Interpreting the role of non-coding genetic variation in human disease

by

Abhishek Sarkar

M.S. Computer Science, Massachusetts Institute of Technology, 2013

B.S. Computer Science, University of North Carolina at Chapel Hill, 2011

Submitted to the Department of Electrical Engineering and Computer Science in partial fulfillment of the requirements for the degree of

Doctor of Philosophy

at the

Massachusetts Institute of Technology

June 2017

Signature redacted

Author____________________________________

Department of Electrical Engineering and Computer Science
May 5, 2017

Signature redacted

Certified by____________________________________

Manolis Kellis
Professor of Electrical Engineering and Computer Science
Thesis Supervisor

Signature redacted

Accepted by____________________________________

Leslie A. Kolodziejski
Professor of Electrical Engineering and Computer Science
Chair, Department Committee on Graduate Students

Interpreting the role of non-coding genetic variation in human disease

by

Abhishek Sarkar

Submitted to the Department of Electrical Engineering and Computer Science
on May 5, 2017 in partial fulfillment of the requirements for the degree of
Doctor of Philosophy in Electrical Engineering and Computer Science

Abstract

One of the fundamental goals of human genetics is to identify the genetic causes of human disease to ultimately design novel therapeutics. However, two challenges have become readily apparent. First, the majority of genomic regions associated with disease do not implicate protein-altering variants but might instead alter gene regulation, making interpretation and validation more difficult. Second, the genomic regions associated with disease explain a fraction of the variance of associated phenotypes, suggesting human diseases are highly polygenic and that many additional regions remain to be discovered and characterized.

Here, we address these challenges by using functional annotation of the human genome spanning diverse data types: epigenomic profiles, gene regulatory circuitry, and biological pathways. We first develop a method to simultaneously select relevant genomic regions not yet associated with disease as well as select relevant functional annotations enriched in those regions. We show that both tissue–specific and shared regulatory regions are enriched for disease associations across eight common diseases.

We then characterize specific genetic variants in the selected regions, the gene regulatory elements they reside in, the cellular contexts in which those elements are active, their upstream regulators, their downstream target genes, and the biological pathways they disrupt across eight common diseases. We show that disease associations are additionally enriched in regulatory motifs of relevant transcription factors and in relevant biological pathways.

We finally investigate why predicted regulatory elements are enriched in disease-associated variants by framing the problem as Bayesian inference of hyperparameters in a structured sparse regression model. We propose an active sampling method to efficiently explore the hyperparameter space and avoid exponential scaling in the dimension of the hyperparameters. We show in simulation that our method can distinguish between possible explanations of the observed enrichments, and we characterize potential biases in the estimates.

Together, our results can help guide the development of new models of disease and gene regulation and discovery of biologically meaningful, but currently undetectable regulatory loci underlying a number of common diseases.

Thesis supervisor: Manolis Kellis

Title: Professor of Electrical Engineering and Computer Science

Contents

Chapter 1

Introduction

The promise of the human genome project was to spark the development of novel therapeutics for the whole gamut of human disease by jump-starting the search for the genes which cause disease[1]. Indeed, 15 years into this endeavor, the field of human genetics has seen a revolutionary explosion in the ability to identify genes which could cause human disease. This scientific revolution required new biological insights into the human genome beyond just the content of its DNA sequence. These insights were spurred by new technologies to gather data at unprecedented scale and detail, and new computational techniques to analyze the data. Specifically, the field produced comprehensive catalogs of genetic variation across the human population, and advanced DNA microarray technology to make measuring millions of genetic variants in the population cost-efficient at the scale of hundreds of thousands of individuals at present.

These biological insights and technological insights enabled one of the most powerful tools to understand human disease: the genome-wide association study[2] (GWAS). The idea of GWAS is to use genetic variation as a marker for genetic loci (regions of the genome) which cause disease. By systematically measuring all genetic variants in the human genome and identifying differences in the patterns of variation across disease cases and healthy controls, we could in principle predict which genes cause disease. To date, over 2,000 genome-wide association studies (GWAS) have been conducted, providing an unprecedented treasure trove of information for understanding the molecular basis of human disease. The largest of these studies spans nearly half a million individuals, and in the near future national biobanks (like the US Precision Medicine Initiative and the UK Biobank) and industry efforts (lead by 23andMe, Mayo Clinic, etc.) will amass data on multiple millions of individuals. However, these massive datasets and pioneering results have not yet led to an avalanche of new therapeutics because we still do not understand the mechanisms through which these thousands of genetic associations act.

Over the past 15 years, the field has increasingly recognized the challenge of translating GWAS findings into novel therapeutic targets:

1. **GWAS cannot directly identify causal variants.** The key insight needed to make performing GWAS possible at scale was that nearby genetic variants show highly correlated patterns of variation in the population, a property known as linkage disequilibrium. This property dramatically reduces the cost of assaying millions of genetic variants in many individuals: instead of directly measuring the genotypes at millions of positions in the genome, we can measure a subset of them (5-10%) and impute (statistically infer) the unobserved data. This

same property makes finding the causal genetic variant a highly challenging open problem[3]. A particular genetic variant identified by GWAS actually implicates a large genomic region containing hundreds of other variants which are highly correlated due to linkage disequilibrium. Ultimately, identifying the causal variant in a locus requires painstaking functional genomic experiments, proposing a mechanism for its action and validating that mechanism in in vitro experiments in e.g. individual cells and in vivo experiments in model organisms. Computational methods modeling or predicting the impact of the human genome on human traits, such as those presented in this thesis must account for this correlation structure (Chapter 2).

2. **A growing body of evidence suggests that human diseases are highly polygenic.** This is the first fundamental observation motivating this thesis: common human diseases involve potentially thousands of additional loci which even the largest meta-analyses to date are underpowered to detect[4–6]. One consequence has been the development of international levels of collaboration in performing GWAS. The key observation by the field has been that as GWAS sample sizes increase, we continue to find more disease-associated loci. This observation has spurred the formation of of dozens of large scale consortia, each spanning multiple countries and individual labs. Each has amassed tens or hundreds of thousands of individuals, giving increased statistical power to detect genetic loci. For example, recent studies of height[7] and schizophrenia[8] have exceeded 100,000 individuals and have now identified hundreds of genetic associations with these phenotypes.

   One of the unusual features of modern GWAS data is that it is often very difficult to acquire genotype observations for these large cohorts of individuals due to data privacy laws and agreements. However, in place of this individual-level data, GWAS data is often summarized as per-variant association statistics, necessitating the development of novel computational and statistical methods to take advantage of this new data type, but opening the door for much more efficient algorithms[9]. In the first part of this thesis (Chapters 2-3), we rely on this summary-level data; in the final part (Chapter 4), we use individual level data.

3. **The vast majority of GWAS associations are non-coding.** This is the second fundamental observation motivating this thesis: one of the great surprises in the wake of the human genome project was just how few genes there actually are (roughly 20,000), and how little of the human genome codes for proteins (1.5%). Unsurprisingly (given our present state of knowledge), 93% of genetic variants marking GWAS loci lie outside protein-coding regions, and 80% of loci do not implicate any protein-altering variant at all[10,11]. Even for protein-coding associations, identifying the causal variant and understanding where and how the disrupted gene acts in cells and organisms is still a challenge. But interpreting non-coding associations requires first understanding what the other 98.5% of the human genome actually does.

   A growing body of evidence suggests that much of the non-coding genome affects gene regulation: the circuitry which controls the levels at which genes are transcribed into RNA and translated into proteins. In parallel to the formation of large scale GWAS consortia, large scale efforts such as the ENCODE consortium[12] and NIH Roadmap Epigenomics project[13] were formed in order to gain insights into the non-coding genome. The broad goal of these projects has been to characterize the non-coding genome by finding and annotating regions which exhibit interesting biochemical activity through systematic experimental profiling across hundreds of human cell types, and predicting the functional role of these regions.

> Over the past five years, the field is increasingly recognizing the utility of these annotations in interpreting non-coding genetic associations. Conceptually, we want to use genetic association with disease to identify the relevant annotations, and simultaneously use the relevant annotations to identify new genetic loci.

The goal of this thesis is to tackle the still unsolved problem in the field: to go beyond merely identifying the relevant annotation and translate non-coding genetic associations into actionable biological insights. For these non-coding associations, we don't know the target gene, let alone the causal variant, the cell type in which it acts, the regulators that control it, the effect on the target gene, or the intermediate phenotypes which mediate its effect on the observed phenotype. In addition to these sorts of mechanistic insights, we seek to investigate the genetic architecture of common diseases: how many causal genetic variants are in the human genome, what is the distribution of their effect sizes, how common are they in the population, and where do they reside in the human genome? The answers to these questions will motivate the design and execution of future genetic studies, whether genotyping-based GWAS of millions of samples, population-based studies of rare variation in whole genome sequences[14], extreme phenotype designs[15], or rich phenotyping studies in prospective or longitudinal cohorts like the UK Biobank.

In this thesis, we develop new computational methods to jointly analyze diverse data such as GWAS, epigenomic profiles, regulatory circuits, gene expression, and intermediate phenotypes in order to answer these questions and elucidate the mechanistic basis and gene-regulatory architecture of human disease. We make a number of contributions:

1. We develop a heuristic for univariate feature selection and a permutation testing procedure for enrichment which allows us to identify not only relevant annotations enriched in weak associations, but implicate a specific set of annotated regions for further investigation. We apply these methods to identify hundreds of genetic loci associated with eight diseases spanning autoimmune, psychiatric, and metabolic disorders (Chapter 2).

2. We exploit a latent representation of the gene regulatory function of the non-coding genome combining diverse regulatory annotations spanning multiple cell types, experimental assays, and computational pipelines to derive relevant biological annotations. We show these annotations account for overlap between directly observed annotations, and identify both tissue-invariant and tissue-specific gene regulatory elements associated with the eight diseases (Chapter 2).

3. We dissect the functional role of the predicted regulatory elements by studying the nearby candidate target genes. We show that these genes are enriched (over-represented) in a number of known biological pathways, but only a small fraction of the genes are already identified by GWAS (Chapter 3)

4. We dissect the mechanism of action of the predicted regulatory elements by identifying upstream master regulator transcription factors whose binding is disrupted. We show that constitutively marked enhancer regions predicted to be tissue-invariant disrupted by weak associations may not be constitutively active due to tissue-specific expression of the upstream transcription factor (Chapter 3).

5. We extend a Bayesian hierarchical model, developing an efficient approximate inference algorithm to learn the genetic architecture of human diseases and interpret regulatory enrichments. We frame the question as a hyperparameter inference problem and extend an active learning scheme for this problem. We demonstrate that our method can distinguish genetic

architectures in simulated data, and characterize the biases of the method (Chapter 4).

Chapter 2

Functional enrichment of enhancer regions

2.1 Background

We have two main goals in this chapter:

1. For each disease of interest, identify the relevant annotations
2. For that disease, identify the relevant genetic loci based on the relevant annotations

To achieve these goals, we will use two fundamental types of data: genome-wide association study (GWAS) summary statistics and epigenomic annotations. Here, we review the concepts underlying these data and their interpretation.

2.1.1 Genome-wide association studies

Genome-wide association studies seek to identify regions of the genome whose patterns of inheritance is correlated with a phenotype of interest[2]. In order to identify these patterns of inheritance, we use naturally occurring genetic variation as markers for regions of the genome, and measure these variants in large samples of unrelated (distantly related) individuals[16]. This strategy is opposed to linkage analysis, a historical approach tracing inheritance of a phenotype and a genetic marker directly through a pedigree or family tree[17,18].

GWAS typically focuses on the two most frequently occurring types of genetic variation in the human genome: single nucleotide polymorphisms (SNPs; order 1 per 100 bases) and small insertions or deletions (indels; order 1 per 1000 bases). These variants typically take one of two alleles (values or states) in the population. We distinguish them by their relative frequency, identifying the major (more frequent) and minor (less frequently) allele. However, each individual has two copies of the varying position (one inherited from their mother and one from their father). Therefore, we can code the genotype of an individual for a variant according to how many copies of the minor allele that individual has inherited.

There are three key advances which underlie modern GWAS:

1. Large scale efforts such as the International HapMap Project[19] and the Thousand Genomes project[20,21] built comprehensive catalogs of genetic variation and also of the patterns of linkage disequilibrium (covariance between nearby genetic variants). These catalogs now span order 10^8 genetic variants, of which order 10^7 commonly occur in the human population (minor allele frequency MAF > 0.01).
2. DNA microarray technology matured, allowing researchers to rapidly and cost-efficiently measure genotypes at order 10^6 genetic variants.
3. Efficient computational models made it possible to impute (statistically infer) genotypes at the remaining unobserved variants (which we describe below). These algorithms motivated efforts to pick a set of representative tag variants which would maximize the number of unobserved variants which could be confidently imputed, and to design DNA microarrays which would measure genetic variation at those tags[22–24].

We conduct a GWAS by collecting a sample of n individuals and genotyping them at p genetic variants. A typical GWAS data set is then a dense $n \times p$ matrix of genotypes X, after using imputation to acquire p on the order 10^7, and an $n \times 1$ vector of phenotypes y (Figure 2.1). For each of the p variants, we fit a univariate generalized linear model regressing y against X_j, the $n \times 1$ column vector for that variant:

$$E[y] = g(X_j\theta_j)$$

We fit linear regression ($g(x) = x$) for continuous phenotypes such as human height or logistic regression ($g(x) = 1/(1 + exp(-x))$) for binary phenotypes such as a particular disease. For linear regression, this leads to a well-known closed form solution:

$$\hat{\theta} = (X'X)^{-1}X'y$$
$$V[\hat{\theta}] = (X'X)^{-1}V[y]$$

In practice, we include additional covariates such as gender, environmental covariates, and population structure (intuitively, the relationships between the individuals based on ancestral relationships) as terms in the generalized linear model. More sophisticated models to account for these covariates have been developed, but are not explored in this thesis.

For each of these p models, we need to answer a decision problem: is the true regression coefficient θ for that variant non-zero? Typically in GWAS, we take a frequentist approach to answer this decision problem by performing a hypothesis test. In the first part of this thesis (Chapters 2-3) we rely heavily on hypothesis testing, and need to explicate important definitional aspects[25]. Briefly, a hypothesis test (following Neyman and Pearson) is defined by:

1. The null hypothesis H_0 and alternative hypothesis H_1, statements about the parameters of the underlying generative process.
2. The test statistic t, a function of the observed data.
3. The null distribution of the test statistic. This is the distribution of t assuming the data were actually generated from H_0. Estimating this distribution (or even representing it, in the case where it is not a standard distribution) is one of the key challenges in designing tests.

4. The significance level. This is the maximum permissible Type 1 error rate (intuitively, false positive rate), the probability of rejecting H_0 when the data actually came from H_0.

5. The rejection region, a region of the support (set of possible values) of the null distribution. If the test statistic falls in this region, we reject the null hypothesis that the data were generated from H_0. The boundary of this region depends on the significance level and the alternative hypothesis.

In the case of GWAS, one typical test for a regression model $y = X\theta + \epsilon$ is the Wald test:

1. The null hypothesis is $H_0 : \theta = 0$ and the alternative hypothesis is $H_0 : \theta \neq 0$

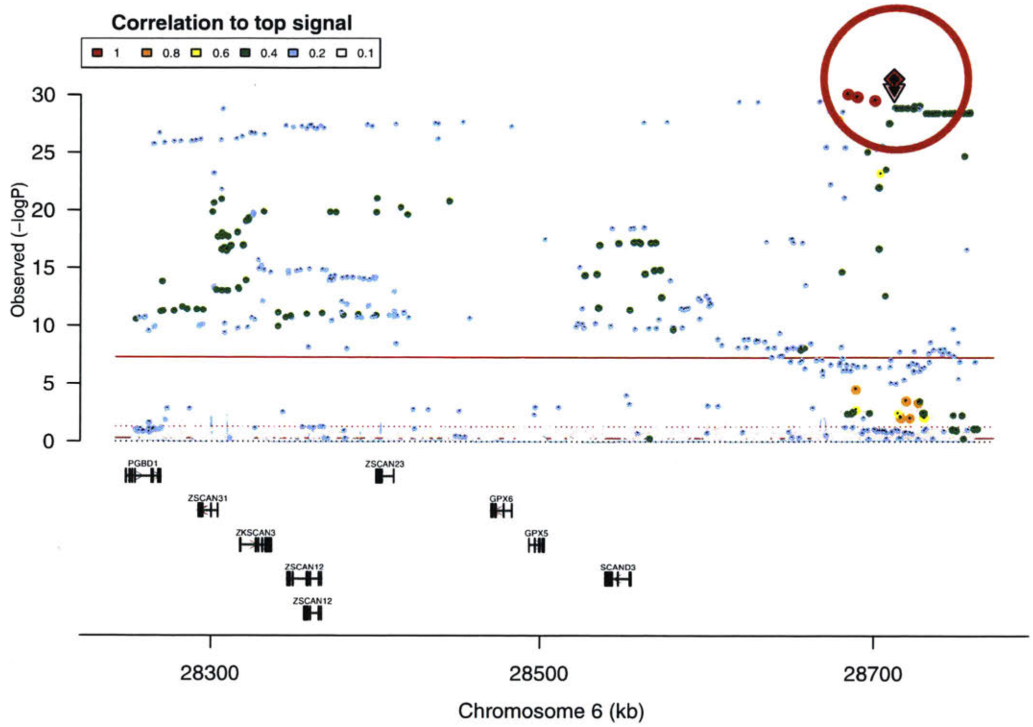
2. The test statistic is the Wald statistic $\chi^2 = (\hat{\theta})^2/V[\hat{\theta}]$. Note that $\hat{\theta}$ is indeed a function of the data only, and not the parameter of interest θ.

3. The null distribution of the Wald statistic is known to be the chi-square distribution with one degree of freedom.

4. A standard significance level is $\alpha = 0.05$.

5. The rejection region is the region above the $1-\alpha = 0.95$ quantile of the chi-square distribution, i.e. $\chi^2 \geq 3.84$. We refer to the threshold as the $1 - \alpha$ critical value. In this case, the region is only one-tail of the distribution because our statistic is non-negative after squaring.

In summary, for each genetic variant we estimate the statistic χ^2 and then make a decision whether or not to reject the null hypothesis H_0. Note that nowhere yet have we introduced the notion of a p-value. Conceptually, the p-value is the probability of observing a test statistic at least as far from the null as we did, given the data was generated from H_0. In the case of the Wald test, this is the area under the probability density function of the chi-square distribution with one degree of freedom to the right of the estimated test statistic χ^2. Then, an equivalent description of the hypothesis test is to compute the p-value and compare it to the desired significance level. If $p < 0.05$, then we reject the null hypothesis.



One challenge in interpreting GWAS is that the co-linearity of predictors due to linkage disequilibrium implies test statistics will be highly correlated to each other. This correlation means that the predictor with the best univariate test statistic is not necessarily most likely to be the true causal signal, and instead only implicates a region of association harboring many other highly correlated variables which could be the causal variant[3]. For example, in Figure 2.1.1 we display a region of chromosome 6 implicated by a schizophrenia association (circled). Just from the GWAS alone, we cannot directly say that alterations in any gene in this region cause schizophrenia.

Another challenge in GWAS is accounting for multiple testing when controlling the false positive rate of hypothesis testing[2]. Conceptually, we set the significance level to control the Type 1 error rate of a single test. However, having fixed that false-positive rate (e.g., at 0.05), we perform order 10^7 hypothesis tests in GWAS. Then, we expect to find 5×10^5 false positives, which in typical GWAS covers all of the rejected hypotheses.

Historically, the field has corrected for multiple testing using a procedure known as Bonferroni correction due to its simplicity and strong guarantees. Bonferroni correction guarantees control of the familywise error rate (FWER): the probability that at least one rejected hypothesis was a Type 1 error (and should not have been rejected). The simplest description of the Bonferroni correction is as follows: if we perform n hypothesis tests, and desire to control the FWER at level 0.05, then we should set the significance level of each individual test to α / n. Understandably, this is a highly conservative correction, but has lead to the replicable results we now rely upon in GWAS. However, as alluded to in the introduction, the field recognizes that this correction is likely over-conservative, and that many additional genetic associations remain to be discovered and characterized. Historically, the significance level required for any single genetic variant in a GWAS was established to be 5×10^{-8}; we refer to this level as genome-wide significance[26]. One of the key problems tackled in this thesis is how to relax this type of correction to discover some of the additional genetic associations which do not reach genome-wide significance.

The most important class of alternative methods to correct for multiple testing control the false discovery rate (FDR): the expected proportion of rejected hypotheses which are Type 1 errors[27]. The field of FDR theory is very rich, and we will rely on many recent theoretical results in this thesis, but here we give a high level overview. There are two main ways to control the FDR:

1. Directly adjusting the observed p-values[27] (equivalently, the significance level)
2. Fitting a Bayesian mixture model where the observed z-scores (equivalently, p-values) are assigned to null or alternative hypotheses, and estimating the q-value, the posterior probability that the observed z-score came from the alternative hypothesis[28].

In this chapter, we will rely on the Benjamini-Hochberg (BH) procedure due to its simplicity. The BH procedure resembles the Bonferroni procedure:

1. Order the n p-values from smallest to largest
2. Find the smallest k such that $p_k \leq \alpha * k/n$
3. Reject all hypotheses $p_1, \ldots, p_k$

Intuitively, the reason the BH procedure is less stringent than Bonferroni correction is that the significance threshold increases as we consider more tests. Some care has to be taken with ties in the ordering of p-values, which is an important problem in this chapter. The simplest strategy is to give the tied p-values the same rank k and leave the algorithm unchanged.

2.1.2 Epigenomic annotations

We have introduced the central computational problem of this thesis as identifying relevant biological annotations for disease. The central biological problem of this thesis is to interpret the role of non-coding genetic variation in human disease. The vast majority of all genetic loci (regions) found by GWAS do not implicate a protein-altering variant, but rather some non-coding genetic variant within the locus[10].

In order to interpret the function of non-coding variation, we need to introduce two ideas:

1. Transcriptional regulation is the process by which genes are activated, repressed, or modulated over time or in reaction to stimuli.
2. The epigenome (literally, "above the genome") is a collection of biochemical modifications to the DNA molecule which are associated with the function of the underlying sequence they are near.

To rephrase the central biological problem of this thesis, we want to identify the genes whose regulation is disrupted by non-coding genetic variation associated with disease and the biological mechanisms by which they are dysregulated. We will focus on these goals in Chapter 3. Here, we outline the most important aspects of regulation as they relate to this chapter (Figure 2.2):

1. Genes are organized into a gene body (containing the coding sequence), beginning with the transcription start site (TSS) and preceded by a region called the promoter.
2. In order for a gene to be transcribed into RNA and ultimately translated into a protein, it must be accessible to the enzyme RNA polymerase (PolII).

3. The promoter is responsible not only for recruiting PolII, but for recruiting other proteins known as transcription factors which are required for the region to become accessible and for PolII to bind.

4. In addition, many genes require distant regions called enhancers to also interact with the promoter. In general, enhancers also recruit transcription factors which must interact with the promoter and PolII to allow transcription to occur. These distant regions are brought into proximity with the promoter region through physical deformation and interaction of the DNA molecule.

The key biological question underlying this chapter is how to identify which parts of the genome are relevant for transcriptional regulation in which human cell types. Every cell in the human body has the same genome; however, different human tissues and cell types exhibit dramatically different phenotypes and perform dramatically different functions. These differences are (partially) explained by transcriptional regulation: the process by which cells differentiate from stem cells into mature adult cells with specialized functions involves a complex developmental program involving activation and repression of specific genes at specific times. Moreover, the functions of the mature cells themselves involve activation and repression of specific genes in response to stimuli.

The key insight required to make some progress in identifying putative transcriptional regulatory elements is that the human epigenome also varies across human cell types and is indicative of the function of the underlying sequence. Historically, epigenetic modifications were defined as non-DNA sequence elements which nevertheless were heritable (passed from parent to offspring). The classical example of such an epigenetic modification is DNA methylation in which a cytosine nucleotide is modified with a methyl group. This modification has long been known to be directly correlated with the expression of the nearby gene (the level at which the gene is transcribed into mRNA; intuitively, its activity). Methylation in the flanking region surrounding the gene is associated with repression of the gene's activity, while methylation within the gene body is associated with active transcription.

Over the past five years, the field has increasingly recognized the importance and utility of histone modifications (also known as chromatin marks), which are modifications to the histone proteins around which the DNA molecule is wrapped. These modifications are not thought to be heritable between generations, and therefore we refer to them as epigenomic modifications rather than epigenetic. Briefly, histone proteins form complexes called nucleosomes which are vital to the organization of DNA within the nucleus. The entire DNA molecule is too large to fit within the cell nucleus, and must be tightly packed into a dense structure called chromatin. However, these proteins admit chemical modifications, such as the addition of methyl groups.

Large-scale sequencing efforts such as the ENCODE project and Roadmap Epigenomics Consortium have made progress in characterizing the patterns of variation in these chromatin marks across dozens of marks and across hundreds of human cell types and tissues. The most important of these marks are:

1. Tri-methylation of histone 3, lysine 4 (H3K4me3): associated with promoters

2. Methylation of histone 3, lysine 4 (H3K4Me1): associated with promoters and enhancers

3. Tri-methylation of histone 3, lysine 36 (H3K36Me3): associated with transcribed genes

4. Tri-methylation of histone 3, lysine 9 (H3K9me3): associated with repetitive elements and heterochromatin (inactive regions)

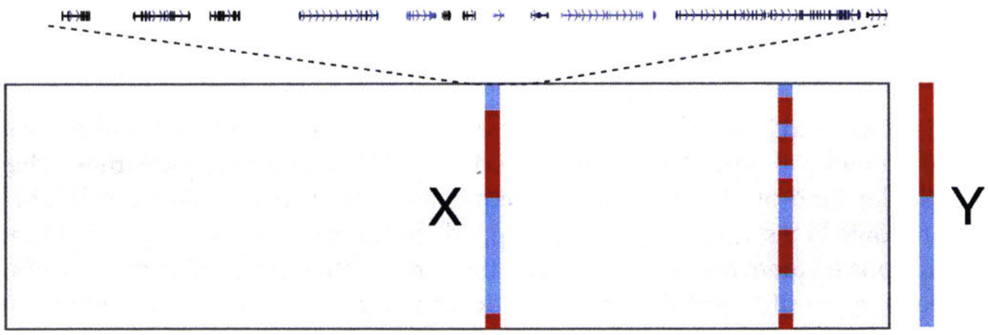


Figure 2.1: Conceptual illustration of genome-wide association studies. We observe $n \times p$ matrix X of genotypes and $n \times 1$ vector of phenotypes y. We fit univariate models regressing phenotype y against genotypes at each variant x_j and perform a hypothesis test to select relevant columns of X. Due to correlations between the columns of X, the hypothesis test implicates a genomic region (locus) rather a single variable, which may contain multiple genes which could be causal for disease.

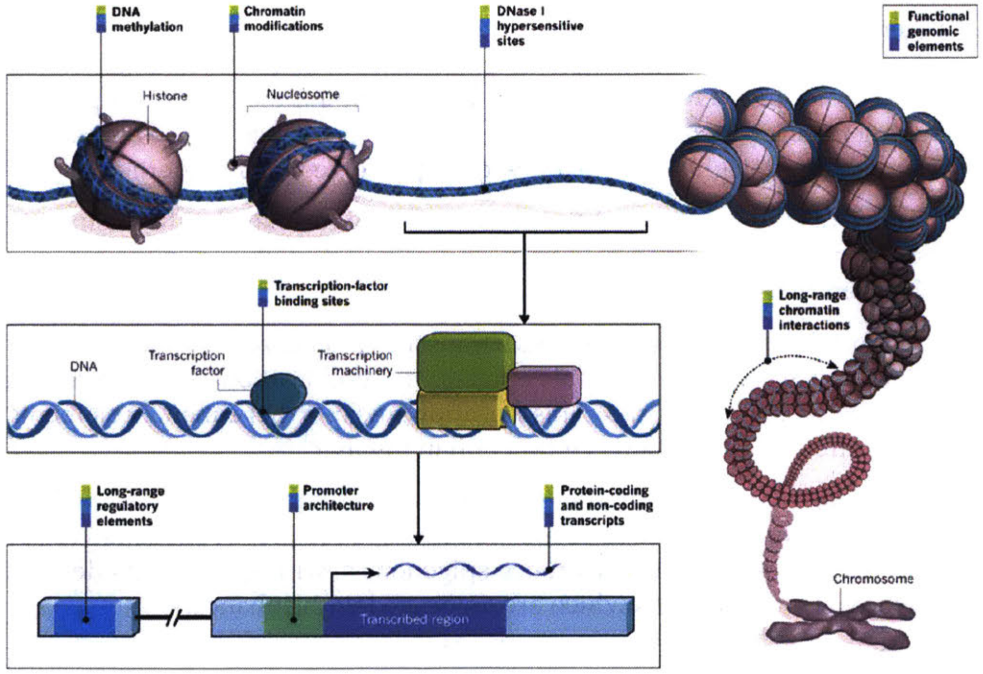

DNA methylation
Chromatin modifications
DNase I hypersensitive sites
Functional genomic elements
Histone
Nucleosome
Transcription-factor binding sites
Long-range chromatin interactions
DNA
Transcription factor
Transcription machinery
Long-range regulatory elements
Promoter architecture
Protein-coding and non-coding transcripts
Transcribed region
Chromosome

Figure 2.2: Illustration of gene regulation and epigenomic modifications. Genes consist of a protein-coding sequence, a promoter sequence, and long-range elements known as enhancers (bottom). In order for transcription to occur, the promoter (and associated enhancers; not shown) must recruit RNA polymerase as well additional transcription factors (middle). In order to predict where promoters, enhancers, and other regulatory elements reside in the genome, we can use epigenomic modifications such as chromatin modifications to the histone backbone of the DNA molecule which are associated with the function of the underlying sequence (top).

5. Tri-methylation of histone 3, lysine 27 (K3K27me3): associated with bivalent enhancers, which show characteristics of both promoters and enhancers

6. Acetylation of histone 3, lysine 27 (K3K27ac): associated with active enhancers

The ENCODE project and Roadmap Epigenomics Consortium have profiled these chromatin marks across diverse human cell types and tissues, producing 127 reference epigenomes. These profiles rely on a particular kind of high throughput sequencing experiment called Chromatin Immunoprecipitation followed by sequencing (ChIP-Seq), which we briefly outline. The fundamental idea of high throughput sequencing is to fragment the target DNA into a library of reads, sequence each of the reads in parallel, and then computationally align the reads back to the position in the original genome they came from.

For the purposes of this discussion, we will simply assume we can perform all of these steps without giving the details. The key idea of ChIP-Seq is that if proteins are bound to the DNA molecule, then after fragmenting the DNA we can use antibodies to select those fragments which are bound, using the same mechanisms which the immune system uses to recognize proteins on the surface of host and pathogen cells. Now, after performing the sequencing and aligning the reads back to original genome, we will only find reads where protein was bound to the genome, and we can analyze the count of reads at each position in the genome as a discrete signal to predict where precisely proteins are physically bound.

To profile epigenomic modifications, we need antibodies which can recognize histone proteins with the modification of interest and need to perform ChIP-Seq experiments for each cell type and modification of interest. Overall, the Roadmap Epigenomics Consortium performed several thousand experiments in order to produce 127 reference epigenomes. These reference epigenomes have lead to a number of insights into the non-coding genome which underlie this thesis:

1. Combinations of epigenomic modifications are highly correlated with the function of the regions they modify. These combinations can be learned de novo using unsupervised methods such as Hidden Markov Models[29].

2. Epigenomic modifications associated with transcribed genes and promoters are largely shared across all reference epigenomes, while epigenomic modifications associated with enhancers vary widely between epigenomes. This variation suggests that enhancer elements are important for cell differentiation[30] and cell function in disease[31,32].

The approach we take to answer the biological question underpinning this chapter is to annotate regions of the genome important for these mechanisms in a diverse set of human cell types. These annotations are summaries of the 127 reference epigenomes, with special consideration to combinations of epigenomic modifications which are associated with enhancers.

2.2 Methods

2.2.1 Genome-wide association summary statistics and regulatory annotations

We downloaded summary statistics for AD from the International Genomics of Alzheimer's Project (see URLs); BIP and SCZ from the Psychiatric Genetics Consortium; CAD from the CARDIOGRAM consortium; CD from the International Inflammatory Bowel Disease Genetics Consortium; RA (`https://www.broadinstitute.org/ftp/pub/rheumatoid_arthritis/Stahl_etal_2010NG/`

RA_GWASmeta2_20090505-results.txt); T1D from the Type 1 Diabetes Genetics Consortium through T1DBase[33]; and T2D from the DIAGRAM Consortium.

We downloaded ChromHMM segmentations from the Roadmap Epigenomics project; clustered regulatory regions from the Regulatory Regions Map (see URLs); genic annotations from the GENCODE project; CAGE–predicted transcription start sites (ftp://genome.crg.es/pub/Encode/data_analysis/TSS/Gencodev10_CAGE_TSS_clusters_May2012.gff.gz); predicted motif instances from the ENCODE project; and motif enrichments (predicted regulators) from the Roadmap Epigenomics project.

2.2.2 Imputation of summary statistics

One key idea underlying the work in this thesis is that using the most descriptive and fine-grained annotation of the non-coding genome will allow us to make the most specific predictions of biological mechanisms. In order to exploit the highest resolution annotations (described below), we need to impute GWAS data to the most comprehensive available catalog of genetic variation. Otherwise, we lose statistical power (equivalently, recall, sensitivity, or true positive rate) to detect the relevant annotations enriched in associations, simply because not enough variants are observed to fall in those annotations.

As described above, imputation of genotypes is possible because genetic variants have a covariance structure called linkage disequilibrium (LD). The Thousand Genomes Consortium (1KG) has sequenced over 1,000 individuals to build up a catalog of genetic variation and an estimate of the covariance structure induced by LD. Of note, 1KG has not measured the genotype of the individuals at each of the genetic variants, but has also statistically inferred the haplotypes of these individuals (intuitively, for each individual, the set of alleles at different genetic variants which were co-inherited from the father, and which from the mother). The covariance structure we need is exactly the covariance between these haplotypes. To perform imputation, we downloaded Thousand Genomes (1KG) reference haplotypes in OXSTATS format (September 2013 version, no singletons).

