Principled "Convergence" Non-coding Rare Variant Association Testing in Complex Disease

by Daniel N. Sosa

S.B., Computer Science/Molecular Biology and Management, M.I.T., 2017

Submitted to the Department of Electrical Engineering and Computer Science in Partial Fulfillment of the Requirements for the Degree of

Master of Engineering in Electrical Engineering and Computer Science

at the

Massachusetts Institute of Technology

May 2017 [June 2017]

Author: Signature redacted
Department of Electrical Engineering and Computer Science
May 22, 2017

Certified by: Signature redacted
Manolis Kellis, Professor, Computer Science, MIT, Thesis Supervisor
May 22, 2017

Accepted by: Signature redacted
Christopher Terman, Chairman, Masters of Engineering Thesis Committee

Principled "Convergence" Non-coding Rare Variant Association Testing in Complex Disease

by Daniel N. Sosa

Submitted to the Department of Electrical Engineering and Computer Science on May 22, 2017 in Partial Fulfillment of the Requirements for the Degree of Master of Engineering in Electrical Engineering and Computer Science

ABSTRACT

Although many genetic loci pertinent to complex diseases have been identified and despite the fact that complex diseases remain an immense burden to healthcare globally, many details about the mechanism of these diseases are still unknown. Thus far, genome-wide association studies (GWAS) have only explained a small proportion of disease heritability, indicating that there is a large number of additional loci that contribute to complex diseases like type 2 diabetes (T2D), which is the primary case study in this work. We overcome some of the limitations of rare variant studies by conducting weighted aggregate association tests in a framework we call "Convergence". We compare potential cell type specific regulatory loci assigned to genes, which serve as the basis for grouping variants and integrated predictors of functional consequence of variants, which serve as variant weights. We demonstrate that this methodology is able to detect significant association to T2D for genes relevant for body weight homeostasis, adipocyte proliferation, and inflammation. As a result, this work provides a principled framework for improving the efficacy of RVAS by successfully converging the abundant epigenetic information available to understand complex disease.

Thesis Supervisor: Manolis Kellis

Title: Professor of Computer Science

Table of Contents

1. Introduction

Today, complex diseases like diabetes, hypertension, and Alzheimer's, are huge, unmet problems in society. One especially troublesome case study, which we considered for this investigation, is type 2 diabetes (T2D), a disease in which the body becomes resistant to insulin, and insulin secretion from pancreatic beta cells is unable to compensate for insulin resistance, resulting in elevated blood glucose levels. In 2012, 29.1 million Americans (9.3% of the population) were diagnosed with T2D. The CDC predicts that upwards of 30% of the American population will be afflicted by 2050. T2D is the seventh leading cause of death in America, leading to complications such as hypertension, kidney disease, and stroke and consists of 80-90% of all reported diabetes cases [1]. Despite the continued impact that T2D has on society, one of the major contemporary public health challenges is to dissect the genetic basis of the disease in order to identify therapeutic targets and enable more effective development of treatments that target the disease's causes rather than its effects.

To test the association of common genomic variants with T2D several international consortia have conducted GWAS of around 130,000 individuals [2,3]. As of 2014, more than 120 variants have been repeatedly demonstrated as being directly linked to T2D via GWAS, but the 98 independent loci derived from these variants only explain about 20% of the heritability of the disease [1], motivating the search for the "missing heritability" of T2D. One leading hypothesis to explain this discrepancy is that T2D is caused by genetic variants that are infrequent in a given population, or rare variants, which cannot be detected by GWAS [4]. Additionally, the push to investigate rare variants is evolutionarily motivated as rarer variants are hypothesized to be more deleterious than common variants on average due to the lessened opportunity for purifying selection.

Because rare variants occur infrequently in the population but are collectively abundant when looking across numerous individuals, the main challenge in analyzing them is insufficient statistical power in single variant association studies [5]. To address this issue, previous work aggregated variants over genes under the assumption that the effect of rare variation is primarily disruption of coding sequences.

Most variants, however, are not located in protein coding regions and likely are regulatory in nature. In addition, computational approaches developed to study somatic mutations in cancer have shown the importance of disruption of regulatory mechanisms in disease [6]. Thus, alternative aggregation criteria considering the relevance of non-coding loci are imperative given the emerging picture of the genetic underpinnings of complex disease. Studying rare variation in non-coding loci with high regulatory potential is therefore a natural next step in the field of GWAS.

Here, we leveraged sequencing data from new large-scale initiatives to explore the Convergence framework in a specific disease of interest. Recently, the GoT2D consortium completed a large scale whole genome sequencing effort to interrogate non-coding variants in 2,850 case and control European subjects in an attempt to shed light on the problem of missing heritability. T2D is an appropriate case study in the Convergence framework because of the abundance of functionally uncharacterized rare variants and the high degree of genetic heterogeneity in the disease.

Unlike coding variants in which the function may be more immediately ascribed to a direct effect on protein structure, interpreting the impact of non-coding variants in gene expression has been historically more challenging. Large-scale epigenomic mapping efforts by the Encyclopedia of DNA Elements (ENCODE) and Roadmap Epigenomics consortia have built epigenomic annotations of regulatory activity across more than a hundred human cell types and

tissues [7,8]. Chromatin state annotations were derived by integrating information motivated by metrics such as accessibility or specific chemical modulation in the DNA. Previous results have already shown enrichment of variants in regions of active regulatory function specific to certain pathologically relevant cell types like stomach mucosa in the case of T2D [8]. Given the active potential regulatory elements in a specific tissue, efforts have been made to relate these regions to specific target genes. The resulting networks can provide the basis to aggregate variants at the gene level.

In addition to regulatory associations between noncoding elements and genes, there are multiple studies that have attempted to measure the functional effect based on upstream regulators and binding sites. One algorithm, known as Intra-Genomic Replicates (IGR), predicts the effect of variation on transcription factor binding affinity by building a model from genome-wide data [9]. Another, known as phastCons, trains an HMM to ultimately calculate a posterior likelihood that a point locus is evolutionarily conserved. We can use epigenomic annotations to learn which regulators are involved in T2D by studying disease enrichment in regulatory variant sets throughout the genome [9].

To meaningfully ground the methodology in statistics, we use the different components to conduct weighted aggregated association tests, whereby the weights correspond to functional predictors of variants and the groups over which to aggregate are the groups of variants in the multiple potential cis-regulatory loci, or "plexus", for each gene. The goal of these tests is to perform meaningful statistical association analyses using groups of variants and leveraging predictions of functionality. This is motivated by the assumption that rare variants with relatively strong association signal are found more often than by chance alone in active regulatory regions of the genome that are associated with disease phenotype, which has previously been shown in many diseases for exome sequencing data. In this application, the principal hypothesis is that

the effect in the function of pathologically relevant genes is likely to occur through perturbations of their upstream regulatory elements, thereby suggesting natural aggregating criteria to test for association. The study aims to validate existing discoveries and find novel genes which contribute to disease under this recontextualization of variant effect and demonstrates the efficacy of this framework to indeed discover biologically meaningful mechanisms underlying diabetes pathogenicity.

2. Methods

2.1. Whole Genome Data

Genomic variants were obtained from GoT2D whole genome data [10]. Genomes were sequenced from 1,326 T2D cases and 1,331 controls. Subjects were from northern and central Europe. Genome sequencing was performed at ~5x coverage. Only single nucleotide variants (SNVs) were included in the study.

2.2. Predicted Functionality Metrics for Non-coding Variants

Two metrics were considered to measure the predicted functional potential, or "weighting", of non-coding variants. To ascertain functionality based on disruption of evolutionarily conserved motifs in the non-coding genome, phastCons was used [11]. PhastCons weights are measures of the degree of evolutionary conservation for each variant range, which range from 0 to 1.

To ascertain the effect of variation on transcription factor binding affinity, the Intragenomic Replicates (IGR) algorithm was used [6]. A functional score was calculated based

on the maximum ratio between average CTCF CHIP-seq binding peaks for reference and alternate alleles for the 8 sequences within an 8-nucleotide sliding window around a given variant. Weights are positive real numbers, whereby larger weights correspond to larger predicted effects in CTCF binding affinity. Variants not passing quality control filters of were assigned weights of 0.

2.3. Enhancer-Gene Assignments

Enhancer loci were assigned to target genes by three methods. First, assignments were derived based on proximal epigenomic co-activity as described in Ernst et al., 2011 [12], which we refer to as "Co-Activity" assignments. Only enhancers annotated as states 6-Genic Enhancers, 7-Enhancers, or 12-Bivalent Enhancers from the Roadmap 15-state model were used. The pathologically relevant and available Roadmap reference epigenomes considered were pancreatic islet (E087) and liver (E066). Only assignments with confidence scores at or above the 90th percentile of confidence scores were considered.

Second, assignments predicted in Wang et al. were used as a means to capture distal enhancer regulation, non-linear activity associations, and more robust regulation by multiple enhancers across cell types. Briefly, assignments were predicted using probabilistic inference by first grouping enhancers into modules based on coordinated activity and linking genes to modules using a Latent Dirichlet Allocation model [13]. We refer to these links as "Module LDA" assignments" The pathologically relevant and available Roadmap cell types considered were pancreatic islet (E087) and liver (E066).

