. Miceli-Richard, S.-Y. Bang, H.-S. Lee, J. Martin, M. A. Gonzalez-Gay, L. Rodriguez-Rodriguez, S. Rantapää-Dahlqvist, L. Ärlestig, H. K. Choi, Y. Kamatani, P. Galan, M. Lathrop, S. Eyre, J. Bowes, A. Barton, N. de Vries, L. W. Moreland, L. A. Criswell, E. W. Karlson, A. Taniguchi, R. Yamada, M. Kubo, J. S. Liu, S.-C. Bae, J. Worthington, L. Padyukov, L. Klareskog, P. K. Gregersen, S. Raychaudhuri, B. E. Stranger, P. L. De Jager, L. Franke, P. M. Visscher, M. A. Brown, H. Yamanaka, T. Mimori, A. Taka-

hashi, H. Xu, T. W. Behrens, K. A. Siminovitch, S. Momohara, F. Matsuda, K. Yamamoto, and R. M. Plenge, "Genetics of rheumatoid arthritis contributes to biology and drug discovery," Nature, vol. 506, pp. 376–381, Dec. 2013.

[35] The ENCODE Project Consortium, "An integrated encyclopedia of DNA elements in the human genome," Nature, vol. 489, pp. 57–74, Sept. 2012.

[36] I. Ezkurdia, D. Juan, J. M. Rodriguez, A. Frankish, M. Diekhans, J. Harrow, J. Vazquez, A. Valencia, and M. L. Tress, "Multiple evidence strands suggest that there may be as few as 19 000 human protein-coding genes," Human Molecular Genetics, vol. 23, pp. 5866–5878, Nov. 2014.

[37] D. Levy, G. B. Ehret, K. Rice, G. C. Verwoert, L. J. Launer, A. Dehghan, N. L. Glazer, A. C. Morrison, A. D. Johnson, T. Aspelund, Y. Aulchenko, T. Lumley, A. Köttgen, R. S. Vasan, F. Rivadeneira, G. Eiriksdottir, X. Guo, D. E. Arking, G. F. Mitchell, F. U. S. Mattace-Raso, A. V. Smith, K. Taylor, R. B. Scharpf, S.-J. Hwang, E. J. G. Sijbrands, J. Bis, T. B. Harris, S. K. Ganesh, C. J. O'Donnell, A. Hofman, J. I. Rotter, J. Coresh, E. J. Benjamin, A. G. Uitterlinden, G. Heiss, C. S. Fox, J. C. M. Witteman, E. Boerwinkle, T. J. Wang, V. Gudnason, M. G. Larson, A. Chakravarti, B. M. Psaty, and C. M. van Duijn, "Genome-wide association study of blood pressure and hypertension," Nature Genetics, vol. 41, pp. 677–687, June 2009.

[38] N. Kato, F. Takeuchi, Y. Tabara, T. N. Kelly, M. J. Go, X. Sim, W. T. Tay, C.-H. Chen, Y. Zhang, K. Yamamoto, T. Katsuya, M. Yokota, Y. J. Kim, R. T. H. Ong, T. Nabika, D. Gu, L.-c. Chang, Y. Kokubo, W. Huang, K. Ohnaka, Y. Yamori, E. Nakashima, C. E. Jaquish, J.-Y. Lee, M. Seielstad, M. Isono, J. E. Hixson, Y.-T. Chen, T. Miki, X. Zhou, T. Sugiyama, J.-P. Jeon, J. J. Liu, R. Takayanagi, S. S. Kim, T. Aung, Y. J. Sung, X. Zhang, T. Y. Wong, B.-G. Han, S. Kobayashi, T. Ogihara, D. Zhu, N. Iwai, J.-Y. Wu, Y. Y. Teo, E. S. Tai, Y. S. Cho, and J. He, "Meta-analysis of genome-wide association studies identifies common variants associated with blood pressure variation in east Asians," Nature Genetics, vol. 43, pp. 531–538, June 2011.

[39] S. Smemo, J. J. Tena, K.-H. Kim, E. R. Gamazon, N. J. Sakabe, C. Gómez-Marín, I. Aneas, F. L. Credidio, D. R. Sobreira, N. F. Wasserman, J. H. Lee, V. Puviindran, D. Tam, M. Shen, J. E. Son, N. A. Vakili, H.-K. Sung, S. Naranjo, R. D. Acemel, M. Manzanares, A. Nagy, N. J. Cox, C.-C. Hui, J. L. Gomez-Skarmeta, and M. A. Nóbrega, "Obesity-associated variants within FTO form long-range functional connections with IRX3," Nature, vol. 507, pp. 371–375, Mar. 2014.

[40] A. B. Paaby and M. V. Rockman, "The many faces of pleiotropy," Trends in Genetics, vol. 29, pp. 66–73, Feb. 2013.

[41] F. W. Stearns, "One Hundred Years of Pleiotropy: A Retrospective," Genetics, vol. 186, pp. 767–773, Nov. 2010.

[42] C. Giambartolomei, D. Vukcevic, E. E. Schadt, L. Franke, A. D. Hingorani, C. Wallace, and V. Plagnol, "Bayesian Test for Colocalisation between Pairs of Genetic Association Studies Using Summary Statistics," PLoS Genetics, vol. 10, p. e1004383, May 2014.

[43] F. Hormozdiari, M. van de Bunt, A. V. Segrè, X. Li, J. W. J. Joo, M. Bilow, J. H. Sul, S. Sankararaman, B. Pasaniuc, and E. Eskin, "Colocalization of GWAS and eQTL Signals Detects Target Genes," American Journal of Human Genetics, vol. 99, pp. 1245–1260, Dec. 2016.

[44] X. Wen, R. Pique-Regi, and F. Luca, "Integrating molecular QTL data into genome-wide genetic association analysis: Probabilistic assessment of enrichment and colocalization," PLOS Genetics, vol. 13, p. e1006646, Mar. 2017.

[45] S. Sivakumaran, F. Agakov, E. Theodoratou, J. G. Prendergast, L. Zgaga, T. Manolio, I. Rudan, P. McKeigue, J. F. Wilson, and H. Campbell, "Abundant Pleiotropy in Human Complex Diseases and Traits," The American Journal of Human Genetics, vol. 89, pp. 607–618, Nov. 2011.

[46] E. A. Boyle, Y. I. Li, and J. K. Pritchard, "An Expanded View of Complex Traits: From Polygenic to Omnigenic," Cell, vol. 169, pp. 1177–1186, June 2017.

[47] A. Zhernakova, C. C. van Diemen, and C. Wijmenga, "Detecting shared pathogenesis from the shared genetics of immune-related diseases," Nature Reviews Genetics, vol. 10, pp. 43–55, Jan. 2009.

[48] D. Ellinghaus, L. Jostins, S. L. Spain, A. Cortes, J. Bethune, B. Han, Y. R. Park, S. Raychaudhuri, J. G. Pouget, M. Hübenthal, T. Folseraas, Y. Wang, T. Esko, A. Metspalu, H.-J. Westra, L. Franke, T. H. Pers, R. K. Weersma, V. Collij, M. D'Amato, J. Halfvarson, A. B. Jensen, W. Lieb, F. Degenhardt, A. J. Forstner, A. Hofmann, S. Schreiber, U. Mrowietz, B. D. Juran, K. N. Lazaridis, S. Brunak, A. M. Dale, R. C. Trembath, S. Weidinger, M. Weichenthal, E. Ellinghaus, J. T. Elder, J. N. W. N. Barker, O. A. Andreassen, D. P. McGovern, T. H. Karlsen, J. C. Barrett, M. Parkes, M. A. Brown, and A. Franke, "Analysis of five chronic inflammatory diseases identifies 27 new associations and highlights disease-specific patterns at shared loci," Nature Genetics, vol. 48, pp. 510–518, Mar. 2016.

[49] S. C. L. Gough and M. C. O'Donovan, "Clustering of metabolic comorbidity in schizophrenia: a genetic contribution?," Journal of Psychopharmacology, vol. 19, pp. 47–55, Nov. 2005.

[50] O. A. Andreassen, S. Djurovic, W. K. Thompson, A. J. Schork, K. S. Kendler, M. C. O'Donovan, D. Rujescu, T. Werge, M. van de Bunt, A. P. Morris, M. I. McCarthy, J. C. Roddey, L. K. McEvoy, R. S. Desikan, and A. M. Dale, "Improved Detection of Common Variants Associated with Schizophrenia by Leveraging Pleiotropy with Cardiovascular-Disease Risk Factors," The American Journal of Human Genetics, vol. 92, pp. 197–209, Feb. 2013.

[51] R. Malik, T. Freilinger, B. S. Winsvold, V. Anttila, J. Vander Heiden, M. Traylor, B. de Vries, E. G. Holliday, G. M. Terwindt, J. Sturm, J. C. Bis, J. C. Hopewell, M. D. Ferrari, K. Rannikmae, M. Wessman, M. Kallela, C. Kubisch, M. Fornage, J. F. Meschia, T. Lehtimaki, C. Sudlow, R. Clarke, D. I. Chasman, B. D. Mitchell, J. Maguire, J. Kaprio, M. Farrall, O. T. Raitakari, T. Kurth, M. A. Ikram, A. P. Reiner, W. T. Longstreth, P. M. Rothwell, D. P. Strachan, P. Sharma, S. Seshadri, L. Quaye, L. Cherkas, M. Schurks, J. Rosand, L. Ligthart, G. B. Boncoraglio, G. Davey Smith, C. M. van Duijn, K. Stefansson, B. B. Worrall, D. R. Nyholt, H. S. Markus, A. M. J. M. van den Maagdenberg, C. Cotsapas, J. A. Zwart, A. Palotie, For the International Headache Genetics Consortium, M. Dichgans, For the METASTROKE Collaboration of the International Stroke Genetics Consortium, U. Thorsteinsdottir, A. L. DeStefano, C. Levi, S. Gretarsdottir, P. Donnelly, I. Barroso, J. M. Blackwell, E. Bramon, M. A. Brown, J. P. Casas, A. Corvin, P. Deloukas, A. Duncanson, J. Jankowski, H. S. Markus, C. G. Mathew, C. N. Palmer, R. Plomin, A. Rautanen, S. J.

sawcer, R. C. Trembath F, A. C. Viswanathan, N. W. Wood, C. C. Spencer, G. Band, C. Bellengues, C. Freeman, G. Hellenthal, E. Giannoulatou, M. Pirinen, R. Pearson, A. Strange, Z. Su, D. Vukcevic, C. Langford, S. E. Hunt, S. Edkind, R. Gwilliam, H. Blackburn, S. J. Bumpstead, S. Dronv, M. Gillman, E. Gray, N. Hammond, A. Jayakumar, O. T. McCann, J. Liddle, S. C. Potter, R. Ravindrarajah, M. Rickette, M. Waller, P. Weston, S. Widaa, P. Whittaker, P. Gormley, F. Bettella, G. McMahon, U. Todt, P. Palta, E. Hamalainen, S. Steinberg, H. Stefansson, M. Farkkila, V. Artto, M. A. Kaunisto, J. Schoenen, R. R. Frants, G. Borck, H. Gobel, A. Heinze, K. Heinze-Kuhn, and B. Muller-Myhsok, "Shared genetic basis for migraine and ischemic stroke: A genome-wide analysis of common variants," Neurology, vol. 84, pp. 2132–2145, May 2015.

[52] C. Cotsapas, B. F. Voight, E. Rossin, K. Lage, B. M. Neale, C. Wallace, G. R. Abecasis, J. C. Barrett, T. Behrens, J. Cho, P. L. De Jager, J. T. Elder, R. R. Graham, P. Gregersen, L. Klareskog, K. A. Siminovitch, D. A. van Heel, C. Wijmenga, J. Worthington, J. A. Todd, D. A. Hafler, S. S. Rich, M. J. Daly, and on behalf of the FOCiS Network of Consortia, "Pervasive Sharing of Genetic Effects in Autoimmune Disease," PLoS Genetics, vol. 7, p. e1002254, Aug. 2011.

[53] M. Parkes, A. Cortes, D. A. van Heel, and M. A. Brown, "Genetic insights into common pathways and complex relationships among immune-mediated diseases," Nature Reviews Genetics, vol. 14, pp. 661–673, Aug. 2013.

[54] B. Han, J. G. Pouget, K. Slowikowski, E. Stahl, C. H. Lee, D. Diogo, X. Hu, Y. R. Park, E. Kim, P. K. Gregersen, S. R. Dahlqvist, J. Worthington, J. Martin, S. Eyre, L. Klareskog, T. Huizinga, W.-M. Chen, S. Onengut-Gumuscu, S. S. Rich, N. R. Wray, and S. Raychaudhuri, "A method to decipher pleiotropy by detecting underlying heterogeneity driven by hidden subgroups applied to autoimmune and neuropsychiatric diseases," Nature Genetics, vol. 48, pp. 803–810, May 2016.

[55] A. Scuteri, S. Sanna, W.-M. Chen, M. Uda, G. Albai, J. Strait, S. Najjar, R. Nagaraja, M. Orr?, G. Usala, M. Dei, S. Lai, A. Maschio, F. Busonero, A. Mulas, G. B. Ehret, A. A. Fink, A. B. Weder, R. S. Cooper, P. Galan, A. Chakravarti, D. Schlessinger, A. Cao, E. Lakatta, and G. R. Abecasis, "Genome-Wide Association Scan Shows Genetic Variants in the FTO Gene Are Associated with Obesity-Related Traits," PLoS Genetics, vol. 3, no. 7, p. e115, 2007.

[56] M. M. Iles, M. H. Law, S. N. Stacey, J. Han, S. Fang, R. Pfeiffer, M. Harland, S. MacGregor, J. C. Taylor, K. K. Aben, L. A. Akslen, M.-F. Avril, E. Azizi, B. Bakker, K. R. Benediktsdottir, W. Bergman, G. B. Scarrà, K. M. Brown, D. Calista, V. Chaudru, M. C. Fargnoli, A. E. Cust, F. Demenais, A. C. de Waal, T. Dębniak, D. E. Elder, E. Friedman, P. Galan, P. Ghiorzo, E. M. Gillanders, A. M. Goldstein, N. A. Gruis, J. Hansson, P. Helsing, M. Hočevar, V. Höiom, J. L. Hopper, C. Ingvar, M. Janssen, M. A. Jenkins, P. A. Kanetsky, L. A. Kiemeney, J. Lang, G. M. Lathrop, S. Leachman, J. E. Lee, J. Lubiński, R. M. Mackie, G. J. Mann, N. G. Martin, J. I. Mayordomo, A. Molven, S. Mulder, E. Nagore, S. Novaković, I. Okamoto, J. H. Olafsson, H. Olsson, H. Pehamberger, K. Peris, M. P. Grasa, D. Planelles, S. Puig, J. A. Puig-Butille, J. Randerson-Moor, C. Requena, L. Rivoltini, M. Rodolfo, M. Santinami, B. Sigurgeirsson, H. Snowden, F. Song, P. Sulem, K. Thorisdottir, R. Tuominen, P. Van Belle, N. van der Stoep, M. M. van Rossum, Q. Wei, J. Wendt, D. Zelenika, M. Zhang, M. T. Landi, G. Thorleifsson, D. T. Bishop, C. I. Amos,

N. K. Hayward, K. Stefansson, J. A. N. Bishop, and J. H. Barrett, "A variant in FTO shows association with melanoma risk not due to BMI," Nature Genetics, vol. 45, pp. 428–432, Mar. 2013.

[57] M. H. Law, D. T. Bishop, J. E. Lee, M. Brossard, N. G. Martin, E. K. Moses, F. Song, J. H. Barrett, R. Kumar, D. F. Easton, P. D. P. Pharoah, A. J. Swerdlow, K. P. Kypreou, J. C. Taylor, M. Harland, J. Randerson-Moor, L. A. Akslen, P. A. Andresen, M.-F. Avril, E. Azizi, G. B. Scarrà, K. M. Brown, T. D?bniak, D. L. Duffy, D. E. Elder, S. Fang, E. Friedman, P. Galan, P. Ghiorzo, E. M. Gillanders, A. M. Goldstein, N. A. Gruis, J. Hansson, P. Helsing, M. Hočevar, V. Höiom, C. Ingvar, P. A. Kanetsky, W. V. Chen, M. T. Landi, J. Lang, G. M. Lathrop, J. Lubiński, R. M. Mackie, G. J. Mann, A. Molven, G. W. Montgomery, S. Novaković, H. Olsson, S. Puig, J. A. Puig-Butille, A. A. Qureshi, G. L. Radford-Smith, N. van der Stoep, R. van Doorn, D. C. Whiteman, J. E. Craig, D. Schadendorf, L. A. Simms, K. P. Burdon, D. R. Nyholt, K. A. Pooley, N. Orr, A. J. Stratigos, A. E. Cust, S. V. Ward, N. K. Hayward, J. Han, H.-J. Schulze, A. M. Dunning, J. A. N. Bishop, F. Demenais, C. I. Amos, S. MacGregor, and M. M. Iles, "Genome-wide meta-analysis identifies five new susceptibility loci for cutaneous malignant melanoma," Nature Genetics, vol. 47, pp. 987–995, Aug. 2015.

[58] R. J. F. Loos and G. S. H. Yeo, "The bigger picture of FTO—the first GWAS-identified obesity gene," Nature Reviews Endocrinology, vol. 10, pp. 51–61, Nov. 2013.

[59] A. Franke, D. P. B. McGovern, J. C. Barrett, K. Wang, G. L. Radford-Smith, T. Ahmad, C. W. Lees, T. Balschun, J. Lee, R. Roberts, C. A. Anderson, J. C. Bis, S. Bumpstead, D. Ellinghaus, E. M. Festen, M. Georges, T. Green, T. Haritunians, L. Jostins, A. Latiano, C. G. Mathew, G. W. Montgomery, N. J. Prescott, S. Raychaudhuri, J. I. Rotter, P. Schumm, Y. Sharma, L. A. Simms, K. D. Taylor, D. Whiteman, C. Wijmenga, R. N. Baldassano, M. Barclay, T. M. Bayless, S. Brand, C. Büning, A. Cohen, J.-F. Colombel, M. Cottone, L. Stronati, T. Denson, M. De Vos, R. D'Inca, M. Dubinsky, C. Edwards, T. Florin, D. Franchimont, R. Gearry, J. Glas, A. Van Gossum, S. L. Guthery, J. Halfvarson, H. W. Verspaget, J.-P. Hugot, A. Karban, D. Laukens, I. Lawrance, M. Lemann, A. Levine, C. Libioulle, E. Louis, C. Mowat, W. Newman, J. Panés, A. Phillips, D. D. Proctor, M. Regueiro, R. Russell, P. Rutgeerts, J. Sanderson, M. Sans, F. Seibold, A. H. Steinhart, P. C. F. Stokkers, L. Torkvist, G. Kullak-Ublick, D. Wilson, T. Walters, S. R. Targan, S. R. Brant, J. D. Rioux, M. D'Amato, R. K. Weersma, S. Kugathasan, A. M. Griffiths, J. C. Mansfield, S. Vermeire, R. H. Duerr, M. S. Silverberg, J. Satsangi, S. Schreiber, J. H. Cho, V. Annese, H. Hakonarson, M. J. Daly, and M. Parkes, "Genome-wide meta-analysis increases to 71 the number of confirmed Crohn's disease susceptibility loci," Nature Genetics, vol. 42, pp. 1118–1125, Dec. 2010.