We used a model called ImpG-Summary to impute summary statistics without access to the underlying genotypes. As mentioned above, this feature of GWAS data sets imputation in this context apart from other applications of imputation (e.g. matrix completion). The key observation of ImpG-Summary is that Pearson correlation between two $2n \times 1$ haplotype vectors denotes not only the level of LD between them, but also the covariance between the GWAS z-scores between those two variants. As a consequence, under the null model of no association, we have:

$$z \sim N(0, R)$$

where where R is the $p \times p$ covariance matrix $X'X/n$. This fact immediately gives an imputation algorithm, exploiting multivariate Gaussian identities:

$$\begin{bmatrix} z_o \\ z_u \end{bmatrix} \sim N\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} R_{oo} & R_{ou} \\ R_{uo} & R_{uu} \end{bmatrix}\right) \tag{2.1}$$

$$z_u \mid z_o \sim N(R_{uo}R_{oo}^{-1}z_o, R_{uu} - R_{uo}R_{oo}^{-1}R_{ou}) \tag{2.2}$$

where z_o denotes the observed z-scores and z_u denotes the unobserved z-scores.

In practice, this algorithm requires some computational tricks:

1. Most obviously, we cannot operate on the entire $p \times p$ matrix. However, another key feature of LD is that it decays with distance, and also has a block structure (within regions high LD, but between regions low LD). This means imputation can proceed in small windows over the genome, overlapping to avoid edge effects.
2. We have to regularize the estimated matrix $\hat{R} = X'X/n + \lambda I$ (analogous to ridge regression) to ensure we can take the necessary inverse R_{oo}^{-1}.
3. Patterns of LD differ between different human ancestry groups, and therefore we need to pick an ancestry-matched set of haplotypes to accurately impute z-scores.

We used ImpG-Summary with default parameters (regularization, window size) and European 1KG samples to impute summary statistics for five of the eight disease we studied (BIP, CAD, RA, T1D, and T2D) into all SNPs with MAF > 0.01 in 1KG European samples.

As described above, imputation requires access to the regression z-scores in order to exploit the properties of the Gaussian distribution. However, for a number of diseases we studied these were not available: instead, only the p-values from the test used in each study were provided. In this case, we need to utilize another identity: if $z \sim N(0, 1)$, then $z^2 \sim \chi^2(1)$. Given a p-value, we can use the inverse survival function (one minus the inverse cumulative density function) to infer the value of z^2. However, we cannot merely take the square-root to derive the original z-score since this loses sign information (which is important since it has to be consistent with the covariance implied by the LD matrix R).

In order to assign signs to z-scores inferred from χ^2 statistics, we additionally need the odds ratio for each test. For a genetic test of association, the odds ratio equals the increase in the log odds of having the phenotype of interest given one additional copy of the effect allele. Note that this means the odds ratio depends how genotypes are coded, specifically which allele is being counted (in the preceding discussion, the minor allele). Also note that the odds ratio only makes sense for a binary (disease) phenotype; in this case we can fit a logistic regression and the estimated coefficient is the log-odds ratio (by the definition of logistic regression). In the case of a continuous phenotype, the regression coefficient itself is the estimated change in the phenotype given one additional copy of the effect allele (see Chapter 4 for further discussion of the units of regression effects).

The interpretation of the odds ratio is straightforward: an odds ratio greater than one means the effect allele increases the chance of getting the disease (positive sign), and an odds ratio less than one means the effect allele reduces the chance (negative sign).

Of note, we used data for T1D which only had p-values and not odds ratios. In order to assign signs of effects for T1D, we directly imputed genotypes for the Wellcome Trust Case Control Consortium study of T1D and took the sign from the single-SNP association test.

We downloaded probe identifiers, hg19 positions, and strand information (`http://www.well.ox.ac.uk/~wrayner/strand/`) to convert positions to hg19 and used GTOOL version 0.7.5 to align all genotypes. We used PLINK version 1.09b to produce hard genotype calls with genotype probability threshold 0.99 and remove all SNPs and samples excluded from the original study. We used SHAPEIT2 v2.r644 (ref.[23]) to exclude unalignable SNPs and phase the case and control cohorts independently for each autosome. We used default values for all model parameters.

We used IMPUTE2 version 2.3.0 (ref.[34]) to impute into all SNPs and indels with MAF in European samples > 0.01. We divided the autosomes into 5 MB windows and threw out windows with fewer than 100 array probes. We used SNPTEST version 2.5.1 (ref.[35]) to compute association β-values using maximum likelihood estimates of an additive model. We included 10 principal components computed using GCTA 1.24 (ref.[36]) on the hard-called array genotypes. We made extensive use of GNU parallel[37] to facilitate the analysis.

2.2.3 Functional enrichment of enhancer annotations

Conceptually, we hypothesize that for each disease, the distribution of GWAS p-values for variants within relevant annotations will be different from the distribution for those outside the annotations. Specifically, we expect that the p-values within relevant annotations will be biased to be smaller (i.e., the estimated effect is less likely to have been observed if there truly no effect) than those outside the relevant annotations.

To relate this concept to the remainder of the thesis, here we are performing a univariate feature selection for a regression of genotype against phenotype, based on the marginal p-value of each feature (genetic variant). The key problem addressed in this chapter is how to pick the p-value threshold for the selected features. We note that in general the method presented in this chapter will not pick the true number of relevant features. One motivation for the Bayesian method presented in Chapter 4 is exactly to estimate this parameter directly from the data. Chapter 3 addresses the problem of interpreting the selected features (genetic variants) and pushing the insights gained in this chapter into testable biological predictions.

The obvious approach to compare the two distributions (inside and outside the annotation of interest) would be to divide the GWAS p-values into two subsets based on the partition induced by the annotation and apply a standard test such as the Mann-Whitney test or Kolmogorov-Smirnov two-sample test. However, these approaches are not valid for this problem because they assume that the two samples being compared are independently drawn, which is not true for a partition of GWAS p-values.

To prove this point, recall that the z-scores (and therefore the p-values) of nearby variants are correlated, and consider pairs of highly correlated variants where one is within the annotation of interest and one is outside. After partitioning the pairs by the annotation, it is obviously not true that the two observed partitions are independent.

Instead of trying to directly compare the distribution of p-values partitioned by the annotation of interest, we instead test for enrichment (over-representation) of associated variants within the annotation of interest. Conceptually, we observe some number of associated variants overlap with the annotation of interest, and seek to estimate how likely this observation would be under the null hypothesis of no enrichment.

Although this general approach has long been established in the field, there are number of problems which have only been recently been appreciated.

1. Historically, enrichment tests have only used the reported variant (with the lowest p-value), and asked if it directly overlapped the annotation of interest, without consideration to the fact that it might only be correlated to the true causal variant.

2. Enrichment tests have also historically only considered the strongest associations after Bonferroni correction. We know Bonferroni correction is over-conservative for GWAS itself, and therefore that enrichment of Bonferroni-corrected statistics will likely also be over-conservative.
3. Estimating the distribution of the statistic under the null distribution of no enrichment remains an open question.

We make two important contributions:

1. We design a procedure which considers the entire set of observed of p-values, and corrects for correlations between the observed p-values, in order to define a heuristic cutoff for association (which is typically much less stringent than Bonferroni correction)
2. We design a non-parametric bootstrapping procedure which controls for other genetic confounders

To pick a heuristic p-value cutoff, we computed enrichment curves for the annotations using an approach inspired by Gene Set Enrichment Analysis[38]. The key idea is that as we consider increasing numbers of GWAS variants (with decreasingly stringent p-value cutoff), we will encounter more overlaps with relevant annotations than with irrelevant annotations. For example, if we use a GWAS data set for Type 1 Diabetes, an auto-immune disorder, we expect that enhancer elements predicted in immune cell types will be more relevant than elements predicted in brain cell types. Indeed, by plotting the cumulative enrichment (**Figure 2.3**), we can order the relative importance of annotations based on the magnitude of the curves and determine a heuristic p-value cutoff based on the first inflection points of the curves.

One challenge in applying this methodology to GWAS p-values is that as we consider increasing number of GWAS variants, we may not necessarily be considering new genetic signals, but rather only finding more variants highly correlated with those we have already considered. In order to account for this possibility, we first pruned the associations, grouping together variants into loci such that the observed associations among those variants are due to correlation to some common causal variant(s) within that locus. In particular, we found a set of tag variants representing these loci with pairwise $r^2 < 0.1$, and then assigned the remaining variants to the best tag variant they are correlated to. We computed pairwise correlations between pairs of variants in the Thousand Genomes European samples within 1 megabase and with $r^2 > 0.1$, using the fact that the entries of the covariance matrix $X'X/n$ of the reference haplotypes are exactly the Pearson correlations. We then pruned to the desired threshold by iteratively picking the top-scoring variant (breaking ties arbitrarily) and removing the tagged variants until no variants remained. We ranked loci using the p-value of the lead SNP.

We scored each locus as the proportion of variants in the locus falling in a functional region, using BEDTools version 2.24 (ref.[39]) to compute overlaps. Intuitively, this score penalizes overlaps which occur in large genetic loci containing many correlated variants. Although the method does not explicitly assume each locus contains a single causal variant (which is a strong assumption, but standard for the field), in essence the score of each locus is the probability that a causal variant falls within the annotation of interest (assuming every variant is observed, which is not strictly true).

To plot the curve, we compared the cumulative observed total score against the expected score for every 100 loci. We computed the expected score as the total genome-wide score (considering all of the loci) multiplied by the cumulative proportion of loci seen so far. We plotted the difference normalized by the total score genome-wide.

In order to use these curves to estimate a heuristic p-value cutoff to take forward in the analysis, we computed the inflection points of the curves. We smoothed the curves, computed second order differences, and then took the point where the second order difference changed sign as the inflection. We took the least stringent p-value cutoff (maximum inflection point) among all annotations as the heuristic cutoff.

Although the enrichment curves computed above can give the relative importance of each annotation, they do not allow us to make a decision as to whether or not the annotation is relevant for the disease. In order to make this decision, we developed a new statistical test which uses the heuristic p-value cutoff described above. As before, to completely define this test:

1. The null hypothesis $H_0 : p_1 = p_0$ is that the proportion of causal variants within the annotation p_1 is equal to the proportion outside the annotation p_0, and the alternate hypothesis $H_1 : p_1 > p_0$ is that causal variants are enriched within the annotation.

2. The test statistic is the number of observed GWAS associations with p-value less than the heuristic cutoff

3. The null distribution of this statistic is not a standard distribution, and requires the non-parametric bootstrap to represent (described below)

4. The significance level for the test is 0.05, but must be adjusted for multiple testing (as described above)

5. The rejection region of the test is easy to compute given the bootstrap distribution described below.

The fundamental challenge which we address in the development of our test is estimating the distribution of the test statistic under H_0. The key idea of our approach is the bootstrap[40]: if we can sample data sets from the generative process assumed by H_0, and we compute the same test statistic on those sampled data sets, then the observed distribution of the computed test statistics is an approximation of the true null distribution of the statistic. These are "bootstrapped" statistics (in the sense of "pulling oneself up by one's bootstraps") because traditionally we use the data to estimate the parameters of H_0, then use the estimated parameters to generate new data sets.

Here, our specific choice of null hypothesis complicates this procedure because of the difficulty of generating new data sets under H_0. To illustrate this difficulty, first consider what it means to actually generate a new data set. The bootstrap procedure prescribes that:

1. We should generate a new vector of causal effects θ under H_0. To generate this vector, we have to randomly sample which variants are causal, and randomly sample their effect size (consistent with the unknown distribution of effect sizes for the disease of interest)

2. We should generate a new genotype matrix X and new phenotypes y for both cases and controls, as determined by the causal effects θ

3. We should recompute the GWAS summary statistics for each bootstrap data set

We describe how to actually perform these generative tasks in Chapter 4, where we directly model the generative process and simulate data from the model to test the calibration of our inference algorithm. However, here it suffices to note that there are a number of fundamental difficulties:

1. Recall that we will be testing for enrichment of m annotations, necessitating multiple testing correction using the BH procedure, and recall that the BH procedure picks hypotheses j such

that $p_j < \alpha * j/m$. This means that the statistical power of the overall procedure is limited by the minimum attainable p-value (equivalently, the size of the rejection region) by any one test. If we bootstrap B datasets, the minimum attainable p-value is $1/B$. As we describe below, our multiple testing burden implies that we will require order 10^5 bootstrap data sets.

2. Actually generating the vector of causal effects θ requires estimating unknown parameters of disease architecture, which is well-studied in the field of human genetics (see Chapter 4)

3. Actually generating new genotype matrices X, conditioned on phenotype, requires a non-trivial algorithm to incorporate the patterns of LD between nearby genetic variants. In principle, generating whole genomes is embarrassingly parallel (we can generate small windows of the genome in parallel); however, in practice using this generative process to bootstrap is computationally infeasible (see Chapter 4).

4. Actually computing GWAS summary statistics is embarrassingly parallel but standard tools require additional engineering and large computational resources to scale from one data set to many bootstrap data sets.

Instead of performing the parametric bootstrap described above, where we estimate the parameters of H_0 from the data and use the estimated parameters to generate new data sets, we perform a non-parametric bootstrap, where we resample from the data itself in order to generate new data sets from H_0. In particular, rather than estimating the proportions of causal variants required to define the null and alternate hypotheses, we will resample from the data itself to generate new datasets with those same parameters. The intuition behind our approach is that under H_0, the annotation has no impact on the distribution of causal variants in the genome, and therefore has no impact on the distribution of p-values partitioned by the annotation (due to correlation of nearby GWAS p-values with the effect size/p-value of the causal variant). However, we know that even without considering annotations, the distribution of GWAS p-values is biased:

1. Genetic variants which are correlated to more variants around them have smaller p-values[41]

2. Genetic variants which are closer to genes have smaller p-values. Intuitively, evolutionary theory predicts that genic variants will have larger effect sizes (since important genes are less tolerant of mutation than intergenic regions of the genome), and the p-value is a function of the effect size.

3. Genetic variants with higher MAF have smaller p-values, because we can observe more individuals with the alternative allele in the population and have increased statistical power to detect subtle differences between case MAF and control MAF.

Therefore, our bootstrap procedure has to preserve these properties of the distribution, while breaking the association between the annotation of interest and the p-value. This procedure has been previously proposed as Variant Set Enrichment, but only controlling for the number of correlated variants[42]. Briefly, our approach is to sample a bootstrap set of genetic variants from outside the tested annotation which matches the properties of the observed set of genetic variants which went into the observed test statistic, and then compute the same test statistic on this bootstrapped set of genetic variants. We resampled variants with replacement (to reduce memory usage) from matched on number of LD partners ($r^2 > 0.1$), minor allele frequency (in bins of width 0.03), and distance to nearest transcription start site (rounded to the nearest kilobase). We then computed p-values by counting the number of bootstrap data sets where the bootstrap test statistic was at least as large as the observed test statistic.

In order to prove that this is a valid hypothesis test, we need to prove that it actually controls the Type 1 error rate at the desired level. In particular, our procedure uses the data twice, once to select loci and once to compute enrichment of those loci. In order to prove this procedure does not lead to an anti-conservative test (i.e., $p < 0.05$ does not imply $P(\text{statisticatleastaslargeasobserved} \mid H_0) < 0.05$), we need to really generate data from H_0, apply our statistical test to the generated data, and then count how many Type 1 errors our method made. We give the full details of the generative model in Chapter 4, but give a brief overview here.

We simulated realistic phenotypes under the null of no enrichment, where variants are sampled uniformly at random independent of regulatory annotation. We used imputed dosages (expected number of minor alleles) in 16,180 individuals from the Wellcome Trust Case Control Consortium[43] and simulated quantitative phenotypes with parameters matching those inferred for rheumatoid arthritis by a Bayesian model[44]. We sampled at most one causal variant uniformly at random from non-overlapping 1MB windows such that causal variants were approximately pairwise independent and the expected number of causal variants was 2,231. We sampled causal effect sizes from a standard Gaussian and generated genetic values from an additive linear model. We then added residual noise to achieve proportion of variance explained (PVE) equal to 0.18 by sampling from a Gaussian with variance $(1/\text{PVE} - 1)V[X\beta]$, where $V[\cdot]$ denotes the sample variance.

2.3 Results

We investigated weak genetic associations for eight well-studied common diseases (having $p < 1.6 \times 10^{-3}$ on average) spanning a variety of etiologies, pathologies, and genetic architectures for which summary statistics are publicly available (Table 2.1): Alzheimer's disease (AD), bipolar disorder (BIP), coronary artery disease (CAD), Crohn's disease (CD), rheumatoid arthritis (RA), schizophrenia (SCZ), Type 1 Diabetes (T1D), and Type 2 Diabetes (T2D).

Here, our terminology needs some clarification: both genome-wide significant associations and those associations which are not yet significantly associated have weak effect sizes, as empirically observed and predicted by evolutionary and population genetic theory. Intuitively, if a genetic variant is common enough in the population to observe at high MAF, then it cannot have a strong deleterious impact on fitness (otherwise, individuals who carried it would not survive to be observed in the population).

Moreover, defining weak association based on p-value induces a dependency on sample size, such that the ranking of variants by p-value involves more than just their MAFs and their effect sizes. Again, empirically we observe that as GWAS sample sizes increase, we find new genetic loci with significant p-values after Bonferroni correction. However, one of the major goals of this thesis, and the prior work it builds upon, is to find new genes which could eventually be discovered by direct genetic evidence of association if the sample size were large enough, but using the sample sizes available today. Note that although the largest GWAS studies to date have surpassed hundreds of thousands of individuals, only a handful of traits (e.g., height[7] and schizophrenia[8]) have been studied at this scale. Computationally, this problem is known as reprioritizing genetic loci, and has been previously studied[45]. The main methodological advance of the method presented in this chapter which allows it to achieve this goal is to select a subset of loci based on significance to dissect further.

We focus on distal enhancer regions because these play a role in transcriptional regulation and are

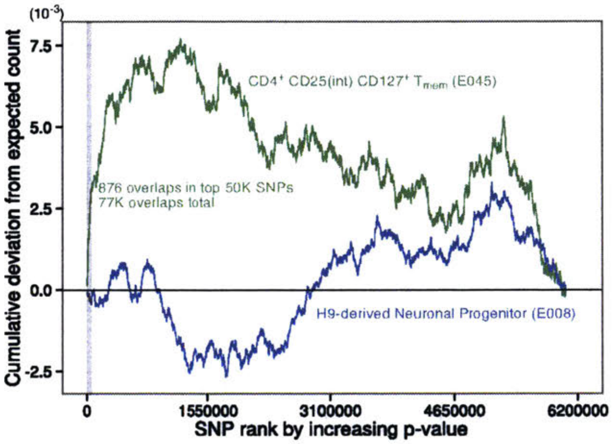


Figure 2.3: Cumulative enrichment curves through the entire list of GWAS p-values for Type 1 Diabetes, considering two different annotations: CD4 T cells (green), and neuronal progenitor cells (blue). The key features of the curves are: (1) relevant annotations are encountered more frequently than irrelevant annotations at the head of the ranked list (the enrichment curve sharply rises), and (2) irrelevant annotations tend to remain around the level of zero enrichment (deviation from the expected count).

Trait	Citation	Cases	Controls
Alzheimer's disease	Lambert et al., Nat Genet 2013[46]	17,008	37,154
Bipolar disorder	PGC Bipolar Disorder Working Group, Nat Genet 2011[47]	7,841	9,250
Coronary artery disease	Schunkert et al., Nat Genet 2011[48]	22,233	64,762
Crohn's disease	Franke et al., Nat Genet 2010[49]	6,333	15,056
Rheumatoid arthritis	Stahl et al., Nat Genet 2010[50]	5,539	20,169
Schizophrenia	Ripke et al., Nat Genet 2013[51]	13,833	18,310
Type 1 Diabetes	Bradfield et al., PLoS Genet 2011[52]	9,934	16,956
Type 2 Diabetes	Morris et al., Nat Genet 2012[53]	12,171	56,862

Table 2.1: References for genome-wide association meta-analyses used in this study.

also dynamic across different cell types, allowing us to propose causal cell types and tissue-specific biological functions which are disrupted. To define putative enhancer regions, we used a 15 chromatin state model[29] summarizing five chromatin marks across 127 reference epigenomes spanning diverse primary cells and tissues from the Roadmap Epigenomics[13] and ENCODE[12] projects (Figure 2.4) and took the union of enhancer-like states.

We removed variants within the Major Histocompatibility Complex (MHC; positions 29.4 – 33 MB of chromosome 6) plus 5 megabases flanking from all analyses. The MHC is a region of the genome important for immune function, encoding the Human Leukocyte Antigens which are necessary for the immune system to distinguish its own cells from pathogens. Although the MHC is known to play significant roles in autoimmune disorders such as T1D, the causal variants in this region are known to be protein-altering variants, leading to auto-immune targeting of other tissues (in the case of T1D, the pancreatic islet cells which produce insulin). In this study, we instead focus on identifying non-coding variation which impacts transcriptional regulation, and therefore do not lose too much power by simply excluding the MHC. Moreover, the MHC region displays unusual long range LD which inflates GWAS test statistics in the flanking regions and would confound our enrichments.

In order to improve our power to detect enrichments, we imputed summary statistics for all studies into the Thousand Genomes reference cohort (if necessary) using ImpG-Summary[54]. The underlying principle of this thesis is that using the most descriptive and fine-grained annotation of the non-coding genome will allow us to make the most specific predictions of biological mechanisms. In order to exploit the highest resolution annotations (described below), we need to impute GWAS data to the most comprehensive available catalog of genetic variation; otherwise, we will simply fail to observe overlaps with the high resolution annotations of interest. The ImpG-Summary model is based on modeling the distribution of GWAS z-scores as a multivariate Gaussian and using standard identities to derive the conditional distribution of the unobserved z-scores, given the observed z-scores.

We visualized enrichment of regulatory annotations using an approach inspired by Gene Set Enrichment Analysis[38]. Briefly, we seek to track the cumulative enrichment in each annotation of interest as we consider increasing numbers of GWAS associations (with increasing p-value). Our visualization allows us to order the relative importance of annotations based on the ordering of the curves and allows us to determine a heuristic p-value cutoff based on the inflection points of the curves. To account for LD between associated variants, we pruned each set of summary statistics to a set of independent loci (pairwise $r^2 < 0.1$), yielding an average of 228,291 loci per disease. We aggregated annotations over all variants within each locus, and found enrichments for relevant cell types which persist even when considering thousands of weak associations (average 1,650 independent loci), equal on average to a p-value cutoff of $p < 1.6 \times 10^{-3}$ (**Fig. 2.5**).

In autoimmune disorders (CD, RA, T1D), we found enhancers active in T cell types showed the strongest enrichment for weak associations. These enrichments are expected given the known role of immune cells in these disorders.

In psychiatric disorders (AD, BIP, SCZ), we also found enrichment of immune cell types, supporting the role of immune pathways in these disorders[55–57]. Interestingly, we found enrichments for B cell enhancers rather than T cell enhancers in AD. In BIP and SCZ, we additionally found enrichment for enhancers in a number of adult brain tissues.

In CAD, we found enrichments in colonic mucosa, which could indicate a role of the intestine and gut microbiome in risk for CAD[58]; however, this finding is speculative and is not further sup-

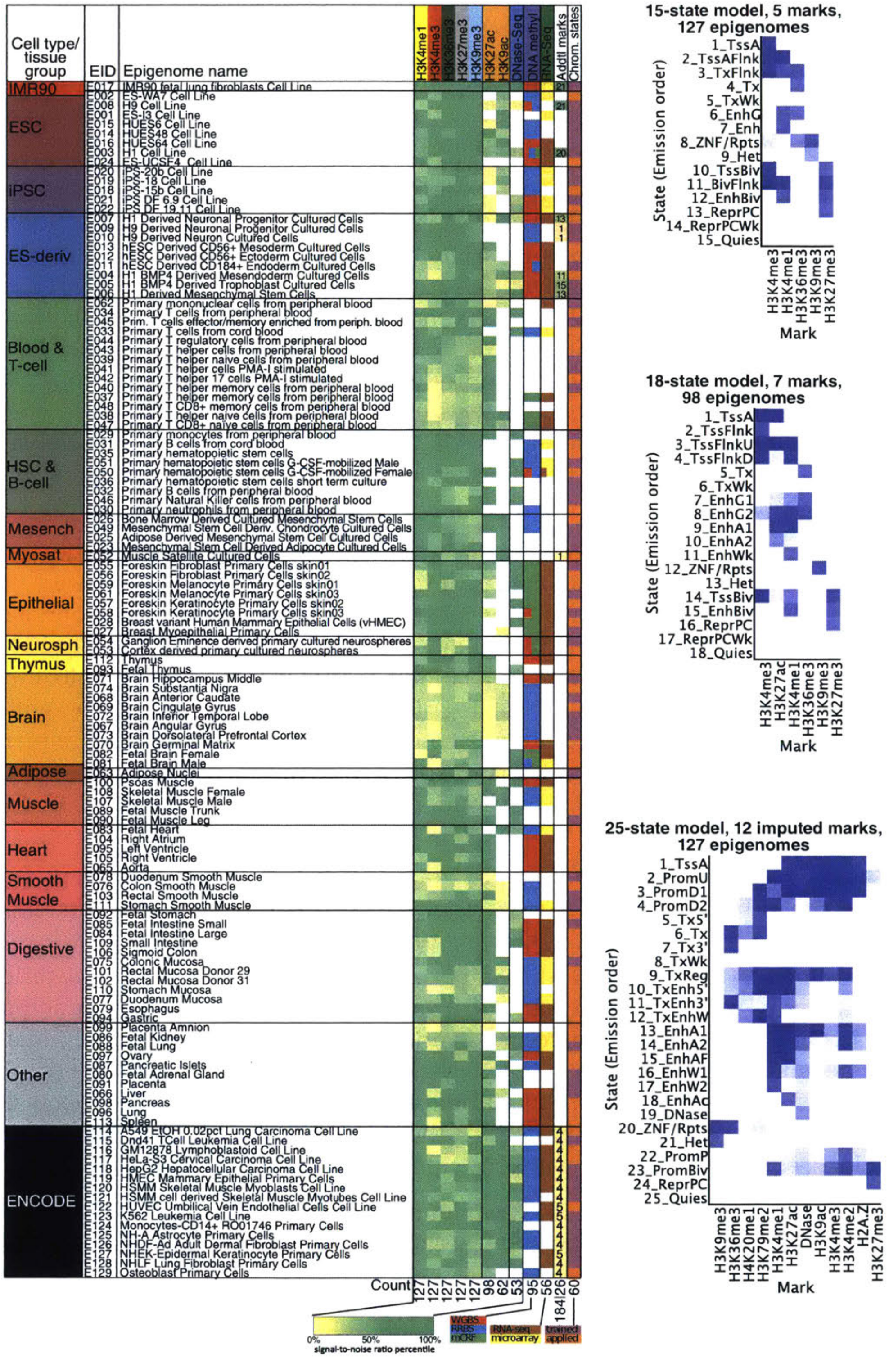


Figure 2.4: Unique identifiers, cell type names, and tissue groups for 127 reference epigenomes. Reproduced from Roadmap Epigenomics Consortium, Nature, 2015[13]. ChromHMM state definitions for the 15-state model trained on observed data and the 25-state model trained on imputed data.

ported by our pathway or motif analysis of the specific enhancer regions (Chapter 3). In particular, we find relevant pathways and motifs enriched primarily in constitutive enhancer rather than mucosa-specific enhancers, as we show below. Surprisingly, we did not find enrichments for aorta enhancers in CAD despite vascaular tissue being relevant a priori.

In T2D, we found enrichments in pancreatic islets, consistent with prior work[59], but additionally in small intestine, consistent with the role of gastrointestinal mucosa in glucose homeostasis[60].

To justify our heuristic p-value, we investigated the false discovery rate (FDR) of genetic associations at the chosen cutoff in each of the eight diseases. One fundamental problem in applying multiple testing procedures to GWAS data is that the collection of test statistics are not mutually independent due to LD. Indeed, the same correlation between z-scores which makes summary statistic imputation possible (as described above) implies that the number of hypotheses we need to correct for is actually less than the number of hypotheses tested. A recent theoretical result proves that the same procedure we used to prune the GWAS data into a list of independent loci can be used to generate a list of independent hypotheses, and that applying the BH procedure to these hypotheses indeed controls the FDR at the desired level[61]. The intuition behind this result is that we need to redefine what constitutes a "discovery" for GWAS. Operationally, having pruned the loci and picked a representative p-value for each locus, we would consider a particular locus to be true positive if we rejected the null hypothesis for the representative of that locus, and there was a true causal variant somewhere in that locus (correlated to the representative). Now, the importance of the result is that if we apply the BH procedure to the set of representatives, we will control the false discovery rate of loci in exactly this sense.

Now, in order to estimate the FDR for a particular set of rejected hypotheses, we cannot use the BH procedure, which produces a set of rejected hypotheses at a desired FDR. Instead, we need to estimate the q-values for the pruned loci: the posterior probability that the z-score for each locus came from the alternate hypothesis. The intuition behind this approach is that having estimated q-values, taking hypotheses with $q < q^*$ controls the FDR at level q^*. Therefore, we just need to estimate the q-value for the locus with p-value matching the cutoff (since we selected the cutoff as the p-value of some locus).

In general, the FDR will depend on a number of parameters, including the study size, heritability (proportion of phenotypic variance explained by genotypes in a linear model), effect size distribution of causal variants, and minor allele frequency distribution of causal variants. Indeed, for the eight diseases we studied, we found the FDR was between 1.5-18% (Table 2.2). However, the fact that the maximum estimated FDR was only 18% at our heuristic p-value cutoff motivates our study of those genetic associations which do not meet genome-wide significance at current sample sizes.

We evaluated the statistical significance of enrichments using a permutation test. Briefly, for each disease and enhancer annotation, we compared the count of associations passing our heuristic p-value cutoff within the annotation against the null distribution of counts of resampled SNPs passing the same cutoff outside the annotation. We resampled SNPs matched on number of LD partners, minor allele frequency, and distance to closest transcription start site. For each phenotype, we used all well-imputed SNPs (mean 7,797,600) to avoid small number effects. We found the enrichments reported above were all statistically significant (permutation test, BH FDR < 0.017, Figure 2.6); however, essentially all cell types showed significant enrichment in all diseases (data not shown), attributable to confounding of constitutive and tissue-specific enhancers as we show below.

Assigning a p-value to the test requires some consideration. For each annotation we tested for

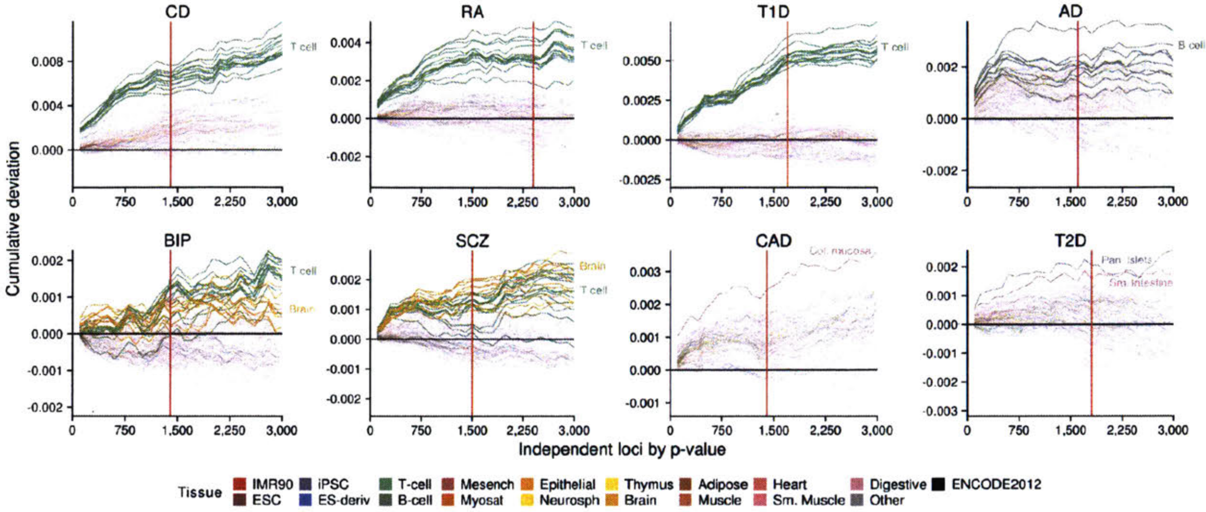


Figure 2.5: Enrichment of independent loci (pairwise $r^2 < 0.1$) across eight diseases in enhancer regions predicted by a 15 chromatin state model learned on observed data for 5 histone modifications across 111 reference epigenomes. Each curve corresponds to enhancer regions predicted in a specific reference epigenome and is colored by tissue group. The black line at zero cumulative deviation indicates no enrichment, and the red vertical line indicates the heuristic p-value cutoff taken forward for the rest of the analysis. Opaque lines denote enrichments highlighted in the results.

Trait	p-value	q-value
AD	0.0023	0.087
BIP	0.0020	0.10
CAD	0.0015	0.088
CD	0.0011	0.14
RA	0.0030	0.11
SCZ	0.00016	0.015
T1D	0.00041	0.018
T2D	0.0025	0.16

Table 2.2: Estimated false discovery rate of independent loci (equivalent to the q-value) at the chosen empirical p-value threshold in eight phenotypes.

enrichment, we additionally tested the estimated null distribution for departure from the Gaussian distribution using the Anderson-Darling test. Intuitively, the test computes the distance between the empirical cumulative density function of the observations (here, the bootstrap null statistics) and the cumulative density function of the target distribution (here, Gaussian. We found that in general the null distribution for all of the annotations we considered was non-Gaussian (Anderson-Darling statistic $> 0.787, p < 0.05$), so we must use empirical p-values counting the number of trials in which the observed statistic was exceeded by a null statistic rather analytical p-values based on the mean and variance of the null distribution. This fact limits the power of the test (essentially, the minimum attainable p-value) to the number of permutations chosen, which is problematic when combined with the fact that correcting for multiple testing of many annotations requires adjusting the significance level downwards (even when controlling the FDR rather than the FWER).

