Understanding the Genetics of Gene Regulation Using Multi-Omics Profiling

by

Arushi Varshney

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
(Human Genetics)
in The University of Michigan
2019

Doctoral Committee:

Associate Professor Stephen C. J. Parker, Chair
Professor Sally A. Camper
Assistant Professor Jacob Kitzman
Professor Mats Ljungman
Associate Professor Cristen J. Willer

"The truth is,
most of us
discover where
we are heading
when we arrive."

Bill Watterson
Calvin and Hobbes

Arushi Varshney

arushiv@umich.edu

ORCID iD: 0000-0001-9177-9707

To my family

ACKNOWLEDGEMENTS

My work has been possible with the mentorship and support of many people and I now take the opportunity to thank them and convey my heartfelt gratitude.

I thank my thesis advisor Prof. Stephen Parker for his exceptional mentorship. He's a brilliant scientist, creative and ambitious and is a genuinely nice person. It has been great working with and learning from him. He looks out for his students, gives constructive feedback and is one of the most optimistic people I know. Of all the things I have learned from him, staying positive is something I'll especially continue to strive for in all walks of life. I thank my fellow lab members Ricardo Albanus, Dr. Venkat Elangovan, John Hensley, Dr. Yoshi Kyono, Dr. Nandini Manickam, Peter Orchard, Dr. Daniel Quang and Vivek Rai for their support, motivation and friendship. It has been an honor to have known them and worked with them.

My thesis committee comprised by Prof. Sally Camper, Prof. Jacob Kitzman, Prof. Mats Ljungman and Prof. Cristen Willer has been just wonderful. All members have given invaluable feedback and support throughout my dissertation work. I thank them for helping me rethink goals and approaches for my final project and helping me stay motivated. I always felt better about my work after my committee meetings. I also thank Jacob for his contributions towards experiments and analyses for my project.

I have been extremely lucky to have had opportunities to work with so many brilliant and kind people through the FUSION, insPIRE and MAGIC groups. Not only did I learn scientific skills from them, but I also acquired other critical skills such

as teamwork and communication. I especially thank Dr. Mike Erdos, Narisu Narisu, Dr. Laura Scott, Dr. Ana Viñuela, Dr. Martijn van de Bunt, Prof. Inês Barroso, Prof. Mark McCarthy, Prof. Karen Mohlke, Prof. Mike Boehnke and Prof. Francis Collins for their kindness and support. It has been my absolute honor to have worked with them all. I especially thank Prof. Karen Mohlke and Prof. Francis Collins for helping with letters of recommendation.

I thank the Human Genetics (HG) department staff who made my life so much easier - Sue Kellogg, Karen Grahl, Molly Martin and Alex Terzian. It was only with Alex's help that my parents and other family members could join my oral thesis defense remotely from India, it really meant a lot to me. I thank all HG faculty members who were always so nice to me. I thank Prof. Stephanie Bielas and Prof. Sundeep Kalantry for happily writing letters of recommendation for many applications. I thank all my Ann Arbor friends and fellow HG graduate students for all their support in so many ways.

I owe thanks to my teachers for their mentorship, especially Mrs. Shilpa Vishnoi and Mrs.Tajili Ansari during my early schooling; Prof. Aiyagari Ramesh and Prof. R. Swaminathan during my undergraduate studies and Prof. Philippe Renaud during my master's studies. They strongly motivated my interest in science and research and greatly supported me as I decided to pursue PhD studies and throughout the application process.

My wonderful little family has been my driving force all these years. My sister Aakriti has been a lovely friend all along. My mother-in-law Mrs. Kanti Deshwali and sister-in-law Dr. Akanksha Deshwali have been nothing by warm and loving. My husband and best friend Dr. Anand Pratap Singh has unconditionally loved and supported me during the years and has been the best partner in all times good and bad. I thank him for being the amazingly calm person he is, listening to me and keeping me sane. My parents Mrs Beena and Mr. Pravish Varshney have always

been unwaveringly supportive. Their love, encouragement and sacrifice has got me here today. I will strive to follow their examples of hard work and integrity in my career and life.

TABLE OF CONTENTS

LIST OF FIGURES

Figure

LIST OF TABLES

Table

LIST OF ABBREVIATIONS

ATAC-seq	assay for transposase accessible chromatin followed by sequencing
BMI	body mass index
ChIP-seq	chromatin immunoprecipitation followed by high-throughput sequencing
DNase-seq	DNase I digestion coupled to DNA sequencing
eQTL	expression quantitative trait locus
ERCC	External RNA Control Consortium
FAIRE-seq	formaldehyde-assisted isolation of regulatory elements followed by sequencing
GO	gene ontology
GTEx	genotype tissue expression
GWAS	Genome wide association studies
H3K27ac	histone H3 lysine 27 acetylation
H3K4me3	histone H3 lysine 4 trimethylation
HMM	hidden markov model
HOT	high occupancy target
insPIRE	integrated network for systematic analysis of pancreatic islet RNA expression
LCL	lymphoblastoid cell line
LD	linkage disequilibrium

lncRNA	long non-coding RNA
MAF	minor allele frequency
MAF	non-Amplified non-Tagging Illumina Cap Analysis of Gene Expression
NSC	Normalized strand cross-correlation
PWM	position weight matrix
RFX	regulatory factor X
ROSE	rank ordering of super enhancers
RSC	Relative strand cross-correlation
SNPs	Single nucleotide polymorphisms
T1D	Type 1 diabetes
T2D	Type 2 diabetes
TF	transcription factor
TSS	transcription start site

ABSTRACT

Type 2 diabetes (T2D) is a complex disease that affects an estimated 415 million people worldwide. Genome wide association studies (GWAS) have identified >240 genetic signals that encode predisposition to this disease and related traits. However, the underlying biological mechanisms driving this predisposition are largely unknown, which is a serious impediment in designing precision therapeutic strategies. The focus of my research is to untangle the genetic complexity of T2D to better understand the biological mechanisms of how disease predisposition is encoded in our DNA. Specifically, I aim to understand how T2D genetic risk variants modulate gene expression in orchestrating disease mechanisms.

I utilize high throughput molecular profiling data in human pancreatic islets and other diverse tissues along with human and rodent cell line model systems and employ computational and experimental approaches to map functional signatures of genetic variants associated with T2D. First, I compared gene regulatory annotations defined using diverse epigenomic data across 4 cell types to compare their cell specificities and genetics of gene expression regulation. I observed that genetic variants in genomic regions with more cell type-specific enhancer chromatin have lower effects on gene expression than variants in genomic regions with more ubiquitous promoter chromatin. However, genetic variants in cell type-specific enhancer regions have higher effects in chromatin accessibility than those in less cell type-specific promoter regions. Second, I integrated GWAS data with various -omics data in islets to nominate bi-

ological mechanisms. I observed that T2D risk variants confluently disrupt DNA binding motifs of the transcription factor (TF) regulatory factor X (RFX) in accessible regions. Third, I describe large scale expression quantitative trait locus (eQTL) mapping efforts along with integration of epigenomic data to describe molecular regulatory mechanisms. Utilizing such large eQTL and integrating information such as chromatin accessibility and TF binding predictions helped elucidate *in vivo* TF activity preferences. Fourth, I describe profiling and analysis of the enhancer transcriptome in islets, which I then integrate with other available epigenomic data to better understand the characteristics of gene regulation.

CHAPTER I

Introduction

T2D is a complex, life-altering and chronic disease, which affects an estimated 30 million Americans and 415 million people worldwide. Large-scale genetic studies have identified numerous independent genetic signals that encode predisposition to this disease and related traits. However, the underlying biological mechanisms driving this predisposition are largely unknown, which is a serious impediment in designing precision therapeutic strategies. The focus of my research is to untangle the genetic complexity of T2D to better understand the biological mechanisms of how disease predisposition is encoded in our DNA.

1.1 T2D pathophysiology

T2D is a heterogeneous syndrome characterized by hyperglycemia (increased plasma glucose levels) and abnormalities in carbohydrate and fat metabolism. The beta cells of the pancreatic islets of langerhans secrete the hormone insulin, which is essential to maintain normal levels of glucose in the body. Insulin secretion from the pancreas normally reduces glucose output by the liver, enhances glucose uptake by skeletal muscle, and suppresses fatty acid release from fat tissue. A combination of factors including resistance to insulin, inadequate insulin secretion, and excessive or inappropriate glucagon secretion contribute towards development of T2D. It has been

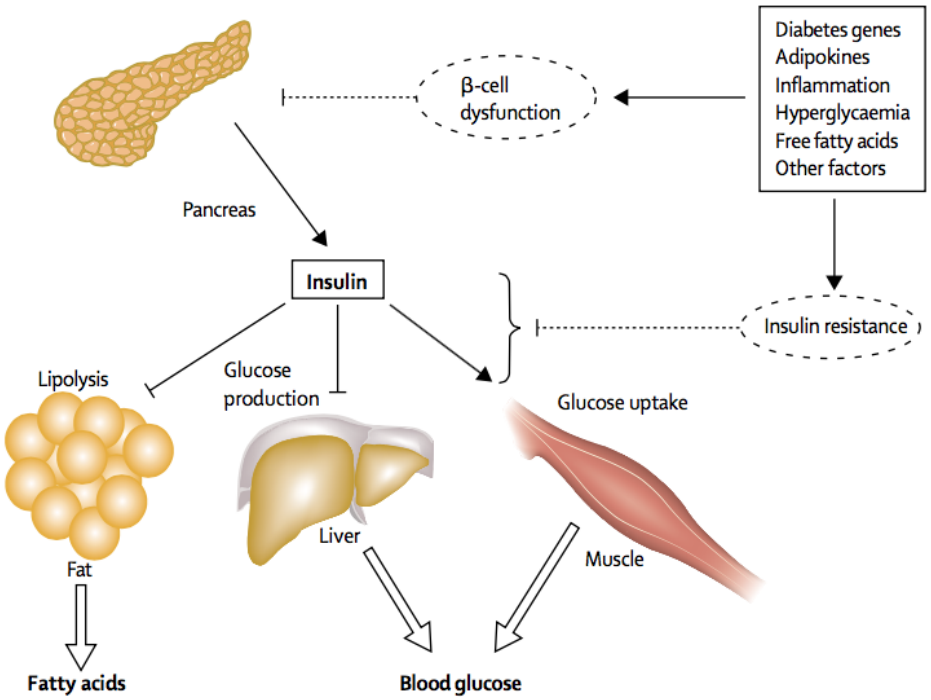


Figure 1.1: Pathophysiology of type 2 diabetes. Insulin resistance in peripheral tissues leads to increased circulating fatty acids and the hyperglycemia. In turn, the raised concentrations of glucose and fatty acids in the bloodstream will feed back to worsen both insulin secretion and insulin resistance. Reprinted from [176]

established that insulin resistance in peripheral tissues such as skeletal muscle, adipose (fat) and liver, which can arise partly due to obesity, results in an increased insulin demand to achieve glucose homeostasis in the body [176] (Fig. 1.1, adapted from [176]). The pancreas can usually compensate for this increased demand with increased insulin levels through an expansion of beta cell mass and/or insulin secretion by the beta cells. However, over time due to glucose toxicity and other factors, islet function decreases and failure to compensate for insulin resistance results in the development of T2D [81]. Conditions such as impaired fasting glucose and impaired glucose tolerance are known to predispose to the development of overt diabetes [176].

1.2 Genetic studies to understand T2D predisposition

T2D is the result of a complex interplay between genetic, epigeneomic and environmental factors. While obesity, diet and lifestyle are strong predictors of T2D, T2D also has a strong genetic component. Individuals with one parent who has T2D have a 40% estimated lifetime risk of developing the disease whereas the risk increases to 70% if both parents are affected. Therefore, identifying these genetic bases can provide crucial insights into T2D pathogenesis.

Numerous studies to date have aimed to identify genetic signatures of T2D. Early family-based linkage studies discovered that variants in the *TCF7L2* gene were associated with T2D [39]. Subsequent fine-mapping efforts indicated that an intronic variant rs7903146 contributed to the original linkage signal [39, 56]. Through more recent high throughput association studies has been confirmed in European, African and Asian populations and it is one of strongest and most consistently replicated genetic association with T2D with an odds ratio of 1.4. Interestingly, a recent large scale (genome wide) study identified seven independent signals at the *TCF7L2* locus [110], highlighting the complexity of the locus. Candidate gene studies elucidated that variants in genes including *KCNJ11*, *PPAR-g*, *ABCC8* among others are associated with T2D. While family-based linkage and candidate gene studies supplemented our understanding of the T2D genetic architecture, these approaches were found to be ultimately limited in the context of T2D. This is because these studies generally had smaller sample sizes and tested a select group of variants based on imperfect understanding of candidate biological pathways. Smaller scale and more focussed approaches have indeed been successful for many mendelian diseases that involve higher effect size and highly penetrant variants, however, emerging evidence posited that T2D had a more complex genetic architecture driven by more commonly occurring variants that would be predicted to have more modest effect sizes. This concept is also known as the common disease, common variant hypothesis (Fig. 1.2), adapted

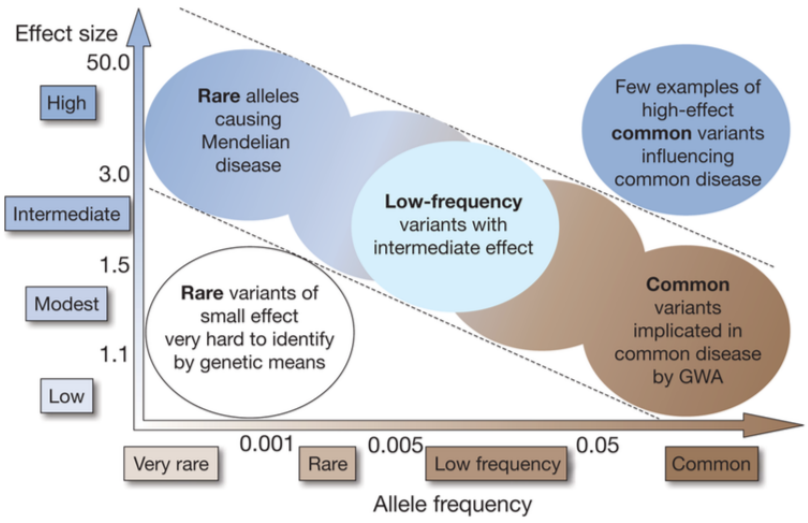


Figure 1.2: Relationship between strengths of effects (effect sizes) and risk variant frequencies. Reprinted from [115]

from [115]. Therefore, it was imperative to cast a wider net to identify the multiple genetic loci associated with the disease. Developments in the genotyping technology and enhanced cataloging of haplotypes (regions of the genomes that are inherited together) of common variants through efforts such as the HapMap [184], 1000 genomes [179] projects enabled testing millions of Single nucleotide polymorphisms (SNPs) for association with T2D. These genome wide association study (GWAS) approaches have now identified numerous loci (>240) associated with T2D and related traits [49, 111, 110]. Multiple GWAS studies have also be combined through meta analyses that can result in increased statistical power and add to the list of new loci. Trans-ethnic GWAS studies which can leverage differences in the genome structure across populations while considering differences in allele frequencies, are also highly effective in identifying loci [150, 181, 101].

The heritability estimates for T2D range from 20% to 70% across various studies [149, 5, 150, 110]. It has been suggested that since chip based GWA studies largely profiled common SNPs, rare variants that could not be profiled initially could explain the missing heritability [35]. However, recent studies with substantially increased

sample sizes and more complete coverage of low frequency variation have not bolstered this hypothesis for T2D [111].

While multitude of studies have demonstrated the potential of GWAS in identifying loci, it is important to note that GWA studies essentially report associations between genomic regions and the disease trait. GWAS, however, do not inform about the underlying causal molecular mechanisms; understanding these necessitate several exhaustive follow up studies. Mechanistic insights from a refined view of T2D genetics are essential to realize the translational value of GWA studies; such efforts may then allow for personalised risk scores [82], stratification of patients by different underlying pathophysiology [191] or towards identifying therapeutic targets.

Understanding the causal molecular mechanisms underlying T2D GWAS associations is quite challenging due to multiple factors. First, T2D and related trait GWA studies have largely implicated common variants that individually have modest effect sizes (odds ratios 1.1-1.5). Moreover, most of the variants occur in non-protein coding regions, suggesting that these do not directly affect protein structure or function. GWAS loci are commonly referred to by the names of genes located close to them for simplicity, however, only a few are close to strong biological candidates. Only occasionally one might find causal SNP candidates with particularly strong biological credentials such as those causing a non-synonymous change. Second, the lead GWAS SNP might not always be the causal SNP. This is because our genome is inherited in blocks such that multiple variants are highly correlated with each other and are said to in linkage disequilibrium (LD). Third, the target genes and how the GWAS SNP risk alleles affect their expression level (increasing/decreasing) are often unknown, as these may be distant. For example, an obesity-associated variant located in the intronic region of the *FTO* gene, does not affect the expression of this gene; instead, it influences the expression of the genes *IRX3* and *IRX5*, which are located over a megabase away from the variant [26]. These factors culminate in scenarios where

GWAS signals tag multiple, mostly non-coding variants where the causal SNP(s) and their target genes are difficult to identify using genetic information alone.

1.3 Flow of biological information from genetic variation to phenotype

To understand how genetic variation influences phenotypic traits, it is critical to consider several layers of molecular domains over which genetic information can propagate. The DNA encodes information which can affect the chromatin landscape and influence gene expression, effects of which can then relay to influence protein and metabolomic networks in eventually affect phenotypic changes (Fig. 1.3, adapted from [24]). As initial molecular control layers, understanding the epigenomic and transcriptomic effects of genetic variation can be highly informative in piecing together molecular mechanisms [92]. For example, genetic variants occurring in regulatory elements may confer risk by altering transcription factor binding sites that propagate signals from upstream transcription factors to influence downstream target gene expression. One of the first steps towards understanding the molecular impact of genetic variation on complex traits is therefore using epigenomic information to identify the regulatory elements through which these act.

1.4 Investigating the epigenomic domain to identify gene regulatory elements

DNA wraps around histone proteins and forms nucleosomes; covalent modifications on these histones and other patterns in this landscape helps establishing and maintaining relevant cell-specific and cell-identity gene expression programs. Different modifications on the histone tails have been observed to be associated with distinct functions. For example, promoters are marked by tri-methylation of histone

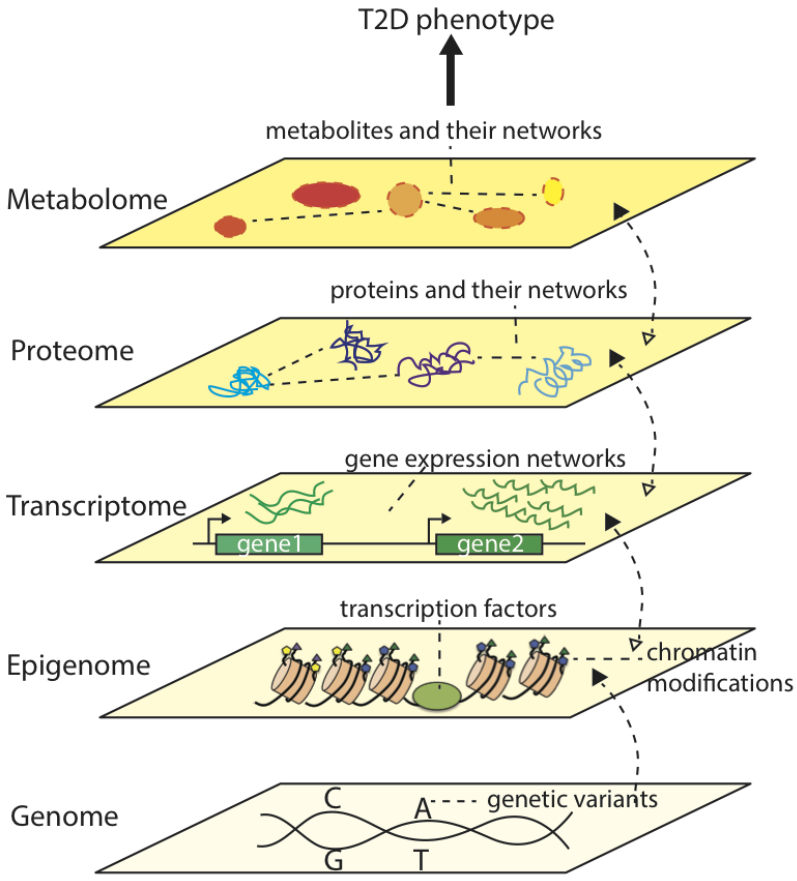


Figure 1.3: Molecular domains propagating genetic information towards phenotype. Adapted from [24]

H3 lysine 4 (H3K4me3) [11, 126, 1], enhancers are marked by mono-methylation of H3K4 (H3K4me1) [66] the acetylation of H3K27 (H3K27ac) mark is associated with both active promoter and enhancer activity [126]. Also, trimethylation of H3K36 (H3K36me3) is associated with transcribed regions; and trimethylation of H3K27 (H3K27me3) is associated with Polycomb repressed regions [66, 210]. These molecular modifications among others have been thoroughly profiled across a multitude of cell and tissue types using chromatin immunoprecipitation followed by sequencing (ChIP-seq) [182, 186]. Patterns of these diverse and informative signals have been distilled using hidden hidden markov model (HMM) method implemented in the ChromHMM tool [43, 41] to segment the genome into chromatin states. Parker, Stitzel and colleagues constructed chromatin state maps for pancreatic islets, and identified islet-specific stretch enhancers (SEs), which are long (3 kb) segments of the genome that are continuously decorated with enhancer-associated histone marks. Importantly, this study revealed that T2D GWAS loci are highly and specifically enriched to occur in islet stretch enhancers. Similar observations were also made by others and these studies collectively represent the first level of functional convergence in which disease-relevant variants across the genome are enriched in a set of large enhancers active in specific tissues [182, 120, 189, 142, 144, 153]. However, while chromatin state analysis is useful for narrowing down the regions of interest to a small subset of regulatory regions (Fig. 1.4, reprinted from [92]), the resolution of analysis is approximately 200 bp (a consequence of the fact that each nucleosome contains about 147 bp of DNA wrapped around the histones), which is still too coarse to pinpoint the underlying sequence motif(s) that could be mediating a genetic regulatory effect.

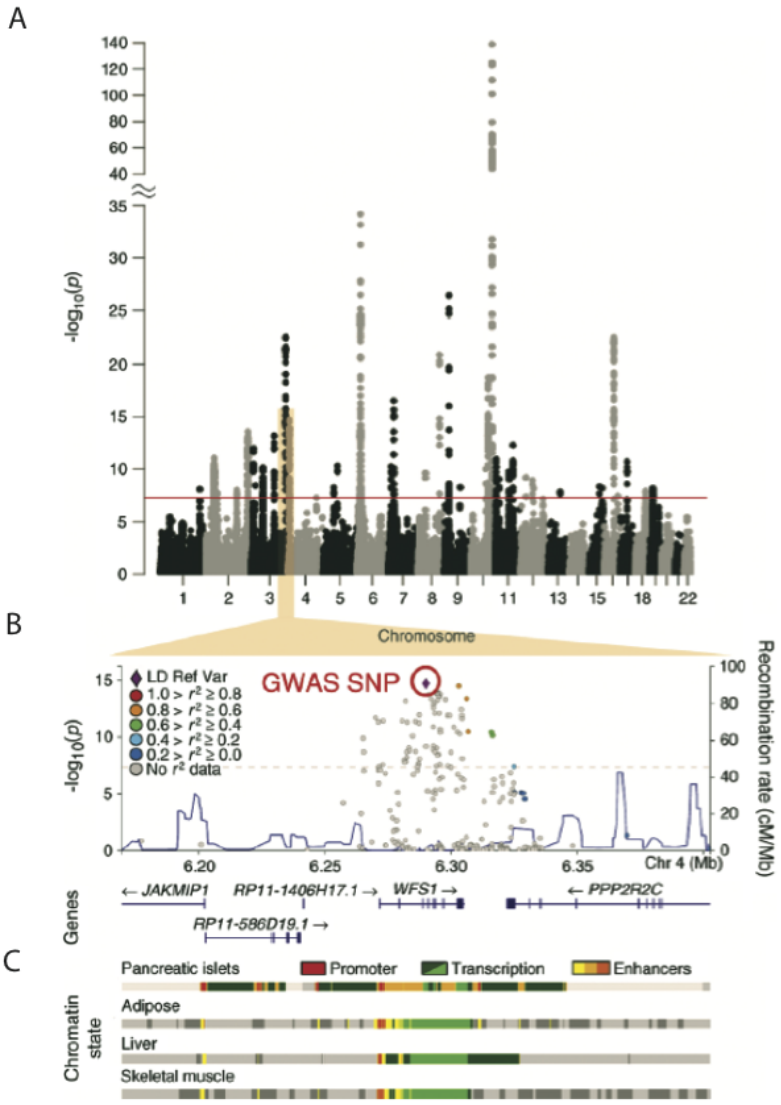


Figure 1.4: Functional mapping of diabetes-associated variants using tissue-specific regulatory maps. GWAS have identified loci associated with risk for type 2 diabetes, with strength of association (-$\log_{10}$ p value) shown throughout the genome in a Manhattan plot (top, data from [180, 178]). Each genome-wide significant region (above the horizontal red line) can then be explored using a locus-zoom plot (B), which shows one of the type 2 diabetes-associated loci (overlapping the gene *WFS1*) as an example [178]. In the locus zoom plot, each dot represents a variant associated with type 2 diabetes, and its colour represents the level of LD, with the lead variant (reference variant [Ref Var]) highlighted in purple. Most SNPs occur in non- coding regions, where chromatin state analyses (C) help identify locations of tissue-specific regulatory regions. While some enhancer regions may be shared across tissues, there are others that are unique. Reprinted from [92]

1.5 Profiling accessible chromatin to identify regulatory elements in high resolution

To identify regulatory segments at a higher resolution, it is imperative to locate the binding sites of TFs. TFs are known to bind in regions of accessible or open chromatin regions, or, conversely, TF binding can create focal changes to chromatin architecture such that nucleosomes are displaced and the surrounding DNA becomes more accessible. Consequently, profiling accessible regions of the genome can help close in to the TF bound regulatory elements. Early genome-wide maps of open chromatin regions in human pancreatic islets used formaldehyde-assisted isolation of regulatory elements followed by sequencing (FAIRE-seq) [51] or DNase I digestion coupled to DNA sequencing (DNase-seq) [172]. By comparing these data to maps from other cell types, these studies identified islet-specific open chromatin regions that coincided with evolutionarily conserved binding sites for key islet transcription factors nearby genes of critical importance in pancreatic islets (e.g. PDX1 and NKX6-1). A more recent open chromatin profiling method, the assay for transposase accessible chromatin followed by sequencing (ATAC-seq) [16], has enabled more routine analysis of scarce samples, such as human pancreatic islets, because of its lower minimum input material requirements (Fig. 1.5A). DNA sequence underlying the highly accessible regions (ATAC-seq peaks) can then be interrogated using the vast TF DNA binding sequence motif (or position weight matrix (PWM)) information databases [83, 118, 158, 77] using specifically designed tools [147] to infer binding sites of these TFs (TF footprint motifs) (Fig. 1.5B, adapted from [92]). Therefore, compared with analysis of histone marks, open chromatin analyses (especially ATAC-seq) have a higher resolution, permitting the identification of specific TF footprint motifs that may be altered by disease risk alleles.

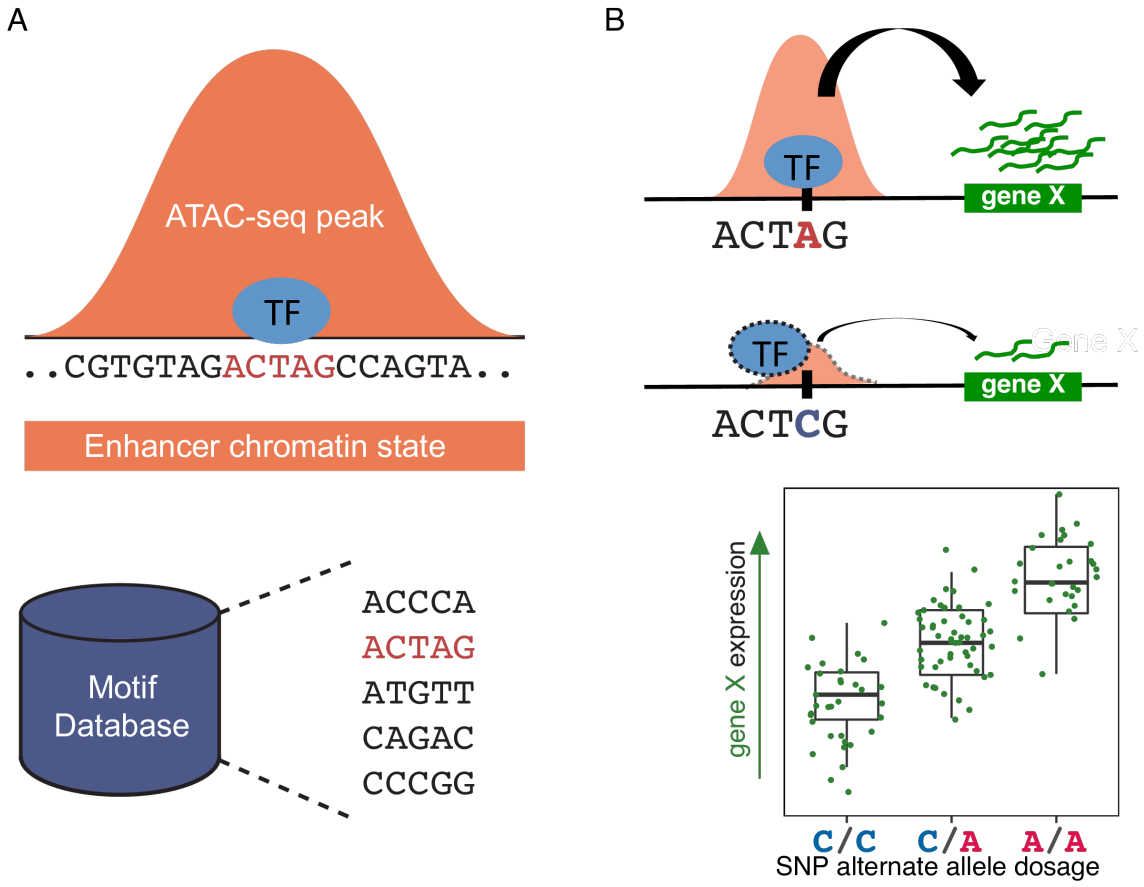


Figure 1.5: Pinpointing individual cis-regulatory elements within broad regulatory regions. A. Open chromatin regions can be identified by assays such as ATAC-seq. TF Motif analysis within open chromatin regions may identify bound by TFs. Searching a TF motifs from available databases in the open chromatin region can nominate TF footprint motif(s) associated with specific TFs. B. eQTL analyses, which use statistical associations between genetic variation and gene expression at a population level, can identify variants that influence expression of downstream target genes, for example, by activating or disrupting transcription factor binding sites. In this example, the blue C allele disrupts an underlying TF footprint motif and is associated with decreased expression of the hypothetical target gene X. Adapted from [92]

1.6 Identifying target genes of regulatory variants

The next step towards attaining a more complete mechanistic insight is identifying the target genes of the regulatory variants. This can be accomplished with eQTL studies, which look at population-level statistical associations between gene expression and genetic variation to assign SNPs to target genes (Fig. 1.5B). Standard eQTL analysis involves a direct association test between markers of genetic variation with gene expression levels typically measured in tens or hundreds of individuals. Several such studies have been conducted across diverse and diabetes-relevant human tissues, such as skeletal muscle [162], adipose [25], liver [58] , islets [44, 193, 196] along with other emerging studies included as a part of this work. Additional layers of regulatory annotation could reveal additional signatures of convergence.

1.7 Thesis outline

In this work, I have analyzed multiple large-scale omics datasets to better understand gene regulatory mechanisms. In chapter 2, I compared gene regulatory annotations defined using diverse epigenomic data across 4 cell types to compare their cell specificities and genetics of gene expression regulation. In chapters 3 and 4, I describe large scale eQTL mapping efforts along with integration of epigenomic data to describe molecular regulatory mechanisms. In chapter 5, I describe profiling and analysis of the enhancer transcriptome in islets, which I then integrate with other available epigenomic data to better understand gene regulatory characteristics. I then delineate the exciting future perspectives that stem from my work and could further contribute to the field of human complex disease genetics.

CHAPTER II

Cell Specificity of Human Regulatory Annotations and Their Genetic Effects on Gene Rxpression

2.1 Abstract

Epigenomic signatures from histone marks and TF binding sites have been used to annotate putative gene regulatory regions. However, a direct comparison of these diverse annotations is missing, and it is unclear how genetic variation within these annotations affects gene expression. Here, we compare five widely-used annotations of active regulatory elements that represent high densities of one or more relevant epigenomic marks: super and typical (non-super) enhancers, stretch enhancers, high-occupancy target (HOT) regions, and broad domains, across the four matched human cell types for which they are available. We observe that stretch and super enhancers cover cell type-specific enhancer chromatin states whereas HOT regions and broad domains comprise more ubiquitous promoter states. eQTL in stretch enhancers have significantly smaller effect sizes compared to those in HOT regions. Strikingly, chromatin accessibility QTL in stretch enhancers have significantly larger effect sizes compared to those in HOT regions. These observations suggest that stretch enhancers could harbor genetically primed chromatin to enable changes in TF binding, possibly to drive cell type-specific response to environmental stimuli. Our results suggest

that current eQTL studies are relatively underpowered or could lack the appropriate environmental context to detect genetic effects in the most cell type-specific regulatory annotations, which likely contributes to infrequent co-localization of eQTL with genome-wide association study (GWAS) signals.

