Characterizing non-coding hits in genome-wide association studies using epigenetic data

by

Abhishek Kulshreshtha Sarkar

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

Master of Science in Electrical Engineering and Computer Science

at the

Massachusetts Institute of Technology

June 2013

Author _____
Department of Electrical Engineering and Computer Science
May 22, 2013

Certified by _____
Manolis Kellis
Associate Professor of Electrical Engineering and Computer Science
Thesis Supervisor

Accepted by _____
Leslie A. Kolodziejski
Chair, Department Committee on Graduate Students

Characterizing non-coding hits in genome-wide association studies using epigenetic data

by

Abhishek Kulshreshtha Sarkar

Submitted to the Department of Electrical Engineering and Computer Science
on May 22, 2013 in partial fulfillment of the requirements for the degree of
Master of Science in Electrical Engineering and Computer Science

Abstract

Understanding the molecular basis of human disease is one of the greatest challenges of our time, and recent explosion in genetic and genomic datasets are finally putting it within reach. In the last ten years, genome-wide association studies have identified thousands of genetic variants associated with disease. However, the majority of these variants fall outside genes making interpreting their role in disease difficult. In parallel, the ENCODE and Roadmap Epigenomics consortia have produced high resolution annotations of the genome which identify large portions with potential regulatory function. We develop methods to interpret genome-wide association studies using these annotations to generate hypotheses about how associated variants contribute to disease mechanism. In particular, we go beyond the usual stringent p-value threshold to investigate variants with small individual effect sizes which current methods do not have power to detect. Evaluating our methods on the Wellcome Trust Case Control Consortium 7 Disease studies, we find associated variants are enriched in a variety of functional categories even after controlling for various biases. We also find an unprecedented number of variants contribute to this enrichment, supporting our hypothesis that the architecture of these diseases involves combinatorial interaction of many variants with small individual effect sizes.

Thesis supervisor: Manolis Kellis

Title: Associate Professor of Electrical Engineering and Computer Science

Contents

Acknowledgments

This thesis would not have been possible without the help and support of friends, colleagues, and family:

Aaron Sidford, whose Not So Great Ideas in Theoretical Computer Science were a welcome diversion.

The MIT Computational Biology Group for their feedback and suggestions.

Wouter Meuleman for his assistance with enhancer clustering and pathway analysis.

Luke Ward for his biological insight and assistance with developing and refining the methods.

Manolis Kellis for his enthusiasm and mentorship.

And my father, without whose support I would have never even made it this far.

Chapter 1

Introduction

Understanding the molecular basis of human disease is one of the greatest challenges of our time, and recent explosion in genetic and genomic datasets are finally putting it within reach. In the last ten years, genome-wide association studies (GWAS) have identified thousands of genetic variants associated with disease. However, the majority of these variants fall outside genes making interpreting their role in disease difficult. The first step in going from GWAS to explaining disease is to generate high quality hypotheses about which disease–associated variants are causal and how they contribute to the disease mechanism. The difficulty up to this point has been a lack of understanding of the non-coding genome.

The ENCODE and Roadmap Epigenomics consortia have now produced high resolution annotations of the genome which identify large portions with potential regulatory function. In particular, ChromHMM learns combinations of chromatin modifications which are enriched in regions with particular function. Annotating the genome with the most likely hidden state at each point gives an unparalleled resource for interpreting non-coding variants. Indeed, current work has begun to show top GWAS hits fall disproportionately in regulatory regions. It is now possible link these regions to their target genes and determine the proteins they recruit to regulate their targets and where they bind. With these rich annotations in hand current work can more confidently identify causal variants and generate highly specific mechanistic hypotheses about their contribution to disease pathology.

However, it is also known that GWAS lacks sufficient power to detect all but the most deleterious variants due to small sample sizes and human population genetic biases. The next key challenge in understanding complex polygenic disease is identifying causal variants in the long tail of the p-value distribution.

The goals of this thesis are three-fold. The first is to develop methods to use regulatory annotations and investigate the whole spectrum of GWAS p-values rather than only the top hits which pass the usual p-value threshold. We hypothesize complex traits arise from large numbers of variants with small individual effect sizes. These variants escape detection because typical p-value cutoffs are too stringent and samples sizes are not large enough. However, the ranking of SNPs by association to trait still gives some partial information which should contribute to genome-wide trends of over-representation in functional regions of the genome.

The second goal is to apply these methods to GWAS data studying Type 1 Diabetes

(T1D) and identify variants which are enriched. In particular, we should identify variants beyond the usual stringent p-value threshold which have smaller effect sizes. We also aim to identify relevant regulatory regions and the cell types which they are active in.

The third goal is to interpret the role of identified variants in disease. We should be able to make mechanistic hypotheses about how the variants we identify contribute to the disease. In particular, we look for gene pathways which are enriched for disease–associated variants.

The contributions of this thesis are two-fold. First, the methods should be generally applicable to complex polygenic traits. They should give some insight which will direct further investigation on the genetic architecture of these traits. Second, our results on T1D should reveal new insights about the disease biology which could be used to develop predictive models, diagnostics, and potential treatments.

Chapter 2

Background information

2.1 Inheritance

Near the turn of the 20th century, researchers rediscovered the work of Gregor Mendel, whose experiments on peas elucidated the particulate nature of inheritance. This work became the basis of the nascent field of genetics. But it took another half century for Watson and Crick to discover the molecular basis of inheritance, the double helical polymer *deoxyribonucleic acid* (DNA) [46]. DNA is a polymer of *nucleotides* (or bases): adenine, thymine, guanine, and cytosine (represented "A", "T", "G", "C"). The four nucleotides appear in two complementary pairs due to hydrogen bonding: adenine with thymine and guanine with cytosine. The primary sequence is paired with its reverse complement in the double helix structure. DNA is further structured into discrete molecules called *chromosomes*.

By the 19th century, scientists had already observed the role of DNA in asexual reproduction. In the process of *mitosis*, enzymes replicate the DNA exactly and divide the two copies among the two daughter cells. In the late 1800s, researchers observed the process of *meiosis* which is required for sexual reproduction. *Diploid* organisms (such as humans) have two *homologous* copies of each chromosome. However, they produce *haploid* cells (which only have one copy) called *gametes* which combine in the process of *fertilization* to produce a new diploid organism. Meiosis is initially identical to mitosis. However, after two diploid cells are produced each immediately divides into two haploid cells.

2.2 Genes and regulation

DNA is often called the "genetic code". In particular, regions of the sequence called *genes* are *transcribed* into ribonucleic acid (RNA). RNA is a single-stranded polymer similar to DNA, but substituting the nucleotide uracil for thymine. While DNA is confined to the nucleus of the cell, RNA can move out of the nucleus. Outside of the nucleus, cell machinery *translates* RNA into proteins. *Codons* (words of length 3) in the DNA/RNA sequence correspond to *amino acids*, the building blocks of proteins.

There are two key observations to make about this code. First, we now know less than 2% of the genome codes for proteins (although significantly more is transcribed into

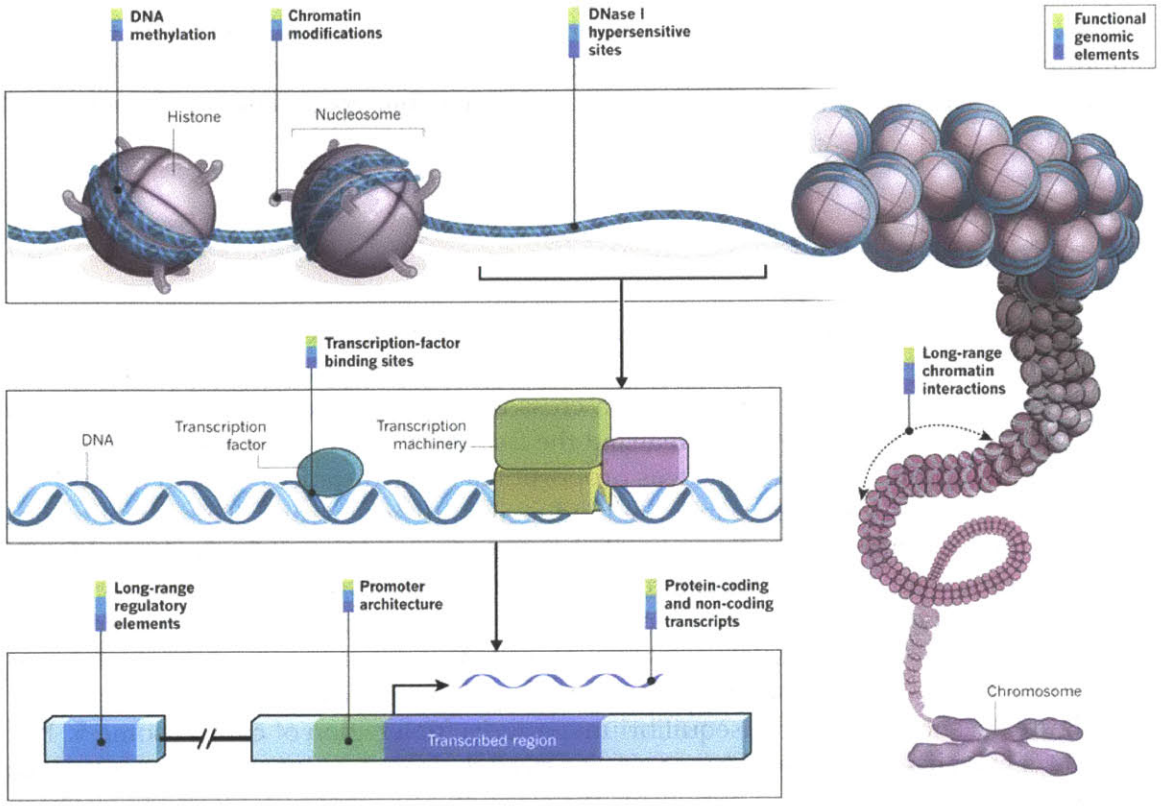

Figure 2.1: Illustration of regulatory regions and mechanisms [7]. The primary sequence (bottom) contains transcribed regions such as genes whose expression is regulated at short-range by promoters and long-range by enhancers. Transcription factors bind to the sequence and can form large complexes (center). The DNA molecule itself undergoes chemical modifications (top left) and changes conformation (top right) to change the accessibility of the primary sequence.

RNA). Second, every cell in complex organisms such as humans contains the same primary sequence of DNA. These observations raise the question of why we observe a huge variety of human cell types with diverse physical traits and functions.

We now know there is more encoded in the DNA sequence than just proteins. Another portion of the genome has regulatory function, modulating the *expression* (level of transcription) of genes. Figure 2.1 shows a variety of regulatory regions. *Promoters* are directly upstream of genes and are responsible for recruiting the *transcription factors* (TFs) which transcribe DNA. The structure of TFs is such that they bind to specific *DNA sequence motifs* (short words of length 8–16). TFs can also recruit other TFs to form larger complexes which are necessary to start transcription of some genes. More distant regions called *enhancers* also recruit TFs through motifs. TFs bound to enhancers can interact with promoters and genes because the DNA molecule can fold to bring these regions close together.

It is important to note genes both produce proteins and are regulated by proteins.

Indeed, there is a complex network of interactions between genes and regulators. The output of one gene may be a regulator which targets another, related gene. Larger *gene pathways* are sets of related or linked genes which together contribute to a coherent cell function.

2.3 Linkage analysis

In 1911, Morgan observed *crossover recombination* in *D. melanogaster* (fruit fly) meiosis. In one phase of meiosis, the chromosomes are arranged in the middle of the cell. During this phase, homologous chromosomes can cross over each other and exchange contiguous subsequences of DNA. Morgan realized the frequency of crossover recombination between two points on a chromosome was proportional to the distance between them. Thus, Mendel's hypothesis genes were inherited independently was wrong. This phenomenon, where genes are inherited together, is called *linkage disequilibrium* (LD). In 1913, Morgan and Sturtevant used this observation to develop the first *linkage map* and localized genes driving fly *phenotypes* (traits). Such a map gives not only the order of genes on the chromosome but also their relative distance to each other. A distance of 1 *centimorgan* corresponds to a .01 probability of recombination between two points on the chromosome.

Figure 2.2 shows linkage disequilibrium patterns for a region of chromosome 5. Pairwise LD is visualized as a lower-triangular matrix, where each point corresponds to a pair of SNPs and the amount of red represents the strength of correlation. Strong LD patterns are outlined in black and demarcate blocks of nearby SNPs which are only rarely separated by recombination.

Linkage analysis is the problem of finding DNA variation which *cosegregates* (is inherited together) with a trait of interest. If the DNA variant is inherited along with the trait, it must lie close or be "linked" to the gene driving the trait because recombination was not observed to separate the two. In the late 1970s, methods for cloning [20] and sequencing [32] DNA made it possible to tie linkage maps to the underlying sequence and clone linked genes. However, building genome-wide linkage maps in human was infeasible because few *genetic markers* (known positions on a chromosome) were known.

In 1980, Botstein proposed using *restriction fragment length polymorphisms* to build linkage maps [3]. *Restriction enzymes* are proteins which cut DNA at specific sequence motifs. Applying a restriction enzyme to the whole genome gives a particular distribution of fragment lengths. Variations in the sequence can either remove or introduce such motif instances, changing this distribution.

With genome-wide linkage maps in hand, *positional cloning* became the paradigm for uncovering the molecular basis of traits. Such experiments identified linked regions, sequenced those regions in cases (subjects who have the trait) and controls (who do not), and thereby identified the causal genes. The first success of this method was the discovery of the mutation driving Huntington's disease [12].

Given the subsequent success of this method in uncovering the basis of *Mendelian* traits arising from a single mutation/gene, geneticists hoped to be able to explain traits like common diseases. They found Mendelian subtypes of some common diseases, but the identified causal genes explained very few of the cases in the population. Indications

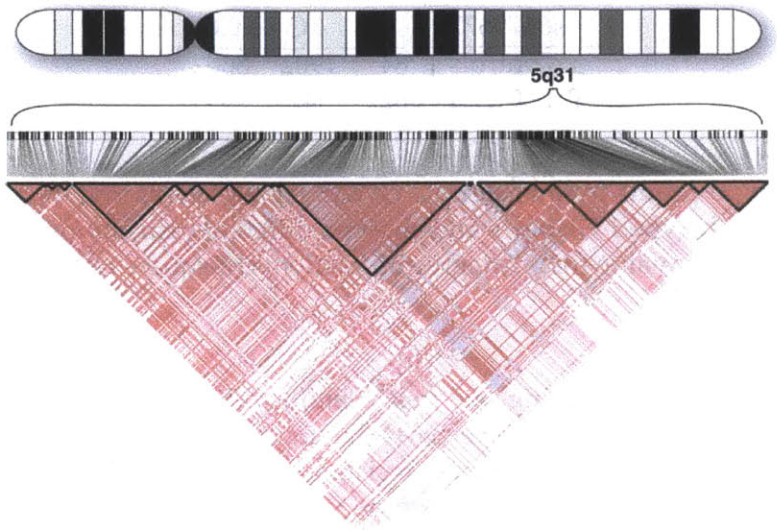

Figure 2.2: Patterns of LD in the genomic region 5q31 [1]. Pairwise correlation is displayed on the lower triangular matrix, where red is $r^2 = 1$. Strong correlation (outlined in black) demarcates blocks of nearby SNPs; however, we also observe long range correlations due to limited recombination.

suggested these traits arise from many genes. Indeed, in 1910 East proposed common traits are polygenic [6] and in 1919 Fisher proposed a model of how many discrete variants could lead to continuous traits [11].