Our multiple testing correction procedure must be stringent enough to account for the fact that we will use enhancer enrichments to pre-screen pathway and motif hypotheses to test (Chapter 3). We made two critical choices:

1. We controlled the FDR of enhancer enrichments for each of the eight diseases separately rather than analyzing all hypotheses together.
2. We controlled the FDR such that the overall false discovery rate for all rejected hypotheses (including gene pathways and regulatory motifs) is 0.05.

To justify the first choice, we note that the key parameter underlying FDR estimation is the proportion of null hypotheses among the set of hypotheses considered, and that for enhancer enrichments this parameter can differ between diseases. Intuitively, in order to estimate the posterior probability that a particular test statistic came from the alternate hypothesis, we need to estimate the prior probability of null hypotheses. The established methodology to estimate this prior probability is an empirical Bayes strategy[62], where the data are used to estimate the required prior, and then combined with that prior to compute the posterior. The prior proportion of null hypotheses could be different per disease because some diseases might be driven by only one cell type, while others may be driven by a combination of many diverse cell types. In this case, jointly analyzing all hypotheses for all diseases considered jointly may lead both to overly conservative conclusions for some diseases and overly liberal conclusions for others.

The key theoretical result which justifies our separate analysis of each disease is that controlling the FDR for each subset of hypotheses (for example, by applying the BH procedure) separately does indeed control the FDR over all of the hypotheses at the same rate[63]. This is opposed to e.g. applying Bonferroni correction for each subset of hypotheses, which does not control the overall FWER. To see this fact, recall that to control the FWER at level 0.05 for a collection of n hypotheses, we need to reject hypotheses with $p < .05/n$. Now, if we partition the hypotheses into two subsets of size $n/2$ and apply Bonferroni correction to each partition by rejecting hypotheses with $p < .05/(n/2)$, the overall FWER will be .1.

Intuitively, the reason this work for the FDR but not the FWER is that the FDR really is a rate which scales correctly for subsets of n hypotheses. The simplest proof sketch is as follows: suppose we have a collection of tests (X, I, Z) where x denotes the subset the test belongs to, i represents whether the test is truly non-null, and z represents the z-score. Then, by definition, $FDR(x, z) = P(i = 1 \mid X = x, Z \geq z)$. Now, suppose that for each possible X, we have some rule $R(X)$ which assigns the test a value $\hat{I}$ denoting whether it is non-null while controlling the FDR at level q. Such a rule compares Z to some threshold value $z(X)$. One such rule is the BH procedure (as described above). Then, by construction $FDR(R, X) = P(\hat{I} = 1 \mid I = 0, X = x) = q$. The key point is that integrating over all possible X (to remove the conditioning):

$$\begin{aligned} FDR(R) &= P(\hat{I} = 1 \mid I = 0) \\ &= \int_X P(\hat{I} = 1 \mid I = 0, X = x, Z \geq z(X)) P(X = x \mid Z \geq z) dx \\ &= \int_X qp(x = x \mid Z \geq z) \\ &= q \end{aligned}$$

To justify the second choice, we note that our hypotheses are arranged as a two-level hierarchy in which a hypothesis at level 2 (pathways/motifs) is tested only if the corresponding hypothesis at level 1 (enhancers) was rejected (refer to Chapter 3). In our preliminary work, we relied on a theoretical result that applying the BH procedure at each level of the hierarchy controls the FDR of the entire hierarchy of hypotheses, but with multiplicative penalty 2.88 to the overall FDR[64]. Intuitively, we have to account for the fact that we might choose to test second level hypotheses based on a first level of hypothesis which was erroneously rejected. Therefore, in the first part of this thesis (Chapters 2-3), we apply the BH procedure with $q = .05/2.28 = 0.017$ at each level such that the overall FDR is 0.05.

One further complication is that our heuristic p-value threshold which was used to define the test statistic was based on the same annotations which we are testing. To account for potential dependencies between the annotations and the threshold used in the permutation test, we verified that the FDR was well-calibrated by simulating realistic phenotypes and repeating the procedure described above. Briefly, we used 16,180 samples from the Wellcome Trust Case Control Consortium[43] to generate phenotypes matching the genetic architecture of rheumatoid arthritis[44]. Over 10 trials, the FDR was well-calibrated, supporting the validity of our testing procedure.

Another potential source of bias is our use of summary statistic imputation. The key statistical parameter required to perform imputation of summary statistics is the $p \times p$ correlation matrix $R = X'X/n$. Recall that the entire reason we rely on summary statistics is that we do not have access to X; therefore, we need to estimate $\hat{R}$ from some other $\tilde{X}$, which typically comes from a reference cohort of individuals such as the Thousand Genomes cohort. It is known that when we impute summary statistics using such out-of-sample LD information, the posterior means of the unobserved z-scores are biased towards zero.

To investigate the impact of deflation of summary statistics imputed using out-of-sample reference LD information on the downstream enrichment results, we held out variants not present in the Hapmap 3 reference panel (comprising 2.4 million variants) and re-imputed summary statistics for the held out variants using ImpG-Summary. We used published summary statistics for schizophrenia[8], for which genotypes were imputed into 8,280,096 variants in the Thousand Genomes reference panel before computing the published marginal association summary statistics. We successfully re-imputed from Hapmap 3 into 8,387,080 variants from the Thousand Genomes panel using ImpG-Summary. We verified that the Pearson correlation was 0.93 between summary statistics derived from genotype-level and summary-level imputation and additionally estimated that the average level of deflation over all variants was 7%. We then repeated the enrichment analysis described above, and found the results were essentially unchanged (data not shown).

The key insight from our initial enrichment analysis is that sharing of elements across tissues confounds the enrichments. We sought to distinguish regions which exhibit enhancer-associated chromatin marks constitutively (in all cell types) from those which are marked in specific tissues. In prior work, stratified LD score regression[65] identified relevant tissue-specific annotations by including both non-specific and tissue-specific annotations in the model, finding that the more complex model including relevant tissues fit the data better through a likelihood ratio test.

Here, we instead used 226 enhancer modules defined as previously described[13] to delineate a biologically meaningful set of disjoint annotations. Briefly, putative enhancers across reference epigenomes are defined as DHSs (in any reference epigenome) labeled by enhancer-like chromatin states in each reference epigenome. Enhancer modules are then defined as k-means clusters of these regions based on their activity profiles (presence/absence) across the reference epigenomes.

We computed enrichments for these enhancer modules and found that constitutive enhancers are significantly enriched for weak association across all eight diseases (permutation test, BH FDR $<$ 0.017, Fig. 2.7). We note that these annotations cover such a small proportion of the genome that we could not use our visualization method to choose a heuristic p-value cutoff specific to these annotations, and instead used the cutoffs described above.

After partitioning regulatory regions into constitutive and tissue-specific modules, we recover much fewer significant tissue-specific annotations. Our enrichments are less noisy not only because we correct for the contribution of constitutive enhancers to all single cell type annotations, but also because we use narrower, higher confidence regions by combining chromatin accessibility and histone modification data. We found that immune-specific enhancers are enriched in both autoimmune (CD, RA, T1D) and psychiatric disorders (AD, BIP, SCZ) and that brain-specific enhancers are enriched in psychiatric disorders. We found that mesenchymal stem cell-specific enhancers are enriched in metabolic disorders (CAD, T2D), but that these enhancers are also predicted to active in relevant adult tissues such as heart and digestive tissues.

2.4 Discussion

In this chapter, we developed methods to study the role of non-coding variants in complex traits by computing enrichments of weak associations (not meeting genome-wide significance) in functional annotations, identifying and correcting for a number of confounders. Across eight complex diseases, we identified relevant regulatory annotations and a specific set of annotated regions to take forward in our analysis (Chapter 3).

Our methodology and results highlight an important distinction in the use of reference epigenomes as proxies for transcriptional regulatory elements to identify and re-prioritize weak associations. Specifically, regulatory annotations predicted on individual reference epigenomes confound constitutive and tissue-specific marking (and activity) of regulatory regions. We showed here that k-means clustering of regulatory regions could deconvolve patterns of histone modification across 111 cell types and tissues.

Our proposed enhancer enrichment method is one of a number of methods which have been proposed for identifying relevant annotations using GWAS data. These methods can be broadly grouped in three classes:

1. Testing for over-representation based on counting overlaps[10,31,66]
2. Regression with enrichment hyperparameters[45,67,68]

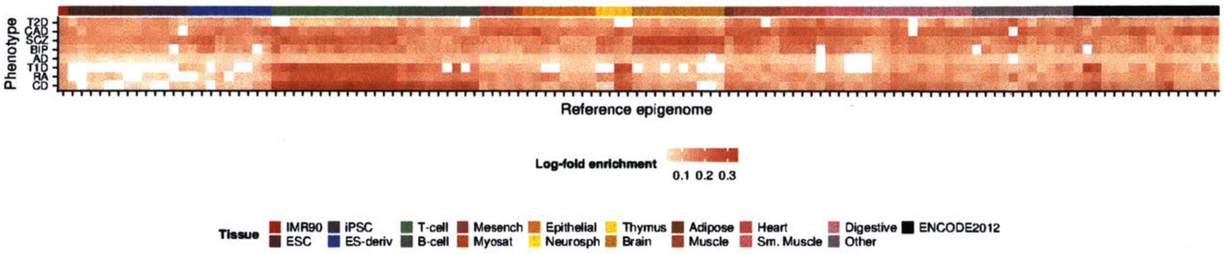


Figure 2.6: Enrichment of enhancers across 127 reference epigenomes in eight diseases. Log-fold enrichment values are shown only if significant (permutation test, BH FDR = 0.017) are shown. In contrast to enhancer modules (Figure 2), enrichment methods for annotations learned on individual cell types count constitutive elements towards every annotation, confounding the enrichments.

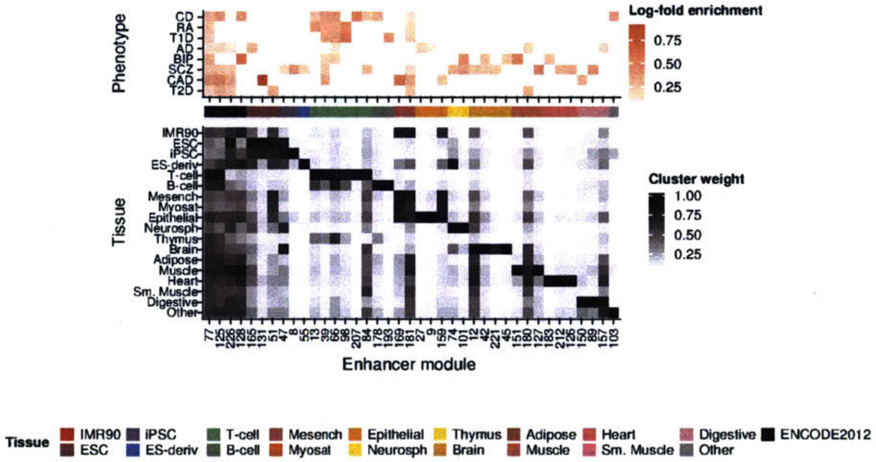


Figure 2.7: Enrichment of weak associations meeting the heuristic p-value threshold in enhancer modules. Log-fold enrichment for 226 enriched enhancer modules corresponding to observed histone modification patterns across 111 reference epigenomes. Only 69 significantly enriched modules (BH FDR < 0.017) are shown. Modules are defined by clustering DHSs labeled as enhancer-like by a 15 chromatin state model learned on observed data for 5 histone modifications across 111 reference epigenomes. Each module is represented by a vector of weights per reference epigenome (proportion of DHSs annotated as enhancer in that reference epigenome). For display, weights are collapsed by tissue group by taking the maximum weight over all reference epigenomes in each tissue group. Modules are ordered by the tissue group with maximum weight. The leftmost four modules are defined as constitutive (having at least 50% of cluster weights greater than 0.25).

3. Partitioned heritability[65,69]

Our method makes a number of advances over existing methods which are based on counting overlaps. The two main advances are: (1) we consider genetic variants which do not achieve genome-wide significance and (2) we prove that our resampling procedure correctly accounts for confounders through simulation.

The approach we developed here is complementary to regression approaches which explicitly model enrichment (which we explore in Chapter 4). As described above, in the framework of regression the method presented in this chapter is a univariate feature selection based tests of significance for each feature. However, here we frame the fundamental inference problem as hypothesis testing rather than parameter estimation. The key difference is that in the approach proposed in this chapter, we do not assume a parametric model which generates the observations; in contrast, the approach we will take in Chapter 4 will indeed try to estimate the parameters from such a problem.

The approach we developed here is also complementary to partitioned heritability approaches, which conceptually seek to estimate the variance explained by annotated variants in a linear model. Of note, one of these methods, stratified LD score regression[65] (LDSC), has been applied to many of the same GWAS we study here. However, we note two important differences in our methodology and results. First, and most importantly, here we are jointly performing feature selection on the genetic variants (through the heuristic p-value cutoff depending on the relevant annotations) as well as the annotations themselves. The fundamental assumption of heritability estimation methods is that all genetic variants explain equal proportion of phenotypic variance in expectation (see Chapter 4), which means that although these methods can identify relevant annotations, they cannot directly identify the relevant regions with those annotations.

Second, we found that enhancer regions predicted by ChromHMM in nearly any single cell type are enriched for each phenotype and that we had to construct a set of annotations summarizing activity across multiple tissues in order to find relevant enrichments. Examining the fitted model (Figure 2.4), ChromHMM learns an enhancer annotation entirely defined by H3K4me1. However, LDSC applied to H3K4me1 peaks measured in individual cell types found only relevant cell types were enriched in these same phenotypes.

This difference is explained by the fact that our statistical test compares the observed number of associated variants overlapping the annotation of interest to the null distribution of counts, where LDSC performs a model comparison between a null model with no tissue-specific features and an alternate model including tissue-specific features. In our methodology, we account for overlap between observed histone modifications by learning a latent representation of the data which distinguishes constitutive from tissue-specific elements by clustering observations across multiple cell types, finding separate enrichments for both constitutive and tissue-specific annotations. In contrast, LDSC performs a model comparison between a null model containing only broad annotations, and an alternate model in which a tissue-specific annotation is added, finding broad enrichments by examining the fitted null model and tissue-specific enrichments through the model comparison. Therefore, our results are still concordant with prior work; however, the constructed enhancer modules we focus on in this study allow us to directly study specific regions predicted to be enhancers with high confidence (Chapter 3).

Our methodology has a number of important limitations with regards to selecting relevant features. As described above, the significance level of a particular genetic variant depends not only on its effect size and minor allele frequency, but also on the sample size of the study. Selecting associations based on significance is common practice in deriving polygenic risk scores which are predictive of phenotype. Historically, very loose thresholds involving many thousands of variants were required to achieve good prediction performance[44]; however, as sample sizes have increased the threshold is becoming more stringent, and approaching genome-wide significance. We observed this trend in recent application of our method to menarche[70], where the heuristic p-value threshold selected by our method was actually more stringent than 5×10^{-8}.

Our method is unbiased in the sense that we consider all annotations rather than restricting to some set of cell types thought to be relevant a priori. However, the panel of 127 reference epigenomes we used is itself biased in representation of tissues, leading to several issues. First, we found unexpected enrichments for intestinal mucosa cell types across a number of the diseases, which will require additional epigenomic profiles to explain. Second, our definition of a constitutively marked enhancer depends on the proportion of reference epigenomes which the enhancer is annotated by an associated chromatin state. Blood cell types make up a large proportion of reference epigenomes considered here, and therefore putative constitutive regions might not actually be constitutive (leaving aside the distinction between enhancer marks and enhancer activity). Third, enhancer modules in lineages other than blood are smaller (cover less of the genome) than either constitutive or blood-specific modules, making it more difficult to find significant enrichments for these annotations.

Chapter 3

Functional characterization of enhancers

3.1 Background

In Chapter 2, we used genetic associations from GWAS to identify relevant classes of regions marked by epigenomic annotations, and we used a heuristic to select a number of specific regions in those classes. It would seem then that we have achieved the main computational goal of this thesis. However, recall that the main biological goal of this thesis was to interpret the role of non-coding genetic variation in human disease. This will require additional development on the computational front. In this chapter, we seek to make progress on this goal, but we first need to explain what exactly interpreting disease-associated genetic variation means.

In the case of a protein-altering variant, the interpretation is clear. Consider the gene cystic fibrosis transmembrane conductance regulator (CFTR). We know that the majority of cystic fibrosis cases have a deletion of three nucleotides in this gene which results in a protein lacking the amino acid phenylalanine in position 508 (of 1,488 amino acids). In these cases, the deletion of the amino acid prevents the protein from folding into the correct structure and destroys its ability to function. Moreover, by examining the structure of CFTR and comparing against other proteins with known functions, researchers inferred it was probably an ion channel, a protein residing in the cell membrane responsible for maintaining concentration gradients (differences) for important molecules between the cell and its environment.

For a non-coding genetic variant, we hypothesize that it changes the transcriptional regulation of a target gene: rather than completely disrupting the function of the target, it modulates the function by changing the level at which the gene is transcribed into mRNA. These changes in transcription result in downstream changes in protein abundance, which result in changes in cell function, which result in changes to tissue, organ, and organ system function, which ultimately result in a organismal-level phenotype such as disease.

Conceptually, in order to interpret non-coding genetic variation we need to expand our definition of "gene". Historically, we have thought of genes as simply the sequence which codes for proteins. However, the study of transcriptional regulation has revealed that surrounding each gene is a constellation of regulatory sequence elements, located both proximal (near) and distal (far; on the order 10^3-10^6 bases away) to that gene (see Figure 2.2).

Moreover, comprehensive profiling of gene expression across different cell types, cell states, and other experimental conditions has revealed that genes co-regulate each other. For example, the protein coded for by one gene might affect the transcription of another gene, through mechanisms like the ones we discuss below. These kinds of relationships imply a complicated transcriptional regulatory network connecting genes and regulatory DNA elements. In particular, in this chapter we will investigate the biological pathways which are disrupted by disease-associated variants. Conceptually, a pathway is a set of genes which are connected by regulatory mechanisms, again bearing in mind that each gene includes its associated regulatory elements.

As we outlined in Chapter 2, a number of mechanisms can change the expression level of a gene. First, the organization and positioning of the DNA around the gene can change whether the protein-coding sequence is accessible to the enzymes required for transcription (such as RNA polymerase). So called chromatin accessibility is known to be highly variable between different human cell types and tissues at genes for specific cell functions, and especially at the regulatory regions important to controlling those genes. Indeed, we used chromatin accessibility as an additional epigenomic modification and an additional line of evidence supporting the putative function of predicted regulatory regions.

More broadly, we know that the epigenomic landscape (including histone modifications, and therefore chromatin states, as outlined in Chapter 2) are highly variable across human cell types and tissues. We hypothesize, and have some evidence based on existing data, that the epigenomic landscape also varies between individuals[71] and through the course of human development and cell differentiation. We additionally know of proteins which are responsible for not only reading but also writing histone modifications, and that writing these modifications can alter the function of the underlying sequence. For example, marking inactive enhancers (which we relied upon in Chapter 2 and describe in more detail below) with H3K27ac has been shown to activate their function in transcriptional regulation[72]. Therefore, the entire epigenomic state of the gene and its ensemble of regulatory regions is an important determinant of the resulting expression level of that gene.

Second, even if the gene were in the correct epigenomic state, the additional transcription factors (TFs) required to bind to the promoter to start transcription might not be bound. TFs are proteins which bind to the DNA through interaction between a groove in the folded protein and the outer backbone of the DNA molecule. These interactions are typically sequence-specific: the amino acids forming the groove present hydrogen atoms in different physical configurations, which can form hydrogen bonds with the outer backbone of the nucleotides making up the DNA molecule (which themselves also present hydrogen atoms in specific configuration). Typically, the sequences which TFs recognize are 8-16 basepairs long, although there are numerous examples of atypical TFs, such as compound TFs which recognize multiple binding sites offset by a certain distance.

Due to the biochemical nature of TF binding, there is some noise permitted in the sequences which they recognize. Essentially, which nucleotides appear in certain positions is more important than the nucleotides which appear in other positions. We can represent the collection of sequences which a TF recognizes as a position weight matrix (PWM). Each entry of the matrix gives the relative frequency of observing that nucleotide in that position of the binding site. We can compute the importance of each position j of the PWM towards determining whether the TF binds as:

$$2 - H(W_j)$$

where $H(j) = \sum_k W_{jk} \, log_2(W_{jk})$ is the entropy of the jth column of the PWM W. The entropy of the distribution of the jth column is maximized when all four entries have equal probability; intuitively, this is when the position doesn't matter for binding. Given a particular subsequence of the genome, we can compute the PWM score as the product of entries of the PWM corresponding to each base; intuitively, this is the probability that a TF binds to that particular subsequence. We typically classify sequences as to whether it is a motif match (i.e., a TF binds) or a mismatch by setting a threshold on the PWM score. We refer to the matrix as a motif and the particular subsequences of the genome which match it as motif instances. Motif instances are computational predictions of transcription factor binding sites, which can be directly observed using ChIP-Seq.

The simplest reason why a TF required for transcription might not be present is that the gene which codes for the upstream protein might itself not be expressed, potentially through some combination of these same mechanisms.

Second, a different, competing transcription factor might recognize a sequence motif present nearby and occupy the space needed for the required TF, preventing transcription. The possibility we focus on in this chapter is that a genetic variant changes the binding site which the TF recognizes, and therefore prevents the TF from binding. Along the same lines, multiple transcription factors might need to cooperatively bind to nearby DNA to begin transcription, and a genetic variant might disrupt just one of the required motifs.

Third, if the gene isn't accessible, certain proteins called pioneer factors might be required to make the relevant part of the chromatin structure accessible. These factors are often also transcription factors themselves, which are capable of displacing nucleosomes (which are used to pack chromatin) and then binding to their target promoter.

Fourth, the gene might require additional distant enhancers to also interact with the promoter through physical deformation and interaction of the DNA molecule. These enhancers might themselves not be accessible, or might harbor genetic variants disrupting the TF motif, leading to changes in downstream target gene expression.

Returning to the conceptual overview of this thesis, to this point we have only used epigenomic annotations in conjunction with disease associations, and we have only identified relevant cell types and tissues based on enrichment of associations within these annotations. However, the primary reason we are interested in epigenomic annotations of the genome is that they are correlated with the function of the underlying sequence. In Chapter 2 we relied upon epigenomic annotations which were associated with enhancer elements specifically because we sought to identify enhancer elements whose disruption could cause disease. In this chapter, we seek to explicitly characterize those candidate enhancer elements, which means:

1. Identify the putative regulatory regions
2. Identify the cellular context in which the regions are functional
3. Identify the causal nucleotides within the putative regulatory regions
4. Identify the target genes of the regulatory regions
5. Identify the upstream regulator which targets the regions
6. Show that altered expression of the gene changes downstream phenotypes

Ultimately, for a particular locus of interest we must prove any statement about these six aspects of its functional impact on disease through experimentation. In particular, (6) cannot be performed computationally. In this thesis, we pursue a more modest biological goal: for each locus of interest, generate the experiments which should be performed (Figure 3.1). Ideally, we want to computationally predict as much of the putative biological mechanism (as outlined above) as possible to design as specific a set of experiments as possible.

In Chapter 2, we described a method to identify more genetic loci of interest which do not have sufficient direct genetic evidence of association, but do have additional evidence in terms of epigenomic modification. We used that method (1) to identify a set of putative regulatory regions within those loci to take forward for further analysis, and (2) characterize the cell types in which those regulatory regions are active.

In this chapter, we take those regions we identified and focus on the problems of (3), (4), and (5). The key idea of this chapter is to jointly analyze many loci associated with a disease of interest to identify common regulatory mechanisms across distant genetic loci. Moreover, we jointly analyze many diseases to compare and contrast the regulatory mechanisms which drive them. The first biological question we need to answer is whether the predicted enhancer regions we identified actually are enhancers, which we approach by showing that they target relevant biological pathways and are regulated by relevant upstream TFs through enrichment analysis. We then need to take the enrichment results, go back to the individual regions, and predict a mechanism for each region by actually finding the putative binding site disrupted by the regulatory variant, and linking the enhancer region to its putative target gene.

3.2 Methods

3.2.1 Pathway enrichment

The first biological question we ask in this chapter is whether non-coding variants associated with disease (as identified in Chapter 2) recurrently disrupt enhancers which target specific biological pathways. We hypothesized that although we could identify hundreds of non-coding variants associated with each of the eight diseases we considered, it is unlikely that each disrupts some distinct biological process. Rather, we expect that regulatory variants will recurrently disrupt some smaller number of important biological processes, by targeting genes which may be spread all over the human genome, but which interact with each other through transcriptional regulation or other pathway mechanisms.

To answer this question, we used Genomic Regions Enrichment of Annotations Tool (GREAT) to test for enrichment of enhancer regions in gene pathways[73]. As before, we phrase this question as a hypothesis test. However, the test we rely upon is Fisher's exact test, which predates the later definitional work on statistical tests by Neyman and Pearson which we explicated in Chapter 2. In particular, this test predates the concept of the alternative hypothesis, which Neyman introduced to answer the question of how to pick the test statistic: pick the statistic which has highest probability of rejecting the null when the data comes from the alternative hypothesis, holding the error rate fixed when the data comes from the null hypothesis.

Conceptually, we consider the set of regions defined by a particular enhancer module (which we found to be enriched in Chapter 2), and consider two possible ways to partition this set of regions into two subsets. First, we use partition the regions according to whether one of the genetic variants we selected in Chapter 2 overlaps each region. Second, we partition the regions according to whether the region overlaps the regulatory domains of genes which have some coherent biological function.

1. The null hypothesis is that the two classifications (partitions) are independent
2. The test statistic is the 2 × 2 contingency table, which gives the counts of regions in each of the partitions (and their intersection).
3. Under the null hypothesis, the counts in the contingency table follow the hypergeometric distribution. Recall that this is probability of drawing n marbles containing k red marbles from an urn containing K red marbles among N total marbles.
4. The significance level for the test is 0.05, but must be adjusted for multiple testing (as described above)
5. The rejection region of the test is the upper tail of the hypergeometric distribution, and requires numerically integrating over all possible counts in the contingency table more extreme than those in the observed data.

Returning to the urn model (3), intuitively we first partition the regions according to whether they are in the regulatory domain of the genes assigned to the pathway we are considering, and then ask whether we would randomly draw as many regions harboring an associated variant by chance.

Defining the regulatory domains of genes is a fundamental problem in pathway enrichment analysis, which is the main reason we rely on GREAT rather than implementing the test ourselves. There are a number of free parameters in pathway analysis (e.g. how large regulatory domains are, how gene identifiers are normalized) which greatly affect the results. In order to make our analysis reproducible, we instead used a publicly available web service which made well-established choices for these parameters. This choice has a number of consequences: in particular, our analysis is limited to the pathway annotations curated by GREAT, and it is not possible to use more sophisticated definitions of regulatory domains (which we also explore in this chapter).

Defining the background set of regions is also a fundamental problem in pathway enrichment analysis. Of note, for this analysis we did not use the whole genome as the background set (which is a typical choice). Instead for each enhancer module, we defined the foreground as the set of regions containing associated SNPs meeting the heuristic p-value cutoff and the background as all regions in the module. As was shown in prior work, the enhancer modules are already enriched in relevant pathways for cell definition and function[13]. Here, we are asking whether a specific set of putatively disrupted enhancers in disease are further enriched in cell type-specific functions which might be indicative of the biological processes for that disease.

In order to compare the genes we found co-enriched for disrupted enhancers and biological pathways, we used Phenotype-Genotype Integrator (PheGenI) to retrieve a list of known genes for each disease and matched linked genes in each enriched pathway to known genes based on gene names. PheGenI integrates the NHGRI GWAS catalog, a listing of published GWAS results, with a number of other biological databases. Of note, PheGenI uses a controlled ontology of disease terms to simplify searching over diseases.

Our use of GREAT limits our pathway analysis to the Gene Ontology (GO), a collection of computationally defined gene sets which are predicted to have coherent functions. Although some of these terms correspond to canonical pathways (e.g., those found in biochemistry textbooks), many contain only loosely connected genes. Of note, GO is arranged as a directed acyclic graph (DAG) of terms, where each node is a GO term (e.g., "regulation of immune process"), and edges connect children to their parents with a relationship label (e.g., "is a").

The hierarchical nature of GO means that pathway enrichments will identify entire paths through the DAG as enriched. In this chapter, we want to identify specific biological processes which are disrupted in each disease. Therefore, we sought to prune the enriched pathways to identify the most specific pathway (deepest in the DAG) which was still enriched To prune enriched pathways, we downloaded the basic version of Gene Ontology in Open Biomedical Ontologies format and built the specified directed acyclic graph connecting terms to their parents. We performed depth-first traversal of the graph starting from enriched terms and took nodes which were never reached from a child node as the most specific enriched terms.

3.2.2 Motif enrichment

The second biological question we ask in this chapter is which transcription factors mediate the regulatory function of the enhancer regions we identified in Chapter 2. Recall that we hypothesize that enhancers modulate the expression of their target gene by recruiting additional transcription factors to distal regions, which are brought into proximity with the transcription start site through folding of the chromatin structure. Here, we seek to identify transcription factors whose binding could be disrupted by disease-associated variants in enhancer regions by combining predicted enhancer regions across 111 human cell types and tissues (assigned to 226 enhancer modules as we described in Chapter 2) with predicted motif instances of 651 transcription factor families.

We used a database of predicted motifs combining known motifs from existing databases (based on experimental data and literature), Transfac and Jaspar, with de novo discovered motifs in 427 ChIP-Seq experiments for 123 transcription factors from ENCODE[74]. These motifs were manually curated into 651 transcription factor families based on similarity of the estimated PWMs and the experimental metadata (when available).

We additionally rely on prediction of active regulators in each enhancer module, as previously described by the Roadmap Epigenomics Consortium[13]. One fundamental problem in using regulatory motif data is that any particular motif will appear tens of thousands of times in the human genome, just by chance. Consider that there are order 10^9 bases in the human genome, and a particular 8-mer motif will appear with probability 2^{-16}, which is order 10^{-5}. In order to address this issue, we can exploit the fact that functional motif instances within enhancer regions will be conserved by evolution (otherwise, they would accumulate mutations and disappear). Conceptually, this means that truly functional PWM matches will be enriched compared to random PWMs. One of the key aspects of prior work we rely upon is the specific null distribution of PWMs, which must account for a number of additional biases beyond the scope of this thesis. In addition, we rely upon filtering of the PWMs based on evolutionary conservation directly. The full data we used is publicly available from the ENCODE and Roadmap Epigenomics Consortium; however, its value has not been widely appreciated by the field. In this chapter, we only considered PWMs with conservation score at least 0.3, and used log_2-fold enrichment > 1.5 as the significance cutoff.

For each combination of enhancer module and predicted regulator, we tested for co-enrichment of disease-associated variants and the regulator within predicted enhancer regions using Fisher's exact test. Conceptually, we hypothesized that although the function of the enhancer might be mediated by a specific TF, the disease-associated variant within that enhancer might not directly disrupt the binding site for that regulator. Instead, it might alter binding indirectly through disruption of a binding site for a different cofactor, which might be required for a number of reasons (described above). In this case, our enrichment test would allow us to find putative disease master regulators: TFs which bind to multiple enhancers throughout the genome, and whose binding is recurrently disrupted in disease. In our preliminary work, we tested for enrichment of disease-associated variants directly within motif instances and directly within, again using Fisher's exact test and found no enrichments.

We constructed a 2 × 2 contingency table counting enhancer regions in that module partitioned by presence of that motif and orthogonally by presence of a disease-associated variant. We restricted the set of regions to the domain on which motifs were discovered (excluding coding regions, 3' UTRs, transposons, and repetitive regions) and additionally to the subset of regions which harbor an imputed SNP for the disease. We applied the Benjamini-Hochberg procedure at level 0.017 to control the overall false discovery rate at level 0.05.

Having identified putative master regulators, we then asked which putative binding sites were directly disrupted by the disease-associated variants in those enhancer regions. We re-scanned regions containing both a motif instance and a weak association for any motif instances overlapping the associated SNP. We again used the manual curating of the motifs to collapse motifs by transcription factor.

We finally sought to understand the cell type-specificity of the regulators which we identified. Conceptually, even though an enhancer region might exhibit the epigenomic modifications associated with regulatory activity in a number of cell types, the region will not be functional unless the upstream TF which mediates its function is actually expressed in those cells. In order to see whether this was the case, we used gene expression as a proxy for protein abundance, noting that expression is in general a poor proxy for abundance due to post-transcriptional regulation (which is beyond the scope of this thesis). We used the transcription factor gene names to visualize the gene expression of the upstream regulators across 57 reference epigenomes. We normalized the expression RPKM by scaling the maximum value to 1 in order to put expression of each TF on the same scale.

3.2.3 Gene-enhancer linking

Finally, we seek to predict specific mechanisms for individual enhancer regions, by combining the motif enrichments and specific disrupted motifs described above with probabilistic predictions of the target genes for each enhancer. Predicting the target gene for enhancers remains an open problem, with the greatest challenge being experimental validation of the candidate links.

In order to predict the target gene of each enhancer, we used a model called Joint-LDA (Wang et al., in preparation). The key idea of this model is to describe the process generating gene expression using Latent Dirichlet Allocation (LDA), and to simultaneously describe the process generating enhancer activity (presence/absence) using the same model hyperparameters.

LDA was initially developed to model the generative process underlying documents to solve problems like finding documents with topics close to some query. Conceptually, the idea of LDA is that each document is generated as a mixture of some finite set of topics, and that each topic generates words at some rate. Then, we model the observed words in each document as having arisen from this generative process, and phrase questions about topics as estimation problems of the model hyperparameters.