2.2 Introduction

Genome-wide association studies (GWAS) have shown that most of the genetic variants associated with disease related traits lie in non protein-coding regions [68]. More importantly, these loci are specifically enriched in enhancer elements of disease-relevant cell types [182, 120, 189, 142, 29, 144, 153]. This suggests that the majority of disease associated genetic variants modulate regulatory elements that can influence gene expression. Therefore, it is essential to identify and understand the genetic signatures and molecular function(s) of gene regulatory regions.

Epigenomic profiling such as chromatin immunoprecipitation followed by high-throughput sequencing (ChIP-seq) of histone modifications or TF that can indicate regulatory activity in vivo have been effectively used to predict the regulatory function of genomic regions. For example, super enhancers have been defined in multiple cell types as regions with high levels of the histone H3 lysine 27 acetylation (H3K27ac) mark [69]. Putative enhancer elements were identified from ChIP-seq peaks, and elements within 12.5 kb of each other were stitched together. After ranking these stitched regions based on the enhancer associated ChIP-seq signal (Fig. 2.1A), a small number (3%) of identified regions that contained a large fraction (>40%) of the ChIP-seq signal, observable as a steep rise in the ChIP-seq signal curve (geometrical inflection point, Fig. 2.1A) [204], were termed super enhancers. These elements were at least an order of magnitude larger in size than the remaining non-super enhancer elements (i.e., typical enhancers). This signal-based approach has been generalized as

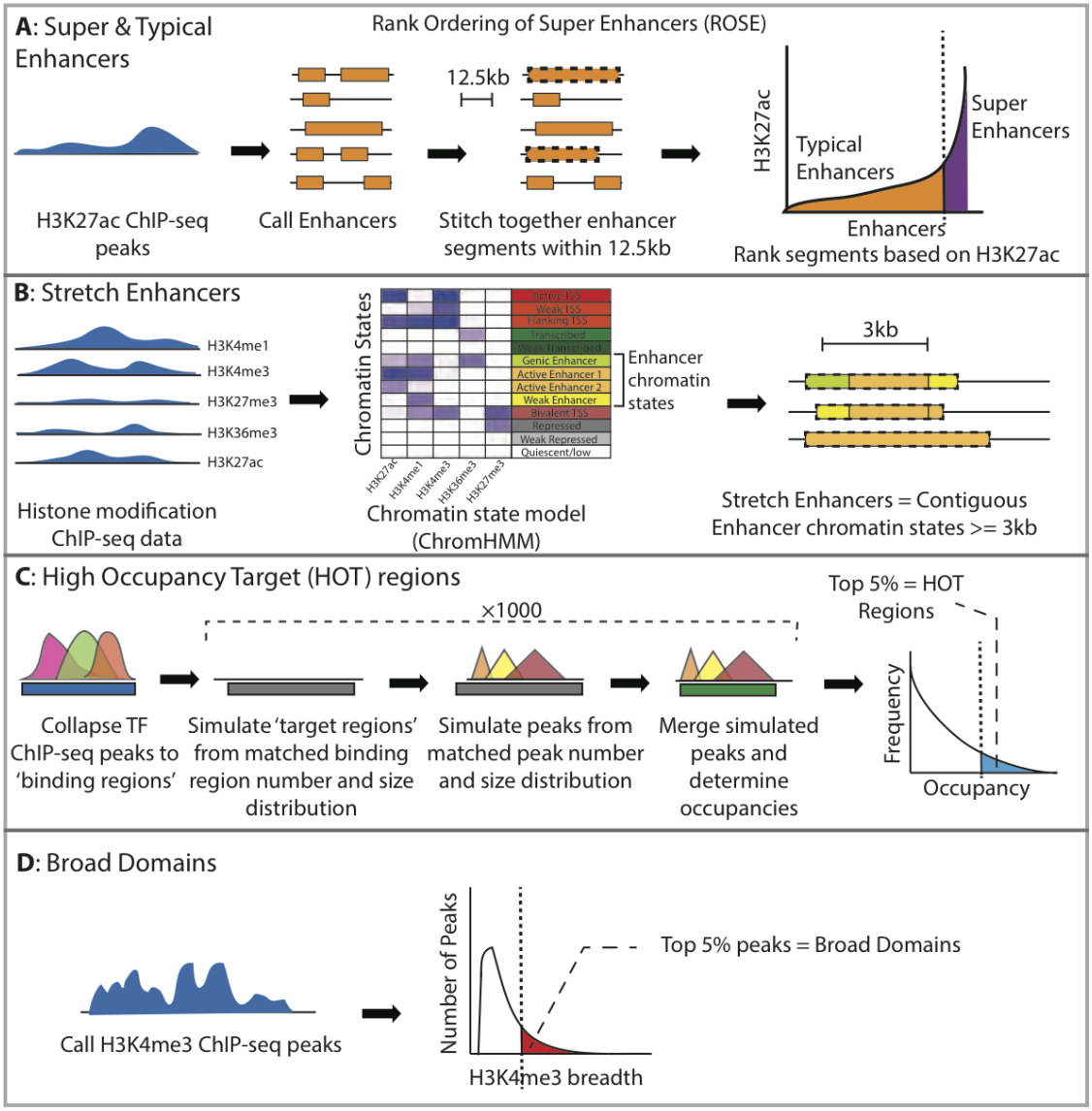


Figure 2.1: Description of the regulatory annotation calling procedures. A: Super/typical enhancers are called by using the H3K27ac mark ChIP-seq to assign enhancer elements, stitching elements within 12.5 kb and ranking the stitched segments based on H3K27ac levels. B: Stretch enhancer calling procedure involves analyzing patterns of multiple histone marks, assigning chromatin state segmentations using ChromHMM, followed by identifying contiguous enhancer chromatin state segments longer than 3 kb. C: HOT regions are defined as regions with higher transcription factor binding occupancies than expected. D: Broad domains are defined as the top 5% of the H3K4me3 ChIP-seq peaks by length.

the rank ordering of super enhancers (ROSE) algorithm [108]; [204] (Fig. 2.1A). Super enhancers are thought to encompass multiple constituent enhancer elements that collectively have high regulatory potential and drive high expression of cell identity regions [204, 69].

In another approach, ChIP-seq data for multiple histone modifications were used to annotate the genome. A HMM based approach identified distinct and recurrent patterns in the ChIP-seq data and segmented the genome into chromatin states [43, 41]. Analyzing chromatin states across diverse cell types and tissues, the authors identified that the longest 10% of contiguous enhancer chromatin states (enhancers $\geq$ 3 kb) were highly cell type-specific, occurred nearby genes with highly cell type-specific gene ontology (GO) terms, and were enriched for cell type relevant disease and trait associated variants [142]. These regions were referred to as stretch enhancers [142] (Fig. 2.1B) and represent substantially large regions of enhancer associated chromatin.

Regulatory annotations have also been defined from TF ChIP-seq profiling. Analysis of such datasets across cell types revealed that more than 50% of TF bound sites occurred in highly occupied clusters that were not randomly distributed across the genome [132, 185, 182, 13]. To identify regions where TF occupancies were higher than expected by chance, one study first collapsed ChIP-seq peaks for multiple TFs as observed binding regions (Fig. 2.1C, blue bar). The expected regions of TF binding or target regions (Fig. 2.1C, gray bars) and individual TF binding sites within these regions (Fig. 2.1C, colored triangles), were then randomly sampled 1000 times, while keeping the number and size distributions equivalent to those observed. Occupancies were scored based on observed and expected collapsed binding sites (Fig. 2.1C, blue and green blocks, respectively); regions with the top 5% occupancies were classified

as high occupancy target (HOT) regions (Fig. 2.1C).

The histone H3 lysine 4 trimethylation (H3K4me3) mark is associated with active and poised promoters [11, 126, 1]. Unusually large regions of the H3K4me3 mark have been observed in multiple cell types across humans, mice and other species, often spanning up to 60 kb [1, 10, 23]. Importantly, the broadest 5% of H3K4me3 domains were found to mark genes with cell type-specific functions [10, 187]. These regions have been termed broad domains (Fig. 2.1D).

These diverse methodologies identify genomic regions with substantially high densities of epigenomic marks known to be associated with gene regulation. These regions denote important classes of regulatory elements, which show cell type-specificity, transcriptional activity in reporter assays, and disease relevance based on GWAS SNP enrichments [91, 142, 69, 10, 13, 102, 12, 31]. Few studies have compared the characteristics for subsets of these annotations, showing some degree of overlap between HOT regions and super enhancers [99] and chromatin interactions between broad domains and super enhancers [187]. However, the functional differences among these annotations, especially how genetic variation in these elements affects target gene expression, are unclear. To fill this gap, we compared diverse characteristics of super-, typical-, stretch-enhancers, HOT regions and broad domains (hereafter collectively referred to as regulatory annotations) in the only four matched human cell types for which they are available: the lymphoblastoid cell line (LCL) GM12878, embryonic stem cell line H1, leukemia cell line K562, and hepatic carcinoma cell line HepG2. We used previously published annotations as these were rigorously generated by respective authors and are widely used. Collectively, these regulatory annotations represent the computational and statistical integration of 245 ChIP-seq data sets (an average of 61 ChIP-seq data sets per cell type). We report annotation summary statistics

and the proportion of overlap with diverse chromatin states in these regions. We measure enrichment for proximity to genes that are expressed in a cell type-specific manner, and integrate genetic regulatory data to measure enrichment for expression quantitative trait loci (eQTL). Finally, as measures of strength of gene and chromatin accessibility regulation, we compare the effect sizes of loci associated with gene expression (eQTL), DNase hypersensitivity (dsQTL), and allelic bias in ATAC-seq data. Comparisons using these metrics allow us to quantify biological properties of these regulatory annotations.

2.3 Results

2.3.1 Genomic distribution, coverage, and overlap of diverse regulatory annotations

To catalogue super, typical, stretch enhancers, HOT regions and broad domains regulatory annotations, we computed the number of distinct segments marked by each annotation, the length distribution of these segments, and the percentage of the genome that is covered by each annotation across the four cell types (Fig. 2.2A-C). Across all cell types, HOT regions comprised the greatest number of segments (Fig. 2.2A). However, they were smaller in size (Fig. 2.2B). Super enhancers comprised the longest segments among all annotations across the studied cell types (Fig. 2.2B), likely due to stitching together H3K27ac peaks that are separated $\leq$ 12.5 kb. All pairwise comparisons between segment lengths for annotations were significant (adjusted $p < 2.2\times10^{-06}$) in each of the cell types from Wilcoxon Rank Sum Test followed by Bonferroni correction, highlighting the differences across annotations. While the percent genome covered by each annotation varied across cell types, these regions consistently covered less than 2% of the genome (Fig. 2.2C).

Next, we calculated the fraction of overlap between all pairs of regulatory annotations. We report the Jaccard statistic (base-pair level intersection/union) for overlap between two annotations (Fig. 2.2D, E). We compare overlaps between different annotations within a cell type (Fig. 2.2D) and between a single annotation (e.g., broad domains) across cell types (Fig. 2.2E). Despite their relatively low genomic coverage (0.5% of the genome), super enhancer segments show considerable overlaps with stretch enhancer segments in the same cell type (Fig. 2.2D), which are significantly enriched (p=0.0001, Fig. 2.3). This is in agreement with both of these annotations representing large domains of active enhancers marked with H3K27ac. HOT regions show extensive overlaps across cell types (Fig. 2.2E), indicating that these regions are less cell type-specific. Broad domains display a similar pattern, though to a less pronounced degree (Fig. 2.2E). Conversely, stretch, super and typical enhancers show low overlaps across cell types, which indicates a higher degree of cell type-specificity (Fig. 2.2E).

2.3.2 Regulatory annotations comprise distinct chromatin states

Most regulatory annotations are defined using histone modification ChIP-seq profiles. However, the differences in their underlying chromatin landscape are unclear. We compared each regulatory annotation with previously reported chromatin state segmentations across all four cell types [196] (Fig. 2.4). Such comparisons are informative because the chromatin states (ChromHMM states) have been generated from an integrative analysis of ChIP-seq data for five diverse histone marks (H3K4me1, H3K4me3, H3K27ac, H3K36me3 and H3K27me3) resulting in 13 chromatin states encompassing active promoter (regions enriched for H3K4me3, H3K27ac marks), enhancer (regions enriched for H3K4me1 and H3K27ac marks), transcribed (regions

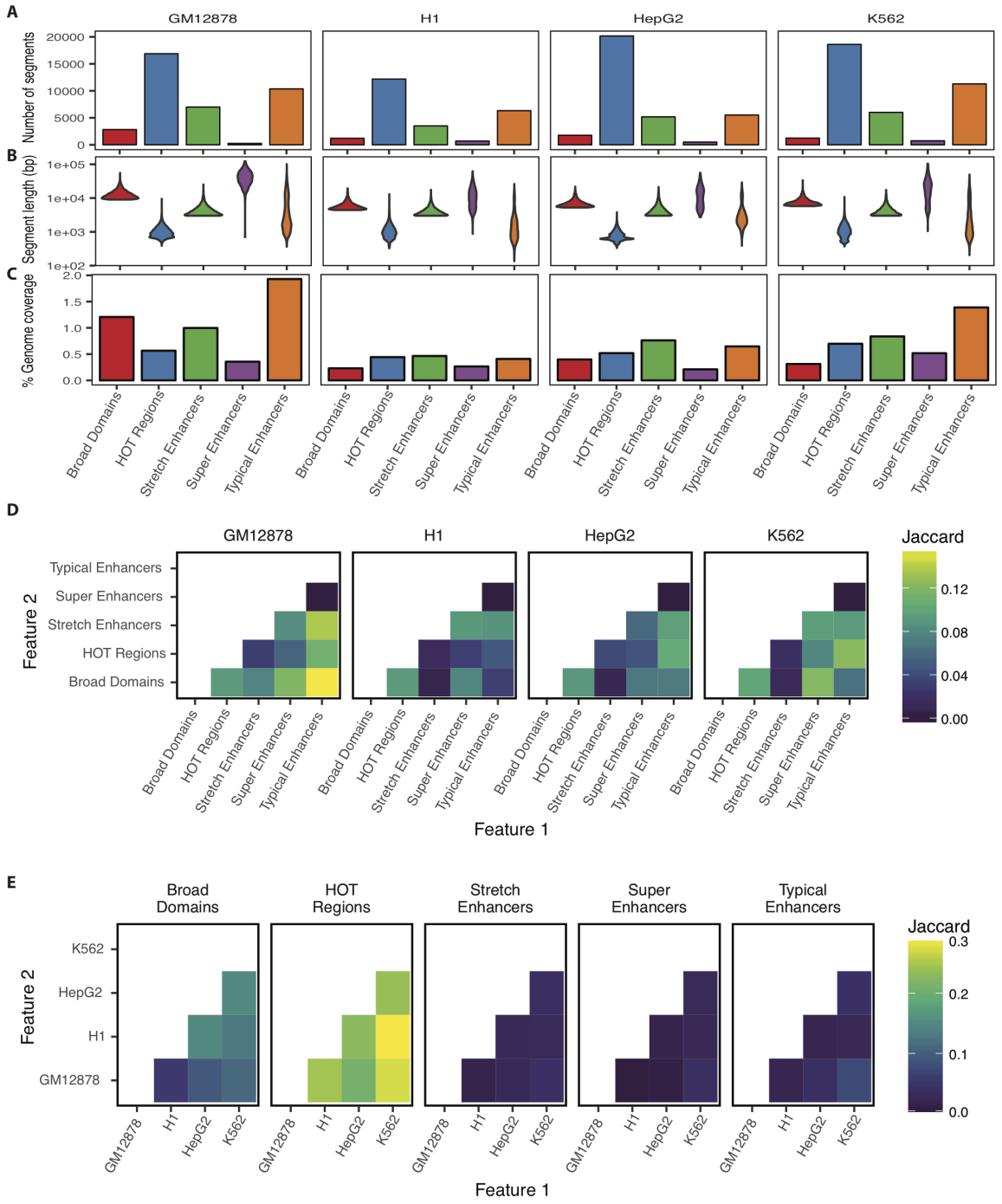


Figure 2.2: Summary statistics and overlaps demonstrate differences in characteristics of regulatory annotations. For each annotation in each cell type considered, shown are number of annotation segments (A), length distribution of segment annotations (B) and percent genomic coverage (C). Jaccard statistic (base-pair level intersection/union) between each pair of annotations is shown within a cell type (D) and across cell types (E).

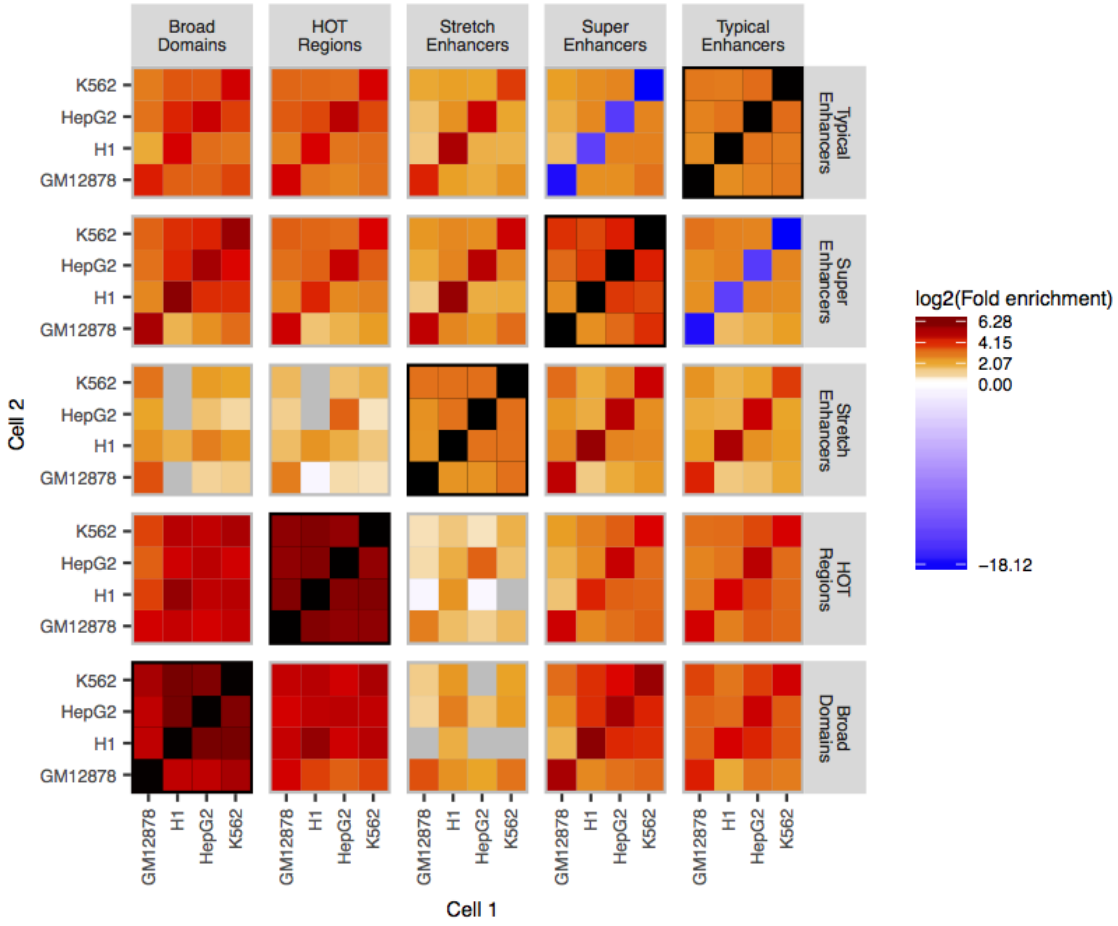


Figure 2.3: Log_2(Fold enrichment) for overlap between each pair of regulatory annotations is shown. Enrichments calculated using GAT [46]. Gray=Not significant after Bonferroni correction. Super and typical enhancers in the same cell type are strongly depleted for overlap since these are disjoint sets. Black tiles on the diagonal represent same cell type and regulatory annotation in the pair.

enriched for H3K36me3), repressed (regions enriched for H3K27me3 marks), and quiescent states (regions lacking marks) [196]. Different enhancer states, such as active enhancer 1 and 2 represent states with different levels of H3K4me1 and H3K27ac mark enrichment and have different genomic coverage [196]. For each regulatory annotation in a particular cell type, we computed the fraction of overlap with chromatin states in the corresponding cell type and across the other three cell types (Fig. 2.4). Generally, HOT regions and broad domains overlap with promoter-related chromatin states consistently across all four cell types, irrespective of which cell type they were called in (Fig. 2.4, facets a1-4, b1-4). In contrast, stretch, super and typical enhancers show a higher fraction of overlap with enhancer-related chromatin states in the corresponding cell type. Notably, stretch/super/typical enhancer regions defined in one cell type constitute mostly non-enhancer chromatin states in other cell types (Fig. 2.4, facets c1-4, d1-4, e1-4), which further reinforces the cell type-specific nature of these annotations.

We then sought to quantify the cell type-specificity of enhancer and promoter chromatin states in each regulatory annotation. For each segment of a regulatory annotation, we computed the ChromHMM posterior probabilities of being called an enhancer or active promoter state averaged over 200bp intervals, denoting chromatin state preference of that segment in each cell of the four cell types. We then computed the information content encoded by these probabilities across cell types (see methods). High information content indicates high specificities of chromatin state. We observe that stretch enhancers constitute high information and high probability enhancer chromatin state (Fig. 2.6A showing GM12878 annotations, Fig. 2.5 showing annotations in all cell types) whereas HOT regions constitute low information and high probability promoter state (Fig. 2.6B showing GM12878 annotations, Fig. 2.7 showing annotations in all cell types). These analyses highlight the differences in the

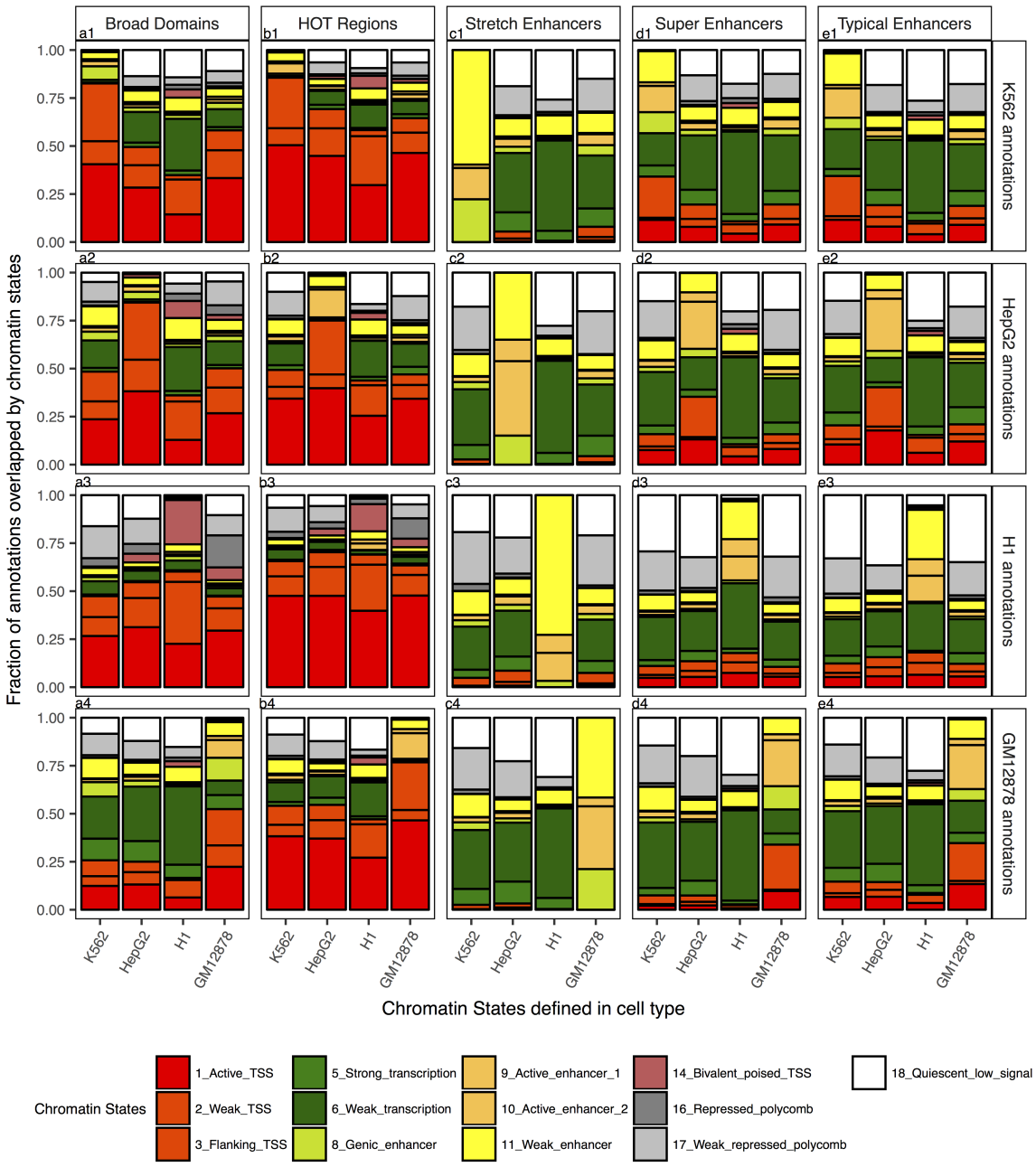


Figure 2.4: Fraction of annotations overlapped by chromatin states. Overlap fractions of each annotation (facet columns) defined in each cell type (facet rows) with chromatin states defined in each cell type (X- axis) is shown. Stretch enhancers were defined using the same chromatin state model for the corresponding cell types.

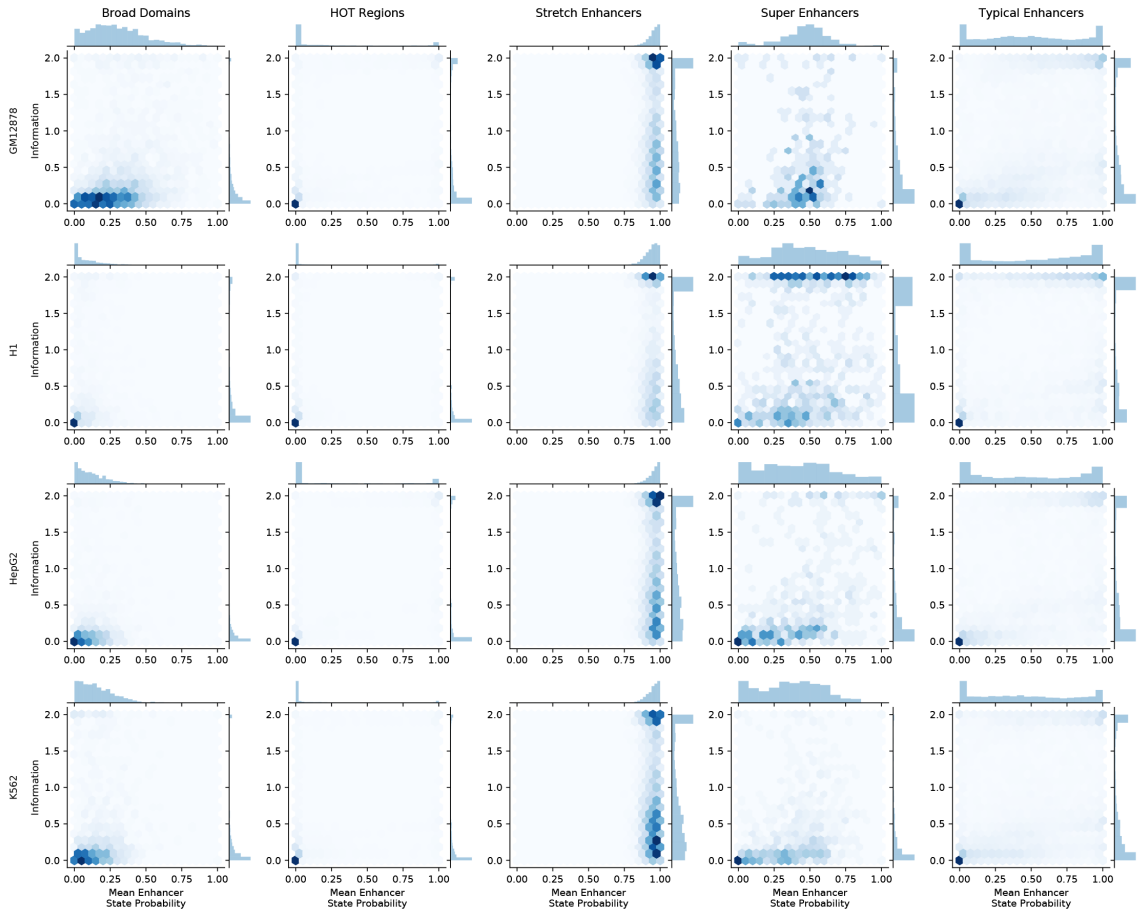


Figure 2.5: Enhancer chromatin state information content for annotations. Average posterior probability for an annotation segment to be called an enhancer chromatin state vs the information content of that feature in the each cell type (facet rows).

underlying chromatin context and cell type-specificities for these annotations.

2.3.3 Regulatory annotations exhibit distinct cell type-specificity of gene regulatory function

Regulatory annotations have been linked to common diseases based on their enrichment to overlap GWAS variants. We directly compared GWAS SNP enrichments for diseases that are relevant to the cell types represented here, such as Crohns disease, rheumatoid arthritis and other autoimmune traits (relevant for lymphoblastoid cell line GM12878), and metabolic traits such as body mass index (BMI) and type 2

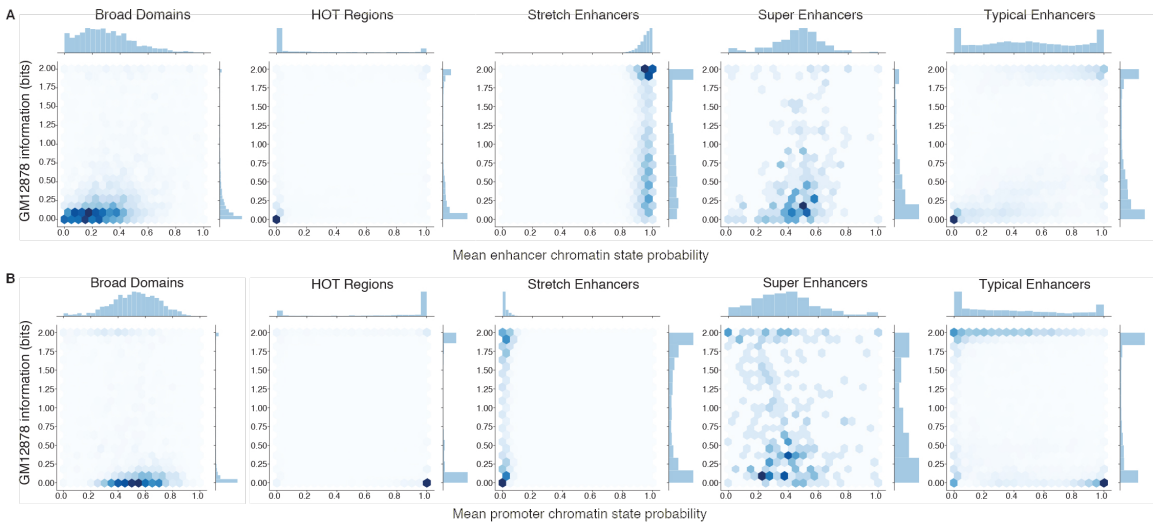


Figure 2.6: Enhancer and promoter chromatin state information content shows cell type-specificity of regulatory annotations. Average posterior probability for an annotation segment to be called an enhancer (A) or promoter (B) chromatin state vs the information content of that feature in the GM12878 cell type calculated by comparing average posterior probabilities across the four cell types.

diabetes (T2D) (relevant for liver hepatocyte cell line HepG2) in each regulatory annotation. Super and stretch enhancers in GM12878 (lymphoblastoid cell line (LCL)) were generally the most enriched for autoimmune related trait GWAS SNPs (Fig. 2.8), whereas stretch enhancers and broad domains in HepG2 were enriched for BMI and T2D GWAS SNPs (Fig. 2.8).