Finally, a third type of enhancer-gene assignment was considered. In order to further reinforce chromatin accessibility information into the module-based linkage of Wang et al., module-based links were aggregated together across all cell types and retained if overlapping

with DNase peaks from Roadmap. For each of these remaining links, three different types of enhancer loci were considered for assignment to genes. "DNase" links used the locus of the DNase peak, "Element" links used the locus of the original module element, and "Extended" links used the locus of the original module element with an additional 20kb flanked on each side. To induce cell-type specificity, three different sets of cell-type specific Roadmap enhancer-like annotations from the 25-state model were used to subset the enhancer locus for each link. Roadmap enhancer state groups considered were referred to as "Loose" (states 1-active TSS, 2-Promoter Upstream TSS, 3-Promoter Downstream TSS1, 4-Promoter Downstream TSS2,9-Transcribed and Regulatory (Prom/Enh), 10-Transcribed 5' preferential and Enh, 11-Transcribed 3' preferential and Enh, 12-Transcribed and Weak Enhancer, 13-Active Enhancer 1, 14-Active Enhancer 2, 15-Active Enhancer Flank, 16-Weak Enhancer 1, 17-Weak Enhancer 2, 18-Primary H3K27ac possible Enhancer, 19-Primary), "Medium" (states 2,9,10,11,12,13,14,15), and "Strict" (states 1,2,3,4,13,14,15,18,19) enhancers.

2.4. Grouped Variant Association Test

For each gene, the group of variants included in the test was defined based on the regulatory plexus of the gene. Genetic association to the binary T2D phenotype was evaluated using a logistic regression over the *m* weighted variants within the group and *k* covariates per subject, as follows:

$$Pr(Y_i = 1) = \frac{e^{\tau\xi^T X_i + \gamma^T Z_i}}{1 + e^{\tau\xi^T X_i + \gamma^T Z_i}}$$

where *i* is the subject, X_i is a binary genotype vector of dimension *m* x 1, τ is a constant, ξ is the variant weight vector of dimension *m* x 1, Z_i is a covariate vector of dimension *k* x 1, γ is the

vector of regression coefficients over covariates with dimension k x 1, and Y_i is binary phenotype. The tested null hypothesis was (H_0: $\tau = 0$) genotype information of variants within the gene plexi, in combination with functional predictors, provide predictive information additional to that provided by covariates alone. Score statistics were calculated as follows:

$$U = \sum_{i=1}^{n} \left(Y_i - \frac{e^{\hat{\gamma}^T Z_i}}{1 + e^{\hat{\gamma}^T Z_i}} \right) \xi^T X_i$$

$$V = \sum_{i=1}^{n} v_i S_i^2 - \left(\sum_{i=1}^{n} \nu_i S_i Z_i \right)^T \left(\sum_{i=1}^{n} \nu_i Z_i Z_i^T \right)^{-1} \left(\sum_{i=1}^{n} \nu_i S_i Z_i \right)$$

where, $\nu_i = \left(\frac{e^{\hat{\gamma}^T Z_i}}{1+e^{\hat{\gamma}^T Z_i}} \right)^2$ and $S_i = \xi^T X_i$, thus yielding the test statistic

$$T = \frac{U}{\sqrt{V}},$$

which is assumed to be asymptotically normally distributed. P-values were calculated with a two-tailed test and adjusted by Bonferroni correction with the number of tests corresponding to the number of genes with at least one variant in a linked non-coding locus. Association testing was implemented in the SCORE-Seq program [14].

3. Results

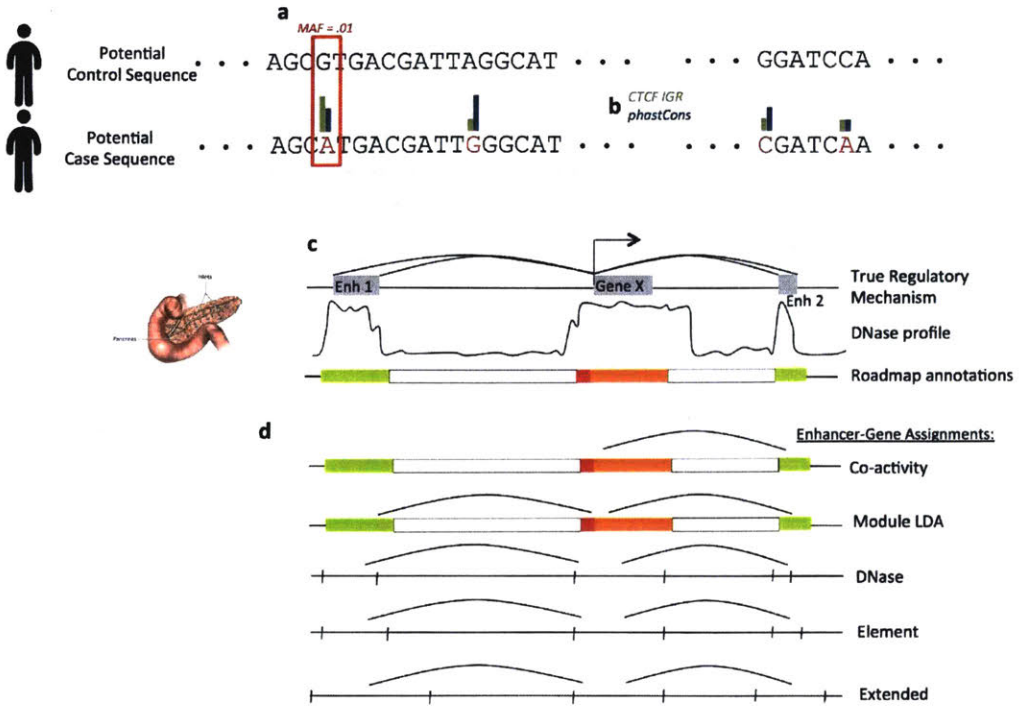


Figure 1 The Convergence framework integrates disparate information about the regulatory potential of non-coding rare variants to detect genetic associations to complex disease. **a)** Variants are filtered by MAF, here only SNVs with MAF < 0.05 are considered for inclusion in the study. **b)** Each SNV is weighted based on the predicted effect of CTCF binding (IGR) or impact based on evolutionary conservation (phastCons). **c)** SNVs are aggregated at the gene level based on their occurrence in enhancers predicted to have a direct regulatory effect on the gene of interest. **d)** Different predictors of regulatory potential were evaluated as approximations to true enhancer activity.

In order to approach the aforementioned problems, we present an aggregate weighted test motivated by the idea of regulatory convergence of non-coding mutations. The Convergence framework for detecting genetic association to complex disease leverages and integrates new data to understand how variation in gene regulation might explain genetic association in a way that common variants or coding variants alone cannot. In the current form,

3.2 Ascertaining Functional Consequences

Subsequently, we computed measures of the potential functional nature of each variant to integrate in the testing. We used two orthogonal proxies for assessing the effects of rare variants in non-coding loci: evolutionary conservation and potential disruption of CTCF binding sites. Disruption of evolutionarily conserved regions has been shown to have greater functional consequences for normal phenotype because conserved regions may have functions that have undergone purifying selection [15]. CTCF was used because it has been demonstrated to affect enhancer regulatory function under perturbations due to chromatin topological properties, such as organization in large domains, and disruption of this activity has been previously shown to result in pathogenic effects like the overexpression of oncogenes [16]. To this end, we calculated PhastCons scores, which measures the degree of conservation across several vertebrate species, and we used the IGR method from CTCF ChIP-Seq data, which measures the relative signal of ChIP-Seq peaks with and without incorporated variants. Conservation-based scores for the studied variants are highly bimodal between the predicted conserved and nonconserved states with there being many fewer variants predicted to be highly conserved (phastCons ≈ 1) (**Figure 2d**). By contrast, of the variants that pass the quality filters for relative signal for ChIP-Seq from reference and variant-incorporated motifs, the distribution of weights is right-skewed and unimodal (**Figure 2e**).

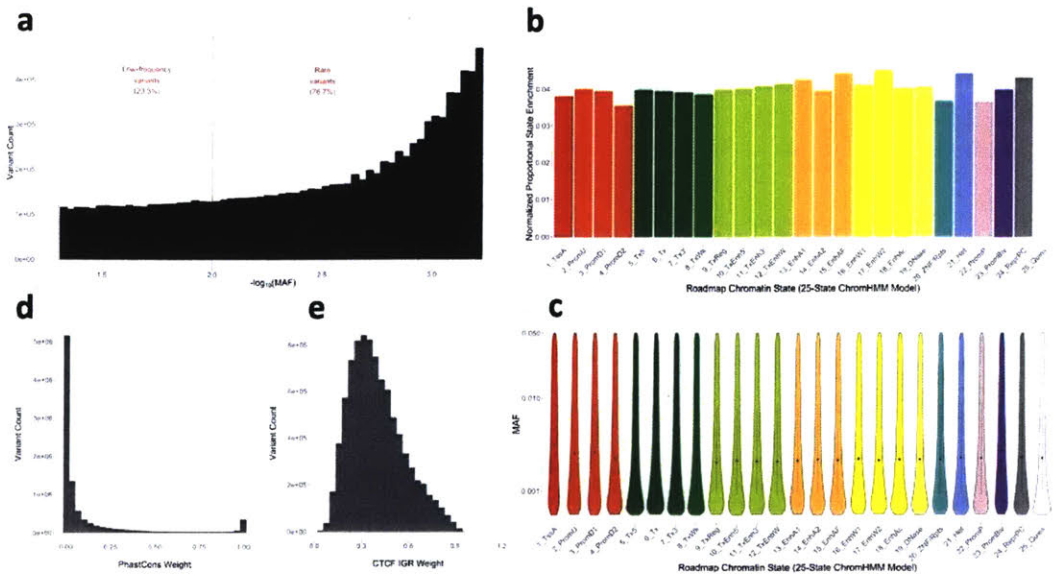


Figure 2 a) Distribution of low-frequency (LF) and rare variants across subjects. Approximately a quarter of variants with MAF < 0.05 fall in the LF range and the rest in the rare range. **b)** Distribution of variants in various Roadmap epigenomic states from the 25-state model, normalized for genome coverage of each state. LF and rare variants are not disproportionately represented in any specific annotation state. **c)** Distribution of MAFs for variants in different epigenomic states. No discernible biases for lower MAFs are present in any particular state. **d)** The distribution for phastCons conservation scores is highly bimodal. **e)** CTCF IGR scores that pass quality filtration are right-skewed and unimodal. Only non-zero values of CTCF IGR are shown above.