The key insight of Joint-LDA is that we can develop an analogue to this generative process for the problem of linking enhancers and genes. Consider first the problem of generating the observed gene expression data across 57 reference epigenomes. We measure gene expression using a sequencing experiment called RNA-Seq. Analogous to ChIP-Seq (described in Chapter 2), in RNA-Seq we extract the mRNA from the cells and sequence it at high throughput, yielding a set of reads which can be mapped back to the reference genome (hopefully, to protein-coding sequences only). Then, for each gene we can quantify the number of reads which were aligned to that gene, and from there quantify the number of mRNA transcripts which were present in the cell. The units are normalized counts (mapped reads per thousand bases per million reads) in order to account for various biases.

Now, define topics to be gene modules, which are defined to be different sets of cell types in which genes might be expressed. For example, we might consider a module of genes which are expressed only in T cells, or only in brain cells, or in all cell types. Topics generate words, which in this case are gene expression levels (counts) for each gene. For example, the expression of a particular gene might be explained by that gene belonging to a T cell gene module, as well as a gene module which contains all cell types. Finally, the observed gene expression data for all genes in each cell type is a document.

Similarly, in order to generate an enhancer activity matrix (presence/absence in each cell type), we define enhancer modules, enhancers, and cell types in the same way. The key idea of Joint-LDA is to fit both models simultaneously, propagating information from one to the other using a diffusion model. Conceptually, for each gene module there should be a corresponding enhancer module, because the genes which are expressed specifically in e.g. T cells should be regulated by elements which are also active specifically in T cells.

In order to estimate links in Joint-LDA, we estimate the module-module linking probability using a diffusion model, then multiply by the probabilities of generating the specific enhancer and gene from the two LDA models. Briefly, each enhancer module is driven by some single representative cell type (with greatest enhancer activity, or topic mixture component in the language of LDA), and that enhancer module is linked to a gene module with probability proportional to the activity (mixture component) of that cell type. In order to calibrate the false discovery rate of the inferred links, we can train Joint-LDA on permuted gene expression matrices.

We took the enhancers which contained a putative master regulator (as described above) as most likely to harbor a causal variant, and used the predicted links (FDR < 0.01) to predict the target genes of those regulatory elements. We then intersected the predicted target genes with known genes from PheGenI to identify promising candidates for functional followup.

3.3 Results

3.3.1 Pathway enrichment

We first investigated the target genes of enriched tissue-specific enhancer modules harboring weak associations. Prior work has used hierarchical modeling to study enrichment of weak associations in gene pathways[75] (see Chapter 4). We used GREAT[73] to test genes near disrupted tissue-specific enhancers (as defined by the enriched modules overlapping variants meeting the heuristic p-value threshold) for enrichment of Gene Ontology (GO) Biological Processes.

We found significant enrichments for a number of known pathways in each of the eight diseases (hypergeometric test, $q < 0.017$, Table 3.1). Recall from Chapter 2 that we pre-screened enhancer modules to test here for pathway enrichment using the BH procedure, and that applying the BH procedure at each level of the hierarchy controls the FDR of the entire hierarchy of hypotheses, but with multiplicative penalty 2.88 to the overall FDR[64]. Therefore, we apply the BH procedure with $q = .05/2.28 = 0.017$ at this level of the hierarchy such that the overall FDR is 0.05.

In autoimmune disorders, we found enrichment for various pathways relating to immune response. However, we identified different specific signaling pathways in each disease: Immunoglobulin E and Interleukin-4 in CD, nuclear factor kappa-B in RA, and Interferon G in T1D. Surprisingly, we found enrichment for MHC class I/II processes in T1D despite excluding the MHC from the analysis. We verified this enrichment was not due to spurious correlations on chromosome 6 by examining which enhancers in the foreground set (harboring an association surpassing our heuristic p-value threshold) were linked to genes in the MHC pathways by GREAT. We found the enrichment is primarily driven by enhancers linked to CIITA, a known regulator of the MHC pathways which resides on chromosome 16.

In psychiatric disorders, we recovered several known signaling pathways important to brain function (cyclic GMP signaling in AD and glucocorticoid signaling in BIP) and brain development (dendritic spine development and neuron migration in SCZ). We additionally found enrichment for immune response in AD, further supporting the role of immune pathways in this disease.

In CAD, we found enrichments for cholesterol and triglyceride biosynthetic processes, but additionally for the Immunoglobulin A pathway. In T2D, we found enrichment for pancreatic β cell apoptosis, a known hallmark of the disease. Of note, we did not find enrichment for pancreatic β cell apoptosis in T1D. Instead, our enrichments support different mechanisms in different tissues eventually leading to β cell loss[76]. In T1D, immune cell activation leads to increased cytokine levels, in particular Interferon G, signaling activation of β cell apoptotic pathways. In T2D, pancreatic β cells change metabolic state in response to increased blood glucose and lipid levels, leading to apoptosis.

We note that we recovered known pathways by considering weak associations which overlap distal regulatory regions rather than genome-wide significant associations which implicate nearby genes in LD. We used Phenotype-Genotype Integrator (PheGenI) to obtain lists of known genes for each disease and found that across the eight diseases, we linked putative disrupted enhancers to only 23 known genes on average (Table 3.2). The remaining genes are potentially new targets for experimental followup; however, a key shortcoming of this approach is that we cannot assign a p-value to any particular gene.

Our approach yielded a large number of enriched GO terms and an average of 395 linked genes in each of the eight diseases, partly due to overlap between general and specific terms (Table 3.3). We used ontology relationships to prune the list of enriched terms to the most specific enriched terms. Briefly, we built a directed acyclic graph where nodes are GO terms and edges are ontology relationships and took all enriched nodes for which no child was enriched. Our approach recovered 121–366 enriched GO terms; however, we still recovered some a priori implausible pathways, possibly due to incorrect linking of enhancers to their target genes.

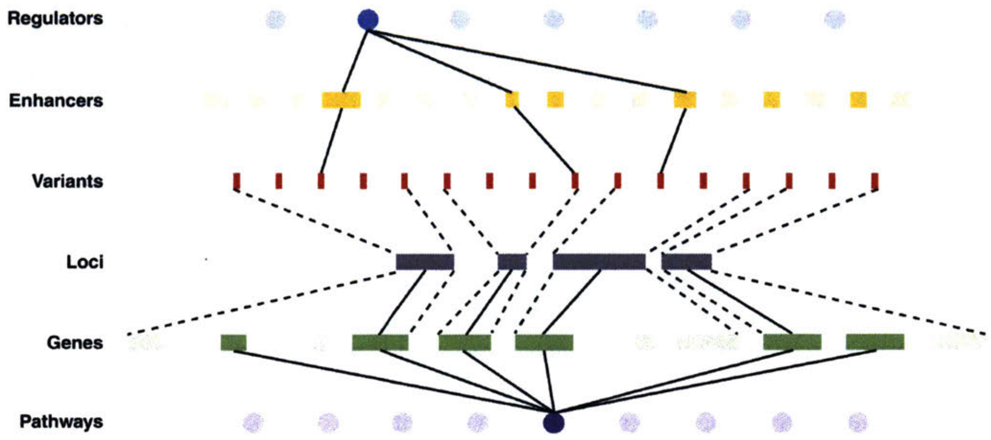


Figure 3.1: Conceptual illustration of the biological problem explored in this chapter. Starting from a set of genomic loci harboring genetic variants enriched within a set of cell type–specific enhancers, we first ask which genes are targeted by those enhancers, and whether those genes have coherent biological functions in some pathway (bottom). Then, we ask whether those enhancers share upstream regulators whose binding is recurrently disrupted by genetic variants (top).

Trait	Known pathways	Total genes	Total pathways
AD	Cyclic GMP signaling, immune response	680	178
BIP	Glucocorticoid signaling	249	154
CAD	Cholesterol/triglyceride biosynthetic process, IgA	171	166
CD	CD8 T cell proliferation, IgE, IL4	717	366
RA	NFKB	299	121
SCZ	Dendritic spine development, neuron migration	411	184
T1D	MHC I/II, JAK-STAT, IFNG	298	176
T2D	Pancreatic β cell apoptosis	334	150

Table 3.1: Pathway enrichments of enhancers harboring weak associations meeting the heuristic p-value threshold. Total gene counts are based on links to weakly associated enhancers across any significantly enriched pathway. Total pathway counts are restricted to GO terms with significant enrichments (FDR $q < 0.017$) for which no child (connected by an ontology relationship) is significantly enriched.

Trait	Known genes
AD	ABCA1, ABCA7, APOC1, APOE, BIN1, CD2AP, CELF2, CHORDC1, CLU, CUBN, EPHA1, ETS1, FARP1, GAB2, IKZF1, LEPREL1, NEDD9, PCSK5, PICALM, PRKCQ, PVRL2, RORA, TFCP2L1, TOMM40
BIP	ACTR3B, CACNA1C, CACNB3, LMAN2L, MAD1L1, MDM1, MSI2, RASGRP1, RASIP1, RYBP, TWIST2, ZMIZ1
CAD	ALDH2, APOA5, CELSR2, CNNM2, COL4A1, COL4A2, CXCL12, IL6R, PPAP2B, PSRC1, RAB23, RAI1, RIMS1, SH2B3, SLC22A3, SMARCA4, TSHZ3, UBE2Z
CD	ADRA1B, ATG16L1, C11orf30, CCL2, CCL7, CSF2, DAB2, FASLG, FGFR1OP, FNDC1, FUT2, ICOSLG, IKZF1, IL12B, IL23R, IL3, IL6R, IRGM, ITLN1, JAK2, KIF21B, KLF6, LRRC32, LRRK2, NKX2-3, NOD2, PTGER4, PTPN2, SLC22A4, SLC22A5, STAT3, TAGAP, TNFSF15, TNFSF18, TRIB1
RA	AFF3, ARID5B, BLK, C1QTNF6, CD247, CDK6, CTLA4, ETS1, FGD4, IL2RA, IL2RB, IRF5, KIAA1109, MAFB, PHF19, PTPN2, PTPN22, RBPJ, REL, RPLP1, SPRED2, TMEM17, TNPO3, TRAF1, UBASH3A
SCZ	ANK3, ANKRD11, CACNA1C, CDK1, CNNM2, DST, ERBB4, KDM4C, MAD1L1, MSRA, MTIF3, NT5C2, PLCB2, PTBP2, RERE, SDCCAG8, SNX19, SPTBN1, TCF4
T1D	AFF3, ASCL2, C1QTNF6, CCR7, CD69, CLEC2D, CTLA4, CTSH, CUX2, IGF2, IL10, IL2, IL2RA, IL2RB, IL7R, INS, PRKCQ, PTPN2, PTPN22, RBPJ, SH2B3, SIRPG, SMARCE1, SSTR3, TH, UBASH3A
T2D	ARAP1, BCL11A, CDKAL1, DMRTA1, FAH, FTO, HMG20A, IDE, IGF2BP2, IRX3, JAZF1, KCNQ1, LIF, NDFIP2, NRG1, PPARG, RBMS1, SPRY2, TCF7L2, TIMP4, TP53INP1, VEGFA, WFS1

Table 3.2: Previously reported genes in the NHGRI GWAS catalog implicated in pathways enriched for links to weakly associated enhancers based on proximity across eight diseases. Genes are taken across all enriched pathways for each disease rather than only known pathways.

Trait	Count of known genes	Total count of genes
AD	24	680
BIP	12	249
CAD	18	171
CD	35	717
RA	25	299
SCZ	19	411
T1D	26	298
T2D	23	334

Table 3.3: Counts of known genes and total genes linked to an enriched Gene Ontology term across eight diseases.

3.3.2 Motif enrichment

We next identified the upstream regulators whose binding may be perturbed by weak associations. Prior work has studied enrichment of regulatory motifs in enhancer regions[13,59]; however, these studies do not consider the impact of SNPs on transcription factor binding affinity at specific motif instances. We studied regulatory motifs curated into 651 families[74] and hypothesized that weak associations may recurrently affect binding of a small number of disease-specific master regulators by disrupting motif instances of co-factors[77].

Briefly, putative regulators are defined as regulatory motifs represented as position weight matrices (PWMs), filtered according to enrichment of PWM matches against shuffled PWM matches, as previously described[74]. We tested for enriched co-occurrence of weak associations and each putative regulator in each enhancer module using Fisher's exact test to identify putative disease master regulators. We finally re-scanned enhancer regions containing both a master regulator motif instance and a weak association to find co-occurring motifs which overlap weakly associated SNPs.

Our approach identified 61 master regulators across the eight diseases (Fisher's exact test, BH FDR < 0.017, Figure 3.2). Only 4 of the 61 regulators have been previously identified by GWAS for the eight diseases and reported in PheGenI: ETS1 and JDP2 in RA and NFKB1 and RUNX2 in SCZ. This result is expected given that the majority of GWAS-identified loci do not implicate protein-coding genes; however, it also illustrates the power of integrating genetic information with knowledge of the transcriptional regulatory network to identify genes whose biological function is disrupted by disease-associated non-coding variation.

Several of the putative master regulators play known roles in related phenotypes, giving orthogonal evidence for their importance in the eight diseases we studied. We identified RXRA in AD, which alters brain cholesterol metabolism[78]. We identified ELF3 in CD, which is over-expressed in ulcerative colitis (UC) cases[79], supporting prior work suggesting CD and UC share common genetic factors[80]. Additionally, several of the putative master regulators have known biological functions which are relevant a priori to the disease they were identified in. We identified REL and ETS1 in multiple diseases, which are known to play a role in immune response[81,82]. We identified SPI1 in AD, consistent with prior work showing an immune basis for AD[83].

We examined the enhancer regions bound by these master regulators and identified a large number of putative co-factors whose binding sites are directly disrupted by disease-associated non-coding variants (Figure 3.3). Moreover, we found that the identified co-factors are specific to both the master regulator and the disease, offering an explanation for how putative master regulators such as AP1 can be shared between very different diseases such as psychiatric and auto-immune disorders.

We note that we identified many master regulators in constitutive enhancers and immune-specific enhancers (Figure 3.4). One explanation for this result is that our method is under-powered to find master regulators in other enhancer modules which cover less of the genome and overlap fewer well-imputed variants. However, even allowing for lack of power in other tissue-specific modules, enrichment in constitutive enhancers runs counter to the hypothesis that different cell type–specific regulators are disrupted in different tissues of action in different complex diseases. We hypothesized that although the enhancers might be constitutively marked, the transcription factors which bind to those enhancers would show cell type–specific patterns of expression, explaining their disease specificity. We used RNA-Seq data across 57 reference epigenomes to study the expression of putative master regulators discovered in constitutively marked enhancers and found that indeed they showed diverse patterns of expression (Figure 3.5). For example, *REL*, *SPI1*, and *ETS1* are predominantly expressed in T cells, consistent with their known tissue-specific functions.

Our results highlight a key distinction between constitutive marking of enhancer-like regions and constitutive activity of distal regulators. However, we found only few master regulators predicted for any disease are clearly expressed in only relevant cell types, possibly due to incomplete profiling of expression across tissues and developmental time points.

We recently applied our method to meta-analysis of over 370,000 women analyzing the genetic determinants of age at menarche (AAM), analyzing a subset of the data excluding nearly 75,000 samples from 23andme due to data use restrictions. In total, we tested 2,382 transcription factor-enhancer module combinations and found sixteen motifs enriched within AAM-associated enhancers (FDR < 0.05). Furthermore, the genes encoding 5 of the 16 enriched transcription factors were also within 1Mb of a genome-wide significant AAM-associated SNP. These transcription factors included notable candidates:

1. Pituitary homeobox 1 (PITX1), is located within 50kb of a genome-wide significant SNP
2. SMAD3 is located within 600kb of an index SNP and its expression in several brain tissues (as assayed by the Gene Tissue Expression Project[84]) is genetically correlated with AAM.
3. RXRB is located within ~500kb of a new locus, and it represents the fifth (out of nine) retinoid-related receptor gene implicated by genome-wide significant AAM variants.

Of note, for a study of this size, our heuristic p-value threshold did not actually include sub-threshold genetic variants. However, we identified these AAM-associated genes using enrichment of binding sites within distal enhancers rather than the genome-wide significant hits in the vicinity of the coding sequence for the upstream TFs.

3.3.3 Gene-enhancer linking

We next sought to complete the proposed biological mechanisms by linking the disrupted enhancer regions harboring putative master regulators to their downstream target genes as predicted by Joint-LDA. In total, we found 99 enhancers, probabilistically linked to 113 target genes in the eight diseases (Table 3.4). We first investigated which of these enhancers could have already been identified by GWAS for each of the eight diseases and found that in total only 8 of the 99 enhancers were in known loci. We then compiled the complete list of predicted gene-enhancer links disrupted by a disease-associated variant (Table 3.5), which are promising candidates for experimental followup.

Trait	Module	Regulator	Chromosome	Start	End	Target gene
AD	22	HNF4A	1	246841402	246841414	SCCPDH
AD	159	NFE2	5	139729147	139729154	ANKHD1, ANKHD1-EIF4EBP3, HBEGF, PFDN1, SLC35A4, SRA1
AD	66	STAT3	6	32572322	32572331	HLA-DQB1
AD	66	NFKB	8	95972805	95972822	TP53INP1
AD	77	BACH1	8	27219982	27219991	PBK
AD	77	ETV7	8	27223659	27223682	EPHX2, PTK2B
AD	178	SPIC	10	11741657	11741671	ECHDC3
AD	180	HLF	11	130767294	130767307	SNX19

Continued on next page

Continued from previous page

Trait	Module	Regulator	Chromosome	Start	End	Target gene
AD	66	ELF1	11	47907629	47907642	FNBP4
AD	157	PPARA	15	85383838	85383849	ALPK3, SEC11A
AD	66	SPI1	16	84784714	84784729	COTL1
AD	66	ETS	17	44342561	44342572	KIAA1267
AD	157	RXRA	19	45242729	45242748	APOC1, APOC2, APOC4, APOE, BCAM, BCL3, CKM, NA, TOMM40
AD	66	ETS	20	48628950	48628966	RNF114, TMEM189, TMEM189-UBE2V1, UBE2V1
BIP	77	AP1	2	106536126	106536137	NCK2
BIP	66	ETS1	7	150179026	150179042	GIMAP2, GIMAP4, GIMAP7
BIP	66	ETS1	11	66048276	66048290	B3GNT1, BRMS1, KLC2, NA, PACS1, RBM14, SF3B2, SSSCA1, TMEM151A
BIP	66	FLI1	11	66048276	66048290	BANF1, CD248, CNIH2, RAB1B, RBM14-RBM4, RBM4, RBM4B, YIF1A
BIP	77	ETS1	11	66648098	66648115	LRFN4, RCE1, RHOD
BIP	77	NFKB	14	102427306	102427326	DYNC1H1
BIP	77	MEF2A	17	65470325	65470336	PITPNC1
BIP	77	MYEF2	17	37833838	37833866	CASC3, ERBB2, GRB7, LASP1, MIEN1, NEUROD2, PGAP3, PNMT, PPP1R1B, PSMD3, STARD3, TCAP
BIP	66	ETS1	19	3478646	3478659	C19orf77, FZR1
BIP	66	ETS1	20	44633111	44633124	MMP9, PCIF1, TNNC2, ZNF335
CAD	213	SPIB	1	3010702	3010714	PRDM16
CAD	77	NFKB	1	154436369	154436386	ATP8B2, S100A14, S100A16, S100A2
CAD	77	ETV7	3	195847490	195847501	TNK2
CAD	77	AP1	10	75669190	75669200	ADK, AP3M1, CAMK2G, NDST2, PLAU, SEC24C, VCL
CAD	125	ETS1	12	111884607	111884622	ALDH2, MYL2
CAD	77	ELF5	13	41558041	41558052	ELF1
CAD	77	BHLHE41	14	68974410	68974429	ACTN1
CAD	77	NFKB1	15	79049404	79049427	ADAMTS7, MORF4L1
CAD	77	SPIC	17	62404446	62404455	CEP95, PSMC5
CAD	77	NFKB	19	18417368	18417380	ARRDC2, JUND
CD	66	RUNX	1	206939895	206939907	FAIM3, IL10, MAPKAPK2, RASSF5
CD	66	RUNX	2	43764431	43764443	ZFP36L2
CD	77	SPIC	2	198121206	198121222	ANKRD44, SF3B1
CD	66	SPI1	3	49423970	49423981	GPX1
CD	66	ETS1	5	131809314	131809331	C5orf56, IRF1
CD	66	FEV	5	40674916	40674931	PTGER4
CD	66	RUNX2	5	131410878	131410894	PDLIM4
CD	66	RUNX3	5	131410878	131410894	CSF2
CD	66	ETS	6	32591206	32591213	HLA-DRA, HLA-DRB1, HLA-DRB5
CD	66	ETS	6	167372973	167372980	RNASET2
CD	66	NFKB1	6	167512471	167512482	CCR6, NA
CD	66	ETS2	7	50257626	50257637	IKZF1
CD	66	NFKB	10	35438331	35438338	CREM
CD	66	RUNX2	13	100027955	100027967	GPR183
CD	66	NFKB	14	69288728	69288744	ZFP36L1

Continued on next page

Continued from previous page

Trait	Module	Regulator	Chromosome	Start	End	Target gene
CD	77	BACH1	14	35834275	35834287	NFKBIA
CD	66	FEV	15	38903657	38903675	RASGRP1
CD	66	EHF	16	50719740	50719749	NOD2
CD	77	SPIC	16	50730445	50730459	CYLD, SNX20
RA	77	BACH1	1	111177853	111177867	CD53, KCNA3
RA	193	SPIB	6	32192511	32192518	GPSM3, RNF5
RA	226	SRF0	6	30720300	30720314	DDX39B, DHX16, NRM, PPP1R18
RA	226	SRF	6	31082127	31082142	IER3, MUC21
RA	77	AP1	6	30738207	30738215	FLOT1, GTF2H4
RA	77	AP1	9	123700183	123700195	TRAF1
RA	77	ETS	9	123659293	123659307	FBXW2, PHF19
RA	226	SRF	11	61637466	61637475	FADS1, PGA5
RA	226	SRF	11	61637466	61637475	C11orf10, PGA4
RA	77	AP1	13	29291422	29291437	POMP
RA	77	BACH1	17	79448661	79448676	DCXR, SLC25A10
RA	77	CTCF	19	19478048	19478066	GATAD2A
SCZ	77	BACH1	1	224019631	224019648	TP53BP2
SCZ	77	ETS	1	154913398	154913406	ADAR, CKS1B, EFNA1, PBXIP1, PMVK, PYGO2, ZBTB7B
SCZ	77	MEF2A	1	205153193	205153210	DSTYK, TMCC2
SCZ	77	MYEF2	5	137073590	137073600	HNRNPA0, KLHL3
SCZ	77	SPIC	7	104601234	104601250	MLL5
SCZ	77	AP1	8	131074007	131074028	FAM49B
SCZ	77	AP1	8	143757590	143757600	PSCA
SCZ	77	AP1	8	143757590	143757602	ARC, SLURP1
SCZ	77	AP1	8	143757590	143757600	LY6K
SCZ	77	ELF4	10	3807121	3807132	KLF6
SCZ	77	MEF2A	10	104941102	104941117	USMG5
SCZ	77	ETS	11	64629267	64629281	ATG2A, C11orf2, CDC42BPG, EHD1, MAP4K2, RASGRP2, SF1
SCZ	77	ETV6	11	63769143	63769156	OTUB1, PPP1R14B, RCOR2, RTN3
SCZ	77	NFKB	11	133817318	133817335	IGSF9B
SCZ	77	MEF2A	12	123591584	123591597	PITPNM2
SCZ	77	CTCF	13	114917151	114917158	RASA3
SCZ	77	CTCF	15	93461360	93461371	CHD2
SCZ	77	CTCF	15	43805981	43805991	MAP1A, TP53BP1
SCZ	77	MYEF2	16	89162335	89162351	ACSF3, CYBA
SCZ	77	NFKB1	16	68113858	68113877	DPEP2, EDC4, NFATC3, PSMB10, SLC7A6, SLC7A6OS, THAP11
SCZ	77	ETS	17	30844380	30844393	CDK5R1, PSMD11, TMEM98, ZNF207
SCZ	77	ETV7	17	78945130	78945149	CHMP6
SCZ	77	MYEF2	17	2167685	2167708	DPH1, MNT, OVCA2, PAFAH1B1, PRPF8, SGSM2, SMG6, SRR, TSR1
SCZ	77	SPIC	17	17861358	17861374	FLII, SREBF1, TOM1L2
SCZ	77	AP1	22	42335620	42335633	CENPM, XRCC6
SCZ	77	SPIB	22	42697445	42697452	C22orf32

Continued on next page

Continued from previous page

Trait	Module	Regulator	Chromosome	Start	End	Target gene
SCZ	77	SPIC	22	42697445	42697452	NDUFA6, TCF20
T1D	77	ETS1	2	64894148	64894159	SLC1A4
T1D	77	BACH1	8	144656974	144656987	EEF1D, FAM83H, GRINA, GSDMD, NAPRT1, PLEC, TSTA3, ZC3H3
T1D	77	ETS	8	144656974	144656987	C8orf73, LY6E
T1D	77	ELF5	10	6094687	6094703	GDI2, IL15RA, IL2RA, PFKFB3, RBM17
T1D	77	NFKB1	16	11652118	11652130	LITAF, RSL1D1, SNN, TXNDC11, ZC3H7A
T1D	77	ETS	17	45961822	45961829	PRR15L, SP2
T1D	77	NFKB1	17	45961822	45961829	CDK5RAP3, NFE2L1, SCRN2
T1D	77	SRF	19	10621108	10621123	AP1M2, ATG4D, CDC37, CDKN2D, ICAM3, ILF3, KEAP1, KRI1, S1PR5, SLC44A2
T1D	77	MEF2A	21	46530097	46530107	ITGB2
T2D	226	TEAD1	6	43818941	43818952	VEGFA
T2D	180	MYEF2	12	121198296	121198312	ACADS, COQ5, COX6A1, DYNLL1, GATC, MLEC, UNC119B
T2D	226	TEAD4	12	66271198	66271219	HMGA2, NA
T2D	226	TEAD4	13	110381681	110381691	IRS2
T2D	51	AP1	15	39505814	39505825	THBS1

Table 3.5: Downstream target genes and upstream regulators linked to predicted enhancers harboring disease associated variants across eight diseases. Target genes are linked at FDR < 0.01.

3.4 Discussion

In this chapter, we showed that putative regulatory regions harboring weak associations target relevant downstream genes and are regulated by relevant upstream transcription factors. We showed that a number of these weak associations targeted known gene pathways but through previously uncharacterized regulatory mechanisms, proving the utility of incorporating transcriptional regulatory network information in interpreting GWAS.

Our results highlight an important distinction in the use of epigenomic annotations as proxies for transcriptional regulatory elements. Regions marked by enhancer-associated histone modifications are not necessarily active distal regulators. Here, we attempted to characterize putative enhancers by linking them to downstream genes and upstream transcription factors. We also used measured expression of predicted upstream regulators to decouple enhancer activity from enhancer marking, and deconvolve enhancer activity across 57 cell types.

Our results also highlight an additional key difference in our methodology and existing methods such as LDSC. The LDSC baseline model includes motifs learned in LCLs; however, LDSC has not been applied to study tissue-specificity of regulatory motifs. Recently, LDSC has additionally been applied to study disease enrichments of sets of strongly expressed genes across 53 different tissues, an approach termed LDSC-SEG[85]. The key difference between the approach we take in this chapter and LDSC-SEG is that LDSC-SEG generates sets of genes which are predicted to be active in cell types, and then compares whether those cell types additionally showed concordant

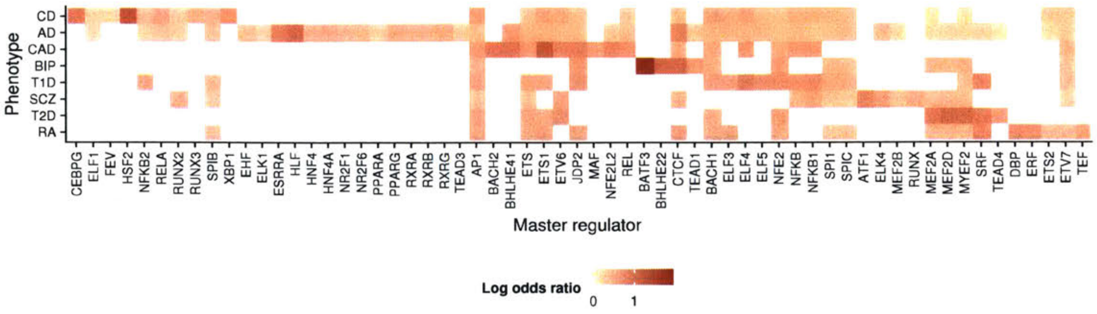


Figure 3.2: Putative master regulators enriched in enhancer regions harboring associations meeting the heuristic p-value threshold. For each phenotype and TF, we take the maximum enrichment (log odds ratio) over the subset of 226 enhancer modules in which associations meeting the heuristic p-value threshold are enriched and in which the TF is predicted to be an active regulator. Only log odds ratios for 61 master regulators with significant enrichment (Fisher's exact test, BH FDR < 0.017) are shown. Phenotypes are represented by a vector of log odds ratios over each of the master regulators and ordered by hierarchical clustering.

Trait	Disrupted enhancers	Total genes	Known loci	Known disease genes
AD	14	32	1	APOC1/APOE/TOMM40
BIP	9	44	0	
SCZ	25	68	0	
CD	18	28	2	NOD2, C5orf56
RA	11	23	1	TRAF1
T1D	7	37	1	IL2RA
CAD	10	23	3	ALDH2, ACTN1, ADAMTS7
T2D	5	12	0	

Table 3.4: Summary of disrupted mechanisms found by our method. Disrupted enhancers are predicted enhancer regions containing a predicted regulatory motif. Total genes are all genes linked to those enhancers (including ambiguous cases). Known disease genes are genes catalogued as associated with the corresponding phenotype in Phenotype-Genotype Integrator (PheGenI). Known GWAS genes are genes associated with any phenotype in PheGenI.

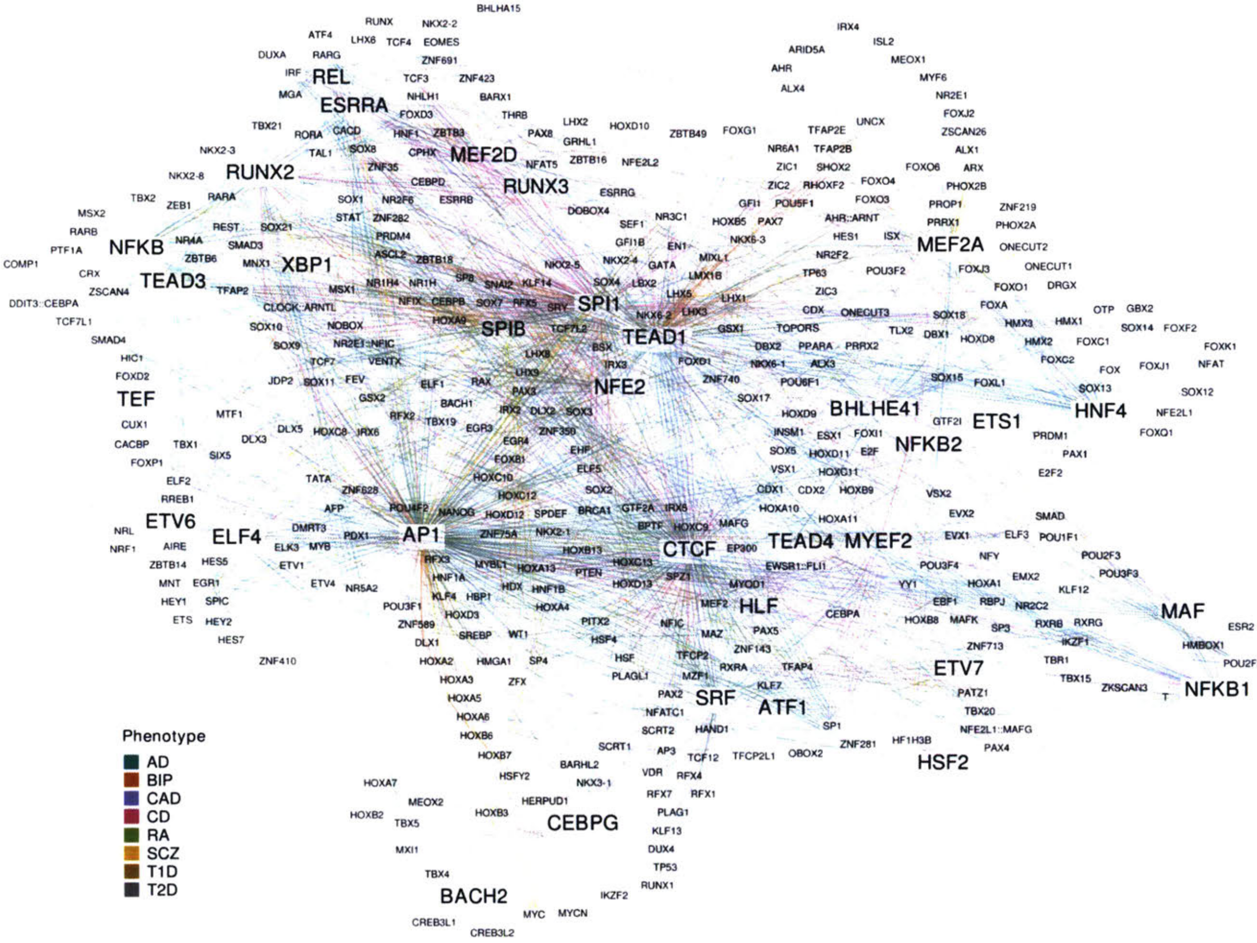


Figure 3.3: Indirect disruptions of master regulators enriched in enhancer regions by associations meeting the heuristic p-value threshold across eight diseases. Master regulator gene names are given in larger text compared to co-factor gene names. Edges connect master regulators to co-factors for which a motif instance overlaps a weakly associated SNP in an enriched enhancer region and are colored by the associated phenotype. Edges are collapsed such that each interaction appears at most once.