[60] M. S. Silverberg, J. H. Cho, J. D. Rioux, D. P. B. McGovern, J. Wu, V. Annese, J.-P. Achkar, P. Goyette, R. Scott, W. Xu, M. M. Barmada, L. Klei, M. J. Daly, C. Abraham, T. M. Bayless, F. Bossa, A. M. Griffiths, A. F. Ippoliti, R. G. Lahaie, A. Latiano, P. Paré, D. D. Proctor, M. D. Regueiro, A. H. Steinhart, S. R. Targan, L. P. Schumm, E. O. Kistner, A. T. Lee, P. K. Gregersen, J. I. Rotter, S. R. Brant, K. D. Taylor, K. Roeder, and R. H. Duerr, "Ulcerative colitis–risk loci on chromosomes 1p36 and 12q15 found by genome-wide association study,"

Nature Genetics, vol. 41, pp. 216–220, Feb. 2009.

[61] D. M. Evans, C. C. A. Spencer, J. J. Pointon, Z. Su, D. Harvey, G. Kochan, U. Opperman, A. Dilthey, M. Pirinen, M. A. Stone, L. Appleton, L. Moutsianis, S. Leslie, T. Wordsworth, T. J. Kenna, T. Karaderi, G. P. Thomas, M. M. Ward, M. H. Weisman, C. Farrar, L. A. Bradbury, P. Danoy, R. D. Inman, W. Maksymowych, D. Gladman, P. Rahman, A. Morgan, H. Marzo-Ortega, P. Bowness, K. Gaffney, J. S. H. Gaston, M. Smith, J. Bruges-Armas, A.-R. Couto, R. Sorrentino, F. Paladini, M. A. Ferreira, H. Xu, Y. Liu, L. Jiang, C. Lopez-Larrea, R. Díaz-Peña, A. López-Vázquez, T. Zayats, G. Band, C. Bellenguez, H. Blackburn, J. M. Blackwell, E. Bramon, S. J. Bumpstead, J. P. Casas, A. Corvin, N. Craddock, P. Deloukas, S. Dronov, A. Duncanson, S. Edkins, C. Freeman, M. Gillman, E. Gray, R. Gwilliam, N. Hammond, S. E. Hunt, J. Jankowski, A. Jayakumar, C. Langford, J. Liddle, H. S. Markus, C. G. Mathew, O. T. McCann, M. I. McCarthy, C. N. A. Palmer, L. Peltonen, R. Plomin, S. C. Potter, A. Rautanen, R. Ravindrarajah, M. Ricketts, N. Samani, S. J. Sawcer, A. Strange, R. C. Trembath, A. C. Viswanathan, M. Waller, P. Weston, P. Whittaker, S. Widaa, N. W. Wood, G. McVean, J. D. Reveille, B. P. Wordsworth, M. A. Brown, and P. Donnelly, "Interaction between ERAP1 and HLA-B27 in ankylosing spondylitis implicates peptide handling in the mechanism for HLA-B27 in disease susceptibility," Nature Genetics, vol. 43, pp. 761–767, July 2011.

[62] Genetic Analysis of Psoriasis Consortium and The Wellcome Trust Case Control Consortium 2, "A genome-wide association study identifies new psoriasis susceptibility loci and an interaction between HLA-C and ERAP1," Nat Genet, vol. 42, pp. 985–990, Nov. 2010.

[63] R. Y. Yu, J. Brazaitis, and G. Gallagher, "The Human IL-23 Receptor rs11209026 A Allele Promotes the Expression of a Soluble IL-23r?Encoding mRNA Species," The Journal of Immunology, vol. 194, pp. 1062–1068, Feb. 2015.

[64] D. J. Cua, J. Sherlock, Y. Chen, C. A. Murphy, B. Joyce, B. Seymour, L. Lucian, W. To, S. Kwan, T. Churakova, S. Zurawski, M. Wiekowski, S. A. Lira, D. Gorman, R. A. Kastelein, and J. D. Sedgwick, "Interleukin-23 rather than interleukin-12 is the critical cytokine for autoimmune inflammation of the brain," Nature, vol. 421, pp. 744–748, Feb. 2003.

[65] S. Hue, P. Ahern, S. Buonocore, M. C. Kullberg, D. J. Cua, B. S. McKenzie, F. Powrie, and K. J. Maloy, "Interleukin-23 drives innate and T cell?mediated intestinal inflammation," The Journal of Experimental Medicine, vol. 203, pp. 2473–2483, Oct. 2006.

[66] N. Chami, M.-H. Chen, A. Slater, J. Eicher, E. Evangelou, S. Tajuddin, L. Love-Gregory, T. Kacprowski, U. Schick, A. Nomura, A. Giri, S. Lessard, J. Brody, C. Schurmann, N. Pankratz, L. Yanek, A. Manichaikul, R. Pazoki, E. Mihailov, W. Hill, L. Raffield, A. Burt, T. Bartz, D. Becker, L. Becker, E. Boerwinkle, J. Bork-Jensen, E. Bottinger, M. O?Donoghue, D. Crosslin, S. de?Denus, M.-P. Dub?, P. Elliott, G. Engstr?m, M. Evans, J. Floyd, M. Fornage, H. Gao, A. Greinacher, V. Gudnason, T. Hansen, T. Harris, C. Hayward, J. Hernesniemi, H. Highland, J. Hirschhorn, A. Hofman, M. Irvin, M. K?h?nen, E. Lange, L. Launer, T. Lehtim?ki, J. Li, D. Liewald, A. Linneberg, Y. Liu, Y. Lu, L.-P. Lyytik?inen, R. M?gi, R. Mathias, O. Melander, A. Metspalu, N. Mononen, M. Nalls, D. Nickerson, K. Nikus, C. O?Donnell, M. Orho-

Melander, O. Pedersen, A. Petersmann, L. Polfus, B. Psaty, O. Raitakari, E. Raitoharju, M. Richard, K. Rice, F. Rivadeneira, J. Rotter, F. Schmidt, A. Smith, J. Starr, K. Taylor, A. Teumer, B. Thuesen, E. Torstenson, R. Tracy, I. Tzoulaki, N. Zakai, C. Vacchi-Suzzi, C. van?Duijn, F. van?Rooij, M. Cushman, I. Deary, D. Velez?Edwards, A.-C. Vergnaud, L. Wallentin, D. Waterworth, H. White, J. Wilson, A. Zonderman, S. Kathiresan, N. Grarup, T. Esko, R. Loos, L. Lange, N. Faraday, N. Abumrad, T. Edwards, S. Ganesh, P. Auer, A. Johnson, A. Reiner, and G. Lettre, "Exome Genotyping Identifies Pleiotropic Variants Associated with Red Blood Cell Traits," The American Journal of Human Genetics, vol. 99, pp. 8–21, July 2016.

[67] R. Jones, M. Pembrey, J. Golding, and D. Herrick, "The search for genenotype/phenotype associations and the phenome scan," Paediatric and Perinatal Epidemiology, vol. 19, pp. 264–275, July 2005.

[68] J. C. Denny, M. D. Ritchie, M. A. Basford, J. M. Pulley, L. Bastarache, K. Brown-Gentry, D. Wang, D. R. Masys, D. M. Roden, and D. C. Crawford, "PheWAS: demonstrating the feasibility of a phenome-wide scan to discover gene-disease associations," Bioinformatics, vol. 26, pp. 1205–1210, May 2010.

[69] S. Pendergrass, K. Brown-Gentry, S. Dudek, E. Torstenson, J. Ambite, C. Avery, S. Buyske, C. Cai, M. Fesinmeyer, C. Haiman, G. Heiss, L. Hindorff, C.-N. Hsu, R. Jackson, C. Kooperberg, L. Le Marchand, Y. Lin, T. Matise, L. Moreland, K. Monroe, A. Reiner, R. Wallace, L. Wilkens, D. Crawford, and M. Ritchie, "The use of phenome-wide association studies (PheWAS) for exploration of novel genotype-phenotype relationships and pleiotropy discovery," Genetic Epidemiology, vol. 35, pp. 410–422, July 2011.

[70] A. Korte, B. J. Vilhjálmsson, V. Segura, A. Platt, Q. Long, and M. Nordborg, "A mixed-model approach for genome-wide association studies of correlated traits in structured populations," Nature Genetics, vol. 44, pp. 1066–1071, Aug. 2012.

[71] X. Zhou and M. Stephens, "Efficient multivariate linear mixed model algorithms for genome-wide association studies," Nature Methods, vol. 11, pp. 407–409, Feb. 2014.

[72] C. Sudlow, J. Gallacher, N. Allen, V. Beral, P. Burton, J. Danesh, P. Downey, P. Elliott, J. Green, and M. Landray, "UK biobank: an open access resource for identifying the causes of a wide range of complex diseases of middle and old age," PLoS medicine, vol. 12, no. 3, p. e1001779, 2015.

[73] Z. Chen, J. Chen, R. Collins, Y. Guo, R. Peto, F. Wu, L. Li, and on behalf of the China Kadoorie Biobank (CKB) collaborative group, "China Kadoorie Biobank of 0.5 million people: survey methods, baseline characteristics and long-term follow-up," International Journal of Epidemiology, vol. 40, pp. 1652–1666, Dec. 2011.

[74] S. J. Wang, B. Middleton, L. A. Prosser, C. G. Bardon, C. D. Spurr, P. J. Carchidi, A. F. Kittler, R. C. Goldszer, D. G. Fairchild, A. J. Sussman, G. J. Kuperman, and D. W. Bates, "A cost-benefit analysis of electronic medical records in primary care," The American Journal of Medicine, vol. 114, pp. 397–403, Apr. 2003.

[75] S. J. Hebbring, "The challenges, advantages and future of phenome-wide association studies," Immunology, vol. 141, pp. 157–165, Feb. 2014.

[76] C. A. Anderson, F. H. Pettersson, G. M. Clarke, L. R. Cardon, A. P. Morris, and K. T. Zondervan, "Data quality control in genetic case-control association

studies," Nature Protocols, vol. 5, pp. 1564–1573, Sept. 2010.
[77] The 1000 Genomes Project Consortium, "A global reference for human genetic variation," Nature, vol. 526, pp. 68–74, Oct. 2015.
[78] World Health Organization, "WHO Tobacco Fact Sheet," June 2017.
[79] C. R. Weinberg, "Toward a Clearer Definition of Confounding," American Journal of Epidemiology, vol. 137, pp. 1–8, Jan. 1993.
[80] L. R. Cardon and L. J. Palmer, "Population stratification and spurious allelic association," The Lancet, vol. 361, pp. 598–604, Feb. 2003.
[81] A. L. Price, N. A. Zaitlen, D. Reich, and N. Patterson, "New approaches to population stratification in genome-wide association studies," Nature Reviews Genetics, vol. 11, pp. 459–463, June 2010.
[82] B. Devlin, K. Roeder, and L. Wasserman, "Genomic Control, a New Approach to Genetic-Based Association Studies," Theoretical Population Biology, vol. 60, pp. 155–166, Nov. 2001.
[83] W. Astle and D. J. Balding, "Population Structure and Cryptic Relatedness in Genetic Association Studies," Statistical Science, vol. 24, no. 4, pp. 451–471, 2009.
[84] B. K. Bulik-Sullivan, P.-R. Loh, H. K. Finucane, S. Ripke, J. Yang, N. Patterson, M. J. Daly, A. L. Price, and B. M. Neale, "LD Score regression distinguishes confounding from polygenicity in genome-wide association studies," Nature Genetics, vol. 47, pp. 291–295, Feb. 2015.
[85] B. Bulik-Sullivan, H. K. Finucane, V. Anttila, A. Gusev, F. R. Day, P.-R. Loh, L. Duncan, J. R. B. Perry, N. Patterson, E. B. Robinson, M. J. Daly, A. L. Price, and B. M. Neale, "An atlas of genetic correlations across human diseases and traits," Nature Genetics, vol. 47, pp. 1236–1241, Sept. 2015.
[86] A. L. Price, N. J. Patterson, R. M. Plenge, M. E. Weinblatt, N. A. Shadick, and D. Reich, "Principal components analysis corrects for stratification in genome-wide association studies," Nature Genetics, vol. 38, pp. 904–909, Aug. 2006.
[87] J. Wakefield, "Bayes factors for genome-wide association studies: comparison with P-values," Genetic Epidemiology, vol. 33, pp. 79–86, Jan. 2009.
[88] G. M. Clarke, C. A. Anderson, F. H. Pettersson, L. R. Cardon, A. P. Morris, and K. T. Zondervan, "Basic statistical analysis in genetic case-control studies," Nature Protocols, vol. 6, pp. 121–133, Feb. 2011.
[89] H. M. Kang, N. A. Zaitlen, C. M. Wade, A. Kirby, D. Heckerman, M. J. Daly, and E. Eskin, "Efficient Control of Population Structure in Model Organism Association Mapping," Genetics, vol. 178, pp. 1709–1723, Feb. 2008.
[90] H. M. Kang, J. H. Sul, S. K. Service, N. A. Zaitlen, S.-y. Kong, N. B. Freimer, C. Sabatti, and E. Eskin, "Variance component model to account for sample structure in genome-wide association studies," Nature Genetics, vol. 42, pp. 348–354, Apr. 2010.
[91] C. Lippert, J. Listgarten, Y. Liu, C. M. Kadie, R. I. Davidson, and D. Heckerman, "FaST linear mixed models for genome-wide association studies," Nature Methods, vol. 8, pp. 833–835, Sept. 2011.
[92] M. Pirinen, P. Donnelly, and C. C. A. Spencer, "Efficient computation with a linear mixed model on large-scale data sets with applications to genetic studies," The Annals of Applied Statistics, vol. 7, pp. 369–390, Mar. 2013.
[93] M. Stephens and D. J. Balding, "Bayesian statistical methods for genetic association studies," Nature Reviews Genetics, vol. 10, pp. 681–690, Oct. 2009.

[94] J. Marchini and G. Band, "SNPTEST."
[95] R. M. Cantor, K. Lange, and J. S. Sinsheimer, "Prioritizing GWAS Results: A Review of Statistical Methods and Recommendations for Their Application," The American Journal of Human Genetics, vol. 86, pp. 6–22, Jan. 2010.
[96] C. Verzilli, T. Shah, J. P. Casas, J. Chapman, M. Sandhu, S. L. Debenham, M. S. Boekholdt, K. T. Khaw, N. J. Wareham, R. Judson, E. J. Benjamin, S. Kathiresan, M. G. Larson, J. Rong, R. Sofat, S. E. Humphries, L. Smeeth, G. Cavalleri, J. C. Whittaker, and A. D. Hingorani, "Bayesian Meta-Analysis of Genetic Association Studies with Different Sets of Markers," The American Journal of Human Genetics, vol. 82, pp. 859–872, Apr. 2008.
[97] P. J. Newcombe, C. Verzilli, J. P. Casas, A. D. Hingorani, L. Smeeth, and J. C. Whittaker, "Multilocus Bayesian Meta-Analysis of Gene-Disease Associations," The American Journal of Human Genetics, vol. 84, pp. 567–580, May 2009.
[98] J. Marchini and B. Howie, "Genotype imputation for genome-wide association studies," Nat Rev Genet, vol. 11, pp. 499–511, July 2010.
[99] The Blue Mountains Eye Study (BMES) and The Wellcome Trust Case Control Consortium 2 (WTCCC2), A. Strange, C. Bellenguez, X. Sim, R. Luben, P. G. Hysi, W. D. Ramdas, L. M. E. van Koolwijk, C. Freeman, M. Pirinen, Z. Su, G. Band, R. Pearson, D. Vukcevic, C. Langford, P. Deloukas, S. Hunt, E. Gray, S. Dronov, S. C. Potter, A. Tashakkori-Ghanbaria, S. Edkins, S. J. Bumpstead, J. M. Blackwell, E. Bramon, M. A. Brown, J. P. Casas, A. Corvin, A. Duncanson, J. A. Z. Jankowski, H. S. Markus, C. G. Mathew, C. N. A. Palmer, R. Plomin, A. Rautanen, S. J. Sawcer, R. C. Trembath, N. W. Wood, I. Barroso, L. Peltonen, P. Healey, P. McGuffin, F. Topouzis, C. C. W. Klaver, C. M. van Duijn, D. A. Mackey, T. L. Young, C. J. Hammond, K.-T. Khaw, N. Wareham, J. J. Wang, T. Y. Wong, P. J. Foster, P. Mitchell, C. C. A. Spencer, P. Donnelly, and A. C. Viswanathan, "Genome-wide association study of intraocular pressure identifies the GLCCI1/ICA1 region as a glaucoma susceptibility locus," Human Molecular Genetics, vol. 22, pp. 4653–4660, Nov. 2013.
[100] M. Nakano, Y. Ikeda, T. Taniguchi, T. Yagi, M. Fuwa, N. Omi, Y. Tokuda, M. Tanaka, K. Yoshii, M. Kageyama, S. Naruse, A. Matsuda, K. Mori, S. Kinoshita, and K. Tashiro, "Three susceptible loci associated with primary open-angle glaucoma identified by genome-wide association study in a Japanese population," Proceedings of the National Academy of Sciences, vol. 106, pp. 12838–12842, Aug. 2009.
[101] A. A. Veroniki, D. Jackson, W. Viechtbauer, R. Bender, J. Bowden, G. Knapp, O. Kuss, J. P. Higgins, D. Langan, and G. Salanti, "Methods to estimate the between-study variance and its uncertainty in meta-analysis," Research Synthesis Methods, vol. 7, pp. 55–79, Mar. 2016.
[102] B. Han and E. Eskin, "Random-Effects Model Aimed at Discovering Associations in Meta-Analysis of Genome-wide Association Studies," The American Journal of Human Genetics, vol. 88, pp. 586–598, May 2011.
[103] S. G. Self and K.-Y. Liang, "Asymptotic Properties of Maximum Likelihood Estimators and Likelihood Ratio Tests Under Nonstandard Conditions," Journal of the American Statistical Association, vol. 82, no. 398, pp. 605–610, 1987.
[104] J. Shi and S. Lee, "A novel random effect model for GWAS meta-analysis and its application to trans-ethnic meta-analysis: A Novel Random Effect Model for GWAS Meta-Analysis and Its Application to Trans-Ethnic Meta-Analysis,"