We next assessed the gene regulatory potential for these annotations using several diverse comparisons. We first measured the distance to nearest protein-coding gene from the ends of each annotation segment and found that broad domain and super enhancer segments tend to occur in closer proximity of gene transcription start sites (TSSs) relative to other annotations (Fig. 2.9). Because a regulatory element does not always target the nearest gene, we next utilized cis-expression quantitative trait loci (cis-eQTL), which unambiguously identify target genes by associating genetic variation (SNPs) with gene expression. We asked if regulatory annotations overlapped cis-eQTL which were previously identified in LCLs in the genotype tis-

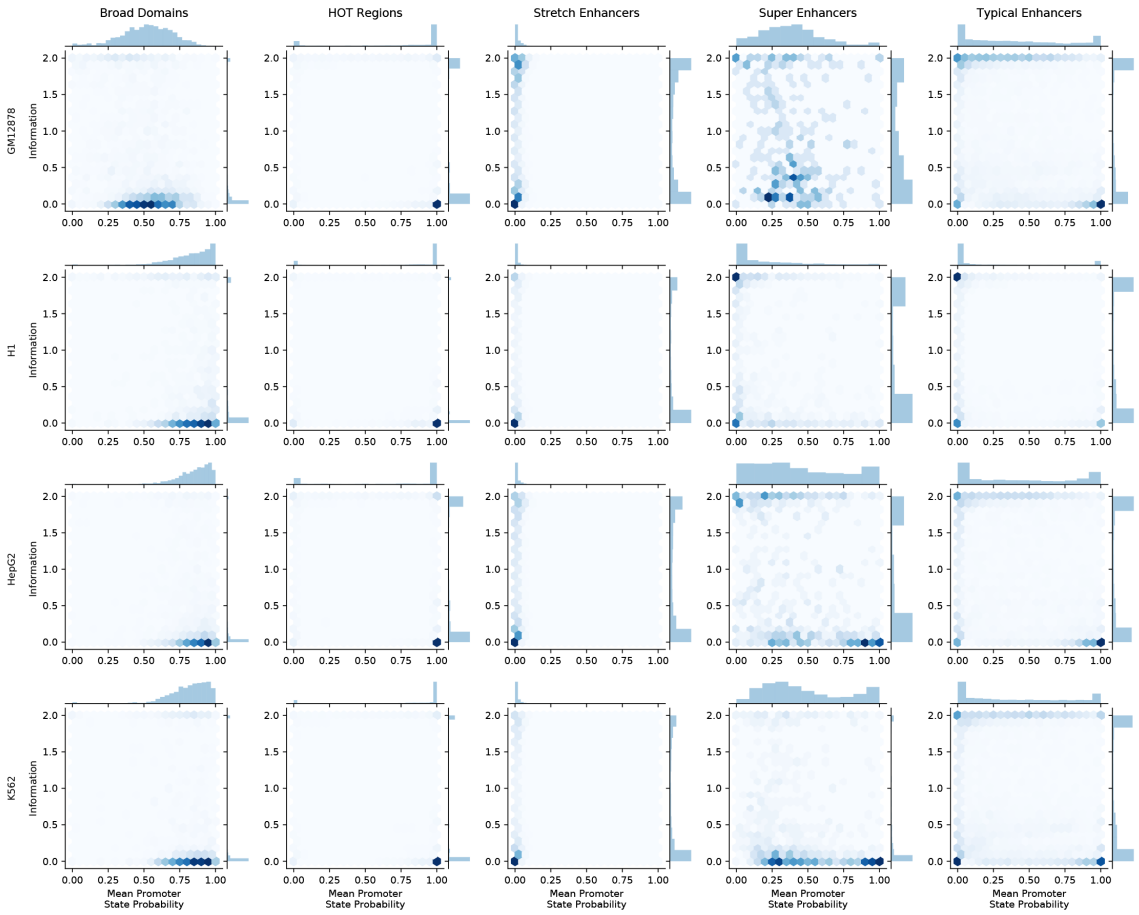


Figure 2.7: Promoter chromatin state information content for annotations. Average posterior probability for an annotation segment to be called an promoter chromatin state vs the information content of that feature in the each cell type (facet rows).

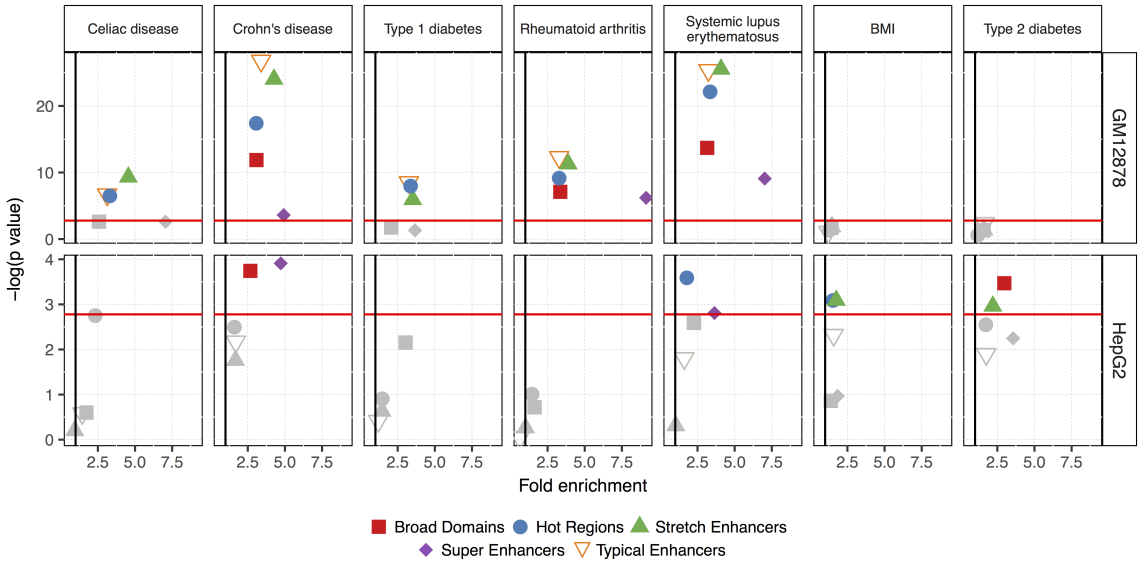


Figure 2.8: Enrichment for annotations in GM12878 and HepG2 to overlap GWAS loci for different traits. Red line = Bonferroni multiple testing correction threshold. Gray = not significant after Bonferroni correction. Annotations overlapping at least 3 GWAS loci for a trait are shown in each panel.

sue expression (GTEx) project [58]. HOT regions in the LCL GM12878 showed the highest enrichment to overlap LCL eQTLs (Fig. 2.10), likely because these represent active promoter regions with high TF binding activity and lie close to protein coding genes (Fig. 2.9). However, HOT regions in control cell types (i.e., non-LCL) were similarly enriched to overlap LCL eQTLs, which highlights the similarity of HOT regions across cell types.

We hypothesized that significant enrichment of LCL eQTLs in regulatory annotations of unrelated cell types is largely driven by eQTLs for more ubiquitously expressed genes. To test this hypothesis, we classified protein-coding genes by their specificity of expression in LCLs using RNA-seq data for 50 diverse tissues from the GTEx project [58] and an information theory approach [161, 64, 162, 196]. We calculated the expression specificities of genes by comparing the relative expression of each gene in LCLs with the entropy of the gene across all 50 tissues in the panel.

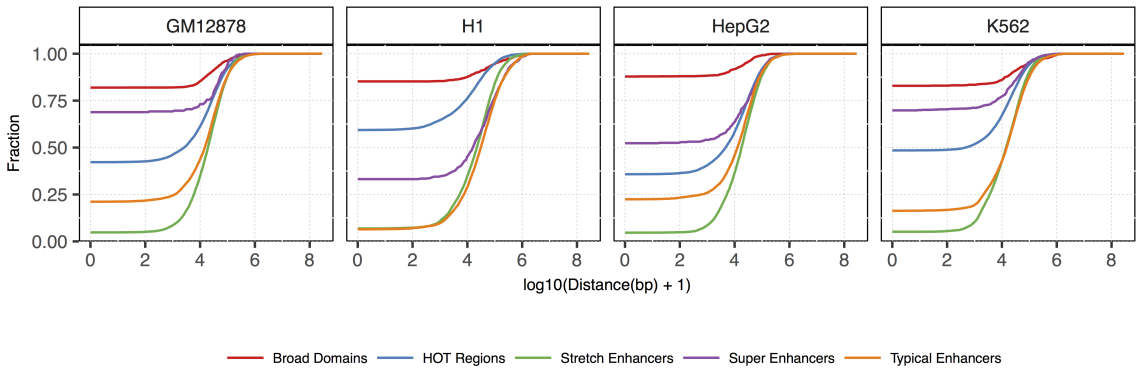


Figure 2.9: Cumulative distribution for distance to nearest TSS (all Gencode V19 protein coding genes) for segments in each regulatory annotation in each cell type.

We defined the LCL expression specificity index (LCL-ESI), which ranges from 0 (i.e., low or ubiquitously expressed genes) to 1 (i.e., highly and specifically expressed genes in LCL). We binned the genes into quintiles based on this LCL-ESI measure; such that bin five represents genes with the highest LCL-ESI scores (Fig. 2.11). We then asked which regulatory annotations occurred closer to cell type-specific genes. We calculated the distance to the nearest TSS for genes in each LCL-ESI bin, which revealed that annotation segments occur closer to genes with higher LCL-ESI (Fig. 2.12, colored lines). To control for the different number of segments in each annotation, we constructed a null expectation by randomly sampling genes from across the five LCL-ESI bins and calculating the distribution of distances to nearest gene TSS (Fig. 2.12, black). We then normalized the observed distance distribution for each LCL-ESI bin gene set with that from the null set and used this as a controlled measure of TSS proximity enrichment (Fig. 2.13A). We observed that all regulatory annotations are depleted from occurring close to non-specific genes (LCL-ESI bin 1) and enriched to occur closer to highly specific genes (LCL-ESI bin 5). Notably, super, stretch, typical enhancers and broad domains were more enriched to occur near the most cell type-specific genes than HOT regions (Fig. 2.13A). As expected, enrich-

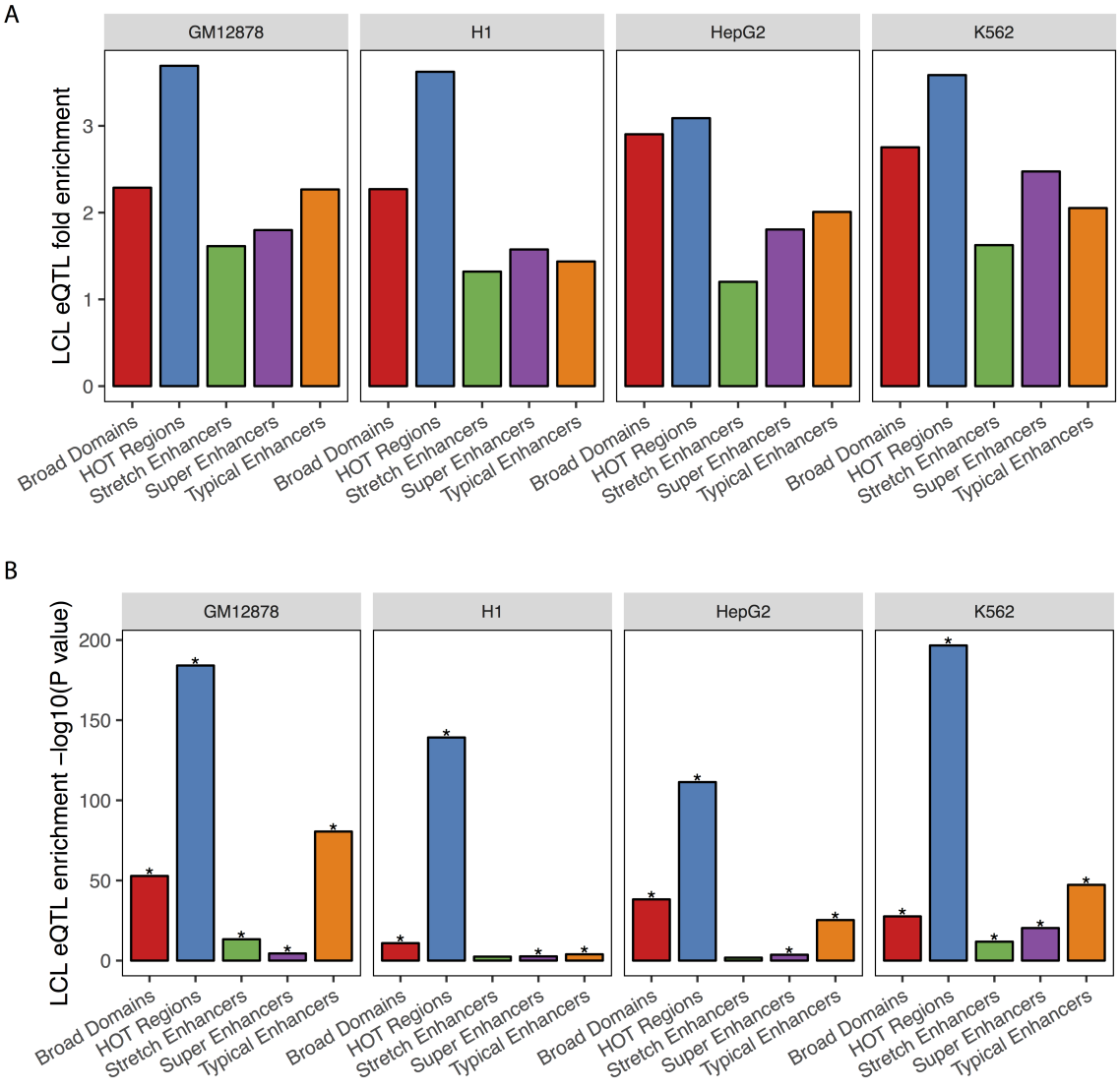


Figure 2.10: Enrichment of regulatory annotations in four cell types to overlap with LCL eQTL (GTEx v7). Fold enrichments are shown in A, -$\log_{10}$(p values) are shown in B. Enrichment p values significant after a Bonferroni correction for 20 tests are marked with *.

ments for all annotations to occur within larger distances to TSS (order of mega bases) converge to 1 (Fig. 2.13A), indicating a properly controlled proximity enrichment test.

We next asked which regulatory annotations were more enriched to overlap eQTL of more cell type-specific genes. We obtained sets of LCL eQTL [58] for genes in each LCL-ESI bin and calculated the enrichment of each eQTL set in the regulatory annotations. Indeed, we observe that GM12878 regulatory annotations were increasingly enriched to overlap eQTLs for highly LCL specific genes (Fig. 2.14) and the fold enrichment for eQTLs in a bin is positively correlated with the LCL-ESI bin number (Fig. 2.13B, GM12878 facet). Notably, stretch enhancers, and in some instances typical enhancers, in non-LCL cell types showed strong negative correlations of LCL eQTL fold enrichment with LCL-ESI bin number (Fig. 2.13B), indicating higher cell type-specificity for stretch enhancers. This is consistent with the previous histone modification based chromatin state analyses (Fig. 2.6, Figs. 2.4, 2.5, 2.7, 2.8), which also highlight the cell type-specificity of stretch enhancers. HOT regions in non-LCL cell types show high enrichments for eQTLs in less cell type-specific LCL-ESI bins 1-3 (Fig. 2.14). This analysis shows that high enrichments of LCL eQTLs in non-LCL annotations (Fig. 2.10) was driven by eQTLs for more ubiquitously expressed genes. These analyses further emphasize the differences in the cell type-specificities of these regulatory annotations.

2.3.4 Patterns of expression and chromatin QTL effect sizes in annotations suggest regulatory buffering

While enriched overlap with eQTLs demonstrates genetic regulatory potential for each annotation (Figs. 2.10, 2.14, 2.13B), this analysis does not distinguish the strength of these genetic effects on gene expression. To understand this, we com-

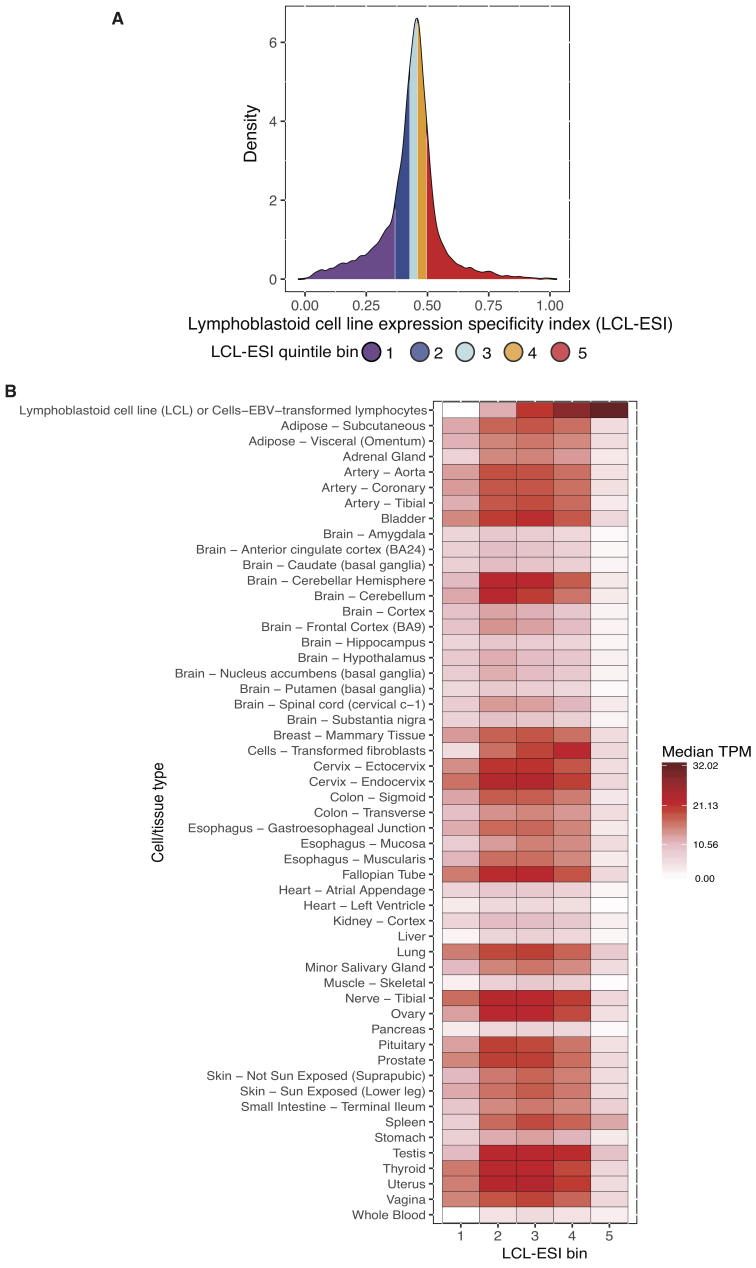


Figure 2.11: Gene expression specificity index in lymphoblastoid cell line (LCL-ESI). A: Distribution of LCL-ESI for protein coding genes with median transcripts per million (TPM) >= 0.15 in LCL. Colors indicate equal sized binning of the genes into quintiles by LCL-ESI. Each bin contained 2753 protein coding genes. B: Median TPM for genes in each LCL-ESI quintile bin across the 50 GTEx tissues analyzed. Lymphoblastoid cell line (LCL) is named as Cells-EBV-transformed lymphocytes in the GTEx dataset.

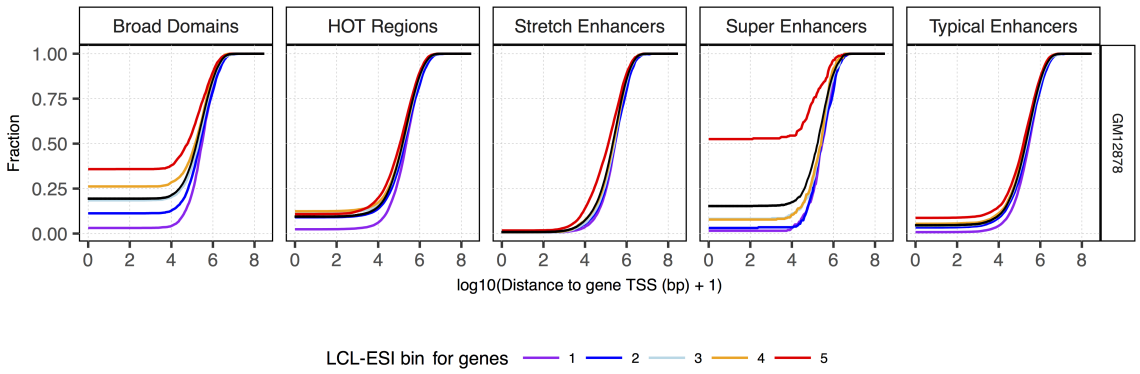


Figure 2.12: Cumulative distribution for distance to nearest TSS (Gencode V19 protein coding genes binned by LCL-ESI, 2753 genes in each bin) for regulatory annotations in GM12878. Black curves represent 10,000 random sub-samplings of 2753 genes from across the five bins.

pared the absolute effect sizes (beta values from the linear regression models) of LCL eQTLs overlapping different GM12878 regulatory annotations. We excluded SNPs with minor allele frequency (MAF) < 0.2, since these SNPs have substantially reduced statistical power and are therefore biased to be detected as eQTL only with higher effect sizes (Fig. 2.15). We observed that LCL eQTLs in GM12878 stretch enhancers have nominally significantly lower (p=0.032) effect sizes than GM12878 HOT regions, however this comparison does not survive a Bonferroni correction accounting for 10 pairwise tests (Fig. 2.16A). To achieve higher power for such an analysis, we utilized the larger GTEx blood eQTL dataset and compared effect sizes in annotations of the blood relevant leukemia cell line K562. Consistent with the LCL analysis, we observed that effect sizes of blood eQTL in K562 stretch enhancers were significantly lower than that of HOT regions (Bonferroni corrected $p = 0.0082$, Fig. 2.17A). We note that the differences in effect sizes for LCL and blood eQTL are largely due to different sample sizes and therefore power to detect eQTL. To further control for potential sources of bias in this analysis, we next asked if this effect size difference was driven by distance to the eQTL target genes TSS or the number of SNPs in high

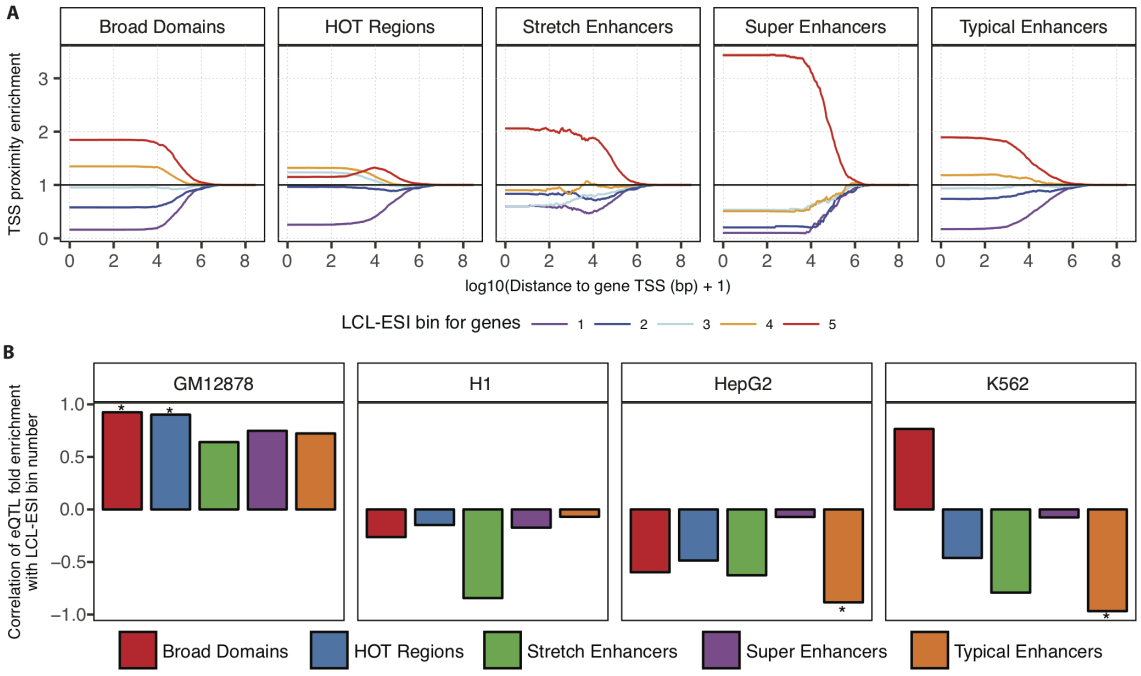


Figure 2.13: Proximity to protein coding genes and enrichment for eQTL highlight functions of regulatory annotations. A: Enrichment for regulatory annotation elements in GM12878 to lie within distances (x-axis) of transcription start site (TSS) of protein coding genes binned by gene expression specificity in lymphoblast cell lines (LCL-ESI). Enrichment calculated in comparison to 10,000 random samplings, 95% confidence intervals shown. B: Pearson correlation of LCL-ESI gene quintile bin numbers (increasing LCL specificity) with the fold enrichment of eQTLs of these genes in regulatory annotations. Positive correlation shows that the eQTLs for more LCL specific genes are more enriched in annotations. Significant ($p < 0.05$) correlations are marked with a *.

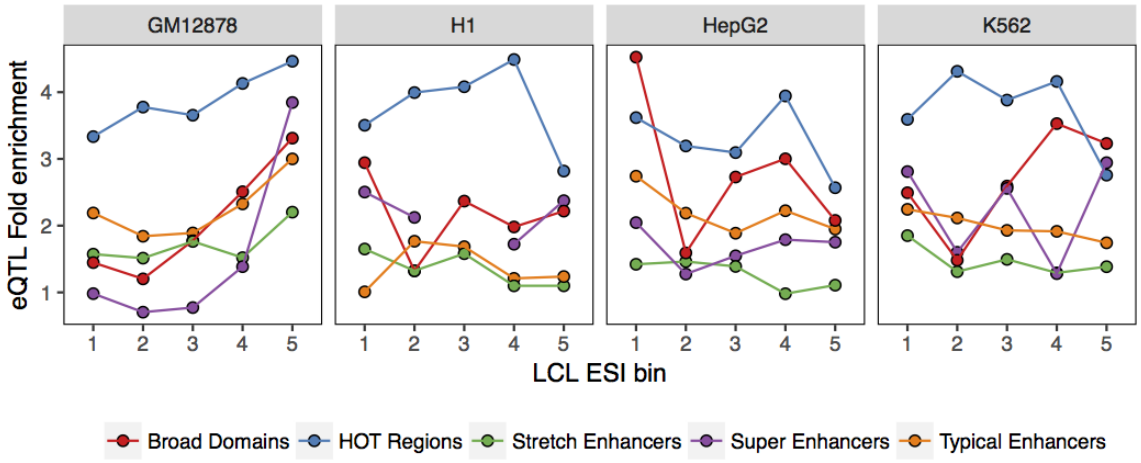


Figure 2.14: Enrichment for regulatory annotations to overlap LCL eQTL (GTEx v7, 10% FDR) binned by LCL-ESI or the eQTL eGene.

LD with the index eQTL SNP. We modeled the eQTL absolute effect size using linear regression including these additional two covariates along with an indicator variable encoding stretch enhancer or HOT region annotation (eQTL overlapping both annotations were not considered). We observed a significant effect on the indicator variable ($p = 0.005$, regression coefficient = -0.0521, Table 2.1), which confirms the smaller effect size of eQTL in stretch enhancers, independent of TSS distance and LD structure.

Differences in effect sizes of eQTLs in stretch enhancers compared to HOT regions directly translates to differences in the statistical power to detect eQTL residing in these regulatory annotations, which have remarkably distinct cell type-specificities. To quantify this, we performed a power calculation for the 10th through 90th percentiles of the eQTL effect size distribution observed in each annotation, keeping other parameters such as sample size, MAF, type 1 error rate, number of tests and the standard deviation of the error term constant. We show that variants in stretch enhancers have nearly uniform lower power to be detected as eQTL across the effect size distribution (Fig. 2.17B, S12B). Indeed, stretch enhancers showed lower enrich-

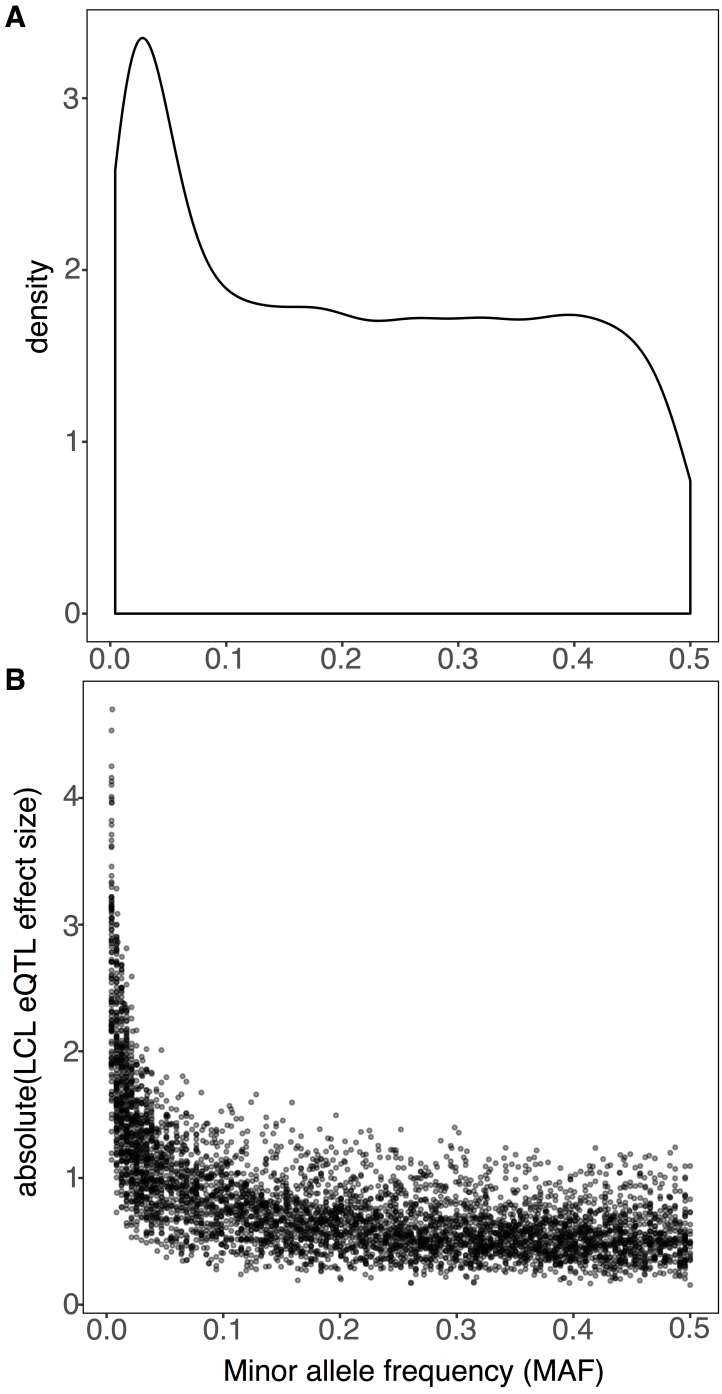


Figure 2.15: Lower minor allele frequency (MAF) variants have higher eQTL effect sizes. A: Distribution of MAF for LCL eQTL (GTEx v7, 10% FDR). B: LCL eQTL absolute effect size (slope of the linear regression) vs minor allele frequency (MAF).

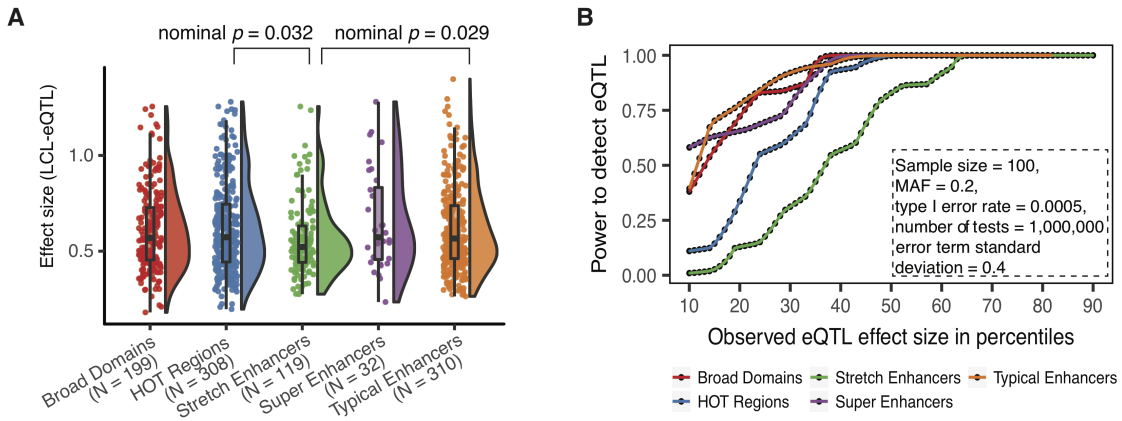


Figure 2.16: Gene expression and chromatin QTL effect size differences in regulatory annotations suggest regulatory buffering. A: Distribution of eQTL effect sizes for LCL eQTL (GTEx v7, 10% FDR) in GM12878 regulatory annotations are shown. Nominal P values < 0.05 are shown. B: Power to detect eQTL after Bonferroni correction at effect sizes corresponding the 10th through 90th percentiles observed for each annotation (shown in A). Other constant parameters for the power calculation are shown in box.