2.4 The common disease, common variants hypothesis

Model organism geneticists working on yeast and fly have the ability to construct large crosses and therefore trace inheritance of *complex* polygenic traits and genetic variants through large pedigrees. Human geneticists did not have this luxury; instead they turned to insights from population and evolutionary genetics.

First, the human population has grown exponentially only recently on an evolutionary time scale. Evolutionary theory thus predicts limited genetic variation in the population. Indeed, *single nucleotide polymorphisms* (SNPs) occur on average only once every thousand base pairs. These naturally occurring DNA variants are single bases which take multiple *alleles* (possibilities) in the population. Most of these variants only take one of two alleles, called *major* and *minor* based on their relative frequency in the population. Moreover, the majority of these variants are common (i.e. the minor allele appears in more than 5% of the population).

Second, many Mendelian diseases arise from rare variants. These variants have large effects on reproductive success, so evolutionary theory predicts selection will drive their frequency to zero. On the other hand, common disease has smaller effect on reproductive fitness and so variants causing such traits could arise to higher frequency in the population. This outcome is facilitated by the recent growth in the human population. Moreover,

rs6679677	AA	AC	CC
Controls	84	1902	8602
T1D Cases	62	541	1359

Table 2.1: Genotype counts for an example GWAS tag SNP. This SNP tags a locus known to be associated with Type 1 Diabetes and shows a different distribution of counts in cases than in controls. The p-value for a 2 degree of freedom chi-square test on this contingency table is $p = 6.5 \times 10^{-39}$.

in some cases *heterozygotes* (having one risk allele) for disease-causing variants have an advantage over their *homozygote* counterparts (having no risk alleles). For example, sickle-cell anemia heterozygotes show increased resistance to malaria. Thus, human geneticists proposed common diseases arise from common variants.

2.5 Genome-wide association studies

To actually find common variants causing common disease in *genome-wide association studies* (GWASs), geneticists had to achieve three goals. First, they had to build comprehensive catalogs of SNPs in the human population. The HapMap consortium has sequenced over one thousand individuals and has published a catalog of over 2 million common SNPs [39, 40, 41]. More recently, the Thousand Genomes Project consortium has sequenced a comparable number of individuals using the latest sequencing technology and published a catalog of 38 million SNPs including rare variants [37].

Second, they had to develop methods to quickly and cheaply genotype these variants in large panels of individuals. Here, they were aided LD. As shown in Figure 2.2, SNPs occur in large blocks in which no recombination has occurred. Thus the genotypes of common SNPs can be inferred from just the genotypes of carefully chosen tag SNPs.

To actually genotype these tag SNPs, they developed DNA microarrays [44]. The array itself is a large library of DNA fragments called *probes* with one end anchored to a chip. The experiment amplifies and fragments the sample DNA, attaches fluorescent tags to the fragments, and hybridizes the fragments to the probes. They recover the original genotypes by observing the light intensity at each probe.

Third, they had to develop the statistics to analyze the patterns of genotypes in the cohorts of cases and controls. At a basic level, GWAS performs one hypothesis test per *locus* (a region of the genome represented by the tag SNP, but also containing all of the common SNPs in LD with the tag). There are two frequently tested null hypotheses: the distribution of allele counts is independent between cases and controls, or the distribution of genotype counts is independent. Table 2.1 shows an example set up for testing the null hypothesis the distributions of genotypes between controls and Type 1 Diabetes cases at a particular tag SNP are independent. One issue is conducting millions of such tests inflates the false positive rate and requires stringent correction. The usual method is Bonferroni correction (dividing the desired false positive rate α by the total number of tests) which is known to be over-conservative.

A more difficult problem is accounting for sources of genetic variation which are known to not be related to phenotype. One obvious confounder is familial relation between subjects, which violates the independence assumption underlying the statistical testing. But another consequence of the recent human population expansion is genome-wide allele frequency differences associated with ethnicity (i.e. human subpopulations). One possible way to account for these biases is to simply choose subjects such that they are unrelated and come from the same ethnic background. Statistically correcting for this bias in cohorts of mixed ethnicities requires more sophisticated methods such as genomic control, principal components analysis, or learning mixed models. Technical artifacts of the genotyping technology are another unavoidable confounder, but the error rates of the technologies are steadily decreasing.

2.6 The missing heritability problem

To date, thousands of loci have been associated with hundreds of traits [15]. Many of these loci are independent, further supporting the polygenic basis of complex traits. However, the vast majority do not lie in or near genes, making interpreting their function difficult. Moreover, those variants which have been independently reproduced in follow-up studies still explain only a fraction of the *heritability* of these traits [23, 24, 8]. Heritability is the proportion of phenotypic variation explained by genetic variation. Obviously, other factors beyond genetics contribute to the instances of common disease in the population. However, the fraction of heritability of common disease explained by known variants does not match the fraction of heritability attributable to genetics based on tracing these diseases through pedigrees.

There are multiple competing hypotheses about where this missing heritability can be found. First, one previously inaccessible source is rare, private mutations. These are not captured by DNA microarrays, but can be efficiently genotyped by current sequencing technology. Second, another potential source is DNA structural variants, such as insertions, deletions, inversions, repeats, and other rearrangements of the sequence. Again, it is only with the latest technology that genome-wide detection of these variants has become feasible.

Third, common variants could indeed explain the heritability of complex traits but have too low individual effect sizes to be detected by current methods. One concern is that the statistical power of genome-wide association studies is limited by the size of the case and control cohorts. Some progress has been made on this front by pooling data across studies and performing meta-analysis. Fourth, variants and genes interact with each other non-additively to contribute to complex traits. Many computational models have been proposed to learn such interactions from genotype data; however, they have to bound the degree of interaction terms to make analysis tractable and still leave much heritability unexplained.

Recent work on the genetic basis of human height lends weight to these last two hypotheses [30, 19]. Height is estimated to be 80% heritable. Although some rare variants have been found which explain extreme values of the phenotype, they explain only a minority of cases. Roughly 50 SNPs have been associated with height in GWASs; however,

they explain only 5% of the heritability. However, by considering the entire GWAS panel of SNPs we can explain roughly 45% of the heritability of height.

2.7 Epigenetics

Parallel to these developments in understanding the role of DNA as genetic code and localizing causal variants for traits, we have also learned there is more to the molecule than simply the primary sequence. In particular, there are heritable molecular traits which are not explained by the primary sequence. One example is *DNA methylation*, in which individual cytosine nucleotides are modified with an additional methyl (CH_3) group. Methylation is now known to serve as a silencer of DNA function, and current work seeks to understand both how the primary sequence and disease traits can change methylation across the genome.

Epigenetics also refers more generally to molecular factors other than the primary DNA sequence which can affect traits. As described previously, the DNA molecule is divided into chromosomes. The structure of chromosomes involves several nested levels of structure as shown in Figure 2.1. At the lowest level, the DNA double helix is wrapped around protein complexes called *nucleosomes*, each constructed of four proteins called *histones*. Histones themselves are accepting of modifications such as methylation and acetylation. Hundreds of these modifications, called *chromatin marks*, have been discovered experimentally, leading to the hypothesis they also encode part of the function of DNA. For example, H3K4Me3 (trimethylation of the 4th amino acid in histone 3) is associated with nearby promoter activity.

Chromatin modifications are measured using *chromatin immunoprecipitation* followed by high-throughput sequencing (ChIP-Seq). The experiment uses restriction enzymes fragment the DNA. Specifically selected antibodies are used to select fragments bound to proteins of interest (such as histones with a particular modification). The fragments are then sequenced and aligned to a reference genome to localize their position.

We do not know *a priori* the function associated with individual marks, nor whether they act independently. Ernst instead used an unsupervised approach to learn "chromatin states", hidden states of a multivariate HMM [9]. The emission alphabet of the HMM is combinations of chromatin marks; the hidden states correspond to biological functions which change the probability of observing particular combinations. By learning the HMM over the whole genome, we get a high resolution map of regions which are likely promoters, enhancers, repressors, transcribed elements, etc. For example, Figure 2.3 shows the annotation of the WLS gene across nine human cell types. This gene is important in eukaryotic development in aligning one axis of symmetry. The gene is is predicted to be transcribed (green) in five of the cell types. In those cell types, we find an active enhancer (yellow). However, in the others the gene is quiescent (not expressed; gray). We find the promoter is *poised* (purple): although transcription factors are bound to it the gene is not being transcribed. Poised regulators are needed for rapid, time-dependent regulation of genes which is necessary in the development of embryos.

Beyond wrapping around histones, the DNA molecule is further condensed into a dense structure called *chromatin*. It is important to note the primary sequence is inacces-

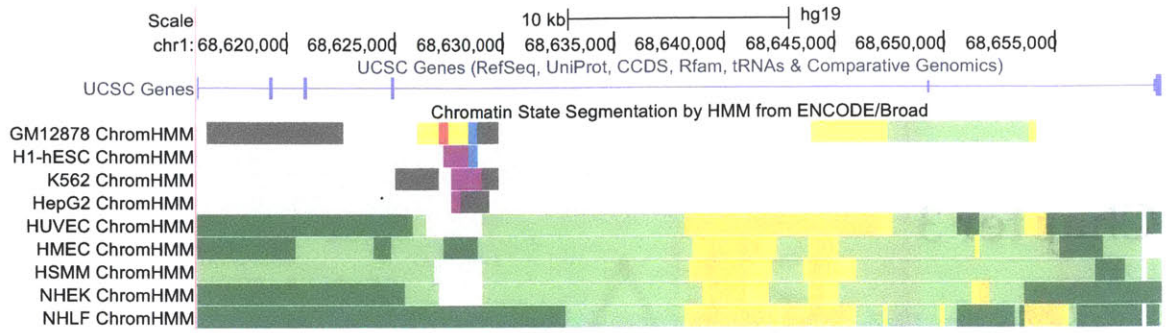

Figure 2.3: Chromatin state annotations of the region containing the WLS gene across nine human cell types. The gene is expressed (green) in five of the cell types. In these cell types we can identify a nearby strong enhancer (yellow). The promoter is poised (purple) but the gene is inactive (gray) in the other cell types.

sible in regions which are so condensed, as shown in Figure 2.1. Thus, whether or not chromatin is open is another indication of regulatory function. There are several experimental approaches to identify open chromatin, such as DNAseI hypersensitivity (DHS) and Digital Genomic Footprints (DGF) [14]. In essence, these methods all use restriction enzymes to cut DNA where it is accessible, sequences the fragments, and aligns them to a reference genome to localize their position.

Epigenetic modifications are an invaluable resource in interpreting the role of the non-coding part of the genome. Towards this end, the ENCODE Consortium has produced chromatin state maps of nine human cell lines and DHS/DGF in ninety [38]. The NIH Roadmap Epigenomics Consortium has produced complete epigenomes of 85 primary human cell types including methylation, gene expression (by extracting and sequencing RNA), and chromatin state maps [2].

Chapter 3

Genome-wide enrichments of functional elements

We have introduced the paradigm of genome-wide association studies and described how potentially many undiscovered loci could escape detection. We also introduced regulatory elements which control gene expression and are known to play some role in disease. Now, we describe our contributions in combining these data to identify large numbers of small effect size, potentially causal variants.

3.1 Methods

3.1.1 Functional annotations

We use ChromHMM annotations for nine ENCODE cell lines downloaded from the UCSC Genome Browser. We also use currently unpublished ChromHMM annotations for 85 Roadmap Epigenomics primary cell types (intersecting replicates) from the Analysis Working Group. We use DHS and DGF annotations for 90 ENCODE cell lines downloaded from the UCSC Genome Browser. We use long poly-A+ RNA-Seq contigs in 9 ENCODE cell lines to produce annotations of discretized expression downloaded from the European Bioinformatics Institute (accessed through the UCSC ENCODE portal).

3.1.2 GWAS Data

We revisit the Wellcome Trust Case Control Consortium 7 Diseases studies [42]. These studies investigate common diseases: bipolar disorder, coronary artery disease, Crohn's disease, hypertension, rheumatoid arthritis, Type 1 Diabetes, and Type 2 Diabetes. Each study involves a cohort of 2,000 cases and a shared set of 3,004 controls. All subjects were of British descent and were chosen to be unrelated. The controls were taken from two sources: the UK National Blood Service and the 1958 British Birth Cohort.

The cohorts were genotyped on the Affymetrix 500K SNP array, capturing just under 500,000 common SNPs. They were then imputed to the HapMap 3 common variants, yielding genotypes for 2.6 million common SNPs. P-values were computed using a 1 degree of freedom chi-square test for independence of allele counts between cases and controls.

We focus on the GWAS for Type 1 Diabetes (T1D). This disease is a classical example of a common, polygenic disease [13]. The disease pathology involves autoimmune destruction of beta cells in the pancreas which produce insulin. Those with the disease must inject insulin into their bloodstream to maintain healthy levels of blood sugars.

There is a Mendelian subtype of T1D which is attributable to a single rare mutation; however, this subtype explains fewer than 1% of cases in the population. Evidence supports the hypothesis hundreds of moderate-effect variants explain the disease making it an ideal disease to study. Moreover, we have regulatory annotations for a large variety of immune cells with various surface markers, giving us the opportunity to find regulators in disease-relevant cell types. These regulators are more likely to have a role in the disease, especially if they are specific to immune cells.

3.1.3 Statistical analysis

Our methods are motivated by two observations. First, the majority (roughly 80%) of trait-associated variants which have been found to date fall outside of genes. One obvious explanation is these variants fall in regulatory regions and contribute to disease mechanism by disrupting gene regulation. Second, evidence suggests disease-associated variants have moderate effect sizes which cannot be captured by current GWASs. They do not have large enough sample sizes to have sufficient power to detect these variants.

In essence, we want to extend the notion of enrichment of top loci to the whole genome. Specifically, we ask whether or not variants falling in regulatory regions are skewed to have lower p-value than variants falling outside regulatory regions. To answer this question we develop a visualization called an *RR plot* [1]. We rank the SNPs in order of increasing GWAS p-value and first compute the total number of SNPs falling in regulatory regions ("hits"). Let T be this total number. As we traverse the list, we keep a running count of how many hits we have observed and how many we would have expected assuming hits were uniformly distributed over the list. Suppose we have seen M of N total SNPs. Then the expected count is equal to:

$$E = \frac{MT}{N}$$

Suppose in the M SNPs there were O SNPs observed to fall in regulatory regions. We define the normalized cumulative deviation as:

$$D = \frac{O - E}{T}$$

We plot this quantity every 1000 SNPs down the list to obtain the RR plot. For example, Figure 3.1 shows an RR plot for enhancers in $CD4^+$ $CD25^-$ $IL17^+$ PMA/Ionomcyin-stimulated T_h17 primary cells in blue and an RR plot for a permuted control set of enhancers in red.