3.3 Developing Diverse Gene Plexi

Because true enhancer-gene interactions are not fully characterized and are dynamic across epigenomic conditions, we tested different enhancer-gene assignment sets reflecting different evidence for regulatory interaction, which would serve as the basis for grouping variants. First, we considered the set of cell-type specific enhancer-gene assignments using the methodology from Ernst et al. for pancreatic islets (E087) and liver (E066) [8]. These assignments, which we dub "Co-activity" assignments, are based on co-activity patterns of cis-regulatory elements and target genes (for details, see [12]). Second, we considered Wang et

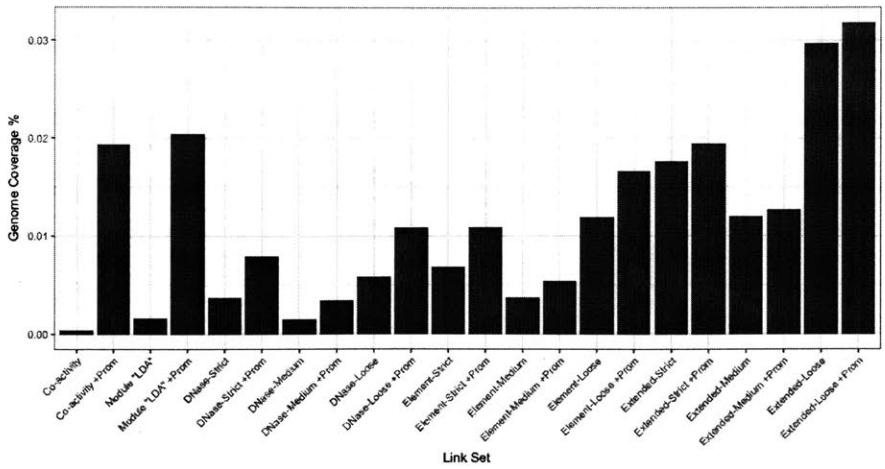


Figure 3 Histogram of genome coverage for different enhancer-gene link sets. Co-activity and Module LDA regulatory loci have low genome coverage (~0.001%) as compared to the loci from DNase (~0.05%), Element (~0.075%), and Extended (~0.02%) links. The DNase-based links have lowest coverage over the "medium" enhancer-like states, then "strict", and "loose" have the greatest coverage. Including promoter regions notably increases genome coverage as well. All links shown are pancreatic islet-specific.

al.'s orthogonal approach to enhancer-gene assignment based on enhancer clusters as defined by H3K4me1 signal being assigned to genes in a topic modeling approach using Latent Dirichlet Allocation (LDA). We refer to these as "Module LDA" assignments and used those derived from pancreatic islets (E087) and liver (E066) [13]. Finally, we considered a third link set derived from the union of the latter across all cell types intersected with DNase-seq peaks, indicative of open chromatin. and further subset by regions defined as enhancer-like from the 25-state ChromHMM Roadmap annotations. In this last setting we generated regulatory loci based on the DNase-seq peak loci, the original regulatory loci from the Module LDA assignments, or the loci from the Module assignments flanked by 20kb on each side. We refer to these assignments as "DNase", "Element", and "Extended", respectively. To induce cell type specificity to these global links, we subset each of these sets of assignments based on three different groups of Roadmap states

that are "enhancer-like", which we dub "loose", "medium", and "strict". For all DNase-based assignments, we examined pancreatic islets (E087), and adipose (E023) cell types.

Different regulatory criteria produce sets of regulatory loci with highly heterogeneous genome coverage (**Figure 3**). The loci from the Co-activity assignments provide lowest genome coverage at ~0.001%. Because the DNase-based links cover a larger subset of the space of possible assignments based on the Module LDA links across cell types, these links have the greatest coverage as expected. In particular, the Extended links have the greatest amount of genome coverage from their predicted regulatory loci at 0.02-0.03%, with the greatest coming from "loose" enhancer states defined.

Next, we ascertained the number of variants per plexus, size of plexus, and density of variants within plexi as a gauge of the scope and potential for capturing variation of each set of regulatory loci used in enhancer-gene assignments (**Figure 4**). As expected, increased genome coverage corresponds an increased number of variants in gene plexi and an increased coverage of regulatory regions per individual gene. At the lowest end of coverage and variants are the Co-activity assignments, with 10 variants or fewer per plexus and plexi in the 1kb total size range being typical. By contrast, the Extended links contained typically in the hundreds of variants with plexi typically on the order of 10kb. The DNase and links are most densely populated with variants in their plexi, reinforcing the claim that interesting variation may be found in areas of open chromatin accessibility. Co-activity links were least dense, perhaps indicating that fewer regulatory loci are recapitulated by these loci and thus fewer variants are present.

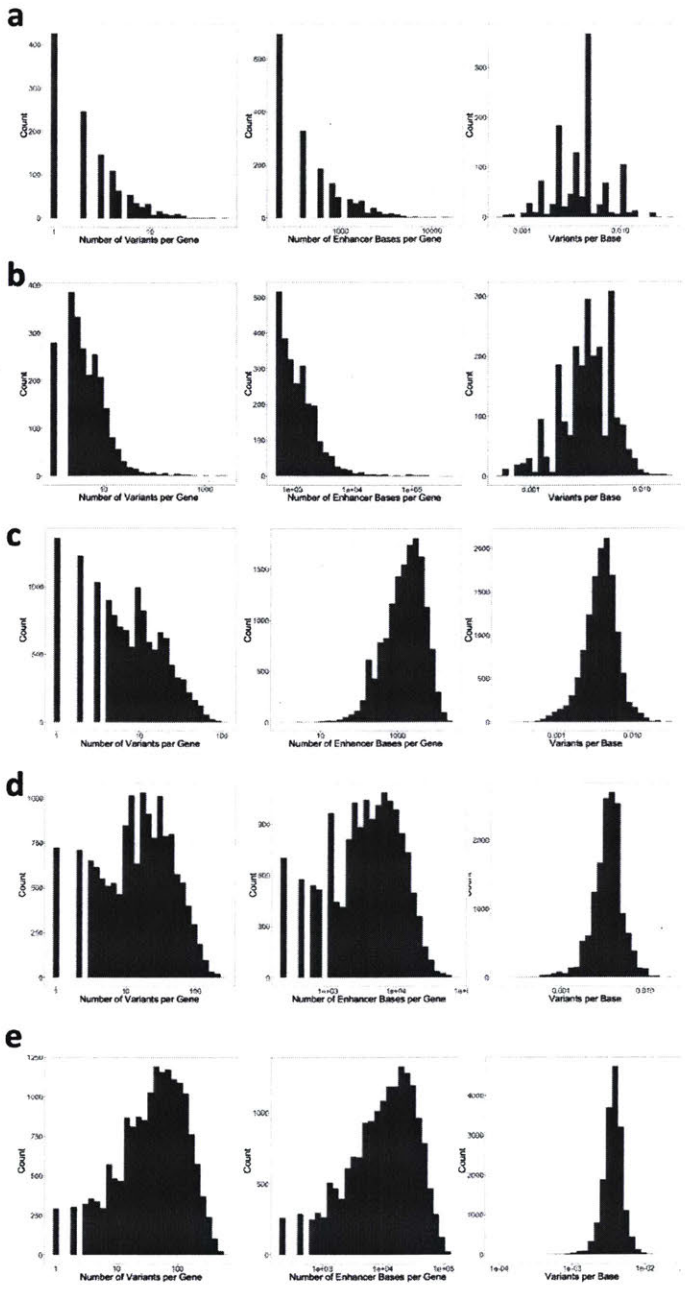


Figure 4 Histograms of number of variants per gene plexus (left), size of plexus in bp (center) and density of variants per plexus (right) in **a)** Co-activity, **b)** Module LDA, **c)** DNase, **d)** Element, and **e)** Extended enhancer-gene assignments. Co-activity links contain the fewest variants and the smallest plexi whereas Extended links have largest plexi and most variants. DNase links are most densely populated with variants per base, whereas Co-activity links are least dense in variants. All histograms are derived from pancreatic islet-based linkage excluding promoters and for the "medium" definition of enhancers for DNase-based links.