OLS Regression Results					
	absolute eQTL				
Dependent Variable:	effect size	R-squared:	0.027	Prob (F-statistic):	0.000172
Model:	OLS	Adj. R-squared:	0.023	Log-Likelihood:	153.18
Method:	Least Squares	F-statistic:	6.748		
No. Observations:	723	AIC:	-298.4		
Df Residuals:	719	BIC:	-280		
Df Model:	3				

	coef	standard error	t	P>|t|	[0.025	0.975]
Intercept	0.3211	0.009	34.394	0	0.303	0.339
absolute eQTL distance from eGene TSS	-0.1638	0.072	-2.281	0.023	-0.305	-0.023
Regulatory annotation binary variable HOT regions = 0; Stretch enhancer = 1	-0.0521	0.019	-2.795	0.005	-0.089	-0.015
Number of SNPs in LD r2>0.99	0.114	0.056	2.021	0.044	0.003	0.225

Omnibus:	210.562	Durbin-Watson:	1.999
Prob(Omnibus):	0	Jarque-Bera (JB):	483.62
Skew:	1.563	Prob(JB):	9.62E-106
Kurtosis:	5.505	Cond. No.	10.2

Table 2.1: Ordinary least squares regression results modeling blood eQTL absolute effect size dependent on K562 HOT regions or stretch enhancer annotation, distance of the eQTL to eGene TSS and number of SNPs in LD r2>0.99

ment to overlap eQTLs than HOT regions (Fig. 2.10). Therefore, identifying eQTLs in cell type-specific stretch enhancers will require larger sample sizes.

Among other mechanisms, eQTL SNPs can influence gene expression in vivo by modulating TF binding. TFs can either bind in nucleosome-depleted regions or bind and displace nucleosomes (pioneer factors) [57, 200, 17]. Therefore, QTL analysis of chromatin accessibility using DNase I hypersensitivity (dsQTL) can assess variant effects on regulatory element activity. Interestingly, we found that LCL dsQTLs [32] in stretch enhancers have significantly higher effect sizes than those in HOT regions (Bonferroni corrected $p=6.2\times10^{-08}$, Fig. 2.17C), which is the opposite of what we observed for eQTL effects (Fig. 2.17A). dsQTL in super enhancers and typical enhancers also have higher effect sizes than those in HOT regions (Bonferroni adjusted $p = 0.013$, 2.2×10^{-05} respectively). To examine the effect of genetic variation on open chromatin at the resolution of an individual sample, we quantified allelic bias in the assay for transposase accessible chromatin followed by sequencing (ATAC-seq) data available in GM12878 [17]. Allelic bias measured by quantifying the ATAC-seq signal over each of the two alleles at a heterozygous site is an indicator of allelic differences in chromatin accessibility at a specific locus. To control for different power to detect allelic bias, we uniformly down-sampled all SNPs to 30x coverage. We included all SNPs from the full range of MAFs with nominally significant allelic bias ($p < 0.05$) since the SNP MAF does not affect the power to detect allelic bias in an individual sample. Consistent with the dsQTL results, we observed that SNPs in stretch enhancers show a significantly larger allelic bias effect size (see Methods) compared to HOT regions (Bonferroni corrected $p = 0.0051$, Fig. 2.17D). This trend remains after removing SNPs with MAF < 0.2, similar to the dsQTL analyses above (Fig. 2.18), indicating that SNP MAF does not confound this analysis. No other pairwise tests were significant. Collectively, these observations show that stretch

enhancers harbor variants that have strong genetic effects on chromatin changes but these are buffered at the level of transcription.

2.4 Discussion

We performed a comparative analysis of five regulatory annotations, all based on diverse epigenomic signatures, to better understand their regulatory capacity and downstream transcriptional effects. We observed that stretch, super and typical enhancers overlap enhancer chromatin states in the corresponding cell type, but overlap non-enhancer chromatin states in unrelated cell types, supporting the cell type-specificity of these regulatory elements. These observations highlight H3K27ac as a good proxy for cell type-specific regulatory function. Annotations based on the H3K4me3 mark (Broad domains) and TF binding (HOT regions) show a large fraction (>40%) of overlaps with promoter chromatin states across different cell types. Consistent with our observations, a recent study in the fly reported that regions bound by large numbers of TFs (such as HOT regions) are less cell type-specific [88]. While the diverse ChIP-seq data used to define regulatory annotations comes from different individuals, we note that future studies using ChIP-seq data from the same individual might have even higher power to detect cell type-specific differences.

Analysis of genetic effects on gene regulatory function of annotations revealed that blood eQTLs in K562 stretch enhancers have significantly lower effect sizes compared to HOT regions. Stretch/super enhancers are known to regulate more cell type-specific genes for which the expression levels may be tightly controlled under basal conditions. Multiple studies have observed redundancy in gene regulation by individual components of super enhancers [63, 166, 133, 207]. Such studies then contested the notion of super/stretch enhancers as a distinct entity, arguing that these annotations are no different than other enhancers. However, here we offer an alternative expla-

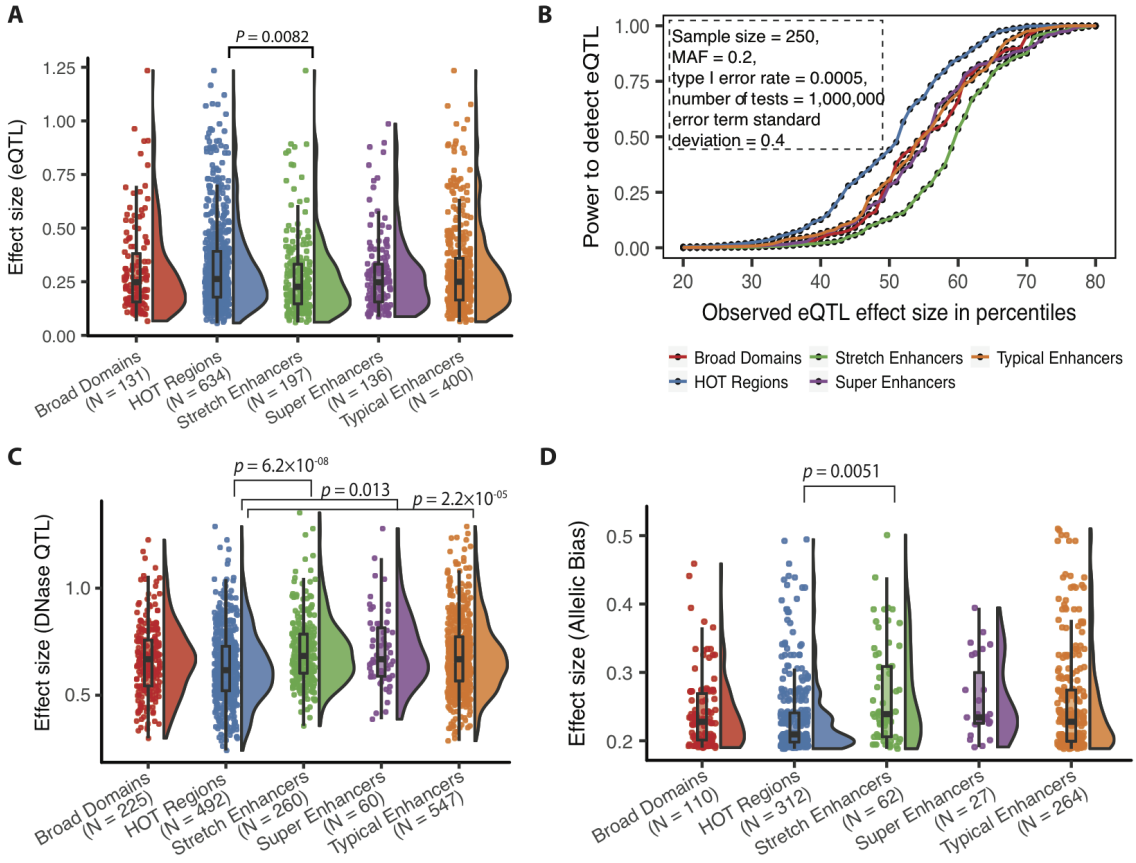


Figure 2.17: Gene expression and chromatin QTL effect size differences in regulatory annotations suggest regulatory buffering. A: Distribution of eQTL effect sizes for blood eQTL (GTEx v7, 10% FDR) in K562 regulatory annotations. B: Power to detect eQTL after Bonferroni correction at effect sizes corresponding the 10th through 90th observed for each annotation (shown in A). Other constant parameters for the power calculation are shown in box. C: Distribution of effect sizes for LCL DNase QTLs in GM12878 regulatory annotations. D: Distribution of effect sizes (deviation from expectation) for SNPs with significant allelic bias in GM12878 ATAC-seq ($p < 0.05$, minimum coverage at SNP=30, reads down-sampled to 30, see methods) in GM12878 regulatory annotations. P values from Wilcoxon rank sum tests, after a Bonferroni correction accounting for 10 pairwise tests. Number of QTLs/allelic biased SNPs overlapping each regulatory annotation is shown in parentheses in A, C and D.

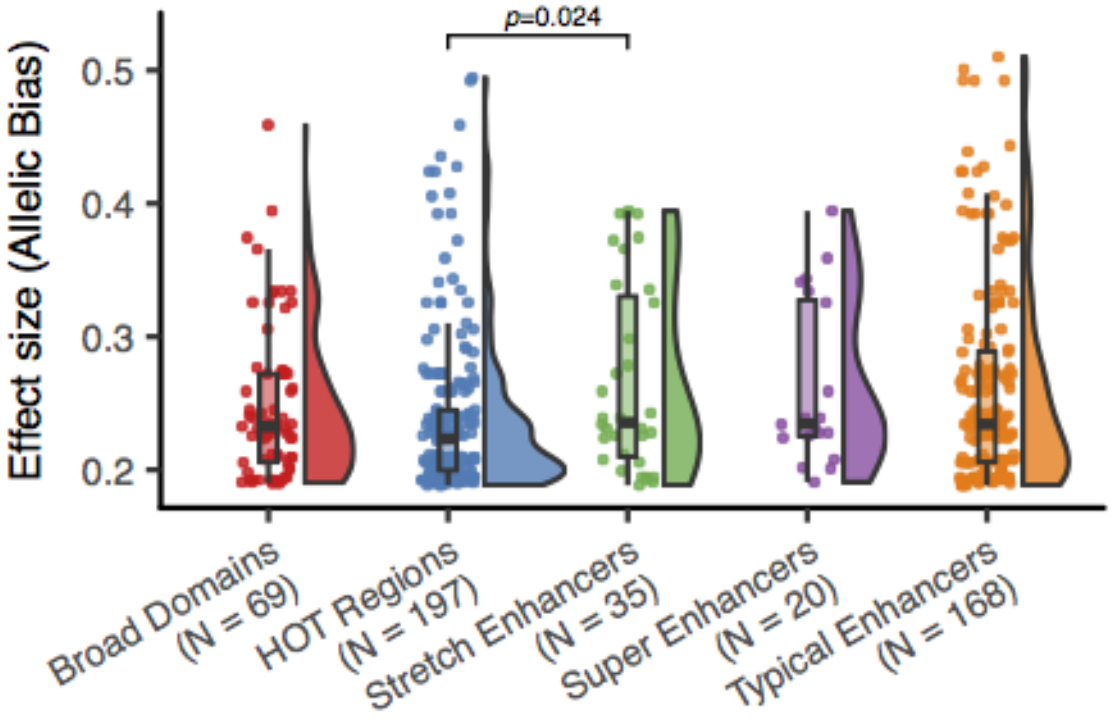


Figure 2.18: Effect sizes for Allelic Bias in GM12878 ATAC-seq after removing low MAF SNPs (consistent with eQTL and dsQTL effect size analyses). SNPs with MAF > 0.2 and allelic bias p value < 0.05 were included for this analysis.

nation - that enhancer buffering which results from functional redundancy could be a mechanism for tighter control of gene expression under basal conditions and would explain the low observed eQTL effect sizes. These regions could encode regulatory plasticity, allowing critical genes to respond to multiple (patho)physiologic stimuli. This would lead to smaller effects in the steady state, whereas each component could contribute to tight but pliable regulation by different signaling pathways. Therefore, the outcome of perturbing enhancer components might be different in response to different environmental stimuli and existing studies that probe basal conditions would not detect such effects.

In contrast, genetic variants associated with open chromatin in stretch enhancers show significantly higher effects than those in HOT regions, both within a single sample (allelic bias in ATAC-seq) and across multiple samples (dsQTL). Our results present an apparent discrepancy in that genetic variants in stretch enhancers display higher chromatin QTL effect sizes and slightly but significantly lower basal expression QTL effect sizes when compared to HOT regions. It is possible that the large constellation of TFs bound in HOT regions [182, 88] maintain more constitutively open chromatin, which would be less susceptible to effects of individual genetic variants. This concept of buffering has been demonstrated previously where a smaller fraction of SNPs in strong DNase peaks showed significant allelic bias compared to those in weak DNase peaks [119]. We reason that chromatin accessibility, which influences TF binding could be a molecular feature of the initial response cascade to propagate gene expression changes under stimulatory conditions. We hypothesize that the larger genetic effects on stretch enhancer chromatin accessibility will propagate to gene expression effects under specific environmental conditions. Under this hypothesis, we expect that many dsQTL will be associated with gene expression under specific stimuli (or response-specific eQTL) rather than steady state (basal eQTL). In support of

this, a recent study in the macrophage model system [3] showed that 60% of eQTLs that manifest upon stimulation are chromatin QTL in the basal state. Unfortunately, currently available response expression or chromatin QTL datasets are underpowered for a comparison of effect sizes in the regulatory annotations analyzed here owing to low sample sizes.

Our observations could help reconcile why many cis-eQTLs are shared across cell types and infrequently co-localize with GWAS signals [103, 74, 58]. We have shown that while stretch enhancers are enriched to overlap GWAS loci for cell type-relevant traits, variants in these regions are underpowered to be identified as eQTL. Current eQTL studies are biased to identify eQTLs for more broadly expressed genes. Our results suggest that larger sample sizes will be needed to identify cell type-specific eQTLs. Additionally, our results suggest the need to perform response eQTL studies under carefully selected environmental conditions.

2.5 Materials and Methods

2.5.1 Regulatory annotation sources

Regulatory annotations for GM12878, H1 hESCs, HepG2 cell types were downloaded from previously published studies for HOT regions [13], Broad domains [10], Stretch Enhancers [196], Super and Typical Enhancers [69].

2.5.2 Summary statistics and overlaps between annotations, chromatin states and ATAC-seq peaks

Summary statistics such as the number of features in each annotation, segment size distribution and percent genome coverage (Fig. 2.2A-C) were calculated using custom scripts (see GitHub). To compute overlap fractions between all pairs of an-

notations shown in Fig. 2.2D,E, we calculated the base pair level overlap between each pair using BEDtools intersect [154]. For each pair of annotation sets, we then calculated the Jaccard statistic by dividing the total length of the intersection region with the total length of the union region. To calculate the fraction of regulatory annotation overlap with chromatin states in Fig. 2.4, we used chromatin states previously defined in the four cell types considered [196] and used BEDtools intersect. Stretch enhancer annotations were also obtained from this previous study [196].

Enrichment for overlap between each pair of regulatory annotations in Fig. 2.3 was calculated using the Genomic Association Tester (GAT) tool [65]. To ask if two sets of regulatory annotations overlap more than that expected by chance, GAT randomly samples segments of one regulatory annotation set from the genomic workspace (hg19 chromosomes) and computes the expected overlaps with the second regulatory annotation set. We used 10,000 GAT samplings for each regulatory annotation. The observed overlap between segments and annotation is divided by the expected overlap and an empirical p-value is obtained.

2.5.3 Chromatin state information content analysis

We first compiled the average posterior probabilities of a regulatory annotation segment to be called an enhancer or promoter chromatin state. We utilized the previously published 13-chromatin state ChromHMM model (also used to define stretch enhancers) [196], which also outputs posterior probabilities for each 200bp genomic segment to be called each of the 13 states in each of the four cell types. We considered the sum of Active enhancer 1 and 2, weak enhancer and genic enhancer posterior probabilities to represent enhancer states, and averaged these values over all the 200bp tiles overlapping each annotation segment. We considered Active TSS, Weak TSS and Flanking TSS states to denote promoter chromatin states. For example, for a

segment in GM12878 broad domains, we obtained the average posterior probabilities for the region being an enhancer or promoter state in a cell $x_{segment,cell}$ for cell $\in$ GM12878, H1, HepG2, and K562. To calculate the information content, we first calculated the relative average posterior probabilities, $p_{segment,cell}$

$$p_{segment,cell} = \frac{x_{segment,cell}}{\sum_{cell=1}^{4} x_{segment,cell}}$$

Next, we calculated entropy of the segment as:

$$Entropy_{segment} = -\sum_{cell=1}^{4} p_{segment,cell} \times log_2(p_{segment,cell})$$

We know that entropy is maximized with all segments have equal relative probabilities, or $p_{segment,cell} = \frac{1}{4}$ for cell $\in$ GM12878, H1, HepG2, and K562

$$Max.Entropy_{segment} = -\sum_{cell=1}^{4} \frac{1}{2} \times log_2(\frac{1}{4}) = 2$$

$$Information\,content_{segment,cell} = p_{segment,cell} \times (Max.Entropy_{segment} - Entropy_{segment})$$

We then compared $x_{segment,cell}$ with $Informationcontent_{segment,cell}$.

While high posterior probabilities for enhancer or promoter states indicate preference for that state, high information content indicates cell type-specificity of that chromatin state preference. For plotting Fig. 2.6, to have the same x axes ranges for all facets for easier comparison (stretch enhancers only show high mean posterior probabilities for enhancer states and low posterior probabilities for promoter states due to their definition), we added one pseudo-count in each corner for all facets.

2.5.4 Distance to nearest gene

We downloaded the Gencode V19 gene annotations from `ftp://ftp.sanger.ac.uk/pub/gencode/Gencode_human/release_19/gencode.v19.annotation.gtf.gz` and obtained the transcription start site (TSS) coordinates for protein coding genes. For each segment in each annotation, we computed the distance to nearest protein coding gene TSS using BEDtools closest [154].

2.5.5 Enrichment of genetic variants in genomic features

Enrichment for genome wide association study (GWAS) variants for different traits and expression quantitative trait loci (eQTL) identified in the lymphoblastoid cell line (LCL) in regulatory annotations was calculated using GREGOR (version 1.2.1) [160]. Since the causal SNP(s) for the traits are not known, GREGOR allows considering the input lead SNP along with SNPs in high linkage disequilibrium (LD) (based on the provided R2THRESHOLD parameter) while computing overlaps with genomic features (regulatory annotations). Therefore, as input to GREGOR, we supplied SNPs that were not in high linkage disequilibrium with each other. We pruned the list of SNPs using the PLINK (v1.9) tool [152]; [22] clump option and 1000 genomes phase 3 vcf files (downloaded from `ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/release/20130502`) as reference. For each input SNP, GREGOR selects 500 control SNPs that match the input SNP for minor allele frequency (MAF), distance to the nearest gene, and number of SNPs in LD. Fold enrichment is calculated as the number of loci at which an input SNP (either lead SNP or SNP in high LD) overlaps the feature over the mean number of loci at which the matched control SNPs (or SNPs in high LD) overlap the same features. This process accounts for the length of the features, as longer features will have more overlap by chance with control SNP sets.

Specific parameters for the GWAS enrichment were: GWAS variants for Rheuma-

toid Arthritis, type 1 diabetes (T1D) and type 2 diabetes (T2D) were obtained from the NHGRI-EBI catalog (`https://www.ebi.ac.uk/gwas/`). We used the following parameters - Pruning to remove SNPs with r2 > 0.2 for European population; GREGOR: r2 threshold = 0.8. LD window size = 1Mb; minimum neighbor number = 500, population = European.

Specific parameters for the LCL eQTL enrichment were: LCL eQTL data from the genotype tissue expression (GTEx V7) study was downloaded from the GTEx website `https://www.gtexportal.org/home/datasets`. We used the following parameters Pruning to remove SNPs with r2 > 0.8 for European population; GREGOR: r2 threshold = 0.99. LD window size = 1Mb; minimum neighbor number = 500, population = European.

We used different r2 threshold for GWAS (r2=0.8) vs eQTL (r2=0.99) enrichment analyses because eQTL analyses measure a molecular feature instead of a complex phenotype and therefore have higher resolution to identify the more likely causal variants.

2.5.6 Analysis of LCL-specific expression (LCL-ESI)

We used an information theory approach [161, 64] to score genes based on LCL expression level and specificity relative to the panel of 50 diverse GTEx tissues, each of which had RNA-seq data for more than 25 samples. We downloaded RNA-seq data from GTEx V7 study from the website `https://www.gtexportal.org/home/datasets` filename `GTEx_Analysis_2016-01-15_v7_RNASeQCv1.1.8_gene_median_tpm.gct.gz`. This data was in the form of median transcripts per million (TPM) for each gene in each tissue. We considered protein-coding genes and removed those that were lowly expressed in LCL (median TPM > 0.15) to avoid potential artifacts. We

calculated the relative expression of each gene (g) in LCL compared to all 50 tissues (t) as p:

$$p_{g,LCL} = \frac{x_{g,LCL}}{\sum_{n=1}^{50} x_{g,t}}$$

We next calculated the entropy for expression of each gene across all 50 tissues as H:

$$H_g = -\sum_{n=1}^{50} p_{g,t} log_2(p_{g,t})$$

Following previous studies [161, 64], we defined LCL tissue expression specificity (Q) for each gene as:

$$Q_{g,LCL} = H_g - log_2(p_{g,LCL})$$

To aid in interpretability, we divided Q for each gene by the maximum observed Q and subtracted this value from 1 and refer to this new score as the LCL expression specificity index (LCL-ESI):

$$LCL - ESI_g = 1 - \frac{Q_{g,LCL}}{Q_{max,LCL}}$$

LCL-ESI scores near zero represent lowly and/or ubiquitously expressed genes and scores near 1 represent genes that are highly and specifically expressed in LCL.

Enrichment for distance to genes based on gene expression specificity in LCL We binned the protein coding genes into quintiles based on LCL-ESI, such that bin 5 included the most LCL specific genes. Each quintile bin contained N=2,753 protein coding genes. We then used BEDtools closest to calculate distance to nearest protein coding gene TSS for each bin, obtaining empirical cumulative distribution functions (ECDFs) for each regulatory annotation in each cell type. Since the regulatory anno-

tations vary in the number of segments, and will therefore have different probabilities to occur nearby TSS, the distance to nearest protein coding gene TSS ECDFs cannot be directly compared. We therefore obtained the expected distance to nearest protein coding gene TSS ECDF for each annotation by randomly sampling N=2,753 genes from across the five bins 10,000 times and calculating the distance to nearest gene. We then calculated the TSS proximity enrichment for each annotation by dividing the observed with the mean expected ECDF. Enrichment therefore denotes the fold change in the observed fraction of annotation segments within a certain distance of protein coding gene TSSs in a specific LCL-ESI bin over the mean fraction of segments at the same distance from the randomly sampled genes. The 95% confidence intervals for the enrichment values were calculated as observed/(mean 1.96*SE), where SE = standard error of the mean expected fraction.

Enrichment to overlap eQTL based on expression specificity of gene We sorted the eQTL SNPs into quintiles based on LCL-ESI of the associated genes (eGene) and grouped them into five equally sized bins, resulting in 585 eQTLs in each bin. Bin numbers represent eQTLs that correspond to increasingly LCL-specific genes where bin number 1 represents the least LCL-specific and bin number 5 represents the most LCL-specific genes. We calculated the enrichment for each eQTL set to overlap regulatory annotations using GREGOR with the same parameters as described above for the bulk set of LCL eQTL. To quantify the trend of LCL eQTL enrichment with LCL eGene expression specificity, we calculated the Spearman correlation of the enrichment effect size expressed as $\log_2$(fold enrichment) with the eQTL bin number using the cor() function from the stats package (v3.5.1) in R(R Core Team).

2.5.7 Gene expression and chromatin accessibility QTL effect sizes in regulatory annotations

We used the beta values or the slope of the linear regression as the effect size of LCL and blood eQTL (GTEx V7) and DNase hypersentivity site QTL (dsQTL) [32]. All of these QTL studies used inverse rank based normalization steps on the molecular features, which enables direct comparison of the effect sizes across the genome. Because low MAF SNPs have low statistical power to be detected as significant QTL at low effect sizes, these SNPs are biased to have large QTL effect sizes. We therefore removed QTL SNPs with MAF < 0.2. We pruned the QTL SNPs to retain SNPs with r2 < 0.8 after sorting by p value of association as described above using PLINK [152]; [22]. Since the causal SNP for the QTL signal is unknown, we also considered SNPs in high linkage disequilibrium at r2 > 0.99 with the lead QTL SNPs which were obtained using vcftools[30] and 1000 genomes phase 3 reference vcf specified above. We observed higher eQTL enrichment in annotations with increasing the r2 thresholds, which is indicative of a higher signal to noise ratio. A previous study analyzing LCL eQTL also showed that functional enrichment decreased rapidly from the best eQTL towards lower ranked eQTL [94]. We compared the absolute QTL effect sizes of loci (QTL index SNP or SNP with r2 > 0.99 with the index SNP) that overlapped each GM12878 annotation. We used the Wilcoxon rank sum test to identify significant differences between effect sizes of eQTL overlapping each annotation.

To test if there may be confounding from other genomic properties such as the distance between the eQTL eSNP to eGene and number of SNPs in high LD with the lead SNP could also influence the eQTL effect size, we calculated the contribution of the underlying regulatory annotation on the effect size while accounting for these factors. We modeled the eQTL effect size in a linear regression using the Python statsmodels library where we included a regulatory annotation indicator variable encoding eQTL overlap by a stretch enhancer or HOT region annotation and the

following two covariates: (1) absolute distance of the eQTL lead SNP to its corresponding eGene TSS and (2) total number of SNPs in high LD (r2 > 0.99 with the lead SNP) that overlapped the annotation. eQTLs that overlapped both annotations were not considered. Summary statistics of this regression model are presented in 2.1.

To calculate the statistical power for eQTL analysis after Bonferroni correction based on a linear regression, we used the powerEQTL.SLR function from the powerEQTL R package [38] (v0.1.3; `https://rdrr.io/cran/powerEQTL/`). For eQTLs overlapping each annotation, we used the eQTL effect sizes representing the 10th to 90th percentile values and calculated power by using the following parameters: MAF = 0.2, type I error rate = 0.0005, total number of tests = 1000,000, standard deviation of the error term = 0.4 and sample size N = 250.

2.5.8 Comparison of allelic bias effect sizes in annotations

To determine SNP allelic bias in GM12878 ATAC-seq data, we used the publicly available data [17]. Adapters were trimmed using cta (v. 0.1.2; `https://github.com/ParkerLab/cta`), and reads mapped to hg19 using bwa mem [96] (default options except for the -M flag; v. 0.7.15-r1140). Bam files were filtered for high-quality autosomal read pairs using samtools [98] view (-f 3 -F 4 -F 8 -F 256 -F 2048 -q 30; v. 1.3.1). WASP [194] (version 0.2.1, commit 5a52185; using python version 2.7.13) was used to adjust for reference mapping bias; for remapping the reads as part of the WASP pipeline, we used the same mapping and filtering parameters described above for the initial mapping and filtering. Duplicates were removed using WASPs rmdup_pe.py script. We used the phased GM12878 VCF file downloaded from `ftp://ftp-trace.ncbi.nlm.nih.gov/giab/ftp/release/NA12878_HG001/NISTv3.3.1/GRCh37/HG001_GRCh37_GIAB_highconf_CG-IllFB-IllGATKHC-Ion-10X-SOLID_CHROM1-X_v.3.3.1_highconf_phased.vcf.gz`. To avoid potential artifacts associated with double-counting

alleles, overlapping read pairs were clipped using bamUtil clipOverlap (v. 1.0.14; `http://genome.sph.umich.edu/wiki/BamUtil`: clipOverlap). The bam files from the samples in Table 2.2 were then merged to create a single GM12878 bam file using samtools merge. We filtered for heterozygous autosomal SNPs with minimum coverage of 30. Since the power to detect allelic bias depends upon the read coverage at the SNP, SNPs with lower coverage are biased toward having higher effect sizes at any given level of statistical significance. To prevent this type of bias, we randomly down-sampled reads at each heterozygous SNP to a total of 30 reads with base quality of at least 20. We then counted the number of reads containing each allele. We used a two-tailed binomial test that accounted for reference allele bias to evaluate the significance of the allelic bias at each SNP (as described previously [196]; implemented in a custom perl script). We did not test SNPs in regions blacklisted by the ENCODE Consortium because of poor mappability (wgEncodeDacMapabilityConsensusExcludable.bed and wgEncodeDukeMapabilityRegionsExcludable.bed). We then selected SNPs that show significant allelic bias at a nominal threshold of binomial test p value < 0.05 and used BEDtools intersect to identify the set of nominally significant SNPs overlapping each annotation. We defined the effect size of allelic bias as the absolute deviation from expectation given by the absolute difference between the observed and expected fraction of reads mapping to the reference allele. We also compared the allelic bias effect sizes while only considering SNPs with MAF > 0.2.

2.6 Data Availability

Supplementary material is available at FigShare. Workflows for analyses as described below were run using Snakemake [86]. All analysis steps and code to facilitate reproducibility of this work are openly shared at GitHub `https://github.com/ParkerLab/regulatoryAnnotations_comparisons`. Static version of scripts and

Table 2.2: GM12878 ATAC-seq sample information

GEO accession	Run accession	Cell type	Cell count	Replicate
GSM1155957	SRR891268	GM12878	50000	rep1
GSM1155958	SRR891269	GM12878	50000	rep2
GSM1155959	SRR891270	GM12878	50000	rep3
GSM1155960	SRR891271	GM12878	50000	rep4
GSM1155961	SRR891272	GM12878	500	rep1
GSM1155962	SRR891273	GM12878	500	rep2
GSM1155963	SRR891274	GM12878	500	rep3

all processed data are deposited to Zenodo `https://zenodo.org/record/1413623#.W8f2x1JRfpB`

2.7 Acknowledgements and publication

The results presented in this chapter have been published in [195]. I thank Steve for his ideas, insight and contributions for this work. I also thank all the other authors for their contributions.

CHAPTER III

Understanding the Genetics of Gene Expression in Human Pancreatic Islets

3.1 Abstract

Genome-wide association studies (GWAS) have identified >400 independent single nucleotide polymorphisms (SNPs) that modulate the risk of type 2 diabetes (T2D) and related traits. However, the pathogenic mechanisms of most of these SNPs remain elusive. Here, we examined genomic, epigenomic, and transcriptomic profiles in human pancreatic islets to understand the links between genetic variation, chromatin landscape, and gene expression in the context of T2D. We first integrated genome and transcriptome variation across 112 islet samples to produce dense cis-expression quantitative trait loci (cis-eQTL) maps. Further integration with chromatin state maps for islets and other diverse tissue types revealed that cis-eQTLs for islet specific genes are specifically and significantly enriched in islet stretch enhancers (SEs). High resolution chromatin accessibility profiling using ATAC-seq in two islet samples enabled us to identify specific transcription factor (TF) footprints embedded in active regulatory elements, which are highly enriched for islet cis-eQTL. Aggregate allelic bias signatures in TF footprints enabled us de novo to reconstruct TF binding affinities genetically, which support the high-quality nature of the TF footprint

predictions. Interestingly, we found that T2D GWAS loci were strikingly and specifically enriched in islet Regulatory Factor X (RFX) footprints. Remarkably, within and across independent loci, T2D risk alleles that overlap with RFX footprints uniformly disrupt the RFX motifs at high information content positions. Together, these results suggest that common regulatory variations have shaped islet TF footprints and the transcriptome, and that a confluent RFX regulatory grammar plays a significant role in the genetic component of T2D predisposition.

3.2 Introduction

Type 2 diabetes (T2D) is a complex disease characterized by pancreatic islet dysfunction and insulin resistance in peripheral tissues. >90% of T2D SNPs identified through genome wide association studies (GWAS) reside in non-protein coding regions and are likely to perturb gene expression rather than alter protein function [129]. In support of this, we and others recently showed that T2D GWAS SNPs are significantly enriched in enhancer elements that are specific to pancreatic islets [142, 144, 189]. The critical next steps to translate these islet enhancer T2D genetic associations into mechanistic biological knowledge are (a) identifying the putative functional SNP(s) from all those that are in tight linkage disequilibrium (LD), (b) localizing their target gene(s), and (c) understanding the direction-of-effect (increased or decreased target gene expression) conferred by the risk allele. Two recent studies performed integrative analyses of genome variation and gene expression variation across human islet samples to identify cis expression quantitative trait loci (cis-eQTL) that linked T2D GWAS SNPs to target genes [44, 193]. However, the transcription factor (TF) molecular mediators of the islet cis-eQTL remain poorly understood and represent important links to upstream pathways that will help untangle the regulatory complexity of T2D.

3.3 Results

3.3.1 Integrated analysis of transcriptome and epigenome islet data

To build links between SNP effects on regulatory element use and gene expression in islets, we performed strand-specific mRNA sequencing of 31 pancreatic islet tissue samples to an average depth of 100M paired end reads. In parallel, we analyzed unstranded mRNA-seq data for 81 islet samples from a previous study [44]. We subjected both datasets to the same quality control and processing. We additionally completed dense genotyping of the 31 islet samples and downloaded genotypes for 81 previously described islet samples [44]. Phasing and imputation yielded a final set of 6,060,203 autosomal SNPs (Methods) present in both datasets with an overall minor allele count (MAC) >10. To identify SNPs affecting gene expression within 1 Mb of the most upstream transcription start site (TSS), we performed separate cis-eQTL analyses for the two sets of islet samples, and combined the cis-eQTL results via meta-analysis. We identified 3,964 unique autosomal cis-eQTL lead SNPs for 3,993 genes at a 5% false discovery rate (FDR).