We make several key observations about these plots. First, the deviation is defined in such a way that the plot starts and ends at zero deviation. Second, if p-values in regulatory regions are skewed to have lower p-values than those outside the RR plot will show

[1] a play on quantile-quantile (QQ) plot

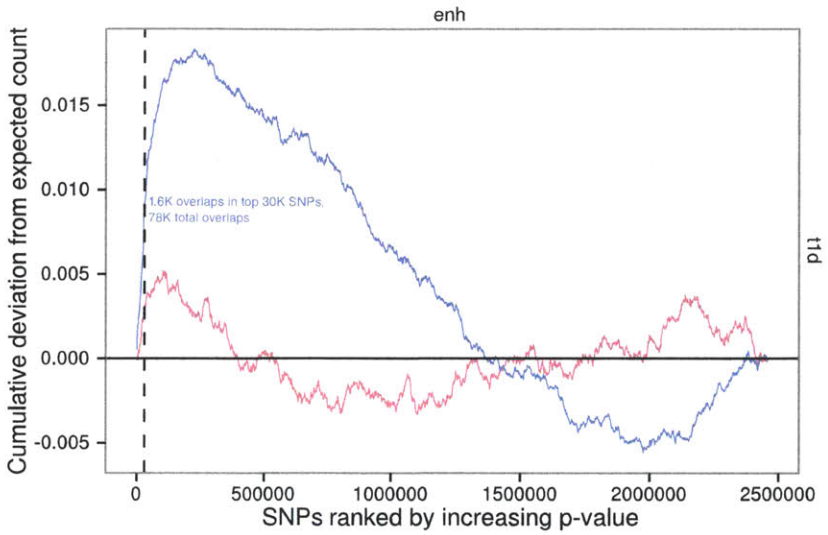

Figure 3.1: Example RR plot. After seeing 30K SNPs we have encountered 1.6K overlaps with T_h17 enhancers where only 900 were expected (based on the total count). When we plot this deviation over the whole ranked list, T_h17 enhancers (blue) are clearly enriched compared to the randomized control regions (red).

larger deviation early in the ranked list. Third, the plots not only allow us to ask whether functional regions are enriched for disease-associated variants but also how far down the ranked list we have to go before we stop observing this enrichment. In particular, we focus on the first part of the curve because although we hypothesize large numbers of variants could contribute to disease we do not believe variants which are not even significant before multiple testing correction are relevant. Therefore, we focus on the first 150,000 SNPs, beyond which p-values stop being significant at the $\alpha = .05$ level.

3.1.4 Related work

There are several methods published to interpret top GWAS loci using regulatory annotations. Ernst revisited published associations for diseases and showed in some cases they overlap more than expected with enhancer regions [10]. Moreover, the implicated enhancers and the genes they target are specifically active only in disease–relevant cell types. For example, reported associations for systemic lupus erythematosis, an autoimmune disease, fall in GM12878 lymphoblastoid–specific enhancers (an immune cell type). One of these associations tags a mutation which disrupts an Ets-1 sequence motif, which disrupts the GM12878–specific activating transcription factor Ets-1. This SNP is therefore hypothesized to disrupt activation of the targeted HLA-DRB1 gene, which is important for recognition of cell surface markers and differentiating one's own cells from invaders.

Trynka used a similar approach to interpret top SNPs and strongly correlated variants in LD in four phenotypes [43]. However, rather than calling chromatin states their

approach examined chromatin marks individually. The method computes a cell type–specificity score by based on the distance to the nearest ChIP-Seq peak in each cell type. To assess significance, the score is computed for LD block size–matched control sets.

Maurano used DHS and DGF to localize GWAS hits [25]. They first used cell type–activity profiles to link DHS regions with their target genes. They next identified variants falling in TF binding sites and showed they were over-represented in binding sites for genes relevant to disease traits. They also proposed hypergeometric tests at increasing p-value thresholds as a method to identify disease–relevant cell types. We do not take this approach because it introduces another multiple testing problem.

The question of whether p-values falling in a set of regions are skewed to be lower than those falling outside the set is a well-studied problem. In particular, this problem formulation is known as the *competitive null hypothesis* in gene set enrichment analysis. Gene set enrichment analysis is the problem of prioritizing sets of genes (usually related by function) for further dissection. Typically, each gene is scored based on the p-value of nearby SNPs either in terms of genomic position or correlation due to LD. The competitive null hypothesis states the scores of genes in the gene set are the same as the scores of genes outside the set. Our method differs from this approach by not assigning scores to regions but rather computing statistics on the p-values of individual SNPs directly.

One obvious approach to answer this question in our setting is treating the two sets of p-values as samples and using established statistical tests. The Mann–Whitney U test is a nonparametric test of the null hypothesis the two samples are the same against the alternative hypothesis one sample is greater than the other. U is defined as the sum for each observation in the first sample of the number of observations in the second sample which have lower rank. This quantity has a closed-form in terms of the sums of ranks of the observations in each sample. It is also approximately normally distributed. However, under this formulation the test is equivalent to one which asks whether difference of the medians of the two samples is nonzero [34]. This quantity does not capture the full distribution of the p-values, making it inappropriate for answering the question we are interested in.

The Kolmogorov–Smirnov two-sample D test is a nonparametric test of the null hypothesis the two samples come from the same probability distribution against the alternative hypothesis they do not. D is defined as a function of the empirical cumulative distribution functions of the two samples. Critical values of D (needed to compute p-values) are tabulated because the statistic does not have a closed-form distribution. This test more directly answers the question of whether the p-values in the first sample (falling in regulatory regions) are skewed to be different from those in the second sample (outside).

However, these two tests make independence assumptions which does not hold. First, they assume individual observations are independent. However, the two samples are p-values of SNPs and we know the genotypes of nearby SNPs are highly correlated due to LD. Therefore, the patterns of genotypes across cases and controls and therefore the p-values of nearby SNPs are also highly correlated.

Second, the they assume the two samples are independent. The patterns of genotypes in regulatory variants across cases and controls will be different from the patterns of other variants due to differential natural selection. Variants outside regulatory regions are less likely to have a role in reproductive fitness and therefore selection is less likely to apply pressure to maintain a particular genotype at those variants. On the other hand, variants

in regulatory regions are more likely to have a role and therefore selection will reduce the variability. Thus, the two samples are not independent because we gain some information about the p-values by conditioning on whether or not the SNP they correspond to falls in a regulatory region.

Statistical pitfalls aside, this approach produces easily interpreted results in identifying which cell types and annotations are relevant. However, the resulting p-values do not give any indication of how many SNPs are contributing to the observed enrichment.

A second approach is to ask whether variants falling in regulatory regions are over-represented at the head of the ranked list. One obvious way to test this approach is to compute Fisher's exact p-values at a variety of cutoffs; however, this method requires further multiple testing correction.

Our approach of keeping a running deviation is used by a number of gene set enrichment methods which we draw inspiration from. In particular, the GSEA algorithm defines an enrichment score as the maximum value achieved by a walk down the ranked list where each overlap counts as $+1$ and each non-overlap counts as -1 [36]. However, GSEA makes the assumption few genes are involved in the trait and exponentially reduces the weight of overlaps in the running sum further down the ranked list. We do not bias our method in this manner in order to compute a new empirical cutoff beyond which we stop seeing enrichment.

3.2 Results

3.2.1 Enhancer regions are enriched for disease–associated variants

We first ask which classes of functional regions are enriched for T1D–associated variants. We compute RR plots for a variety of annotations in the GM12878 lymphoblastoid cell line:

- Promoter chromatin states
- Enhancer chromatin states
- Transcribed chromatin states
- Repressed chromatin states
- Other chromatin states
- Expressed regions (intersection of poly-A+ RNA-Seq contigs with transcribed chromatin states)

Figure 3.2 shows promoters, enhancers, transcribed regions, and expressed regions are all enriched for T1D–associated variants. We expect transcribed and expressed regions to be enriched because mutations in these regions are more likely to be *non-synonymous* (changing an amino acid) and therefore deleterious. We also expect promoters to be enriched because mutations in these regions are more likely to disrupt binding sites which are necessary for the transcription of proteins, directly affecting gene expression. Moreover, variants in promoters are more likely to be in LD with variants falling in the gene

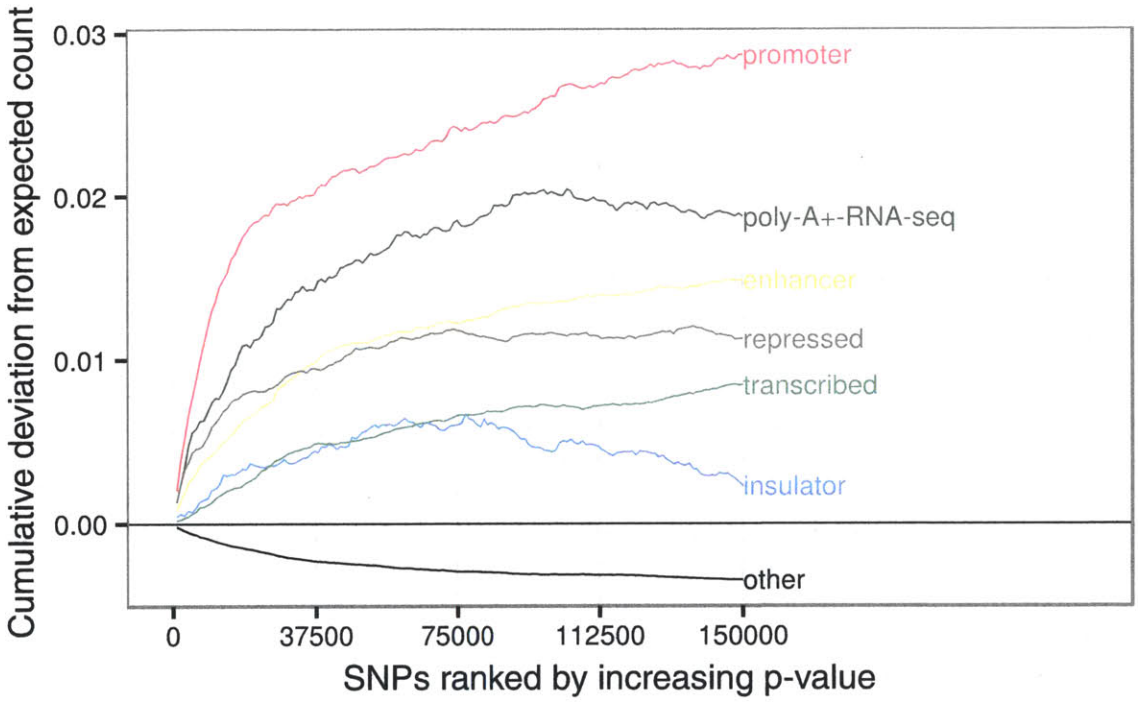

Figure 3.2: Enrichment of functional regions in GM12878 lymphoblastoid. We expect promoters and expressed regions to be enriched for T1D–associated variants due to the obvious role of these regions in cell function. However, we also find enhancer regions are enriched suggesting disregulation plays a role in the disease.

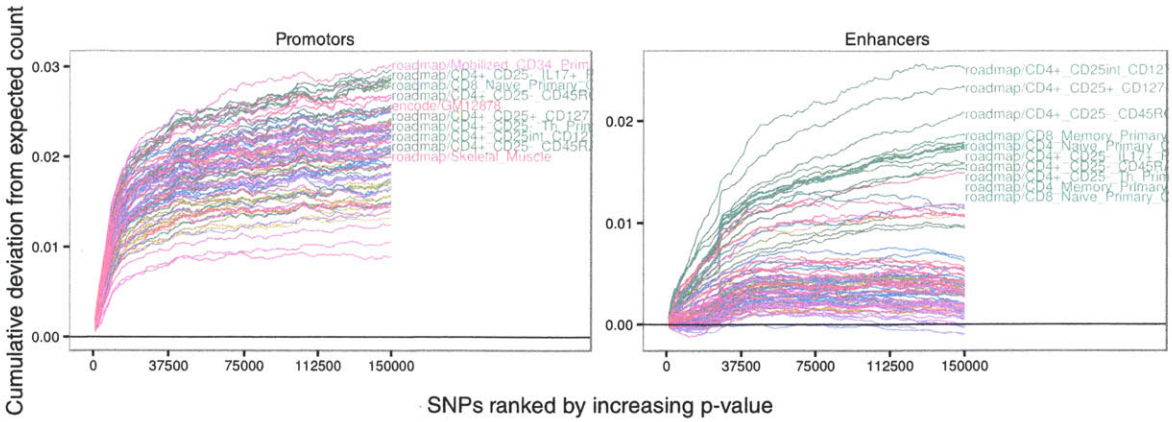

Figure 3.3: Comparison of enrichment of promoters and enhancers across cell types. Promoters (left) are equally enriched in all cell types. In contrast, enhancers (right) show cell type–specific enrichment in disease-relevant immune cell types.

they target meaning their p-values are more likely to be correlated and equally skewed. However, we also find enhancers are enriched, suggesting distal regulation of genes plays a significant role in the disease.

Next, we look at enrichment of these various functional classes across the ENCODE and Roadmap cell types. Figure 3.3 shows promoters in all cell types are equally enriched for T1D–associated variants. This result is supported by the fact promoters are conserved across cell types, i.e. regions of the genome which are promoters in one cell type are also highly likely to be promoters in another cell type. We find enrichment of promoters is uninformative in picking out regulatory variants which are cell type–specific.

Unlike promoters, enhancers are dynamic across different cell types. They show much greater variability in activity across cell types and are important in processes like cell type differentiation. This fact suggests enrichment of enhancer regions should show more cell type–specificity. Indeed, we find enhancers in T and B immune cell lines with a variety of surface markers show the greatest enrichment of all cell types. These cell types are the most relevant to the autoimmune nature of T1D out of those cell types for which we have chromatin state annotations.

Another key observation to make is the number of SNPs to which the observed enrichment continues past. We find enrichment of enhancers in immune cell lines even beyond 30,000 SNPs. This number is two orders of magnitude greater than the largest estimates of the number of SNPs involved in complex traits such as height. One obvious question is how many of these top 30,000 SNPs are actually involved in T1D. We can begin to answer this question by first looking at how many of the top 30,000 SNPs actually fall in functional regions. Table 3.1 shows these counts for functional regions in GM12878 lymphoblastoid. Only roughly 2,000 of these SNPs fall in enhancers, and even fewer fall in promoters and transcribed regions. But this result raises another question: why do we have to traverse 30,000 SNPs in the ranked list before we pick up all of the 2,000 the enhancers which could contribute to the enrichment. One potential explanation is we pick up SNPs in LD with causal variants as we walk down the ranked list and therefore dilute the signal.

Region type	Count (top)	Count (genome)
Promoter	876	27641
Enhancer	2463	117406
Insulator	249	15631
Repressed	1848	86762
Transcribed	5566	337297
Other	18998	1870718
poly-A+ RNA-Seq	379	14085

Table 3.1: Counts of top T1D–associated variants in functional regions in GM12878 lymphoblastoid. In the top 30,000 variants, only a small fraction fall in promoters, enhancers, or coding regions (either predicted by chromatin marks or actually expressed).