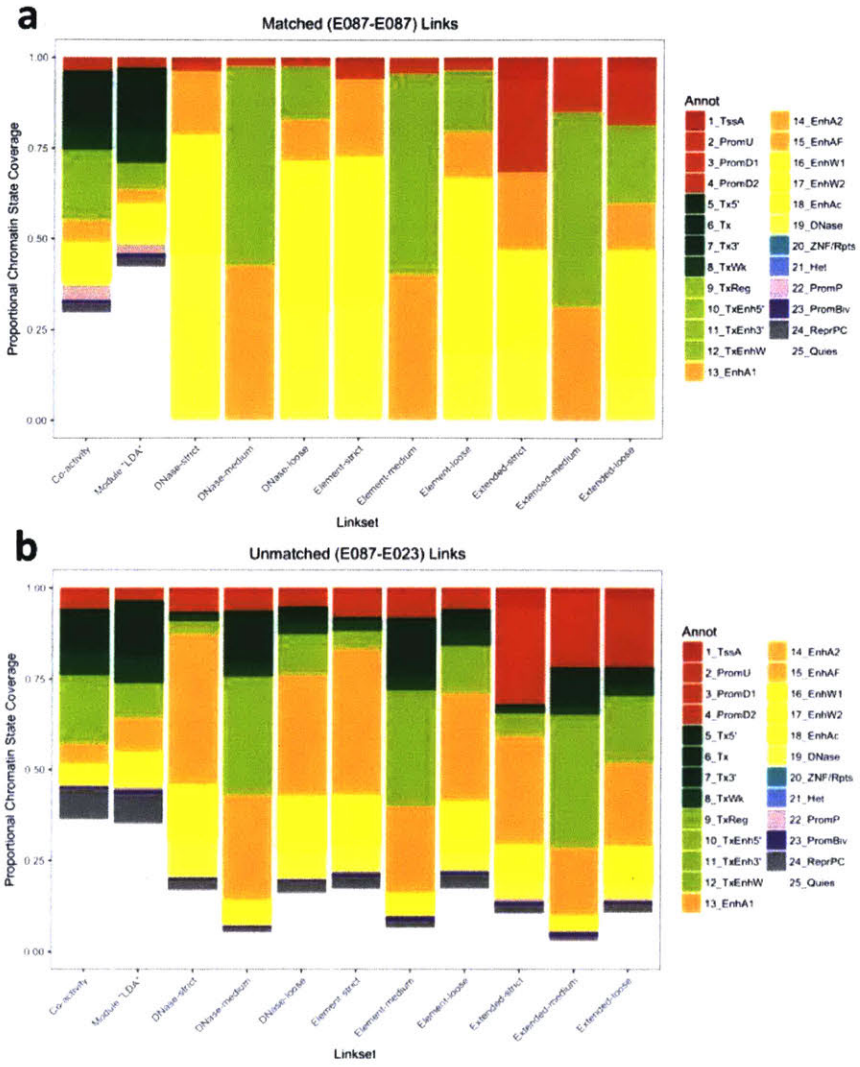


Figure 5 In order to scrutinize the regulatory potential of the regulatory loci from different link sets, we assessed the correspondence of regulatory loci with cell type specific Roadmap state annotations. **a**) Comparing pancreatic islet regulatory loci with pancreatic islets Roadmap states shows a great degree of representation from enhancer-like states as expected, noting that the Co-activity and Module LDA links have a baseline amount of quiescent regions represented. **b**) Comparing islet predicted loci with the unmatched adipose Roadmap states shows a great increase in quiescent and repressed polycomb regions for all links, indicating that the predicted loci are indeed recapitulating cell type-specific regulatory potential in regulatory loci.

To evaluate the ability of different links to recapitulate predicted cell type-specific enhancer states via Roadmap, we compared the overlap between enhancer loci and 25-state Roadmap annotations for the matched and an unmatched cell type (**Figure 5**). For the matched islet links with islet states, co-activity and module-based assignments capture similar distributions over epigenomic states with enhancer-like states being overrepresented as expected. When comparing islet links with liver states, an increased abundance of repressed polycomb and quiescent regions is present, which is consistent with specificity, since non-active reaches are picked up in the non-target cell types.

In summary, we found that there is a heterogeneity regarding the representation of variants in gene plexi across different plexus definitions. DNase-based plexi are able to capture more variants, affirming the motivation for chromatin accessibility as a basis of regulatory significance. Additionally, we have demonstrated that cell type specific links capture more active regulatory regions within plexi than for non-target cell types. All of this shows that it is potentially useful to define links in different ways because mutations are captured heterogeneously. Given the diversity of plexus information, we tested all link sets to ascertain if different genetic associations could be detected from differing variant groupings.

3.4 Recapitulating Pathologically Relevant Mechanisms

For each gene, we conducted weighted aggregate tests over variants in the gene plexus with predicted functional weighting to produce a *p* value. To check that the distribution of *p* values for each condition were not inflated because of latent population substructures or other covariates not being considered, we generated QQ plots for each condition to ascertain if the observed distribution of *p* values roughly matches the expected (normal) distribution of *p* values below the top percentiles. Based on these plots and examining the inflation factor, it seems that

the covariates are sufficiently accounted for the Convergence p value and therefore the genotypic information itself is the true underlying factor driving association with T2D under the model considered. Across all conditions, the tests are well calibrated for the convergence statistic as evidenced from each parameter set's representative QQ plot. For many conditions, we observed interesting results to be analyzed more in depth in future studies, several of which vary for different conditions indicating that different plexus definitions may capture different regulatory action. In what follows, we selected a subset of salient findings for further analysis due to their relevance to diabetes-related function, namely the following genes: TM4SF1, CEBPA, RP11-475A13.1, RP1-68D18.4, and MAPKAP1. A full table of top gene results per condition is available in **Supplementary Tables 1** and **2**.

In pancreatic islets, TM4SF1 was implicated as genome-wide significant in Co-activity links with CTCF-IGR weights ($p = 1.69 \times 10^{-7}$). TM4SF1 overexpression has previously been demonstrated as being associated with adipose hypertrophy. This association is noteworthy because enlarged adipocytes have been demonstrated to be linked with insulin resistance, a feature of T2D [17]. A potential mechanism for action is that dysregulation of TM4SF1 may lead to overexpression causing hypertrophy by increasing adipocyte proliferation, although further experimental follow-up would be required for validation of the effect of these variants on expression.

CEBPA evaluated as statistically significant in liver Module LDA links with CTCF IGR weighting ($p = 2.06 \times 10^{-5}$). CEBPA is a transcription factor known to modulate the expression of genes regulating cell cycle, body weight homeostasis, and lipid storage. In the liver, CEBPA also helps to regulate gluconeogenesis and lipogenesis. Previously, a study demonstrated that in diabetes, where immune cell recruitment from bone marrow does not occur and instead dysfunctional myeloid cells are excessively recruited to injury sites, CEBPA transcription was

diminished. Ultimately, this led to failure of both monocyte and granulocytes to mature, an activity which was restored with normal CEBPA expression [18]. This important role in preventing chronic inflammation may have been recapitulated from our study, whereby dysregulation may lead to diabetes symptoms. In the case that these variants downregulate the expression of CEBPA, the rate of body weight homeostasis may decrease, although again further validation would be required. This further demonstrates that rare non-coding variation may have "tuning" effects on gene activity through expression in contrast to coding mutations that are more likely to affect protein function directly.

RP11-475A13.1 tested as near-significant using the Co-activity and Module LDA links ($p = 3.63 \times 10^{-6}$) when integrating promoters and when weighted by CTCF IGR weights (**Figure 6d, 7d**). This lincRNA lies approximately 500kb downstream of the gene MEIS2, a gene critical for [19] pancreatic development, which is indicative of a potential interaction effect. Because of RP11-475A13.1's proximity to MEIS2, it is plausible that this lincRNA has an important regulatory interaction with MEIS2, whereby disrupting that lincRNA function leads to disruption of normal MEIS2 activity, which disrupts normal development of pancreatic cells. Further, MEIS genes have been shown to regulate beta-cell survival [20], another function whose disruption may lead to the pathogenic activity. This example highlights the potential of this method to identify perturbations potentially affecting regulation of non-coding genes, which is difficult to analyze through other means.

Similarly, RP1-68D18.4 is a lincRNA downstream of SLC1A2, a gene known to be involved in fasting [21]. It is also contained within CD44, a gene associated inflammation in diabetes [22]. RP1-68D18.4 and two protocadherins, PCDHGA7 and PCDHGA8, reached significance in Extended links intersected with "strict" enhancer states and CTCF IGR weights ($p = 2.19 \times 10^{-10}$, 3.72×10^{-10}, and 4.68×10^{-8}, respectively) (**Figure 6e, 7e**). Protocadherins

malfunction may have association with childhood obesity [23]. Given all these functions, the disruption of the RP1-68D18.4, which may act directly or serve as a proxy, could affect propensity to obesity and other relevant pathogenic side effects that could increase the T2D disease liability.

MAPKAP1 was evaluated as significant ($p = 1.58 \times 10^{-6}$) in DNase links with strictly defined enhancer-like states and CTCF IGR weights. The gene is particularly notable for its role in recruiting mTORC2, a master regulator that interacts with other kinases that regulate cell growth and survival. Additionally, mTORC2 is known for regulation of glucose homeostasis in adipocytes. Recent studies have shown how dysregulation of these pathways is integral to the onset and progression of diabetes among other diseases [24]. Thus, inability of MAPKAP1 to properly recruit mTORC2 and initiate these signal pathways may result in pathogenic effects observed in T2D.