Biometrics, pp. n/a–n/a, Feb. 2016.
[105] A. P. Morris, "Transethnic meta-analysis of genomewide association studies," Genetic Epidemiology, vol. 35, pp. 809–822, Dec. 2011.
[106] R. A. S. Fisher, Statistical methods for research workers. Edinburgh Oliver & Boyd, fourth ed. - revised and enlarged ed., 1932.
[107] J. Huang, A. D. Johnson, and C. J. O'Donnell, "PRIMe: a method for characterization and evaluation of pleiotropic regions from multiple genome-wide association studies," Bioinformatics, vol. 27, pp. 1201–1206, May 2011.
[108] J. K. Pickrell, T. Berisa, J. Z. Liu, L. Segurel, J. Y. Tung, and D. A. Hinds, "Detection and interpretation of shared genetic influences on 42 human traits," Nat Genet, vol. 48, pp. 709–717, July 2016.
[109] X. Wen and M. Stephens, "Bayesian methods for genetic association analysis with heterogeneous subgroups: From meta-analyses to gene–environment interactions," The Annals of Applied Statistics, vol. 8, pp. 176–203, Mar. 2014.
[110] J. L. Asimit, K. Panoutsopoulou, E. Wheeler, S. I. Berndt, the GIANT consortium, the arcOGEN consortium, H. J. Cordell, A. P. Morris, E. Zeggini, and I. Barroso, "A Bayesian Approach to the Overlap Analysis of Epidemiologically Linked Traits," Genetic Epidemiology, vol. 39, pp. 624–634, Dec. 2015.
[111] S. Bhattacharjee, P. Rajaraman, K. B. Jacobs, W. A. Wheeler, B. S. Melin, P. Hartge, GliomaScan Consortium, M. Yeager, C. C. Chung, S. J. Chanock, and N. Chatterjee, "A Subset-Based Approach Improves Power and Interpretation for the Combined Analysis of Genetic Association Studies of Heterogeneous Traits," The American Journal of Human Genetics, vol. 90, pp. 821–835, May 2012.
[112] B. Han and E. Eskin, "Interpreting Meta-Analyses of Genome-Wide Association Studies," PLoS Genetics, vol. 8, p. e1002555, Mar. 2012.
[113] S. van der Sluis, D. Posthuma, and C. V. Dolan, "TATES: Efficient Multivariate Genotype-Phenotype Analysis for Genome-Wide Association Studies," PLoS Genetics, vol. 9, p. e1003235, Jan. 2013.
[114] J. J. Yang, J. Li, L. K. Williams, and A. Buu, "An efficient genome-wide association test for multivariate phenotypes based on the Fisher combination function," BMC Bioinformatics, vol. 17, Dec. 2016.
[115] L. Jostins and G. McVean, "Trinculo: Bayesian and frequentist multinomial logistic regression for genome-wide association studies of multi-category phenotypes," Bioinformatics, vol. 32, pp. 1898–1900, June 2016.
[116] A. Majumdar, T. Haldar, S. Bhattacharya, and J. Witte, "An efficient Bayesian meta-analysis approach for studying cross-phenotype genetic associations," bioRxiv, Jan. 2017.
[117] R. Mägi, Y. V. Suleimanov, G. M. Clarke, M. Kaakinen, K. Fischer, I. Prokopenko, and A. P. Morris, "SCOPA and META-SCOPA: software for the analysis and aggregation of genome-wide association studies of multiple correlated phenotypes," BMC Bioinformatics, vol. 18, Dec. 2017.
[118] P. Turley, R. K. Walters, O. Maghzian, A. Okbay, J. J. Lee, M. A. Fontana, T. A. Nguyen-Viet, N. A. Furlotte, P. Magnusson, S. Oskarsson, M. Johannesson, P. M. Visscher, D. Laibson, D. Cesarini, B. Neale, and D. J. Benjamin, "MTAG: Multi-Trait Analysis of GWAS," bioRxiv, Mar. 2017.
[119] J. E. Taylor, K. J. Worsley, and F. Gosselin, "Maxima of discretely sampled random fields, with an application to 'bubbles'," Biometrika, vol. 94, pp. 1–18,

Feb. 2007.
[120] R Core Team, "R: A language and environment for statistical computing.," 2013.
[121] S. Bhattacharjee, N. Chatterjee, and W. Wheeler, "ASSET: An R package for subset-based association analysis of heterogeneous traits and subtypes.," 2015.
[122] Q. Yang, H. Wu, C.-Y. Guo, and C. S. Fox, "Analyze multivariate phenotypes in genetic association studies by combining univariate association tests," Genetic Epidemiology, vol. 34, pp. 444–454, July 2010.
[123] A. P. Morris, C. M. Lindgren, E. Zeggini, N. J. Timpson, T. M. Frayling, A. T. Hattersley, and M. I. McCarthy, "A powerful approach to sub-phenotype analysis in population-based genetic association studies," Genetic Epidemiology, vol. 34, pp. 335–343, May 2010.
[124] B. J. Becker and M.-J. Wu, "The Synthesis of Regression Slopes in Meta-Analysis," Statistical Science, vol. 22, pp. 414–429, Aug. 2007.
[125] C. Bellenguez, S. Bevan, A. Gschwendtner, C. C. A. Spencer, A. I. Burgess, M. Pirinen, C. A. Jackson, M. Traylor, A. Strange, Z. Su, G. Band, P. D. Syme, R. Malik, J. Pera, B. Norrving, R. Lemmens, C. Freeman, R. Schanz, T. James, D. Poole, L. Murphy, H. Segal, L. Cortellini, Y.-C. Cheng, D. Woo, M. A. Nalls, B. Müller-Myhsok, C. Meisinger, U. Seedorf, H. Ross-Adams, S. Boonen, D. Wloch-Kopec, V. Valant, J. Slark, K. Furie, H. Delavaran, C. Langford, P. Deloukas, S. Edkins, S. Hunt, E. Gray, S. Dronov, L. Peltonen, S. Gretarsdottir, G. Thorleifsson, U. Thorsteinsdottir, K. Stefansson, G. B. Boncoraglio, E. A. Parati, J. Attia, E. Holliday, C. Levi, M.-G. Franzosi, A. Goel, A. Helgadottir, J. M. Blackwell, E. Bramon, M. A. Brown, J. P. Casas, A. Corvin, A. Duncanson, J. Jankowski, C. G. Mathew, C. N. A. Palmer, R. Plomin, A. Rautanen, S. J. Sawcer, R. C. Trembath, A. C. Viswanathan, N. W. Wood, B. B. Worrall, S. J. Kittner, B. D. Mitchell, B. Kissela, J. F. Meschia, V. Thijs, A. Lindgren, M. J. Macleod, A. Slowik, M. Walters, J. Rosand, P. Sharma, M. Farrall, C. L. M. Sudlow, P. M. Rothwell, M. Dichgans, P. Donnelly, and H. S. Markus, "Genome-wide association study identifies a variant in HDAC9 associated with large vessel ischemic stroke," Nature Genetics, vol. 44, pp. 328–333, Feb. 2012.
[126] M. W. Fagerland, S. Lydersen, and P. Laake, "The McNemar test for binary matched-pairs data: mid-p and asymptotic are better than exact conditional," BMC Medical Research Methodology, vol. 13, Dec. 2013.
[127] E. I. George and R. E. McCulloch, "Variable Selection Via Gibbs Sampling," Journal of the American Statistical Association, vol. 88, no. 423, pp. 881–889, 1993.
[128] G. Malsiner-Walli and H. Wagner, "Comparing spike and slab priors for Bayesian variable selection," Austrian Journal of Statistics, vol. 40, no. 4, pp. 241–264, 2011.
[129] P. Duchesne and P. Lafaye De Micheaux, "Computing the distribution of quadratic forms: Further comparisons between the Liu–Tang–Zhang approximation and exact methods," Computational Statistics & Data Analysis, vol. 54, pp. 858–862, Apr. 2010.
[130] V. E. Johnson and D. Rossell, "On the use of non-local prior densities in Bayesian hypothesis tests," Journal of the Royal Statistical Society: Series B (Statistical Methodology), vol. 72, pp. 143–170, Mar. 2010.
[131] D. V. Zaykin and D. O. Kozbur, "P-value based analysis for shared controls design in genome-wide association studies," Genetic Epidemiology, vol. 34,

pp. 725–738, Nov. 2010.

[132] R. DerSimonian and N. Laird, "Meta-analysis in clinical trials," Controlled Clinical Trials, vol. 7, pp. 177–188, Sept. 1986.

[133] M. Parkes, J. C. Barrett, N. J. Prescott, M. Tremelling, C. A. Anderson, S. A. Fisher, R. G. Roberts, E. R. Nimmo, F. R. Cummings, D. Soars, H. Drummond, C. W. Lees, S. A. Khawaja, R. Bagnall, D. A. Burke, C. E. Todhunter, T. Ahmad, C. M. Onnie, W. McArdle, D. Strachan, G. Bethel, C. Bryan, C. M. Lewis, P. Deloukas, A. Forbes, J. Sanderson, D. P. Jewell, J. Satsangi, J. C. Mansfield, Wellcome Trust Case Control Consortium, L. Cardon, and C. G. Mathew, "Sequence variants in the autophagy gene IRGM and multiple other replicating loci contribute to Crohn's disease susceptibility," Nature Genetics, vol. 39, pp. 830–832, July 2007.

[134] J. A. Todd, N. M. Walker, J. D. Cooper, D. J. Smyth, K. Downes, V. Plagnol, R. Bailey, S. Nejentsev, S. F. Field, F. Payne, C. E. Lowe, J. S. Szeszko, J. P. Hafler, L. Zeitels, J. H. M. Yang, A. Vella, S. Nutland, H. E. Stevens, H. Schuilenburg, G. Coleman, M. Maisuria, W. Meadows, L. J. Smink, B. Healy, O. S. Burren, A. A. C. Lam, N. R. Ovington, J. Allen, E. Adlem, H.-T. Leung, C. Wallace, J. M. M. Howson, C. Guja, C. Ionescu-Tîrgovişte, Genetics of Type 1 Diabetes in Finland, M. J. Simmonds, J. M. Heward, S. C. L. Gough, Wellcome Trust Case Control Consortium, D. B. Dunger, L. S. Wicker, and D. G. Clayton, "Robust associations of four new chromosome regions from genome-wide analyses of type 1 diabetes," Nature Genetics, vol. 39, pp. 857–864, July 2007.

[135] P. C. A. Dubois, G. Trynka, L. Franke, K. A. Hunt, J. Romanos, A. Curtotti, A. Zhernakova, G. A. R. Heap, R. Ádány, A. Aromaa, M. T. Bardella, L. H. van den Berg, N. A. Bockett, E. G. de la Concha, B. Dema, R. S. N. Fehrmann, M. Fernández-Arquero, S. Fiatal, E. Grandone, P. M. Green, H. J. M. Groen, R. Gwilliam, R. H. J. Houwen, S. E. Hunt, K. Kaukinen, D. Kelleher, I. Korponay-Szabo, K. Kurppa, P. MacMathuna, M. Mäki, M. C. Mazzilli, O. T. McCann, M. L. Mearin, C. A. Mein, M. M. Mirza, V. Mistry, B. Mora, K. I. Morley, C. J. Mulder, J. A. Murray, C. Núñez, E. Oosterom, R. A. Ophoff, I. Polanco, L. Peltonen, M. Platteel, A. Rybak, V. Salomaa, J. J. Schweizer, M. P. Sperandeo, G. J. Tack, G. Turner, J. H. Veldink, W. H. M. Verbeek, R. K. Weersma, V. M. Wolters, E. Urcelay, B. Cukrowska, L. Greco, S. L. Neuhausen, R. McManus, D. Barisani, P. Deloukas, J. C. Barrett, P. Saavalainen, C. Wijmenga, and D. A. van Heel, "Multiple common variants for celiac disease influencing immune gene expression," Nature Genetics, vol. 42, pp. 295–302, Apr. 2010.

[136] Y. Okada, C. Terao, K. Ikari, Y. Kochi, K. Ohmura, A. Suzuki, T. Kawaguchi, E. A. Stahl, F. A. S. Kurreeman, N. Nishida, H. Ohmiya, K. Myouzen, M. Takahashi, T. Sawada, Y. Nishioka, M. Yukioka, T. Matsubara, S. Wakitani, R. Teshima, S. Tohma, K. Takasugi, K. Shimada, A. Murasawa, S. Honjo, K. Matsuo, H. Tanaka, K. Tajima, T. Suzuki, T. Iwamoto, Y. Kawamura, H. Tanii, Y. Okazaki, T. Sasaki, P. K. Gregersen, L. Padyukov, J. Worthington, K. A. Siminovitch, M. Lathrop, A. Taniguchi, A. Takahashi, K. Tokunaga, M. Kubo, Y. Nakamura, N. Kamatani, T. Mimori, R. M. Plenge, H. Yamanaka, S. Momohara, R. Yamada, F. Matsuda, and K. Yamamoto, "Meta-analysis identifies nine new loci associated with rheumatoid arthritis in the Japanese popu-

lation," Nature Genetics, vol. 44, pp. 511–516, Mar. 2012.

[137] J. Z. Liu, S. van Sommeren, H. Huang, S. C. Ng, R. Alberts, A. Takahashi, S. Ripke, J. C. Lee, L. Jostins, T. Shah, S. Abedian, J. H. Cheon, J. Cho, N. E. Dayani, L. Franke, Y. Fuyuno, A. Hart, R. C. Juyal, G. Juyal, W. H. Kim, A. P. Morris, H. Poustchi, W. G. Newman, V. Midha, T. R. Orchard, H. Vahedi, A. Sood, J. Y. Sung, R. Malekzadeh, H.-J. Westra, K. Yamazaki, S.-K. Yang, International Multiple Sclerosis Genetics Consortium, International IBD Genetics Consortium, J. C. Barrett, B. Z. Alizadeh, M. Parkes, T. Bk, M. J. Daly, M. Kubo, C. A. Anderson, and R. K. Weersma, "Association analyses identify 38 susceptibility loci for inflammatory bowel disease and highlight shared genetic risk across populations," Nature Genetics, vol. 47, pp. 979–986, Sept. 2015.

[138] J. C. Barrett, D. G. Clayton, P. Concannon, B. Akolkar, J. D. Cooper, H. A. Erlich, C. Julier, G. Morahan, J. Nerup, C. Nierras, V. Plagnol, F. Pociot, H. Schuilenburg, D. J. Smyth, H. Stevens, J. A. Todd, N. M. Walker, and S. S. Rich, "Genome-wide association study and meta-analysis find that over 40 loci affect risk of type 1 diabetes," Nature Genetics, vol. 41, pp. 703–707, June 2009.

[139] E. A. Stahl, S. Raychaudhuri, E. F. Remmers, G. Xie, S. Eyre, B. P. Thomson, Y. Li, F. A. S. Kurreeman, A. Zhernakova, A. Hinks, C. Guiducci, R. Chen, L. Alfredsson, C. I. Amos, K. G. Ardlie, A. Barton, J. Bowes, E. Brouwer, N. P. Burtt, J. J. Catanese, J. Coblyn, M. J. H. Coenen, K. H. Costenbader, L. A. Criswell, J. B. A. Crusius, J. Cui, P. I. W. de Bakker, P. L. De Jager, B. Ding, P. Emery, E. Flynn, P. Harrison, L. J. Hocking, T. W. J. Huizinga, D. L. Kastner, X. Ke, A. T. Lee, X. Liu, P. Martin, A. W. Morgan, L. Padyukov, M. D. Posthumus, T. R. D. J. Radstake, D. M. Reid, M. Seielstad, M. F. Seldin, N. A. Shadick, S. Steer, P. P. Tak, W. Thomson, A. H. M. van der Helm-van Mil, I. E. van der Horst-Bruinsma, C. E. van der Schoot, P. L. C. M. van Riel, M. E. Weinblatt, A. G. Wilson, G. J. Wolbink, B. P. Wordsworth, C. Wijmenga, E. W. Karlson, R. E. M. Toes, N. de Vries, A. B. Begovich, J. Worthington, K. A. Siminovitch, P. K. Gregersen, L. Klareskog, and R. M. Plenge, "Genome-wide association study meta-analysis identifies seven new rheumatoid arthritis risk loci," Nature Genetics, vol. 42, pp. 508–514, June 2010.

[140] S. Sawcer, G. Hellenthal, M. Pirinen, C. C. A. Spencer, N. A. Patsopoulos, L. Moutsianas, A. Dilthey, Z. Su, C. Freeman, S. E. Hunt, S. Edkins, E. Gray, D. R. Booth, S. C. Potter, A. Goris, G. Band, A. Bang Oturai, A. Strange, J. Saarela, C. Bellenguez, B. Fontaine, M. Gillman, B. Hemmer, R. Gwilliam, F. Zipp, A. Jayakumar, R. Martin, S. Leslie, S. Hawkins, E. Giannoulatou, S. D'alfonso, H. Blackburn, F. Martinelli Boneschi, J. Liddle, H. F. Harbo, M. L. Perez, A. Spurkland, M. J. Waller, M. P. Mycko, M. Ricketts, M. Comabella, N. Hammond, I. Kockum, O. T. McCann, M. Ban, P. Whittaker, A. Kemppinen, P. Weston, C. Hawkins, S. Widaa, J. Zajicek, S. Dronov, N. Robertson, S. J. Bumpstead, L. F. Barcellos, R. Ravindrarajah, R. Abraham, L. Alfredsson, K. Ardlie, C. Aubin, A. Baker, K. Baker, S. E. Baranzini, L. Bergamaschi, R. Bergamaschi, A. Bernstein, A. Berthele, M. Boggild, J. P. Bradfield, D. Brassat, S. A. Broadley, D. Buck, H. Butzkueven, R. Capra, W. M. Carroll, P. Cavalla, E. G. Celius, S. Cepok, R. Chiavacci, F. Clerget-Darpoux, K. Clysters, G. Comi, M. Cossburn, I. Cournu-Rebeix, M. B. Cox, W. Cozen, B. A. C. Cree, A. H. Cross, D. Cusi, M. J. Daly, E. Davis, P. I. W. de Bakker,

M. Debouverie, M. B. D'hooghe, K. Dixon, R. Dobosi, B. Dubois, D. Ellinghaus, I. Elovaara, F. Esposito, C. Fontenille, S. Foote, A. Franke, D. Galimberti, A. Ghezzi, J. Glessner, R. Gomez, O. Gout, C. Graham, S. F. A. Grant, F. Rosa Guerini, H. Hakonarson, P. Hall, A. Hamsten, H.-P. Hartung, R. N. Heard, S. Heath, J. Hobart, M. Hoshi, C. Infante-Duarte, G. Ingram, W. Ingram, T. Islam, M. Jagodic, M. Kabesch, A. G. Kermode, T. J. Kilpatrick, C. Kim, N. Klopp, K. Koivisto, M. Larsson, M. Lathrop, J. S. Lechner-Scott, M. A. Leone, V. Leppä, U. Liljedahl, I. Lima Bomfim, R. R. Lincoln, J. Link, J. Liu, Å. R. Lorentzen, S. Lupoli, F. Macciardi, T. Mack, M. Marriott, V. Martinelli, D. Mason, J. L. McCauley, F. Mentch, I.-L. Mero, T. Mihalova, X. Montalban, J. Mottershead, K.-M. Myhr, P. Naldi, W. Ollier, A. Page, A. Palotie, J. Pelletier, L. Piccio, T. Pickersgill, F. Piehl, S. Pobywajlo, H. L. Quach, P. P. Ramsay, M. Reunanen, R. Reynolds, J. D. Rioux, M. Rodegher, S. Roesner, J. P. Rubio, I.-M. Rückert, M. Salvetti, E. Salvi, A. Santaniello, C. A. Schaefer, S. Schreiber, C. Schulze, R. J. Scott, F. Sellebjerg, K. W. Selmaj, D. Sexton, L. Shen, B. Simms-Acuna, S. Skidmore, P. M. A. Sleiman, C. Smestad, P. S. Sørensen, H. B. Søndergaard, J. Stankovich, R. C. Strange, A.-M. Sulonen, E. Sundqvist, A.-C. Syvänen, F. Taddeo, B. Taylor, J. M. Blackwell, P. Tienari, E. Bramon, A. Tourbah, M. A. Brown, E. Tronczynska, J. P. Casas, N. Tubridy, A. Corvin, J. Vickery, J. Jankowski, P. Villoslada, H. S. Markus, K. Wang, C. G. Mathew, J. Wason, C. N. A. Palmer, H.-E. Wichmann, R. Plomin, E. Willoughby, A. Rautanen, J. Winkelmann, M. Wittig, R. C. Trembath, J. Yaouanq, A. C. Viswanathan, H. Zhang, N. W. Wood, R. Zuvich, P. Deloukas, C. Langford, A. Duncanson, J. R. Oksenberg, M. A. Pericak-Vance, J. L. Haines, T. Olsson, J. Hillert, A. J. Ivinson, P. L. De Jager, L. Peltonen, G. J. Stewart, D. A. Hafler, S. L. Hauser, G. McVean, P. Donnelly, and A. Compston, "Genetic risk and a primary role for cell-mediated immune mechanisms in multiple sclerosis," Nature, vol. 476, pp. 214–219, Aug. 2011.