3.2.2 Implicated enhancers appear genome-wide

This concern about LD is significant because regulatory regions are physically clustered. This fact follows naturally from the fact protein-protein interaction is the mechanism of transcriptional regulation. Proteins bind to sequence motifs in regulatory regions, which are close to each other either due to being close in terms of genetic (base pair) position or being close to each other in three-dimensional position (because the DNA molecule folds on itself). Variants which are nearby are in stronger LD and therefore their genotypes and p-values are more highly correlated The concern then is the 2,000 enhancers we find are all physically clustered and in LD with each other, so only a few of them are actually causal and the rest simply have correlated p-value.

Another concern is the enhancers we find all fall in the *Major Histocompatability Complex* (MHC). Although genome-wide association studies consider mutations all over the genome, the top p-values are often highly localized to the MHC. This region of the genome in chromosome 6 contains many genes related to the function of recognizing cell surface markers to distinguish cells belonging to oneself versus cells which are invaders. It is highly variable across individuals and human subpopulations due to its role in immune response. It also shows atypical LD patterns compared to the rest of the genome such as long-range LD (i.e., abnormally large blocks in which no recombination occurs). These features of the MHC violate usual assumptions used in GWAS statistical testing and therefore variants in these regions often show highly significant p-values regardless of their relevance to the trait in question. Although in the case of T1D we expect to find hits in the MHC due to the autoimmune disease pathology, we also expect to find hits outside the MHC.

To visualize where in the genome the hits cluster, we map individual chromosomes to *Hilbert curves*. Hilbert curves are space filling curves defined by David Hilbert in 1891 which map the one-dimensional line to the two-dimensional plane [5]. The key property of these curves we exploit is preservation of locality. If two points are close to each other on the line, they will remain close to each other when mapped to the Hilbert curve. Although the converse is not true (points which are far apart on the line may be mapped close to each other on the Hilbert curve), we are mainly concerned about overlaps which are close

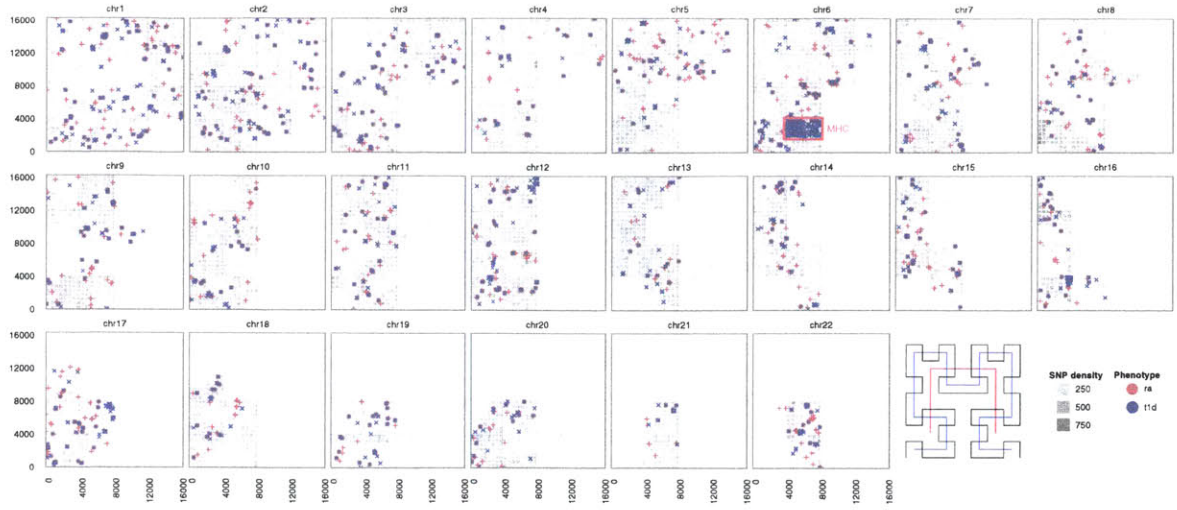

Figure 3.4: Overlaps with GM12878 enhancers in the top 30,000 T1D–associated variants and rheumatoid arthritis–associated variants mapped to Hilbert curves. No more than half the overlaps localize to the MHC (boxed). There are hundreds of independent clusters of enhancers, including some specific to each disease.

to each other.

Figure 3.4 shows this visualization for top enhancers in both T1D and rheumatoid arthritis (RA), another autoimmune disease. Indeed, we find a large cluster of roughly half the hits in the MHC. However, the remainder of the hits are scattered all over the genome in small clusters giving further evidence these enhancers are independent. Moreover, while we observe many clusters are shared between the two diseases, we also find many clusters which are disease-specific. Although both T1D and RA are autoimmune, they attack different parts of the human body (pancreas versus connective tissue). Accordingly, disease-specific clusters of enhancers are potentially related to disease-specific disregulation which causes differential pathology.

3.2.3 Implicated enhancers are independent of known associated loci

Another concern is the enhancers we find are linked by LD to known loci and therefore we are not actually finding novel associations. In the case of T1D, 91 loci are listed in the T1DBase as reliably associated with the disease [33]. To address this issue, we subtract out these loci (considering tagged SNPs with $r^2 > .8$) from foreground and background and redo the analysis. We address whether these enhancers are linked to nearby genes by subtracting out TSS–proximal regions. We account for linkage to non-synonymous variants by subtracting out loci tagged by those variants (again requiring $r^2 > .8$) We also do the same for loci overlapping the MHC.

Figure 3.5 shows we continue to find enrichment of enhancers even after subtracting out all of these potential confounders, although to a lower magnitude. We continue to see separation of immune cell types from other cell types giving further evidence we are finding novel associations.

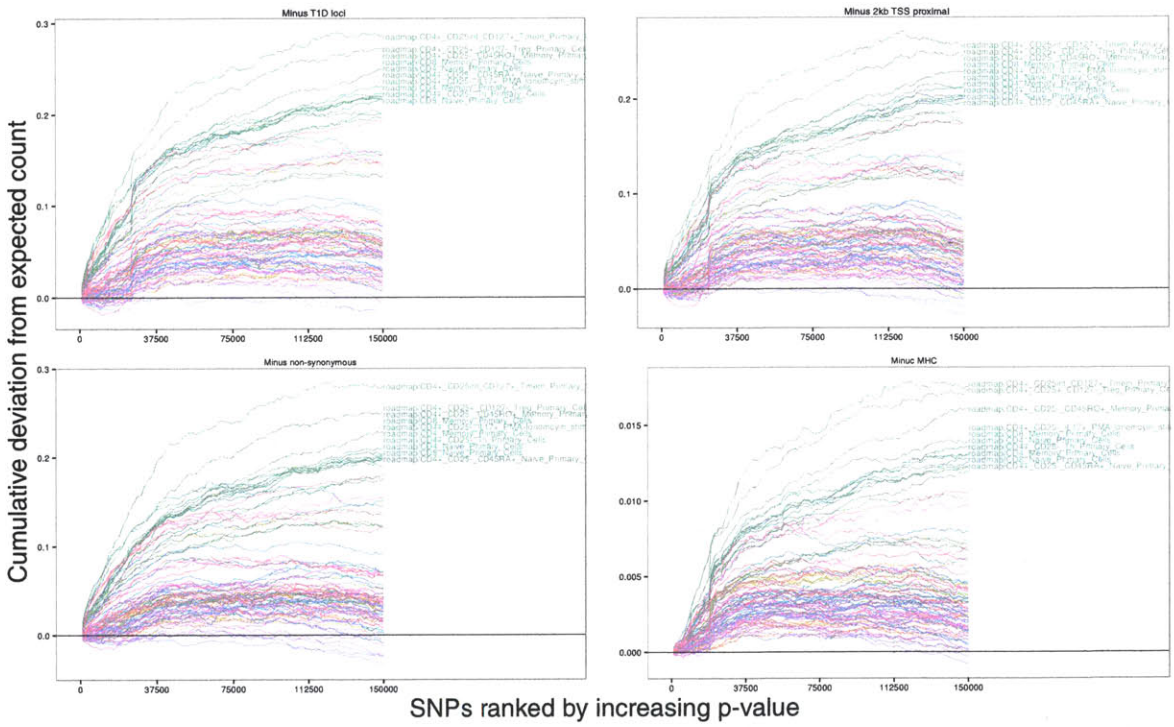

Figure 3.5: Enrichment of enhancers persists after subtracting out potential confounders such as known loci, nearby genes, coding variants, and the MHC. Immune cell types continue to be separated from other cell types even up to 30,000 SNPs.

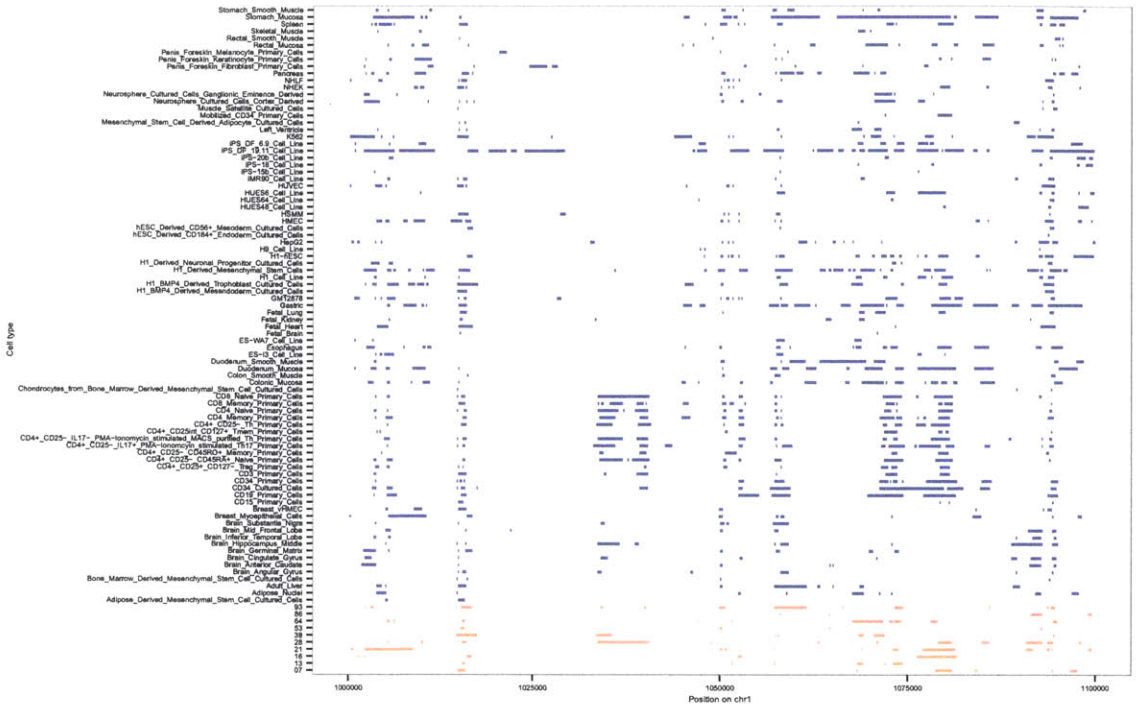

Figure 3.6: Generation of permuted enhancer tracks (orange) from real enhancer tracks (blue). The permutation procedure samples from real elements and therefore preserves their properties. Sampling is more likely to pick constitutive enhancers because they appear more often in the population of all elements, allowing us to investigate the contribution of these enhancers. We destroy the association between the identity of the cell type and which enhancers are assigned to it and hypothesize permuted tracks will show less enrichment than real tracks.

3.2.4 Implicated enhancers are cell type–specific

Next, we look at whether the enhancers we found are actually specific to the immune cell types or are *constitutive* (constant across all cell types). We first permute elements across cell types to generate 100 new randomized cell types as shown in Figure 3.6. We first sample a total number of elements from the distribution of total number of elements across real cell types, then sample that number of elements from all elements across all cell types. We hypothesize the identity of which enhancers are in a specific cell type is the important quality. Therefore, our permutation procedure destroys this quality while preserving other properties of the elements (such as the distribution of their sizes and distances to the closest gene).

Figure 3.7 shows RR plots computed for these permuted cell types. We find all of the randomized cell types show moderate enrichment early in the ranked list suggesting this quantity is the contribution of constitutive enhancers. However, the enrichment is of lesser magnitude than the observed enrichment for actual cell types. This result suggests when we look at real cell types we are finding enrichment beyond just that of constitutive

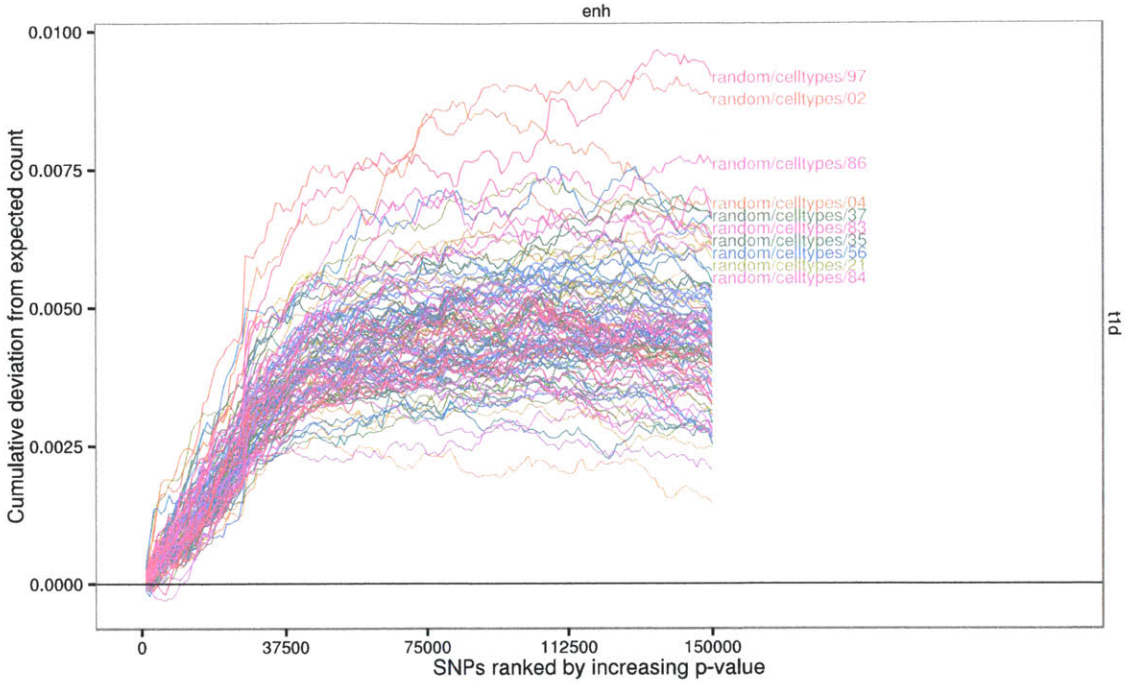

Figure 3.7: Enrichment of permuted enhancer tracks. As in the case of promoters, permuted tracks show some enrichment in all cell types, suggesting they represent the contribution of constitutive enhancers. However, the observed enrichment is less than that for real enhancer tracks.

enhancers. Moreover, as in the case of promoters none of the cell types is separated from the rest when considering enrichment of constitutive enhancers.

We next look at *enhancer clusters*. To perform the clustering, we sweep a line over the concatenated genome. For each intersected enhancer, we take the union of enhancers intersecting that enhancer across all the other cell types. For that region, we compute a binary vector specifying whether the region overlaps an enhancer in each cell type. We cluster these activity vectors using k-means clustering, iteratively picking optimal k. For each cluster, we generate a new pseudo-cell type containing the regions corresponding to the activity vectors assigned to that cluster.