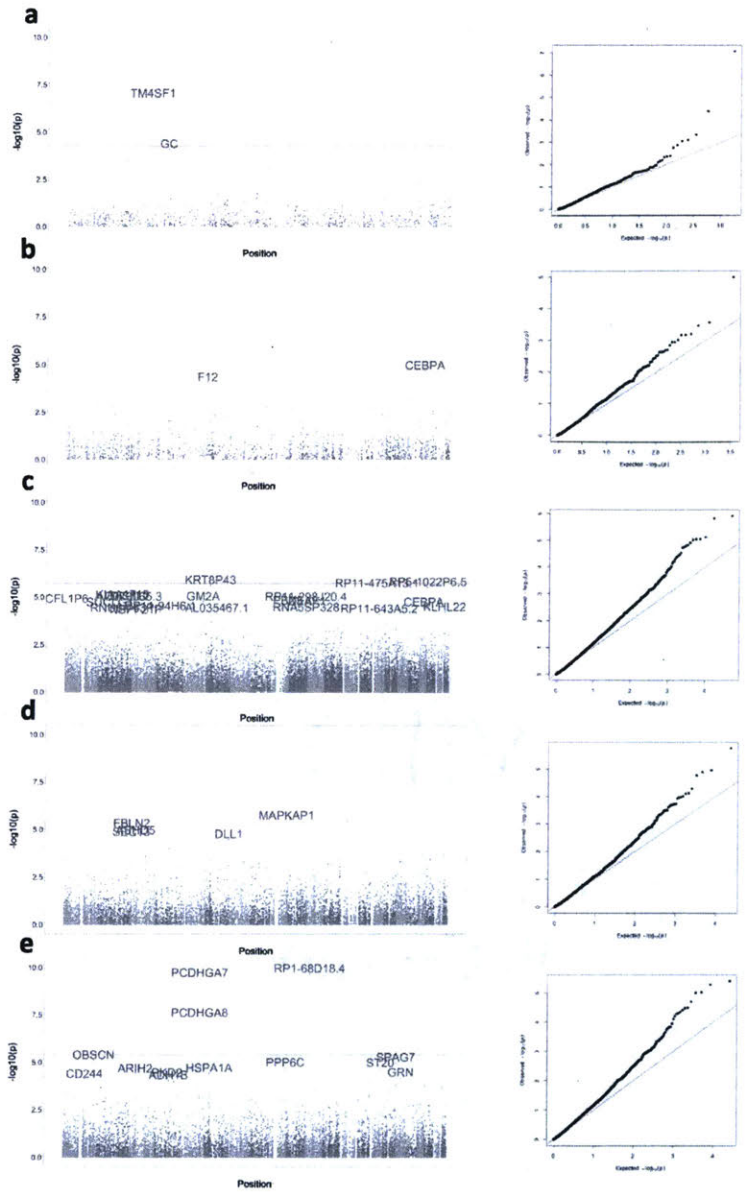


Figure 6 Manhattan plots (left) and QQ plots (right) from **a)** Co-activity pancreatic islet links with **b)** Module LDA liver links, **c)** Module LDA +Promoter liver links, **d)** Extended strict adipose links, and **e)** DNase strict links adipose links, all with CTCF IGR weighting. In all cases, distributions of *p*-values are well calibrated with modest inflation at most. From these conditions, TM4SF1, CEBPA, RP11-475A13.1, RP1-68D18.4, and MAPKAP1, respectively, reach genome-wide significance or near significance and were analyzed further.

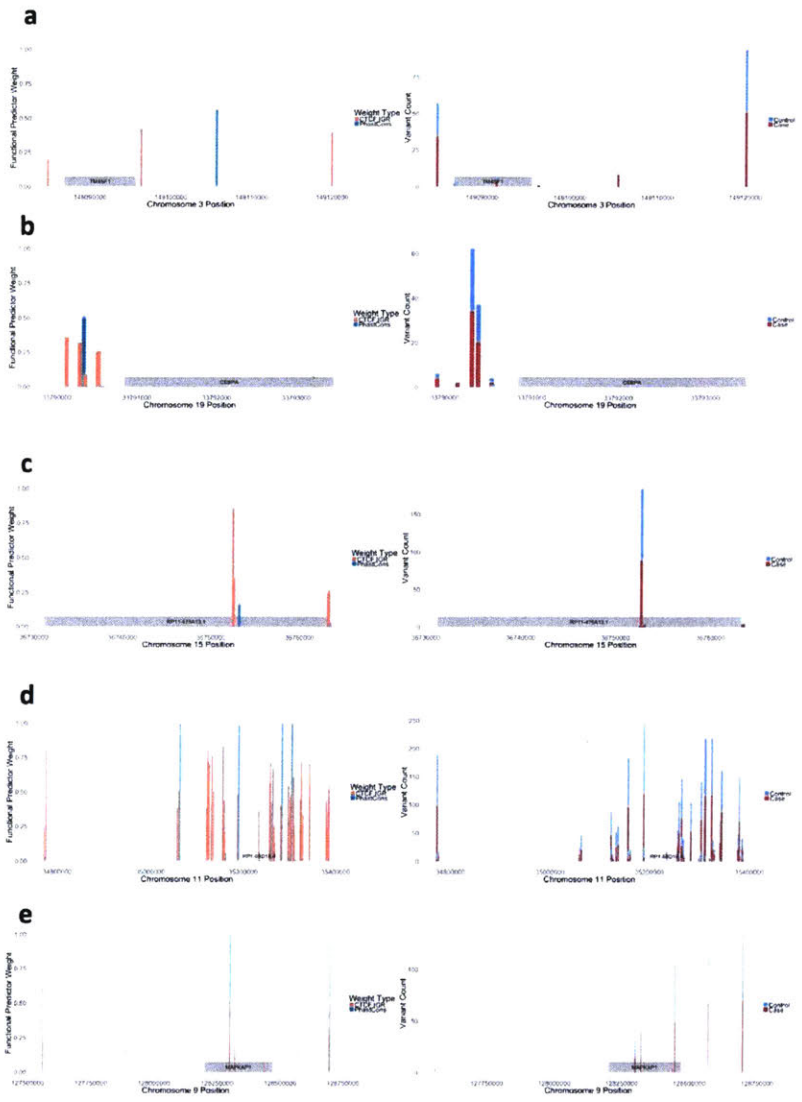


Figure 7 Pictured are the loci of the genes **a)** TM4SF1, **b)** CEBPA, **c)** RP11-475A13.1, **d)** RP1-68D18.4, and **e)** MAPKAP1. Left are the different weights (CTCF IGR and PhastCons) assigned to each variant falling within the gene's plexus for a given link set. Right are the different distributions of the variants in cases versus controls.

4. Discussion

In this work, we have identified novel gene targets and reaffirmed known genetic associations with T2D via an orthogonal approach to GWAS based on grouping rare variants by attempting to understand their potential regulatory effect on normal gene expression. These results were deduced even in the absence of prior enrichment of variants in regulatorily relevant epigenomic regions and without bias towards especially rare variants in any epigenetic state, a testament to the ability of this framework to uncover latent genetic association to complex disease. Grouping variants based on distal regulatory interactions enables discovery of potential regulatory mechanism disrupted in disease or differential activity across cases and controls.

Further, we have demonstrated that some findings are robust to the testing parameters, namely the enhancer-gene assignment set used and the functional prediction metrics for variants. Multiple genes evaluate as being significantly associated with either PhastCons or CTCF IGR weighting, suggesting a convergence of information that can be gleaned from these metrics and a validity to using either approach to ascertain the functional nature of variants. The significant of CTCF weighting could indicate that potential mechanism underlying the effect on the gene targets could be due to the alteration of chromosomal domain boundaries or loopings. In several other cases, however, different genes were determined to be associated with T2D using different assignment sets, suggesting that different sets of regulatory loci may capture different aspects of regulatory potential and that the differences enable motivation of different, somewhat orthogonal approaches to understanding how variants might disrupt regulatory potential at different scales. Despite their promise, the predicted regulatory loci are still imperfect and undoubtedly contain false positives for regulatory potential, which undoubtedly affects the model's ability to discover genetic associations. This highlights the need for further research

regarding the distinctive features of noncoding regions with regulatory potential and their corresponding target genes.

Another interesting finding of this study is that some of the variants with the largest predicted functional consequences were overburdened in controls rather than cases as might have been expected. This surprising result is suggestive that rather than variants acting by disrupting normal regulatory function in a way that causes dysregulation of genes in a deleterious manner, some variants may have a compensatory effect in that dysregulation of certain genes may offset other nascent causes that may otherwise have lead to T2D, such as environmental factors like diet or indirect genetic effect. Already we have demonstrated that these some genes found from this framework have clear implications in biology of relevant tissues for systemic processes involved in homeostasis, inflammation, and proliferation. The cases of RP11-475A13.1 and RP1-68D18.4 show the power of this framework to detect intergenic regulation of non-coding genes and how regulation of these regulators might have major downstream pathological consequences, an indirect effect that would remain otherwise elusive.