[141] J. Bentham, D. L. Morris, D. S. C. Graham, C. L. Pinder, P. Tombleson, T. W. Behrens, J. Martín, B. P. Fairfax, J. C. Knight, L. Chen, J. Replogle, A.-C. Syvänen, L. Rönnblom, R. R. Graham, J. E. Wither, J. D. Rioux, M. E. Alarcón-Riquelme, and T. J. Vyse, "Genetic association analyses implicate aberrant regulation of innate and adaptive immunity genes in the pathogenesis of systemic lupus erythematosus," Nature Genetics, vol. 47, pp. 1457–1464, Dec. 2015.

[142] J. C. Barrett, S. Hansoul, D. L. Nicolae, J. H. Cho, R. H. Duerr, J. D. Rioux, S. R. Brant, M. S. Silverberg, K. D. Taylor, M. M. Barmada, A. Bitton, T. Dassopoulos, L. W. Datta, T. Green, A. M. Griffiths, E. O. Kistner, M. T. Murtha, M. D. Regueiro, J. I. Rotter, L. P. Schumm, A. H. Steinhart, S. R. Targan, R. J. Xavier, C. Libioulle, C. Sandor, M. Lathrop, J. Belaiche, O. Dewit, I. Gut, S. Heath, D. Laukens, M. Mni, P. Rutgeerts, A. Van Gossum, D. Zelenika, D. Franchimont, J.-P. Hugot, M. de Vos, S. Vermeire, E. Louis, L. R. Cardon, C. A. Anderson, H. Drummond, E. Nimmo, T. Ahmad, N. J. Prescott, C. M. Onnie, S. A. Fisher, J. Marchini, J. Ghori, S. Bumpstead, R. Gwilliam, M. Tremelling, P. Deloukas, J. Mansfield, D. Jewell, J. Satsangi, C. G. Mathew, M. Parkes, M. Georges, and M. J. Daly, "Genome-wide association defines more than 30 distinct susceptibility loci for Crohn's disease," Nature Genetics, vol. 40, pp. 955–962, Aug. 2008.

[143] P. L. De Jager, X. Jia, J. Wang, P. I. W. de Bakker, L. Ottoboni, N. T. Aggarwal, L. Piccio, S. Raychaudhuri, D. Tran, C. Aubin, R. Briskin, S. Romano, S. E. Baranzini, J. L. McCauley, M. A. Pericak-Vance, J. L. Haines, R. A. Gibson, Y. Naeglin, B. Uitdehaag, P. M. Matthews, L. Kappos, C. Polman, W. L. McArdle, D. P. Strachan, D. Evans, A. H. Cross, M. J. Daly, A. Compston, S. J. Sawcer, H. L. Weiner, S. L. Hauser, D. A. Hafler, and J. R. Oksenberg, "Meta-analysis of genome scans and replication identify CD6, IRF8 and TNFRSF1a as new multiple sclerosis susceptibility loci," Nature Genetics, vol. 41, pp. 776–782, July 2009.

[144] R. P. Nair, K. C. Duffin, C. Helms, J. Ding, P. E. Stuart, D. Goldgar, J. E. Gudjonsson, Y. Li, T. Tejasvi, B.-J. Feng, A. Ruether, S. Schreiber, M. Weichenthal, D. Gladman, P. Rahman, S. J. Schrodi, S. Prahalad, S. L. Guthery, J. Fischer, W. Liao, P.-Y. Kwok, A. Menter, G. M. Lathrop, C. A. Wise, A. B. Begovich, J. J. Voorhees, J. T. Elder, G. G. Krueger, A. M. Bowcock, G. R. Abecasis, R. P. Nair, K. C. Duffin, C. Helms, J. Ding, P. E. Stuart, D. Goldgar, J. E. Gudjonsson, Y. Li, T. Tejasvi, J. Paschall, M. J. Malloy, C. R. Pullinger, J. P. Kane, J. Gardner, A. Perlmutter, A. Miner, B.-J. Feng, R. Hiremagalore, R. W. Ike, H. W. Lim, E. Christophers, T. Henseler, S. Schreiber, A. Franke, A. Ruether, M. Weichenthal, D. Gladman, P. Rahman, S. J. Schrodi, S. Prahalad, S. L. Guthery, J. Fischer, W. Liao, P.-Y. Kwok, A. Menter, G. M. Lathrop, C. Wise, A. B. Begovich, J. J. Voorhees, J. T. Elder, G. G. Krueger, A. M. Bowcock, and G. R. Abecasis, "Genome-wide scan reveals association of psoriasis with IL-23 and NF-κB pathways," Nature Genetics, vol. 41, pp. 199–204, Feb. 2009.

[145] V. Gateva, J. K. Sandling, G. Hom, K. E. Taylor, S. A. Chung, X. Sun, W. Ortmann, R. Kosoy, R. C. Ferreira, G. Nordmark, I. Gunnarsson, E. Svenungsson, L. Padyukov, G. Sturfelt, A. Jönsen, A. A. Bengtsson, S. Rantapää-Dahlqvist, E. C. Baechler, E. E. Brown, G. S. Alarcón, J. C. Edberg, R. Ramsey-Goldman, G. McGwin, J. D. Reveille, L. M. Vilá, R. P. Kimberly, S. Manzi, M. A. Petri, A. Lee, P. K. Gregersen, M. F. Seldin, L. Rönnblom, L. A. Criswell, A.-C. Syvänen, T. W. Behrens, and R. R. Graham, "A large-scale replication study identifies TNIP1, PRDM1, JAZF1, UHRF1bp1 and IL10 as risk loci for systemic lupus erythematosus," Nature Genetics, vol. 41, pp. 1228–1233, Nov. 2009.

[146] G. S. Cooper, M. L. Bynum, and E. C. Somers, "Recent insights in the epidemiology of autoimmune diseases: Improved prevalence estimates and understanding of clustering of diseases," Journal of Autoimmunity, vol. 33, pp. 197–207, Nov. 2009.

[147] M. Nikpay, A. Goel, H.-H. Won, L. M. Hall, C. Willenborg, S. Kanoni, D. Saleheen, T. Kyriakou, C. P. Nelson, J. C. Hopewell, T. R. Webb, L. Zeng, A. Dehghan, M. Alver, S. M. Armasu, K. Auro, A. Bjonnes, D. I. Chasman, S. Chen, I. Ford, N. Franceschini, C. Gieger, C. Grace, S. Gustafsson, J. Huang, S.-J. Hwang, Y. K. Kim, M. E. Kleber, K. W. Lau, X. Lu, Y. Lu, L.-P. Lyytikäinen, E. Mihailov, A. C. Morrison, N. Pervjakova, L. Qu, L. M. Rose, E. Salfati, R. Saxena, M. Scholz, A. V. Smith, E. Tikkanen, A. Uitterlinden, X. Yang, W. Zhang, W. Zhao, M. de Andrade, P. S. de Vries, N. R. van Zuydam, S. S. Anand, L. Bertram, F. Beutner, G. Dedoussis, P. Frossard, D. Gauguier, A. H. Goodall, O. Gottesman, M. Haber, B.-G. Han, J. Huang, S. Jalilzadeh, T. Kessler, I. R. König, L. Lannfelt, W. Lieb, L. Lind, C. M. Lindgren, M.-

L. Lokki, P. K. Magnusson, N. H. Mallick, N. Mehra, T. Meitinger, F.-U.-R. Memon, A. P. Morris, M. S. Nieminen, N. L. Pedersen, A. Peters, L. S. Rallidis, A. Rasheed, M. Samuel, S. H. Shah, J. Sinisalo, K. E. Stirrups, S. Trompet, L. Wang, K. S. Zaman, D. Ardissino, E. Boerwinkle, I. B. Borecki, E. P. Bottinger, J. E. Buring, J. C. Chambers, R. Collins, L. A. Cupples, J. Danesh, I. Demuth, R. Elosua, S. E. Epstein, T. Esko, M. F. Feitosa, O. H. Franco, M. G. Franzosi, C. B. Granger, D. Gu, V. Gudnason, A. S. Hall, A. Hamsten, T. B. Harris, S. L. Hazen, C. Hengstenberg, A. Hofman, E. Ingelsson, C. Iribarren, J. W. Jukema, P. J. Karhunen, B.-J. Kim, J. S. Kooner, I. J. Kullo, T. Lehtimäki, R. J. F. Loos, O. Melander, A. Metspalu, W. März, C. N. Palmer, M. Perola, T. Quertermous, D. J. Rader, P. M. Ridker, S. Ripatti, R. Roberts, V. Salomaa, D. K. Sanghera, S. M. Schwartz, U. Seedorf, A. F. Stewart, D. J. Stott, J. Thiery, P. A. Zalloua, C. J. O'Donnell, M. P. Reilly, T. L. Assimes, J. R. Thompson, J. Erdmann, R. Clarke, H. Watkins, S. Kathiresan, R. McPherson, P. Deloukas, H. Schunkert, N. J. Samani, and M. Farrall, "A comprehensive 1,000 Genomes-based genome-wide association meta-analysis of coronary artery disease," Nature Genetics, vol. 47, pp. 1121–1130, Oct. 2015.

[148] F. R. Schumacher, S. L. Schmit, S. Jiao, C. K. Edlund, H. Wang, B. Zhang, L. Hsu, S.-C. Huang, C. P. Fischer, J. F. Harju, G. E. Idos, F. Lejbkowicz, F. J. Manion, K. McDonnell, C. E. McNeil, M. Melas, H. S. Rennert, W. Shi, D. C. Thomas, D. J. Van Den Berg, C. M. Hutter, A. K. Aragaki, K. Butterbach, B. J. Caan, C. S. Carlson, S. J. Chanock, K. R. Curtis, C. S. Fuchs, M. Gala, E. L. Giovannucc, E. L. Giocannucci, S. M. Gogarten, R. B. Hayes, B. Henderson, D. J. Hunter, R. D. Jackson, L. N. Kolonel, C. Kooperberg, S. Küry, S. Kury, A. LaCroix, C. C. Laurie, C. A. Laurie, M. Lemire, M. Lemire, D. Levine, J. Ma, K. W. Makar, C. Qu, D. Taverna, C. M. Ulrich, K. Wu, S. Kono, D. W. West, S. I. Berndt, S. Bezieau, H. Brenner, P. T. Campbell, A. T. Chan, J. Chang-Claude, G. A. Coetzee, D. V. Conti, D. Duggan, J. C. Figueiredo, B. K. Fortini, S. J. Gallinger, W. J. Gauderman, G. Giles, R. Green, R. Haile, T. A. Harrison, M. Hoffmeister, J. L. Hopper, T. J. Hudson, E. Jacobs, M. Iwasaki, S. H. Jee, M. Jenkins, W.-H. Jia, A. Joshi, L. Li, N. M. Lindor, K. Matsuo, V. Moreno, B. Mukherjee, P. A. Newcomb, J. D. Potter, L. Raskin, G. Rennert, S. Rosse, G. Severi, R. E. Schoen, D. Seminara, X.-O. Shu, M. L. Slattery, S. Tsugane, E. White, Y.-B. Xiang, B. W. Zanke, W. Zheng, L. Le Marchand, G. Casey, S. B. Gruber, and U. Peters, "Genome-wide association study of colorectal cancer identifies six new susceptibility loci," Nature Communications, vol. 6, p. 7138, July 2015.

[149] J. V. Raelson, R. D. Little, A. Ruether, H. Fournier, B. Paquin, P. Van Eerdewegh, W. E. C. Bradley, P. Croteau, Q. Nguyen-Huu, J. Segal, S. Debrus, R. Allard, P. Rosenstiel, A. Franke, G. Jacobs, S. Nikolaus, J.-M. Vidal, P. Szego, N. Laplante, H. F. Clark, R. J. Paulussen, J. W. Hooper, T. P. Keith, A. Belouchi, and S. Schreiber, "Genome-wide association study for Crohn's disease in the Quebec Founder Population identifies multiple validated disease loci," Proceedings of the National Academy of Sciences of the United States of America, vol. 104, pp. 14747–14752, Sept. 2007.

[150] A. Zhernakova, E. A. Stahl, G. Trynka, S. Raychaudhuri, E. A. Festen, L. Franke, H.-J. Westra, R. S. N. Fehrmann, F. A. S. Kurreeman, B. Thomson, N. Gupta, J. Romanos, R. McManus, A. W. Ryan, G. Turner, E. Brouwer, M. D.

Posthumus, E. F. Remmers, F. Tucci, R. Toes, E. Grandone, M. C. Mazzilli, A. Rybak, B. Cukrowska, M. J. H. Coenen, T. R. D. J. Radstake, P. L. C. M. van Riel, Y. Li, P. I. W. de Bakker, P. K. Gregersen, J. Worthington, K. A. Siminovitch, L. Klareskog, T. W. J. Huizinga, C. Wijmenga, and R. M. Plenge, "Meta-analysis of genome-wide association studies in celiac disease and rheumatoid arthritis identifies fourteen non-HLA shared loci," PLoS genetics, vol. 7, p. e1002004, Feb. 2011.

[151] D. P. B. McGovern, A. Gardet, L. Törkvist, P. Goyette, J. Essers, K. D. Taylor, B. M. Neale, R. T. H. Ong, C. Lagacé, C. Li, T. Green, C. R. Stevens, C. Beauchamp, P. R. Fleshner, M. Carlson, M. D'Amato, J. Halfvarson, M. L. Hibberd, M. Lördal, L. Padyukov, A. Andriulli, E. Colombo, A. Latiano, O. Palmieri, E.-J. Bernard, C. Deslandres, D. W. Hommes, D. J. de Jong, P. C. Stokkers, R. K. Weersma, NIDDK IBD Genetics Consortium, Y. Sharma, M. S. Silverberg, J. H. Cho, J. Wu, K. Roeder, S. R. Brant, L. P. Schumm, R. H. Duerr, M. C. Dubinsky, N. L. Glazer, T. Haritunians, A. Ippoliti, G. Y. Melmed, D. S. Siscovick, E. A. Vasiliauskas, S. R. Targan, V. Annese, C. Wijmenga, S. Pettersson, J. I. Rotter, R. J. Xavier, M. J. Daly, J. D. Rioux, and M. Seielstad, "Genome-wide association identifies multiple ulcerative colitis susceptibility loci," Nature Genetics, vol. 42, pp. 332–337, Apr. 2010.

[152] R. C. Betz, L. Petukhova, S. Ripke, H. Huang, A. Menelaou, S. Redler, T. Becker, S. Heilmann, T. Yamany, M. Duvic, M. Hordinsky, D. Norris, V. H. Price, J. Mackay-Wiggan, A. de Jong, G. M. DeStefano, S. Moebus, M. Böhm, U. Blume-Peytavi, H. Wolff, G. Lutz, R. Kruse, L. Bian, C. I. Amos, A. Lee, P. K. Gregersen, B. Blaumeiser, D. Altshuler, R. Clynes, P. I. W. de Bakker, M. M. Nöthen, M. J. Daly, and A. M. Christiano, "Genome-wide meta-analysis in alopecia areata resolves HLA associations and reveals two new susceptibility loci," Nature Communications, vol. 6, p. 5966, Jan. 2015.

[153] M. C. O'Donovan, N. Craddock, N. Norton, H. Williams, T. Peirce, V. Moskvina, I. Nikolov, M. Hamshere, L. Carroll, L. Georgieva, S. Dwyer, P. Holmans, J. L. Marchini, C. C. A. Spencer, B. Howie, H.-T. Leung, A. M. Hartmann, H.-J. Möller, D. W. Morris, Y. Shi, G. Feng, P. Hoffmann, P. Propping, C. Vasilescu, W. Maier, M. Rietschel, S. Zammit, J. Schumacher, E. M. Quinn, T. G. Schulze, N. M. Williams, I. Giegling, N. Iwata, M. Ikeda, A. Darvasi, S. Shifman, L. He, J. Duan, A. R. Sanders, D. F. Levinson, P. V. Gejman, S. Cichon, M. M. Nöthen, M. Gill, A. Corvin, D. Rujescu, G. Kirov, M. J. Owen, N. G. Buccola, B. J. Mowry, R. Freedman, F. Amin, D. W. Black, J. M. Silverman, W. F. Byerley, C. R. Cloninger, and Molecular Genetics of Schizophrenia Collaboration, "Identification of loci associated with schizophrenia by genome-wide association and follow-up," Nature Genetics, vol. 40, pp. 1053–1055, Sept. 2008.