Next, we integrated chromatin immunoprecipitation followed by sequencing (ChIP-seq) data for 5 histone modifications across islets [142, 162] and 30 diverse tissues with publicly available datasets (Table 3.1) [186, 125, 17] using ChromHMM [40, 43, 41]. This produced 13 unique and recurrent chromatin states (Fig. 3.1A, chromatin state tracks; Fig. 3.2) including promoter, enhancer, transcribed, and repressed regions, which were annotated after analyzing overlap enrichments with the chromatin states reported by the Roadmap Epigenome Consortium [186] for matching cell types. To identify specific regulatory element sites within these chromatin states, we profiled open chromatin in two islets using the assay for transposase-accessible chromatin-sequencing (ATAC-seq) [147] (Fig. 3.1A, ATAC-seq tracks). Our high-depth ATAC-seq data (>1.4B reads across both islets) allowed us to identify transcription factor

Cell type	H3K27ac (NSC)	H3K4me3 (NSC)	H3K27ac (RSC)	H3K4me3 (RSC)	QC status used to learn ChromHMM model
Islets	2	2.3	2.21	1.33	Passed
GM12878	1.43	1.24	1.28	0.85	Passed
H1	1.1	1.59	1.13	1.1	Passed
HepG2	1.81	2.42	1.35	1.3	Passed
HMEC	1.56	1.97	1.27	1.19	Passed
HSMM	1.73	2.7	1.34	1.35	Passed
Huvec	1.76	2	1.49	1.15	Passed
K562	1.67	1.3	1.54	1.07	Passed
NHEK	1.73	1.64	1.3	1.07	Passed
NHLF	1.74	1.92	1.42	1.22	Passed
Adipose	1.15	1.6	0.98	1.16	Passed
Anterior caudate	1.2	1.14	1.18	1.09	Passed
CD34-PB	1.5	1.69	1.47	1.07	Passed
Colonic mucosa	1.28	1.72	1.11	1.14	Passed
ES-HUES6	1.1	1.78	1.13	1.3	Passed
Liver	1.38	1.3	1.45	0.96	Passed
Mid-frontal lobe	1.14	1.1	0.88	1.01	Passed
Rectal mucosa	1.59	1.36	1.52	1.11	Passed
Rectal smooth muscle	1.31	1.25	1.28	1.1	Passed
Skeletal muscle	1.37	1.24	1.29	1.07	Passed
Stomach smooth muscle	1.11	1.51	0.83	1.16	Passed
hASC-t1	2.18	2.36	1.4	1.24	Passed
hASC-t2	1.94	2.36	1.36	1.23	Passed
hASC-t3	1.84	2.42	1.41	1.27	Passed
hASC-t4	1.7	2.44	1.31	1.26	Passed
Cingulate gyrus	1.1	1.12	0.61	1.06	Failed
Duodenum mucosa	1.13	1.24	0.6	1.02	Failed
ES-HUES64	1.04	1.88	0.74	1.25	Failed
Hippocampus middle	1.29	1.07	1.32	0.89	Failed
Inferior temporal lobe	1.13	1.14	0.45	1.04	Failed
Substantia nigra	1.07	1.07	0.8	0.86	Failed

Table 3.1: NSC and RSC scores for H3K27ac and H3K4me3 datasets used in this study

(TF) DNA footprints using the CENTIPEDE algorithm (13) (Methods). We assigned regulatory state and TF footprint status to every islet cis-eQTL based on the annotation of SNPs with r2 >0.8 with the lead SNP (Fig. 3.1B). We used iterative conditional analyses [162] to identify 28 T2D and related quantitative trait GWAS SNPs that could be islet cis-eQTL signals (Fig. 3.1C; Dataset S1: Conditional analysis results for GWAS SNP-gene pairs that were significant in the original eQTL analysis; Dataset S2: Epigenetic annotation for SNPs in r2 $\geq$ 0.8 with the GWAS variants included in Dataset S1). Given the modest eQTL signals at most of these loci, conditional analysis in larger islet samples cis-eQTL signals will likely change this list.

As an example, T2D GWAS index SNP rs1535500 occurs at the KCNK16 locus and the risk allele results in a glutamate substitution at alanine 277 (A277E).

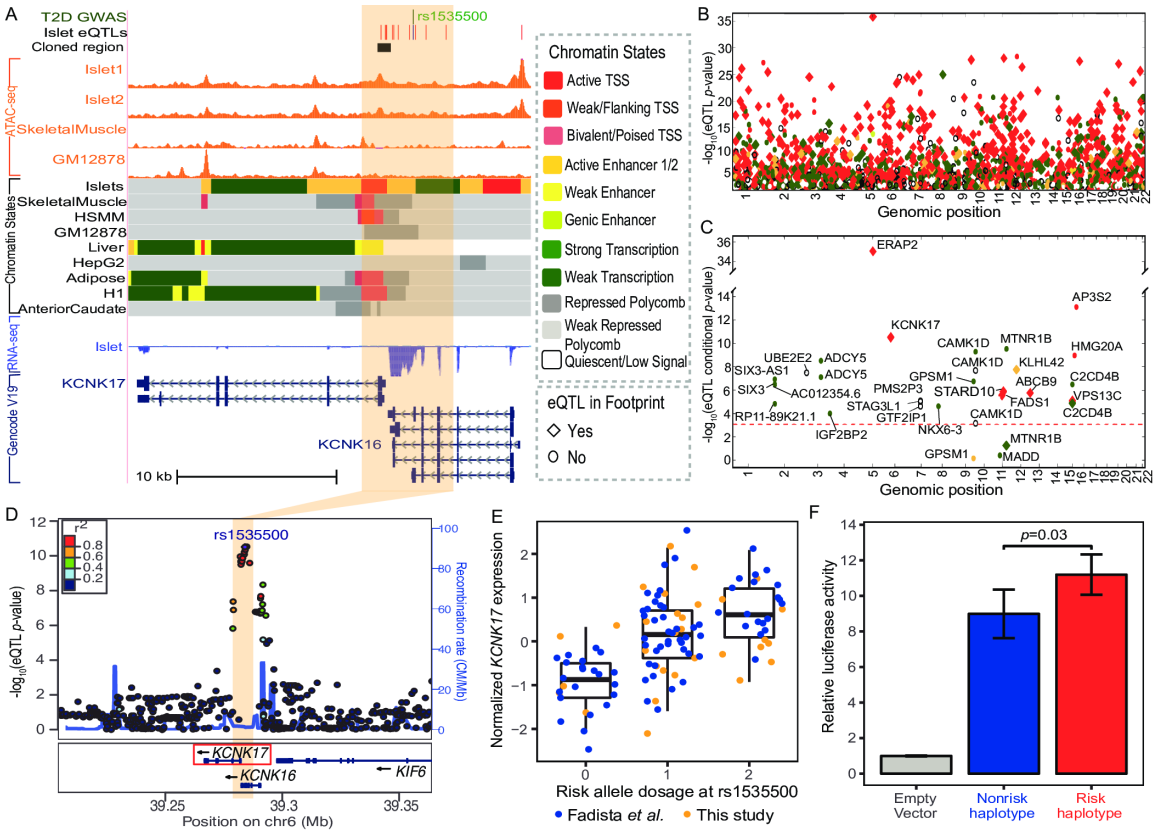
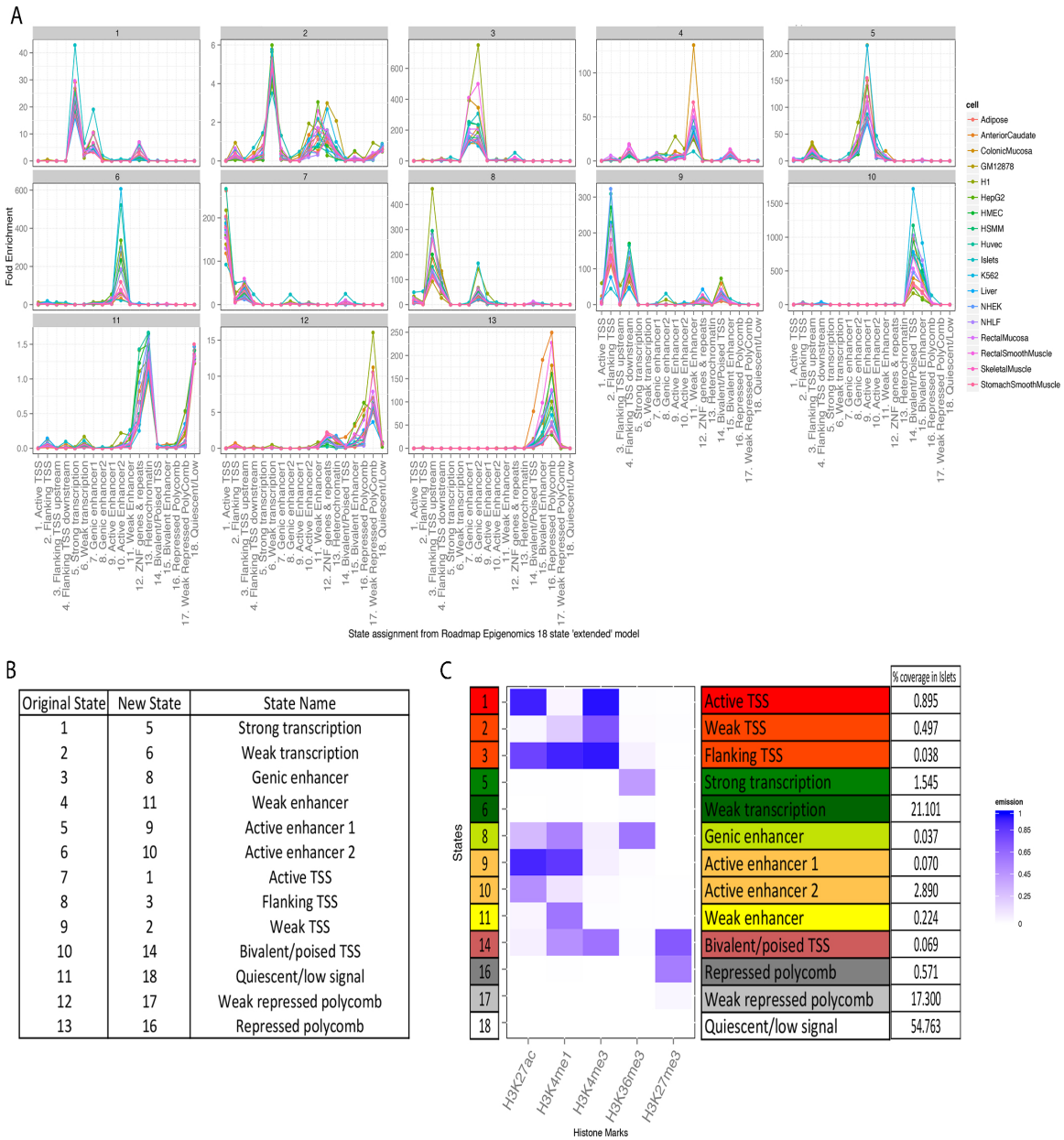


Figure 3.1: Integrated genomic, epigenomic, and transcriptomic analyses of human pancreatic islets. (A) An overview of diverse molecular profiling data types used in this study. Integrative molecular profiling (open chromatin, ATAC-seq; chromatin states; RNA-seq) highlights islet-specific signatures at the KCNK17 locus. (B) Plot of strength of association (y axis) for significant islet cis-eQTLs colored by chromatin-state annotation (A) by chromosomal location (x axis); diamonds indicate SNPs overlapping ATAC-seq footprints. An interactive version of this plot can be found at theparkerlab.org/tools/isleteqtl/. (C) Plot of strength of islet cis-eQTL association for T2D and related trait GWAS SNPs after conditional analysis to identify variants likely independent of stronger cis- eQTL signals for the same gene by chromosomal position and annotated as in B. The plot includes all GWAS SNPgene pairs with FDR < 0.05 in original cis- eQTL analysis. The dotted red line represents the P value threshold for FDR < 0.05 based on the conditional analysis. (D) Islet cis-eQTL associated with KCNK17 expression highlighted for comparison with molecular profiling tracks in A. (E) Plot of normalized KCNK17 expression in islet samples and cis-eQTL risk allele dosage. (F) Functional validation of KCNK17 cis-eQTL at its promoter region. The haplotype containing alleles associated with T2D risk and increased KCNK17 expression (rs10947804-C, rs12663159-A, rs146060240-G, and rs34247110-A) shows higher transcriptional activity than the haplotype with nonrisk alleles. The cloned region is indicated at the top of A. Relative luciferase activity is given as mean SD of four to five independent clones per haplotype normalized to empty vector. Significance was evaluated using a two-sided t test.



Original State	New State	State Name
1	5	Strong transcription
2	6	Weak transcription
3	8	Genic enhancer
4	11	Weak enhancer
5	9	Active enhancer 1
6	10	Active enhancer 2
7	1	Active TSS
8	3	Flanking TSS
9	2	Weak TSS
10	14	Bivalent/poised TSS
11	18	Quiescent/low signal
12	17	Weak repressed polycomb
13	16	Repressed polycomb

		% coverage in Islets
1	Active TSS	0.895
2	Weak TSS	0.497
3	Flanking TSS	0.038
5	Strong transcription	1.545
6	Weak transcription	21.101
8	Genic enhancer	0.037
9	Active enhancer 1	0.070
10	Active enhancer 2	2.890
11	Weak enhancer	0.224
14	Bivalent/poised TSS	0.069
16	Repressed polycomb	0.571
17	Weak repressed polycomb	17.300
18	Quiescent/low signal	54.763

Figure 3.2: Thirteen-chromatin-state model built from histone modification ChIP-seq data generated using ChromHMM [43] for 33 cell types. (A) Each graph represents the overlap enrichment for 18 cell types of each of our 13 generated chromatin states with the Roadmap Epigenomics [186] reported states. (B) Renaming of generated 13 states (Original State) according to Roadmap Epigenomics overlap enrichments (New State) in A. (C) State numbers, histone mark emission probabilities, state names, and percentage genomic coverage of each chromatin state in human islets.

This change was implicated in increasing the KCNK16 basal channel activity and cell surface localization when tested in a mouse model [197]. Our analysis revealed that rs1535500 is not associated with KCNK16 expression (Fig. 3.3). Interestingly, the rs1535500 risk allele is associated with increased expression of the neighboring potassium channel gene KCNK17 (Fig. 3.1D, E). We observed that rs1535500 is in high linkage disequilibrium (LD) (r2>0.95) with four SNPs (rs10947804, rs12663159, rs146060240 and rs34247110) that are located in an islet promoter chromatin state; all but rs34247110 are located in an ATAC-seq peak (Fig. 3.1A). Motivated by the overlap with islet regulatory annotations, we cloned two different copies of the 473 bp DNA sequence surrounding these SNPs, one containing the T2D risk alleles for each of the 4 SNPs (risk haplotype), and the other containing the non-risk alleles (non-risk haplotype). We performed luciferase reporter assays in the mouse insulinoma (MIN6) beta cell line to test the transcriptional activity of these two clones. Both clones exhibited promoter activity (10-fold increased compared to empty vector) but the T2D risk haplotype showed significantly greater (24%, p=0.03) transcriptional activity than the non-risk haplotype (Fig. 3.1F). This suggests that one or more of these T2D risk variants cause increased regulatory activity in islets. We note these findings highlight a complex functional genetic architecture for a single haplotype that results in regulatory activity linked to one gene (KCNK17) and coding variation in another (KCNK16). Together, these results illustrate how integrated analyses help to identify potential causal SNPs associated with islet expression and T2D risk. To enable easy, in-depth exploration of our integrated genome-wide results, we created an interactive islet cis-eQTL and chromatin state browser (`http://theparkerlab.org/tools/isleteqtl/`).

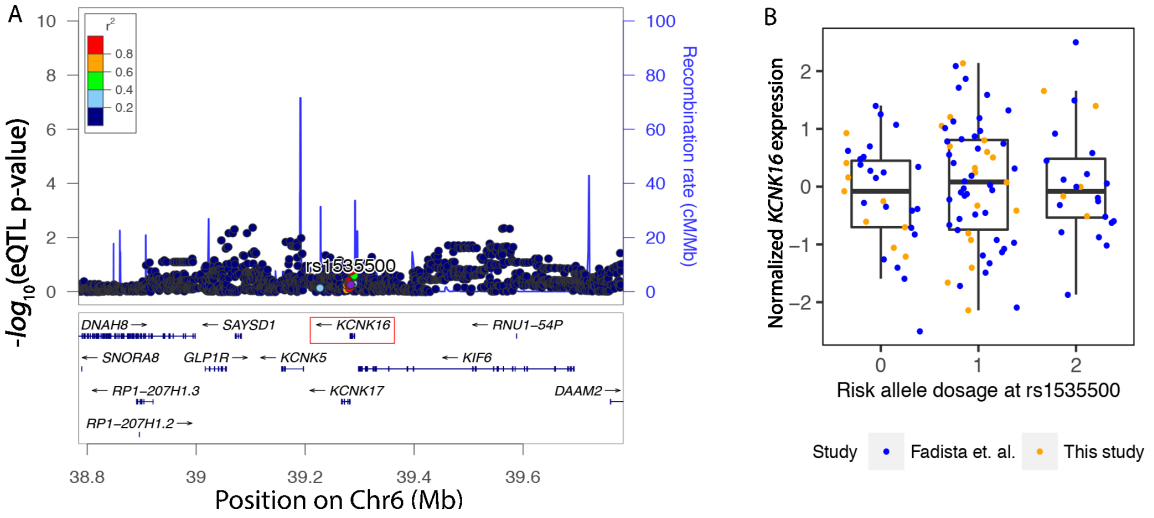


Figure 3.3: (A) LocusZoom plot showing that a T2D GWAS SNP (`rs1535500/chr6:39284050`, hg19, purple; other variants in LD colored according to r2) is not associated with KCNK16 expression in islets. (B) Plot for normalized KCNK16 expression and rs1535500 risk allele dosage from mRNA-seq and genotyping data in islet samples.

3.3.2 Common and islet-specific gene eQTLs are enriched in different chromatin states

To understand the regulatory architecture of islet eQTLs, we measured their co-occurrence with different classes of chromatin states across diverse tissues, including stretch enhancers (SEs), defined as enhancer chromatin states ≥3kb long. These tend to mark cell identity regions and have been shown to harbor tissue-specific GWAS SNPs [142, 153]. We calculated genome wide enrichment for cis-eQTL overlaps with these features while controlling for minor allele frequency, distance to TSS, and the number of SNPs in LD [160]. cis-eQTLs were enriched in active chromatin states such as promoter, and genic enhancer in islets, while inactive states such as polycomb repressed were depleted for such overlaps across multiple tissues (Fig. 3.4). Reasoning that this common enrichment pattern across diverse tissues may be largely driven by cis-eQTLs of commonly expressed housekeeping genes, we sought to classify cis-eQTLs by the islet expression specificity of their associated genes. To measure gene expression specificity in islets, we analyzed RNA-seq data from 16 additional tissues

from the Illumina Human Body Map 2.0 project. We used an information theory approach to define the islet expression specificity index (iESI, 3.5, also see Methods) as in our previous skeletal muscle study [162]. iESI values near zero represent lowly and/or ubiquitously expressed genes whereas values near one represent genes that are highly and specifically expressed in islets. We divided genes into quintiles based on the ascending iESI (3.5). We assigned eQTLs for these genes to their respective iESI quintile and estimated enrichment of each set in chromatin annotations. Interestingly, while eQTLs across iESI quintile bins were similarly enriched in islet promoter states, eQTL enrichment in active and SE states increased concomitantly with the islet specificity (3.6). As an example, we found that the cis-eQTL for the KCNA6 gene (3.7A) which is expressed in islets with high specificity (KCNA6 iESI = 0.78), overlapped islet-specific enhancer states (3.7B). We note that this cis-eQTL locus does not overlap a known T2D GWAS locus. When we restricted our enrichment analysis to ATAC-seq peaks in SE states, we saw a stronger trend toward increasing enrichment by iESI quintile (3.6). These results indicate a strong link between active regulatory chromatin architecture and the genetic control of cell-specific gene expression.

To further identify and dissect regulatory regions critical for islet-specific gene expression, we sought to distinguish between shared and tissue specific enhancer chromatin states. We performed k-means clustering for active enhancer chromatin states across 31 cells/tissues (Methods). This method segregated enhancer regions based on activity across diverse tissues; for example, cluster 13 is islet specific, while cluster 3 is liver specific (Fig. 3.7C). We compared these enhancer clusters with SE annotations across tissues and found that tissue specific clusters, such as the islet-specific cluster 13, indeed displayed high enrichment for islet SEs (Fig. 3.7D). Likewise, in other tissues, tissue-specific enhancer clusters were enriched for the corresponding tissues SEs (Fig. 3.7D). Next, we asked if islet cis-eQTL were enriched in specific enhancer

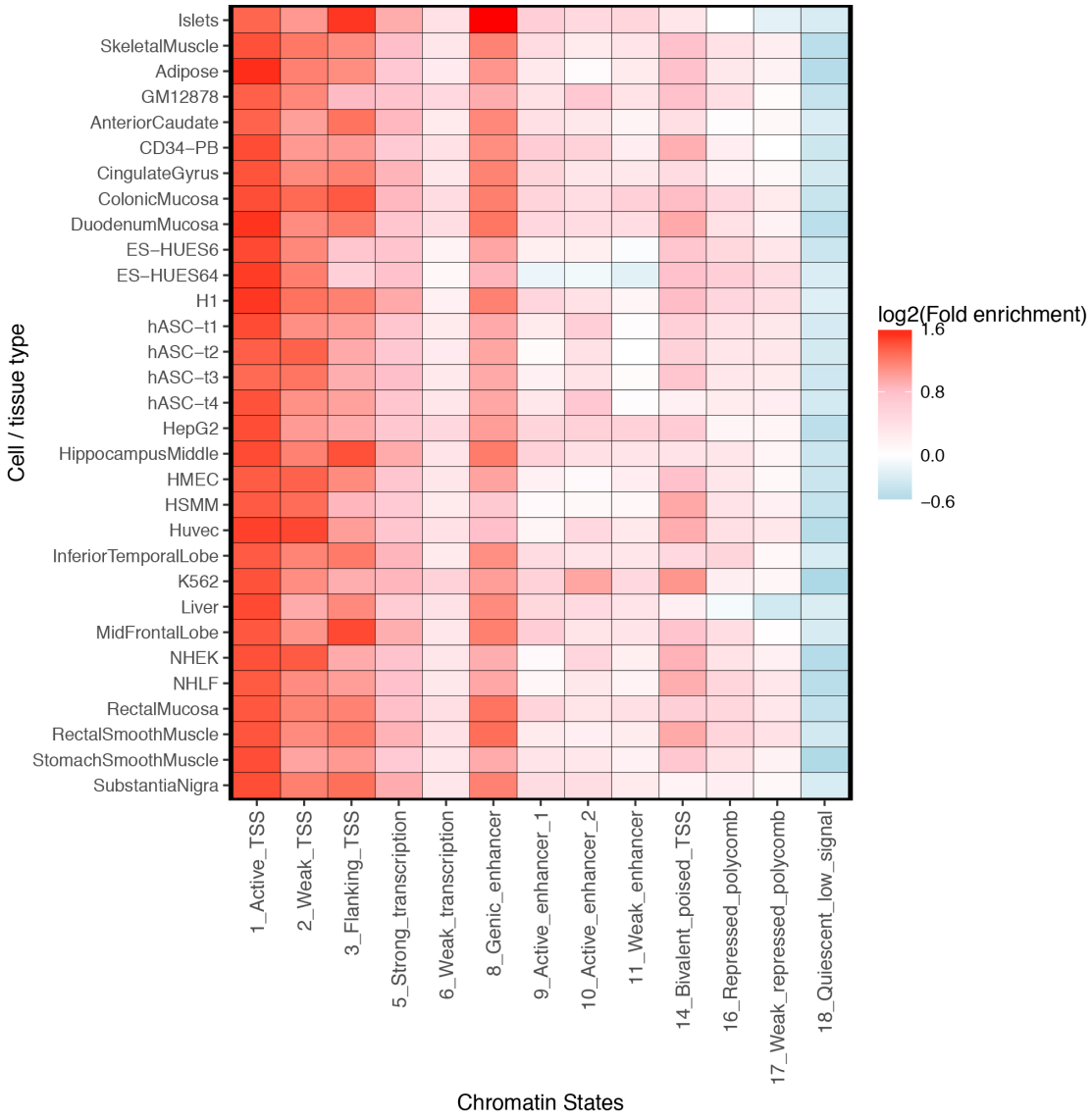


Figure 3.4: Fold enrichment of islet eQTLs in chromatin states across cells/tissues.

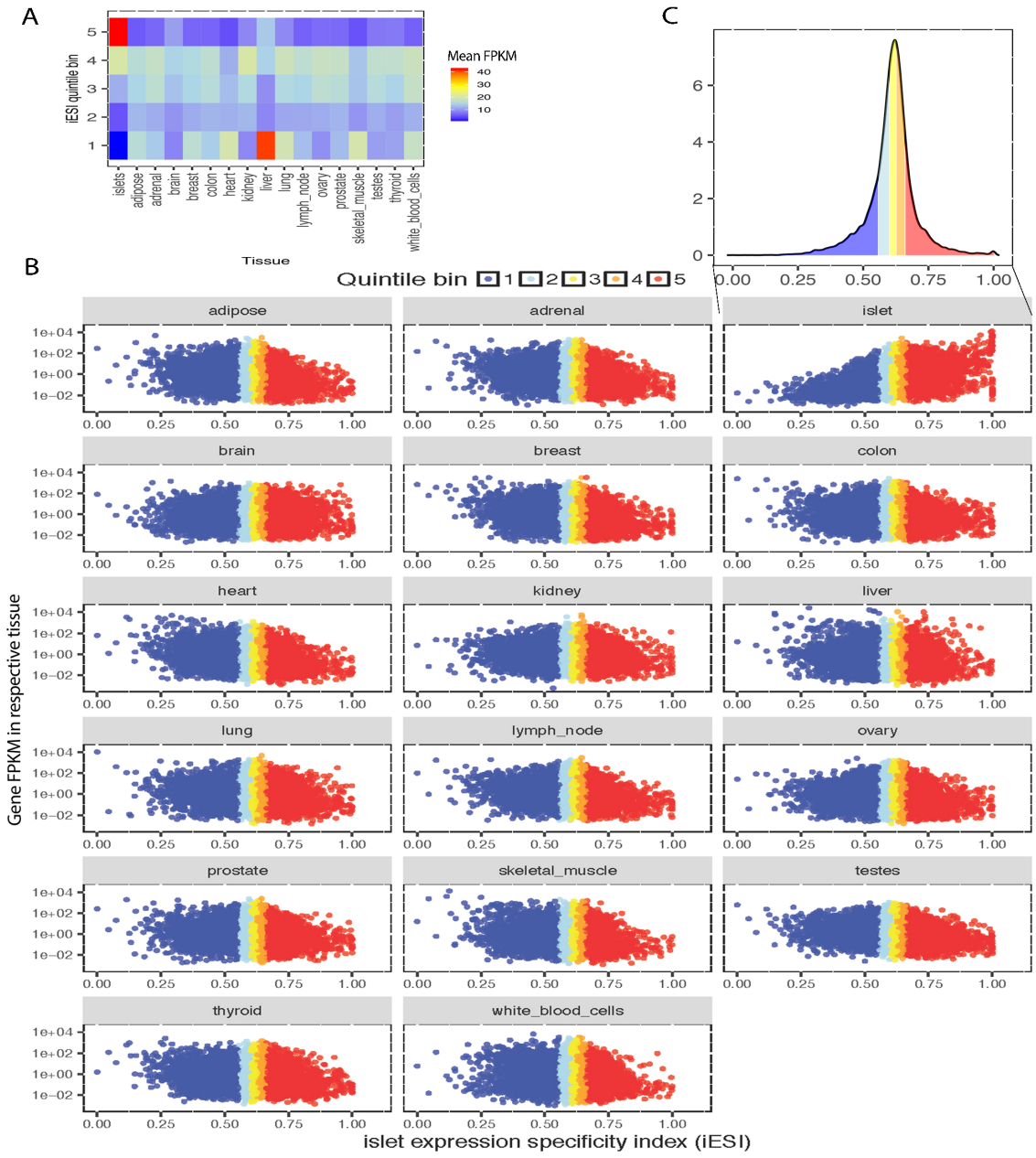


Figure 3.5: iESI. (A) Heat map showing mean FPKM for genes expressed in different tissues when binned by iESI quintiles. (B) Scatterplots showing FPKM for genes expressed in different tissues vs. the iESI. (C) Distribution of iESI by quintile of expression.

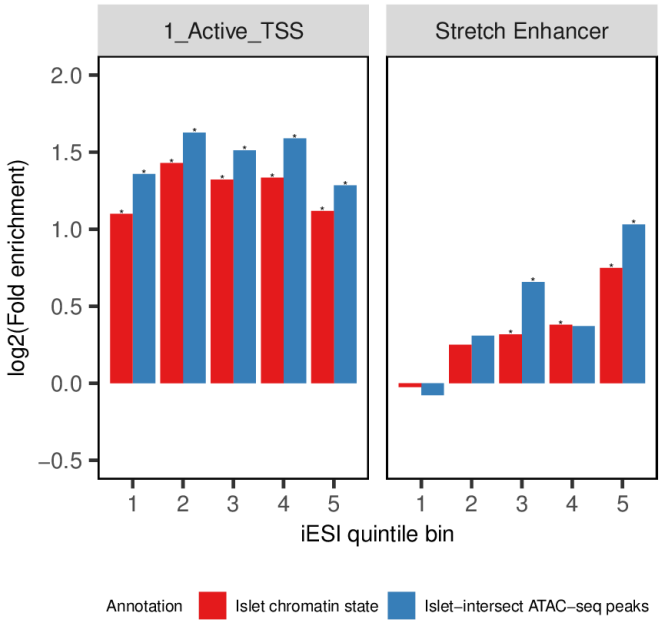


Figure 3.6: Enrichment of islet cis-eQTLs binned into quintiles by target gene iESI in islet active TSS and stretch enhancer chromatin states (red) and consensus islet intersect ATAC-seq peaks (present in both islet samples) in these states (blue). $*$ P < 0.05 from GREGOR analysis.

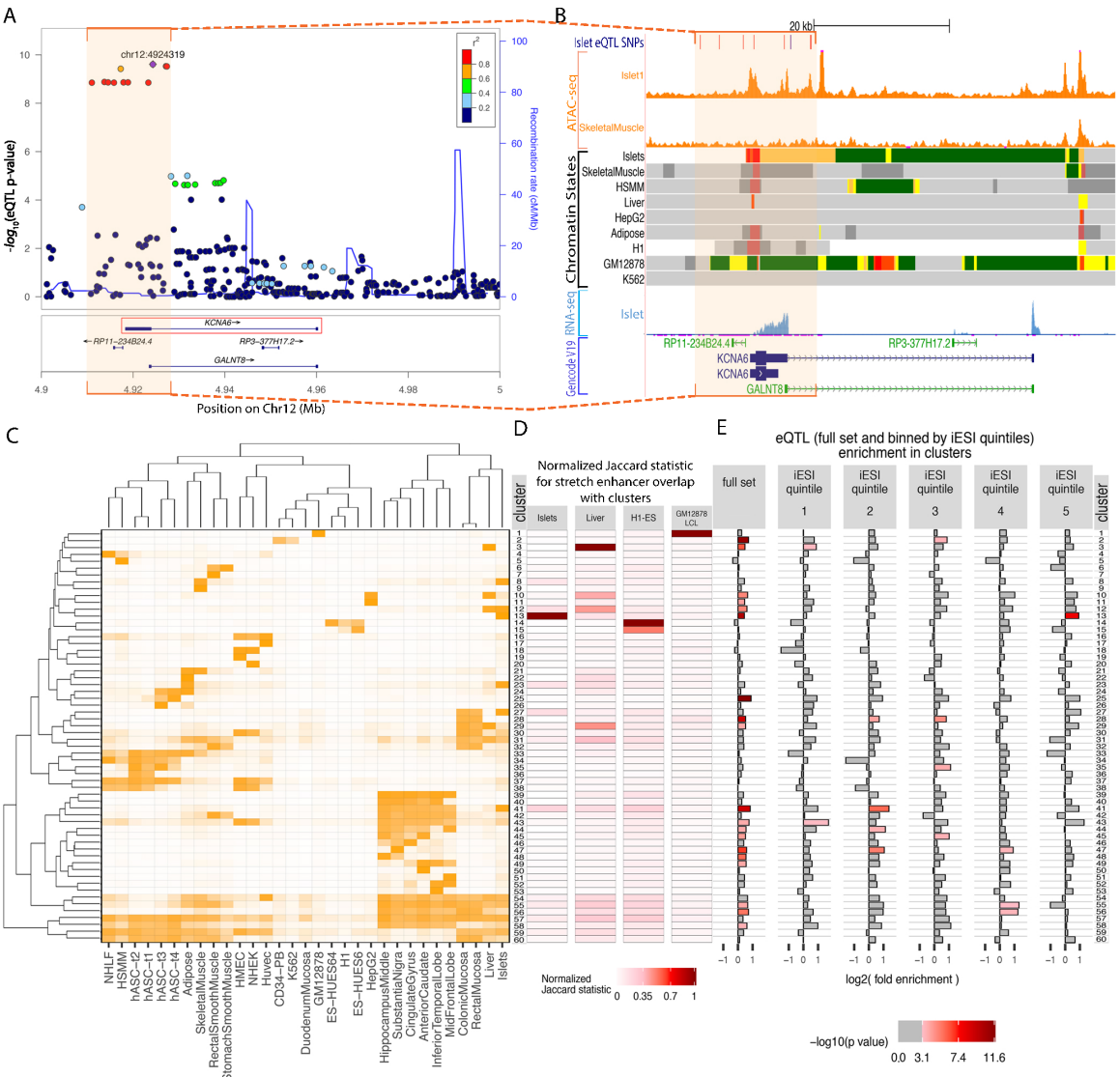


Figure 3.7: Common and islet-specific gene eQTLs are enriched in different chromatin states. (A) LocusZoom plot of an islet cis-eQTL in the KCNA6 locus. (B) The cis-eQTL for KCNA6, which is in the top quintile of the iESI (iESI 5), overlaps an islet-specific enhancer state. (C) Active enhancer clustering (y axis) across cell types (x axis) reveals cell-specific enhancer regions. Cluster 13 is islet-specific. (D) Degree of overlap of enhancer clusters with stretch enhancers from four cell types. Islet stretch enhancers show the strongest overlap with islet-specific enhancer cluster 13, whereas GM12878 stretch enhancers show the strongest overlap with GM12878-specific enhancer cluster 1. The Jaccard statistic was normalized per column, so that values range from zero (no overlap) to one (maximum observed overlap). (E) Enrichment of islet eQTLs across enhancer clusters reveals that the full set of eQTLs (column 1) is enriched across multiple enhancer clusters, whereas eQTLs for islet-specific genes (iESI quintile 5; column 5) are enriched in the islet-specific enhancer cluster 13. Gray bars indicate nonsignificant after Bonferroni correction.

clusters and observed enrichment in multiple clusters (Fig. 3.7E). We then stratified the cis-eQTLs by iESI quintile and repeated this analysis. Notably, islet cis-eQTLs for genes in iESI quintile 5 only showed significant enrichment in the islet specific enhancer cluster 13 (p-value = 1.2×10^{-8}, fold enrichment = 1.91, Fig. 3.7E). Together, these results demonstrate that islet tissue-specific genetic regulatory architecture is enriched in islet-specific enhancers and SEs.