The novelty of this research lies in the principled aggregate testing of regulatory rare variants through extensive annotation of gene plexi. First the principal novelty of this work is the development the most comprehensive regulatory plexi that have ever been leveraged to test for detecting regulatory rare variant association in complex disease. This will be the first application of multiple-enhancer aggregation of regulatory variants in the context of GWAS, whereby several different approximations of regulatory potential were tested in the context of a specific complex disease. Second, this research utilizes novel metrics for ascertaining the functional consequence of a non-coding rare variant for weights, namely the IGR algorithm, an algorithm developed to predict the affinity of binding between transcription factors (TF) and their

associated binding sites, and phastCons, a metric which serves as a proxy for the likelihood that a locus is evolutionary conserved. Lastly, this work compares several different approximations of regulatory potential were tested in the context of a specific complex disease. The primary thesis of this work was that principled aggregate testing of rare variants across different functionality metrics and by building up rich regulatory networks based on the logic that disruption of tissue-specific distal regulators may have genetic consequences would lead to unprecedented power to detect rare variant-driven association in complex diseases.

Some limitations of this study are due to the data used, the restriction to enhancers as the primary regulatory vehicle, and the means of evaluating functional prediction. For the data, we were limited by the size of the GoT2D study in number of genomes available and the depth of sequencing, two issues which will likely be ameliorated in further whole genome studies as the price of sequencing continues to decrease. For this work, we focused our investigation on understanding variants that fell in regions predicted to be enhancers interacting with specific genes. While this provides narrow focus and already has demonstrated fruitful results, a more holistic approach to regulatory variation considering alternative mechanisms for regulatory action could be considered. In future investigation, one might also consider integrating regions that are topologically associated with genes, for instance from Hi-C data, into gene plexi. Further, future studies should focus on analyzing the different structural properties of plexi and how those may provide biological insights into the association between modes of gene regulation and function. Finally, this work considers individual prediction metrics for variant effects independently, but future work may consider developing prediction metrics for weighting variants considering regressions over vectors of different weights. Ultimately, successful integrative analyses such as this one will improve our understanding of the biological

mechanism of T2D and open the door to unraveling the mechanisms of complex traits in general.

5. Acknowledgements

There are many people I have to thank for making it so through this incredible MIT adventure. First and foremost, I am eternally grateful to have such loving, supportive parents as I do. I would hardly be the person I am today without the top-notch education they have provided me, as an academic student, but more importantly, as their son. I am also incredibly grateful to have my sister Sarah, whose enthusiasm for my future endeavors and empathy for our shared experiences are unmatched. Of course, I would be remiss not to thank all of my close friends who make every day unique and exciting. To my friends Brandon, Robby, Dalsin, Smitty, and the rest of the Minnesota gang, you guys are lifelong friends since Day 1, and I can always count on you guys for some unique antics back in flavortown. To Edgar, Orneels, Tito, Raul, Quinhas, Dlernz, John de Jesus, Jerbear, JC, Nosh, OG, Isaac, Javier, Zmills, Egar, Mgrim, Casanch, Oso, Fitz, Mael, Pasquacks, Phat, Reymundo, Dadvar, Glowmez, Serge, and all the rest of the brothers I love and respect so much, you have made my undergraduate (and Master's) experience absolutely unforgettable. I have grown and matured so much these past five years, and I know now to treasure every moment. Thank you.

Additionally, I have had the great privilege of having so many incredible mentors throughout my time here. It has been a great honor to work with Manolis these past two years on such an urgent, exciting project. I have learned so much from him both in and out of the classroom, and I look forward to bringing my knowledge to my future ventures. His generous support for my development has been instrumental in my success. Other mentors who have had a huge impact in my life academically are Jason Flannick and Seymour de Picciotto, who trained me in different research areas and showed me two paths in the breadth of biology. I have also worked with phenomenal advisors in industry, namely Joe Manfredonia and Brian Bettencourt, who both inspired me to pursue

the computational and analytical aspects of bioinformatics with their great advisorship even beyond my internship experiences. Lastly, I would like to thank Denny Freeman, my incredible advisor, and all of the MIT faculty and staff, especially in Course 6, who have provided me with direction and an unparalleled education. I am eternally humbled by the opportunity.

6. References

1. Prasad, R. B., & Groop, L. (2015). Genetics of Type 2 Diabetes—Pitfalls and Possibilities. Genes 6(1), 87123. http://doi.org/10.3390/genes6010087

2. Mahajan, R., & Gupta, K. (2014). Prevention and management of type 2 diabetes: Potential role of genomics. International Journal of Applied & Basic Medical Research, 4(Suppl 1), S1. http://doi.org/10.4103/2229-516X.140704

3. Tattersall, R. (1998). Maturity-onset diabetes of the young: a clinical history. Diabetic Medicine, 15(1), 1114.http://doi.org/10.1002/(SICI)1096-9136(199801)15:111::AID-DIA5613.0.CO;2-0

4. Zuk, O., Hechter, E., Sunyaev, S. R., & Lander, E. S. (2012). The mystery of missing heritability: Genetic interactions create phantom heritability. Proceedings of the National Academy of Sciences, 109(4), 11931198. http://doi.org/10.1073/pnas.1119675109

5. Thormaehlen, A. S., Schuberth, C., Won, H.-H., Blattmann, P., Joggerst-Thomalla, B., Theiss, S., Runz, H. (2015). Systematic Cell-Based Phenotyping of Missense Alleles Empowers Rare Variant Association Studies: A Case for LDLR and Myocardial Infarction. PLoS Genet, 11(2), e1004855. http://doi.org/10.1371/journal.pgen.1004855

6. Cowper-Sallari, R., Zhang, X., Wright, J. B., Bailey, S. D., Cole, M. D., Eeckhoute, J., Lupien, M. (2012). Breast cancer risk-associated SNPs modulate the affinity of chromatin for FOXA1 and alter gene expression. Nature Genetics, 44(11), 11911198. http://doi.org/10.1038/ng.2416

7. ENCODE Project Consortium. (2012). An integrated encyclopedia of DNA elements in the human genome. Nature, 489(7414), 5774. http://doi.org/10.1038/nature11247

8. Roadmap Epigenomics Consortium, Kundaje, A., Meuleman, W., Ernst, J., Bilenky, M., Yen, A., Kellis, M. (2015). Integrative analysis of 111 reference human epigenomes. Nature, 518(7539), 317330. http://doi.org/10.1038/nature14248

9. Cowper-Sallari, R., et al. Convergence of dispersed regulatory mutations reveals candidate driver genes in prostate cancer.

10. Fuchsberger, C., Flannick, J., Teslovich, T. M., Mahajan, A., Agarwala, V., Gaulton, K. J., McCarthy, M. I. (2016). The genetic architecture of type 2 diabetes. Nature, 536(7614), 4147. http://doi.org/10.1038/nature18642

11. Siepel, A., Bejerano, G., Pedersen, J. S., Hinrichs, A. S., Hou, M., Rosenbloom, K., Haussler, D. (2005). Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. Genome Research, 15(8), 10341050. http://doi.org/10.1101/gr.3715005

12. Ernst, J., Kheradpour, P., Mikkelsen, T. S., Shoresh, N., Ward, L. D., Epstein, C. B., ... Bernstein, B. E. (2011). Mapping and analysis of chromatin state dynamics in nine human cell types. *Nature*, *473*(7345), 43–49. https://doi.org/10.1038/nature09906

13. Wang, J., Kundaje, A., Kellis, M. Statistical inference of enhancer-gene interactions in 56 human cell and tissue types.

14. Lin, D. Y. and Tang, Z. Z. (2011). A General Framework for Detecting Disease Associations With Rare Variants in Sequencing Studies. American Journal of Human Genetics, 89, 354-367.12

15. Samocha, K. E., Robinson, E. B., Sanders, S. J., Stevens, C., Sabo, A., McGrath, L. M., ... Daly, M. J. (2014). A framework for the interpretation of de novo mutation in human disease. *Nature Genetics*, *46*(9), 944–950. https://doi.org/10.1038/ng.3050

16. Hnisz, D., Day, D. S., & Young, R. A. (2016). Insulated Neighborhoods: Structural and Functional Units of Mammalian Gene Control. *Cell*, *167*(5), 1188–1200. https://doi.org/10.1016/j.cell.2016.10.024

17. Jernås, M., Palming, J., Sjöholm, K., Jennische, E., Svensson, P.-A., Gabrielsson, B. G., ... Lönn, M. (2006). Separation of human adipocytes by size: hypertrophic fat cells display distinct gene expression. *The FASEB Journal*, *20*(9), 1540–1542. https://doi.org/10.1096/fj.05-5678fje

18. Wicks, K., Torbica, T., Umehara, T., Amin, S., Bobola, N., & Mace, K. A. (2015). Diabetes Inhibits Gr-1+ Myeloid Cell Maturation via Cebpa Deregulation. *Diabetes*, *64*(12), 4184–4197. https://doi.org/10.2337/db14-1895

19. Zhang, X., Rowan, S., Yue, Y., Heaney, S., Pan, Y., Brendolan, A., ... Maas, R. L. (2006). Pax6 is regulated by Meis and Pbx homeoproteins during pancreatic development. *Developmental Biology*, *300*(2), 748–757. https://doi.org/10.1016/j.ydbio.2006.06.030