[154] S. Ripke, C. O'Dushlaine, K. Chambert, J. L. Moran, A. K. Kähler, S. Akterin, S. E. Bergen, A. L. Collins, J. J. Crowley, M. Fromer, Y. Kim, S. H. Lee, P. K. E. Magnusson, N. Sanchez, E. A. Stahl, S. Williams, N. R. Wray, K. Xia, F. Bettella, A. D. Borglum, B. K. Bulik-Sullivan, P. Cormican, N. Craddock, C. de Leeuw, N. Durmishi, M. Gill, V. Golimbet, M. L. Hamshere, P. Holmans, D. M. Hougaard, K. S. Kendler, K. Lin, D. W. Morris, O. Mors, P. B. Mortensen, B. M. Neale, F. A. O'Neill, M. J. Owen, M. P. Milovancevic, D. Posthuma, J. Powell, A. L. Richards, B. P. Riley, D. Ruderfer, D. Rujescu, E. Sigurdsson, T. Silagadze, A. B. Smit, H. Stefansson, S. Steinberg, J. Suvisaari, S. Tosato,

M. Verhage, J. T. Walters, Multicenter Genetic Studies of Schizophrenia Consortium, D. F. Levinson, P. V. Gejman, K. S. Kendler, C. Laurent, B. J. Mowry, M. C. O'Donovan, M. J. Owen, A. E. Pulver, B. P. Riley, S. G. Schwab, D. B. Wildenauer, F. Dudbridge, P. Holmans, J. Shi, M. Albus, M. Alexander, D. Campion, D. Cohen, D. Dikeos, J. Duan, P. Eichhammer, S. Godard, M. Hansen, F. B. Lerer, K.-Y. Liang, W. Maier, J. Mallet, D. A. Nertney, G. Nestadt, N. Norton, F. A. O'Neill, G. N. Papadimitriou, R. Ribble, A. R. Sanders, J. M. Silverman, D. Walsh, N. M. Williams, B. Wormley, Psychosis Endophenotypes International Consortium, M. J. Arranz, S. Bakker, S. Bender, E. Bramon, D. Collier, B. Crespo-Facorro, J. Hall, C. Iyegbe, A. Jablensky, R. S. Kahn, L. Kalaydjieva, S. Lawrie, C. M. Lewis, K. Lin, D. H. Linszen, I. Mata, A. McIntosh, R. M. Murray, R. A. Ophoff, J. Powell, D. Rujescu, J. Van Os, M. Walshe, M. Weisbrod, D. Wiersma, Wellcome Trust Case Control Consortium 2, P. Donnelly, I. Barroso, J. M. Blackwell, E. Bramon, M. A. Brown, J. P. Casas, A. P. Corvin, P. Deloukas, A. Duncanson, J. Jankowski, H. S. Markus, C. G. Mathew, C. N. A. Palmer, R. Plomin, A. Rautanen, S. J. Sawcer, R. C. Trembath, A. C. Viswanathan, N. W. Wood, C. C. A. Spencer, G. Band, C. Bellenguez, C. Freeman, G. Hellenthal, E. Giannoulatou, M. Pirinen, R. D. Pearson, A. Strange, Z. Su, D. Vukcevic, P. Donnelly, C. Langford, S. E. Hunt, S. Edkins, R. Gwilliam, H. Blackburn, S. J. Bumpstead, S. Dronov, M. Gillman, E. Gray, N. Hammond, A. Jayakumar, O. T. McCann, J. Liddle, S. C. Potter, R. Ravindrarajah, M. Ricketts, A. Tashakkori-Ghanbaria, M. J. Waller, P. Weston, S. Widaa, P. Whittaker, I. Barroso, P. Deloukas, C. G. Mathew, J. M. Blackwell, M. A. Brown, A. P. Corvin, M. I. McCarthy, C. C. A. Spencer, E. Bramon, A. P. Corvin, M. C. O'Donovan, K. Stefansson, E. Scolnick, S. Purcell, S. A. McCarroll, P. Sklar, C. M. Hultman, and P. F. Sullivan, "Genome-wide association analysis identifies 13 new risk loci for schizophrenia," Nature Genetics, vol. 45, pp. 1150–1159, Oct. 2013.

[155] M. A. R. Ferreira, M. C. O'Donovan, Y. A. Meng, I. R. Jones, D. M. Ruderfer, L. Jones, J. Fan, G. Kirov, R. H. Perlis, E. K. Green, J. W. Smoller, D. Grozeva, J. Stone, I. Nikolov, K. Chambert, M. L. Hamshere, V. L. Nimgaonkar, V. Moskvina, M. E. Thase, S. Caesar, G. S. Sachs, J. Franklin, K. Gordon-Smith, K. G. Ardlie, S. B. Gabriel, C. Fraser, B. Blumenstiel, M. Defelice, G. Breen, M. Gill, D. W. Morris, A. Elkin, W. J. Muir, K. A. McGhee, R. Williamson, D. J. MacIntyre, A. W. MacLean, C. D. St, M. Robinson, M. Van Beck, A. C. P. Pereira, R. Kandaswamy, A. McQuillin, D. A. Collier, N. J. Bass, A. H. Young, J. Lawrence, I. N. Ferrier, A. Anjorin, A. Farmer, D. Curtis, E. M. Scolnick, P. McGuffin, M. J. Daly, A. P. Corvin, P. A. Holmans, D. H. Blackwood, H. M. Gurling, M. J. Owen, S. M. Purcell, P. Sklar, N. Craddock, and Wellcome Trust Case Control Consortium, "Collaborative genome-wide association analysis supports a role for ANK3 and CACNA1c in bipolar disorder," Nature Genetics, vol. 40, pp. 1056–1058, Sept. 2008.

[156] W. Yang, H. Tang, Y. Zhang, X. Tang, J. Zhang, L. Sun, J. Yang, Y. Cui, L. Zhang, N. Hirankarn, H. Cheng, H.-F. Pan, J. Gao, T. L. Lee, Y. Sheng, C. S. Lau, Y. Li, T. M. Chan, X. Yin, D. Ying, Q. Lu, A. M. H. Leung, X. Zuo, X. Chen, K. L. Tong, F. Zhou, Q. Diao, N. K. C. Tse, H. Xie, C. C. Mok, F. Hao, S. N. Wong, B. Shi, K. W. Lee, Y. Hui, M. H. K. Ho, B. Liang, P. P. W. Lee, H. Cui, Q. Guo, B. H.-Y. Chung, X. Pu, Q. Liu, X. Zhang,

C. Zhang, C. Y. Chong, H. Fang, R. W. S. Wong, Y. Sun, M. Y. Mok, X.-P. Li, Y. Avihingsanon, Z. Zhai, P. Rianthavorn, T. Deekajorndej, K. Suphapeetiporn, F. Gao, V. Shotelersuk, X. Kang, S. K. Y. Ying, L. Zhang, W. H. S. Wong, D. Zhu, S. K. S. Fung, F. Zeng, W. M. Lai, C.-M. Wong, I. O. L. Ng, M.-M. Garcia-Barceló, S. S. Cherny, N. Shen, P. K.-H. Tam, P. C. Sham, D.-Q. Ye, S. Yang, X. Zhang, and Y. L. Lau, "Meta-analysis followed by replication identifies loci in or near CDKN1b, TET3, CD80, DRAM1, and ARID5b as associated with systemic lupus erythematosus in Asians," American Journal of Human Genetics, vol. 92, pp. 41–51, Jan. 2013.

[157] J.-W. Han, H.-F. Zheng, Y. Cui, L.-D. Sun, D.-Q. Ye, Z. Hu, J.-H. Xu, Z.-M. Cai, W. Huang, G.-P. Zhao, H.-F. Xie, H. Fang, Q.-J. Lu, J.-H. Xu, X.-P. Li, Y.-F. Pan, D.-Q. Deng, F.-Q. Zeng, Z.-Z. Ye, X.-Y. Zhang, Q.-W. Wang, F. Hao, L. Ma, X.-B. Zuo, F.-S. Zhou, W.-H. Du, Y.-L. Cheng, J.-Q. Yang, S.-K. Shen, J. Li, Y.-J. Sheng, X.-X. Zuo, W.-F. Zhu, F. Gao, P.-L. Zhang, Q. Guo, B. Li, M. Gao, F.-L. Xiao, C. Quan, C. Zhang, Z. Zhang, K.-J. Zhu, Y. Li, D.-Y. Hu, W.-S. Lu, J.-L. Huang, S.-X. Liu, H. Li, Y.-Q. Ren, Z.-X. Wang, C.-J. Yang, P.-G. Wang, W.-M. Zhou, Y.-M. Lv, A.-P. Zhang, S.-Q. Zhang, D. Lin, Y. Li, H. Q. Low, M. Shen, Z.-F. Zhai, Y. Wang, F.-Y. Zhang, S. Yang, J.-J. Liu, and X.-J. Zhang, "Genome-wide association study in a Chinese Han population identifies nine new susceptibility loci for systemic lupus erythematosus," Nature Genetics, vol. 41, pp. 1234–1237, Nov. 2009.

[158] L. Jostins, S. Ripke, R. K. Weersma, R. H. Duerr, D. P. McGovern, K. Y. Hui, J. C. Lee, L. P. Schumm, Y. Sharma, C. A. Anderson, J. Essers, M. Mitrovic, K. Ning, I. Cleynen, E. Theatre, S. L. Spain, S. Raychaudhuri, P. Goyette, Z. Wei, C. Abraham, J.-P. Achkar, T. Ahmad, L. Amininejad, A. N. Ananthakrishnan, V. Andersen, J. M. Andrews, L. Baidoo, T. Balschun, P. A. Bampton, A. Bitton, G. Boucher, S. Brand, C. Büning, A. Cohain, S. Cichon, M. D'Amato, D. De Jong, K. L. Devaney, M. Dubinsky, C. Edwards, D. Ellinghaus, L. R. Ferguson, D. Franchimont, K. Fransen, R. Gearry, M. Georges, C. Gieger, J. Glas, T. Haritunians, A. Hart, C. Hawkey, M. Hedl, X. Hu, T. H. Karlsen, L. Kupcinskas, S. Kugathasan, A. Latiano, D. Laukens, I. C. Lawrance, C. W. Lees, E. Louis, G. Mahy, J. Mansfield, A. R. Morgan, C. Mowat, W. Newman, O. Palmieri, C. Y. Ponsioen, U. Potocnik, N. J. Prescott, M. Regueiro, J. I. Rotter, R. K. Russell, J. D. Sanderson, M. Sans, J. Satsangi, S. Schreiber, L. A. Simms, J. Sventoraityte, S. R. Targan, K. D. Taylor, M. Tremelling, H. W. Verspaget, M. De Vos, C. Wijmenga, D. C. Wilson, J. Winkelmann, R. J. Xavier, S. Zeissig, B. Zhang, C. K. Zhang, H. Zhao, International IBD Genetics Consortium (IIBDGC), M. S. Silverberg, V. Annese, H. Hakonarson, S. R. Brant, G. Radford-Smith, C. G. Mathew, J. D. Rioux, E. E. Schadt, M. J. Daly, A. Franke, M. Parkes, S. Vermeire, J. C. Barrett, and J. H. Cho, "Host-microbe interactions have shaped the genetic architecture of inflammatory bowel disease," Nature, vol. 491, pp. 119–124, Nov. 2012.

[159] C. A. Anderson, G. Boucher, C. W. Lees, A. Franke, M. D'Amato, K. D. Taylor, J. C. Lee, P. Goyette, M. Imielinski, A. Latiano, C. Lagacé, R. Scott, L. Amininejad, S. Bumpstead, L. Baidoo, R. N. Baldassano, M. Barclay, T. M. Bayless, S. Brand, C. Büning, J.-F. Colombel, L. A. Denson, M. De Vos, M. Dubinsky, C. Edwards, D. Ellinghaus, R. S. N. Fehrmann, J. A. B. Floyd, T. Florin, D. Franchimont, L. Franke, M. Georges, J. Glas, N. L. Glazer, S. L.

Guthery, T. Haritunians, N. K. Hayward, J.-P. Hugot, G. Jobin, D. Laukens, I. Lawrance, M. Lémann, A. Levine, C. Libioulle, E. Louis, D. P. McGovern, M. Milla, G. W. Montgomery, K. I. Morley, C. Mowat, A. Ng, W. Newman, R. A. Ophoff, L. Papi, O. Palmieri, L. Peyrin-Biroulet, J. Panés, A. Phillips, N. J. Prescott, D. D. Proctor, R. Roberts, R. Russell, P. Rutgeerts, J. Sanderson, M. Sans, P. Schumm, F. Seibold, Y. Sharma, L. A. Simms, M. Seielstad, A. H. Steinhart, S. R. Targan, L. H. van den Berg, M. Vatn, H. Verspaget, T. Walters, C. Wijmenga, D. C. Wilson, H.-J. Westra, R. J. Xavier, Z. Z. Zhao, C. Y. Ponsioen, V. Andersen, L. Torkvist, M. Gazouli, N. P. Anagnou, T. H. Karlsen, L. Kupcinskas, J. Sventoraityte, J. C. Mansfield, S. Kugathasan, M. S. Silverberg, J. Halfvarson, J. I. Rotter, C. G. Mathew, A. M. Griffiths, R. Gearry, T. Ahmad, S. R. Brant, M. Chamaillard, J. Satsangi, J. H. Cho, S. Schreiber, M. J. Daly, J. C. Barrett, M. Parkes, V. Annese, H. Hakonarson, G. Radford-Smith, R. H. Duerr, S. Vermeire, R. K. Weersma, and J. D. Rioux, "Meta-analysis identifies 29 additional ulcerative colitis risk loci, increasing the number of confirmed associations to 47," Nature Genetics, vol. 43, pp. 246–252, Mar. 2011.

[160] J. C. Barrett, J. C. Lee, C. W. Lees, N. J. Prescott, C. A. Anderson, A. Phillips, E. Wesley, K. Parnell, H. Zhang, H. Drummond, E. R. Nimmo, D. Massey, K. Blaszczyk, T. Elliott, L. Cotterill, H. Dallal, A. J. Lobo, C. Mowat, J. D. Sanderson, D. P. Jewell, W. G. Newman, C. Edwards, T. Ahmad, J. C. Mansfield, J. Satsangi, M. Parkes, C. G. Mathew, P. Donnelly, L. Peltonen, J. M. Blackwell, E. Bramon, M. A. Brown, J. P. Casas, A. Corvin, N. Craddock, P. Deloukas, A. Duncanson, J. Jankowski, H. S. Markus, C. G. Mathew, M. I. McCarthy, C. N. A. Palmer, R. Plomin, A. Rautanen, S. J. Sawcer, N. Samani, R. C. Trembath, A. C. Viswanathan, N. Wood, C. C. A. Spencer, J. C. Barrett, C. Bellenguez, D. Davison, C. Freeman, A. Strange, P. Donnelly, C. Langford, S. E. Hunt, S. Edkins, R. Gwilliam, H. Blackburn, S. J. Bumpstead, S. Dronov, M. Gillman, E. Gray, N. Hammond, A. Jayakumar, O. T. McCann, J. Liddle, M. L. Perez, S. C. Potter, R. Ravindrarajah, M. Ricketts, M. Waller, P. Weston, S. Widaa, P. Whittaker, P. Deloukas, L. Peltonen, C. G. Mathew, J. M. Blackwell, M. A. Brown, A. Corvin, M. I. McCarthy, C. C. A. Spencer, A. P. Attwood, J. Stephens, J. Sambrook, W. H. Ouwehand, W. L. McArdle, S. M. Ring, and D. P. Strachan, "Genome-wide association study of ulcerative colitis identifies three new susceptibility loci, including the HNF4a region," Nature Genetics, vol. 41, pp. 1330–1334, Dec. 2009.

[161] D. Chang and A. Keinan, "Principal Component Analysis Characterizes Shared Pathogenetics from Genome-Wide Association Studies," PLoS Computational Biology, vol. 10, p. e1003820, Sept. 2014.

[162] A. Cortes, J. Hadler, J. P. Pointon, P. C. Robinson, T. Karaderi, P. Leo, K. Cremin, K. Pryce, J. Harris, S. Lee, K. B. Joo, S.-C. Shim, M. Weisman, M. Ward, X. Zhou, H.-J. Garchon, G. Chiocchia, J. Nossent, B. A. Lie, Ø. Førre, J. Tuomilehto, K. Laiho, L. Jiang, Y. Liu, X. Wu, L. A. Bradbury, D. Elewaut, R. Burgos-Vargas, S. Stebbings, L. Appleton, C. Farrah, J. Lau, T. J. Kenna, N. Haroon, M. A. Ferreira, J. Yang, J. Mulero, J. L. Fernandez-Sueiro, M. A. Gonzalez-Gay, C. Lopez-Larrea, P. Deloukas, P. Donnelly, P. Bowness, K. Gafney, H. Gaston, D. D. Gladman, P. Rahman, W. P. Maksymowych, H. Xu, J. B. A. Crusius, I. E. van der Horst-Bruinsma, C.-T. Chou, R. Valle-

Oñate, C. Romero-Sánchez, I. M. Hansen, F. M. Pimentel-Santos, R. D. Inman, V. Videm, J. Martin, M. Breban, J. D. Reveille, D. M. Evans, T.-H. Kim, B. P. Wordsworth, and M. A. Brown, "Identification of multiple risk variants for ankylosing spondylitis through high-density genotyping of immune-related loci," Nature Genetics, vol. 45, pp. 730–738, June 2013.

[163] The UK Parkinson's Disease Consortium and The Wellcome Trust Case Control Consortium 2, C. C. A. Spencer, V. Plagnol, A. Strange, M. Gardner, C. Paisan-Ruiz, G. Band, R. A. Barker, C. Bellenguez, K. Bhatia, H. Blackburn, J. M. Blackwell, E. Bramon, M. A. Brown, M. A. Brown, D. Burn, J.-P. Casas, P. F. Chinnery, C. E. Clarke, A. Corvin, N. Craddock, P. Deloukas, S. Edkins, J. Evans, C. Freeman, E. Gray, J. Hardy, G. Hudson, S. Hunt, J. Jankowski, C. Langford, A. J. Lees, H. S. Markus, C. G. Mathew, M. I. McCarthy, K. E. Morrison, C. N. A. Palmer, J. P. Pearson, L. Peltonen, M. Pirinen, R. Plomin, S. Potter, A. Rautanen, S. J. Sawcer, Z. Su, R. C. Trembath, A. C. Viswanathan, N. W. Williams, H. R. Morris, P. Donnelly, and N. W. Wood, "Dissection of the genetics of Parkinson's disease identifies an additional association 5' of SNCA and multiple associated haplotypes at 17q21," Human Molecular Genetics, vol. 20, pp. 345–353, Jan. 2011.

[164] K. Zhou, C. Bellenguez, C. C. A. Spencer, A. J. Bennett, R. L. Coleman, R. Tavendale, S. A. Hawley, L. A. Donnelly, C. Schofield, C. J. Groves, L. Burch, F. Carr, A. Strange, C. Freeman, J. M. Blackwell, E. Bramon, M. A. Brown, J. P. Casas, A. Corvin, N. Craddock, P. Deloukas, S. Dronov, A. Duncanson, S. Edkins, E. Gray, S. Hunt, J. Jankowski, C. Langford, H. S. Markus, C. G. Mathew, R. Plomin, A. Rautanen, S. J. Sawcer, N. J. Samani, R. Trembath, A. C. Viswanathan, N. W. Wood, L. W. Harries, A. T. Hattersley, A. S. F. Doney, H. Colhoun, A. D. Morris, C. Sutherland, D. G. Hardie, L. Peltonen, M. I. McCarthy, R. R. Holman, C. N. A. Palmer, P. Donnelly, and E. R. Pearson, "Common variants near ATM are associated with glycemic response to metformin in type 2 diabetes," Nature Genetics, vol. 43, pp. 117–120, Feb. 2011.