Each cluster captures a set of enhancers which is specific to a set of cell types. Figure 3.9a shows for each cluster the cell types which the set of enhancers is active in. Red indicates strong enhancer activity, orange weak, and purple poised activity. For example, cluster 19 captures constitutive enhancers which are active in all cell types whereas cluster 5 captures enhancers which are active only in a small set of T helper and memory T cells.

Figure 3.8 shows the enrichment of these clusters for T1D–associated variants. We find several clusters are enriched, showing clear separation from other clusters. As shown in Figure 3.9b, the most enriched clusters are largely specific to immune cell types. One cluster represents the contribution of constitutive enhancers and shows strong enrichment. However, the cell type–specific clusters contain enhancers active in only two broad types

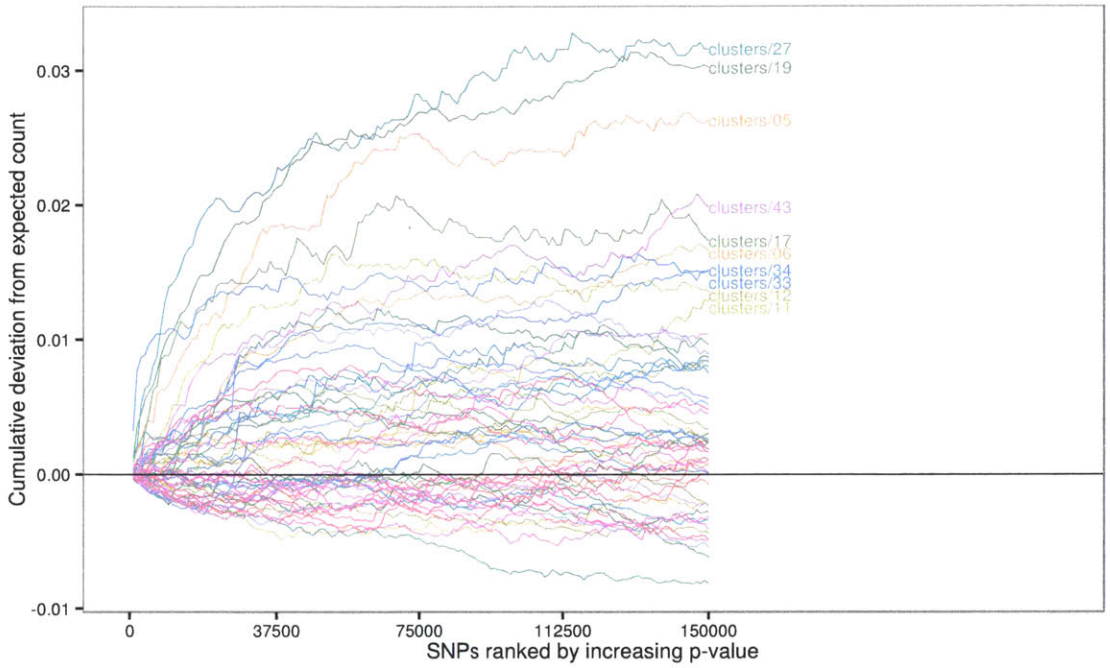

Figure 3.8: Enrichment of enhancer clusters for T1D–associated variants. Several clusters are clearly separated from the others. Moreover, we continue to see a greater magnitude of enrichment persisting to tens of thousands of SNPs.

of cells. The first are memory T cells, which recognize and respond to invasion. The second are T helper cells, which are initially programmable to respond to new invaders. After encountering a new *antigen* (cell surface marker which can be used to identify invaders), they mature into either memory cells, effector cells which increase immune response when exposed to the same antigen, or regulatory cells which decrease immune response. Misclassification of one's own cells and disregulation of the immune response at the tissue level are central to autoimmune disorders such as T1D. These results suggest we are indeed finding enhancers which are specifically active in exactly the disease-relevant cell types. They could play a role in disregulation at the molecular level, modulating the expression of important genes and function of important pathways which could give rise to the observed tissue-level disregulation.

3.2.5 Open chromatin is enriched for disease-associated variants

We have chromatin states for 9 ENCODE cell types and 85 Roadmap cell types. However, these cell types are still only a fraction of the full spectrum of human cell types. For other cell types we do not have this rich annotation of regulatory activity. However, for 90 different ENCODE cell types we have experimental assays of open chromatin, another proxy for regulatory activity. Specifically, we have annotations of DNAseI hypersensitive sites (DHS) and Digital Genomic Footprints (DGF).

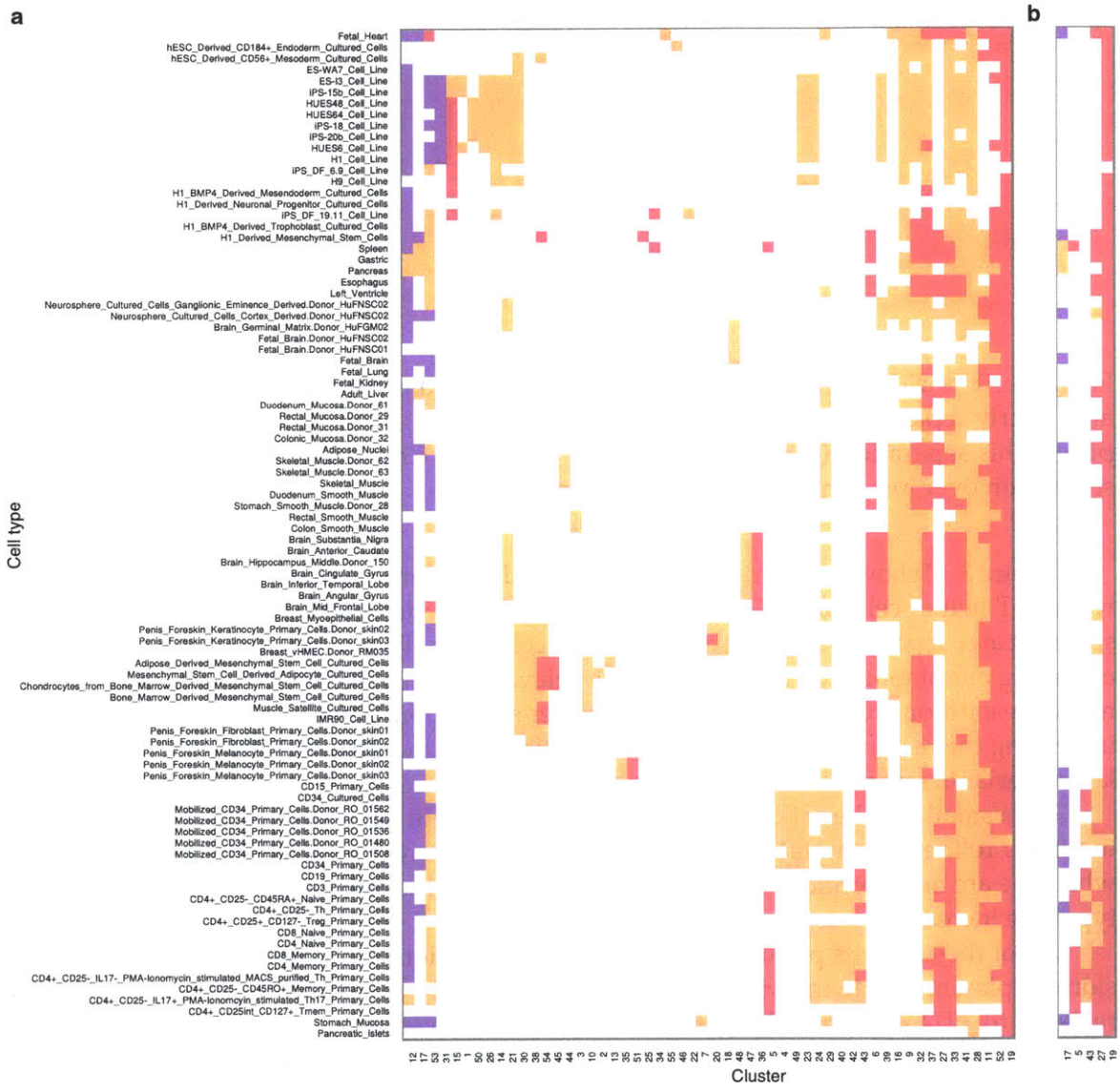

Figure 3.9: Discretized activity profiles of enhancer clusters. a) Profiles across all clusters. Red, orange, and purple denote strong, weak, and poised enhancers respectively. White represents inactive or missing data. b) Focusing on the top clusters enriched for T1D–associated variants, we find one constitutive cluster. The others show activity more specifically in immune cell types.

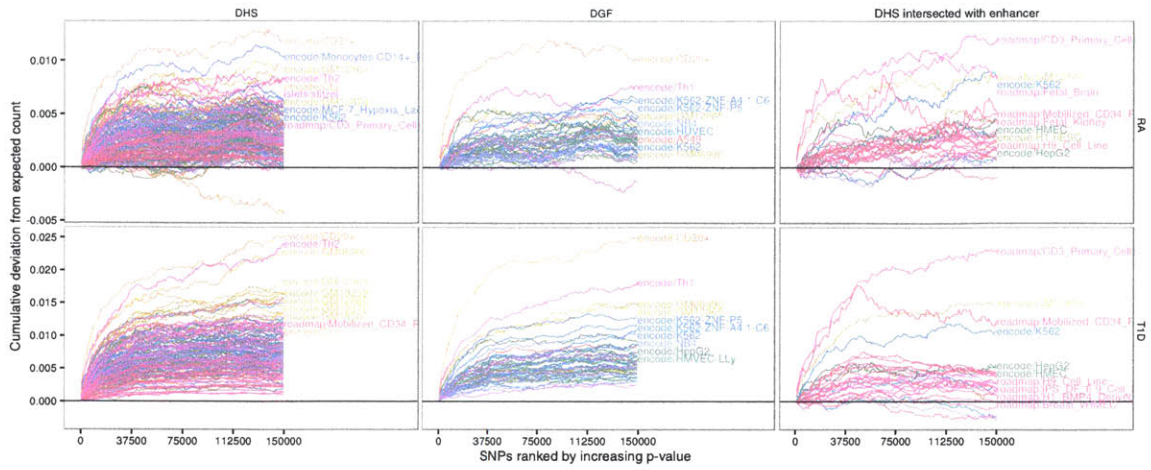

Figure 3.10: Enrichment of open chromatin. Enrichment of DHS for T1D–associated variants clearly separates immune cell types from others (left). The higher resolution DGF improves this separation (center). Intersecting DHS with ChromHMM enhancers improves separation over just considering enhancers (right).

Figure 3.10 shows the enrichment of DHS and DGF for T1D–associated variants. Again we find T helper cells and $CD20^+$ B cells are enriched. However, we find these cell types show greater separation from other irrelevant cell types. This result shows the importance of having annotations for the most disease-relevant cell types. Moreover, we find that the higher-resolution DGF annotation shows better separation between cell types, suggesting increasing the resolution of regulatory annotations will increase our power to detect regulatory variants relevant to disease.

For a small number of cell types, we have both chromatin states and DHS, allowing us to further refine the annotation. By intersecting enhancer regions with DHS, we get regions which we are more confident have regulatory function. Indeed, we again find immune cell types are increasingly separated from other cell types.

All of these results point to molecular level regulation (i.e., regulation of gene expression) as playing a role in Type 1 Diabetes. We find only enhancers in immune cell types relevant to T1D show enrichment for T1D–associated variants. This enrichment persists to tens of thousands of SNPs, suggesting the genetic architecture of T1D involves many more common variants than previously thought. We separate the contribution of these enhancers from the linked contributions of nearby genes and find it is indeed the regulatory regions which are contributing to the observed enrichment. Moreover, we separate the contribution of constitutive enhancers from the contribution of enhancers specific to the enriched cell types. We find it is precisely those enhancers specific to disease-relevant immune cell types which show the strongest enrichment. Thus, these variants could contribute to molecular level disregulation, which in turn gives rise to tissue level disregulation of the immune response and the autoimmune disease pathology.

Chapter 4

Pathway analysis

In the previous chapter we identified a large set of of variants which are associated with disease in enhancer regions. Now we develop mechanistic hypotheses about how those variants actually contribute to disease. In particular, we want to find gene pathways which these variants and regions are over-represented in.

4.1 Methods

4.1.1 Region-based tests for enrichment of pathways

We use the Genomic Enrichments of Annotations Tool (GREAT) to test whether regulatory regions are over-represented in regulatory domains of genes [26]. We use the default settings, which estimates the regulatory domain of a gene as 5 kb upstream and 1 kb downstream of the TSS, extended to a maximum of 1 Mb upstream. However, the GREAT web service (`http://great.stanford.edu`) also uses some experimentally validated regulatory domains. It incorporates twenty gene ontologies and has been shown to produce robust and specific results.

GREAT uses two tests for over-representation of input regions in regulatory domains. The first is based on a binomial model. For a given ontology term, the parameter p is the fraction of the genome covered by the regulatory domains of genes with that term. The parameter n is the number of background regions (the foreground is required to be a subset of the background). Then, the observed number of overlaps is $\text{Binom}(n, p)$ and the probability of observing at least that many overlaps can be computed using the probability mass function.

The second test is based on a hypergeometric model. A 2×2 contingency table partitions genes into those with a given term and those without and genes whose regulatory domain intersects an input region and those which do not. The observed counts follow the hypergeometric distribution, for which a p-value can be computed using Fisher's exact test. By computing both tests, GREAT separates truly enriched pathways from pathways with highly enriched genes (which drives the binomial p-value down) or biased regulatory domains (which drives the hypergeometric p-value down).

4.1.2 SNP–based exact rank sum test for enrichment of pathways

We are also interested in identifying gene sets for which the associated variants are skewed to have lower p-values. Rather than scoring and computing statistics on genes, we want to annotate and compute statistics on variants. We assign each variant to the closest gene because for most cell types we do not have gene expression data. We then ask for each pathway in the Molecular Signatures Database [36] whether variants annotated with that pathway are skewed to have lower p-value than variants which are not.

We previously described the application of the Mann–Whitney U and Kolmogorov–Smirnov two-sample D tests to perform this analysis We also described how these tests are inappropriate for answering this question due to the overly strong independence assumptions they make. In particular, variants which are close by are more strongly linked by LD and therefore have more strongly correlated p-values. Also, variants which are close by are more likely to be assigned the same pathway annotation.

Instead, we develop a permutation test based on the Mann–Whitney U test to compute exact p-values. The test statistic itself is simply the sum of ranks of variants with the pathway annotation. Suppose we have a vector of ranks $\mathbf{r} \in \mathbb{Z}^n$ of variants arranged in genomic order (concatenating chromosomes) and a vector of binary (0, 1) annotations $\mathbf{a}$. Then the statistic is:

$$X = \mathbf{r} \cdot \mathbf{a}$$

The permutation procedure preserves the clustering properties of the p-values (ranks) and annotations, but destroys the association between the two. For the ith rotation of the vector $\mathbf{a}$, we compute the statistic X_i in the same fashion. Then, the exact p-value is

$$p = \frac{1}{\|\mathbf{r}\|} \sum_{i=1}^{n} \mathbb{1}(X_i < X_0)$$

4.1.3 Related work

There are two broad categories of pathway analysis: self-contained tests and competitive tests [45]. The self-contained null hypothesis is the gene set is not associated with the phenotype. The competitive null hypothesis is the genes in a gene set are not more associated with phenotype than genes not in the gene set. As in the previous chapter, we focus on the competitive null hypothesis.