20. Liu, J., Wang, Y., Birnbaum, M. J., & Stoffers, D. A. (2010). Three-amino-acid-loop-extension homeodomain factor Meis3 regulates cell survival via PDK1. *Proceedings of the National Academy of Sciences*, *107*(47), 20494–20499. https://doi.org/10.1073/pnas.1007001107

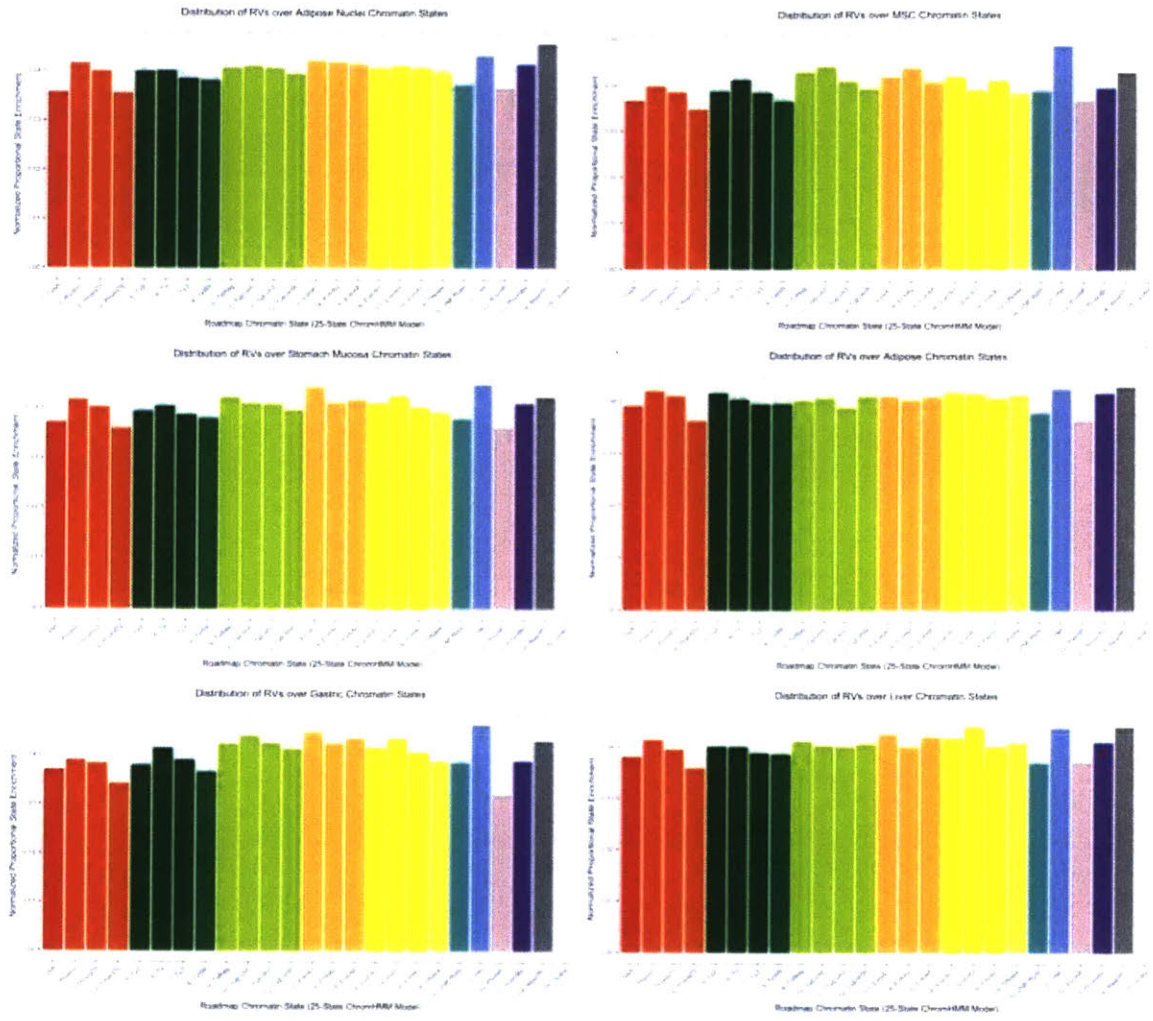
21. Berger, U. V., & Hediger, M. A. (2006). Distribution of the glutamate transporters GLT-1 (SLC1A2) and GLAST (SLC1A3) in peripheral organs. *Anatomy and Embryology*, *211*(6), 595–606. https://doi.org/10.1007/s00429-006-0109-x

22. Kodama, K., Horikoshi, M., Toda, K., Yamada, S., Hara, K., Irie, J., ... Butte, A. J. (2012). Expression-based genome-wide association study links the receptor CD44 in adipose tissue with type 2 diabetes. *Proceedings of the National Academy of Sciences*, *109*(18), 7049–7054. https://doi.org/10.1073/pnas.1114513109

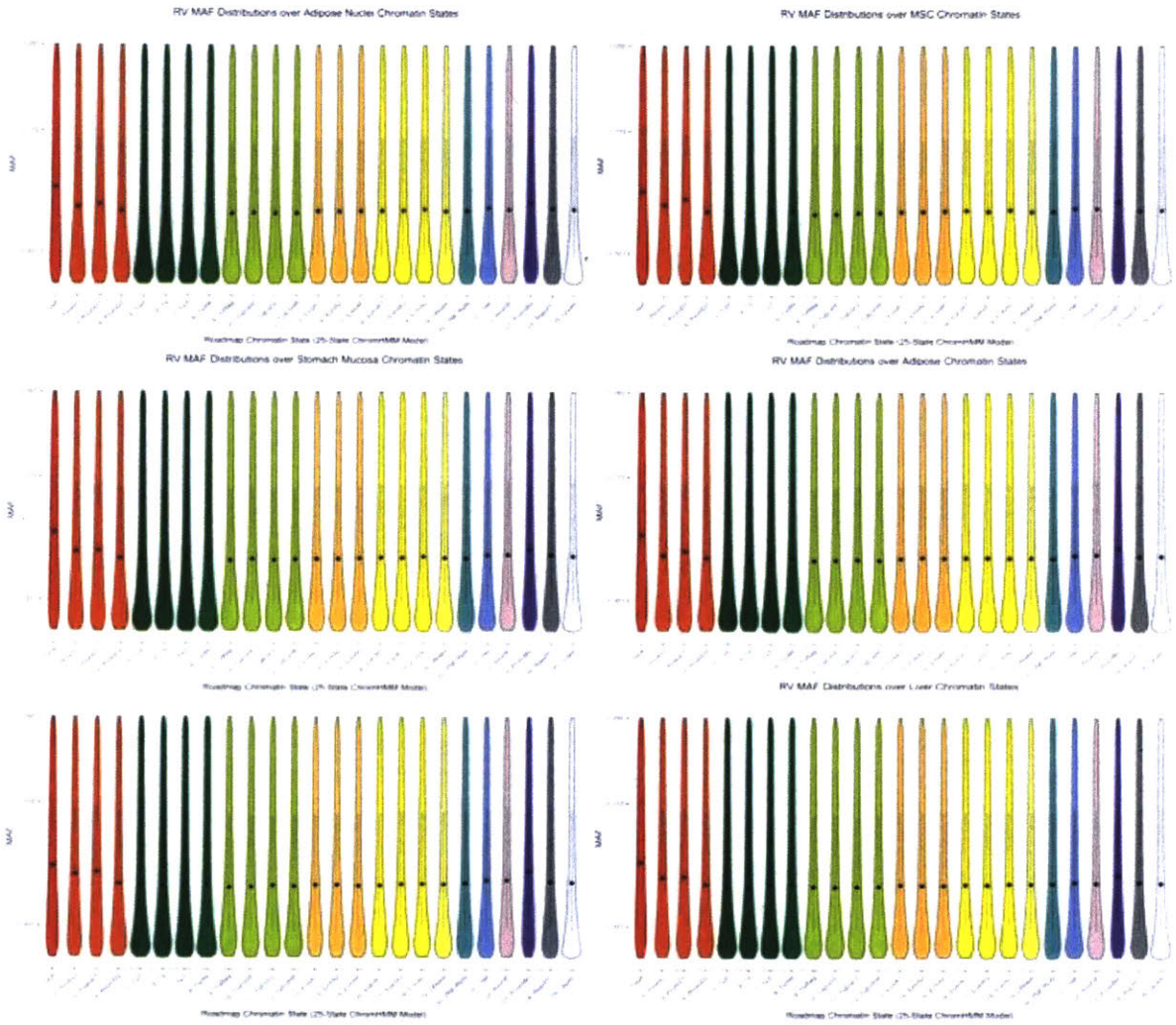
23. Moon, S., Hwang, M. Y., Jang, H. B., Han, S., Kim, Y. J., Hwang, J.-Y., ... Kim, B.-J. (2017). Whole-exome sequencing study reveals common copy number variants in protocadherin genes associated with childhood obesity in Koreans. *International Journal of Obesity*, *41*(4), 660–663. https://doi.org/10.1038/ijo.2017.12

24. Zoncu, R., Efeyan, A., & Sabatini, D. M. (2011). mTOR: from growth signal integration to cancer, diabetes and ageing. *Nature Reviews. Molecular Cell Biology*, *12*(1), 21–35. https://doi.org/10.1038/nrm3025

7. Supplementary Figures



Supplementary Figure 1 Distribution of variants in various Roadmap epigenomic states from the 25-state model, normalized for genome coverage of each state in (clockwise from the top left) adipose nuclei, MSC, adipose, liver, gastric, and stomach mucosa cell types. LF and rare variants are not disproportionately represented in any specific annotation state.