[165] Z. Su, L. J. Gay, A. Strange, C. Palles, G. Band, D. C. Whiteman, F. Lescai, C. Langford, M. Nanji, S. Edkins, A. van der Winkel, D. Levine, P. Sasieni, C. Bellenguez, K. Howarth, C. Freeman, N. Trudgill, A. T. Tucker, M. Pirinen, M. P. Peppelenbosch, L. J. W. van der Laan, E. J. Kuipers, J. P. H. Drenth, W. H. Peters, J. V. Reynolds, D. P. Kelleher, R. McManus, H. Grabsch, H. Prenen, R. Bisschops, K. Krishnadath, P. D. Siersema, J. W. P. M. van Baal, M. Middleton, R. Petty, R. Gillies, N. Burch, P. Bhandari, S. Paterson, C. Edwards, I. Penman, K. Vaidya, Y. Ang, I. Murray, P. Patel, W. Ye, P. Mullins, A. H. Wu, N. C. Bird, H. Dallal, N. J. Shaheen, L. J. Murray, K. Koss, L. Bernstein, Y. Romero, L. J. Hardie, R. Zhang, H. Winter, D. A. Corley, S. Panter, H. A. Risch, B. J. Reid, I. Sargeant, M. D. Gammon, H. Smart, A. Dhar, H. McMurtry, H. Ali, G. Liu, A. G. Casson, W.-H. Chow, M. Rutter, A. Tawil, D. Morris, C. Nwokolo, P. Isaacs, C. Rodgers, K. Ragunath, C. MacDonald, C. Haigh, D. Monk, G. Davies, S. Wajed, D. Johnston, M. Gibbons, S. Cullen, N. Church, R. Langley, M. Griffin, D. Alderson, P. Deloukas, S. E. Hunt, E. Gray, S. Dronov, S. C. Potter, A. Tashakkori-Ghanbaria, M. Anderson, C. Brooks, J. M. Blackwell, E. Bramon, M. A. Brown, J. P. Casas, A. Corvin, A. Duncanson, H. S. Markus, C. G. Mathew, C. N. A. Palmer, R. Plomin, A. Rautanen, S. J. Sawcer,

R. C. Trembath, A. C. Viswanathan, N. Wood, G. Trynka, C. Wijmenga, J.-B. Cazier, P. Atherfold, A. M. Nicholson, N. L. Gellatly, D. Glancy, S. C. Cooper, D. Cunningham, T. Lind, J. Hapeshi, D. Ferry, B. Rathbone, J. Brown, S. Love, S. Attwood, S. MacGregor, P. Watson, S. Sanders, W. Ek, R. F. Harrison, P. Moayyedi, J. de Caestecker, H. Barr, E. Stupka, T. L. Vaughan, L. Peltonen, C. C. A. Spencer, I. Tomlinson, P. Donnelly, and J. A. Z. Jankowski, "Common variants at the MHC locus and at chromosome 16q24.1 predispose to Barrett's esophagus," Nature Genetics, vol. 44, pp. 1131–1136, Sept. 2012.

[166] Irish Schizophrenia Genomics Consortium and Wellcome Trust Case Control Consortium 2, "Genome-Wide Association Study Implicates HLA-C*01:02 as a Risk Factor at the Major Histocompatibility Complex Locus in Schizophrenia," Biological Psychiatry, vol. 72, pp. 620–628, Oct. 2012.

[167] M. Fakiola, A. Strange, H. J. Cordell, E. N. Miller, M. Pirinen, Z. Su, A. Mishra, S. Mehrotra, G. R. Monteiro, G. Band, C. Bellenguez, S. Dronov, S. Edkins, C. Freeman, E. Giannoulatou, E. Gray, S. E. Hunt, H. G. Lacerda, C. Langford, R. Pearson, N. N. Pontes, M. Rai, S. P. Singh, L. Smith, O. Sousa, D. Vukcevic, E. Bramon, M. A. Brown, J. P. Casas, A. Corvin, A. Duncanson, J. Jankowski, H. S. Markus, C. G. Mathew, C. N. A. Palmer, R. Plomin, A. Rautanen, S. J. Sawcer, R. C. Trembath, A. C. Viswanathan, N. W. Wood, M. E. Wilson, P. Deloukas, L. Peltonen, F. Christiansen, C. Witt, S. M. B. Jeronimo, S. Sundar, C. C. A. Spencer, J. M. Blackwell, and P. Donnelly, "Common variants in the HLA-DRB1–HLA-DQA1 HLA class II region are associated with susceptibility to visceral leishmaniasis," Nature Genetics, vol. 45, pp. 208–213, Jan. 2013.

[168] Psychosis Endophenotypes International Consortium and Wellcome Trust Case-Control Consortium 2, "A Genome-wide Association Analysis of a Broad Psychosis Phenotype Identifies Three Loci for Further Investigation," Biological Psychiatry, vol. 75, pp. 386–397, Mar. 2014.

[169] O. S. P. Davis, G. Band, M. Pirinen, C. M. A. Haworth, E. L. Meaburn, Y. Kovas, N. Harlaar, S. J. Docherty, K. B. Hanscombe, M. Trzaskowski, C. J. C. Curtis, A. Strange, C. Freeman, C. Bellenguez, Z. Su, R. Pearson, D. Vukcevic, C. Langford, P. Deloukas, S. Hunt, E. Gray, S. Dronov, S. C. Potter, A. Tashakkori-Ghanbaria, S. Edkins, S. J. Bumpstead, J. M. Blackwell, E. Bramon, M. A. Brown, J. P. Casas, A. Corvin, A. Duncanson, J. A. Z. Jankowski, H. S. Markus, C. G. Mathew, C. N. A. Palmer, A. Rautanen, S. J. Sawcer, R. C. Trembath, A. C. Viswanathan, N. W. Wood, I. Barroso, L. Peltonen, P. S. Dale, S. A. Petrill, L. S. Schalkwyk, I. W. Craig, C. M. Lewis, T. S. Price, P. Donnelly, R. Plomin, and C. C. A. Spencer, "The correlation between reading and mathematics ability at age twelve has a substantial genetic component," Nature Communications, vol. 5, July 2014.

[170] A. Rautanen, M. Pirinen, T. C. Mills, K. A. Rockett, A. Strange, A. W. Ndungu, V. Naranbhai, J. J. Gilchrist, C. Bellenguez, C. Freeman, G. Band, S. J. Bumpstead, S. Edkins, E. Giannoulatou, E. Gray, S. Dronov, S. E. Hunt, C. Langford, R. D. Pearson, Z. Su, D. Vukcevic, A. W. Macharia, S. Uyoga, C. Ndila, N. Mturi, P. Njuguna, S. Mohammed, J. A. Berkley, I. Mwangi, S. Mwarumba, B. S. Kitsao, B. S. Lowe, S. C. Morpeth, I. Khandwalla, J. M. Blackwell, E. Bramon, M. A. Brown, J. P. Casas, A. Corvin, A. Duncanson, J. Jankowski, H. S. Markus, C. G. Mathew, C. N. Palmer, R. Plomin, S. J. Sawcer, R. C. Trembath, A. C. Viswanathan, N. W. Wood, P. Deloukas, L. Peltonen, T. N. Williams,

J. A. G. Scott, S. J. Chapman, P. Donnelly, A. V. Hill, and C. C. Spencer, "Polymorphism in a lincRNA Associates with a Doubled Risk of Pneumococcal Bacteremia in Kenyan Children," The American Journal of Human Genetics, vol. 98, pp. 1092–1100, June 2016.

[171] C. J. Willer, E. K. Speliotes, R. J. F. Loos, S. Li, C. M. Lindgren, I. M. Heid, S. I. Berndt, A. L. Elliott, A. U. Jackson, C. Lamina, G. Lettre, N. Lim, H. N. Lyon, S. A. McCarroll, K. Papadakis, L. Qi, J. C. Randall, R. M. Roccasecca, S. Sanna, P. Scheet, M. N. Weedon, E. Wheeler, J. H. Zhao, L. C. Jacobs, I. Prokopenko, N. Soranzo, T. Tanaka, N. J. Timpson, P. Almgren, A. Bennett, R. N. Bergman, S. A. Bingham, L. L. Bonnycastle, M. Brown, N. P. Burtt, P. Chines, L. Coin, F. S. Collins, J. M. Connell, C. Cooper, G. D. Smith, E. M. Dennison, P. Deodhar, P. Elliott, M. R. Erdos, K. Estrada, D. M. Evans, L. Gianniny, C. Gieger, C. J. Gillson, C. Guiducci, R. Hackett, D. Hadley, A. S. Hall, A. S. Havulinna, J. Hebebrand, A. Hofman, B. Isomaa, K. B. Jacobs, T. Johnson, P. Jousilahti, Z. Jovanovic, K.-T. Khaw, P. Kraft, M. Kuokkanen, J. Kuusisto, J. Laitinen, E. G. Lakatta, J. Luan, R. N. Luben, M. Mangino, W. L. McArdle, T. Meitinger, A. Mulas, P. B. Munroe, N. Narisu, A. R. Ness, K. Northstone, S. O'Rahilly, C. Purmann, M. G. Rees, M. Ridderstråle, S. M. Ring, F. Rivadeneira, A. Ruokonen, M. S. Sandhu, J. Saramies, L. J. Scott, A. Scuteri, K. Silander, M. A. Sims, K. Song, J. Stephens, S. Stevens, H. M. Stringham, Y. C. L. Tung, T. T. Valle, C. M. Van Duijn, K. S. Vimaleswaran, P. Vollenweider, G. Waeber, C. Wallace, R. M. Watanabe, D. M. Waterworth, N. Watkins, J. C. M. Witteman, E. Zeggini, G. Zhai, M. C. Zillikens, D. Altshuler, M. J. Caulfield, S. J. Chanock, I. S. Farooqi, L. Ferrucci, J. M. Guralnik, A. T. Hattersley, F. B. Hu, M.-R. Jarvelin, M. Laakso, V. Mooser, K. K. Ong, W. H. Ouwehand, V. Salomaa, N. J. Samani, T. D. Spector, T. Tuomi, J. Tuomilehto, M. Uda, A. G. Uitterlinden, N. J. Wareham, P. Deloukas, T. M. Frayling, L. C. Groop, R. B. Hayes, D. J. Hunter, K. L. Mohlke, L. Peltonen, D. Schlessinger, D. P. Strachan, H.-E. Wichmann, M. I. McCarthy, M. Boehnke, I. Barroso, G. R. Abecasis, and J. N. Hirschhorn, "Six new loci associated with body mass index highlight a neuronal influence on body weight regulation," Nature Genetics, vol. 41, pp. 25–34, Jan. 2009.

[172] Psychiatric GWAS Consortium Steering Committee, "A framework for interpreting genome-wide association studies of psychiatric disorders," Molecular Psychiatry, vol. 14, pp. 10–17, Jan. 2009.

[173] W. Rayner, "Strand Data," May 2017.

[174] G. Band, "Inthinnerator," Apr. 2017.

[175] S. Raychaudhuri, E. F. Remmers, A. T. Lee, R. Hackett, C. Guiducci, N. P. Burtt, L. Gianniny, B. D. Korman, L. Padyukov, F. A. S. Kurreeman, M. Chang, J. J. Catanese, B. Ding, S. Wong, A. H. M. van der Helm-van Mil, B. M. Neale, J. Coblyn, J. Cui, P. P. Tak, G. J. Wolbink, J. B. A. Crusius, I. E. van der Horst-Bruinsma, L. A. Criswell, C. I. Amos, M. F. Seldin, D. L. Kastner, K. G. Ardlie, L. Alfredsson, K. H. Costenbader, D. Altshuler, T. W. J. Huizinga, N. A. Shadick, M. E. Weinblatt, N. de Vries, J. Worthington, M. Seielstad, R. E. M. Toes, E. W. Karlson, A. B. Begovich, L. Klareskog, P. K. Gregersen, M. J. Daly, and R. M. Plenge, "Common variants at CD40 and other loci confer risk of rheumatoid arthritis," Nature Genetics, vol. 40, pp. 1216–1223, Oct. 2008.

[176] G. Orozco, S. Viatte, J. Bowes, P. Martin, A. G. Wilson, A. W. Morgan, S. Steer,

P. Wordsworth, L. J. Hocking, UK Rheumatoid Arthritis Genetics Consortium, Wellcome Trust Case Control Consortium, Biologics in Rheumatoid Arthritis Genetics and Genomics Study Syndicate Consortium, A. Barton, J. Worthington, and S. Eyre, "Novel Rheumatoid Arthritis Susceptibility Locus at 22q12 Identified in an Extended UK Genome-Wide Association Study: New RA Risk Locus at 22q12," Arthritis & Rheumatology, vol. 66, pp. 24–30, Jan. 2014.

[177] N. Yamamoto, J. Nakayama, K. Yamakawa-Kobayashi, H. Hamaguchi, R. Miyazaki, and T. Arinami, "Identification of 33 polymorphisms in the adipocyte-derived leucine aminopeptidase (ALAP) gene and possible association with hypertension," Human Mutation, vol. 19, pp. 251–257, Mar. 2002.

[178] M. P. Johnson, L. T. Roten, T. D. Dyer, C. E. East, S. Forsmo, J. Blangero, S. P. Brennecke, R. Austgulen, and E. K. Moses, "The ERAP2 gene is associated with preeclampsia in Australian and Norwegian populations," Human Genetics, vol. 126, pp. 655–666, Nov. 2009.

[179] A. H. Beecham, N. A. Patsopoulos, D. K. Xifara, M. F. Davis, A. Kemppinen, C. Cotsapas, T. S. Shah, C. Spencer, D. Booth, A. Goris, A. Oturai, J. Saarela, B. Fontaine, B. Hemmer, C. Martin, F. Zipp, S. D'Alfonso, F. Martinelli-Boneschi, B. Taylor, H. F. Harbo, I. Kockum, J. Hillert, T. Olsson, M. Ban, J. R. Oksenberg, R. Hintzen, L. F. Barcellos, C. Agliardi, L. Alfredsson, M. Alizadeh, C. Anderson, R. Andrews, H. B. Søndergaard, A. Baker, G. Band, S. E. Baranzini, N. Barizzone, J. Barrett, C. Bellenguez, L. Bergamaschi, L. Bernardinelli, A. Berthele, V. Biberacher, T. M. C. Binder, H. Blackburn, I. L. Bomfim, P. Brambilla, S. Broadley, B. Brochet, L. Brundin, D. Buck, H. Butzkueven, S. J. Caillier, W. Camu, W. Carpentier, P. Cavalla, E. G. Celius, I. Coman, G. Comi, L. Corrado, L. Cosemans, I. Cournu-Rebeix, B. A. C. Cree, D. Cusi, V. Damotte, G. Defer, S. R. Delgado, P. Deloukas, A. di Sapio, A. T. Dilthey, P. Donnelly, B. Dubois, M. Duddy, S. Edkins, I. Elovaara, F. Esposito, N. Evangelou, B. Fiddes, J. Field, A. Franke, C. Freeman, I. Y. Frohlich, D. Galimberti, C. Gieger, P.-A. Gourraud, C. Graetz, A. Graham, V. Grummel, C. Guaschino, A. Hadjixenofontos, H. Hakonarson, C. Halfpenny, G. Hall, P. Hall, A. Hamsten, J. Harley, T. Harrower, C. Hawkins, G. Hellenthal, C. Hillier, J. Hobart, M. Hoshi, S. E. Hunt, M. Jagodic, I. Jelčić, A. Jochim, B. Kendall, A. Kermode, T. Kilpatrick, K. Koivisto, I. Konidari, T. Korn, H. Kronsbein, C. Langford, M. Larsson, M. Lathrop, C. Lebrun-Frenay, J. Lechner-Scott, M. H. Lee, M. A. Leone, V. Leppä, G. Liberatore, B. A. Lie, C. M. Lill, M. Lindén, J. Link, F. Luessi, J. Lycke, F. Macciardi, S. Männistö, C. P. Manrique, R. Martin, V. Martinelli, D. Mason, G. Mazibrada, C. McCabe, I.-L. Mero, J. Mescheriakova, L. Moutsianas, K.-M. Myhr, G. Nagels, R. Nicholas, P. Nilsson, F. Piehl, M. Pirinen, S. E. Price, H. Quach, M. Reunanen, W. Robberecht, N. P. Robertson, M. Rodegher, D. Rog, M. Salvetti, N. C. Schnetz-Boutaud, F. Sellebjerg, R. C. Selter, C. Schaefer, S. Shaunak, L. Shen, S. Shields, V. Siffrin, M. Slee, P. S. Sorensen, M. Sorosina, M. Sospedra, A. Spurkland, A. Strange, E. Sundqvist, V. Thijs, J. Thorpe, A. Ticca, P. Tienari, C. van Duijn, E. M. Visser, S. Vucic, H. Westerlind, J. S. Wiley, A. Wilkins, J. F. Wilson, J. Winkelmann, J. Zajicek, E. Zindler, J. L. Haines, M. A. Pericak-Vance, A. J. Ivinson, G. Stewart, D. Hafler, S. L. Hauser, A. Compston, G. McVean, P. De Jager, S. J. Sawcer, and J. L. McCauley, "Analysis of immune-related loci identifies 48 new susceptibility variants for multiple sclerosis," Nature Genetics, vol. 45,

pp. 1353–1360, Sept. 2013.

[180] J. M. Blackwell, M. Fakiola, M. E. Ibrahim, S. E. Jamieson, S. B. Jeronimo, E. N. Miller, A. Mishra, H. S. Mohamed, C. S. Peacock, M. Raju, S. Sundar, and M. E. Wilson, "Genetics and visceral leishmaniasis: of mice and man," Parasite Immunology, vol. 31, pp. 254–266, May 2009.

[181] M. He, M. Xu, B. Zhang, J. Liang, P. Chen, J.-Y. Lee, T. A. Johnson, H. Li, X. Yang, J. Dai, L. Liang, L. Gui, Q. Qi, J. Huang, Y. Li, L. S. Adair, T. Aung, Q. Cai, C.-Y. Cheng, M.-C. Cho, Y. S. Cho, M. Chu, B. Cui, Y.-T. Gao, M. J. Go, D. Gu, W. Gu, H. Guo, Y. Hao, J. Hong, Z. Hu, Y. Hu, J. Huang, J.-Y. Hwang, M. K. Ikram, G. Jin, D.-H. Kang, C. C. Khor, B.-J. Kim, H. T. Kim, M. Kubo, J. Lee, J. Lee, N. R. Lee, R. Li, J. Li, J. Liu, J. Longe, W. Lu, X. Lu, X. Miao, Y. Okada, R. T.-H. Ong, G. Qiu, M. Seielstad, X. Sim, H. Song, F. Takeuchi, T. Tanaka, P. R. Taylor, L. Wang, W. Wang, Y. Wang, C. Wu, Y. Wu, Y.-B. Xiang, K. Yamamoto, H. Yang, M. Liao, M. Yokota, T. Young, X. Zhang, N. Kato, Q. K. Wang, W. Zheng, F. B. Hu, D. Lin, H. Shen, Y. Y. Teo, Z. Mo, T. Y. Wong, X. Lin, K. L. Mohlke, G. Ning, T. Tsunoda, B.-G. Han, X.-O. Shu, E. S. Tai, T. Wu, and L. Qi, "Meta-analysis of genome-wide association studies of adult height in East Asians identifies 17 novel loci," Human Molecular Genetics, vol. 24, pp. 1791–1800, Mar. 2015.

[182] M. R. Irvin, D. M. Rotroff, S. Aslibekyan, D. Zhi, B. Hidalgo, A. Motsinger-Reif, S. Marvel, V. Srinivasasainagendra, S. A. Claas, J. B. Buse, R. J. Straka, J. M. Ordovas, I. B. Borecki, X. Guo, I. Y. D. Chen, J. I. Rotter, M. J. Wagner, and D. K. Arnett, "A genome-wide study of lipid response to fenofibrate in Caucasians: a combined analysis of the GOLDN and ACCORD studies," Pharmacogenetics and Genomics, vol. 26, no. 7, pp. 324–333, 2016.

[183] L. E. Mitchell, A. Agopian, A. Bhalla, J. T. Glessner, C. E. Kim, M. D. Swartz, H. Hakonarson, and E. Goldmuntz, "Genome-wide association study of maternal and inherited effects on left-sided cardiac malformations," Human Molecular Genetics, vol. 24, pp. 265–273, Jan. 2015.