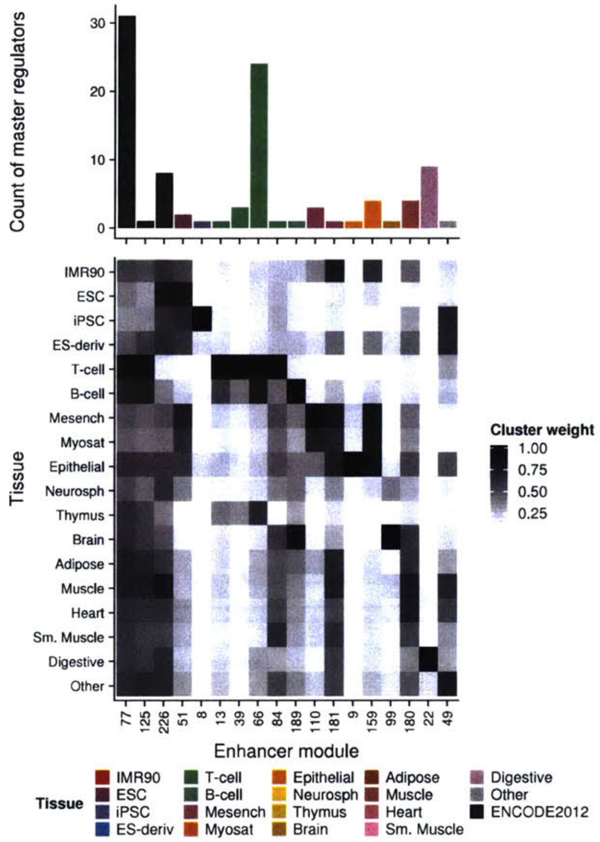


Figure 3.4: Counts of putative master regulators identified across any of the eight diseases in 226 enhancer modules comprising patterns of observed histone modification across 111 reference epigenomes. Only counts for 18 enhancer modules in which a master regulator was discovered in any phenotype are shown. The leftmost three modules are defined as constitutive (having at least 50% of cluster weights greater than 0.25).

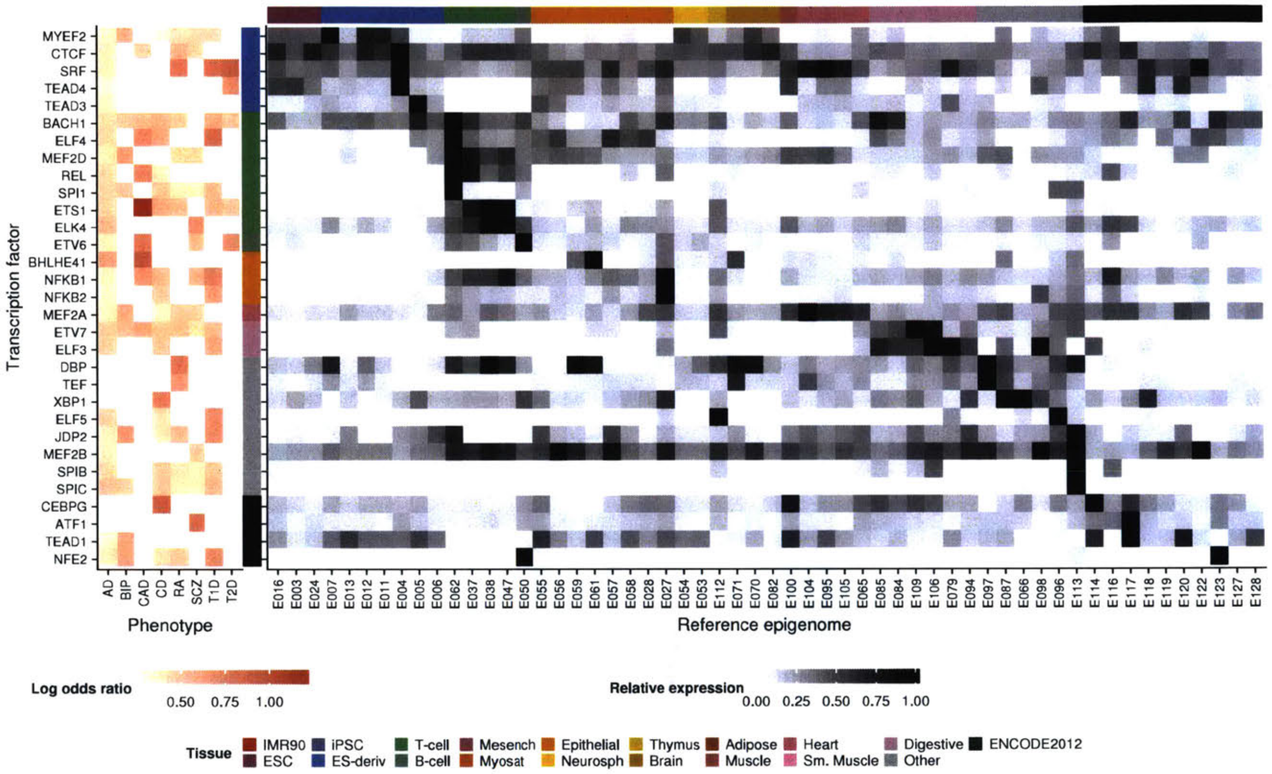


Figure 3.5: Scaled expression (such that the maximum equals one) across 57 reference epigenomes of putative master regulators predicted in constitutively marked enhancers in eight diseases. Individual genes are assigned to tissue groups based on the tissue group of the reference epigenome with maximum expression.

enrichments based on epigenomic annotations. In contrast, we analyze a set of genes predicted to be a coherent biological pathway based on a Gene Ontology term, and then test whether the genes in that pathway are targeted by putative enhancers predicted to be active in particular sets of cell types. Further, our incorporation of gene-enhancer linking in interpreting sub-threshold non-coding loci has not been previously studied. Therefore, our methodology incorporates additional data types not previously studied, and reveals biological insights which could not be found by existing methods.

The methodology presented in this chapter has several limitations which should be addressed in future work. Most importantly, our methodology finds excesses of associations, linked genes, and motifs in specific annotations and pathways but does not naturally provide measures of confidence for particular loci, genes, or master regulators. We used the BH procedure throughout to control the FDR of rejected hypotheses over the entire study. However, this procedure does not estimate a local FDR for each hypothesis, and a novel empirical Bayes model would be required to estimate local FDR in our hierarchical setting, which is beyond the scope of this study. Thus, our results should be interpreted as identifying putative enhancer regions, genes, and transcription factors whose role in disease mechanism needs to be confirmed by experimental followup.

In principle, we should use the gene-enhancer links to perform pathway enrichment, which might improve the quality of the results by reducing the number of *a priori* irrelevant enrichments. However, the fundamental challenge is that we don't know the false positive rate of the predicted links (in the sense of how many would replicate in a targeted experiment). In our preliminary work, incorporating Joint-LDA links into the enrichment analysis greatly reduced the number of pathways found, especially for enhancer regions predicted to be active in non-immune cells. Conceptually, the modeling choices made in Joint-LDA bias the predicted links to immune cell gene and enhancer modules (Wang, personal communication). The intuition behind why this is the case is that immune cells are best represented both in the expression data and the enhancer data, and therefore the immune cell module is also best represented. The diffusion model which links enhancer modules to gene modules starts from the cell type most likely to have been generated by each enhancer module, which just by chance will be an immune cell type due to the distribution of cell types among the different tissue groups in the Roadmap Epigenomics project.

We could further improve upon the gene-enhancer linking by scanning the associated promoter regions for interacting transcription factors, incorporating protein-protein interaction information. Although this method is a straightforward extension of the work presented in this thesis, the main shortcoming of this approach is not incorporating uncertainty in which transcription factor is likely to be functional in the enhancer and the promoter, and which of many potential TF-TF interactions is most likely to be functional. In order to properly account for these different possibilities, we really need a model incorporating these these diverse data types.

Along the same lines, the field has begun to scale up experimental techniques to directly observe physical interactions between DNA elements. So called *chromatin conformation capture* experiments are high throughput sequencing experiments in which interacting DNA is chemically bound together. Then, after fragmenting the genome and selecting for interacting fragments, we can sequence the ends of the fragments and align the ends to the reference genome, yielding a contact frequency map of all positions against all other positions on each chromosome. To date, we have constructed reference contact maps for dozens of human cell types and tissues[86], providing a rich resource for building models of the transcriptional regulatory network and integrating that network with generative models for disease.

We analyzed several million well-imputed variants in each of the eight diseases, but we also used higher resolution, higher confidence predictions of regulatory regions, making it more difficult to find significant enrichments. Although we initially found thousands of loci, they implicate only hundreds of putative enhancer regions of which only a fraction either harbor an enriched motif or target a gene in an enriched pathway. Future work will need to use more comprehensive panels of variants, better predictions of transcription factor binding sites, and better predictions of distal targets to increase the number of high-confidence testable hypotheses to carry forward to experimental followup.

More broadly, our methods use annotations of regulatory regions, genes, pathways, and transcription factor binding sites produced by a number of published computational pipelines. These annotations could be sensitive to choices of thresholds and filtering used in each of the pipelines, and therefore our results could also be sensitive to such choices. Further work will be needed to characterize the error rates in regulatory annotations and the impact of errors on downstream analyses.

Chapter 4

Dissecting non-infinitesimal architectures

4.1 Background

In Chapters 2 and 3, we proved the utility of epigenomic annotations in interpreting the role of non-coding variation in disease. In particular, we used enrichments in diverse annotations to not only identify the relevant tissues, but also identify specific regulatory pathways and biological mechanisms. However, a more fundamental question remains to be answered: why are predicted regulatory elements enriched for disease-associated variants?

As we noted in Chapter 2, the methodology developed in this thesis is closely related to a number of other methods for identifying relevant annotations using GWAS data. To motivate the model presented in this chapter, we explain in more detail the most important of the alternative methods, known as heritability partitioning[65].

Intuitively, the heritability of a phenotype quantifies how similar two individuals will be, given they share some proportion of genotypes. For example, Francis Galton observed that human height was highly heritable, and specifically that the height of offspring tended to regress towards the mean from the heights of their parents.

Historically, the concept of heritability was introduced by Sir Ronald Fisher, who was interested in how Mendelian inheritance of discrete characters (it was not known what these discrete units were yet) could give rise to continuous phenotypes such as height, and how Darwinian natural selection could change the distribution of these continuous phenotypes in the population. Fisher's insight was that as the number of discrete characters (which were later thought of as genes, and which we think of now as genetic variants) which drove phenotype increased to infinity, and if their effects added up to the final phenotype, then Mendelian inheritance of those discrete characters would give rise to a continuously distributed phenotype in the population.

Mathematically, there are two ways to define heritability h^2, leading to two different study designs and methods of estimating heritability:

1. Starting from the viewpoint that heritability quantifies phenotypic similarity as a function of genotypic similarity, we can define:

$$\mathbb{E}\left[p_i p_j\right] = h^2 \mathbb{E}\left[G_i G_j\right]$$

2. Starting from a specific generative model of phenotype and genotype $y = X\theta + \epsilon$, we can define

$$h^2 = V[X\theta]/V[y]$$

These definitions depend on a number of assumptions which are typical in the field, but nonetheless are known not to hold in real data. The two most important of these assumptions are that genetic effects only add and do not interact with each other, and that genetic and environmental effects only add and do not interact. We do not explore violations of these assumptions in this chapter.

In the first definition, we write the phenotypic correlation of two individuals i and j in terms of their genotypic correlation. To gain some insight into this equation, consider first the case of two monozygotic twins who share their entire genome, i.e. $\mathbb{E}\left[G_i G_j\right] = 1$. In this case, if the phenotype is completely heritable, we expect the twins to have the same phenotype, and as the heritability of the trait reduces, we expect the twins to have increasingly different phenotypes.

The main challenge in using the first definition to estimate the heritability is accurately estimating $\mathbb{E}\left[G_i G_j\right]$. Historically, the field has used pedigrees (family trees) to estimate this quantity, using the fact that direct relatives share known proportions of the genome in expectation. For example, monozygotic twins share 100%, dyzygotic twins share 50%, parents and children share 50%, and siblings share 25%. Human genetic studies of monozygotic twins have revealed that many human phenotypes are heritable; for example, the heritability of human height is estimated to be 80-90% in twin studies[6].

One of the recent key insights has been that we can still estimate $\mathbb{E}\left[G_i G_j\right]$ in unrelated (really, distantly related) individuals in the population using a comprehensive catalog of genetic variants. In this case, we can compute a genetic relatedness matrix (GRM) \(G = X X' / p), such that the entries of this matrix are the covariance between the individuals. Intuitively, rather than estimating the proportion of the genome which is shared through inheritance from a common ancestor, this method estimates the proportion of the genome which has the same state, regardless of where it was inherited from. For the purposes of estimating heritability, these two quantities are in principle equivalent, although there is still intense debate in the field about whether these two are equivalent in practice, which is beyond the scope of this thesis.

In the second definition, we directly posit an additive generative model for phenotype. Under the assumptions that this model is correctly specified, heritability can be defined as the proportion of variance explained (PVE) by the fitted model. In other words, how much of the variation in the phenotype is explained by variation in genotype? In this chapter, we start from this second definition, and develop a Bayesian method to directly fit this generative model.

Perhaps surprisingly, existing methods do not estimate this quantity by fitting the regression model $y = X\theta + \epsilon$. The key challenge in fitting this model is that it is ill-posed: we have more predictors than data points, known as the $p \gg n$ problem. The dominant approach to estimate this quantity is instead a mixed effects approach. Here, "mixed" refers to incorporating both directly observed fixed effects and unobserved random effects in a linear model, leading to a linear mixed model.

In order to apply this approach to estimate heritability, we have to make the infinitesimal assumption: that every genetic variant has a random, unobserved effect on phenotype. If we assume that these effects are drawn from a Gaussian distribution, then we can marginalize out the effects and get a variance components estimation problem. This problem can be solved by maximum likelihood or the method of moments. For example, PCGC regression uses the method of moments to derive the equation $\mathbb{E}\left[p_i p_j\right] = h^2 G_{ij}$ as described above. In order to estimate h^2, we just have to fit a linear regression of phenotype correlations $p_i p_j$ against genetic correlations G_{ij} for each pair of individuals.

Importantly, these estimators are frequentist. Although we posit that random effects come from some distribution, this distribution is not a prior distribution, and the inference task is not to estimate a posterior. Instead, the key idea is to model the vector θ as some unobserved realization of a random generative process, and then marginalize out that process to get a problem with few parameters.

The key insight behind partitioning heritability is that starting from the same linear model, we can ask about the proportion of variance explained by specific subsets of predictors. Technically, this relies on a further assumption that causal variants are in linkage equilibrium, although in practice this does not appear to be an issue:

$$\mathbb{V}\left[X\theta\right] = \sum_k \sum_{j \in [k]} \theta_j^2 \mathbb{V}\left[X_j\right]$$

In the same way that we studied epigenomic annotations by asking whether genetic associations are enriched in regulatory regions, we can ask whether the genetic variants which overlap those regulatory regions explain more variance of the phenotype than expected by chance. Typically, we compare the variance explained per variant (again, following the infinitesimal assumption) against the null that every variant explains equal proportion of variance, regardless of whether it overlaps a regulatory region.

Heritability partitioning has been applied to the same traits we study here, with largely concordant results (as outlined in the discussions of the previous chapters). In this chapter, we are interested in the more fundamental question of why regulatory regions appear to be enriched, and to gain some insight into the cases where heritability-based methods and enrichment methods as explored in this thesis could disagree.

There are multiple possible explanations for why regulatory regions are enriched for associated variants and explain more phenotypic variance than expected by chance:

1. Regulatory elements harbor more causal variants for disease than the rest of the genome
2. The causal variants within regulatory elements have larger effect sizes than those outside
3. The causal variants within regulatory elements have different minor allele frequencies than those outside
4. Regulatory regions show different levels of LD than other regions of the genome

To motivate the contributions of this chapter, we consider a simple simulation where we generated synthetic genotypes in linkage equilibrium and phenotypes from a Gaussian linear model. We divided the synthetic genome into two equal sized regions, and considered two non-infinitesimal genetic architectures: (1) equal number of causal variants are drawn per region, but effects in the first region are drawn from a distribution with half the precision (twice the variance), and (2) all causal effects are drawn from the same distribution, but twice as many are drawn from the first region. We then estimated the heritability explained by each half of the genome using PCGC regression, and normalized by the number of SNPs in order to compute the per-SNP heritability. We then normalized by the total heritability per SNP over the whole genome to estimate the heritability enrichment.

As expected, PCGC regression estimates that the heritability enrichment is equal in both cases and therefore cannot distinguish between these two architectures (Figure 2.3). The fundamental reason why existing mixed model approaches (GREML, PCGC, LDSC) cannot distinguish these two architectures is that all of these methods assume each genetic variant explains equal proportion of variance in expectation.

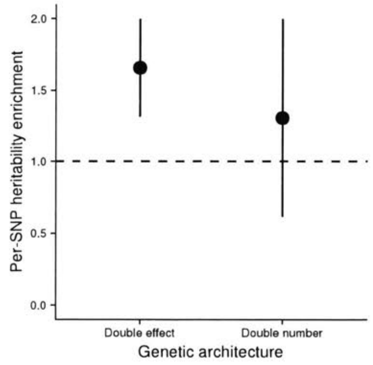


Figure 4.1: Heritability enrichment under two simulated non-infinitesimal architectures: effect sizes drawn from a distribution with double variance, or twice as many causal variants in the enriched annotation.

In Chapter 2, we developed a heuristic to perform univariate feature selection on GWAS variants, and then performed post-hoc analysis of the selected features. In this chapter, we will pursue a Bayesian approach, estimating model parameters which correspond directly to the parameters of the biological question we posed. The key computational goal of this chapter is to relax the infinitesimal assumption and simultaneously perform hyperparameter estimation and sparse regression coefficient estimation in the Bayesian model.

4.2 Methods

4.2.1 Model specification

We fit a multivariate generalized linear model, leaving aside for now the choice of the choice of likelihood $p(y \mid X, \theta)$ or equivalently, the link function (Figure 4.2). Recall that a GLM describes the expected response y given observed covariates X, through some (possibly non-linear) link function g:

$$\mathbb{E}\left[y\right] = g^{-1}(X\theta)$$

For example, for linear regression $g^{-1}(x) = x$, and for logistic regression $g^{-1}(x) = 1/(1+exp(-x))$. Although for linear regression we can derive a closed form for the maximum likelihood estimator of θ, in general the MLE must be found using an algorithm called iterative reweighted least squares. In this chapter, we will instead take a Bayesian approach to fitting the GLM, estimating the posterior distribution of the parameters θ, as well as the posterior distribution of the hyperparameters controlling the prior on θ (described below).

Our ultimate goal is to model binary phenotypes, since we are primarily interested in modeling human disease. For linear regression, centering the predictors and response is typical practice; however, this preprocessing does not make sense for binary phenotypes. Instead, we fit an additional intercept term, yielding the GLM $\mathbb{E}\left[y\right] = g^{-1}(X\theta + \theta_0)$. Of note, we do not follow the typical practice of including the intercept in the design matrix X as a column of entries equal to one, due to our use of a specialized prior on θ which does not apply to θ_0.

Specifically, we impose the spike and slab prior (point-normal mixture) on the regression coefficients. Intuitively, under the prior effects are either exactly equal to 0 (the spike), or come from a Gaussian distribution (the slab). This prior distribution regularizes the regression, making it possible to fit the model for $p \gg n$ as in GWAS, and also reflects biologically relevant prior information that only a fraction of variants have non-zero effect. The key idea of our model is to generalize to a group spike and slab (GSS) prior, allowing groups of predictors to have different levels of sparsity, which will be reflected in the posterior distribution of the hyperparameters. In order to represent our GSS prior, we introduce a vector of indicator variables z_j which denote the causal status of each variant. Intuitively, if $z_j = 0$, then the effect size $\theta_j = 0$ with probability 1; otherwise, θ_j is assumed to be generated from the Gaussian distribution.

We additionally impose hyperpriors on the spike and slab distributions which depend on the annotations A_j of variant j. The idea is that the annotation of the variants affects both the level of sparsity of non-zero coefficients within the annotated group variants, as well as the variance of the Gaussian distribution which the non-zero coefficients are drawn from (intuitively, the average effect size).

$$\begin{aligned}
X &= [x_{ij}]_{n \times p} \\
y &= (y_1, \ldots, y_n) \\
A &= [a_{jk}]_{p \times m} \\
p(\cdot) &= p(y \mid x, \theta) \prod_j p(\theta_j \mid z_j, A_j \tau) p(z_j \mid A_j \pi) \\
p(\theta_j \mid z_j = 1, \tau) &= N(0, A_j \tau^{-1}) \\
p(\theta_j \mid z_j = 0, \tau) &= 0 \\
p(z_j \mid \pi) &= B(A_j \pi)
\end{aligned}$$

Here, $N(\cdot, \cdot)$ denotes the Gaussian density parameterized by mean and precision (inverse variance) and $B(\cdot)$ denotes the Bernoulli density.

We make several critical modeling assumptions:

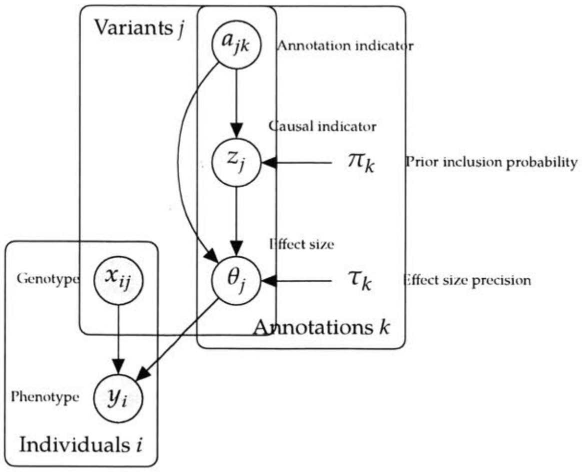


Figure 4.2: Graphical model representation of the proposed model

1. We assume the grouping of variables is known, and infer the parameters of the group-specific slab distributions. Specifically, we assume the annotation matrix A contains a one-hot encoding of annotations $\{1 \ldots, m\}$. In other words, A consists of binary entries, where each variant has exactly one annotation. As we described in Chapters 2-3, we have constructed such biological annotations, grouping together genetic variants which we have prior belief are coherently functional in disease and motivating this assumption.

   We noted previously that annotations which are typically studied overlap, and the typical approach to handle overlapping annotations in this class of models is to re-parameterize the prior inclusion probability as another generalized linear model $logit(\pi_j) = A_j w$, where the weights w_k denote the log-odds increase in the variant being causal given it is in annotation k.

   However, this re-parameterization makes the computation more difficult, and doesn't make sense in terms of the biological goals of the specific motivating question of this chapter. The main computational challenge in fitting this re-parameterized model is handling the resulting dense matrix A, which has order of magnitude comparable to X. For example, the 226 enhancer modules which we constructed summarize order 10^3 ChIP-Seq and DNAse-Seq datasets, and we might fit the model on order 10^4 individuals. But the annotations and the genotypes are observed at the same number of genetic variants, essentially increasing the order of magnitude of the total problem size.

   Biologically, our interpretation of the enhancer modules is that the thousands of different epigenomic assays are noisy observations generated by some latent state. That latent state corresponds to the functional elements of the genome and their activities in each of the different human cell types and tissues, and naturally we want to perform inference on this latent state rather than the noisy observations.

   Conceptually, reparameterizing the model $logit(\pi_j) = A_j w$ corresponds to simultaneously learning the latent state w while learning a predictive model for disease. However, in this parameterization the latent state conflates the functional role of different parts of the genome with the relevance of those functional roles in the disease being modeled. Therefore, we would instead prefer to learn a latent state A and then annotate variants using those latent states before incorporating the phenotype association information.

2. We allow each group to have a distinct level of sparsity, generalizing previous work which assumes each group of predictor was either completely included or excluded from the model[87]. In essence, prior work put a multivariate Gaussian prior on each group of regression coefficients (equivalent to ridge regression), but assigned an indicator variable to each group of predictors, and put a single hyperparameter controlling the prior number of groups with non-zero coefficients. Here, we assign an indicator to each variable, and use a vector of hyperparameters to control the prior number of coefficients within each group with non-zero value.

3. We assume the hyperprior factorizes as $p(\pi, \tau) = p(\pi)p(\tau \mid \pi, h^2)$. Prior work proposed this dependent prior distribution in order to constrain the model to sparse solutions which are consistent with heritability[88], using the parameterization

$$h_k^2 \approx \pi_k \tau_k \sum_{j \in [k]} \mathbb{V}\left[X_j\right]$$

4. We allow each group to have a distinct slab precision hyperparameter, generalizing prior work[89]. Sharing the slab precision τ between all causal variants can be done based on a subjective prior derived from real GWAS data, as in prior work, or based on modifying the equation in (3). When sharing τ, we have

$$h^2 \approx \tau \sum_k \pi_k \sum_j A_{jk} \mathbb{V}\left[X_j\right]$$

Our inference task is to estimate the full posterior distribution of π and τ given the data:

$$\begin{aligned} p(\pi, \tau \mid x, y, a) &\propto p(x, y, a \mid \pi, \tau) p(\pi, \tau) \\ &= \iint p(y \mid x, \theta) \prod_j p(\theta_j \mid z_j, A_j \tau) p(z_j \mid A_j \pi) \, dz \, d\theta \, p(\pi, \tau) \end{aligned}$$

However, this problem is intractable for a number of reasons:

1. The prior is non-conjugate to the likelihood, so the posterior $\prod_j p(\theta_j \mid z_j, A_j \tau) p(z_j \mid A_j \pi)$ cannot be computed analytically, and requires time exponential in the number of features p to estimate numerically

2. Accurately estimating estimating model evidences or marginals of the components of the hyperparameter vectors requires a number of samples exponential in the number of annotations m.

We address these issues as follows:

1. We fix h_k^2 to its moment estimate using PCGC regression[90]. This approximation is justified by the empirical observation that for a variety of simulated genetic architectures the moment estimate accurately captures the true value. In this case, τ can be derived from (π, h^2), eliminating half of the hyperparameters. For fixed π, we have $\tau_k = (1 - h_k^2) \pi_k \sum_j \mathbb{V}\left[X_j\right] / h_k^2$.

2. We replace the intractable model evidence term $p(X, y, A \mid \pi, \tau)$ with the best lower bound obtained by a mean-field variational approximation. The key idea is to approximate $p(\theta, z \mid X, y, A, \pi, \tau)$, which could have complicated correlations between θ and z, with a different distribution $q(\theta, z \mid \alpha, \beta, \gamma)$ where all variables are mutually independent.

   We find this approximating distribution using Variational Bayes (VB), which recasts the problem as an optimization problem. The objective function of this optimization is a lower bound on the model evidence, which we can use in place of the model evidence.

3. We explore two techniques to efficiently estimate the posterior distribution of the hyperparameters: Bayesian quadrature (BQ) and Stochastic Gradient Variational Bayes (SGVB).

   In BQ, we perform Bayesian inference inference over the model evidence as a function of the hyperparameters[91]. The key idea is to impose a Gaussian process (GP) prior on the model evidence itself. We can then express the mean and variance of the evidence as expectations over the GP, and actively pick the next point in hyperparameter space to evaluate. We can then normalize the active samples and use them as an approximation of the target posterior $p(\pi, \tau \mid \cdot)$.

   In SGVB, we simultaneously find the best approximating $q(\theta, z \mid \cdot)$ and the best approximating $q(\pi, \tau \mid \cdot)$ by re-parameterizing the objective function and performing stochastic gradient ascent[92]. We can then recover expectations of the hyperparameters as simple closed forms over the parameters of the best approximating variational distribution.

4.2.2 Variational Bayes

Recall that we seek to estimate $p(x, y, a \mid \pi, \tau)$. The key challenge in estimating this distribution is that we need to marginalize over the model parameters (θ, z). In the process of sampling to estimate the model evidence, we can additionally estimate the posterior distribution $p(\theta_j \mid z_j, A_j \tau) p(z_j \mid A_j \pi)$, which is needed to actually use the fitted model to identify specific genetic variants and predict phenotypes from genotypes. Although estimating $p(\theta_j \mid z_j, A_j \tau) p(z_j \mid A_j \pi)$ and performing the associated inference tasks are not the primary goal of this chapter, they are vital to evaluating the quality of the fitted models.

The typical Markov Chain Monte Carlo (MCMC) approach for this problem is to draw samples $(\theta^{(s)}, z^{(s)})$ and evaluate the joint likelihood of the model for each sample. However, the number of possible values of z is exponential in the number of genetic variants p making a naive MCMC approach infeasible.

Prior work developed sophisticated sampling schemes to efficiently explore the parameter space, avoiding the large number of configurations z which are unlikely to have high posterior probability. Intuitively, if most variants are non-causal then most configurations z contain many non-causal variants which should have zero effect on phenotype. One method which has been proposed to avoid configurations which are unlikely to contribute to the posterior density is to rank the variants by their univariate regression z-scores (equivalently, p-values), and use the Metropolis-Hastings algorithm to produce a new configuration given the current configuration[88,93]. The typical MH updates are to add a new predictor to the configuration (with probability proportional to its univariate z-scores) or removing a predictor from the configuration.

However, these sampling schemes have widely varying convergence properties on real problems, motivating our use of a different approximate Bayesian inference algorithm[94]. Rather than approximating some intractable integral over (θ, z), we would instead like to find some surrogate distribution $q(\theta, z \mid \cdot)$ which is easy to integrate over. Ideally, we would like closed form approximate solutions for $p(\theta_j \mid z_j, A_j\tau)p(z_j \mid A_j\pi)$. Now, we replace the problem of approximating an intractable integral with finding an optimal surrogate distribution.

A natural definition of "optimal" surrogate distribution is one which is "close" to the target posterior distribution $p(\theta_j \mid z_j, A_j\tau)p(z_j \mid A_j\pi)$. One natural measure of the distance between two probability distribution is the Kullback-Leibler (KL) divergence. There are two main approaches in the field for finding optimal surrogate distributions in KL divergence: expectation propagation (EP)[95] and Variational Bayes (VB). Mathematically, the two methods are minimizing $KL(p\|q)$ and $KL(q\|p)$, respectively. However, the KL divergence is not a proper distance measure, and so these two divergences are not necessarily equal. There is a deep connection between these two methods: the two divergences are actually special cases of the α-divergence[96].

Conceptually, both methods seek to find a surrogate distribution which has high density where the target posterior has high density (Figure 4.3). However, the key difference between the two methods is what they find as the optimal surrogate to a multi-modal target posterior. The key feature of VB (which differentiates it from EP) which we exploit here is that the optimal surrogate distribution should have zero density where the target posterior has zero density. This feature means that VB picks one mode of the multi-modal posterior, which is important for our problem due to the correlation structure between the columns of X.

Therefore, in this chapter we reformulate the problem of inferring the posterior $p(\theta_j \mid z_j, A_j\tau)p(z_j \mid A_j\pi)$ to minimizing the Kullback-Leibler divergence between the mean-field approximation $q(\theta, z \mid \alpha, \beta, \gamma)$ and $p(\theta_j \mid z_j, A_j\tau)p(z_j \mid A_j\pi)$.

$$
\begin{aligned}
q(\theta, z \mid \alpha, \beta, \gamma) &= \prod_j q(\theta_j \mid z_j, \beta_j, \gamma_j) q(z_j \mid \alpha_j) \\
q(\theta_j \mid z_j = 1, \beta_j, \gamma_j) &= N(\beta_j, \gamma_j^{-1}) \\
q(\theta_j \mid z_j = 0, \beta_j, \gamma_j) &= 0 \\
q(z_j \mid \alpha_j) &= \alpha_j
\end{aligned}
$$

Our algorithm must actually solve two coupled inference problems:

1. For each sampled (π, τ), estimate $p(X, y, A \mid \pi, \tau)p(\pi, \tau)$
2. Find the best approximation $q(\theta, z \mid \cdot)$, and use the optimal value of the objective function to bound $p(X, y, A \mid \pi, \tau)$

After normalizing the solution from (1), we can recover expectations and variances of $q(\theta, z \mid X, y, A)$ by taking weighted averages over the weighted samples (π, τ).

In this chapter, we will apply new algorithms for each of these problems (Bayesian quadrature and Stochastic Gradient Variational Bayes, respectively), and compare them to the prior work (Importance sampling and coordinate ascent, respectively). We abbreviate each of these combinations BQ-CA, BQ-SGVB, IS-CA, IS-SGVB.

Under the mean-field approximation, the parameters (θ, z) are mutually independent, which means that the approximate posterior means and variances have closed-form solutions:

$$\begin{aligned}\mathbb{E}_q\left[\theta_j\right] &= \alpha_j \beta_j \\ \mathbb{V}_q\left[\theta_j\right] &= \alpha_j \gamma_j^{-1} + \alpha_j(1-\alpha_j)\beta_j^2\end{aligned}$$

These analytical solutions are the main reason to use the variational approximation. Assuming we can estimate the optimal parameters α, β, γ, posterior inference is trivial. However, we know there are true correlations in the predictors, so we need to demonstrate through simulation that the mean-field approximation does not bias the results.