3.3.3 Islet eQTL are enriched in islet ATAC-seq peaks and DNA footprints

Chromatin state maps identify regulatory regions such as promoters and enhancers but lack the resolution to pinpoint specific sites that may be bound and regulated by a TF. To refine the link between genetic variation, TF binding sites, and gene expression, we leveraged the high-resolution ATAC-seq data to identify in vivo putative TF binding sites. We employed CENTIPEDE as previously described to predict whether TF motif occurrences in ATAC-seq peaks are bound or unbound by the corresponding TF [162, 147]. This approach detected high-quality footprints for many TFs, including the general CCCTC-binding factor (CTCF) and the more islet specific TF Regulatory Factor X (RFX) (Fig. 3.8A, B). Fig. 3.8A depicts the integrated chromatin and transcriptional landscape at RFX6, a gene that exhibits islet-specific expression (iESI = 0.94, iESI quintile bin = 5). Notably, we detect RFX footprints in islet SEs near RFX6 (Fig. 3.8A), suggesting an autoregulatory mechanism that, based on recent studies [144, 157], may indicate RFX6 is an islet core transcriptional regulatory gene. We detected periodic, nucleosome-sized spikes in the ATAC-seq signal adjacent to the CTCF footprint regions (Fig. 3.8B), consistent with its reported nucleosome-phasing properties [47]. Comparing ATAC-seq profiles from islets to those of skeletal muscle tissue [162], adipose tissue [4] and a lymphoblastoid cell line (GM12878) [17], we found that islet ATAC-seq peaks occurred preferentially in

islet promoter and enhancer chromatin states (Fig. 3.9). Islet cis-eQTLs were highly enriched in multiple TF footprint motifs (p-value range: 5.1×10^{-5} - 5.0×10^{-23}; fold enrichment range: 2.02-6.67) but not in non-footprint motifs (Fig. 3.8C, Dataset S3, Methods). These results suggest a strong link between SNPs at TF binding sites in relevant tissues and gene regulation.

To detect motif occurrences that could be altered by the presence of alleles not in the reference genome, we developed a personalized phased SNP-aware genome motif scanning procedure to determine where a motif occurs (Methods). This allowed us to identify motif instances even when multiple non-reference alleles occur within a few base pairs of each other. We observed significant enrichment for islet cis-eQTLs in the set of TF footprint motifs identified only from this haplotype phase-aware scanning approach (that is, the motifs are missed even when a single SNP-aware motif scanning approach is used) in islet samples (p-values: Islet1=4.6×10^{-5}, Islet2=4.0×10^{-7}; Islet-intersect=0.0014; 3.10). Given the informative chromatin accessibility allelic analyses in recent studies (20, 21), we next asked if we could recreate known TF position weight matrices (PWMs) (Fig 3.8D, row 1) based on the allele-specific bias at heterozygous SNPs within TF footprint motifs. To do this, we identified every heterozygous site in a given TF footprint motif, calculated the allelic bias in ATAC-seq signal at these positions, and retained all SNPs with significant bias (Fig. 3.8D, row 3; Methods). We constructed the PWM using the degree of allelic bias for the over-represented alleles (Fig. 2D, row 2). This allelic bias-based PWM (Fig. 3.8D row 2) closely matched the canonical PWM for the corresponding TF (Fig. 3.8D, row 1), providing an in vivo verification of the cognate PWM. We consider this a genetically reconstructed PWM. There was a larger difference in the PWM score for the two alleles of allelic bias SNPs than for the two alleles of matched 1000G SNPs occurring in the same motif (RFX p-value=0.018; CTCF p-value=0.023) (3.11). To further verify that the allelic bias-based genetically reconstructed PWMs were not

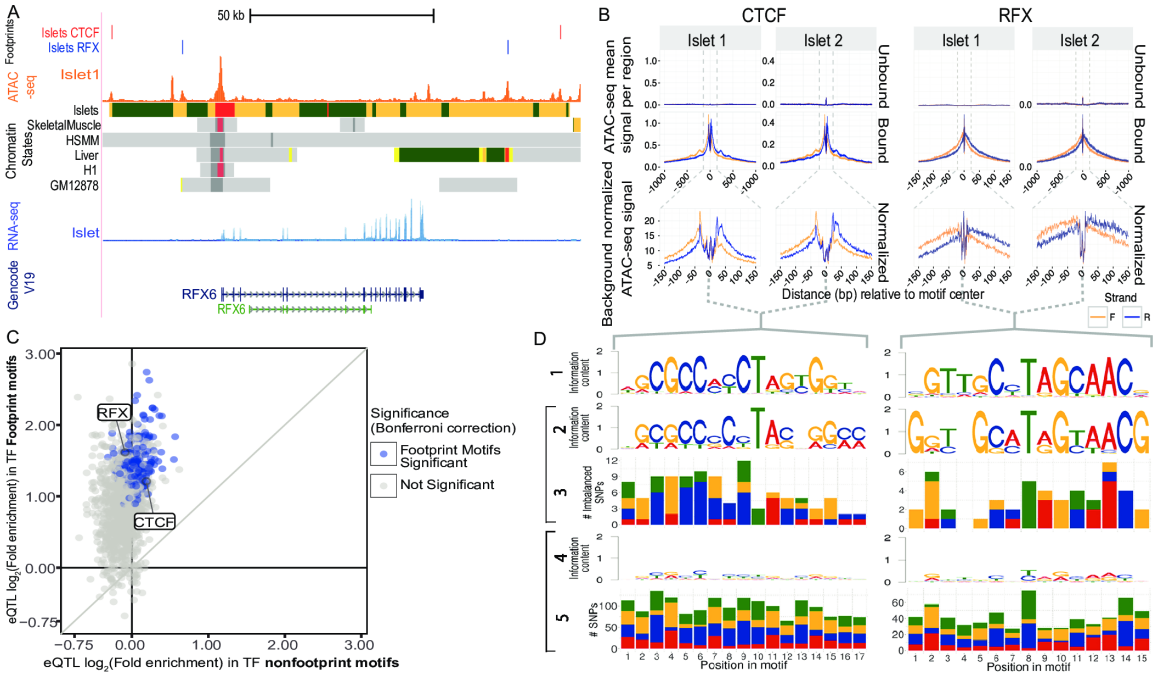


Figure 3.8: Nucleotide resolution islet ATAC-seq profiling nominates regulatory mechanisms. (A) RFX6 locus with expression (RNA-seq), chromatin states, open chromatin (ATAC-seq), and footprints for CTCF and RFX in islets. (B) Density plots indicating normalized sequence coverage of ATAC-seq from two human islet samples at sites overlapping CTCF (motif = CTCF_known2) and RFX (motif = RFX2_4) motifs. (C) Log twofold enrichment of islet cis-eQTLs in TF footprint motifs compared with their enrichment in TF nonfootprint motifs. TFs for which footprint and nonfootprint motifs overlap four or more eQTL SNPs are shown. Blue shows significant enrichment in footprints only (Bonferroni corrected $P < 0.05$). No significant enrichment was observed in any TF nonfootprint motif. (D) Reconstruction of CTCF (motif = CTCF_known2) and RFX (motif = RFX2_4) motifs using ATAC-seq TF footprint allelic bias data. Row 1: original motif PWM. Row 2: PWM genetically reconstructed using the overrepresented alleles (and extent of overrepresentation) for SNPs with significant ATAC-seq allelic bias. Row 3: count of nucleotides in SNPs with significant allelic bias. Row 4: PWM reconstructed using the count of nucleotides for heterozygous SNPs in the TF footprint. Row 5: count of nucleotides in heterozygous SNPs in the TF footprint.

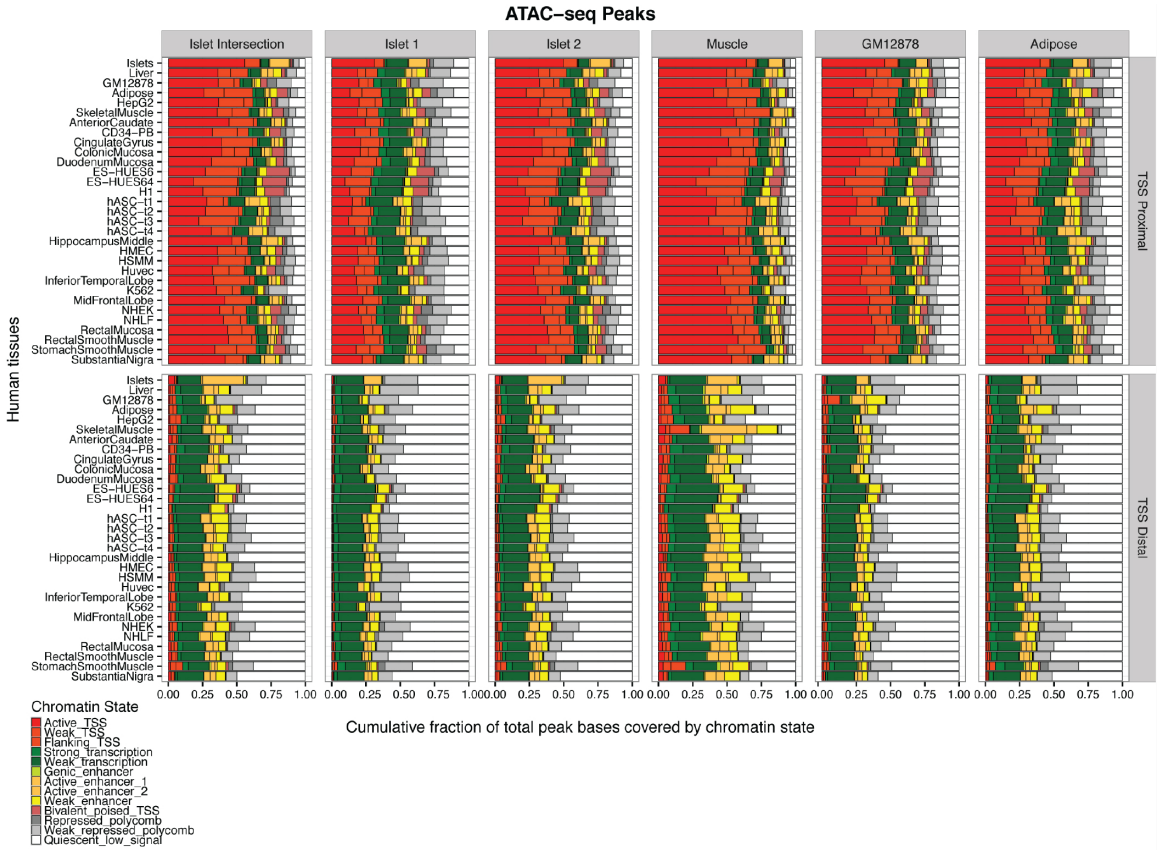


Figure 3.9: Enrichment of islet, muscle, GM12878, and adipose ATAC-seq peaks (columns) in chromatin states across diverse tissues (y axis). Consensus (islet intersection) and individual (islets 1 and 2) islet ATAC-seq peaks show enrichment for active chromatin states in islets, which is more pronounced at TSS-distal (>5 kb from TSS) regions. Muscle (column 4), GM12878 (column 5), and adipose (column 6) ATAC-seq peak calls show similar trends with chromatin states from matched tissues. Note that TSS-distal ATAC-seq peaks from the islet intersect dataset overlap islet active enhancers more than any other chromatin state in islets. Note also that the level of islet enhancer overlap is larger than enhancer overlap in any other tissue.

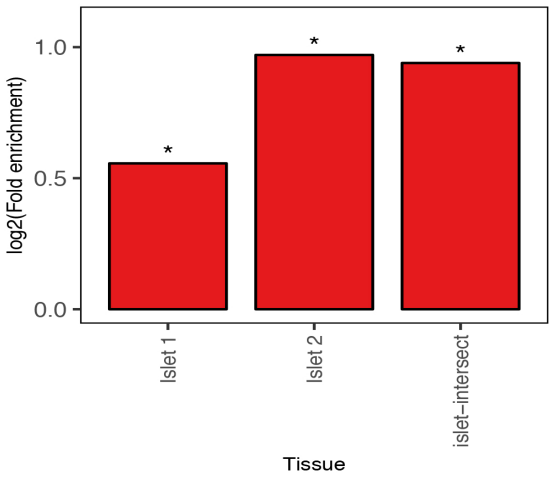


Figure 3.10: Enrichment of islet cis-eQTLs (5% FDR) in ATAC-seq TF footprints that are only detected using phased SNP-aware scans (Materials and Methods). ∗ P <0.05 from GREGOR analysis.

simply reflecting the allelic composition of SNPs in the motifs, we constructed the PWMs using the allele count for all TF footprint heterozygous SNPs observed at each position (where each observed SNP contributed two alleles) and found the resulting PWMs had little information and little similarity to the cognate motifs used to scan across the genome (Fig. 3.8D, rows 4 and 5). Collectively, these results reinforce the potential of ATAC-seq and allelic footprinting analyses to identify relevant and potentially causal TF binding changes in the genetic control of gene expression.

3.3.4 T2D GWAS loci are enriched in RFX footprints and T2D risk alleles disrupt the motifs at independent locations

Given the strong enrichment for islet cis-eQTL in diverse TF footprints, we next sought to identify T2D GWAS SNPs that might regulate gene expression by modulating TF binding. We found that T2D-associated SNPs were significantly enriched in islet RFX TF footprints (p-value range 1.5×10^{-5} - 8.3×10-9 for five RFX TF family members fold enrichment range 7.46 - 30.06 (Fig. 3.12A, Dataset S4)). In contrast, we did not see significant enrichment of T2D associated SNPs in islet non-footprint

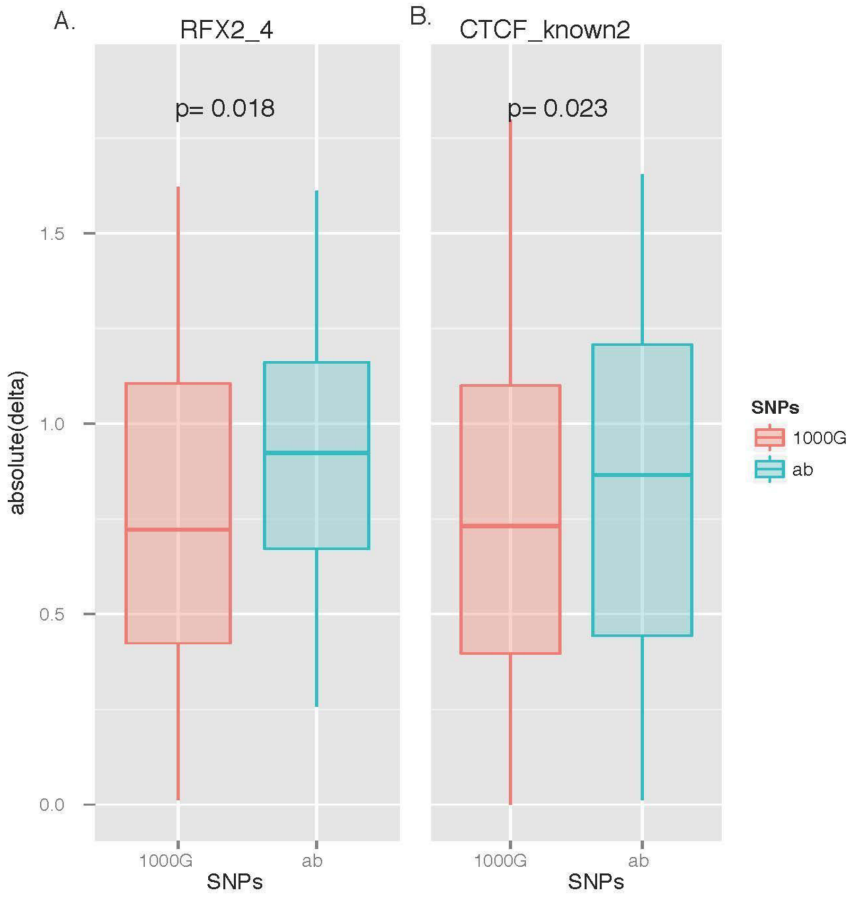
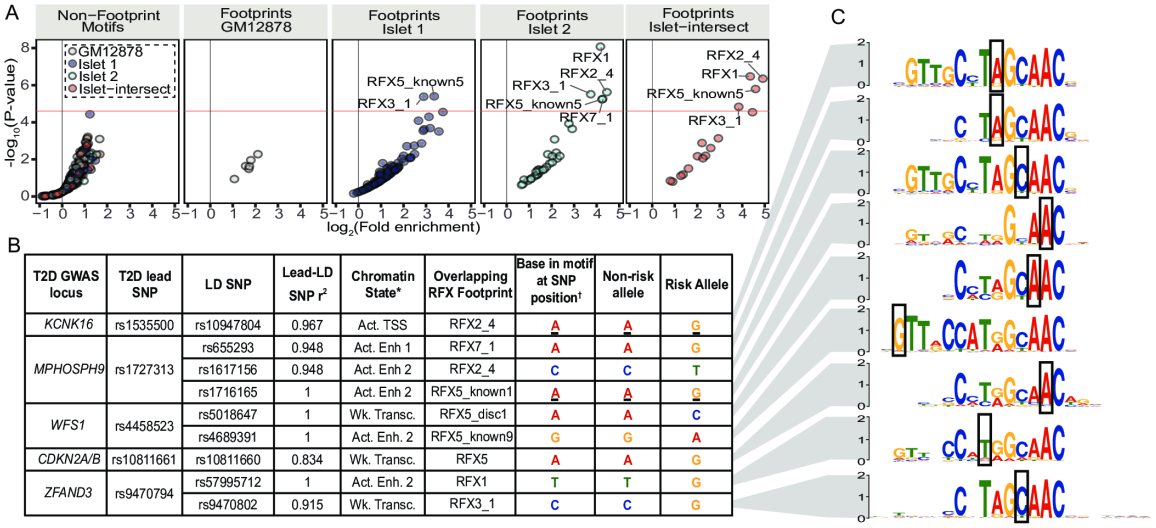


Figure 3.11: SNPs that show allelic bias in ATAC-seq data (ab; blue box plot) exhibit larger effects on the predicted TF binding site motifs compared with randomly sampled 1000G SNPs (1000G; red box plot) overlapping the same footprint in islet 1. The y axis shows absolute value of the delta score [delta = $\log_{10}$(FIMO P value of alternate sequence) ($\log_{10}$(FIMO P value of reference sequence))]. P values of the comparisons were determined by the Wilcoxon rank sum of test. (A) Footprints motif = RFX2_4. (B) Footprint motif = CTCF_known2.

RFX TF motifs (Methods) or in GM12878 TF footprints (Fig. 3.12A). The RFX family of transcription factors recognize X-box motifs and have highly evolutionarily conserved DNA binding domains (22), which may explain why similar motifs from many RFX family members are enriched. Gaulton et al. (2015) observed enrichment of T2D GWAS SNPs in FOXA2 ChIP-seq binding sites in islet tissues (23). We observed 13.5 fold enrichment of T2D-associated SNPs in islet FOX TF footprints ($p=2.8\times10^{-5}$, slightly less significant than the Bonferroni threshold of 2.5×10^{-5}, Dataset S4).

Studies of autoimmune disease have found that disease associated variants often occur near but not in TF motifs (24). We therefore asked if T2D-associated SNPs were enriched in regions flanking RFX footprints motifs (excluding footprint itself) (n=22). We found that regions flanking RFX footprint motifs were enriched for T2D associated SNPs and that the enrichment decreased with increasing distance from footprint motif (Fig. 3.13). The flanking enrichment was lower than in the RFX TF footprints. In contrast, we did not see enrichment of T2D associated SNPs in, non-footprint RFX TF motifs or in the regions flanking the non-footprint RFX TF motifs (Fig. 3.13).

We next assessed the potential effects of the risk and non-risk alleles for 9 T2D associated SNPs at 5 independent loci on RFX TF binding (Methods) (Fig. 3.12B). For each SNP, the non-risk allele was the highest probability nucleotide in the RFX position weight matrix (PWM), and thus the risk allele was predicted to disrupt the motif (Fig. 3.12B and C black boxes). At two of the five loci, the T2D GWAS risk alleles were associated with significantly increased gene expression in our conditional eQTL analysis: KCNK17 (KCNK16 locus, Fig. 3.1B, C, E) and ABCB9 (PITPNM2 locus, Fig. 3.1C). Other loci might have not been detectable as eQTLs due to state specific regulation or small effect sizes. The observation that T2D risk alleles at multiple loci confluently disrupt RFX motifs provides a hypothesis that could explain



T2D GWAS locus	T2D lead SNP	LD SNP	Lead-LD SNP r^2	Chromatin State*	Overlapping RFX Footprint	Base in motif at SNP position†	Non-risk allele	Risk Allele
KCNK16	rs1535500	rs10947804	0.967	Act. TSS	RFX2_4	A	A	G
MPHOSPH9	rs1727313	rs655293	0.948	Act. Enh 1	RFX7_1	A	A	G
		rs1617156	0.948	Act. Enh 2	RFX2_4	C	C	T
		rs1716165	1	Act. Enh 2	RFX5_known1	A	A	G
WFS1	rs4458523	rs5018647	1	Wk. Transc.	RFX5_disc1	A	A	C
		rs4689391	1	Act. Enh. 2	RFX5_known9	G	G	A
CDKN2A/B	rs10811661	rs10811660	0.834	Wk. Transc.	RFX5	A	A	G
ZFAND3	rs9470794	rs57995712	1	Act. Enh. 2	RFX1	T	T	G
		rs9470802	0.915	Wk. Transc.	RFX3_1	C	C	G

Figure 3.12: T2D GWAS enrichment at islet footprints reveals confluent RFX motif disruption. (A) T2D GWAS SNPs are significantly enriched in RFX motifs in islet footprints but not in control motifs or footprints from a nondisease-relevant cell type (GM12878). TF motifs for which footprints overlap four or more T2D GWAS SNPs are shown. The red line indicates Bonferroni multiple testing threshold. (B) T2D-associated SNPs that overlap high information content (>1 bit) positions in RFX motifs. The highest scoring RFX footprints are reported for each T2D GWAS SNP. Act. Enh., active enhancer; Act. TSS, active TSS; Wk. Transc., weak transcribed. ∗ Chromatin-state annotation overlapping the SNP. † Because RFX motifs in C are organized by alignment to the longest RFX3_1 motif, motifs overlapping rs10947804 and rs1716165 correspond to the reverse complement sequence. Therefore, risk and nonrisk alleles are also reported as reverse complement relative to the plus strand sequence. (C) Alignment of highest scoring RFX footprint at each SNP; the boxes indicate the SNP overlap positions. Note that, in every case, the risk allele disrupts that motif.

the mechanism of a subset of T2D associated variants.

3.4 Discussion

We have integrated genome, epigenome, and transcriptome variation and created maps to better understand the genetic control of islet gene expression. Comparison of these maps with T2D GWAS SNPs has helped identify potential disease mechanisms. For example, the risk allele of the coding SNP rs1535500 has been implicated to

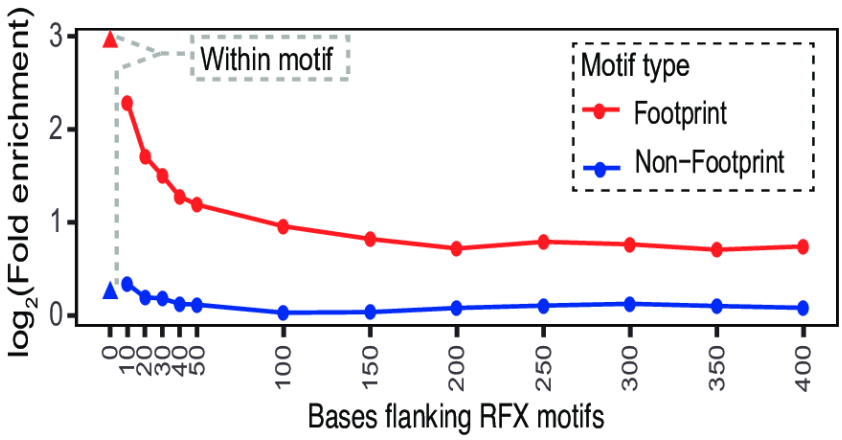


Figure 3.13: Enrichment for T2D GWAS SNPs in regions flanking merged RFX footprint (red) and nonfootprint (blue) motifs.

increase KCNK16 activity and cell surface localization in a mouse model [197]. Other risk alleles in SNPs in high LD with rs153550 are associated with increased expression of the neighboring potassium channel gene KCNK17. KCNK16 and KCNK17 are two pore domain background K+ channels, members of the TWIK related alkaline pH activated K+ channel (TALK) family [54, 106]. Both genes are expressed in islets with high specificity (KCNK16 iESI = 0.98; KCNK17 iESI = 0.76). KCNK16 has been implicated in regulating electrical excitability and glucose stimulated insulin secretion (GSIS) [197]. The KCNK17 gene is not present in the mouse genome. It is possible that the T2D risk haplotype at this locus may have multiple effects that collectively disrupt beta cell K+ signaling and glucose stimulated insulin secretion by simultaneously over-activating KCNK16 and over-expressing KCNK17.

We find that T2D GWAS associated SNPs are significantly enriched in RFX TF footprint motifs and are almost significantly enriched in FOX TF footprint motifs. We find consistent disruption of islet RFX footprint motifs by T2D risk alleles, including at the KCNK17 locus. Lizio et al. found that knockdown of RFX6 results in increased expression of KCNK17 [104] which is consistent with the T2D risk allele disrupting TF binding and increasing target gene expression. At other T2D GWAS loci, such as the

MPHOSPH9 locus, (index SNP rs1727313), two or three T2D GWAS SNPs in high LD are each predicted to have risk alleles that coordinately disrupt independent RFX footprint motifs (Fig. 3.12B, C). We and others [142, 29, 59] previously described the presence of multiple SNPs in enhancers at individual GWAS loci. Our current results build on this concept to include the possibility of multiple confluent disruptions of similar TF motifs in the same locus. Collectively, these results indicate that T2D risk may in part be propagated through genetic modulation of RFX binding in islets. Indeed, our study shortlists only a subset of T2D associated variants as candidates which should be functionally dissected in vivo.

Among the RFX TFs, RFX6 is expressed in islets with high specificity (iESI = 0.94; 3.14) and is involved in pancreatic progenitor specification, endocrine cell differentiation, maintaining beta cell functional identity, and controlling glucose homeostasis [212, 167, 169]. Beta cell-specific deletion of RFX6 results in impaired insulin secretion [145, 21]. Individuals that are heterozygous for a frameshift mutation in RFX6 have increased 2-hour glucose levels [75]. Importantly, rare autosomal recessive mutations that alter DNA-contacting amino acids in the DNA binding domain of RFX6 result in Mitchell-Riley syndrome, which is characterized by neonatal diabetes [167]. Although RFX6 was not in our motif database, a recent report found it to be highly similar to the other RFX family motifs [104], consistent with the expectation for highly conserved DNA binding domains [2]. Our findings could represent a novel connection between rare coding variation in the islet master TF RFX6 [169, 145] and common non-coding variation in multiple target sites for this TF. The impact of these variations mirror the expected effect on organismal physiology, with coding variants that result in neonatal diabetes and non-coding variants that result in later onset T2D. This study for the first time implicates impaired RFX-dependent transcriptional responses in genetic susceptibility to T2D and nominates new mechanistic hypotheses about the molecular genetic pathogenesis of this complex disease.

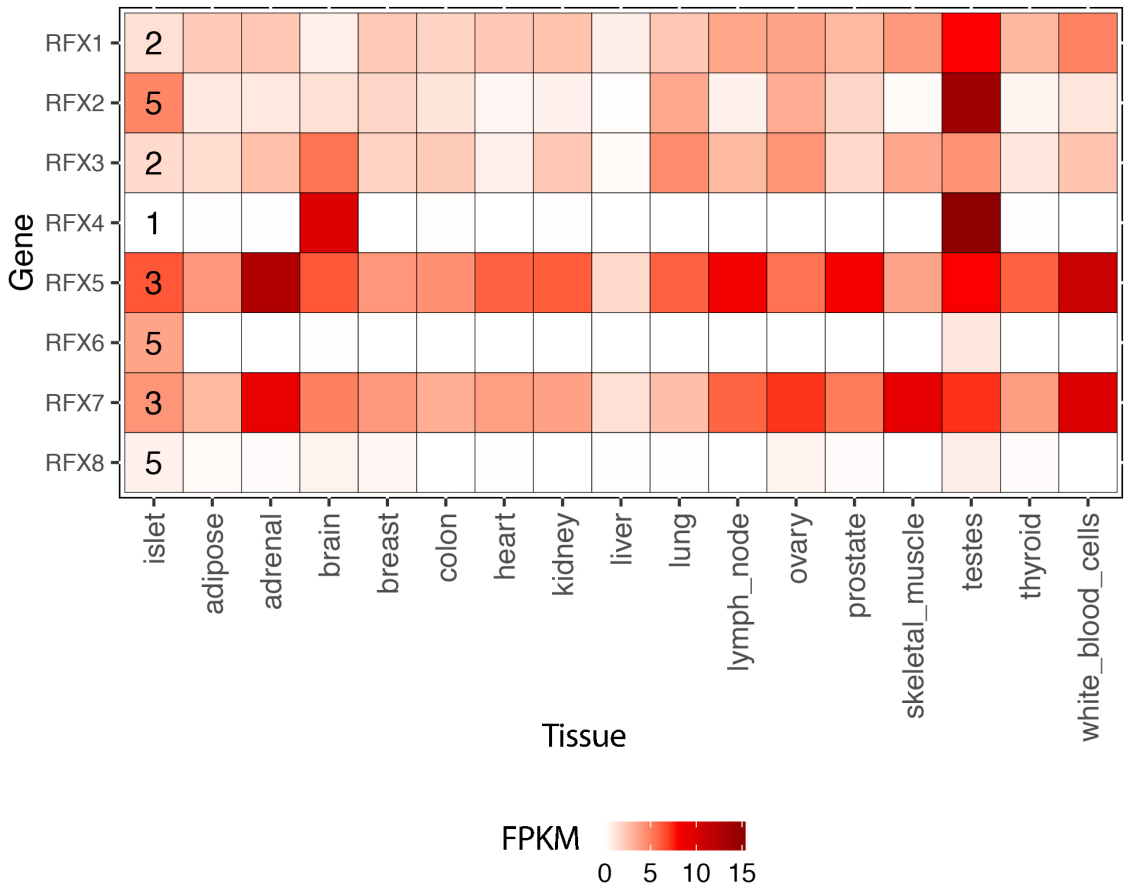


Figure 3.14: *RFX* gene expression (FPKM) across islets and 16 Illumina Body map 2.0 tissues. The iESI quintile for each *RFX* gene is labeled in the islet columns. RFX6 has the highest iESI (0.94) among all *RFX* TF genes.