[184] H. Stefansson, A. Helgason, G. Thorleifsson, V. Steinthorsdottir, G. Masson, J. Barnard, A. Baker, A. Jonasdottir, A. Ingason, V. G. Gudnadottir, N. Desnica, A. Hicks, A. Gylfason, D. F. Gudbjartsson, G. M. Jonsdottir, J. Sainz, K. Agnarsson, B. Birgisdottir, S. Ghosh, A. Olafsdottir, J.-B. Cazier, K. Kristjansson, M. L. Frigge, T. E. Thorgeirsson, J. R. Gulcher, A. Kong, and K. Stefansson, "A common inversion under selection in Europeans," Nature Genetics, vol. 37, pp. 129–137, Feb. 2005.

[185] M. C. Zody, Z. Jiang, H.-C. Fung, F. Antonacci, L. W. Hillier, M. F. Cardone, T. A. Graves, J. M. Kidd, Z. Cheng, A. Abouelleil, L. Chen, J. Wallis, J. Glasscock, R. K. Wilson, A. D. Reily, J. Duckworth, M. Ventura, J. Hardy, W. C. Warren, and E. E. Eichler, "Evolutionary toggling of the MAPT 17q21.31 inversion region," Nature Genetics, vol. 40, pp. 1076–1083, Sept. 2008.

[186] L. Skipper, K. Wilkes, M. Toft, M. Baker, S. Lincoln, M. Hulihan, O. A. Ross, M. Hutton, J. Aasly, and M. Farrer, "Linkage Disequilibrium and Association of MAPT H1 in Parkinson Disease," The American Journal of Human Genetics, vol. 75, pp. 669–677, Oct. 2004.

[187] J. E. Tobin, J. C. Latourelle, M. F. Lew, C. Klein, O. Suchowersky, H. A. Shill, L. I. Golbe, M. H. Mark, J. H. Growdon, G. F. Wooten, B. A. Racette, J. S. Perlmutter, R. Watts, M. Guttman, K. B. Baker, S. Goldwurm, G. Pezzoli,

C. Singer, M. H. Saint-Hilaire, A. E. Hendricks, S. Williamson, M. W. Nagle, J. B. Wilk, T. Massood, J. M. Laramie, A. L. DeStefano, I. Litvan, G. Nicholson, A. Corbett, S. Isaacson, D. J. Burn, P. F. Chinnery, P. P. Pramstaller, S. Sherman, J. Al-hinti, E. Drasby, M. Nance, A. T. Moller, K. Ostergaard, R. Roxburgh, B. Snow, J. T. Slevin, F. Cambi, J. F. Gusella, and R. H. Myers, "Haplotypes and gene expression implicate the MAPT region for Parkinson disease: The GenePD Study," Neurology, vol. 71, pp. 28–34, July 2008.

[188] K. M. Steinberg, F. Antonacci, P. H. Sudmant, J. M. Kidd, C. D. Campbell, L. Vives, M. Malig, L. Scheinfeldt, W. Beggs, M. Ibrahim, G. Lema, T. B. Nyambo, S. A. Omar, J.-M. Bodo, A. Froment, M. P. Donnelly, K. K. Kidd, S. A. Tishkoff, and E. E. Eichler, "Structural diversity and African origin of the 17q21.31 inversion polymorphism," Nature Genetics, vol. 44, pp. 872–880, July 2012.

[189] J. Simón-Sánchez, C. Schulte, J. M. Bras, M. Sharma, J. R. Gibbs, D. Berg, C. Paisan-Ruiz, P. Lichtner, S. W. Scholz, D. G. Hernandez, R. Krüger, M. Federoff, C. Klein, A. Goate, J. Perlmutter, M. Bonin, M. A. Nalls, T. Illig, C. Gieger, H. Houlden, M. Steffens, M. S. Okun, B. A. Racette, M. R. Cookson, K. D. Foote, H. H. Fernandez, B. J. Traynor, S. Schreiber, S. Arepalli, R. Zonozi, K. Gwinn, M. van der Brug, G. Lopez, S. J. Chanock, A. Schatzkin, Y. Park, A. Hollenbeck, J. Gao, X. Huang, N. W. Wood, D. Lorenz, G. Deuschl, H. Chen, O. Riess, J. A. Hardy, A. B. Singleton, and T. Gasser, "Genome-wide association study reveals genetic risk underlying Parkinson's disease," Nature Genetics, vol. 41, pp. 1308–1312, Dec. 2009.

[190] N. Kouri, O. A. Ross, B. Dombroski, C. S. Younkin, D. J. Serie, A. Soto-Ortolaza, M. Baker, N. C. A. Finch, H. Yoon, J. Kim, S. Fujioka, C. A. McLean, B. Ghetti, S. Spina, L. B. Cantwell, M. R. Farlow, J. Grafman, E. D. Huey, M. Ryung Han, S. Beecher, E. T. Geller, H. A. Kretzschmar, S. Roeber, M. Gearing, J. L. Juncos, J. P. G. Vonsattel, V. M. Van Deerlin, M. Grossman, H. I. Hurtig, R. G. Gross, S. E. Arnold, J. Q. Trojanowski, V. M. Lee, G. K. Wenning, C. L. White, G. U. Höglinger, U. Müller, B. Devlin, L. I. Golbe, J. Crook, J. E. Parisi, B. F. Boeve, K. A. Josephs, Z. K. Wszolek, R. J. Uitti, N. R. Graff-Radford, I. Litvan, S. G. Younkin, L.-S. Wang, N. Ertekin-Taner, R. Rademakers, H. Hakonarsen, G. D. Schellenberg, and D. W. Dickson, "Genome-wide association study of corticobasal degeneration identifies risk variants shared with progressive supranuclear palsy," Nature Communications, vol. 6, p. 7247, June 2015.

[191] A. U. Bayer, O. N. Keller, F. Ferrari, and K.-P. Maag, "Association of glaucoma with neurodegenerative diseases with apoptotic cell death: Alzheimer's disease and Parkinson's disease," American Journal of Ophthalmology, vol. 133, pp. 135–137, Jan. 2002.

[192] S.-W. Lai, C.-L. Lin, and K.-F. Liao, "Glaucoma correlates with increased risk of Parkinson's disease in the elderly: a national-based cohort study in Taiwan," Current Medical Research and Opinion, pp. 1–18, Apr. 2017.

[193] I.-C. Lin, Y.-H. Wang, T.-J. Wang, I.-J. Wang, Y.-D. Shen, N.-F. Chi, and L.-N. Chien, "Glaucoma, Alzheimer's Disease, and Parkinson's Disease: An 8-Year Population-Based Follow-Up Study," PLoS ONE, vol. 9, p. e108938, Oct. 2014.

[194] T. E. Fingerlin, E. Murphy, W. Zhang, A. L. Peljto, K. K. Brown, M. P. Steele, J. E. Loyd, G. P. Cosgrove, D. Lynch, S. Groshong, H. R. Collard, P. J. Wolters,

W. Z. Bradford, K. Kossen, S. D. Seiwert, R. M. du Bois, C. K. Garcia, M. S. Devine, G. Gudmundsson, H. J. Isaksson, N. Kaminski, Y. Zhang, K. F. Gibson, L. H. Lancaster, J. D. Cogan, W. R. Mason, T. M. Maher, P. L. Molyneaux, A. U. Wells, M. F. Moffatt, M. Selman, A. Pardo, D. S. Kim, J. D. Crapo, B. J. Make, E. A. Regan, D. S. Walek, J. J. Daniel, Y. Kamatani, D. Zelenika, K. Smith, D. McKean, B. S. Pedersen, J. Talbert, R. N. Kidd, C. R. Markin, K. B. Beckman, M. Lathrop, M. I. Schwarz, and D. A. Schwartz, "Genome-wide association study identifies multiple susceptibility loci for pulmonary fibrosis," Nature Genetics, vol. 45, pp. 613–620, Apr. 2013.

[195] A. J. Myers, A. M. Pittman, A. S. Zhao, K. Rohrer, M. Kaleem, L. Marlowe, A. Lees, D. Leung, I. G. McKeith, R. H. Perry, C. M. Morris, J. Q. Trojanowski, C. Clark, J. Karlawish, S. Arnold, M. S. Forman, V. Van Deerlin, R. de Silva, and J. Hardy, "The MAPT H1c risk haplotype is associated with increased expression of tau and especially of 4 repeat containing transcripts," Neurobiology of Disease, vol. 25, pp. 561–570, Mar. 2007.

[196] W. J. Kent, C. W. Sugnet, T. S. Furey, K. M. Roskin, T. H. Pringle, A. M. Zahler, and a. D. Haussler, "The Human Genome Browser at UCSC," Genome Research, vol. 12, pp. 996–1006, May 2002.

[197] M. Germain, D. I. Chasman, H. de Haan, W. Tang, S. Lindström, L.-C. Weng, M. de Andrade, M. C. H. de Visser, K. L. Wiggins, P. Suchon, N. Saut, D. M. Smadja, G. Le Gal, A. van Hylckama Vlieg, A. Di Narzo, K. Hao, C. P. Nelson, A. Rocanin-Arjo, L. Folkersen, R. Monajemi, L. M. Rose, J. A. Brody, E. Slagboom, D. Aïssi, F. Gagnon, J.-F. Deleuze, P. Deloukas, C. Tzourio, J.-F. Dartigues, C. Berr, K. D. Taylor, M. Civelek, P. Eriksson, Cardiogenics Consortium, B. M. Psaty, J. Houwing-Duitermaat, A. H. Goodall, F. Cambien, P. Kraft, P. Amouyel, N. J. Samani, S. Basu, P. M. Ridker, F. R. Rosendaal, C. Kabrhel, A. R. Folsom, J. Heit, P. H. Reitsma, D.-A. Trégouët, N. L. Smith, and P.-E. Morange, "Meta-analysis of 65,734 individuals identifies TSPAN15 and SLC44a2 as two susceptibility loci for venous thromboembolism," American Journal of Human Genetics, vol. 96, pp. 532–542, Apr. 2015.

[198] S. Kugathasan, R. N. Baldassano, J. P. Bradfield, P. M. A. Sleiman, M. Imielinski, S. L. Guthery, S. Cucchiara, C. E. Kim, E. C. Frackelton, K. Annaiah, J. T. Glessner, E. Santa, T. Willson, A. W. Eckert, E. Bonkowski, J. L. Shaner, R. M. Smith, F. G. Otieno, N. Peterson, D. J. Abrams, R. M. Chiavacci, R. Grundmeier, P. Mamula, G. Tomer, D. A. Piccoli, D. S. Monos, V. Annese, L. A. Denson, S. F. A. Grant, and H. Hakonarson, "Loci on 20q13 and 21q22 are associated with pediatric-onset inflammatory bowel disease," Nature Genetics, vol. 40, pp. 1211–1215, Oct. 2008.

[199] Australo-Anglo-American Spondyloarthritis Consortium (TASC), J. D. Reveille, A.-M. Sims, P. Danoy, D. M. Evans, P. Leo, J. J. Pointon, R. Jin, X. Zhou, L. A. Bradbury, L. H. Appleton, J. C. Davis, L. Diekman, T. Doan, A. Dowling, R. Duan, E. L. Duncan, C. Farrar, J. Hadler, D. Harvey, T. Karaderi, R. Mogg, E. Pomeroy, K. Pryce, J. Taylor, L. Savage, P. Deloukas, V. Kumanduri, L. Peltonen, S. M. Ring, P. Whittaker, E. Glazov, G. P. Thomas, W. P. Maksymowych, R. D. Inman, M. M. Ward, M. A. Stone, M. H. Weisman, B. P. Wordsworth, and M. A. Brown, "Genome-wide association study of ankylosing spondylitis identifies non-MHC susceptibility loci," Nature Genetics, vol. 42, pp. 123–127, Feb. 2010.

[200] J. Spada, M. Scholz, H. Kirsten, T. Hensch, K. Horn, P. Jawinski, C. Ulke, R. Burkhardt, K. Wirkner, M. Loeffler, U. Hegerl, and C. Sander, "Genome-wide association analysis of actigraphic sleep phenotypes in the LIFE Adult Study," Journal of Sleep Research, vol. 25, pp. 690–701, Dec. 2016.

[201] J. R. Korenberg, H. Kawashima, S. M. Pulst, T. Ikeuchi, N. Ogasawara, K. Yamamoto, S. A. Schonberg, R. West, L. Allen, and E. Magenis, "Molecular definition of a region of chromosome 21 that causes features of the Down syndrome phenotype," American Journal of Human Genetics, vol. 47, pp. 236–246, Aug. 1990.

[202] P. Capkova, N. Misovicova, and D. Vrbicka, "Partial trisomy and tetrasomy of chromosome 21 without down syndrome phenotype and short overview of genotype-phenotype correlation. a case report," Biomedical Papers, Oct. 2013.

[203] L. C. Tsoi, S. L. Spain, J. Knight, E. Ellinghaus, P. E. Stuart, F. Capon, J. Ding, Y. Li, T. Tejasvi, J. E. Gudjonsson, H. M. Kang, M. H. Allen, R. McManus, G. Novelli, L. Samuelsson, J. Schalkwijk, M. Ståhle, A. D. Burden, C. H. Smith, M. J. Cork, X. Estivill, A. M. Bowcock, G. G. Krueger, W. Weger, J. Worthington, R. Tazi-Ahnini, F. O. Nestle, A. Hayday, P. Hoffmann, J. Winkelmann, C. Wijmenga, C. Langford, S. Edkins, R. Andrews, H. Blackburn, A. Strange, G. Band, R. D. Pearson, D. Vukcevic, C. C. A. Spencer, P. Deloukas, U. Mrowietz, S. Schreiber, S. Weidinger, S. Koks, K. Kingo, T. Esko, A. Metspalu, H. W. Lim, J. J. Voorhees, M. Weichenthal, H. E. Wichmann, V. Chandran, C. F. Rosen, P. Rahman, D. D. Gladman, C. E. M. Griffiths, A. Reis, J. Kere, K. C. Duffin, C. Helms, D. Goldgar, Y. Li, J. Paschall, M. J. Malloy, C. R. Pullinger, J. P. Kane, J. Gardner, A. Perlmutter, A. Miner, B. J. Feng, R. Hiremagalore, R. W. Ike, E. Christophers, T. Henseler, A. Ruether, S. J. Schrodi, S. Prahalad, S. L. Guthery, J. Fischer, W. Liao, P. Kwok, A. Menter, G. M. Lathrop, C. Wise, A. B. Begovich, A. Onoufriadis, M. E. Weale, A. Hofer, W. Salmhofer, P. Wolf, K. Kainu, U. Saarialho-Kere, S. Suomela, P. Badorf, U. Hüffmeier, W. Kurrat, W. Küster, J. Lascorz, R. Mössner, F. Schürmeier-Horst, M. Ständer, H. Traupe, J. G. M. Bergboer, M. d. Heijer, P. C. van de Kerkhof, P. L. J. M. Zeeuwen, L. Barnes, L. E. Campbell, C. Cusack, C. Coleman, J. Conroy, S. Ennis, O. Fitzgerald, P. Gallagher, A. D. Irvine, B. Kirby, T. Markham, W. H. I. McLean, J. McPartlin, S. F. Rogers, A. W. Ryan, A. Zawirska, E. Giardina, T. Lepre, C. Perricone, G. Martín-Ezquerra, R. M. Pujol, E. Riveira-Munoz, A. Inerot, Å. T. Naluai, L. Mallbris, K. Wolk, J. Leman, A. Barton, R. B. Warren, H. S. Young, I. Ricano-Ponce, G. Trynka, F. J. Pellett, A. Henschel, M. Aurand, B. Bebo, C. Gieger, T. Illig, S. Moebus, K.-H. Jöckel, R. Erbel, P. Donnelly, L. Peltonen, J. M. Blackwell, E. Bramon, M. A. Brown, J. P. Casas, A. Corvin, N. Craddock, A. Duncanson, J. Jankowski, H. S. Markus, C. G. Mathew, M. I. McCarthy, C. N. A. Palmer, R. Plomin, A. Rautanen, S. J. Sawcer, N. Samani, A. C. Viswanathan, N. W. Wood, C. Bellenguez, C. Freeman, G. Hellenthal, E. Giannoulatou, M. Pirinen, Z. Su, S. E. Hunt, R. Gwilliam, S. J. Bumpstead, S. Dronov, M. Gillman, E. Gray, N. Hammond, A. Jayakumar, O. T. McCann, J. Liddle, M. L. Perez, S. C. Potter, R. Ravindrarajah, M. Ricketts, M. Waller, P. Weston, S. Widaa, P. Whittaker, R. P. Nair, A. Franke, J. N. W. N. Barker, G. R. Abecasis, J. T. Elder, and R. C. Trembath, "Identification of 15 new psoriasis susceptibility loci highlights the role of innate immunity," Nature Genetics, vol. 44, pp. 1341–1348, Nov. 2012.

[204] E. Ellinghaus, D. Ellinghaus, P. E. Stuart, R. P. Nair, S. Debrus, J. V. Raelson, M. Belouchi, H. Fournier, C. Reinhard, J. Ding, Y. Li, T. Tejasvi, J. Gudjonsson, S. W. Stoll, J. J. Voorhees, S. Lambert, S. Weidinger, B. Eberlein, M. Kunz, P. Rahman, D. D. Gladman, C. Gieger, H. E. Wichmann, T. H. Karlsen, G. Mayr, M. Albrecht, D. Kabelitz, U. Mrowietz, G. R. Abecasis, J. T. Elder, S. Schreiber, M. Weichenthal, and A. Franke, "Genome-wide association study identifies a psoriasis susceptibility locus at TRAF3ip2," Nature Genetics, vol. 42, pp. 991–995, Nov. 2010.

[205] T. H. Pers, P. Timshel, and J. N. Hirschhorn, "SNPsnap: a Web-based tool for identification and annotation of matched SNPs," Bioinformatics, vol. 31, pp. 418–420, Feb. 2015.

[206] A. Fabregat, K. Sidiropoulos, P. Garapati, M. Gillespie, K. Hausmann, R. Haw, B. Jassal, S. Jupe, F. Korninger, S. McKay, L. Matthews, B. May, M. Milacic, K. Rothfels, V. Shamovsky, M. Webber, J. Weiser, M. Williams, G. Wu, L. Stein, H. Hermjakob, and P. D'Eustachio, "The Reactome pathway Knowledgebase," Nucleic Acids Research, vol. 44, pp. D481–D487, Jan. 2016.

[207] W.-M. Chu, "Tumor necrosis factor," Cancer Letters, vol. 328, pp. 222–225, Jan. 2013.

[208] I. Dunham, E. Kulesha, V. Iotchkova, S. Morganella, and E. Birney, "FORGE: A tool to discover cell specific enrichments of GWAS associated SNPs in regulatory regions," F1000Research, Jan. 2015.

[209] B. E. Bernstein, J. A. Stamatoyannopoulos, J. F. Costello, B. Ren, A. Milosavljevic, A. Meissner, M. Kellis, M. A. Marra, A. L. Beaudet, J. R. Ecker, P. J. Farnham, M. Hirst, E. S. Lander, T. S. Mikkelsen, and J. A. Thomson, "The NIH Roadmap Epigenomics Mapping Consortium," Nature Biotechnology, vol. 28, pp. 1045–1048, Oct. 2010.