Prior work derived analytical expression for the KL-divergence in the case of Gaussian models (where the posterior distribution can be derived analytically even though the prior is non-conjugate). Further, prior work derived an approximate analytical expression for the ELBO for the case of binomial models (logistic regression)[75]. From these analytical expression, prior work derived coordinate ascent updates to solve the optimization problem. These coordinate ascent updates are straightforward to generalize to the case where π, τ are multivariate. We refer to this method as IS-CA.

$$\begin{aligned}\gamma_j &= \frac{(X'X)_{jj} + A_j \tau}{\sigma^2} \\ \beta_j &= \frac{1}{\gamma_j \sigma^2}\left((X'y)_j - \sum_{k \neq j}(X'X)_{jk}\alpha_j \beta_j\right) \\ \frac{\alpha_j}{1-\alpha_j} &= \frac{A_j \pi}{1 - A_j \pi}\frac{A_j \tau}{\gamma \sigma} exp(\beta_j^2 \gamma_j^2)\end{aligned}$$

4.2.3 Bayesian quadrature

The most obvious way to estimate the posterior distribution $p(\pi, \tau \mid x, y, a)$ is importance sampling. From Bayes rule,

$$p(\pi, \tau \mid x, y, a) \propto p(x, y, a \mid \pi, \tau) p(\pi, \tau)$$

Therefore, we can estimate the left-hand side by proposing values of (π, τ), evaluating the right-hand side, and then re-normalizing the discrete set of evaluated values. Conceptually, we can represent a probability distribution as a collection of sampled realizations drawn from the distribution, and as the number of samples increases to infinity, the histogram of samples will approach the true density function of the distribution.

Prior work used this strategy by proposing (π, τ) on a uniform grid in one dimension, which we refer to as IS. The challenge is that this grid approach does not scale to high dimensional hyperparameter spaces. Intuitively, if we put our samples on a uniform grid, we require exponentially many samples to achieve equal resolution on each dimension of the hyperparameter space. The problem is that the typical posterior distribution of the hyperparameters for this problem is highly spiked: there is a small region of high density, and most of the hyperparameter space has zero density.

We instead sought to use active learning to efficiently explore the hyperparameter space. We assume that h^2 is known; therefore, the hyperparameters (π, τ) are completely determined by π and we have $p(x, y, a \mid \pi, \tau) = f(\pi)$. The key idea is take a Bayesian approach and model the function f, putting a prior on it, and using the posterior distribution of f to find the next point to evaluate. Here, we will focus on the problem of modeling and integrating over f, so we will simplify notation by defining $x = logit(\pi)$. Then, the model evidence is:

$$Z = \int f(x)p(x)dx$$

As described above, we are unable to derive an analytical form for the function f, and evaluating $f(x)$ requires optimizing a variational objective. Taking a Bayesian approach, we write a generative model which generates ordered pairs $D = (x_i, f(x_i))$ from a latent function f. We assume a Gaussian process (GP) prior on the latent function f and integrate over the uncertainty in the latent function values. This approach, termed Bayesian quadrature[98,99], has two key features:

1. The model evidence Z itself is a random variable, so we can estimate its posterior mean and posterior variance. By using a GP prior on f with the squared exponential covariance function (described below) and a Gaussian prior on x, we can derive analytical solutions for the posterior mean and variance[100].
2. We can estimate the level of uncertainty in the function given our current observations, and use this information to actively sample the next point in hyperparamater space to evaluate. For an appropriate definition of uncertainty (described below), this strategy automatically balances exploitation (sampling where the function value is estimated to be large) and exploration (sampling where uncertainty is large).

A GP is a stochastic process parameterized by a mean function $m(x)$ and covariance function $K(x, x')$ which generates functions as finite sets of ordered pairs $(x_i, y_i = f(x_i))$ with the following property[101]:

$$y \sim \mathcal{N}(m(x), K(x, x'))$$

The kernel function K controls the properties of the latent function f. Here, we choose the squared exponential covariance function (in other settings, also known as the radial basis function kernel)

$$K(x, x') = \lambda^2 exp\left\{-\frac{1}{2}(x - x')'A^{-1}(x - x')\right\}$$

$$\begin{aligned}
\mathbb{E}\,[Z] &= \iint f(\Theta)p(f)df\; p(\Theta)d\Theta \\
f &\sim GP(0, K(\cdot, \cdot)) \\
\mathbb{V}\,[Z] &= \iint K(x, x')p(x)p(x')dx\,dx'
\end{aligned}$$

There are two challenges in applying BQ to model probability densities: (1) densities do not have mean 0, and (2) densities are non-negative, which GPs do not capture. These properties can give rise to nonsense results such as negative model evidences[100], motivating a more recent approach called Warped Sequential Approximate Bayesian Inference (WSABI)[91]. The idea of WSABI is to assume a GP prior on $\sqrt{2f}$ and instead perform inference on the resulting chi-square process prior on f. The chi-square process is itself intractable, requiring a Taylor expansion around $m_{f|D}$, the posterior mean of the GP given the observed function values.

We make several critical choices in our use of WSABI:

1. We use the first-order Taylor approximation (called WSABI-L), which has the property that the posterior mean of the latent function outside of the range of observed samples observed is 0. In our preliminary simulation experiments, we found that the target function is unimodal and highly spiked, and that therefore this approximation is correct.

2. We use the squared exponential kernel with a single lengthscale parameter (isotropic covariance for $m > 1$). In our preliminary simulation experiments, we found that the target function is indeed isotropic about the mode, even in the case where predictors are correlated to each other (using real genotypes).

3. We constrain the lengthscale parameter to avoid overfitting the GP when optimizing the GP hyperparameters with Type II maximum likelihood. In our experiments, unconstrained optimization often leads to degenerate solutions where lengthscales go to infinity, leading to nonsense estimates. In our preliminary simulations, we found that typical lengthscales were less than 1 on the logit-scale.

The posterior mean and variance of the model evidence have closed form solutions:

$$
\begin{aligned}
z_i &= \lambda^2 |B|^{-1/2} exp\{-1/2(x_i - b)'(A+B)^{-1}(x_i - b)\} \\
w_i &= \lambda^4 |2A^{-1}B + I|^{-1/2} exp\{-1/2(x_i - b)'(A/2+B)^{-1}(x_i - b)\} \\
E[Z] &= w'(K^{-1}x)^2 \\
\Sigma &= A^{-1} + (A+B)^{-1} \\
\Lambda &= \Sigma^{-1} + (A^{-1} + B^{-1})^{-1} \\
V[Z] &= \lambda^2 z'z|(2A^{-1} + B)\Lambda|^{-1/2} - w'K^{-1}w
\end{aligned}
$$

In principle, we could use the estimated $\mathbb{E}\ [Z]$ to normalize the distribution $p(x, y, a \mid \pi, \tau)p(\pi, \tau)$ and compute an expectation directly over the target distribution $p(\pi, \tau \mid x, y, a) = \frac{1}{Z}p(x, y, a \mid \pi, \tau)p(\pi, \tau)$.

$$
E[g(x)] = \iint g(x)\frac{f(x)}{Z}p(x)df dx
$$

However, in our preliminary experiments this method does not work due to numerical scaling problems. We instead take a Monte Carlo integral over the active samples, which we refer to as WSABI.

$$
\begin{aligned}
E[x] &\approx w_i / \textstyle\sum_i w_i \\
w_i &= p(x, y, a \mid \pi, \tau)
\end{aligned}
$$

4.2.4 Modeling ascertained data

To model binary phenotypes, we use logistic regression. However, this is not the typical generative model for binary phenotypes in human genetics. Usually we assume the liability threshold model generates the phenotypes (see Section 4.2.6). The fundamental reason we use logistic regression here is that in GWAS, samples are ascertained to achieve balanced case-control design.

Ascertainment is known to cause model mis-specification for generalized linear models of binary responses. Specifically, the data is generated from the retrospective likelihood $p(x \mid y, \theta)$; however, we typically analyze the data as though it arose from the prospective likelihood $p(y \mid x, \theta)$ for two reasons. First, standard GLM algorithms assume the data are described the prospective likelihood. Second, we write our generative models in terms of the prospective likelihood. In other words, we model how the response variable (disease) is generated from the predictors (genetic variants, environmental exposures, etc.).

Of course, we could incorporate the prior distribution $p(x)$ to translate between the two likelihoods:

$$p(y \mid x, \theta) = \frac{p(x \mid y, \theta)}{p(x)}$$

However, the prior distribution $p(x)$ is intractable for our problem because it describes the prior probability of observing each genotype vector. In order to represent this prior, we require a multinomial distribution which has number of values exponential in the number of predictors p. A priori we actually know that most configurations of x are not observed in the population; however, even with this simplifying assumption, representing the prior is extremely challenging. We could in principle use an improper prior $p(x) = 1$ to simplify the inference, but it is unclear what impact this improper prior would have on the posterior inference when it clearly does not describe the data.

The maximum likelihood estimates of probit regression (closely related to the liability threshold model) are known to be biased under this model mis-specification. However, MLEs for logistic regression coefficients (except the intercept) are unbiased[102]. A more recent idea which generalizes this proof to the Bayesian case is the Poisson-multinomial transformation[103]. Using this argument, Bayesian posteriors with respect to $p(y \mid x, \theta)$ are equivalent to posteriors with respect to $p(x \mid y, \theta)$, after marginalizing over nuisance parameters with appropriate priors[104].

The key insight is that for logistic regression, the retrospective likelihood is a product of multinomial likelihoods and the prospective likelihood is a product of binomial likelihoods. Both are parameterized in terms of the log odds ratio of the predictors, and we can recover both likelihoods as special cases of a Poisson likelihood (hence Poisson-multinomial transform). In the MLE case, we obtain the prospective/retrospective likelihood by maximizing nuisance parameters in Poisson; for the Bayesian case we integrate over nuisance parameters. In either case, the only difference is the order in which we handle nuisance parameters, proving that the two are equivalent.

For the purposes of estimating posterior inclusion probabilities and regression effect sizes, this result suggests that applying any algorithm to fit logistic regression to ascertained data is sufficient. However, in order to estimate the posterior distribution of the hyperparameters of interest in this chapter, or to perform Bayesian model averaging over the hyperparameters to estimate the PIP and regression effect sizes, we need to accurately estimate the model evidence $p(x, y, a \mid \pi, \tau)$, which means we need to accurately estimate the intercept and the likelihood.

In order to get an unbiased estimator of the intercept, the key result we need is that generalized linear models are closed under ascertainment[105]. Intuitively, if we introduce a variable S_i indicating that sample i was included, ascertainment changes $p(y_i = 1 \mid X_i, S_i\backslash)$, in such a way that the change can be written in terms of $p(y_i = 1 \mid X_i)$ as a modified link function, which admits the same algorithms.

$$\begin{aligned} E[y_i \mid x_i, S] &= h^{-1}\left(E[y_i \mid x_i]\right) \\ &= h^{-1}(g^{-1}(\eta_i)) \end{aligned}$$

Specifically, we assume that the case-control study is totally ascertained, such that $p(s_i = 1 \mid y_i = 1) = 1$ and $p(s_i = 1 \mid y_i = 0) = \frac{K(1-P)}{P(1-K)}$, where K is the prevalence of the phenotype in the population and P is the study proportion of cases. Then,

$$\begin{aligned} h(x) &= \frac{x}{r - (r-1)x} \\ r &= \frac{P(1-K)}{K(1-P)} \end{aligned}$$

4.2.5 Stochastic Gradient Variational Bayes

As described above, prior work developed an analytical lower bound to the evidence lower bound for logistic regression models, which could be optimized using coordinate ascent. Rather than deriving new coordinate ascent updates by hand to incorporate the modified link function above, we used an approach known as Stochastic Gradient Variational Bayes[92]. This approach was independently developed as Doubly Stochastic Variational Inference[106] and stochastic backpropagation[107]. The key advantage of this approach is that the same generic algorithm can be used to combine any likelihood function of interest with our group spike-and-slab prior prior with no changes to the core inference algorithm. We outline the algorithm below, and then describe how to incorporate the modified link function described above.

Recall that in optimizing the variational objective, we cannot write an analytical expression for the KL-divergence. We instead need to solve a dual problem: minimizing the KL-divergence is equivalent to maximizing the evidence lower bound (ELBO) $\mathcal{L}$:

$$\begin{aligned} \mathcal{L} &= E_{q(\theta_j \mid z_j, \beta_j, \gamma_j) q(z_j \mid \alpha_j)}[ln\, p(\cdot)] - E_{q(\cdot)}[ln\, q(\cdot)] \\ &= E_{q(\theta_j \mid z_j, \beta_j, \gamma_j) q(z_j \mid \alpha_j)}[ln\, p(y \mid X, \theta)] - \mathcal{KL}(q(\theta, z \mid \Psi) \| p(\theta, z \mid X, y, a)) \end{aligned}$$

Now, the objective function consists of a negative reconstruction error $E_{q(\theta_j|z_j,\beta_j,\gamma_j)q(z_j|\alpha_j)}[ln\, p(y \mid X, \theta)]$ and a regularizer $\mathcal{KL}(q(\theta, z \mid \Psi) \| p(\theta, z \mid X, y, a))$. Intuitively, these two parts of the objective function trade off moving regression coefficients away from zero to explain the data (and reduce the negative reconstruction error), and moving coefficients towards the prior (to reduce the KL divergence penalty). If the prior is strong enough, then most coefficients will go to 0, except for those coefficients which are required to explain the data.

For our choice of prior and surrogate family, we can derive an analytical expression for the regularizer (generalizing prior work):

$$
\begin{aligned}
\mathbb{E}_{q(\theta_j|z_j,\beta_j,\gamma_j)q(z_j|\alpha_j)}\left[ln\, p(\theta_j \mid z_j, A_j\tau)\right] &= \mathbb{E}_{q(z_j|\alpha_j)}\left[\mathbb{E}_{q(\theta_j|z_j,\beta_j,\gamma_j)}\left[ln\, p(\theta_j \mid z_j, A_j\tau)\right] \mid z\right] \\
&= \alpha_j \mathbb{E}_{q(\theta_j|z_j=1)}\left[ln\, p(\theta_j \mid z_j, A_j\tau)\right] \\
&= \frac{\alpha_j}{2}\left(ln\, \tau - ln\, c - \tau \mathbb{E}_{q(\theta_j|z_j=1)}\left[\theta_j^2\right]\right) \\
\mathbb{E}_{q(\theta_j|z_j=1)}\left[\theta_j^2\right] &= V_{q(\cdot)}\theta + \mathbb{E}_{q(\cdot)}\left[\theta \mid z = 1\right]^2 \\
&= \gamma_j^{-1} + \beta^2 \\
\mathbb{E}_{q(z_j|\alpha_j)}\left[ln\, p(z_j \mid A_j\pi)\right] &= \alpha_j\, ln\, \pi_{a_j} + (1 - \alpha_j)\, ln(1 - \pi_{a_j}) \\
\mathbb{E}_{q(\theta_j|z_j,\beta_j,\gamma_j)}\left[ln\, q(\theta_j \mid z_j, \beta_j, \gamma_j)\right] &= \frac{\alpha_j}{2}\left(ln\, \gamma_j - ln\, c - \gamma_j \mathbb{E}_{q(\theta_j|z_j=1)}\left[(\theta_j - \beta_j)^2\right]\right) \\
\mathbb{E}_{q(\theta_j|z_j=1)}\left[(\theta_j - \beta_j)^2\right] &= \gamma_j^{-1} \\
H(\cdot) &= \frac{1}{2}\sum_j \alpha_j \left(1 + ln\, \tau - ln\, \gamma_j - \tau(\gamma_j^{-1} + \beta^2)\right) \\
&\quad - \sum_j \left(a_j\, ln\left(\frac{a_j}{\pi_{a_j}}\right) + (1 - \alpha_j)\, ln\left(\frac{1 - \alpha_j}{1 - \pi_{a_j}}\right)\right)
\end{aligned}
$$

We require the parameters α to lie in the interval $[0, 1]$ and the parameters γ to be non-negative. Therefore, we re-parameterize them in terms of unconstrained variables and appropriate link functions which map the real line to the required co-domain. We chose the sigmoid and softplus nonlinearities due to their empirical convergence properties in small-scale experiments.

$$
\begin{aligned}
\alpha &= (1 + exp(-\tilde{\alpha}))^{-1} \\
\gamma &= \gamma_{\min} + log(1 + exp(\tilde{\gamma}))
\end{aligned}
$$

The challenge is that the reconstruction error does not have an analytical form for binomial likelihoods (logistic regression), so we cannot even write down the objective function, never mind optimize it. The problem is that the likelihood involves the sigmoid function, which does not have an analytical integral. The key idea of SGVB is to re-parameterize the objective function to make it differentiable. Intuitively, we replace the expectation with a Monte Carlo estimator which is a polynomial; however, this estimator is still not differentiable because of the dependency between the samples $\theta^{(s)}$ and the distribution q we are taking the expectation over.

$$\mathbb{E}_{q(\cdot)}\left[F(\cdot)\right] \approx S^{-1} \sum_{s=1}^{S} ln\, p(y \mid x, \theta^{(s)}); \theta^{((s))} \sim q(\cdot)$$

The key insight is that for a wide class of distributions, we can re-write sampling from the distribution q as some transformation of samples from a standard distribution. Of particular importance, if we want to sample from some Gaussian distribution $\theta \sim N(\mu, \nu)$, we can sample $\epsilon \sim N(0,1)$ and transform $\theta^{(s)} = \mu + \epsilon^{(s)}\sqrt{\nu}$. With this re-parameterization, we can take gradients of the reconstruction error with respect to the variational parameters because we can backpropagate (apply the chain rule) through $\partial\theta/\partial\mu$ and $\partial\theta/\partial\nu$, treating the noise vector ϵ as constants. We automatically compute the gradients of $\mathcal{L}$ using Theano[108,109] and perform stochastic gradient ascent to optimize the objective function. We tuned the learning rate (fixed at 0.001) and number of full batch training epochs (by default, 1,000) on our small-scale simulations.

To efficiently estimate the stochastic gradient in each step, we make another re-parameterization, which has been previously studied in the context of variational auto-encoders[110].

$$\begin{aligned}
\eta_i &= \sum_j X_{ij}\theta_j \\
\mathbb{E}_{q(\cdot)}\left[\eta_i\right] &= \sum_j X_{ij}\mathbb{E}_{q(\cdot)}\left[\theta_j\right] = \sum_j X_{ij}\alpha_j\beta_j \\
\mathbb{V}_{q(\cdot)}\left[\eta_i\right] &= \sum_j X_{ij}^2\mathbb{V}_{q(\cdot)}\left[\theta_j\right] \\
\zeta^{(s)} &\sim N(0, I_n) \\
\eta_i^{(s)} &= \mathbb{E}_{q(\cdot)}\left[\eta_i\right] + \zeta^{(s)}\sqrt{\mathbb{V}_{q(\cdot)}\left[\theta_j\right]}
\end{aligned}$$

This local re-parameterization improves the computational cost of optimizing the variational objective for two reasons: (1) We avoid sampling from $q(\theta_j \mid z_j, \beta_j, \gamma_j)q(z_j \mid \alpha_j)$ (which scales linearly in p), and (2) the resulting stochastic gradient estimator has lower variance, reducing the convergence time.

Incorporating the modified link function from the previous section is trivial in SGVB: we simply modify the symbolic expression representing $E_q[ln\, p(\cdot)]$ and automatically backpropagate. In principle, we could use this same $h(\cdot)$ to fit a probit regression to the data; however, typically in genetics we use the liability threshold model. The main difference between the liability threshold model and the probit model that the decision boundary is at some non-zero value of the latent liability; this means that the likelihood is described by truncated Gaussian distributions. Therefore, we do not explore liability threshold models in this study.

4.2.6 Simulation study

The generative model for phenotypes is the liability threshold model:

$$\begin{aligned} y_i &= \mathbb{1}(l_i > \Phi^{-1}(K)) \\ l_i &= \sum_j X_{ij}\theta_j + \epsilon_i \\ \theta_j \mid z_j = 1 &\sim N(0, \tau_{a_j}^{-1}) \\ \epsilon_i &\sim N\left(0, \left(\frac{1}{h^2} - 1\right) \sum_j 2f_j(1 - f_j)\theta_j^2\right) \\ f_j &\sim U(0.01, 0.5) \end{aligned}$$

Here, $U(\cdot, \cdot)$ denotes the uniform density. We make several important choices in the specification of this generative model:

1. We do not normalize genotypes to have variance one. Therefore, effect sizes are in units of one allele substitution rather than one standard deviation of alleles. This choice equates to a prior assumption that effect size is independent of minor allele frequency (MAF). This is a strong assumption which might lead to unrealistic simulations, andsimulating realistic effect size and MAF distributions should be explored in future work.

2. The population value of the genetic variance (based on the sampled MAFs and effect sizes) is used to determine the residual variance. This is opposed to using the realized value of the genetic variance (the sample variance $\hat{\mathbb{V}}\ [X\theta]$) or the expected genetic variance (in the one-component case, the PVE by construction).

3. The variant parameters f, a, and θ are fixed, but the genotypes X are random, allowing us to plug in different sampling schemes for X.

To speed up sampling synthetic case-control genotypes for (1), we adapted the simCC algorithm[90] to relax the assumptions as in our generative model. Under the assumption of linkage equilibrium, genotypes can be drawn i.i.d. sequentially with no additional auxiliary storage (i.e., this is a state space model where transitions are fixed). We used this algorithm to quickly sample from the distribution of genotypes conditioned on binary phenotype.

$$\begin{aligned} p(x_1, \dots, x_j \mid y) &\propto p(y \mid x_1, \dots, x_j)p(x_1, \dots, x_j) \\ p(y \mid x_1, \dots, x_j) &= N(x_{1..j}\theta, V[y] - V[x_{j..p}\theta]) \\ p(x_1, \dots, x_j) &= p(x_j \mid x_1, \dots, x_{j-1}) \\ p(y \mid l) &= p(l < t), y = 0; p(l > t), y = 1 \end{aligned}$$

To evaluate the accuracy of our method over simulation replicates, we bootstrapped the entire simulation, generating new MAFs, causal variants, causal effect sizes, genotypes (for simulations with synthetic genotypes), and phenotypes.

4.3 Results

4.3.1 Hierarchical model of genetic architectures

In order to infer the parameters of the genetic architectures of complex traits and distinguish between different non-infinitesimal architectures, we propose a hierarchical model regressing phenotype on genotype. We relax the infinitesimal assumption and instead assume a spike and slab prior (point-normal mixture) on the regression coefficients[75]. Our choice of prior distribution regularizes the regression, making it possible to fit the model directly for $p \gg n$, and also reflects biologically relevant prior information that only a fraction of variants have non-zero effect. In order to represent this prior, we introduce a vector of indicators z which denote the causal status of each variant.

We assume hyperpriors on the hyperparameters of the spike and slab distributions, controlling the prior probability of each variant being causal π_k and the prior effect size precision of causal variants τ_k depending on the annotation a_{jk} of variant j. We assume annotations are non-overlapping (for each j, there is only one non-zero a_{jk}) and that they can independently affect both the probability a variant is causal and its effect size. However, we assume that the hyperpriors $p(\pi), p(\tau)$ are dependent, in order to be consistent with the estimated partitioned heritability h^2. Our inference task is to estimate the posterior distribution of (π, τ) given the observed data.

4.3.2 One group problems

We first considered the simplest case, where all variants are in one group, and asked whether the model could accurately estimate the posterior mean of the scalar hyperparameters π, τ. We simulated synthetic genotypes in an idealized scenario where all causal variants are observed and irrelevant variants are in linkage equilibrium with the causal variants. We simulated problems with $p = 10000$, fixed 1% of the variants to be causal, and generated Gaussian phenotypes with heritability 20%. We then applied IS-CA to these problem instances and surprisingly found that the method underestimated the posterior mean proportion of causal variants π (Figure 4.4). The reason is that IS-CA is highly sensitive to the samples which are actually chosen. We compared the result to IS-CA on a hand-tuned grid which included the true value of π which generated the data as one of the samples, and found that the model was properly calibrated (Figure 4.5).

Rather than hand-tuning the grid, which is problem instance-specific, we sought to learn the optimal samples from the data. This strategy will also allow us to efficiently generalize the approach to multiple dimensions, which would require an exponential number of samples in the grid search approach. We based our approach on an active learning scheme called Warped Sequential Approximate Bayesian Inference (WSABI). The key idea is to model the likelihood as an unknown function, putting a Gaussian Process prior on the function, and then exploiting the posterior distribution of the unknown function values given the observed function values. We use this model to pick the optimal next hyperparameter setting to fix and evaluate the evidence lower bound.

We first sought to understand the performance of this method on the single group problem and and found that the method performed reasonably well, but still underestimated the posterior mean proportion of causal variants π. Worryingly, the number of samples required to accurately estimate π from the data appeared to grow linearly in the number of variants p.

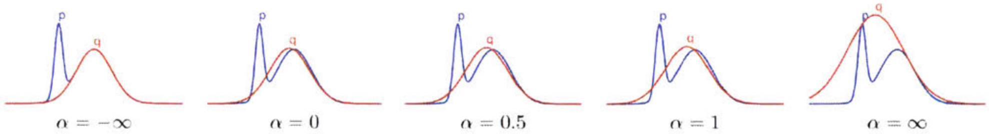


Figure 4.3: Toy example approximating distributions q for a complicated posterior p minimizing α-divergence. Variational Bayes corresponds to $\alpha = 0$ and Expectation Propagation corresponds to $\alpha = 1$. Reproduced from Minka, 2005[97].

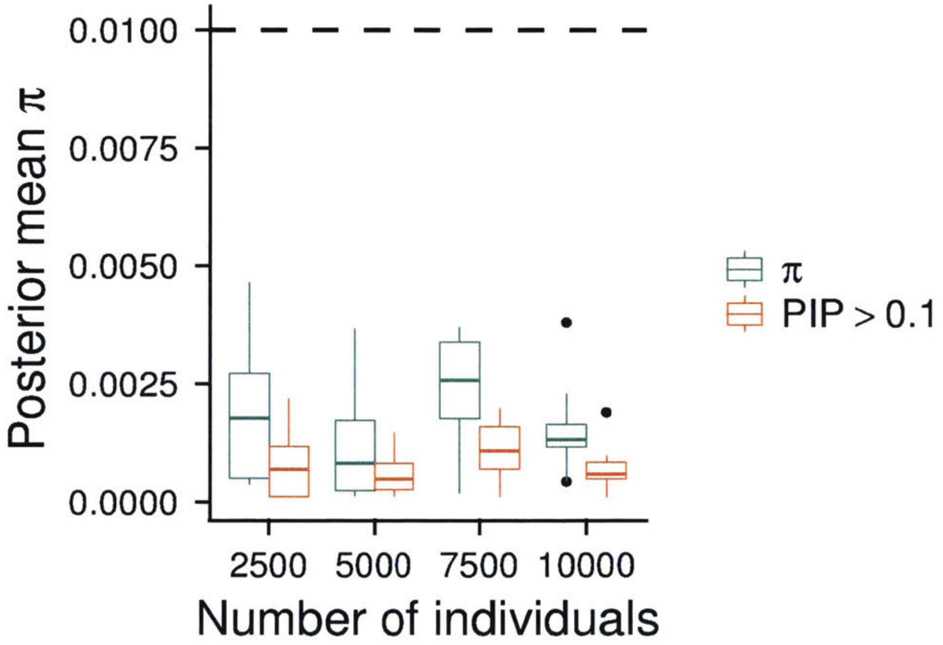


Figure 4.4: Estimated posterior mean π and number of selected variables (posterior inclusion probability > 0.1) for linear regression model on Gaussian phenotypes under the idealized scenario using importance sampling.

To investigate why the method systematically underestimated *pi* for $n < p$, we traced the evolution of the active samples for an example simulation. We found that the method quickly learned the correct shape of the likelihood (Figure 4.7), which is highly spiked around the true answer $\pi = 0.01; logit(\pi) \approx -2$. However, the active sampler concentrated samples around the spike, in order to accurately characterize its boundary. We relied upon the simple Monte Carlo estimator to compute the expectation of π, and as we showed in the case of grid search, the accuracy of the estimator depends on which samples were actually chosen.

We next investigated the performance of the method on binary phenotypes. In prior work, IS-CA used coordinate ascent to fit a variational lower bound to the evidence lower bound for logistic regression. We propose an alternative method, IS-SGVB, which can directly optimize the evidence lower bound, and also incorporate a correction for ascertainment of binary phenotypes.

We simulated binary phenotypes from our model under a non-infinitesimal architecture in which 10% of the variants are causal (have a non-zero effect), holding the total PVE fixed at $h^2 = 0.2$. We first fixed the population prevalence at 0.5 and sampled a balanced number of cases and controls, such that there was no ascertainment. As before, we found that IS-SGVB underestimated the number of causal variants (Figure 4.8). The estimated proportion tended to be consistent with the number of variants which had high posterior inclusion probability (PIP), but with non-negligible error (Figure 4.9).

In all of these simulations, we found that the model failed to find many true causal variants. There are a number of reasons why the model might fail to assign high PIP to true causal variants:

1. The variant doesn't explain enough genetic variance. In this case, the model regularizes the coefficient to the prior (essentially, π) and explains the variance using the error term (for logistic regression, the Bernoulli likelihood)

2. The variant doesn't have a large enough effect size. This is closely related to the first point, as the variance explained by the jth variant is $\mathbb{V}\left[X_j\theta_j\right] = \theta_j^2\mathbb{V}\left[X_j\right] = 2f_j(1-f_j)\theta_j^2$. However, even for cases where the MAF f_j is high enough, the causal effect size might be too small.

   One open question, which is beyond the scope of this thesis, is how to generate combinations of MAF and causal effects which are plausible. Conceptually, realistically simulating the minor allele frequency and effect size of causal variants for disease requires forward simulating the evolution of a population of genomes[111], but this is infeasible at scale.

   We could try to use an empirical procedure to estimate the distribution of plausible effect sizes based on GWAS; however, this strategy has two shortcomings: effect sizes are likely to be smaller than estimated in GWAS (due to the winner's curse) and most small effect sizes have not been reported as GWAS associations. Future work should investigate whether the combinations which are chosen in simplified sampling schemes like the one proposed in this chapter are plausible, and whether the case where the model fails are plausible.

3. With insufficient samples, we might not observe significant differences in MAF between disease cases and controls at the causal variants just due to finite sampling.

4. With insufficient samples, we might observe spurious correlations between causal variants due to finite sampling which might cause the model to regularize correlated variables to the prior.

We investigated each of the four possibilities for the simulation trial shown in Figure 4.9, which failed to recover the correct answer. Surprisingly, we found that none of these explanations completely explained how the the model failed to find one of the causal variants. For example, we observed a scenario where the model was able to find causal variant #8, which explained small proportion of genetic variance, but not #7, which in fact explained more variance.

Of note, SGVB failed to converge in a number of simulation trials due to the sensitivity of SGD to parameters like the learning rate, minibatch size, and iteration count. In the case of simulated data where we know the correct answer, we can of course tune these parameters by hand to get the correct answer. However, even without knowing the real answer, we found in our experiments that we could detect failure to convergence by inspecting the variational surrogate to the posterior inclusion probability (PIP) α_j for irrelevant variables.

In order to understand how we could detect failure to converge from the estimated PIPs, we traced the evolution of the PIPs in SGVB. We initialize $\alpha_j = 0.5$, and during the course of the optimization, the objective function pulls $\alpha_j \rightarrow 1$ for relevant variables and $\alpha_j \rightarrow \pi$ for irrelevant variables. After taking an expectation over the hyperparameter samples (via grid search or active sampling), the result is that the PIP tends to sharply classify variants: either the PIP is close to 1, or it is close to 0. Therefore, even without knowing the true answer, we can inspect the estimated PIPs to detect whether the model converged. In our preliminary experiments, we compared a number of settings for these parameters, and found that the optimal setting was not only specific to the problem size, but also the particular problem instance.

It would be desirable to automatically detect failure to converge, and automatically tune learning parameters, which has been the topic of recent methodological development. The most important class of methods for tuning learning parameters automatically in a data-driven fashion is Bayesian optimization (BO). The key idea of BO is to model the performance of a fitted model (as quantified by some loss function) as a function of the learning parameters, put a GP prior on this function, and then use the GP to actively sample a point which is likely to improve the current optimum value observed.

BO is closely related to Bayesian quadrature (BQ), which we use in this chapter to perform Bayesian inference on the hyperparameters of our hierarchical model. Both methods model an unknown function as a Gaussian Process posterior over samples, and actively sample to find the optimal next point for the inference task; however, The fundamental difference is that BO seeks to find a local minimum of the unknown function, and BQ seeks to integrate over the unknown function.

In principle, we could use BO to automatically tune the learning rate and number of full-batch epochs for our inference algorithm. The natural choice of loss function for binary classification is the log loss function:

$$\sum_i y_i \, log(\hat{y}_i) + (1 - y_i) \, log(\hat{y}_i)$$

We investigated the log loss of our fitted model for a particular example in which we knew the model did not converge to the correct answer based on our simulation (in Figure 4.8, the outlier at $n = p, \pi = 0.0175$). We first compared the fitted model parameters to the ground truth simulation parameters and found that the model did not sufficiently regularize irrelevant variables, leading to incorrect estimation of the lower bound and therefore incorrect estimation of the target posterior. When we examined the training and validation log loss of the model, we found the model actually seemed to underfit the data (Table 4.1).

Minibatch size	Learning rate	Epochs	Training log loss	Validation log loss	Posterior mean π	#(PIP > 0.1)	Mean non-causal PIP
100	1×10^{-3}	1000	0.681	0.671	0.018	1	0.21
100	1×10^{-3}	4000	0.686	0.675	0.018	2	0.20
100	5×10^{-3}	4000	0.685	0.673	0.018	1	0.10
100	1×10^{-2}	4000	0.678	0.677	0.001	3	0.05
100	1×10^{-2}	6000	0.689	0.691	0.001	1	0.11

Table 4.1: Impact of learning parameters on model performance and estimates for a specific case. Top panel reflects default values of the training parameters tuned on one small-scale simulation problem; bottom panel reflects l_1-regularized logistic regression. Validation log loss is computed on an independent sample of 500 cases and 500 controls. Non-causal variants are assumed to be known (specified in the simulation). The optimal model based on posterior mean calibration (bold) is not the model with best validation log loss. PIP threshold 0.1 is uncalibrated with respect to false discovery rate, but corresponds to the assumption that when the model is fully regularized, the PIP of irrelevant variables is equal to the prior.

Based on the observation that the model was not sufficiently regularized, we manually tuned the learning rate and maximum number of epochs to produce a fitted model which more accurately estimated the PIP (specifically, regularizing irrelevant variables to 0) and posterior mean π. However, we needed to use not only the validation log loss, but also the training loss and summaries of the fitted model parameters to determine acceptable hyperparameter settings. These results suggest that naive application of BO to automatically tune the learning procedure for this problem will not work. Future work should investigate the sensitivity of BO to the problem size for this class of dense high-dimensional regression problems, which is not typically studied in the field of machine learning.