Supplementary Figure 2 Distribution of MAFs for variants in different epigenomic states for (clockwise from top left) adipose nuclei, MSC, adipose, liver, gastric, and stomach mucosa cell types. No discernible biases for lower MAFs are present in any particular state.

8. Supplementary Tables

	Pancreatic Islets			Liver		
	Name	-log10(*p*)	Sig	Name	-log10(*p*)	Sig
Co-activity - CTCF IGR	TM4SF1	6.77272672	***	SNHG15	3.86683546	
	GC	4.09648551	-	USP36	3.74555039	
	MSI2	3.64230427		PVR	3.71176397	
	AOX1	3.52313507		AKR1C2	3.57987235	
	RP11-422N1(	3.03993039		GPR126	3.4531664	
Co-activity - phastCons	UGT8	3.59566928		C17orf70	3.69430856	
	MAN1A1	2.75955269		PVR	3.55387056	
	FKBP1B	2.56798437		ZCCHC2	3.44988623	
	PDE8B	2.38498611		CFLAR	3.19111808	
	AOX1	2.29533257		LINS	3.15472951	
Co-activity+prom - CTCF IGR	RN7SL103P	6.29324332	*	KRT8P43	5.60864635	-
	KRT8P43	5.60864635	-	RP5-1022P6.	5.52115107	
	RP5-1022P6.	5.52115107		RP11-475A1:	5.44435663	
	RP11-475A1:	5.44435663		RAB6A	4.89896055	
	CTC-369A16.	4.85611935		CTC-369A16.	4.85611935	
Co-activity+prom - phastCons	CCDC149	6.53971564	**	CCDC149	6.53971564	**
	Y_RNA	6.29324332	*	Y_RNA	6.29324332	*
	KRT8P43	5.64223943	-	KRT8P43	5.64223943	-
	RNA5SP360	5.43670112		RAB6A	5.50027471	
	LCP2	5.06610506		RNA5SP360	5.43670112	
Module LDA - CTCF IGR	CBLL1	4.14344179		CEBPA	4.68587748	*
	PCF11	3.25470114		F12	4.06859162	
	RIC3	3.18610927		DHX8	3.26042961	
	PBX3	3.16666624		SLC39A11	3.15715795	
	RP11-423H2.	2.97781353		SUCLG1	2.89462515	
Module LDA - phastCons	PBX3	4.08899772		CEBPA	4.56734259	*
	PDCD4	3.99209807		IGFBP4	4.08195861	
	FAM102A	3.2616794		PC	3.18687429	
	PCF11	3.25470114		PLIN2	2.88746778	
	NEMF	3.14367013		FEZ1	2.76281711	
Module LDA+prom - CTCF IGR	KRT8P43	5.60864635	-	KRT8P43	5.60864635	-
	RP5-1022P6.	5.52115107		RP5-1022P6.	5.52115107	-
	RP11-475A1:	5.44435663		RP11-475A1:	5.44435663	-
	CTC-369A16.	4.85611935		KIAA1715	4.83209566	
	DIS3L2	4.74886901		DIS3L2	4.74886901	
Module LDA+prom - phastCons	Y_RNA	6.29324332	*	Y_RNA	6.29324332	*
	KRT8P43	5.64223943	-	KRT8P43	5.64223943	-
	RNA5SP360	5.43670112		RNA5SP360	5.43670112	
	LCP2	5.06610506		LCP2	5.06610506	
	AC073065.3	4.73488695		RN7SL103P	5.03362169	

Supplementary Table 1 Top five genes per condition using Co-activity or Module LDA enhancer-gene assignments.

	Pancreatic Islets			Adipose		
	Name	-log10(p)	Sig	Name	-log10(p)	Sig
Dnase - loose - CTCF IGR	TP63	6.17787698	**	MAPKAP1	5.451399765	*
	ALG9	4.6674844		FBLN2	5.034301328	
	SNRK	4.65737539		ABHD5	4.657375395	
	ABHD5	4.65737539		SEC13	4.58286567	
	NOP58	3.78627851		DLL1	4.470202271	
Dnase - loose - prom - CTCF IGR	TP63	6.17787698	*	ERAP1	5.256326545	-
	UAP1	5.06707501		DIS3L2	4.748869013	
	MAPKAP1	5.06066436		MYBPC1	4.617262736	
	DIS3L2	4.74886901		SLC28A2	4.586652003	
	CTC-369A16.3	4.55750541		TBCCD1	4.547053514	
Dnase - strict - CTCF IGR	TP63	6.17787698	**	B3GNT1	5.320247565	-
	PBX3	4.92502572		UTP18	4.862117534	
	LNPEP	4.89761407		NME1	4.862117534	
	AMBN	4.86147373		CTC-506B8.1	4.748484672	
	ERAP1	4.78702746		POP4	4.688983459	
Dnase - strict - prom				MAPKAP1	5.79902116	*
				UTP18	4.862117534	
				NME1	4.862117534	
				ABHD5	4.657375395	
				WFS1	4.592305665	
Dnase - medium	CCNL1	4.88163031				
	LIX1	4.74515721				
	LNPEP	4.74188649				
	ERAP2	4.74188649				
	AC008865.1	4.74188649				
Dnase - medium - prom	LIX1	4.74481873		RNU6-11S4P	4.966972615	
	AC008865.1	4.74202862		PBX3	4.807506545	
	LNPEP	4.74188649		MYL4	4.454548404	
	NEFH	4.37523583		MYBPC1	4.152629557	
	RNF166	3.9730452		RPL24P8	4.086584894	
Element - loose - CTCF IGR	RP11-452F19.3	6.15980209	**	ENO3	4.977870346	
	CMPK1	4.97287341		PPARGC1A	4.284076252	
	SNRK	4.65737539		RP5-894D12.3	4.200610368	
	ANO10	4.65737539		GATM	4.053863574	
	ABHD5	4.65737539		RP11-109D20.2	3.998428938	
Element - loose - prom - CTCF IGR	RP11-436I9.5	4.93543763		ACVRL1	5.739785616	*
	ABHD5	4.82942514		TRMT2A	4.900681795	
	TEX2	4.82940235		DIS3L2	4.748869013	
	SNRK	4.80364447		DUOXA1	4.634947042	
	DIS3L2	4.74886901		IER5	4.604228124	
Element - strict - CTCF IGR	DUOX2	5.61211199	*	MAPKAPK5-AS1	4.344713521	
	DUOX1	5.04485649		KRT121P	4.263479203	
	CTD-2033D15.1	4.74296911		SNRPF	3.959488716	
	LIX1	4.74188649		NDUFA4L2	3.86437364	
	RNPEPL1	4.67064257		AC073610.5	3.832105837	
Element - strict - prom	EIF3I	6.00601815	*	TMEM203	5.649587229	*
	RP11-54O7.3	5.68111146	*	AC005104.3	5.314137157	-
	DIS3L2	4.74886901		RP11-531A24.5	4.96415143	
	LIX1	4.74188649		UBLCP1	4.504997205	
	PLCL2	4.69161879		MASP2	4.46891252	
Element - medium	LIX1	4.74188649		CAPN2	5.761461405	*
	NEFH	4.51381498		MAPKAP1	4.75216248	
	ZMAT5	4.48235934		RP11-288I21.1	4.459941287	
	LINC00948	4.29125763		NR6A1	4.129128253	
	ZFAND2A	4.26081562		FAM214A	4.11912695	
Element - medium - prom	ZMAT5	4.48463764		SNORD87	5.398781834	-
	C9orf156	4.43674157		ANKRD28	4.660957741	
	KB-1615E4.2	4.1746474		RHOF	4.630635399	
	NANS	3.955678		SPC25	4.463456302	
	AC092634.2	3.95052406		IGFBP2	4.007915944	
Extended - loose - CTCF IGR	ENO3	4.97787035		RP1-68D18.4	9.662376332	***
	QPCTL	4.89344141		PCDHGA7	9.434428854	***
	AMBN	4.8614706		PCDHGA8	7.333951177	***
	SNORA48	4.8098277		OBSCN	5.09854385	
	ABHD5	4.65737539		SPAG7	4.977870346	
Extended - loose - prom - CTCF IGR				FIZ1	5.683867225	*
				ISOC1	5.640283046	*
				FAM99A	5.593254506	*
				PIGY	5.248994567	-
				ZC3H18	5.156265115	
Extended - strict - CTCF IGR	TCEB1P19	3.24627088				
	PPP6C	2.8912556				
	USE1	2.85871885				
	TRIM72	2.83645326				
	CYP4F2	2.47271109				
Extended - strict - prom						
Extended - medium	RIOK2	5.30689867	-			
	PRICKLE1	5.04155491				
	QARS	4.99026039				
	SP9	4.97318541				
	SCN5A	4.70754255				
Extended - medium - prom	SP9	4.98952272				
	IMPDH2	4.85514984				
	FBXO46	4.81699659				
	FLJ00104	4.71846444				
	LNPEP	4.54794256				

Supplementary Table 2 Top five genes per condition using DNase-based enhancer-gene assignments with CTCF IGR weighting.