Following up on the reported loci to functionally validate this hypothesis could help in better understanding T2D mechanisms. Given that most other GWAS SNPs are non-coding, this approach could be used to identify other master TF and multiple target site relationships.

3.5 Materials and Methods

Islet procurement and processing We procured the islet samples used in this study from the Integrated Islet Distribution Program (IIDP), the National Disease Re-

search Interchange (NDRI), or ProdoLabs. 31 islets (age=45.3 +/- 11.9 years; 52% male; BMI=27.2+/-4.3 kg/m2) passed mRNA-seq and genotype QC steps (see below). Islets were shipped overnight from the distribution centers. Upon receipt, we pre-warmed islets to 37 C in shipping media for 1-2 hours prior to harvest. Approximately 2500-5000 islet equivalents (IEQ) from each organ donor were harvested for RNA isolation. We transferred 500-1000 IEQ to tissue culture-treated flasks and cultured as in (1); genomic DNA isolated from islet explant cultures was used for genotyping.

3.5.1 SNP genotyping, sample and genotype QC

Genomic DNA was genotyped at the Genetic Resources Core Facility (GRCF) of the Johns Hopkins Institute of Genetic Medicine on the HumanOmni2.5-4v1_H BeadChip array (Illumina, San Diego, CA, USA): minimum call rate was 97.14%. We mapped the Illumina array probe sequences to the hg19 genome assembly using BWA. We excluded SNPs with ambiguous probe alignments, SNPs with 1000 Genomes phase 1 variants with minor allele frequency of $\geq$ 1% or greater within 7bp of the 3' end of probes, or call rates <95%. All alleles were oriented relative to the reference.

We identified no individuals with $\geq$ 3rd degree relatedness using KING [112]. We performed principal components analysis using PLINK 1.9(`http://www.cog-genomics.org/plink2/general_usage`) on 60,714 SNPs with minor allele count (MAC) >5 and r2 <0.2, after excluding SNPs from regions of high LD [151]. 33 self-reported Caucasian samples and one sample of unknown ethnicity clustered together by principal components analysis (PCA). One sample self-reported to be of Caucasian ancestry did not cluster with the others and was excluded for eQTL analyses.

Fadista et al. graciously provided genotypes of their 89 islet samples for the Illumina HumanOmniExpress 12v1C BeadChip [44]. We processed the probes and genotypes as described above. We identified no individuals with $\geq$ 3rd degree relat-

edness. We performed PCA as described above using 86,502 LD pruned SNPs. All 89 samples clustered together in the PCA analysis.

3.5.2 RNA isolation, mRNA-seq library preparation and mRNA sequencing

We extracted and purified total RNA from 2000-3000 islet equivalents (IEQ) using Trizol (Life Technologies). RNA quality was confirmed with Bioanalyzer 2100 (Agilent); samples with RNA integrity number (RIN) >6.5 were prepared for mRNA sequencing. We added External RNA Control Consortium (ERCC) spike-in controls (Life Technologies) to one microgram of total RNA. We generated PolyA+, stranded mRNA RNA-sequencing libraries for each islet using the TruSeq stranded mRNA kit according to manufacturers protocol (Illumina). Each islet RNA-seq library was barcoded, pooled into 12-sample batches, and sequenced over multiple lanes of HiSeq 2000 to obtain an average depth of 100 million 2 x 101 bp sequences.

3.5.3 mRNA-seq processing and QC

We retained RNA-seq reads passing the Illumina chastity filter and mapped reads to a reference sequence composed of ERCC control fragments and all chromosomes and contigs from hg19, excluding alternate haplotypes, replacing the mitochondrial sequence (chrM) with the Cambridge Reference Sequence, and masking the pseudoautosomal region on chromosome Y. We aligned reads using STAR (version 2.3.1y) [37] with default parameters and a splice junction catalog based on Gencode v19 [61]. Non-uniquely mapping reads and read pairs with unpaired alignments were discarded. Duplicate read pairs, i.e. those mapping to the same coordinates, were retained.

RNA-seq QC was performed at the level of readgroups (i.e. a library on a lane) using QoRTs [62]. We inspected the comprehensive set of QC metrics generated by QoRTs for outlying libraries, lanes, and sequencing runs. We used the 92 ERCC RNA

spike-in controls and in-house scripts to assess library quality and batch effects, and to check the accuracy of the strand-specific protocol [76]. We also performed a principal components analysis (PCA) on the matrix of expression data. The QoRTs and PCA processes revealed two outlying sample libraries. One sample showed extreme 3 bias in gene body coverage, and the other showed low gene diversity and was a strong outlier by PCA. In addition, we excluded one sample that was reportedly Caucasian, but was an outlier in the genotype PCA (see above). These three libraries were removed, leaving 31 islet samples for analysis.

To confirm sample identity and check for contamination, we compared SNP chip genotypes to RNA-seq alignments in annotated exonic regions using verifyBamID [78], using the –maxDepth 100 option to avoid having highly-expressed genes bias the estimate of contamination. No sample showed contamination >0.78%.

We aligned the non-strand-specific RNA-seq reads from Fadista et al. [44] with the same version of STAR to the same hybrid reference genome. Again, we discarded non-uniquely mapped reads and read pairs with unpaired alignment, and we retained duplicate pairs. We performed QC using QoRTs, and PCA of the expression data, as described above, and identified one outlier library. In addition, comparison of SNP chip genotypes to RNA-seq alignments with verifyBamID identified two swapped samples, and five samples that had greater than 2% estimated contamination in the RNA-Seq sample. We removed all eight samples, leaving 81 samples to be analyzed.

3.5.4 Expression quantification

To study regulatory variation, we performed analyses at the gene level. Definitions for all transcriptome features were based on GENCODE v19 [61], which annotates a total of 57,820 genes: 20,345 protein-coding, 13,870 long non-coding RNAs, and 14,206 pseudogenes. We ignored pseudogenes for all downstream analyses. We counted fragments mapping to genes using htseq-count v0.5.4 [6] (`http://www-`

huber.embl.de/users/anders/HTSeq/doc/count.html) and calculated FPKM values for each gene.

We processed data from Fadista et al. in the same way, except that counts of genes were performed in a non-strand-specific manner, consistent with the RNA-seq libraries.

3.5.5 Imputation

We excluded SNPs with: MAC <1, Hardy-Weinberg equilibrium p-value $< 10^{-6}$, absolute alternate allele frequency difference >0.2 compared to the 1000G EUR sample and A/T or C/G SNPs with MAF >0.2. This left 2,057,703 autosomal SNPs for subsequent imputation. We performed autosomal SNP imputation using a two-step strategy [71] with the haplotypes from 1000G phase3 v5 [179] as the reference panel. To improve phasing quality given the small number of islet samples, we pre-phased our islet samples together with the 2,504 reference panel samples using ShapeIT version 2 [33]. We then imputed genotypes with Minimac2 [48]. We retained 8,377,422 imputed variants with a MAC $\geq$ 1 and r2 $\geq$ 0.3.

For the 81 Fadista et al. islet samples, we removed SNPs, and prephased and imputed genotypes as described above. We used 692,118 SNPs for imputation. We retained 9,758,857 imputed variants with a MAC $\geq$ 1 and r2 $\geq$ 0.3.

3.5.6 cis-eQTL meta-analysis

We performed separate cis expression quantitative trait (cis-eQTL) analysis for our islets (n=31) and Fadista et al. islets (n=81) and combined the results using meta-analysis. We performed PCA (using the same procedure described in the genotype QC section) separately on these two sets of samples. We considered for analysis 6,060,203 SNPs that were present in both studies and had a combined MAC $\geq$ 10. We tested SNPs within 1 Mb of the most upstream TSS of each gene using Matrix eQTL [165].

We included in the analysis 19,360 genes present in both sets of samples (out of 26,845 genes present in our islets and 19,650 present in Fadista et al. islets). For individual i and gene j, to generate the gene expression value Y_{ij}, we inverse normalized $FPKM_{ji}$ for each gene j. We then performed factor analysis via PEER [170, 171] on the inverse normalized FPKM (specifying from 1 to 60 factors to optimize the detection of cis-eQTLs (below) and including age, sex, the top 2 genotype-based principal components and and for our islet samples only experimental batch , as covariates in the model), and inverse normalized the resulting residuals. We used the linear regression model with an additive genetic effect:

$$Y_{ij} = \alpha + \beta_{js} G_{is} + \epsilon_{ij}$$

where is the intercept, G_{is} is the imputed allele count for SNP s for individuals i, $_{js}$ is the regression coefficient of the imputed allele count for SNP s on transformed gene expression Y_{ij} and $_{ij}$ is a normally distributed error term with mean 0 and variance σ^2.

We used false discovery rate (FDR) [174] to account for multiple testing and considered as significant associations with FDR $\leq$ 5%. We expect that removing technical and biological variation via PEER will increase power to detect cis-eQTLs [162]. For each study we report results using number of PEER factors that maximized the number of eQTLs on chromosome 20 at FDR $\leq$ 5%: 30 for our islets and 32 for Fadista et al.

For each SNP-gene pair we combined the results from our islet samples and Fadista et al. using a sample-sized weighted meta-analysis [205] and report p-values based on this analysis. In addition, we performed a fixed-effects inverse variance-weighted meta-analysis [205] and report eQTL effect sizes from this analysis. We do not report p-values from this analysis as we found the p-values were consistently inflated. We present results for SNPs present in both studies (MAC$\geq$1) and with MAC $\geq$ 10 in

the combined study.

Gene-based cis-eQTLs for GWAS variants for T2D and related traits We compiled a list of 225 SNPs with p-value <5 x 10-8 in GWAS (GWAS SNPs) for T2D, fasting glucose, fasting glucose adjusted for BMI, fasting insulin, fasting insulin adjusted for BMI, 2-hr glucose, 2-hr glucose adjusted for BMI, and fasting proinsulin from the NHGRI GWAS catalog [203] and carried out manual curation of the literature to create a comprehensive list that was up-to-date as of May 2014. Of these 225 GWAS SNPs, 214 were tested in our cis-eQTL analysis for a total of 3,995 GWAS SNP-gene pairs. To identify GWAS variant cis-eQTLs that may be independent of other stronger cis-eQTLs for the same gene, we performed iterative conditional analysis on each of the 3,995 GWAS SNP-gene pairs. For each GWAS SNP-gene pair and study we used the linear regression model with an additive genetic effect:

$$Y_{ij} = \alpha + \beta_{jGWAS} G_{iGWAS} + \epsilon_{ij}$$

where G_{iGWAS} is the imputed allele count for the GWAS SNP for individuals i, $_{jGWAS}$, is the regression coefficient of the imputed allele count for the GWAS SNP, and G_{is} is the set of SNPs within 1Mb of the most upstream TSS. We combined the results from our samples and Fadista et al. islets using meta-analysis as described above. If ≥ 1 SNP had a meta-analysis p-value $<$ 1.2 x 10-4 (corresponding to the p-value threshold for gene-based cis-eQTLs with FDR <5%) we retained the SNP with the most significant p-value in the model and repeated the procedure until no added SNP had a p-value $<1.2 \times 10^{-4}$. This procedure corresponds to performing step-wise forward selection of SNPs within 1Mb of the most upstream TSS based on the results of the meta-analysis at each step (using a stopping threshold p-value of 1.2×10^{-4}). The conditional p-value for a given GWAS SNP is the p-value for βtextsubscriptjGWAS from the final model. We considered as significant conditional associations with FDR $\leq$ 5% based on the 3,995 GWAS SNP-gene pairs.

3.5.7 Functional validation of eQTL variant activity and direction of effect

We maintained the MIN6 mouse insulinoma beta cell line [128] as previously described [89]. We amplified a 473-bp genomic region containing rs10947804, rs12663159, rs146060240 and rs34247110 from human DNA (primers: 5'-GCCAGGTAAGCCAGGTA-3' and 5'-GAGTGCGGTTTCCAGAAGTC-3') and cloned it into the pGL4.10 promoterless vector (Promega) as previously described [89]. The region was cloned in the forward orientation with respect to KCNK17 transcription and includes the promoter, 5UTR, and the first 34 codons of KCNK17. We performed site-directed mutagenesis with the QuikChange Lightning Site-Directed Mutagenesis Kit (Agilent) to change the KCNK17 start codon from ATG to AGG to prevent interference with translation and function of the luciferase protein. We amplified the KCNK17-increasing and decreasing haplotypes. The haplotype of alleles associated with higher KCNK17 expression (risk haplotype) includes rs10947804-C, rs12663159-A, rs146060240-G, and rs34247110-A. The haplotype associated with lower KCNK17 expression (non-risk haplotype) includes rs10947804-T, rs12663159-C, rs146060240-deletion, and rs34247110-G. We performed luciferase assays as previously described [89]. We plated 200,000 cells per well in a 24-well plate and transfected after 24 hours. We co-transfected 250 ng of haplotype plasmid and 80 ng Renilla plasmid in duplicate wells using Lipofectamine LTX (Life Technologies) and assayed luciferase activity 48 hours post-transfection. For the KCNK17-increasing haplotype, we transfected 4 independent clones and for the KCNK17-decreasing haplotype we transfected 5 independent clones. Each independent clone was transfected in duplicate. Quantified luciferase activity was normalized to empty vector (EV) and we tested for difference in luciferase activity between haplotypes using a two-sided t-test. We observed the same transcriptional effect in three separate experiments.

3.5.8 Analysis of islet-specific expression

We used an information theory approach [161, 64] to score genes based on islet expression level and specificity relative to the panel of 16 diverse Illumina Human Body Map 2.0 tissues. We first calculated expression (x) in FPKM values for all Gencode v19 genes across a representative islet sample and each of the 16 tissues in the Body Map 2.0 data. We calculated the relative expression of each gene (g) in islets compared to all 17 tissues (t) as p:

$$p_{g,islet} = \frac{x_{g,islet}}{\sum_{n=1}^{17} x_{g,t}}$$

We next calculated the entropy for expression of each gene across all 17 tissues as H:

$$H_g = -\sum_{n=1}^{17} p_{g,t} log_2(p_{g,t})$$

Following previous studies (24, 25), we defined islet tissue expression specificity (Q) for each gene as:

$$Q_{g,islet} = H_g - log_2(p_{g,islet})$$

To aid in interpretability, we divided Q for each gene by the maximum observed Q and subtracted this value from 1 and refer to this new score as the islet expression specificity index (iESI):

$$iESI_g = 1 - \frac{Q_{g,islet}}{Q_{max,islet}}$$

iESI scores near zero represent lowly and/or ubiquitously expressed genes and scores near 1 represent genes that are highly and specifically expressed in islets. We divided genes and subsequently the associated eQTL variants associated into quintiles

as shown in Fig. 3.5, 3.6, 3.7E) based on the iESI score. The division by quintile provided an average sample size of 618 lead eQTL variants in each bin, which we then used to compute enrichments in genomic features. In our previous analysis of skeletal muscle, we used a decile approach as we detected lead eQTLs in >90% of testable genes [162], whereas here we detected a lower number of eQTLs, 3,964, and thus used a quintile approach. To depict a higher-resolution partitioning of genes based on the iESI score, we used a deciles in the interactive eQTL browser (`http://theparkerlab.org/tools/isleteqtl/`).

3.5.9 Chromatin state analyses

We collected cell/tissue ChIP-seq reads for H3K27ac, H3K27me3, H3K36me3, H3K4me1, and H3K4me3, and input from a diverse set of publically available data [186, 43, 125, 142]. Collectively, these data represent 31 cells/tissues (shown in Fig. S6C), as well as 8 additional human and rodent datasets included for other ongoing projects. We performed read mapping and integrative chromatin state analyses in a manner similar to our previous reports (20, 29), and followed quality control procedures reported by the Roadmap Epigenomics study [186]. Briefly, we mapped reads using BWA [97] (version 0.5.8c), removed duplicates using samtools, and filtered for mapping quality score of at least 30. To assess the quality of each dataset, we performed strand cross correlation analysis using phantompeakqualtools (v2.0, `http://code.google.com/p/phantompeakqualtools`) [186]. To select cells/tissues for ChromHMM to learn chromatin states, and following the Roadmap Epigenomics practices, for each tissue we performed QC on the most well-defined peak datasets, H3K27ac and H3K4me3. We required each of these two marks within a tissue/cell to have normalized strand cross-correlation (NSC) > 0.8 and relative strand cross-correlation (RSC) score >1.1. Islets and 32 other cell/tissue types out of 39 passed this criteria. The failed samples are consistent with the Roadmap Epigenomics study

analyses: the 5 brain tissues and ES-HUES64 did not pass this criteria. To more uniformly represent data sets with different sequencing depths, we randomly subsampled each data set containing >20M mapped reads to a depth of 20M. Chromatin states were learned jointly from the 33 cell/tissues that passed QC by applying the ChromHMM (version 1.10) hidden Markov model (HMM) algorithm at 200 bp resolution to the 5 chromatin marks and input [43, 41, 40]. We ran ChromHMM with a range of possible states, and selected a 13 state model as it most accurately captured information from higher state models and provided sufficient resolution to identify biologically meaningful patterns in a reproducible way. We have used this state selection procedure in previous analyses [162, 142]. To assign biological function names to our states that are consistent with previously published states, we performed enrichment analyses in ChromHMM comparing our states to the states reported by Roadmap Epigenomics (in their extended 18 state model) [186] for 18 matched cells/tissues (Fig. S1). We assigned the name of the Roadmap state that was most strongly enriched in each of our states. We then applied our chromatin state model to obtain chromatin state segmentations for the 6 cell/tissue types which were not used to learn the model using ChromHMM MakeSegmentation.

3.5.10 Clustering by enhancer states across tissues

To identify patterns of active enhancer chromatin state calls across cell and tissues, we performed k-means clustering using 200bp genomic windows where ChromHMM posterior probability for active enhancer state 1 or 2 was greater than 0.95 in at least one cell/tissue type used in this study. We identified an optimal number of clusters by plotting the within group sum of squares versus number of clusters for a range of k, and selected k=60 which corresponded to the elbow in the plot. We performed k-means clustering using the Hartigan-Wong algorithm with 10,000 iterations and 50 random starts.

3.5.11 Overlap of enhancer clusters with stretch enhancers

We called stretch enhancers for all cells/tissues in our chromatin state segmentations as in our previous work [162, 142] by merging adjacent enhancer states (Active Enhancer 1 and 2, Weak Enhancer and Genic Enhancer) in a given tissue and identifying contiguous regions $\geq$ 3kb. We quantified the overlap between each enhancer cluster and stretch enhancers for islets, liver, H1 and GM12878 using the Jaccard statistic in BEDtools [154]. In Fig. 3.7C, we normalized the Jaccard statistic within each column such that the maximum is set to 1.

Enrichment of genetic variants in genomic features We calculated the enrichment of lead islet cis-eQTL or lead T2D GWAS SNPs (including SNPs in r2 $\geq$ 0.8 with the lead SNP (SNPs in LD)) in features such as chromatin states, stretch enhancers, enhancer clusters, transcription factor footprint or non-footprint motifs using GREGOR [160]. TF non-footprint motifs (shown in Fig. 3.8C and 3.12A and B), are defined as TF motifs that are not called as footprints in either of the islet samples. For eQTL enrichment, we included the lead cis-eQTL SNP for genes significant at a given FDR threshold. The enrichment trends were consistent across different FDR thresholds (5%, 1%, and 0.1%), with more stringent sets having slightly more pronounced trends. We report here the results for the FDR $\leq$ 5% set. For T2D GWAS SNP enrichment, we aimed to use independent T2D association signals, i.e. reported lead T2D SNPs that were not in LD with each other. We sorted the list of lead GWAS SNPs (defined in section Gene-based cis-eQTLs for GWAS variants for T2D and related traits) by p-value of association with T2D and sequentially removed SNPs with r2 $>$ 0.2 with a higher ranked SNP.

For each input SNP, 500 control SNPs were selected that matched the input SNP for MAF, distance to the gene, and number of SNPs in r2 $\geq$ 0.8. Fold enrichment is calculated as the number of loci at which the index SNP (or SNP in LD) overlaps the feature over the mean number of loci at which the matched control SNPs (or

SNPs in LD) overlap the same feature. This process accounts for the length of the features as longer features will have more overlap by chance with control SNP sets. We used the following parameters in GREGOR: r2 threshold (for inclusion of SNPs in linkage disequilibrium (LD) with the lead eQTL or T2D GWAS SNP) = 0.8, LD window size = 1Mb, and minimum neighbor number = 500. For both eQTL and GWAS SNP enrichment of TF footprint and non-footprint motifs, we report results for SNP-feature overlaps ≥ 4 to avoid artifacts due to low overlaps.

3.5.12 Open chromatin profiling (ATAC-seq)

We profiled chromatin accessibility in islets from 2 human organ donor samples, which were genotyped using methods identical to the other samples (see above), using the assay for transposase-accessible chromatin-sequencing (ATAC-seq). Approximately 50-100 islet equivalents from each sample were transposed in triplicate following the methods in [17]. ATAC-seq replicates were barcoded and sequenced 2 x 125 bp on a HiSeq 2000 to combined total depths of >831M reads for islet 1 and >585M reads for islet 2.

For each library, we performed read alignment, duplicate removal, and filtering as described in our previous study [162]. We next pooled all replicates for each sample and called peaks using MACS2 (`https://github.com/taoliu/MACS`), version 2.1.0, with flags -g hs –nomodel –shift -100 –extsize 200 -B –broad –keep-dup all, retaining all peaks that satisfied a 5% FDR.

3.5.13 Haplotype-aware PWM scans

To detect potential TFBSs in a haplotype-aware manner, we generated personalized diploid genomes from the phased, imputed genotypes for each of two islet samples, using vcf2diploid (v0.2.6a, [156]). We scanned each haplotype using FIMO with PWMs from ENCODE [83], JASPAR [118], and Jolma et al. [77]. We ran

FIMO using the observed nucleotide frequencies from the hg19 reference (40.9% GC) and the default p-value cutoff (1x10-4). We converted the resulting hits to reference coordinates using chainSwap and liftOver with -minMatch=0.1, and merged the results from the two haplotypes into a single set of results per motif per sample. As an example, for islet 1, this procedure produced a total of 2.16B motif matches from our motif database. Of these, 610,544 (0.0283%) are not detected in a single-SNP aware motif scanning procedure.

3.5.14 ATAC-seq footprints

We used CENTIPEDE [147] to call footprints in the islets ATAC-seq data. Briefly, for each PWM scan result, we built a matrix encoding the number of TN5 integration events at a region 100 bp from each motif occurrence. To increase the amount of information given as input for the algorithm, we split the ATAC-seq signal into three different categories based on the diverse fragment length distribution: 36-149bp, 150-324bp, and 325-400bp. We considered any given motif occurrence bound if both the CENTIPEDE posterior probability was ≥ 0.99 and its coordinates were fully contained within an ATAC-seq peak.

3.5.15 Genetic reconstruction of position weight matrices using ATAC-seq footprint allelic bias data

Previous studies have identified signatures of allelic bias in chromatin accessibility data at TF footprints [119, 134]. Motivated by this, we used the heterozygous genotype calls from our islet ATAC-seq samples and the alleles observed in the reads to quantify allelic bias in regions of open chromatin (ATAC-seq TF footprints). To diminish reference allele mapping bias of our mapped ATAC-seq reads, we used the WASP mapping pipeline and duplicate removal tool [194] (downloaded from GitHub on Feb. 19, 2016). To avoid double-counting alleles that may be covered by each read

in a pair as a result of occurring on a short fragment, we clipped overlapping read pairs using the ClipOverlap function of BamUtil. We included properly paired and mapped reads with mapping quality of ≥ 30 and base quality of ≥ 20. We restricted our analyses to the set of heterozygous SNPs calls within each sample (see above for genotype information). For each SNP we counted the number of reads containing each allele. Because we did not have sufficient statistical power to call allelic bias at SNPs with low coverage, we included only SNPs with $\geq$ 20x coverage to reduce the multiple testing burden. To help protect against mapping artifacts, we excluded SNPs with $\leq$ 2x coverage for either allele.

We used a two-tailed binomial test that accounted for reference allele bias to evaluate the significance of the allelic bias at each SNP in each sample. We estimated the allelic bias expected under the null for each sample and reference-alternate allele pair as previously described [94]. Briefly, for each sample and for each reference-alternate allele pair (e.g. AG and GA are separate ref-alt allele pairs) we calculated the expected fraction of reference alleles (fracRef) as the sum of the reference allele count divided by the sum of the total allele count for SNPs of a given reference-alternate allele pair. To prevent SNPs of high coverage from biasing the fracRef, we downsampled SNPs with coverage in the top 25th percentile to 30x coverage and used the downsampled reference allele and total count. To prevent SNPs of low coverage from biasing the mean fracRef, only SNPs with a total read coverage ≥ 30 were used. We used the observed sample- and allele-pair-specific fracRef as the true fracRef under the null hypothesis of no ASE in the binomial test. We did not test SNPs in regions blacklisted by the ENCODE Consortium due to poor mappability (wgEncodeDacMapability-ConsensusExcludable.bed and wgEncodeDukeMapabilityRegionsExcludable.bed) [9]. We performed the binomial test using Rs binom.test and multiple testing correction using the 'qvalue' command in Bioconductors qvalue R package (version 2.2.2, `http://github.com/jdstorey/qvalue`; [174]. We considered SNPs with q-value $<$

0.05 as having significant allelic bias.

For each motif, we reconstructed the PWM using variants with significant ATAC-seq footprint allelic bias. To create the PWM for each motif, we took all significant allelic bias SNPs at position j with over-represented nucleotide i, and summed their absolute allelic deviations from the adjusted expected fracRef (sample-specific and allele pair-specific, as calculated in the previous section). The resulting matrix of values for nucleotide i at position j is a reflection of the number of allelic biased SNPs of nucleotide i at position j and the unevenness of their imbalance toward nucleotide i. We summed the values in the matrices for the two islet samples and used them to create a PWM, so that the genetically reconstructed motifs represent the combined data from both samples.

As a control, at each motif, we also reconstructed the PWM by summing the counts of nucleotide i at position j for all SNPs (biased + unbiased) in the motif.

3.5.16 Effect of ATAC-seq footprint SNPs with allelic bias on predicted TFBS strength for CTCF and RFX motifs

Given that we were able to reconstruct PWMs with ATAC-seq allelic bias results, we sought to address whether the two alleles from SNPs with significant allelic bias had larger differences in their PWM score than the alleles of randomly chosen SNPs occurring within the same footprints. We calculated PWM scores for the reference and alternate allele version of each sequence using the FIMO tool as described above. For each SNP, we used the FIMO p-value to calculate a SNP effect score (delta) as follows: delta = -$\log_{10}$(p-value of alternate sequence) - (-$\log_{10}$(p-value of reference sequence)). We then measured the delta score for all allelic bias SNPs overlapping a TF footprint for CTCF_known2 and RFX2_4 motifs. We constructed a null set of SNPs by choosing a random set of 1000G SNPs with matching MAF and TSS-distance that also overlap the same footprints. We evaluated the difference in the

absolute (delta score) distributions with a Wilcoxon rank sum test. These results are shown in Supplementary figure S9.

3.5.17 T2D GWAS loci overlap with RFX footprints

We performed enrichment analysis for T2D associated SNPs (T2D GWAS SNP and SNPs in r2 $\geq$ 0.8 with the GWAS SNP) to overlap with TF footprint and non-footprint motifs as described above. We selected TF motifs with less than 100,000 footprint occurrences genome wide in either of the Islet samples or GM12878 to help ensure specificity of binding; 1995 out of 2,870 TF motifs passed this criteria. In Fig. 3.12C, we show T2D associated SNPs that occur at high information content (> 1) positions in their respective RFX PWM. For each shown T2D associated SNP, we used our phased genotype calls to determine the T2D associated SNP risk allele (given the T2D GWAS SNP risk allele). Multiple RFX footprints can be called at the same SNP due to motif similarity; we report the motif from the highest scoring PWM. In Fig. 3.12D we used TOMTOM [60] to align the different RFX motifs using the longest (RFX3_1) as the seed.

3.6 Acknowledgements and publication

The results presented in this chapter have been published in [196]. My specific contributions towards this project were analyzing the epigenomic data including chromatin state, ATAC-seq and TF footprint enrichments for eQTL and GWAS loci and the analyses of T2D GWAS loci overlapping TF footprint motifs. This work was a large team effort, and I am lucky to have had the opportunity to work with such smart, motivated and helpful group of people. I thank Drs. Laura Scott, Mike Erdos and Ryan Welch for their significant contributions towards this work. I also thank my lab mates Ricardo Albanus, Peter Orchard and John Hensley for their contributions for this work. Many thanks to the members of the FUSION team for their feedback

during this project. I especially thank Prof. Steve Parker for his guidance, insight and mentorship throughout this project.

CHAPTER IV

Analyses of Islet eQTL and T2D GWAS Along with Epigenomic Data to Elucidate Gene Regulatory Mechanisms

4.1 Abstract

One of the main difficulties for the identification of functional processes through which genomics regions implicated in complex diseases identified via genome-wide studies (GWAS) resides in the limited access to relevant tissues or the difficulties to define good proxy tissues for genetic studies. We generated expression quantitative trait loci (eQTL) in aggregated published and newly generated human islet RNA-Seq data (n=420), to provide a detailed landscape of the genetic regulation of gene expression in a tissue relevant for Type 2 Diabetes (T2D) development. Thorough integration with eQTLs from GTEx, we report an enrichment of T2D and glycemic GWAS variants associated to beta-cell dysfunction in islets compare to other tissues and variants associated to insulin resistance. The integration with islet chromatin states derived from histone modification data, identified a high proportion (80%) of islet eQTLs overlap in islet ATAC-seq peaks and islet active TSS chromatin states and revealed a relationship between TF footprint motif and the effect sizes of eQTLs. Integrating T2D and glycemic traits GWAS information, we also identify 23 loci with

evidence for co-localization of islet eQTLs, including *TCF7L2*, *HMG20A*, *MADD* and two independent signals at *DGKB*. Together, our findings illustrate the advantages of functional and regulatory studies in tissues relevant for diseases, while expanding our machinist insights into complex traits association loci activity with an expanded list of putative transcripts implicated in T2D development.

4.2 Introduction

Over the past decade, analysis of GWAS data has generated a growing inventory of genomic regions implicated in T2D predisposition and variation in diabetes-related glycemic traits. However, progress in defining the mechanisms whereby these associated variants mediate their impact on disease-risk has been relatively slow. One major reason is that at least 90% of the associated signals map to non-coding sequence. This complicates efforts to connect T2D-associated variants with the transcripts and networks through which they exert their effects [180, 181, 49, 164, 110].

For many common, multifactorial diseases, one valuable approach for addressing this variant-to-function challenge is to use expression quantitative trait loci (eQTL) mapping to characterize the impact of disease-associated regulatory variants on the expression of nearby genes [50]. Demonstrating that a disease-risk variant co-localizes with a *cis*-eQTL signal is consistent with a causal role for the transcript concerned in disease development. Such hypotheses can then be subject to more direct evaluation, for example, by perturbing the gene in suitable cellular or animal models. However, eQTL signals often show marked tissue-specificity. The power to detect mechanistically-informative expression effects is therefore, at least in part, dependent on assaying expression data from sufficient numbers of samples across the range of disease-relevant tissues [50]. Appropriate interpretation of GWAS-eQTL signal co-localization analyses also needs to consider physiological and genomic data which may

point to the specific tissue most likely to be mediating disease-risk at a given locus [110, 111].

The pathogenesis of T2D involves dysfunction across multiple tissues, most obviously pancreatic islets, adipose, muscle, liver, and brain. Risk variants that influence T2D predisposition through processes active in each of these tissues have been reported (e.g. MC4R in brain [192], KLF14 in adipose, TBC1D4 in muscle [130], ADCY5 in islets [188], GCKR in liver [159]). However, a range of physiological and genomic analyses consistently point to islet dysfunction as having the greatest contribution to T2D risk [110, 36, 206]. For example, genome-wide enrichment analyses highlight the particularly strong relationship between T2D-risk variants and regulatory elements active in human islets [142, 144, 196, 188].

Research access to human pancreatic islet material is largely limited to samples accessible from a subset of cadaveric organ donors, and consequent scarcity has compromised efforts to characterize the regulation of human islet expression. The human pancreatic samples examined in GTEx14 are of limited value, since islets constitute only 1% of total pancreas. Previous studies have demonstrated the potential of islet gene expression information to characterize T2D effector genes such as MTNR1B and ADCY5 [196, 44, 193], but the sample sizes examined to date have been modest: the largest published human islet RNA-Seq data includes only 118 samples [193].

We constituted the integrated network for systematic analysis of pancreatic islet RNA expression (insPIRE) consortium as a vehicle for the aggregation and joint analysis of available human islet RNA-Seq data [196, 44, 193]. Here, we report on analyses of 420 human islet preparations which provide a detailed landscape of the genetic regulation of gene expression in this key tissue, and its relationship to mechanisms of

T2D predisposition. This research addresses a series of questions with relevance beyond the specific example of T2D. When a disease-relevant tissue is missing from reference datasets such as GTEx, what additional value accrues from dedicated expression profiling from that missing tissue? What is the impact of tissue heterogeneity (cellular heterogeneity within the tissue of interest, and contamination with cells that are not of direct interest) on the interpretation of eQTL data? What does the synthesis of tissue specific epigenomic and expression data tell us about the coordination of upstream transcription factor regulators of gene expression? And, finally, what information do tissue-specific eQTL analyses provide about the regulatory mechanisms mediating disease predisposition?