Next, we investigated the effect of increasing case-control ascertainment on the inferred hyperparameters. We repeated the idealized simulation above, varying the population prevalence of the binary phenotype, and found that in most scenarios the correction did not improve the calibration of the posterior mean when compared to the same algorithm fit using a model which ignored the ascertainment (Figure 4.10). We additionally compared our method to the prior work which directly estimated a logistic regression without the additional correction for ascertainment (Figure 4.14). Of note, the estimated posterior was equal to the prior in many of the trials (as evidenced by posterior mean $\pi \approx 0.1$).

4.3.3 Two group problems

We generated synthetic annotations by dividing the synthetic genomes into two halves. We then sampled causal variants and effects under different non-infinitesimal architectures conditioned on the annotations, and generated Gaussian phenotypes according to a linear, additive model. For speed of the simulation, we simulated problems with $p = 1000$ variants. We considered two non-infinitesimal architectures: (1) equal effect sizes across groups, but one group had three times as many causal variants; (2) equal proportion of causal variants across groups, but one group had causal effects drawn from a Gaussian with twice the variance. In both cases, the ratios were chosen such that the partitioned heritability was $(0.05, 0.15)$ in expectation.

We used the active sampler to pick hyperparameter samples, and used coordinate ascent updates to optimize the variational objective for speed. For this idealized case, we found that the method could accurately estimate the proportion of causal variants and distinguish between the two architectures, but only when both groups had sufficient genetic variance (Figure 4.11). When a group had insufficient genetic variance, the model found a degenerate solution for that group ($\pi_k \rightarrow 0, \theta \rightarrow 0$). As we observed above, this degeneracy is due to the failure of the model to find the causal variants in the degenerate group.

We next investigated the effect of increasing the sample size on avoiding degeneracy. We simulated a simplified version of the problem, based on the following insight: the hyperparameters for each group of predictors are uncorrelated to each other by construction. Then, when the predictors themselves are also uncorrelated (therefore groups are uncorrelated), the posterior distributions of the effect sizes θ are also uncorrelated. This means we can consider just one group of predictors, and put the variance explained by the remaining predictors into the residual (which is assumed to be uncorrelated in GLMs).

With this insight, we simulated a simplified problem by fixing $n = 5000, p = 100$, taking real array genotypes to include a realistic level of co-linearity. As an aside, using synthetic genotypes would only change the rotation/scaling of the posterior contours: for uncorrelated variables, the shape of the posterior is isotropic. We put all of the predictors in one group, but simulated exactly one non-zero coefficient. We further fixed (π, τ) to their true values.

In order to estimate the posterior contours, we evaluated the posterior $p(\theta_1, \theta_2 \mid x, y, \pi, \tau)$ on a grid to estimate contours. As before, we need to marginalize over the parameter z; however, to estimate contours we only require the posterior up to a constant factor. Marginalizing over $(\theta_3, \dots, \theta_{100}, z_3, \dots, z_{100})$ yields a constant which does not depend on (θ_1, θ_2); therefore, we can simply ignore this constant. We still have to marginalize over (z_1, z_2); however, there are only four possible configurations of these variables so the estimation is straightforward. In this simplified simulation, we found that as $h^2 \to 0$, the prior dominates the likelihood in the posterior density, moving the posterior mean towards 0 (Figure 4.12).

The important challenge is that in real data partitioned by regulatory annotations, we are dealing with h^2 on the order of 0.01 or smaller. We investigated the change in the posterior mean θ as the sample size increased by fixing $h^2 = 0.01$ in the same simulation and bootstrapping individuals (sampling with replacement) to achieve large n. We found that to accurately estimate the posterior mean requires a number of individuals n only present in the largest extant data sets (Figure 4.13).

The consequence of this degeneracy is that the method might not detect extremely weak effects in real data sets. However, even in the case of degeneracy, the fitted model will still perform well in predicting the phenotype of an independently generated validation set, since the degenerate/incorrect part of the model did not explain much genetic variance.

4.3.4 Comparison to existing methods

The models proposed in this chapter are closely related to two existing approaches for estimating sparse regression effects in large-scale biological models: (1) Variational inference for Bayesian variable selection (which we referred to earlier as IS-CA)[75,89], and (2) Scalable Functional Bayesian Association (SFBA)[112].

There are a number of key differences between the models proposed here and VARBVS:

1. To optimize the intractable variational objective for logistic regression, SGVB performs stochastic gradient ascent. In contrast, VARBVS uses a second variational lower bound, and uses analytic gradients to derive coordinate ascent updates.
2. VARBVS fits standard logistic regression, marginalizing over the bias term, where SGVB incorporates a specialized GLM for ascertainment and fits a bias.
3. Here, we frame the fundamental inference problem as parameter estimation, where VARBVS frames the problem as model comparison.
4. VARBVS shares τ across all groups, using a subjective prior based on real GWAS data, where we use a dependent prior $p(\pi)p(\tau \mid \pi, h^2)$.
5. VARBVS uses a hand-tuned grid over hyperparameters, where WSABI uses active sampling.

The challenge in optimizing the variational objective for logistic regression is that the term $E_Q[ln\, P]$ is non-analytic. VARBVS uses a variational lower bound on $E_Q[ln\, P]$, which has an analytic form[113], and then takes analytic gradients to derive coordinate ascent updates for logistic regression. Coordinate ascent converges quickly because at every step it makes an optimal update; however, this property also makes it sensitive to the initial starting point of the optimization and also to the level of co-linearity in the problem. In practice, VARBVS solves each variational problem (for each candidate setting of the hyperparameters) twice: once to find an optimal initialization (over all candidates), and once to compute the optimum from a warm start. Moreover, coordinate ascent steps are generally too large in the setting of highly co-linear variables (e.g. imputed dosages), and so the convergence time for the algorithm dramatically increases. In contrast, SGVB relies on stochastic gradient descent, which requires more time per optimization but does not degrade in performance on highly co-linear problems.

As was the case for SGVB, VARBVS underestimated the posterior mean proportion of causal variants, while still correctly finding the causal variants based on the estimated posterior inclusion probabilities (Figure 4.14). However, the coordinate ascent algorithm converged more consistently than SGVB, suggesting further improvement needs to be made in the stochastic optimization. SGVB will still be of interest because we can use minibatches to reduce the space requirement for the optimization and improve the convergence time, and use GPU code generation in libraries such as Theano[108,109] to speed up the optimization.

We investigated the performance of SGVB when τ is shared across groups, modifying the hyperprior accordingly: when sharing τ, we have $h^2 \approx \tau \sum_k \pi_k \sum_j A_{jk} \mathbb{V}\left[X_j\right]$. Intuitively, sharing the effect size precision τ between groups corresponds to assuming that the effect size distributions are not very different between groups, such that a single precision parameter can adequately explain the data. As expected, this model can accurately estimate posterior means when the data is generated from the model, but cannot accurately estimate π when the model is mis-specified and causal effects are drawn from different distributions (Figure 4.15).

Of note, a model with shared slab precision can also avoid degenerate solutions in the case where one of the groups has very few causal variants which the model failed to find. Intuitively, sharing the slab precision between groups changes the plausible prior values of π, such that the setting of π_2 contributes the most to the posterior density.

Unlike the approach we proposed here, VARBVS performs model comparison in order to test whether annotations are enriched, defined as having more causal variants. In this chapter, we assumed that we already know that annotations are enriched (from a method such as the one proposed in Chapter 2), and focus on parameter estimation in order to explain the enrichment. The challenge in performing model comparison is efficiently estimating the marginal likelihood.

Our active sampling scheme is based on minimizing the posterior variance of the model evidence, which allows us to rapidly estimate Bayes factors. We performed a null simulation by simulating idealized data from the model with $m = 2$ but equal number of causal variants in the two groups with effect sizes drawn from the same distribution (the left-most case of Figure 4.11b). For speed, we used coordinate ascent on a Gaussian model. We found that in this null simulation, our model sometimes gave Bayes factors which would suggest enrichment where there was none (Table 4.2).

We inspected the fitted models for trial #2, where the model clearly overestimated the proportion of causal variants π_2 and found that the estimated Bayes factor was highly sensitive to the active samples chosen by our method. Indeed, restarting the active sampling with a different initialization lead to wildly different Bayes factors between 10^{-5} and 115.

From the true genetic variances which generated the data, it is clear that defining enrichment as different in proportion of causal variants is not sufficient. The particular non-infinitesimal architecture we simulated in trial #2 had the feature that one annotation truly had lower per-SNP heritability than the other annotation, due to differences in minor allele frequency (Figure 4.16). Fu-

Trial	Samples	Bayes factor	π_1	π_2	Training r^2	Validation r^2
1	16	6.94×10^{-6}	0.0154	0.0146	0.183	0.201
2	9	4.02	0.009	0.0406	0.228	0.156
3	12	3.03	0.00778	0.0106	0.204	0.148
4	15	6.47×10^{-6}	0.00137	0.0481	0.256	0.145
5	15	8.29	0.0238	0.00878	0.206	0.137
6	9	0.374	0.0284	0.0267	0.266	0.136
7	16	5.71×10^{-7}	0.018	0.0144	0.244	0.169
8	9	22.7	0.0345	0.0108	0.188	0.111
9	9	0.0786	0.0277	0.0338	0.238	0.151
10	13	0.00236	0.0266	0.0214	0.204	0.155

Table 4.2: Comparison of Bayes factors for a null simulation in the idealized case.

ture work should explore the power of the method to detect variables with high posterior inclusion probability as a function of minor allele frequency and effect size.

We then performed a positive simulation where $\pi_1 = 0.01$ and $\pi_2 = 0.03$ and computed Bayes factors of the alternate model with $k = 2$ against the null model $k = 1$. We found multiple cases where the Bayes factor was less than one, indicating the data were more likely to have been generated under the null model than the alternate model, even though we simulated problems where there was true enrichment (Table 4.3). Surprisingly, the training and validation score of the alternate and null models were nearly identical, even for cases where the Bayes factor was incorrect, suggesting that these were cases in which the model found a degenerate solution for one group. We examined the fitted Gaussian process model for the trials which failed and found that in many of the trials the fitted surface indeed reflected that the posterior was equal to the prior. For example, in trial #9, we found that conditioned on π_2, the alternate model likelihood was constant, leading to the marginal posterior of π_1 being equal to the marginal prior $p(\pi_1)$ (Figure 4.17).

Analogous to our generalization of VARBVS, SFBA generalizes Bayesian variable selection regression (BVSR)[88]. There are three key differences between SGVB and SFBA:

1. SGVB can fit local hyperparameters, where SFBA assumes hyperparameters are globally shared

2. SFBA assumes (π, τ) are independent, where SGVB (and BVSR) do not

3. SGVB uses the variational approximation where SFBA uses MCMC. In our notation, SFBA relies on a combination of local MCMC updates to the latent variables (θ, z) and global expectation-maximization updates to the hyperparameters (π, τ). In contrast, SGVB relies on variational inference over (θ, z) and active learning over (π, τ) in each local window. Our results show our approach can estimate the local hyperparameters and that local hyperparameters give fine-grained biological insights into disease.

In our preliminary experiments, SFBA failed to converge for our simplest simulated datasets due to sensitivity to the initialization and low Metropolis-Hastings acceptance probabilities. This results underscores the difficulty of the underlying inference problem, and the necessity for problem-specific tuning. Methods such as SFBA can accurately estimate genome-wide parameters at GWAS scale, but cannot be used unmodified to solve small sparse regression problems. In contrast, methods such as SGVB can accurately estimate model hyperparameters in small-scale experiments, but require tricky algorithmic modifications to operate on GWAS scale data. On the other hand, SGVB failed to detect causal variants in single chromosome analysis, necessitating further development to handle whole genome imputed dosage data.

4.4 Discussion

In this chapter, we proposed a hierarchical, large-scale sparse regression model to infer fundamental parameters of non-infinitesimal genetic architectures and an approximate Bayesian inference method. We showed through simulation that our model could estimate these parameters from observed data and distinguish between different explanations of observed enrichments. Recall that in Chapter 2, we used a heuristic to find a number of of relevant loci and attempted to identify enriched annotations, genes, and regulators without explicitly imposing parametric assumptions about the disease model or causal cell types. Here, we proposed a parametric approach where the

k	Trial	Bayes factor	π_1	π_2	Training r^2	Validation r^2
2	1	58.2	0.0159	0.0135	0.204	0.081
	2	0.00094	0.00811	0.0388	0.215	0.084
	3	1.88×10^{-7}	0.00782	0.0769	0.212	0.101
	4	0.000931	0.144	0.0794	0.202	0.143
	5	42	0.0279	0.0615	0.273	0.121
	6	47.6	0.0238	0.0325	0.255	0.100
	7	$5.84 \times 10^{+4}$	0.0155	0.0449	0.205	0.150
	8	4.48×10^{-10}	0.101	0.0678	0.209	0.121
	9	6.06×10^{-6}	0.0216	0.0496	0.235	0.109
	10	$2.73 \times 10^{+4}$	0.0425	0.0723	0.240	0.114
1	1				0.253	0.087
	2				0.246	0.100
	3				0.233	0.093
	4				0.212	0.096
	5				0.292	0.152
	6				0.284	0.128
	7				0.283	0.095
	8				0.164	0.140
	9				0.196	0.118
	10				0.250	0.090

Table 4.3: Estimated posterior means, training and validation set prediction performance, and Bayes factors for positive simulation $\pi_1 = 0.01, \pi_2 = 0.03$ in the idealized scenario. $k = 2$ denotes the alternate model, and $k = 1$ denotes the corresponding null model

parameters directly correspond to the biological question and framed the problem as parameter estimation rather than hypothesis testing.

Our choice to fix the proportion of variance explained by each component to its moment estimate reduces the computational burden of fitting the model and is motivated by the observation that the point estimates are empirically accurate in a variety of simulations. In principle, we could estimate the full posterior $p(\pi, \tau, h^2 \mid x, y, a)$ by integrating over a prior on h^2. However, our Bayesian quadrature approach relies on Type II maximum likelihood estimation of the GP hyperparameters, and in our experiments we found that it was prone to overfit the data unless we constrained the optimization. The key assumption underlying our constraint was that the typical lengthscale of the target posterior distribution is constant over all dimensions, which we verified for multi-dimensional $logit(\pi)$ in simulation. However, if we were to include components of $logit(h^2)$ as additional dimensions of the hyperparameter space, it is not obvious that this assumption still holds. One possible solution would be to marginalize over the GP hyperparameters; however, in general this strategy requires MCMC which should be explored in future work.

Our assumption that each variant has exactly one annotation simplifies the inference algorithm and is justified here by the construction of disjoint annotations (Chapter 2). Generalizing the algorithm to support overlapping annotations requires a re-parameterization, which is straightforward to do in our SGVB framework due to our use of automatic differentiation. However, conceptually we advocate instead for learning the latent state underlying the overlapping observations, and use these latent states (which are disjoint by construction) as more interpretable biological annotations.

Above, we described how re-parameterizing the model hyperparameters as a generalized linear model $logit(\pi_j) = A_j w$ corresponded to simultaneously learning a latent state w of the biological annotations and a latent state (θ, z) of the disease model. The conceptual challenge with this approach is that the latent state w conflates the functional role of different parts of the genome with the relevance of those functional roles in the disease being modeled. Instead, we really want to generate the latent states A, which should be independent of any disease. To connect this idea to the approach presented in this thesis, rather than using an ad-hoc approach combining a Hidden Markov Model and clustering to produce a set of clusters as proxies for latent functional classes, we should write a full generative model to learn them. Then, we can integrate samples from those latent classes into integrative models (like the class presented in this chapter) in order to make principled inferences over the biological states. Beyond that, it would be desirable to have a model which simultaneously described the enhancer regions, predicted active regulators, and downstream target genes for each enhancer module. The difference between such a model and the intersection of annotations which we used in Chapter 2 is the ability to quantify the uncertainty in the parameters of this model.

One promising extension of the work would be to use SGVB to simultaneously fit $q(\pi, \tau \mid \cdot)$ with $q(\theta, z \mid, \cdot)$. This would avoid exponential blow up in the number of annotations in the importance sampling approach, as well as problems of initialization and GP hyperparameter tuning in the active sampling approach. Our initial experiments suggest that although fitting this model is possible, the convergence time for the hyperparameters is much longer than for the parameters, likely due to the geometry of the non-convex objective function. Surprisingly, we found in our preliminary experiments that the fitted model finds an acceptable sparse solution (in terms of training loss, validation loss, and selected variables) long before the sparsity hyperparameter has converged to its true value, suggesting that the method could be used as a generic building block in larger generative models (Figure 4.4).

This approach leads to a computationally intractable problem due to the use of non-conjugate priors. An alternative approach with less computational burden would be to fit a regularized regression such as elastic net to the data. However, the main challenge to applying this approach to the problem of estimating the level of sparsity is that we cannot interpret the regularization penalty in the same way that we interpret the posterior inclusion probability in the Bayesian approach. We argue that the main reason to prefer the approximate Bayesian sparse regression method in place of a traditional regularization approach is that we can simultaneously estimate the parameters and their associated error bars.

In the case of regularized regression, we would have to perform post-hoc analysis of the fitted model to determine which coefficients are non-zero. Although statistical tests for coefficients selected by Lasso exist, for more general regularization schemes they do not. Then, the main challenge is that the post-hoc analysis would require non-parametric bootstrapping to perform hypothesis tests on each regression coefficient. This bootstrapping would defeat the performance gain achieved by not approximating an intractable model.

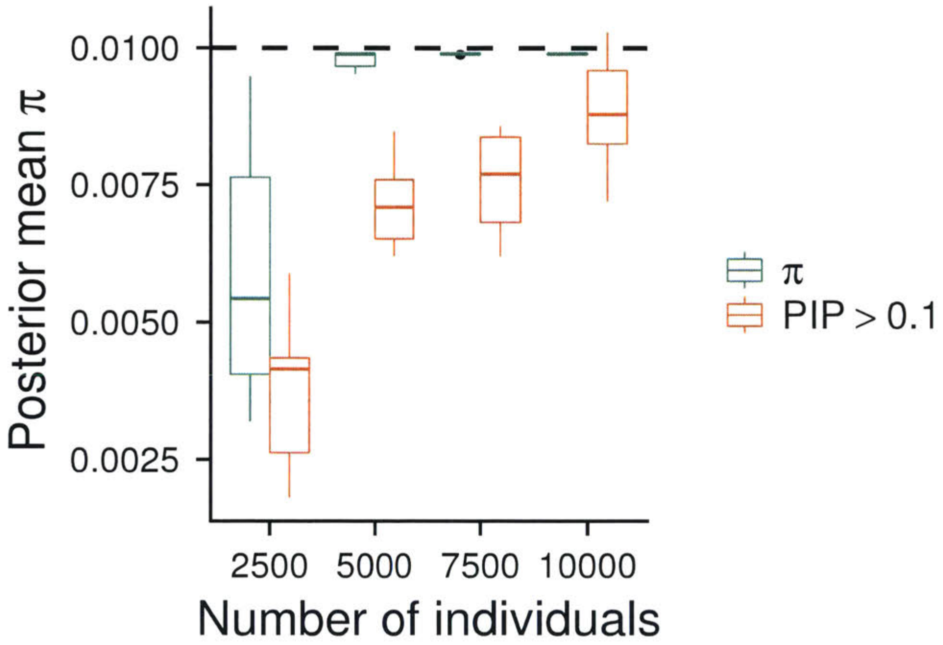


Figure 4.5: Estimated posterior mean π for linear regression model on Gaussian phenotypes under the idealized scenario using the reference implementation.

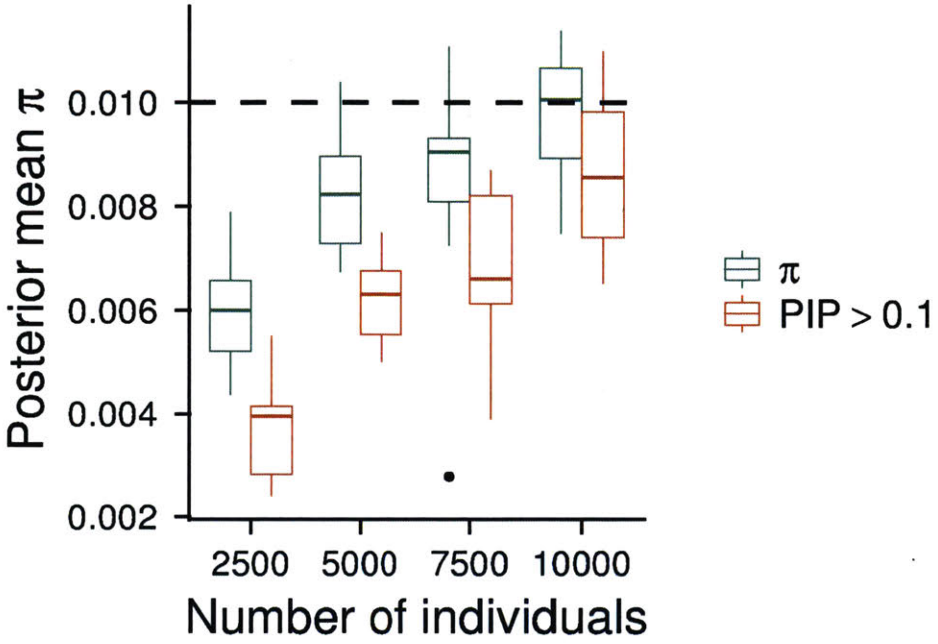


Figure 4.6: Estimated posterior mean π and number of selected variables (posterior inclusion probability > 0.1) for linear regression model on Gaussian phenotypes under the idealized scenario using active sampling.

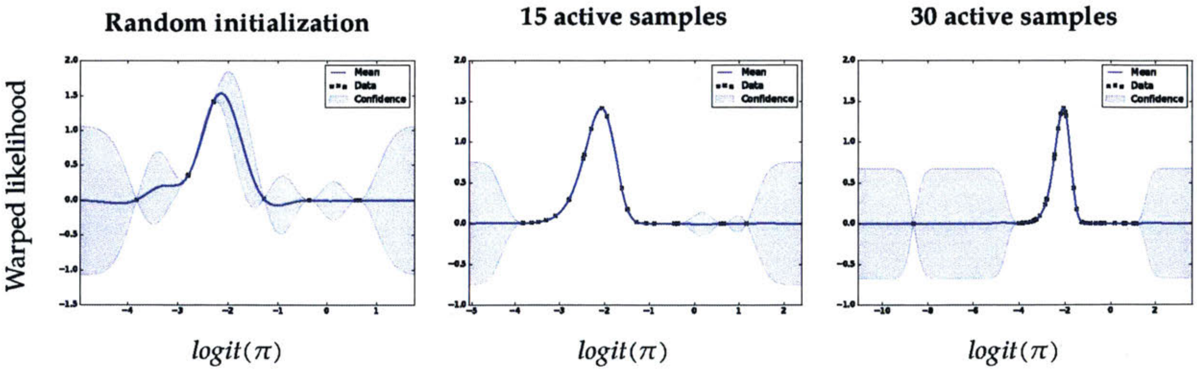


Figure 4.7: Example evolution of active sampling scheme on one group problem with proportion of causal variants 0.01

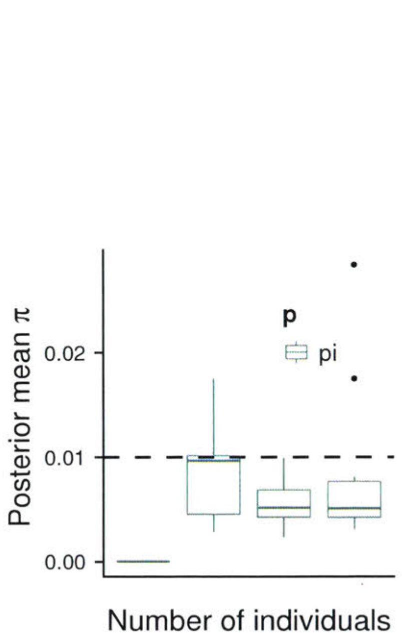


Figure 4.8: Estimated posterior mean π for logistic regression model on unascertained binary data under the idealized scenario.

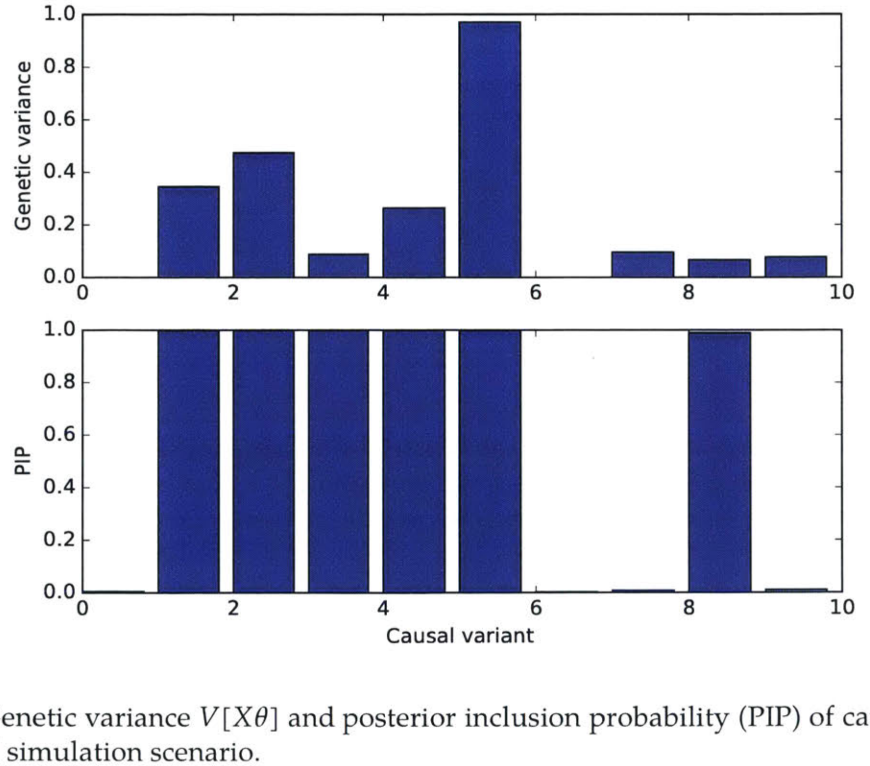


Figure 4.9: Genetic variance $V[X\theta]$ and posterior inclusion probability (PIP) of causal variants in the idealized simulation scenario.

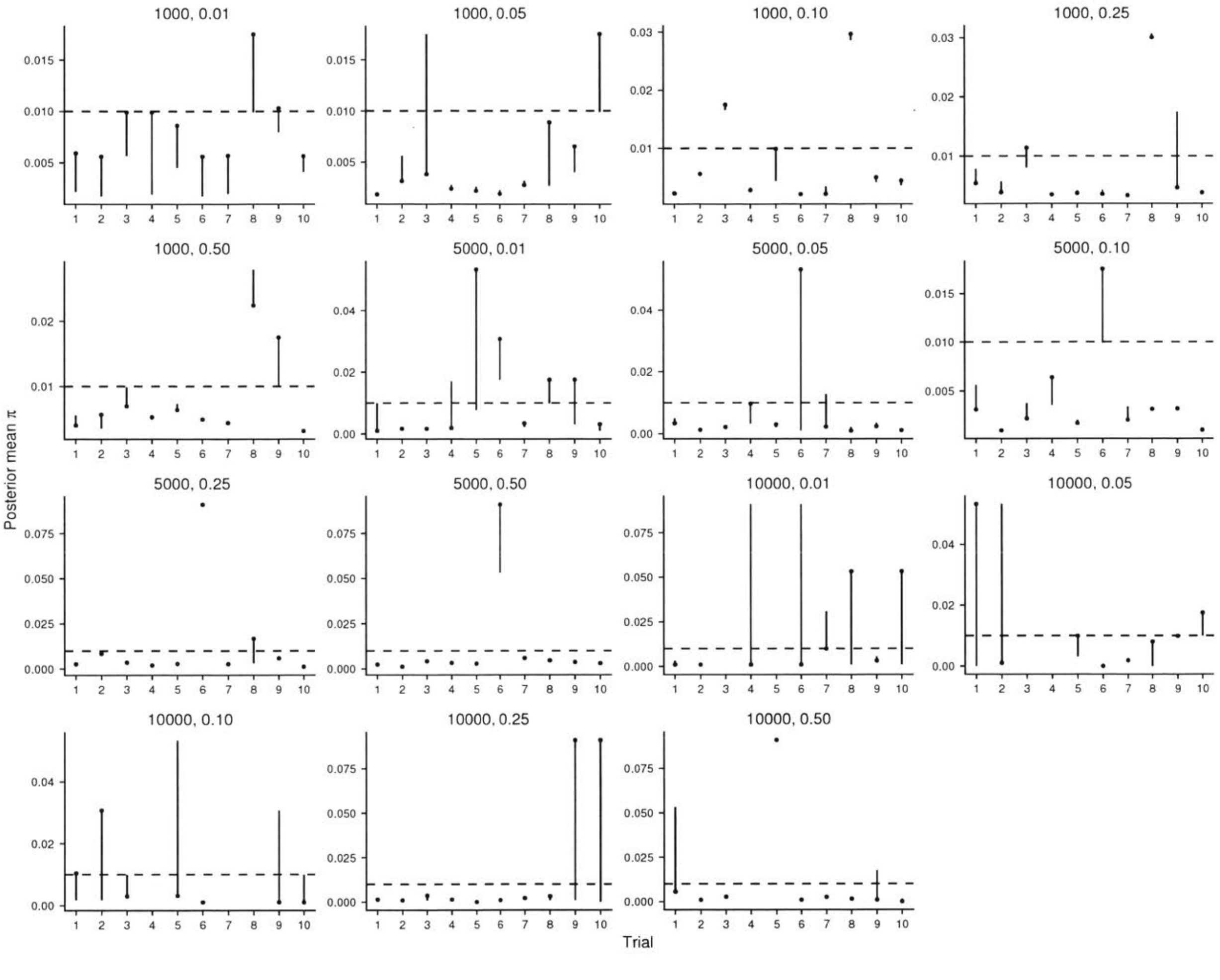


Figure 4.10: Estimated posterior mean π for logistic regression model on binary data under the idealized scenario, varying the population level prevalence. Estimates using the modified GLM are given as points, and estimates assuming the prevalence equals the study case proportion are given as offsets.