There are two main approaches for testing the competitive null hypothesis: gene-based tests and SNP-based tests. We use a SNP-based approach, assigning gene ontology terms to SNPs and computing statistics on SNPs. There are several SNP-based tests in the literature. GSEA-SNP is based on GSEA and computes a similar running-sum enrichment score based on walking down the ranked list of p-values [27]. The significance of this score is computed using an exact test based on permuting phenotype labels, recomputing association p-values, and then recomputing the enrichment score. Unlike GSEA-SNP, our method does not down-weight the contribution of SNPs as their rank increases. Our method is also less computationally expensive than permuting phenotype labels. SNPtoGO [4] and GESBAP [17] use Fisher's exact test to test the over-representation of genes in a pathway at extreme ends of the ranked list. However, such a strategy requires

one hypothesis test per cutoff and therefore more stringent multiple testing correction. We only perform one hypothesis test per pathway–cell type combination.

4.2 Results

4.2.1 Enrichment of T1D–associated variants

We first asked which pathways are assigned SNPs with lower p-values. We approximated the exact rank sum test by computing z-scores for the statistic computed under 1,000 randomly chosen rotations of the annotation vector. However, only few pathways are significantly enriched, and none remain after Benjamini–Hochberg FDR correction (FDR $< .05$).

4.2.2 Enrichment of enhancer regions

We next investigated whether enhancer regions were enriched for pathways. We used GREAT to test enrichment of enhancer tracks in the 9 ENCODE and 90 Roadmap cell types using the default settings (maximum extension of regulatory domains). GREAT failed to complete using a background of all enhancer regions in all cell types, so we used the whole genome as the background.

We restrict to the GO biological process terms. We throw out cell type–pathway combinations for which the FDR q-values for either the binomial and hypergeometric tests is greater than .001, leaving 2,817 combinations. We clustered cell types and pathways independently using hierarchical clustering with a Euclidean distance metric. Figure 4.1 shows selected portions of the heatmap of enrichment for cell type–pathway combinations.

We find modest enrichment of a large number of pathways across all cell types. The strongest enrichments are in a cluster of generic categories including transcriptional regulation, macromolecular disassembly, and viral translation.

We expect transcriptional regulation to be enriched because we are considering enhancer regions. Macromolecular disassembly is a generic GO term which includes translational termination. Post-translational modification has been implicated in the development of T1D autoimmune response [29]. We find further support for the enrichment of the these categories in a cluster of enriched pathways in a cluster of immune cells. In particular, we find enrichment for the term protein modification by small protein conjugation.

We also find modest enrichment of relevant immune pathways in the top immune cell types identified through genome-wide enrichments. Specifically, we find enrichment for terms such as positive regulation of immune response, T cell activation, and leukocyte activation. Moreover, we find enrichment for more specific terms such as immune response-activating cell surface receptor signaling pathway.

4.2.3 Enrichment of T1D–associated regulatory variants

Given enhancer regions are enriched for relevant high-level categories, we next asked whether restricting to these regions could improve our SNP–based results. We used the exact rank sum test against a background set of variants which fall in enhancer regions

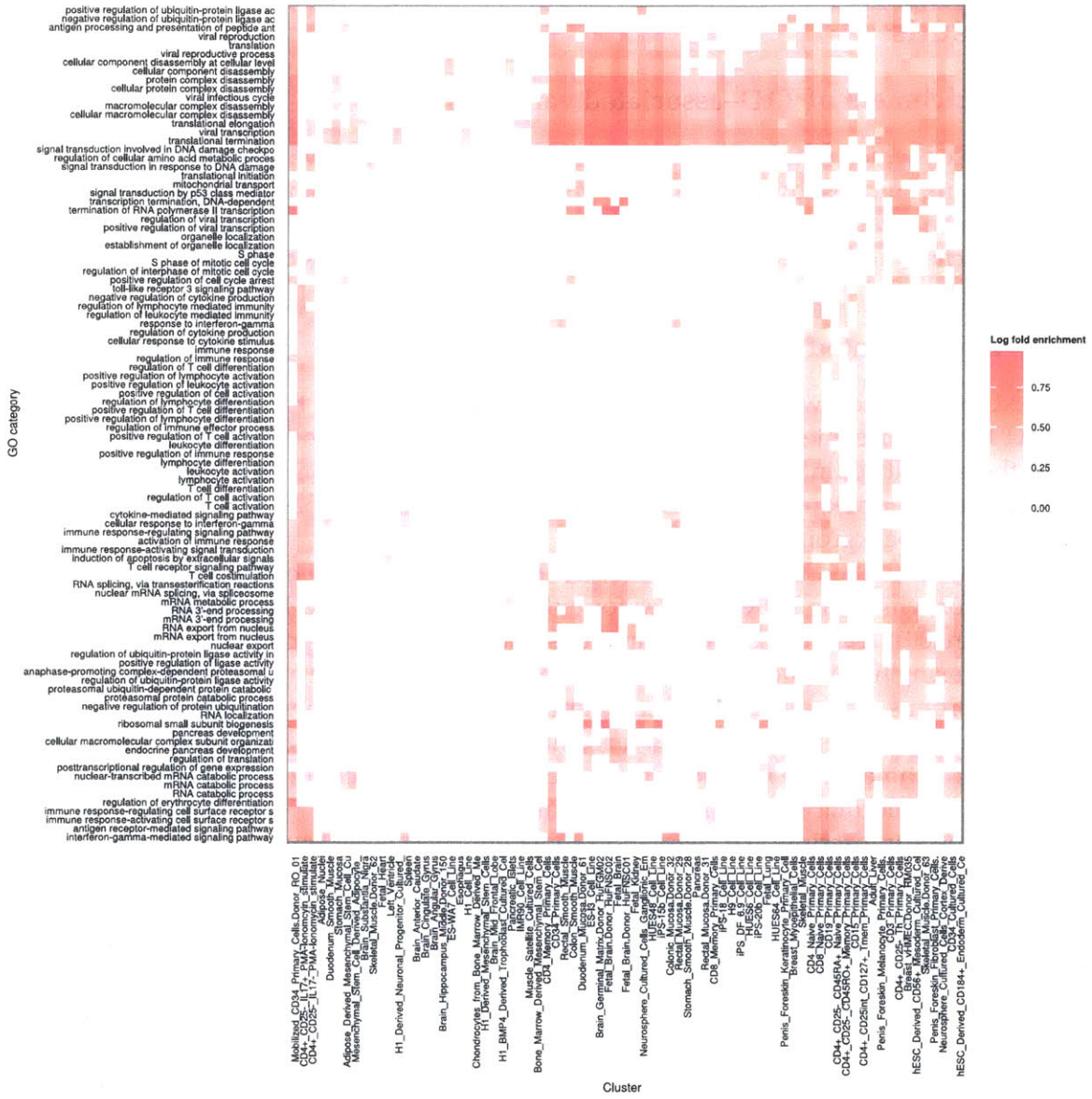

Figure 4.1: Enrichment of enhancers for selected GO terms. Cell types and terms were independently hierarchically clustered based on patterns of fold enrichment. We find large blocks for terms related to transcription (top right), immune response (center right), and post-transcriptional regulation (bottom right). We also find a cluster of T_{reg} cell types with various surface markers are enriched for all of these terms (left).

in any cell type. We found few pathways were enriched after FDR correction. However, those pathways were not obviously relevant to regulatory or disease function.

We then asked whether we could improve the region-based enrichment by restricting to enhancers containing T1D–associated variants. We took the top 30,000 SNPs (our empirical cutoff), and for each cell type used GREAT to test enrichment of only those enhancers containing one of those top SNPs. We again restricted to cell type–pathway combinations with binomial and hypergeometric $q < .01$, leaving 294 combinations. Figure 4.2 shows when we consider this subset of disease relevant regulatory elements, we find a large cluster of immune cell types is enriched in a large cluster of immune response–related GO categories. In addition to the immune pathways found before we find 300-fold enrichment for antigen processing and 30-fold enrichment for interferon gamma–mediated signaling pathway. The gene *PTPN2* which is associated with T1D is thought to modulate pancreatic cell apoptosis in response to interferon gamma [28]. Pancreatic beta cells are responsible for the production of insulin; their destruction is part of the known T1D pathology.

We again find a cluster of weaker enrichments in the top immune cell types for terms positive regulation of T cell, leukocyte, and lymphocyte proliferation; B cell activation; and positive regulation of immune response. However, we also find 4-fold enrichment for regulation of interleukin production. Disregulation of the IL-2 pathway is a central defect in T1D pathology [16]. Low dosage of IL-2 has been shown to offer protection from the disease; however, high doses increase the autoimmune response.

They key observation here is we identified these pathways starting from non-coding variants rather than coding variants. We know a large fraction of the top 30,000 T1D hits are not near coding regions at all and therefore tag novel regulatory variants. We are not merely recapitulating known T1D–relevant pathways in the literature by identifying non-synonymous mutations in the genes contained in those pathways. Rather, we are potentially identifying novel disease-relevant links in the regulatory networks of these pathways.

4.2.4 Enrichment of enhancer clusters

Next, we asked whether enhancer clusters would show these same enrichments for specific terms. In the previous chapter we showed top T1D–associated variants were overrepresented in cell type–specific enhancers. The question is whether we gain anything by starting from the variant associations instead of going straight to *a priori* disease-relevant cell types.

Restricting to binomial and hypergeometric $q < 1e8$, we find 6,221 enriched cluster–pathway combinations. These combinations show modest enrichments (order 10-fold) in a variety of high level terms. Moreover, we do recapitulate pathways we found by using top T1D SNPs to filter enhancers as shown in Figure 4.3. If we look specifically at the clusters which showed enrichment for T1D–associated variants in the previous chapter, we find many of the same enrichments for specific disease-relevant terms. However, the fold enrichment is lower because the number and size of regions considered is higher.

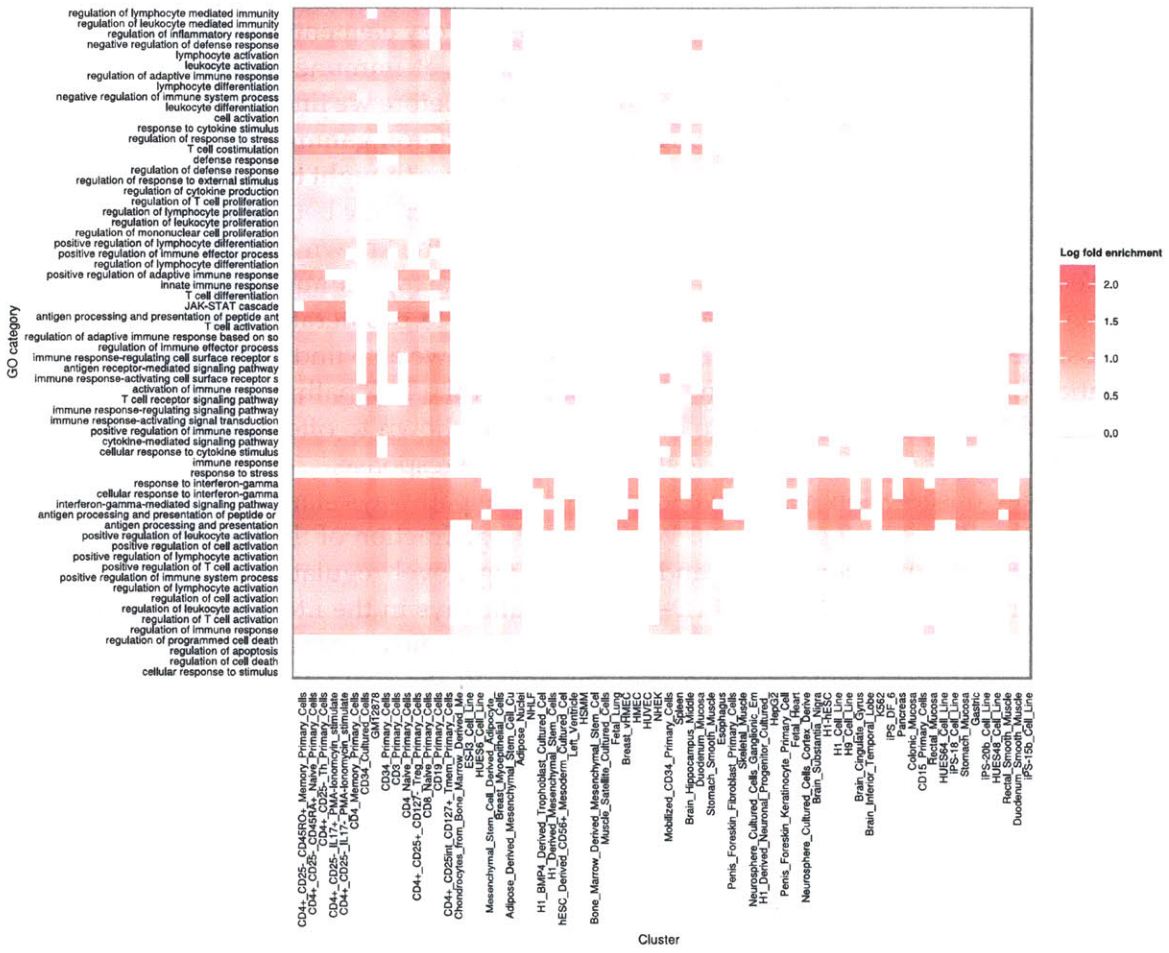

Figure 4.2: Enrichment of enhancers containing top T1D–associated variants for selected GO terms. Enrichments are an order of magnitude greater when restricted to this subset of regions. We find a larger block of stronger enrichments (left) for immune cell types in the previously described pathways. We also find strong enrichment for response to interferon gamma across all cell types.