[210] J. K. Chan, C. S. Ng, and P. K. Hui, "A simple guide to the terminology and application of leucocyte monoclonal antibodies," Histopathology, vol. 12, pp. 461–480, May 1988.

[211] P. Engel, L. Boumsell, R. Balderas, A. Bensussan, V. Gattei, V. Horejsi, B.-Q. Jin, F. Malavasi, F. Mortari, R. Schwartz-Albiez, H. Stockinger, M. C. van Zelm, H. Zola, and G. Clark, "CD Nomenclature 2015: Human Leukocyte Differentiation Antigen Workshops as a Driving Force in Immunology," The Journal of Immunology, vol. 195, pp. 4555–4563, Nov. 2015.

[212] The Immunology Link, "CD Antigens," Aug. 2017.

[213] C. A. Rietveld, S. E. Medland, J. Derringer, J. Yang, T. Esko, N. W. Martin, H.-J. Westra, K. Shakhbazov, A. Abdellaoui, A. Agrawal, E. Albrecht, B. Z. Alizadeh, N. Amin, J. Barnard, S. E. Baumeister, K. S. Benke, L. F. Bielak, J. A. Boatman, P. A. Boyle, G. Davies, C. de Leeuw, N. Eklund, D. S. Evans, R. Ferhmann, K. Fischer, C. Gieger, H. K. Gjessing, S. Hagg, J. R. Harris, C. Hayward, C. Holzapfel, C. A. Ibrahim-Verbaas, E. Ingelsson, B. Jacobsson, P. K. Joshi, A. Jugessur, M. Kaakinen, S. Kanoni, J. Karjalainen, I. Kolcic, K. Kristiansson, Z. Kutalik, J. Lahti, S. H. Lee, P. Lin, P. A. Lind, Y. Liu, K. Lohman, M. Loitfelder, G. McMahon, P. M. Vidal, O. Meirelles, L. Milani, R. Myhre, M.-L. Nuotio, C. J. Oldmeadow, K. E. Petrovic, W. J. Peyrot, O. Polasek, L. Quaye, E. Reinmaa, J. P. Rice, T. S. Rizzi, H. Schmidt, R. Schmidt, A. V. Smith, J. A. Smith, T. Tanaka, A. Terracciano, M. J. H. M. van der Loos, V. Vitart, H. Volzke, J. Wellmann, L. Yu, W. Zhao, J. Allik, J. R. Attia,

S. Bandinelli, F. Bastardot, J. Beauchamp, D. A. Bennett, K. Berger, L. J. Bierut, D. I. Boomsma, U. Bultmann, H. Campbell, C. F. Chabris, L. Cherkas, M. K. Chung, F. Cucca, M. de Andrade, P. L. De Jager, J.-E. De Neve, I. J. Deary, G. V. Dedoussis, P. Deloukas, M. Dimitriou, G. Eiriksdottir, M. F. Elderson, J. G. Eriksson, D. M. Evans, J. D. Faul, L. Ferrucci, M. E. Garcia, H. Gronberg, V. Guthnason, P. Hall, J. M. Harris, T. B. Harris, N. D. Hastie, A. C. Heath, D. G. Hernandez, W. Hoffmann, A. Hofman, R. Holle, E. G. Holliday, J.-J. Hottenga, W. G. Iacono, T. Illig, M.-R. Jarvelin, M. Kahonen, J. Kaprio, R. M. Kirkpatrick, M. Kowgier, A. Latvala, L. J. Launer, D. A. Lawlor, T. Lehtimaki, J. Li, P. Lichtenstein, P. Lichtner, D. C. Liewald, P. A. Madden, P. K. E. Magnusson, T. E. Makinen, M. Masala, M. McGue, A. Metspalu, A. Mielck, M. B. Miller, G. W. Montgomery, S. Mukherjee, D. R. Nyholt, B. A. Oostra, L. J. Palmer, A. Palotie, B. W. J. H. Penninx, M. Perola, P. A. Peyser, M. Preisig, K. Raikkonen, O. T. Raitakari, A. Realo, S. M. Ring, S. Ripatti, F. Rivadeneira, I. Rudan, A. Rustichini, V. Salomaa, A.-P. Sarin, D. Schlessinger, R. J. Scott, H. Snieder, B. St Pourcain, J. M. Starr, J. H. Sul, I. Surakka, R. Svento, A. Teumer, The LifeLines Cohort Study, H. Tiemeier, F. J. A. van Rooij, D. R. Van Wagoner, E. Vartiainen, J. Viikari, P. Vollenweider, J. M. Vonk, G. Waeber, D. R. Weir, H.-E. Wichmann, E. Widen, G. Willemsen, J. F. Wilson, A. F. Wright, D. Conley, G. Davey-Smith, L. Franke, P. J. F. Groenen, A. Hofman, M. Johannesson, S. L. R. Kardia, R. F. Krueger, D. Laibson, N. G. Martin, M. N. Meyer, D. Posthuma, A. R. Thurik, N. J. Timpson, A. G. Uitterlinden, C. M. van Duijn, P. M. Visscher, D. J. Benjamin, D. Cesarini, and P. D. Koellinger, "GWAS of 126,559 Individuals Identifies Genetic Variants Associated with Educational Attainment," Science, vol. 340, pp. 1467–1471, June 2013.

[214] A. Okbay, J. P. Beauchamp, M. A. Fontana, J. J. Lee, T. H. Pers, C. A. Rietveld, P. Turley, G.-B. Chen, V. Emilsson, S. F. W. Meddens, S. Oskarsson, J. K. Pickrell, K. Thom, P. Timshel, R. de Vlaming, A. Abdellaoui, T. S. Ahluwalia, J. Bacelis, C. Baumbach, G. Bjornsdottir, J. H. Brandsma, M. Pina Concas, J. Derringer, N. A. Furlotte, T. E. Galesloot, G. Girotto, R. Gupta, L. M. Hall, S. E. Harris, E. Hofer, M. Horikoshi, J. E. Huffman, K. Kaasik, I. P. Kalafati, R. Karlsson, A. Kong, J. Lahti, S. J. v. d. Lee, C. deLeeuw, P. A. Lind, K.-O. Lindgren, T. Liu, M. Mangino, J. Marten, E. Mihailov, M. B. Miller, P. J. van der Most, C. Oldmeadow, A. Payton, N. Pervjakova, W. J. Peyrot, Y. Qian, O. Raitakari, R. Rueedi, E. Salvi, B. Schmidt, K. E. Schraut, J. Shi, A. V. Smith, R. A. Poot, B. St Pourcain, A. Teumer, G. Thorleifsson, N. Verweij, D. Vuckovic, J. Wellmann, H.-J. Westra, J. Yang, W. Zhao, Z. Zhu, B. Z. Alizadeh, N. Amin, A. Bakshi, S. E. Baumeister, G. Biino, K. Bønnelykke, P. A. Boyle, H. Campbell, F. P. Cappuccio, G. Davies, J.-E. De Neve, P. Deloukas, I. Demuth, J. Ding, P. Eibich, L. Eisele, N. Eklund, D. M. Evans, J. D. Faul, M. F. Feitosa, A. J. Forstner, I. Gandin, B. Gunnarsson, B. V. Halldórsson, T. B. Harris, A. C. Heath, L. J. Hocking, E. G. Holliday, G. Homuth, M. A. Horan, J.-J. Hottenga, P. L. de Jager, P. K. Joshi, A. Jugessur, M. A. Kaakinen, M. Kähönen, S. Kanoni, L. Keltigangas-Järvinen, L. A. L. M. Kiemeney, I. Kolcic, S. Koskinen, A. T. Kraja, M. Kroh, Z. Kutalik, A. Latvala, L. J. Launer, M. P. Lebreton, D. F. Levinson, P. Lichtenstein, P. Lichtner, D. C. M. Liewald, L. Cohort Study, A. Loukola, P. A. Madden, R. Mägi,

T. Mäki-Opas, R. E. Marioni, P. Marques-Vidal, G. A. Meddens, G. McMahon, C. Meisinger, T. Meitinger, Y. Milaneschi, L. Milani, G. W. Montgomery, R. Myhre, C. P. Nelson, D. R. Nyholt, W. E. R. Ollier, A. Palotie, L. Paternoster, N. L. Pedersen, K. E. Petrovic, D. J. Porteous, K. Räikkönen, S. M. Ring, A. Robino, O. Rostapshova, I. Rudan, A. Rustichini, V. Salomaa, A. R. Sanders, A.-P. Sarin, H. Schmidt, R. J. Scott, B. H. Smith, J. A. Smith, J. A. Staessen, E. Steinhagen-Thiessen, K. Strauch, A. Terracciano, M. D. Tobin, S. Ulivi, S. Vaccargiu, L. Quaye, F. J. A. van Rooij, C. Venturini, A. A. E. Vinkhuyzen, U. Völker, H. Völzke, J. M. Vonk, D. Vozzi, J. Waage, E. B. Ware, G. Willemsen, J. R. Attia, D. A. Bennett, K. Berger, L. Bertram, H. Bisgaard, D. I. Boomsma, I. B. Borecki, U. Bültmann, C. F. Chabris, F. Cucca, D. Cusi, I. J. Deary, G. V. Dedoussis, C. M. van Duijn, J. G. Eriksson, B. Franke, L. Franke, P. Gasparini, P. V. Gejman, C. Gieger, H.-J. Grabe, J. Gratten, P. J. F. Groenen, V. Gudnason, P. van der Harst, C. Hayward, D. A. Hinds, W. Hoffmann, E. Hyppönen, W. G. Iacono, B. Jacobsson, M.-R. Järvelin, K.-H. Jöckel, J. Kaprio, S. L. R. Kardia, T. Lehtimäki, S. F. Lehrer, P. K. E. Magnusson, N. G. Martin, M. McGue, A. Metspalu, N. Pendleton, B. W. J. H. Penninx, M. Perola, N. Pirastu, M. Pirastu, O. Polasek, D. Posthuma, C. Power, M. A. Province, N. J. Samani, D. Schlessinger, R. Schmidt, T. I. A. Sørensen, T. D. Spector, K. Stefansson, U. Thorsteinsdottir, A. R. Thurik, N. J. Timpson, H. Tiemeier, J. Y. Tung, A. G. Uitterlinden, V. Vitart, P. Vollenweider, D. R. Weir, J. F. Wilson, A. F. Wright, D. C. Conley, R. F. Krueger, G. Davey Smith, A. Hofman, D. I. Laibson, S. E. Medland, M. N. Meyer, J. Yang, M. Johannesson, P. M. Visscher, T. Esko, P. D. Koellinger, D. Cesarini, and D. J. Benjamin, "Genome-wide association study identifies 74 loci associated with educational attainment," Nature, vol. 533, pp. 539–542, May 2016.

[215] S. Sniekers, S. Stringer, K. Watanabe, P. R. Jansen, J. R. I. Coleman, E. Krapohl, E. Taskesen, A. R. Hammerschlag, A. Okbay, D. Zabaneh, N. Amin, G. Breen, D. Cesarini, C. F. Chabris, W. G. Iacono, M. A. Ikram, M. Johannesson, P. Koellinger, J. J. Lee, P. K. E. Magnusson, M. McGue, M. B. Miller, W. E. R. Ollier, A. Payton, N. Pendleton, R. Plomin, C. A. Rietveld, H. Tiemeier, C. M. van Duijn, and D. Posthuma, "Genome-wide association meta-analysis of 78,308 individuals identifies new loci and genes influencing human intelligence," Nature Genetics, vol. 49, pp. 1107–1112, May 2017.

[216] Immunobase Team, Juvenile Diabetes Research Foundation, Wellcome Trust Diabetes and Inflammation Laboratory, J. Todd, O. Burren, E. Schofield, T. Carver, and P. Achuthan, "Immunobase," Aug. 2017.

[217] G. Trynka, K. A. Hunt, N. A. Bockett, J. Romanos, V. Mistry, A. Szperl, S. F. Bakker, M. T. Bardella, L. Bhaw-Rosun, G. Castillejo, E. G. de la Concha, R. C. de Almeida, K.-R. M. Dias, C. C. van Diemen, P. C. A. Dubois, R. H. Duerr, S. Edkins, L. Franke, K. Fransen, J. Gutierrez, G. A. R. Heap, B. Hrdlickova, S. Hunt, L. P. Izurieta, V. Izzo, L. A. B. Joosten, C. Langford, M. C. Mazzilli, C. A. Mein, V. Midah, M. Mitrovic, B. Mora, M. Morelli, S. Nutland, C. Núñez, S. Onengut-Gumuscu, K. Pearce, M. Platteel, I. Polanco, S. Potter, C. Ribes-Koninckx, I. Ricaño-Ponce, S. S. Rich, A. Rybak, J. L. Santiago, S. Senapati, A. Sood, H. Szajewska, R. Troncone, J. Varadé, C. Wallace, V. M. Wolters, A. Zhernakova, B. K. Thelma, B. Cukrowska, E. Urcelay, J. R. Bilbao, M. L. Mearin, D. Barisani, J. C. Barrett, V. Plagnol, P. Deloukas, C. Wijmenga, and

D. A. van Heel, "Dense genotyping identifies and localizes multiple common and rare variant association signals in celiac disease," Nature Genetics, vol. 43, pp. 1193–1201, Nov. 2011.

[218] A. Hinks, J. Cobb, M. C. Marion, S. Prahalad, M. Sudman, J. Bowes, P. Martin, M. E. Comeau, S. Sajuthi, R. Andrews, M. Brown, W.-M. Chen, P. Concannon, P. Deloukas, S. Edkins, S. Eyre, P. M. Gaffney, S. L. Guthery, J. M. Guthridge, S. E. Hunt, J. A. James, M. Keddache, K. L. Moser, P. A. Nigrovic, S. Onengut-Gumuscu, M. L. Onslow, C. D. Rosé, S. S. Rich, K. J. A. Steel, E. K. Wakeland, C. A. Wallace, L. R. Wedderburn, P. Woo, J. F. Bohnsack, J. P. Haas, D. N. Glass, C. D. Langefeld, W. Thomson, and S. D. Thompson, "Dense genotyping of immune-related disease regions identifies 14 new susceptibility loci for juvenile idiopathic arthritis," Nature Genetics, vol. 45, pp. 664–669, Apr. 2013.

[219] S. Eyre, J. Bowes, D. Diogo, A. Lee, A. Barton, P. Martin, A. Zhernakova, E. Stahl, S. Viatte, K. McAllister, C. I. Amos, L. Padyukov, R. E. M. Toes, T. W. J. Huizinga, C. Wijmenga, G. Trynka, L. Franke, H.-J. Westra, L. Alfredsson, X. Hu, C. Sandor, P. I. W. de Bakker, S. Davila, C. C. Khor, K. K. Heng, R. Andrews, S. Edkins, S. E. Hunt, C. Langford, D. Symmons, P. Concannon, S. Onengut-Gumuscu, S. S. Rich, P. Deloukas, M. A. Gonzalez-Gay, L. Rodriguez-Rodriguez, L. Ärlsetig, J. Martin, S. Rantapää-Dahlqvist, R. M. Plenge, S. Raychaudhuri, L. Klareskog, P. K. Gregersen, and J. Worthington, "High-density genetic mapping identifies new susceptibility loci for rheumatoid arthritis," Nature Genetics, vol. 44, pp. 1336–1340, Nov. 2012.

[220] S. Onengut-Gumuscu, W.-M. Chen, O. Burren, N. J. Cooper, A. R. Quinlan, J. C. Mychaleckyj, E. Farber, J. K. Bonnie, M. Szpak, E. Schofield, P. Achuthan, H. Guo, M. D. Fortune, H. Stevens, N. M. Walker, L. D. Ward, A. Kundaje, M. Kellis, M. J. Daly, J. C. Barrett, J. D. Cooper, P. Deloukas, J. A. Todd, C. Wallace, P. Concannon, and S. S. Rich, "Fine mapping of type 1 diabetes susceptibility loci and evidence for colocalization of causal variants with lymphoid gene enhancers," Nature Genetics, vol. 47, pp. 381–386, Mar. 2015.

[221] J. Z. Liu, M. A. Almarri, D. J. Gaffney, G. F. Mells, L. Jostins, H. J. Cordell, S. J. Ducker, D. B. Day, M. A. Heneghan, J. M. Neuberger, P. T. Donaldson, A. J. Bathgate, A. Burroughs, M. H. Davies, D. E. Jones, G. J. Alexander, J. C. Barrett, R. N. Sandford, and C. A. Anderson, "Dense fine-mapping study identifies new susceptibility loci for primary biliary cirrhosis," Nature Genetics, vol. 44, pp. 1137–1141, Sept. 2012.

[222] D. Lee, T. B. Bigdeli, B. P. Riley, A. H. Fanous, and S.-A. Bacanu, "DIST: direct imputation of summary statistics for unmeasured SNPs," Bioinformatics, vol. 29, pp. 2925–2927, Nov. 2013.

[223] B. Pasaniuc, N. Zaitlen, H. Shi, G. Bhatia, A. Gusev, J. Pickrell, J. Hirschhorn, D. P. Strachan, N. Patterson, and A. L. Price, "Fast and accurate imputation of summary statistics enhances evidence of functional enrichment," Bioinformatics, vol. 30, pp. 2906–2914, Oct. 2014.

[224] L. Jostins, "Immunobase data processing," June 2017. Personal correspondence.

[225] R. Horton, L. Wilming, V. Rand, R. C. Lovering, E. A. Bruford, V. K. Khodiyar, M. J. Lush, S. Povey, C. C. Talbot, M. W. Wright, H. M. Wain, J. Trowsdale, A. Ziegler, and S. Beck, "Gene map of the extended human MHC," Nature Reviews Genetics, vol. 5, pp. 889–899, Dec. 2004.

[226] S. Kitada, T. Hayashi, and H. Kishino, "Empirical Bayes procedure for es-

timating genetic distance between populations and effective population size," Genetics, vol. 156, pp. 2063–2079, Dec. 2000.
[227] M. A. Beaumont and B. Rannala, "The Bayesian revolution in genetics," Nature Reviews Genetics, vol. 5, pp. 251–261, Apr. 2004.
[228] W. E. Johnson, C. Li, and A. Rabinovic, "Adjusting batch effects in microarray expression data using empirical Bayes methods," Biostatistics, vol. 8, pp. 118–127, Jan. 2007.
[229] The Wellcome Trust Case Control Consortium, "Overview - WTCCC," May 2017.
[230] B. M. Psaty, C. J. O'Donnell, V. Gudnason, K. L. Lunetta, A. R. Folsom, J. I. Rotter, A. G. Uitterlinden, T. B. Harris, J. C. Witteman, E. Boerwinkle, and on Behalf of the CHARGE Consortium, "Cohorts for Heart and Aging Research in Genomic Epidemiology (CHARGE) Consortium: Design of Prospective Meta-Analyses of Genome-Wide Association Studies From 5 Cohorts," Circulation: Cardiovascular Genetics, vol. 2, pp. 73–80, Feb. 2009.
[231] C. Hans, A. Dobra, and M. West, "Shotgun Stochastic Search for "Large p " Regression," Journal of the American Statistical Association, vol. 102, pp. 507–516, June 2007.
[232] R. Plomin, C. M. A. Haworth, and O. S. P. Davis, "Common disorders are quantitative traits," Nature Reviews Genetics, vol. 10, p. 872, Dec. 2009.