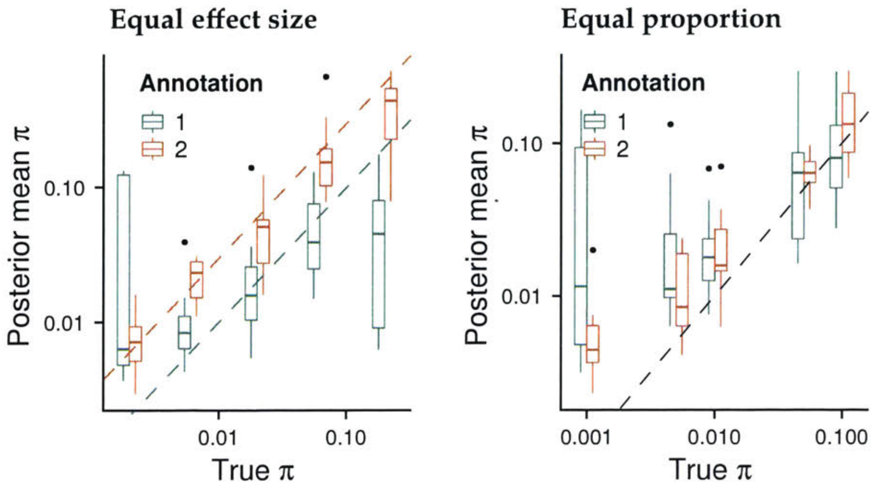


Figure 4.11: Estimated posterior mean π under two non-infinitesimal architectures

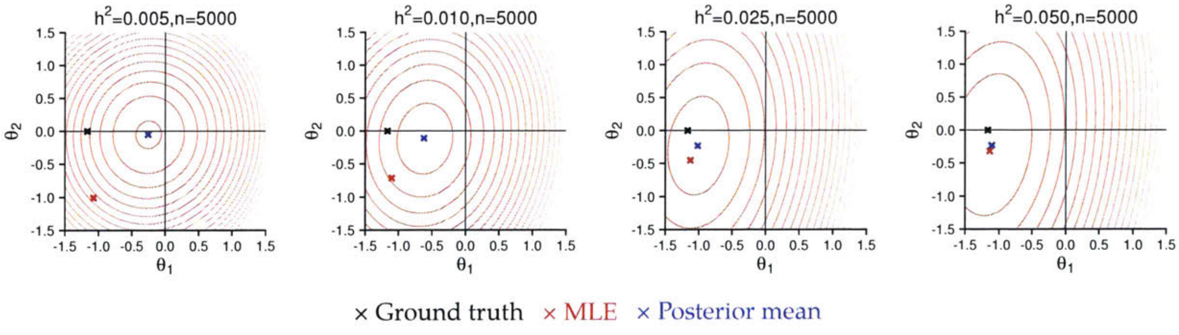


Figure 4.12: Posterior contours of θ as a function of h^2

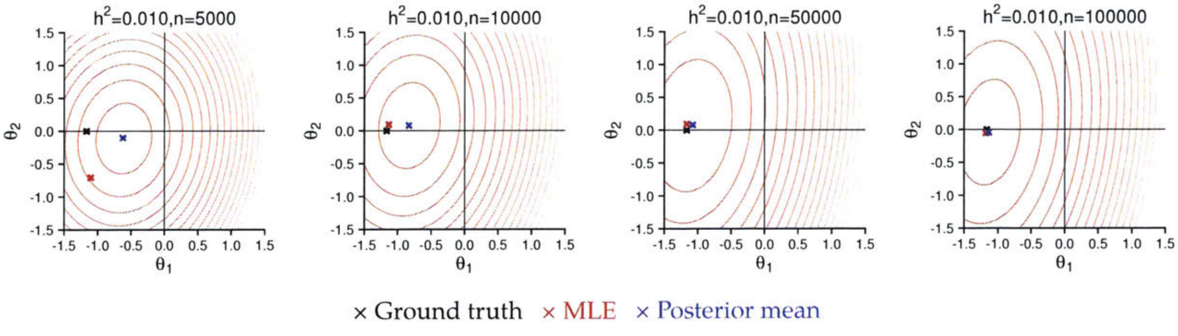


Figure 4.13: Posterior contours of θ as a function of n

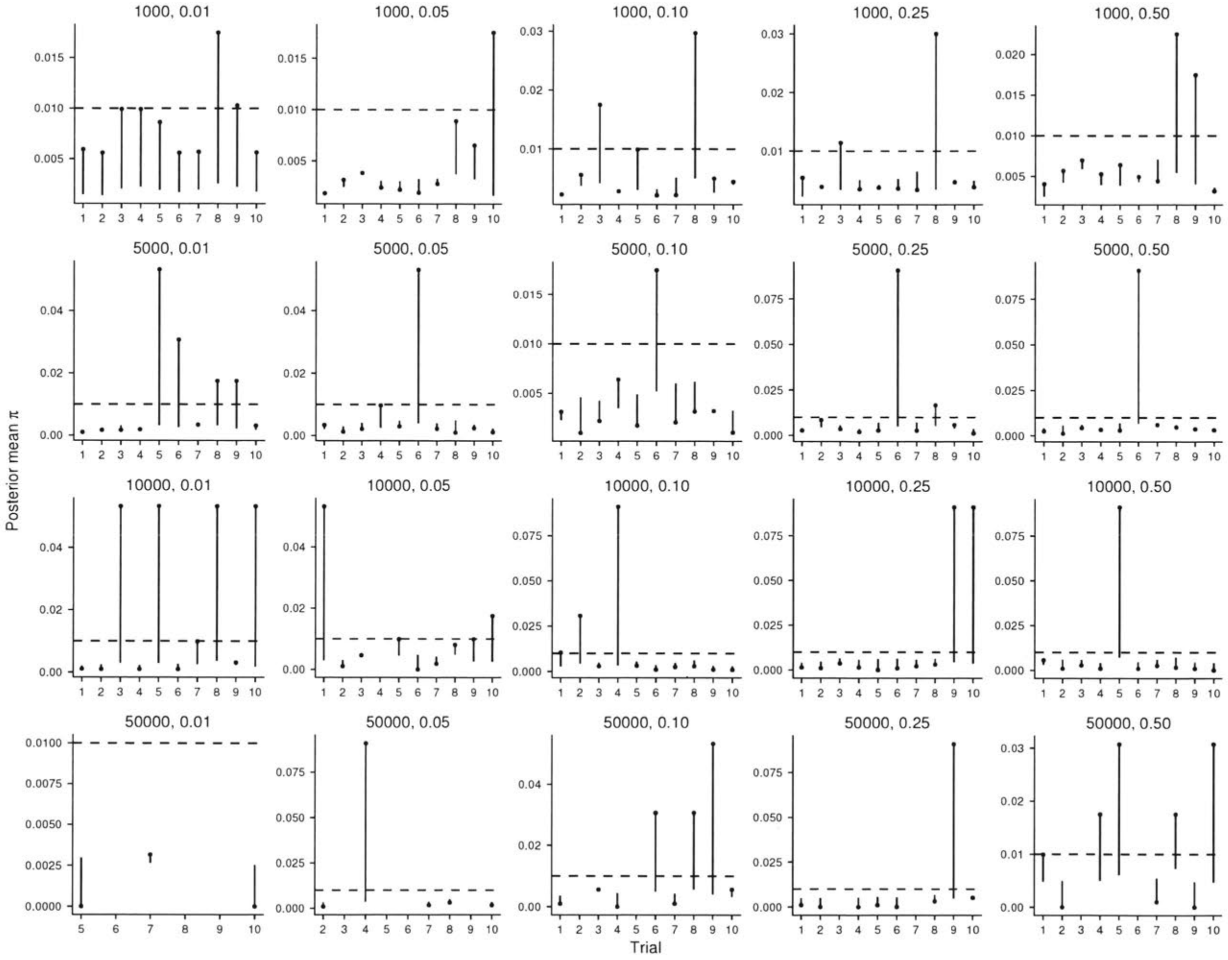


Figure 4.14: SGVB and VARBVS estimated posterior mean π for logistic regression model on binary data under the idealized scenario, varying the population level prevalence. SGVB estimates are given by points, and VARBVS estimates are given as an offset.

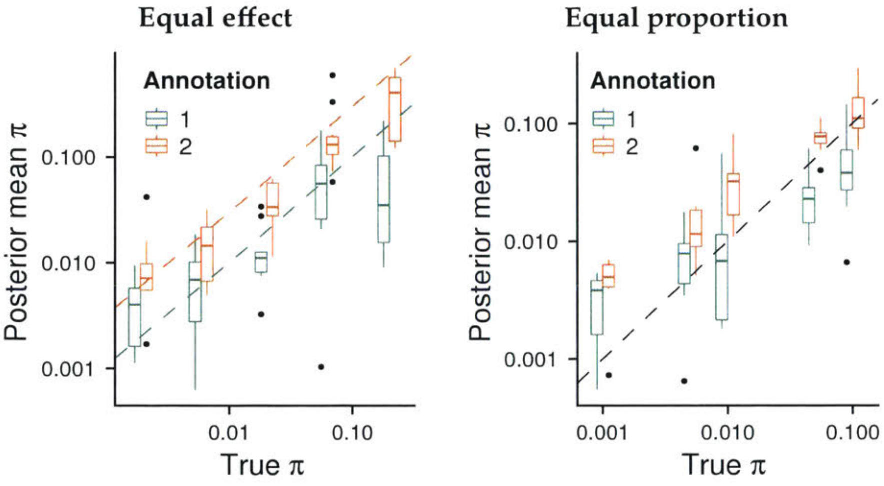


Figure 4.15: Posterior mean π for a model pooling τ across annotations under two non-infinitesimal architectures in the idealized scenario.

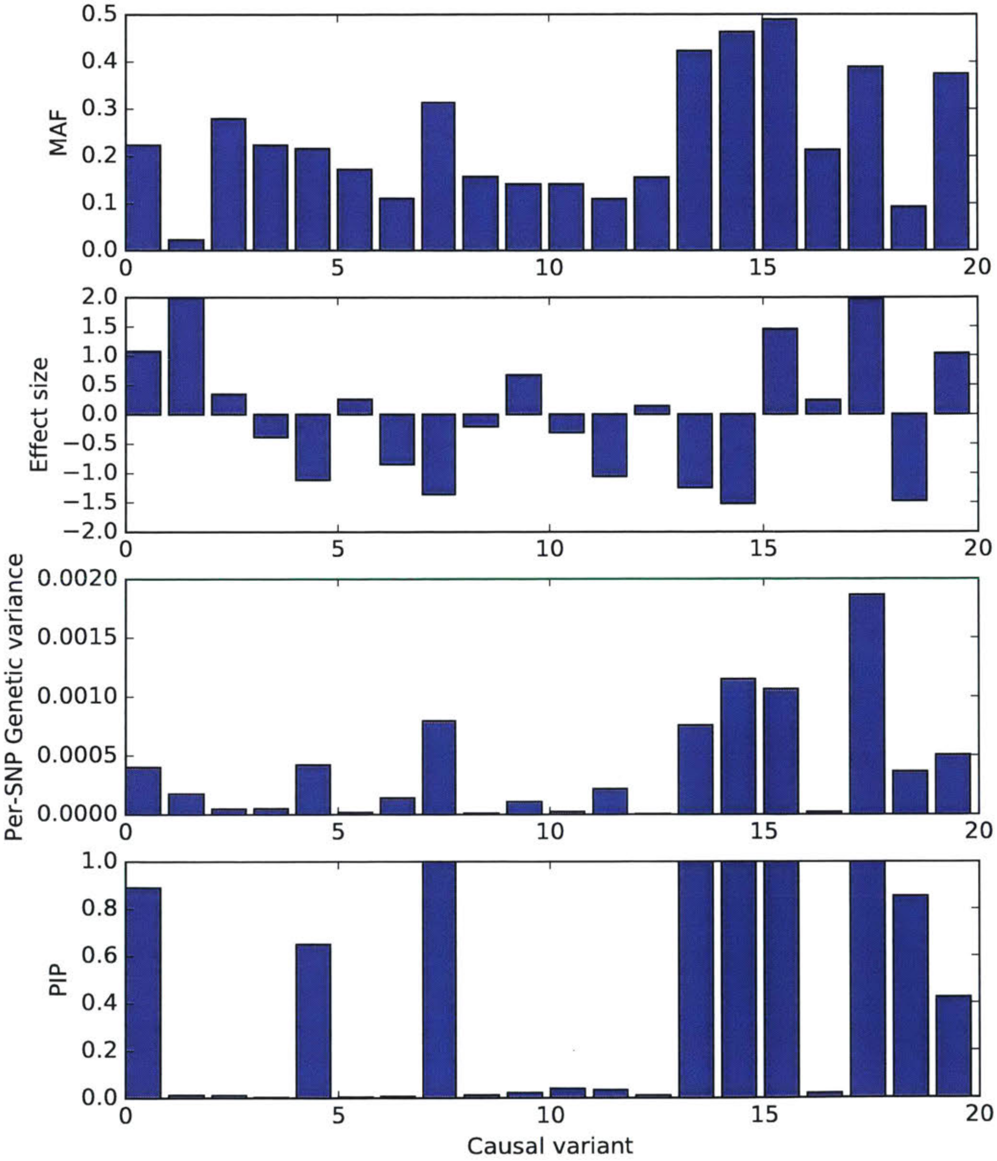


Figure 4.16: Simulation parameters and fitted model for null simulation trial #2 (Table 4.2)

Warped likelihood

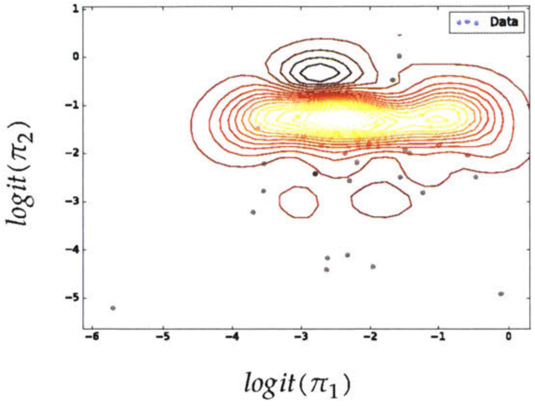
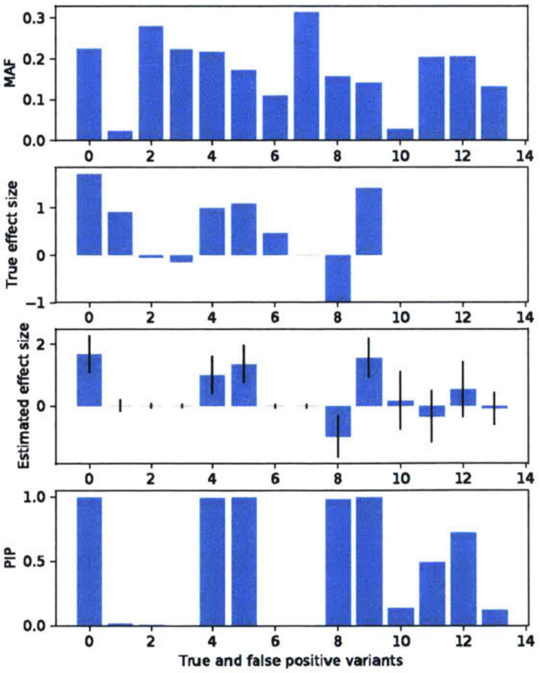


Figure 4.17: Warped likelihood surface for positive simulation trial 9 estimated using Gaussian process regression on active samples.



Chapter 5

Conclusion

In this thesis, we sought to go beyond merely identifying relevant biological annotations of the genome for disease, and move towards translating non-coding genetic associations into mechanistic biological insights. In Chapter 2, we develop new computational methods to analyze GWAS data and epigenomic profiles to identify cell type-specific regulatory elements associated with disease. In Chapter 3, we used regulatory circuits and gene expression to characterize the joint functional impact of those regulatory elements on biological pathways. Finally, in Chapter 4, we framed the problem as a Bayesian regression problem and investigated approximate inference techniques to estimate the relevant parameters.

Together, our results illustrate an approach to interpret weak non-coding variation and characterize disrupted biological pathways in complex diseases, and motivate a number of future directions:

1. The key idea underlying the work in this thesis is that exploiting more complex models of transcriptional regulation integrating diverse data types will lead to more specific and higher confidence biological predictions. However, in this thesis we do not explore the question of building such models de novo from existing data such as epigenomic profiles, regulatory motifs, and gene pathways we describe in Chapters 2 and 3.

   Building these models will require techniques such as the recent advances in approximate Bayesian inference described in Chapters 4. However, a further challenge will be to integrate these models with human genetic association data at the scale of GWAS (which we also explore in Chapter 4).

2. In Chapter 3, we highlight the different roles of transcription factor binding and transcription factor expression, specific mechanisms of gene regulation, in the diseases we study in this thesis. However, there is a pressing need to distinguish between different modes of tissue-specific gene regulatory action in order to accurately predict the role of genetic variation on transcriptional regulation, and through regulation on disease phenotypes.

   In order to fit models which are capable of distinguishing these mechanisms, we need the techniques presented in Chapter 4, as well as large scale expression data sets such as the Gene Tissue Expression Project[84]. Of particular interest are recent developments in single cell expression profiling, allowing us to gain insight into intracellular variation in expression, and the impact of genetics on that variation.

3. Throughout this thesis, we exploit reference annotations of epigenomic signatures and other

biochemical activity generated by large-scale sequencing efforts. However, one open question is the relative importance of of inter-individual variation in these annotations, and more broadly in intermediate phenotypes (gene expression, cell functions, metabolite levels, etc.) which mediate the impact of genetic variation on downstream disease phenotypes.

Answering this question requires novel data generation in diverse panels of individuals and novel computational algorithms to deconvolve inter-individual variation, intra-individual variation, and technical noise.

4. In this thesis, we rely heavily on previously published computational algorithms/pipelines and models which produced biological annotation of the non-coding genome. However, these computational methods typically produce point estimates without associated uncertainties. Worse, even when methods can estimate uncertainties, we often need novel techniques to propagate uncertainties into downstream computations.

   Bayesian inference is one natural way to incorporate uncertainty into models; however, the necessary computations quickly become intractable and necessitate the design of sophisticated approximation algorithms of the sort presented in Chapter 4.

Bibliography

1. Lander, E. S. Initial impact of the sequencing of the human genome. Nature **470,** 187–197 (2011).
2. Altshuler, D., Daly, M. J. & Lander, E. S. Genetic Mapping in Human Disease. Science **322,** 881–888 (2008).
3. Spain, S. L. & Barrett, J. C. Strategies for fine-mapping complex traits. Hum Mol Genet **24.** 26157023[pmid], R111–R119 (2015).
4. Manolio, T. A. et al. Finding the missing heritability of complex diseases. Nature **461,** 747–753 (2009).
5. Hill, W. G., Goddard, M. E. & Visscher, P. M. Data and Theory Point to Mainly Additive Genetic Variance for Complex Traits. PLoS Genet **4,** e1000008 (2008).
6. Polderman, T. J. C. et al. Meta-analysis of the heritability of human traits based on fifty years of twin studies. Nat Genet **47.** Analysis, 702–709 (2015).
7. Wood, A. R. et al. Defining the role of common variation in the genomic and biological architecture of adult human height. Nat Genet **46.** Article, 1173–1186 (2014).
8. Schizophrenia Working Group of the Psychiatric Genomics Consortium. Biological insights from 108 schizophrenia-associated genetic loci. Nature **511.** Article, 421–427 (2014).
9. Pasaniuc, B. & Price, A. L. Dissecting the genetics of complex traits using summary association statistics. Nat Rev Genet **18.** Review, 117–127 (2017).
10. Maurano, M. T. et al. Systematic Localization of Common Disease-Associated Variation in Regulatory DNA. Science **337,** 1190–1195 (2012).
11. Welter, D. et al. The NHGRI GWAS Catalog, a curated resource of SNP-trait associations. Nucleic Acids Research **42,** D1001–D1006 (2014).
12. Consortium, T. E. P. An integrated encyclopedia of DNA elements in the human genome. Nature **489,** 57–74 (2012).
13. Roadmap Epigenomics Consortium et al. Integrative analysis of 111 reference human epigenomes. Nature **518.** Article, 317–330 (2015).
14. Zuk, O. et al. Searching for missing heritability: Designing rare variant association studies. Proceedings of the National Academy of Sciences **111,** E455–E464 (2014).
15. Li, D., Lewinger, J. P., Gauderman, W. J., Murcray, C. E. & Conti, D. Using extreme phenotype sampling to identify the rare causal variants of quantitative traits in association studies. Genetic Epidemiology **35,** 790–799 (2011).

16. Botstein, D., White, R. L., Skolnick, M. & Davis, R. W. Construction of a genetic linkage map in man using restriction fragment length polymorphisms. Am J Hum Genet **32,** 314–331 (3 1980).

17. Ott, J., Wang, J. & Leal, S. M. Genetic linkage analysis in the age of whole-genome sequencing. Nat Rev Genet **16.** Review, 275–284 (2015).

18. Ott, J., Kamatani, Y. & Lathrop, M. Family-based designs for genome-wide association studies. Nat Rev Genet **12,** 465–474 (2011).

19. The International HapMap Consortium. A second generation human haplotype map of over 3.1 million SNPs. Nature **449,** 851–861 (2007).

20. The 1000 Genomes Project Consortium. An integrated map of genetic variation from 1,092 human genomes. Nature **491,** 56–65 (2012).

21. 1000 Genomes Project Consortium, T. A global reference for human genetic variation. Nature **526.** Article, 68–74 (2015).

22. Stephens, M. & Balding, D. J. Bayesian statistical methods for genetic association studies. Nat Rev Genet **10,** 681–690 (2009).

23. Howie, B., Fuchsberger, C., Stephens, M., Marchini, J. & Abecasis, G. R. Fast and accurate genotype imputation in genome-wide association studies through pre-phasing. Nat Genet **44,** 955–959 (2012).

24. Howie, B., Marchini, J. & Stephens, M. Genotype Imputation with Thousands of Genomes. G3 **1,** 457–470 (2011).

25. Spanos, A. Probability Theory and Statistical Inference: Econometric Modeling with Observational Data (Cambridge University Press, 1999).

26. Pe'er, I., Yelensky, R., Altshuler, D. & Daly, M. J. Estimation of the multiple testing burden for genomewide association studies of nearly all common variants. Genetic Epidemiology **32,** 381–385 (2008).

27. Benjamini, Y. & Yekutieli, D. The Control of the False Discovery Rate in Multiple Testing under Dependency. The Annals of Statistics **29,** 1165–1188 (2001).

28. Storey, J. D. The positive false discovery rate: a Bayesian interpretation and the q-value. Ann. Statist. **31,** 2013–2035 (2003).

29. Ernst, J. & Kellis, M. ChromHMM: automating chromatin-state discovery and characterization. Nat Meth **9,** 215–216 (2012).

30. Whyte, W. A. et al. Master Transcription Factors and Mediator Establish Super-Enhancers at Key Cell Identity Genes. Cell **153,** 307–319 (2013).

31. Ernst, J. et al. Mapping and analysis of chromatin state dynamics in nine human cell types. Nature **473,** 43–49 (2011).

32. Hnisz, D. et al. Super-Enhancers in the Control of Cell Identity and Disease. Cell **155,** 934–947 (2013).

33. Burren, O. S. et al. T1DBase: update 2011, organization and presentation of large-scale data sets for type 1 diabetes research. Nucleic Acids Res **39** (2011).

34. Howie, B. N., Donnelly, P. & Marchini, J. A Flexible and Accurate Genotype Imputation Method for the Next Generation of Genome-Wide Association Studies. PLoS Genet **5,** e1000529 (2009).

35. Marchini, J., Howie, B., Myers, S., McVean, G. & Donnelly, P. A new multipoint method for genome-wide association studies by imputation of genotypes. Nat Genet **39,** 906–913 (2007).

36. Yang, J., Lee, S. H., Goddard, M. E. & Visscher, P. M. GCTA: A Tool for Genome-wide Complex Trait Analysis. Am J Hum Genet **88,** 76–82 (2011).

37. Tange, O. GNU Parallel - The Command-Line Power Tool. The USENIX Magazine **36,** 42–47 (2011).

38. Subramanian, A. et al. Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. Proc Natl Acad Sci USA **102,** 15545–15550 (2005).

39. Quinlan, A. R. & Hall, I. M. BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics **26,** 841–842 (2010).

40. Efron, B. Bootstrap Methods: Another Look at the Jackknife. Ann. Statist. **7,** 1–26 (1979).

41. Bulik-Sullivan, B. K. et al. LD Score regression distinguishes confounding from polygenicity in genome-wide association studies. Nat Genet **47.** Technical Report, 291–295 (2015).

42. Cowper-Sal·lari, R. et al. Breast cancer risk-associated SNPs modulate the affinity of chromatin for FOXA1 and alter gene expression. Nat Genet **44,** 1191–1198 (2012).

43. The Wellcome Trust Case Control Consortium. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. Nature **447,** 661–678 (2007).

44. Stahl, E. A. et al. Bayesian inference analyses of the polygenic architecture of rheumatoid arthritis. Nat Genet **44,** 483–489 (2012).

45. Pickrell, J. K. Joint Analysis of Functional Genomic Data and Genome-wide Association Studies of 18 Human Traits. Am J Hum Genet **94,** 559–573 (2014).

46. Lambert, J.-C. et al. Meta-analysis of 74,046 individuals identifies 11 new susceptibility loci for Alzheimer's disease. Nat Genet **45.** Letter, 1452–1458 (2013).

47. Psychiatric GWAS Consortium Bipolar Disorder Working Group. Large-scale genome-wide association analysis of bipolar disorder identifies a new susceptibility locus near ODZ4. Nat Genet **43,** 977–983 (2011).

48. Schunkert, H. et al. Large-scale association analysis identifies 13 new susceptibility loci for coronary artery disease. Nat Genet **43,** 333–338 (2011).

49. Franke, A. et al. Genome-wide meta-analysis increases to 71 the number of confirmed Crohn's disease susceptibility loci. Nat Genet **42,** 1118–1125 (2010).

50. Stahl, E. A. et al. Genome-wide association study meta-analysis identifies seven new rheumatoid arthritis risk loci. Nat Genet **42,** 508–514 (2010).

51. Ripke, S. et al. Genome-wide association analysis identifies 13 new risk loci for schizophrenia. Nat Genet **45.** Article, 1150–1159 (2013).

52. Bradfield, J. P. et al. A Genome-Wide Meta-Analysis of Six Type 1 Diabetes Cohorts Identifies Multiple Associated Loci. PLoS Genet **7,** e1002293 (2011).

53. Morris, A. P. et al. Large-scale association analysis provides insights into the genetic architecture and pathophysiology of type 2 diabetes. Nat Genet **44,** 981–990 (2012).

54. Pasaniuc, B. et al. Fast and accurate imputation of summary statistics enhances evidence of functional enrichment. Bioinformatics (2014).

55. Heneka, M. T., Kummer, M. P. & Latz, E. Innate immune activation in neurodegenerative disease. Nat Rev Immunol **14.** Review, 463–477 (2014).

56. Rege, S. & Hodgkinson, S. J. Immune dysregulation and autoimmunity in bipolar disorder: Synthesis of the evidence and its clinical application. Australian and New Zealand Journal of Psychiatry **47,** 1136–1151 (2013).

57. Sekar, A. et al. Schizophrenia risk from complex variation of complement component 4. Nature **530.** Article, 177–183 (2016).

58. Rogler, G. & Rosano, G. The heart and the gut. European Heart Journal **35,** 426–430 (2014).

59. Pasquali, L. et al. Pancreatic islet enhancer clusters enriched in type 2 diabetes risk-associated variants. Nat Genet (2014).

60. Drucker, D. J. The role of gut hormones in glucose homeostasis. The Journal of Clinical Investigation **117,** 24–32 (2007).

61. Brzyski, D. et al. Controlling the Rate of GWAS False Discoveries. Genetics. eprint: `http://www.genetics.org/content/early/2016/10/26/genetics.116.193987.full.pdf` (2016).

62. Efron, B., Tibshirani, R., Storey, J. D. & Tusher, V. Empirical Bayes Analysis of a Microarray Experiment. Journal of the American Statistical Association **96,** 1151–1160 (2001).

63. Efron, B. Simultaneous inference: When should hypothesis testing problems be combined? Ann Appl Stat **2,** 197–223 (2008).

64. Yekutieli, D. Hierarchical False Discovery Rate–Controlling Methodology. J Am Stat Assoc **103,** 309–316 (2008).

65. Finucane, H. K. et al. Partitioning heritability by functional annotation using genome-wide association summary statistics. Nat Genet **advance online publication.** Analysis (2015).

66. Trynka, G. et al. Chromatin marks identify critical cell types for fine mapping complex trait variants. Nat Genet **45,** 124–130 (2013).

67. Kichaev, G. et al. Integrating Functional Data to Prioritize Causal Variants in Statistical Fine-Mapping Studies. PLoS Genet **10,** e1004722 (2014).

68. Li, Y. & Kellis, M. Joint Bayesian inference of risk variants and tissue-specific epigenomic enrichments across multiple complex human diseases. Nucleic Acids Research. eprint: `http://nar.oxfordjournals.org/content/early/2016/07/12/nar.gkw627.full.pdf+html` (2016).

69. Gusev, A. et al. Partitioning Heritability of Regulatory and Cell-Type-Specific Variants across 11 Common Diseases. Am J Hum Genet **95,** 535–552 (2015).

70. Day, F. et al. Genomic analyses for age at menarche identify 389 independent signals and indicate BMI-independent effects of puberty timing on cancer susceptibility. bioRxiv. eprint: `http://biorxiv.org/content/early/2016/09/23/076794.full.pdf` (2016).

71. Kasowski, M. et al. Extensive Variation in Chromatin States Across Humans. Science **342,** 750–752 (2013).

72. Stasevich, T. J. et al. Regulation of RNA polymerase II activation by histone acetylation in single living cells. Nature **516.** Letter, 272–275 (2014).

73. McLean, C. Y. et al. GREAT improves functional interpretation of cis-regulatory regions. Nat Biotech **28,** 495–501 (2010).

74. Kheradpour, P. & Kellis, M. Systematic discovery and characterization of regulatory motifs in ENCODE TF binding experiments. Nucleic Acids Res (2013).

75. Carbonetto, P. & Stephens, M. Scalable Variational Inference for Bayesian Variable Selection in Regression, and Its Accuracy in Genetic Association Studies. Bayesian Anal **7,** 73–108 (2012).

76. Cnop, M. et al. Mechanisms of Pancreatic β-Cell Death in Type 1 and Type 2 Diabetes. Diabetes **54,** S97–S107 (2005).

77. Claussnitzer, M. et al. Leveraging Cross-Species Transcription Factor Binding Site Patterns: From Diabetes Risk Loci to Disease Mechanisms. Cell **156,** 343–358 (2014).

78. Kölsch, H. et al. RXRA gene variations influence Alzheimer's disease risk and cholesterol metabolism. Journal of Cellular and Molecular Medicine **13,** 589–598 (2009).

79. Li, L. et al. Epithelial-specific ETS-1 (ESE1/ELF3) regulates apoptosis of intestinal epithelial cells in ulcerative colitis via accelerating NF-κB activation. Immunologic Research **62,** 198–212 (2015).

80. Bulik-Sullivan, B. et al. An atlas of genetic correlations across human diseases and traits. Nat Genet **advance online publication.** Analysis (2015).

81. Senger, K. et al. Immunity Regulatory DNAs Share Common Organizational Features in Drosophila. Molecular Cell **13,** 19–32 (2004).

82. Grenningloh, R., Kang, B. Y. & Ho, I.-C. Ets-1, a functional cofactor of T-bet, is essential for Th1 inflammatory responses. The Journal of Experimental Medicine **201,** 615–626 (2005).

83. Gjoneska, E. et al. Conserved epigenomic signals in mice and humans reveal immune basis of Alzheimer's disease. Nature **518.** Letter, 365–369 (2015).

84. The GTEx Consortium. The Genotype-Tissue Expression (GTEx) pilot analysis: Multitissue gene regulation in humans. Science **348,** 648–660 (2015).

85. Finucane, H. et al. Heritability enrichment of specifically expressed genes identifies disease-relevant tissues and cell types. bioRxiv. eprint: `http://biorxiv.org/content/early/2017/01/25/103069.full.pdf` (2017).

86. Schmitt, A. D. et al. A Compendium of Chromatin Contact Maps Reveals Spatially Active Regions in the Human Genome. Cell Reports **17,** 2042–2059 (2017).

87. Hernández-Lobato, D., Hernández-Lobato, J. M. & Dupont, P. Generalized spike-and-slab priors for Bayesian group feature selection using expectation propagation. Journal of Machine Learning Research **14,** 1891–1945 (2013).

88. Guan, Y. & Stephens, M. BAYESIAN VARIABLE SELECTION REGRESSION FOR GENOME-WIDE ASSOCIATION STUDIES AND OTHER LARGE-SCALE PROBLEMS. English. Ann Appl Stat **5,** 1780–1815 (2011).

89. Carbonetto, P. & Stephens, M. Integrated Enrichment Analysis of Variants and Pathways in Genome-Wide Association Studies Indicates Central Role for IL-2 Signaling Genes in Type 1 Diabetes, and Cytokine Signaling Genes in Crohn's Disease. PLoS Genet **9** (2013).

90. Golan, D., Lander, E. S. & Rosset, S. Measuring missing heritability: Inferring the contribution of common variants. Proc Natl Acad Sci U S A **111,** E5272–E5281 (2014).

91. Gunter, T., Osborne, M. A., Garnett, R., Hennig, P. & Roberts, S. J. Sampling for Inference in Probabilistic Models with Fast Bayesian Quadrature in NIPS (2014), 2789–2797.

92. Kingma, D. P. & Welling, M. Auto-Encoding Variational Bayes. CoRR **abs/1312.6114** (2013).

93. Zhou, X., Carbonetto, P. & Stephens, M. Polygenic Modeling with Bayesian Sparse Linear Mixed Models. PLoS Genet **9,** e1003264 (2013).

94. Park, Y., Sarkar, A. K., Bhutani, K. & Kellis, M. Multi-tissue polygenic models for transcriptome-wide association studies. bioRxiv. eprint: `http://biorxiv.org/content/early/2017/02/10/107623.full.pdf` (2017).

95. Minka, T. P. Expectation Propagation for approximate Bayesian inference in UAI (Morgan Kaufmann, 2001), 362–369.

96. Hernández-Lobato, J. M. et al. Black-box α-divergence Minimization. ArXiv e-prints. arXiv: `1511.03243 [stat.ML]` (2015).

97. Minka, T. Divergence Measures and Message Passing tech. rep. (2005), 17.

98. O'Hagan, A. Bayes–Hermite quadrature. Journal of Statistical Planning and Inference **29,** 245–260 (1991).

99. Osborne, M. A. et al. Active Learning of Model Evidence Using Bayesian Quadrature in NIPS (2012), 46–54.

100. Rasmussen, C. E. & Ghahramani, Z. Bayesian Monte Carlo in NIPS (MIT Press, 2002), 489–496.

101. Rasmussen, C. E. & Williams, C. K. I. Gaussian Processes for Machine Learning (Adaptive Computation and Machine Learning) (The MIT Press, 2005).

102. Prentice, R. L. & Pyke, R. Logistic disease incidence models and case-control studies. Biometrika **66,** 403–411 (1979).

103. Baker, S. G. The Multinomial-Poisson Transformation. J Royal Stat Soc **43,** 495–504 (4 1994).

104. Seaman, S. R. & Richardson, S. Equivalence of prospective and retrospective models in the Bayesian analysis of case-control studies. Biometrika **91,** 15–25 (2004).

105. Neuhaus, J. M. Closure of the class of binary generalized linear models in some non-standard settings. Journal of the Royal Statistical Society: Series B (Statistical Methodology) **62,** 839–846 (2000).

106. Titsias, M. K. & Lázaro-Gredilla, M. Doubly Stochastic Variational Bayes for non-Conjugate Inference in Proceedings of the 31th International Conference on Machine Learning, ICML 2014, Beijing, China, 21-26 June 2014 **32** (JMLR.org, 2014), 1971–1979.

107. Rezende, D. J., Mohamed, S. & Wierstra, D. Stochastic Back-propagation and Variational Inference in Deep Latent Gaussian Models. CoRR **abs/1401.4082** (2014).

108. Bergstra, J. et al. Theano: a CPU and GPU Math Expression Compiler in Proceedings of the Python for Scientific Computing Conference (SciPy) Oral Presentation (Austin, TX, 2010).

109. Bastien, F. et al. Theano: new features and speed improvements Deep Learning and Unsupervised Feature Learning NIPS 2012 Workshop. 2012.

110. Kingma, D. P., Salimans, T. & Welling, M. Variational Dropout and the Local Reparameterization Trick. CoRR **abs/1506.02557** (2015).

111. Agarwala, V., Flannick, J., Sunyaev, S., GoT2D Consortium & Altshuler, D. Evaluating empirical bounds on complex disease genetic architecture. Nat Genet **45.** Analysis, 1418–1427 (2013).

112. Yang, J., Fritsche, L. G., Zhou, X. & Abecasis, G. A scalable Bayesian method for integrating functional information in genome-wide association studies. bioRxiv. eprint: `http://biorxiv.org/content/early/2017/01/19/101691.full.pdf` (2017).

113. Jaakkola, T. S. & Jordan, M. I. Bayesian parameter estimation via variational methods. Statistics and Computing **10,** 25–37 (2000).