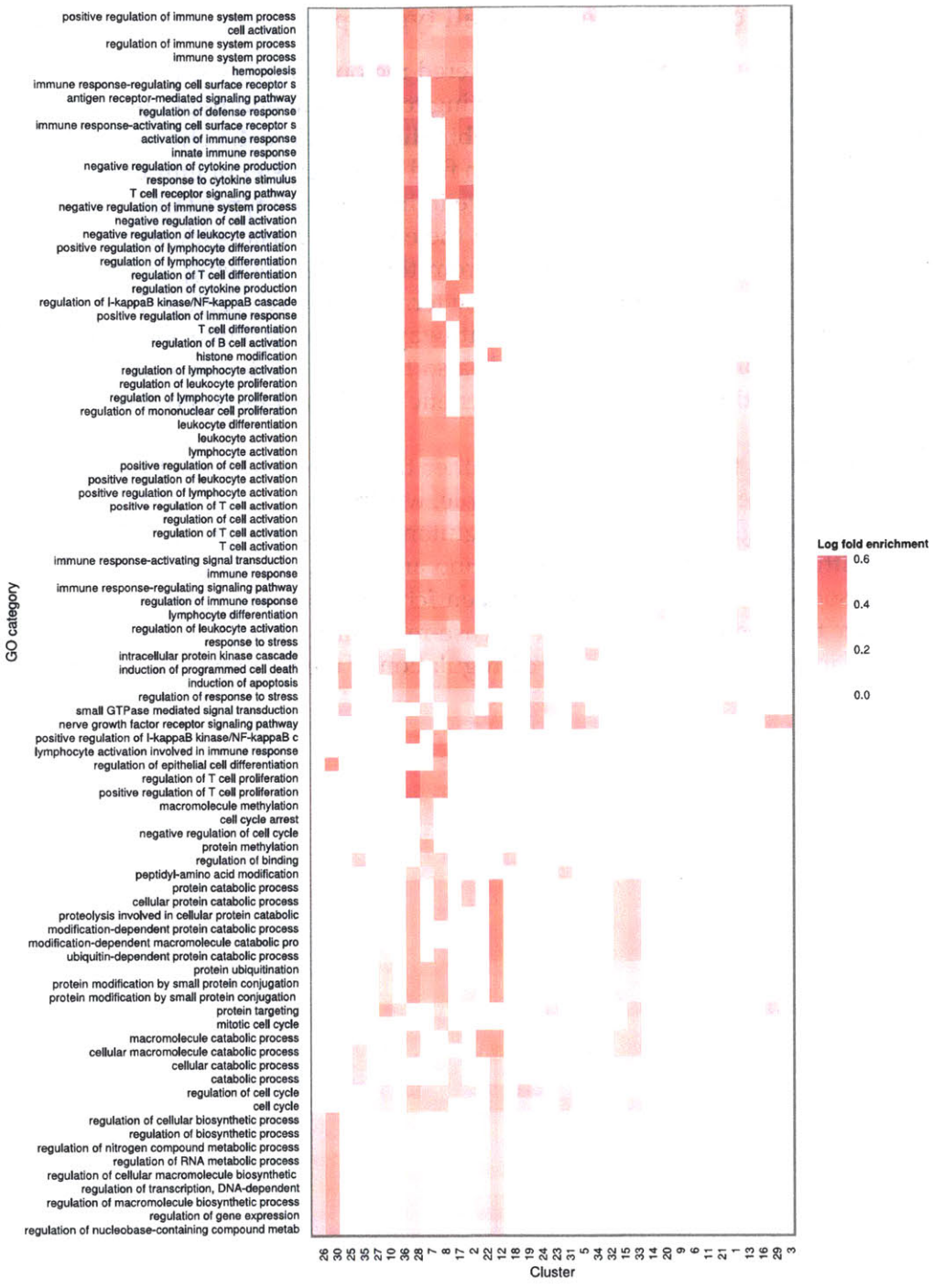

Figure 4.3: Enrichment of clustered enhancers for selected GO terms. Many cluster–pathway combinations show just over 1 fold enrichment (not shown). The strongest enrichments are again in a large cluster of immune response pathways, but they show relatively low fold enrichment.

4.2.5 Enrichment of T1D–associated cell type–specific regulatory variants

Finally, we asked whether restricting to clustered regions containing T1D–associated variants could improve the enrichment. Restricting to binomial and hypergeometric $q < .01$, we find 294 enriched cluster–pathway combinations. Figure 4.4 shows the heatmap of enrichment for these combinations. We again find a large cluster of immune pathways enriched in clusters which are specifically active in immune cell types (although some show weak activity in all cell types). Focusing on the enriched clusters identified in the previous chapter (Figure 3.9b), we find regions from the cluster showing constitutive activity (19) does not show enrichment in any of these terms. However, regions from the immune cell type–specific clusters are enriched for pathways involving leukocyte and T cell differentiation, proliferation, and activation. We also find one cluster enriched for chromatin modification, suggesting regulation of epigenetic modifications could also play a role in the disease.

By starting from the empirical cutoff rather than from enriched clusters directly, we also find some novel enrichments. For example, we find over 1,000-fold enrichment of the term regulation of L-glutamate transport. L-glutamate is a known target of autoantibodies associated with pancreatic beta cell destruction and onset of T1D [31] Again, by starting from T1D–associated variants we find novel enrichments which suggest disruption of regulatory rather than coding function in disease-relevant pathways. This molecular level disregulation in turn could give rise to disregulation of the immune response and the disease pathology.

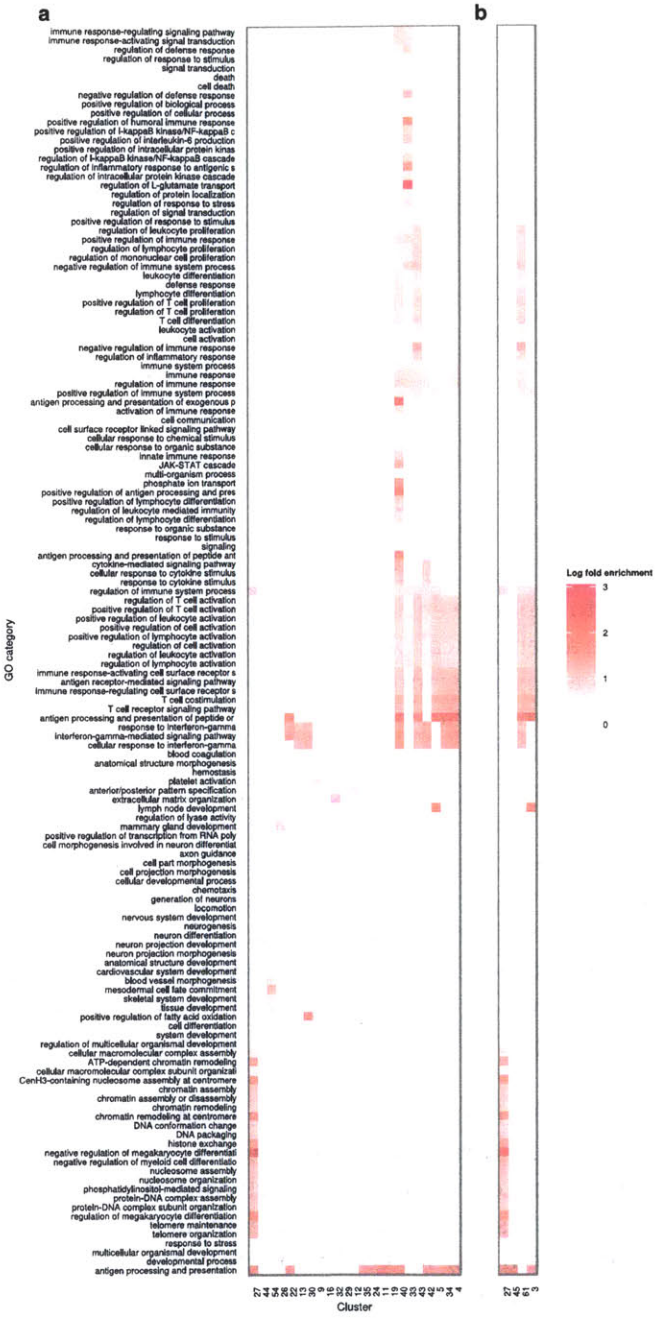

Figure 4.4: Enrichment of clustered enhancer regions containing top T1D–associated variants. a) Restricting to this subset of regions, we find much stronger enrichment for immune pathways. We also find strong enrichment of L-glutamate transport, which we did not find using any other set of regions. b) Focusing on clusters showing genome-wide enrichment for T1D–associated variants identified in the previous chapter, we find the immune–specific clusters show enrichment in more relevant pathways than the constitutive cluster.

Chapter 5

Discussion

The first goal of this thesis was to develop methods to identify common regulatory variants below the usual stringent GWAS p-value cutoffs. Towards this end, we developed the RR plot visualization and exact ranksums test. The RR plot uses a novel definition of cumulative deviation to separate enriched cell types from non-enriched cell types while also allowing us to gauge an empirical p-value cutoff beyond which enrichment stops. The exact ranksums test uses a novel permutation procedure (rotating vectors of annotations) to preserve local clustering properties of the genome while destroying the association between regulatory regions and variants.

These methods have some weaknesses which future work needs to address. The definition of cumulative deviation is a nonstandard one which is harder to interpret at a glance. In particular, the magnitude of deviation is not directly comparable between classes of functional elements due to its dependence on the total number of overlaps. Moreover, the magnitude is only comparable between cell types assuming there are roughly the same number of overlaps per cell type (which does appear to be true). One natural alternative statistic is fold enrichment; however, this statistic is also sensitive to the total number of overlaps and is moreover highly variable early in the curve due to the small number of SNPs seen.

To improve the interpretation of RR plots, we need to attach some notion of significance to the enrichment. One strategy for computing significance is to develop sets of matched controls and recompute the curves for each control set. Control sets would match the real overlaps for properties such as minor allele frequency, LD block size, and distance to closest TSS. Based on the curves, we could compute z-scores for each point on the RR plot, which are both directly comparable across features/cell types and easier to intuit significance from. However, the major weakness of this strategy is we do not know *a prior* which properties of the genome are necessary to match. Moreover, as the list of properties grows the number of matching sets goes down, potentially reducing the number of unique control sets which can be produced.

A more robust strategy is to permute phenotype labels, recompute association statistics, and then recompute the curves. This approach produces a more obviously meaningful null distribution. The trial we are computing a statistic over consists of generating a vector of association statistics and then intersecting it with a vector of annotations. If we were to actually repeat the trial, we would only generate a new vector of association statistics;

everything else would remain constant. Moreover, this strategy automatically preserves all properties of the genome since only phenotype labels change. However, this method is more computationally expensive and also is sensitive to the choice of association statistic.

The second goal was to apply these methods to the WTCCC1 T1D GWAS and evaluate their efficacy. Indeed, we find enrichment of enhancers persists to 30,000 top SNPs and beyond, dwarfing usual estimates of the number of common variants involved in complex disease by two orders of magnitude. Unlike enrichment of promoters, enrichment of enhancers is highly dynamic across cell types and separates disease–relevant cell types (immune cells) from irrelevant cell types. Moreover, this enrichment persists after correcting for various confounders such as known T1D risk loci, nearby genes, non-synonymous variants, and the MHC.

Surprisingly, of the top 30,000 SNPs, fewer than 2,000 actually fall in enhancer regions in any single cell type. Even fewer fall in promoters and coding regions, raising the following question: why do we have to consider 30,000 SNPs before enrichment stops when only a small proportion of those SNPs have potential function? One possibility is non-robust p-values disrupts the ranking of SNPs. The current lack of GWAS power due to sample sizes and population genetic biases could cause such non-robustness. Another possibility is these variants are functional in cell types whose epigenomes have not yet been profiled.

Future work in this area should address three concerns. First, we should attach significance to the observed enrichments as described above.

Second, we should expand our catalog of common SNPs from HapMap 3 to Thousand Genomes. This expansion requires re-imputation of the data to the Thousand Genomes reference panel. Thousand Genomes includes over double the common SNPs as HapMap. By imputing to 5.6 million common SNPs, we increase the likelihood we find causal variants directly instead of just tag SNPs.

Along these lines, we must also consider the role of rare, private mutations in disease such as T1D. Imputation to the full Thousand Genomes panel is one step towards this goal. On another front, the reduction in price of the latest generation of sequencing technologies has made full genome sequencing and exome sequencing in large cohorts of cases and controls possible. In particular, the Genetics of Type 2 Diabetes consortium has conducted such a study in 2,700 individuals [18]. This study will not only give us unparalleled power to detect causal variants, but will also give us the ability to query the genetic architecture of complex disease. In particular, we can start to ask questions about the combinatorial interaction of variants in disease, and learning functions which describe these interactions.

Third, we should apply these methods to larger cohorts with richer phenotyping. The T1D Exchange (`http://www.t1dexchange.org`) is a non-profit organization bringing together data on tens of thousands of patients across 65 American clinics. The data collection was designed specifically for downstream analysis aimed at identifying causal variants. Towards this end, patients are genotyped at high density in the MHC as well as other relevant parts of the genome. They are also richly phenotyped for a large number of metabolic and molecular traits as well as environmental and behavioral covariates. This resource will give us unparalleled power in our analyses, as well as the opportunity to study problems such as improving the power of association by considering multiple correlated phenotypes.

The third goal of this thesis was to interpret the function of the identified variants by

identifying pathways they were enriched in. Using GREAT, we found enhancer regions alone are already enriched for some disease–relevant pathways. However, by subsetting to regions containing one of the top 30,000 T1D–associated SNPs, we found greater enrichment in more specific GO terms. Moreover, by starting from associated variants falling in regulatory regions, we specifically implicate regulatory mechanisms rather than coding mutations.

Future work in this area should first investigate enrichment of transcription factor motifs for T1D–associated variants. Preliminary work showed no enrichment for variants falling in sequence motifs directly; however, current work shows enhancer regions containing motifs are enriched for reported cardiovascular disease–associated loci. By applying our methods for genome-wide enrichment, we should be able to attach more specific function to individual regions which we previously implicated.

Second, we should use expression data to link enhancers to the genes they target. It is not clear what proportion of naive enhancer links (assigning to the closest gene) are actually correct, which could severely impact the results of our SNP–based test for gene set enrichment. GREAT makes the same assumptions about the regulatory domains of genes, which may impact the region-based results as well. Instead, we should link enhancers to gene targets by correlating enhancer activity across cell types with expression across cell types. By using such higher confidence assignments we gain higher confidence hypotheses about how variants in those regions affect gene pathways.

Third, we should study methylation as another potential mechanism by which variants affect gene expression. Methylation is associated with repression of genes, and it is known that variants can affect methylation status. Recently, an epigenome-wide association study found methylation mediates the association between genotype and rheumatoid arthritis (RA) [22]. Along these lines, we should correlate the regulatory regions and pathways identified based on SNP associations in the WTCCC1 RA GWAS alone to those identified by methylation associations for the same disease. By doing so, we can generate more specific hypotheses about how non-coding variants contribute to complex disease.

We have begun to shed some light on the genetic architecture of complex disease, but there is still much work to be done. In particular, we need to estimate the additional heritability explained by the regulatory variants we have identified. The usual heritability estimate involves studies of inheritance of the disease in cohorts of families; however, in our case we do not have such data.

There are two types of approaches to estimating heritability from GWAS data directly. The first is using restricted maximum likelihood estimation to fit a linear model from genotypic variance (characterized by a genetic relationship, or kinship, matrix) to phenotypic variance [19]. When applied to the 500,000 genotype probes in the WTCCC1 T1D GWAS, the model estimates the SNPs explain 37% of the variance in phenotype [21]. However, the method is sensitive to confounding sources of genetic variation such as genotyping error, population structure, and cryptic relatedness between individuals. This estimate is obtained after throwing out over half the probes and one third of the samples and is a very conservative estimate. A second approach learns a polygenic model to estimate heritability [35]. When applied to T1D, it estimates 47% of the heritability is explained by common SNPs.

These methods have additional pitfalls when applied to imputed data. In particu-

lar, non-perfect LD ($r^2 < 1$) affects the observed distribution of minor allele frequencies and confounds estimates of genetic variance. One potential approach is to restrict to independent tags and assign the union of annotations to these tags. However, this approach assumes only one causal variant per locus which would be surprising if true.

There is a long road from studies of genetics to novel diagnostics and therapeutics. However, the first step is identifying variants and generating hypotheses about their function. Interpreting GWAS has been the major challenge of the last 10 years. In parallel, the study of epigenomics has shown the genome is rich in function beyond just the small fraction which codes for proteins. This thesis contributes methods for combining genetic and epigenetic data to generate hypotheses and the results of evaluating them on T1D. All indications suggest the regulatory variants we have identified play a role in the disease and could present novel diagnostic and treatment targets.

Bibliography

[1] ALTSHULER, D., DALY, M. J., AND LANDER, E. S. Genetic mapping in human disease. *Science 322*, 5903 (2008), 881–888.

[2] BERNSTEIN, B. E., STAMATOYANNOPOULOS, J. A., COSTELLO, J. F., REN, B., MILOSAVLJEVIC, A., MEISSNER, A., KELLIS, M., MARRA, M. A., BEAUDET, A. L., ECKER, J. R., FARNHAM, P. J., HIRST, M., LANDER, E. S., MIKKELSEN, T. S., AND THOMSON, J. A. The NIH Roadmap Epigenomics Mapping Consortium. *Nature Biotechnology 28*, 10 (Oct 2010), 1045–1048.

[3] D.BOTSTEIN, R.L.WHITE, M.SKOLNICK, AND R.W.DAVIS. Construction of a genetic linkage map in man using restriction fragment length polymorphisms. *American Journal of Human Genetics 32* (May 1980), 314–331.

[4] D.F.SCHWARZ, O.HADICKE, J.ERDMANN, A.ZIEGLER, D.BAYER, AND S.MOLLER. Snptogo: characterizing snps by enriched go terms. *Bioinformatics 24* (2008), 146–148.

[5] D.HILBERT. ber die stetige abbildung einer linie auf ein flchenstck. *Mathematische Annalen 38* (1891), 459–460.

[6] EAST, E. A mendelian interpretation of variation that is apparently continuous. *The American Naturalist 44*, 518 (Feb 1910), 65–82.

[7] ECKER, J. R., BICKMORE, W. A., BARROSO, I., PRITCHARD, J. K., GILAD, Y., AND SEGAL, E. Genomics: Encode explained. *Nature 489*, 7414 (Sep 2012), 52–55.

[8] EICHLER, E. E., FLINT, J., GIBSON, G., KONG, A., LEAL, S. M., MOORE, J. H., AND NADEAU, J. H. Missing heritability and strategies for finding the underlying causes of complex disease. *Nature Reviews Genetics 11*, 6 (2010), 446–450.

[9] ERNST, J., AND KELLIS, M. Discovery and characterization of chromatin states for systematic annotation of the human genome. *Nature Biotechnology 28*, 8 (2010), 817–825.

[10] ERNST, J., KHERADPOUR, P., MIKKELSEN, T. S., SHORESH, N., WARD, L. D., EPSTEIN, C. B., ZHANG, X., WANG, L., ISSNER, R., COYNE, M., KU, M., DURHAM, T., KELLIS, M., AND BERNSTEIN, B. E. Mapping and analysis of chromatin state dynamics in nine human cell types. *Nature 473*, 7345 (2011), 43–49.

[11] FISHER, R. The correlation between relatives on the supposition of mendelian inheritance. *Earth and Environmental Science Transactions of the Royal Society of Edinburgh 52* (0 1919), 399–433.

[12] GUSELLA, J. F., WEXLER, N. S., CONNEALLY, P., NAYLOR, S. L., ANDERSON, M. A., TANZI, R. E., WATKINS, P. C., OTTINA, K., WALLACE, M. R., SAKAGUCHI, A. Y., YOUNG, A. B., SHOULSON, I., BONILLA, E., AND MARTIN, J. B. A polymorphic dna marker genetically linked to huntington's disease. *Nature 306*, 5940 (Nov 1983), 234–238.

[13] HEMMINKI, K., LI, X., SUNDQUIST, J., AND SUNDQUIST, K. Familial association between type 1 diabetes and other autoimmune and related diseases. *Diabetologia 52* (2009), 1820–1828. 10.1007/s00125-009-1427-3.

[14] HESSELBERTH, J. R., CHEN, X., ZHANG, Z., SABO, P. J., SANDSTROM, R., REYNOLDS, A. P., THURMAN, R. E., NEPH, S., KUEHN, M. S., NOBLE, W. S., FIELDS, S., AND STAMATOYANNOPOULOS, J. A. Global mapping of protein-dna interactions in vivo by digital genomic footprinting. *Nature Methods 6*, 4 (Apr 2009), 283–289.

[15] HINDORFF, L. A., SETHUPATHY, P., JUNKINS, H. A., RAMOS, E. M., MEHTA, J. P., COLLINS, F. S., AND MANOLIO, T. A. Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. *Proceedings of the National Academy of Sciences 106*, 23 (2009), 9362–9367.

[16] HULME, M. A., WASSERFALL, C. H., ATKINSON, M. A., AND BRUSKO, T. M. Central role for interleukin-2 in type 1 diabetes. *Diabetes 61*, 1 (2012), 14–22.

[17] I.MEDINA, D.MONTANER, N.BONIFACI, M.A.PUJANA, J.CARBONELL, J.TARRAGA, F.AL-SHAHROUR, AND J.DOPAZO. Gene set-based analysis of polymorphisms: finding pathways or biological processes associated to traits in genome-wide association studies. *Nucleic Acids Research 37* (2009), W340–W344.

[18] KANG, H., K.GAULTON, VOIGHT, B., C.FUCHSBERGER, PEARSON, R., J.MAGUIRE, T.TESLOVICH, Y.CHEN, J.FLANNICK, SCOTT, L., L.MOUTSIANAS, R.POPLIN, P.CHINES, J.PERRY, C.SIDORE, M.RIVAS, T.BLACKWELL, I.PROKOPENKO, T.FENNELL, G.JUN, T.FRAYLING, N.BURTT, ABECASIS, G., P.DONNELLY, L.GROOP, M.BOEHNKE, M.MCCARTHY, AND D.ALTSHULER. Sequencing and Genotyping Thousands of European Genomes and Exomes to Better Understand the Genetic Architecture of Type 2 Diabetes: the GoT2D Study. In *Proceedings of the 12th International Congress of Human Genetics/61st Annual Meeting of The American Society of Human Genetics, Montreal, Canada* (October 2011).

[19] LANGO ALLEN, H., ET AL. Hundreds of variants clustered in genomic loci and biological pathways affect human height. *Nature 467*, 7317 (2010), 832–838.

[20] L.CLARKE, AND J.CARBON. Isolation of the centromere-linked cdc10 gene by complementation in yeast. *Proceedings of the National Academy of Science 77* (April 1980), 2173–2177.

[21] LEE, S., NAOMIR.WRAY, MICHAELE.GODDARD, AND PETERM.VISSCHER. Estimating missing heritability for disease from genome-wide association studies. *The American Journal of Human Genetics 88*, 3 (2011), 294 – 305.

[22] LIU, Y., ARYEE, M. J., PADYUKOV, L., FALLIN, M., HESSELBERG, E., RUNARSSON, A., REINIUS, L., ACEVEDO, N., TAUB, M., RONNINGER, M., SHCHETYNSKY, K., SCHEYNIUS, A., KERE, J., ALFREDSSON, L., KLARESKOG, L., EKSTROM, T. J., AND FEINBERG, A. P. Epigenome-wide association data implicate dna methylation as an intermediary of genetic risk in rheumatoid arthritis. *Nat Biotech advance online publication* (Jan 2013).

[23] MAHER, B. Personal genomes: The case of the missing heritability. *Nature 456*, 7218 (Nov 2008), 18–21.

[24] MANOLIO, T. A., COLLINS, F. S., COX, N. J., GOLDSTEIN, D. B., HINDORFF, L. A., HUNTER, D. J., MCCARTHY, M. I., RAMOS, E. M., CARDON, L. R., CHAKRAVARTI, A., CHO, J. H., GUTTMACHER, A. E., KONG, A., KRUGLYAK, L., MARDIS, E., ROTIMI, C. N., SLATKIN, M., VALLE, D., WHITTEMORE, A. S., BOEHNKE, M., CLARK, A. G., EICHLER, E. E., GIBSON, G., HAINES, J. L., MACKAY, T. F., MCCARROLL, S. A., AND VISSCHER, P. M. Finding the missing heritability of complex diseases. *Nature 461*, 7265 (2009), 747–753.

[25] MAURANO, M. T., HUMBERT, R., RYNES, E., THURMAN, R. E., HAUGEN, E., WANG, H., REYNOLDS, A. P., SANDSTROM, R., QU, H., BRODY, J., SHAFER, A., NERI, F., LEE, K., KUTYAVIN, T., STEHLING-SUN, S., JOHNSON, A. K., CANFIELD, T. K., GISTE, E., DIEGEL, M., BATES, D., HANSEN, R., NEPH, S., SABO, P. J., HEIMFELD, S., RAUBITSCHEK, A., ZIEGLER, S., COTSAPAS, C., SOTOODEHNIA, N., GLASS, I., SUNYAEV, S. R., KAUL, R., AND STAMATOYANNOPOULOS, J. A. Systematic localization of common disease-associated variation in regulatory dna. *Science 337*, 6099 (2012), 1190–1195.

[26] MCLEAN, C. Y., BRISTOR, D., HILLER, M., CLARKE, S. L., SCHAAR, B. T., LOWE, C. B., WENGER, A. M., AND BEJERANO, G. Great improves functional interpretation of cis-regulatory regions. *Nat Biotech 28*, 5 (May 2010), 495–501.

[27] M.HOLDEN, S.DENG, L.WOJNOWSKI, AND B.KULLE. Gsea-snp: applying gene set enrichment analysis to snp data from genome-wide association studies. *Bioinformatics 24* (2008), 2784–2785.

[28] MOORE, F., COLLI, M., CNOP, M., ESTEVE, M., CARDOZO, A., CUNHA, D., BUGLIANI, M., MARCHETTI, P., AND EIZIRIK, D. Ptpn2, a candidate gene for type 1 diabetes, modulates interferon-gamma-induced pancreatic beta-cell apoptosis. *Diabetes 58* (Jun 2009), 1283–91.

[29] M.WGNER, A., CLOOS, P., BERGHOLDT, R., EISING, S., BRORSSON, C., STALHUT, M., CHRISTGAU, S., NERUP, J., AND POCIOT, F. Posttranslational protein modifications in type 1 diabetes - genetic studies with pcmt1, the repair enzyme protein isoaspartate methyltransferase (pimt) encoding gene. *The Review of Diabetics Studies 5* (2008), 225–231.

[30] P.MCEVOY, B., AND M.VISSCHER, P. Genetics of human height. *Economics & Human Biology 7*, 3 (2009), 294 – 306.

[31] RICHTER, W., SHI, Y., AND BAEKKESKOV, S. Autoreactive epitopes defined by diabetes-associated human monoclonal antibodies are localized in the middle and c-terminal domains of the smaller form of glutamate decarboxylase. *Proceedings of the National Academy of Sciences 90*, 7 (1993), 2832–2836.

[32] SANGER, F., NICKLEN, S., AND COULSON, A. Dna sequencing with chain-terminating inhibitors. *Proceedings of the National Academy of Sciences 74*, 12 (1977), 5463–5467.

[33] S.BURREN, O., C.ADLEM, E., ACHUTHAN, P., CHRISTENSEN, M., COULSON, R. M., AND A.TODD, J. T1DBase: update 2011, organization and presentation of large-scale data sets for type 1 diabetes research. *Nucleic Acids Research 39* (Jan 2011).

[34] SHESKIN, D. *Handbook of Parametric and Nonparametric Statistical Procedures*, 5th ed. ed. Chapman & Hall, Boca Raton, Florida, 2011.

[35] STAHL, E. A., WEGMANN, D., TRYNKA, G., GUTIERREZ-ACHURY, J., DO, R., VOIGHT, B. F., KRAFT, P., CHEN, R., KALLBERG, H. J., KURREEMAN, F. A., KATHIRESAN, S., WIJMENGA, C., GREGERSEN, P. K., ALFREDSSON, L., SIMINOVITCH, K. A., WORTHINGTON, J., DE BAKKER, P. I., RAYCHAUDHURI, S., AND PLENGE, R. M. Bayesian inference analyses of the polygenic architecture of rheumatoid arthritis. *Nat Genet 44*, 5 (May 2012), 483–489.

[36] SUBRAMANIAN, A., TAMAYO, P., MOOTHA, V. K., MUKHERJEE, S., EBERT, B. L., GILLETTE, M. A., PAULOVICH, A., POMEROY, S. L., GOLUB, T. R., LANDER, E. S., AND MESIROV, J. P. Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. *Proceedings of the National Academy of Sciences of the United States of America 102*, 43 (2005), 15545–15550.

[37] THE 1000 GENOMES PROJECT CONSORTIUM. An integrated map of genetic variation from 1,092 human genomes. *Nature 491*, 7422 (Nov 2012), 56–65.

[38] THE ENCODE PROJECT CONSORTIUM. The ENCODE (ENCyclopedia Of DNA Elements) project. *Science 306*, 5696 (2004), 636–640.

[39] THE INTERNATIONAL HAPMAP CONSORTIUM. The International HapMap Project. *Nature 426*, 6968 (Dec 2003), 789–796.

[40] THE INTERNATIONAL HAPMAP CONSORTIUM. A second generation human haplotype map of over 3.1 million snps. *Nature 449*, 7164 (Oct 2007), 851–861.

[41] THE INTERNATIONAL HAPMAP CONSORTIUM. Integrating common and rare genetic variation in diverse human populations. *Nature 467*, 7311 (Sep 2010), 52–58.

[42] 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*, 7145 (Jun 2007), 661–678.

[43] TRYNKA, G., SANDOR, C., HAN, B., XU, H., STRANGER, B. E., LIU, X. S., AND RAYCHAUDHURI, S. Chromatin marks identify critical cell types for fine mapping complex trait variants. *Nat Genet 45*, 2 (Feb 2013), 124–130.

[44] WANG, D. G., FAN, J.-B., SIAO, C.-J., BERNO, A., YOUNG, P., SAPOLSKY, R., GHANDOUR, G., PERKINS, N., WINCHESTER, E., SPENCER, J., KRUGLYAK, L., STEIN, L., HSIE, L., TOPALOGLOU, T., HUBBELL, E., ROBINSON, E., MITTMANN, M., MORRIS, M. S., SHEN, N., KILBURN, D., RIOUX, J., NUSBAUM, C., ROZEN, S., HUDSON, T. J., LIPSHUTZ, R., CHEE, M., AND LANDER, E. S. Large-scale identification, mapping, and genotyping of single-nucleotide polymorphisms in the human genome. *Science 280*, 5366 (1998), 1077–1082.

[45] WANG, L., JIA, P., D.WOLFINGER, R., CHEN, X., AND ZHAO, Z. Gene set analysis of genome-wide association studies: Methodological issues and perspectives. *Genomics 98*, 1 (2011), 1 – 8.

[46] WATSON, J., AND CRICK, F. Molecular structure of nucleic acids: A structure for deoxyribose nucleic acid. *Nature 171*, 4356 (Apr 1953), 737–738.