4.3 Results

4.3.1 Characterization of genetic regulation of gene expression in islets

We combined pancreatic islet RNA-Seq and dense genome-wide genotype data from 420 individuals. Data from 196 of these individuals have been reported previously [196, 44, 193]. We aggregated, and then jointly mapped and reprocessed, all samples (median sequence-depth per sample, 60M reads) to generate exon- and gene-level quantifications, using principal component methods to correct for technical and batch variation (see Methods and Fig. 4.2).

To characterize the regulation of gene expression for the 17,914 protein coding and long non-coding RNA (lncRNA) genes with quantifiable expression in these samples, we performed eQTL analysis (fastQTL [139]) on both exon and gene-level expression measures, using all 8.05×10^6 variants that pass quality control (QC) (see Methods). This joint analysis of all 420 individuals identified 4,312 genes (eGenes) with significant *cis*-eQTLs at the gene level (FDR<1%; *cis* defined as within 1Mb of the

transcription start site (TSS)). Results of this joint analysis were highly-correlated with those obtained from a fixed-effects meta-analysis of the four component studies, indicating appropriate control for the technical differences between the studies (Methods). The complementary exon-level analysis, which can capture the impact of variants influencing splicing as well as expression, detected 6,039 eGenes (FDR<1%, Fig. 4.3) [131, 94]. Stepwise regression analysis (after conditioning on the lead variant) identified a further 1,702 independent eQTLs (involving 1,291 eGenes), giving a total of 7,741 islet exon-level eQTLs. At the 1,291 eGenes with at least two independent exon-eSNPs, although primary eSNPs (that is, the most significant signals for each eGene) tended to be localized closer to the canonical TSS than secondary eSNPs (Wilcoxon test $P=6.3\times10^{-30}$), there were 503 (39.0%) of these genes for which the second eSNP identified during stepwise conditional analysis was more proximal to the TSS (Fig. 4.3).

4.3.2 Tissue specific regulatory variation in islets

For many complex traits of biomedical interest, the cell types considered most relevant to disease development are either entirely absent from large-scale eQTL datasets or represent a minor component of the cellular content of assayed tissues. The value of targeting the specific cell-types of interest for dedicated eQTL discovery - as opposed to relying on existing eQTL data from more accessible tissues remains unclear. To examine this, we considered the degree to which the set of 6,039 exon-level islet eGenes overlapped with eQTLs from 44 tissues (for which sample size >70) in the version 6p release of GTEx [58]. To allow direct comparison with the InsPIRE data, we reprocessed GTEx sequence data to generate exon-level eQTLs (Methods). Approximately 5% (337) of the 6,039 islet eGenes had no significant eQTLs (in exon- or gene-level analyses) in any of the 44 GTEx tissues suggesting the islets eQTL were strong enough to be detected with 420 samples, but maybe not active or strong

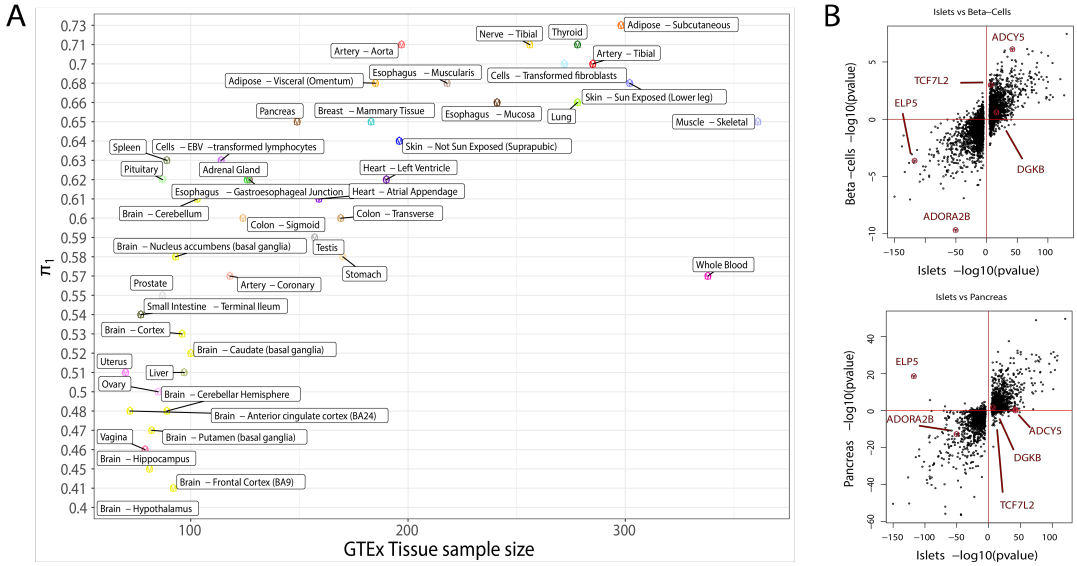


Figure 4.1: Islet eQTL discovery. A: Proportion of islet eQTLs active in GTEx tissues using P-value enrichment analysis ($\pi 1$ estimate for replication). B: Comparison between eQTLs discovered in islets and their pvalues in beta-cells (top figure, N=26) and whole pancreas tissue from GTEX (bottom figure, n=149). The axes show the -$\log_{10}$ Pvalue of the eQTL associations adjusted by the eQTL direction of effect with respect to the reference allele.

enough to be significant in other tissues with the current sample size of GTEx. Therefore, and rather than defining 'tissue-shared' effects based on arbitrary thresholds, we estimated the proportion of islet eQTLs active in other GTEx tissues using P-value enrichment analysis ($\pi 1$ [173]): the proportion of islet-eQTLs shared with other GTEx tissues ranged from 40% (brain) to 73% (adipose). As previously reported14, there was a positive linear relationship between these $\pi 1$ measures and sample sizes for the respective tissues in GTEx (Fig. 4.1A). However, $\pi 1$ estimates for shared eQTL effects reached only 65% and 57% (respectively) for skeletal muscle (n=361), and whole blood (n=338), the tissues with the largest representation in this version of GTEx.

These data demonstrate that there is a substantial component of tissue-specific genetic regulation that could, at these sample sizes, only be detected in islets, illus-

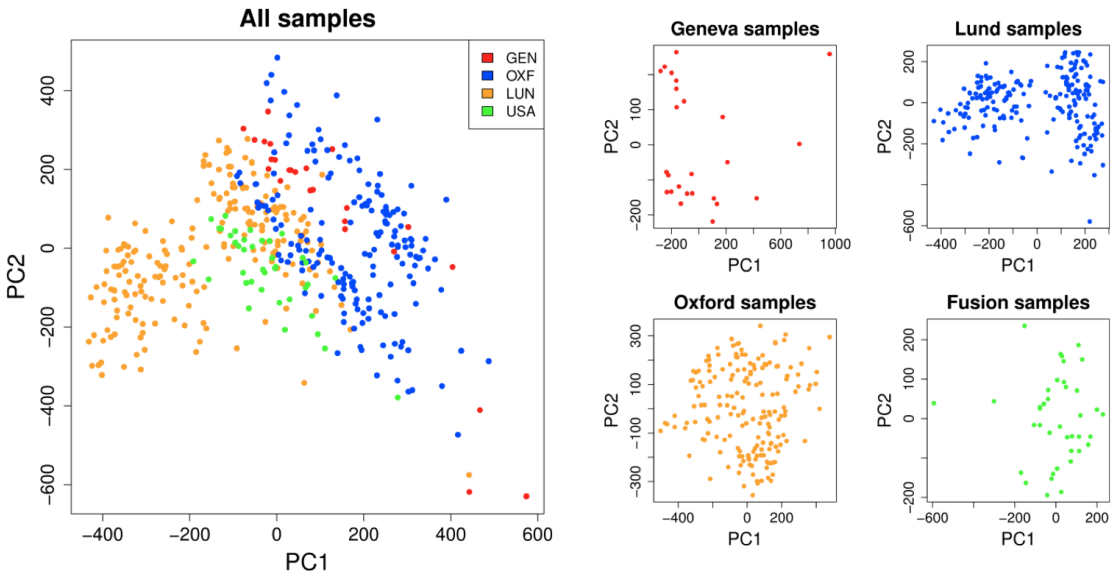


Figure 4.2: Principal component analysis (PCA) of the exon expression profiles per sample included in the InsPIRE project. Samples were re-quantified and normalized together to account for differences in the data production. The samples showed in the PCA analysis the differences due to experimental processing differences, with internal batch effects.

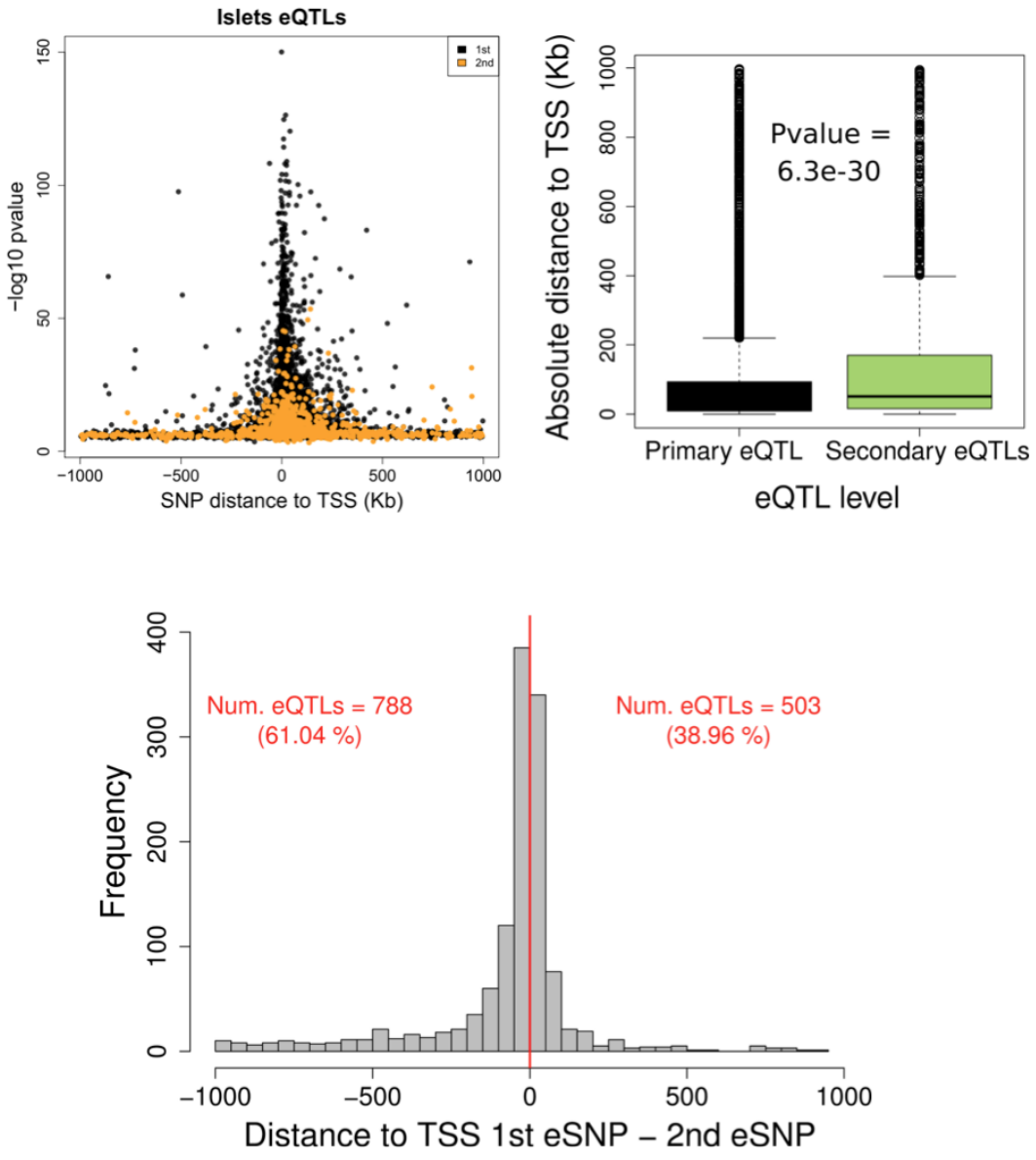


Figure 4.3: eQTL analysis: Top left figure shows the $\log_{10}$ pvalue distribution of the lead eSNP per gene around the transcription start site (TSS) of the genes in black. Yellow values show the secondary signals discovered after conditional analysis. Both the primary and secondary sSNPs show smaller pvalues around the TSS, however, the secondary signals are significantly futher away from the TSS (top right plot). The bottom plot shows the distance of the eSNPs around the TSS for those genes with 2 indepdnetn eQTLs (N = 1,290). The difference in the Kb distance between primary SNP (1st) and secondary SNP (2nd SNP as the highest variance explained in expression) independent eSNPs significantly affecting the expression of the same gene is expressed in negative values (left) if the primary eSNPs is closer to the TSS than the secondary eSNPs (N = 788). Positive values identify those eGenes in which the secondary eSNP is closer to the TSS than the primary (N = 503)

trating the value of extending current expression profiling efforts to additional tissues and cell-types of particular biomedical importance.

4.3.3 Cellular heterogeneity

The human islets analyzed in this, and other, studies include a mixture of cell types, including the hormone-producing α, β and δ cells, and, given the limitations of physical islet isolation, a variable amount of adherent exocrine material. From the perspective of T2D pathogenesis, the transcriptomes of the former (particularly α and β-cells) are of most interest. However, the eQTLs identified could have their origins from any of the cellular components. We used a number of approaches to address interpretative challenges resulting from this cellular heterogeneity.

First, we performed tissue deconvolution analysis to estimate the proportion of exocrine contamination across the 420 InsPIRE islet samples: these analyses were performed prior to the principal component adjustment used to generate the main eQTL results. We used reference expression signatures for: (a) exocrine tissue (GTEx pancreas data) [58]; (b) beta-cell; and (c) non-beta cells (the last two from a set of 26 human islet preparations which had been FAC-sorted using the zinc-binding dye Newport Green to separate the beta-cell fraction). Estimates of the proportion of exocrine pancreas contamination ranged from 1.8% to 91.8% (median 33.5%). These measures of exocrine contamination were significantly correlated ($r=0.50$, $P=2.8\times10^{-15}$) with independent estimates of exocrine content obtained at the time of islet collection by dithizone staining of the preparations (available for 232 samples) (Fig. 4.4). Within the endocrine fraction of the islet preparations, median estimates of beta-cell (58.8%, IQR 43.2-66.9%) and non-beta-cell (41.2%, 33.1-56.8%) fractions correspond well to estimates obtained through morphometric assessment [84]. In 34 samples from donors annotated as having T2D, median estimates of beta-cell composition were lower than

those from donors annotated as non-diabetic (n=330) (median 31.8% vs. 35.6%, P=4.5×10^{-4}, Fig. 4.4). This analysis provides independent confirmation, based on transcriptomic signatures, of evidence from morphometric and physiological studies that the functional mass of beta-cells is reduced in T2D [122, 19].

Second, we investigated the proportion of eQTL signals from InsPIRE islet RNA profiles also active in GTEx tissues (see above) and confirmed that whole pancreas, often naively-used as a surrogate for the T2D-relevant islet component, represents an imperfect proxy for islet ($\pi1$=0.67 with our human islet data). The extent of this eQTL overlap depends on study sample sizes and with GTEx v6p, the whole pancreas overlap is on a par with other tissues such as photo protected skin ($\pi1$=0.67) and spleen ($\pi1$=0.61) This suggests pancreas and other tissues are equally useful to infer genetic regulatory effects on expression, with the expectation that larger studies will reduce the overlap across tissues while increasing the detection of tissue specific regulatory effects.

The principal component adjustments we used to control for unwanted technical variation during eQTL analysis were designed to account for some of the impact of variation in sample purity. However, by correlating the data-generated PCs with cell proportion estimates, we observed that, even when adjusting using 25 PCs, only 30% of the variance attributable to variation in exocrine or beta-cell composition was regressed out, requiring more than 107 PCs to remove 50% of the variance. This suggests that some of the eQTLs here attributed to pancreatic islets may, in fact, reflect exocrine pancreatic contamination. To evaluate this further, we compared the sets of eQTLs identified in the InsPIRE islet samples with the highest and lowest proportions of exocrine contamination (n=100 for each) and 100 randomly-selected GTEx whole pancreas samples. Overlap between whole pancreas and islet eQTLs

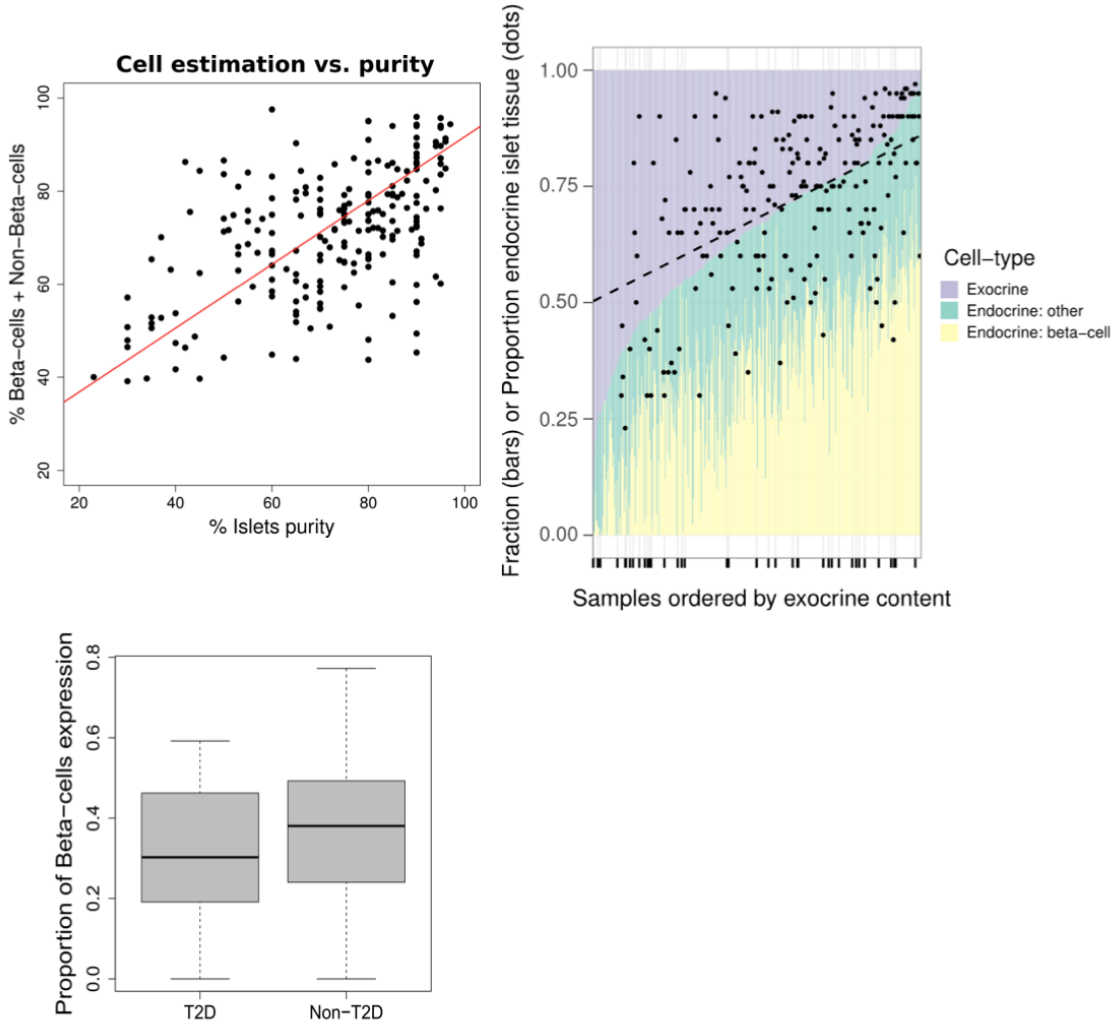


Figure 4.4: Cell deconvolution analysis. Top right plot shows the estimates of the different types of cell considered in the 420 islets samples processed. The beta-cells proportion composition form per sample corresponded to a median os 58.8%, and 41.2% for non-beta-cell fractions. Top left plot shows the percentage of purity for islets as measured in dithizone staining of the 232 samples compare to the estimated proportion of (beta-cells + other non-exocrine cell)/ total cell content in islets. The correlation between measured values of purity was ρ = -0.5 (P=2.8×10^{-15}). Bottom plot shows the Percentage of Beta-cells expression detected in islets samples from individuals identified as diabetics (T2D), compare to non-T2D individuals.

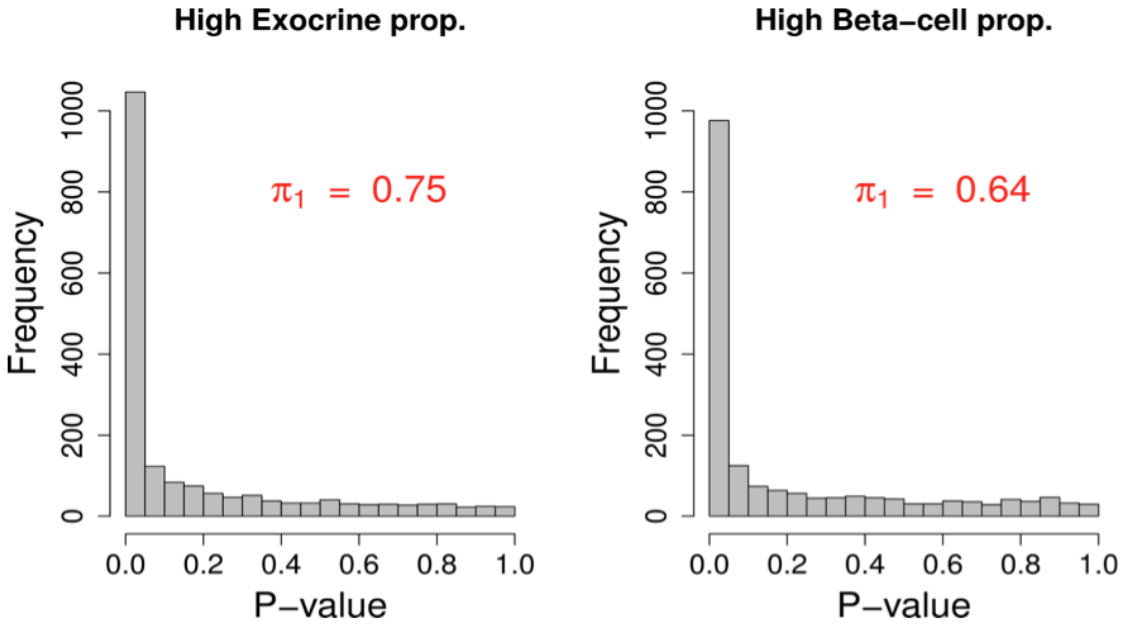


Figure 4.5: Replication rate of pancreas eQTLs in 100 islets with high proportion of exocrine expression (left) and in 100 islets with high proportion of beta-cells expression (right).

(using $\pi 1$ [173]) was greater in islet samples with the highest exocrine contamination (75% vs 64%) (Fig. 4.5). Although shared regulatory processes between acinar and islet tissue are to be expected [168], these data suggest that apparent overlap in regulatory signals between islet and whole pancreas may partly reflect the consequence of the inadvertent exocrine contamination of islet data.

Of the 420 InsPIRE samples, beta-cell enriched transcriptomes were available for 26 following FAC-sorting. These data allowed us to look for evidence of cell-type-specific eQTL effects and attempt to identify the cellular source of the eQTLs detected in the islet material. With this limited sample size, the only eQTL reaching significance (and then only, at a less stringent threshold of FDR<5%) was at *ADORA2B* ($P=3.8\times10^{-10}$, beta = -1.207): this signal was also detected in the InsPIRE islets ($P=3.9\times10^{-51}$, beta = -0.656) and in GTEx pancreas ($P=1.6\times10^{-16}$, beta = -0.737) (Fig. 4.6). Of the 7,741 independent SNP-exon pairs significant in islets, 227 were

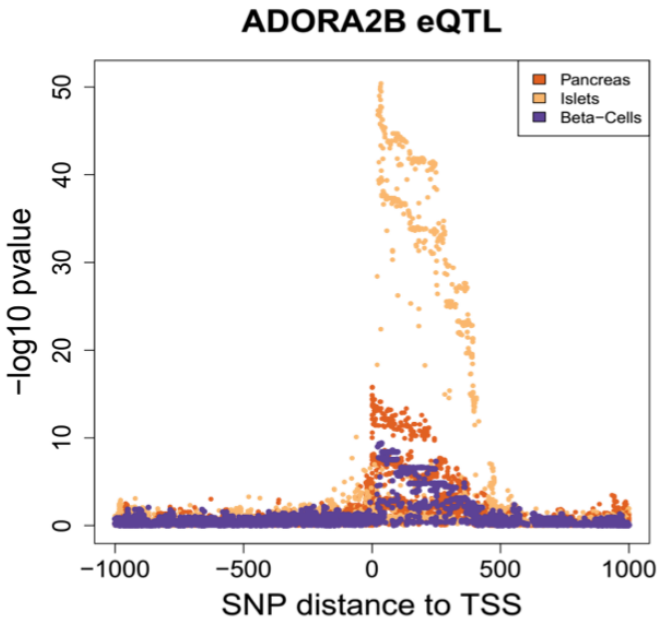


Figure 4.6: eQTL for *ADORA2B* gene in islets, beta-cells and pancreas samples. Each dot represent a SNPs in the *cis* window of *ADORA2B* and their distance in kb to the TSS. The y-axis shows the -$\log_{10}$ of the P value for the association between a given SNP and the expression of the same exon in *ADORA2B*. For all tissues, at least one SNP was significant after multiple testing (FDR = 5%).

also significant in beta-cells at FDR<1%. By comparing the p-value distributions of the eQTLs in islets vs beta-cells, we estimate that 46% of islet-eQTLs are active in beta-cells (Fig. 4.1B).

To identify specific genes with cell-type-specific regulatory effects, we tested for interactions between genotype and the beta-cell or exocrine cellular fraction estimates, controlling for technical variables (Methods). We identified 18 islet *cis*-eQTLs with a genotype-by-beta-cell proportion interaction and 8 with a genotype-by-exocrine cell proportion interaction (FDR<1%). The former group included *ADCY5*, a member of the adenylate cyclase family implicated as a T2D GWAS effector transcript by several islet-eQTL studies [196, 155] and CCL2 (also known as MCP-1), a cytokine implicated in type 1 diabetes (T1D) development [87].

We conclude that a substantial proportion of the regulation of gene expression detected in pancreatic islets is derived from cell-specific effects. Ongoing efforts to

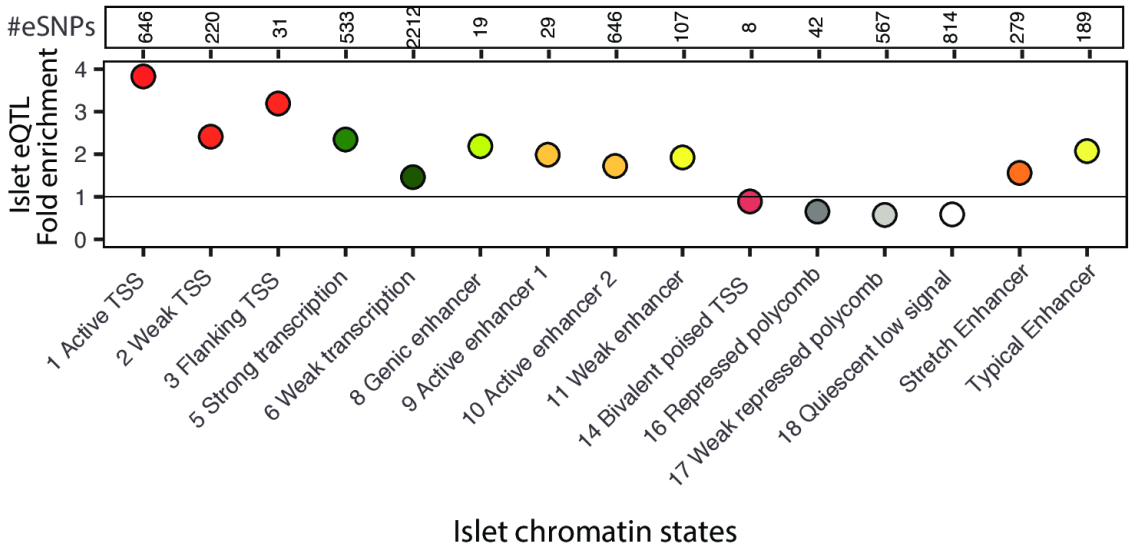


Figure 4.7: eQTL in enrichment chromatin states. Islet eQTL overlap with chromatin states and stretch/typical enhancers. Top: Number of islet eQTL in 13 islet chromatin states and stretch and typical enhancers. Bottom: Fold enrichment of islet eQTL in chromatin states calculated using GREGOR [160].

develop a single-cell view of islet transcriptional signatures should help to inform these analyses, although the limited sample size of current single-cell transcriptomic studies [202] and their lack of genotype information means they offer little direct insight into the relationship between genetic variation and cell-type-specific transcript abundance.

4.3.4 Functional properties of islet genetic regulatory signals

Using previously-published islet chromatin states derived from histone modification data12, we observed a significant enrichment of islet eSNPs in active islet chromatin states including active TSS (fold enrichment = 3.84, P = 5.5×10^{-206}), active enhancers (fold enrichment > 1.73, P < 4.8×10^{-04} between two enhancer states) and stretch enhancers (fold enrichment = 1.57, P = 2.7×10^{-13}), with concomitant depletion of eSNPs in repressed and quiescent islet chromatin states (fold enrichment < 0.66, Fig. 4.7). This recapitulates the enrichment observed for T2D GWAS signals within active islet chromatin Fig. 4.8 [142, 144, 196, 188].

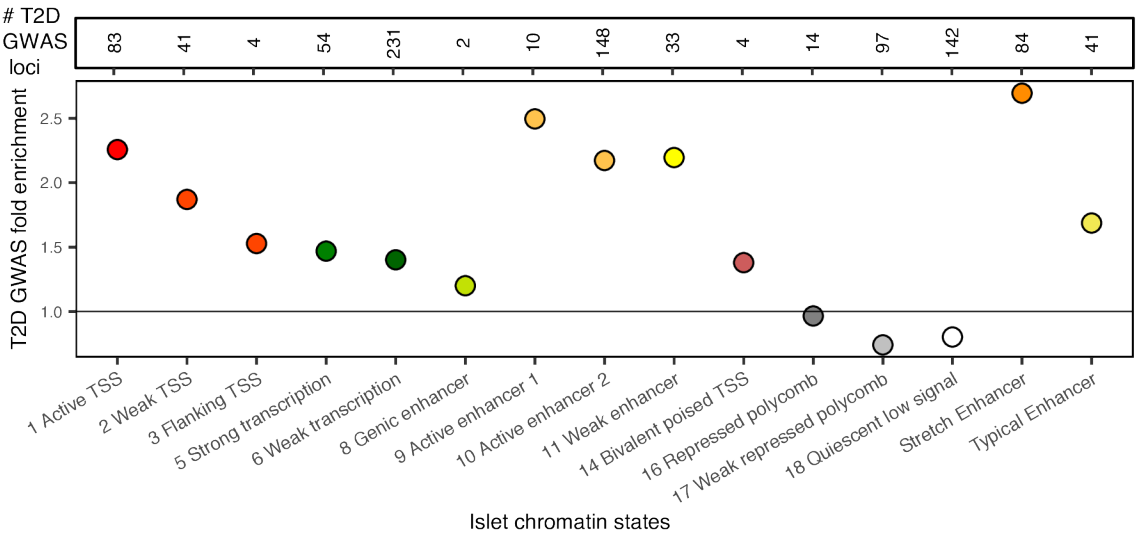


Figure 4.8: Enrichment for T2D GWAS loci to overlap islet chromatin states. Top: Number of T2D GWAS loci occurring in each of the 13 islet chromatin states along with stretch and typical enhancers. Bottom: Fold enrichment of T2D GWAS in chromatin states calculated using GREGOR [160].

To explore further the chromatin context of islet eSNPs, we first asked whether eSNP effect sizes (measured as the slope of the linear regression) were uniform across the different underlying chromatin contexts in which they occur. We found a non-uniform range of distributions (Fig. 4.9A): for example, eSNPs that overlap active TSS chromatin states had significantly larger effects than those that overlap repressed or weak-repressed polycomb chromatin states (Wilcoxon Rank Sum Test P=0.039).

Because chromatin states represent integrated histone mark patterns, and transcription factors (TFs) are more likely to bind in open accessible DNA, we next focused on regions of accessible chromatin within each of the chromatin states, using previously-published ATAC-seq (assay for transposase accessible chromatin followed by sequencing) data from human islets [196]. As expected, a disproportionately high proportion (80%) of islet eQTLs (based on the lead eSNP itself or proxy SNPs with LD r^2>0.99) overlap islet ATAC-seq peaks in islet-active TSS chromatin states. More specifically, of the 646 islet eSNPs that overlap islet active TSS chromatin, 522

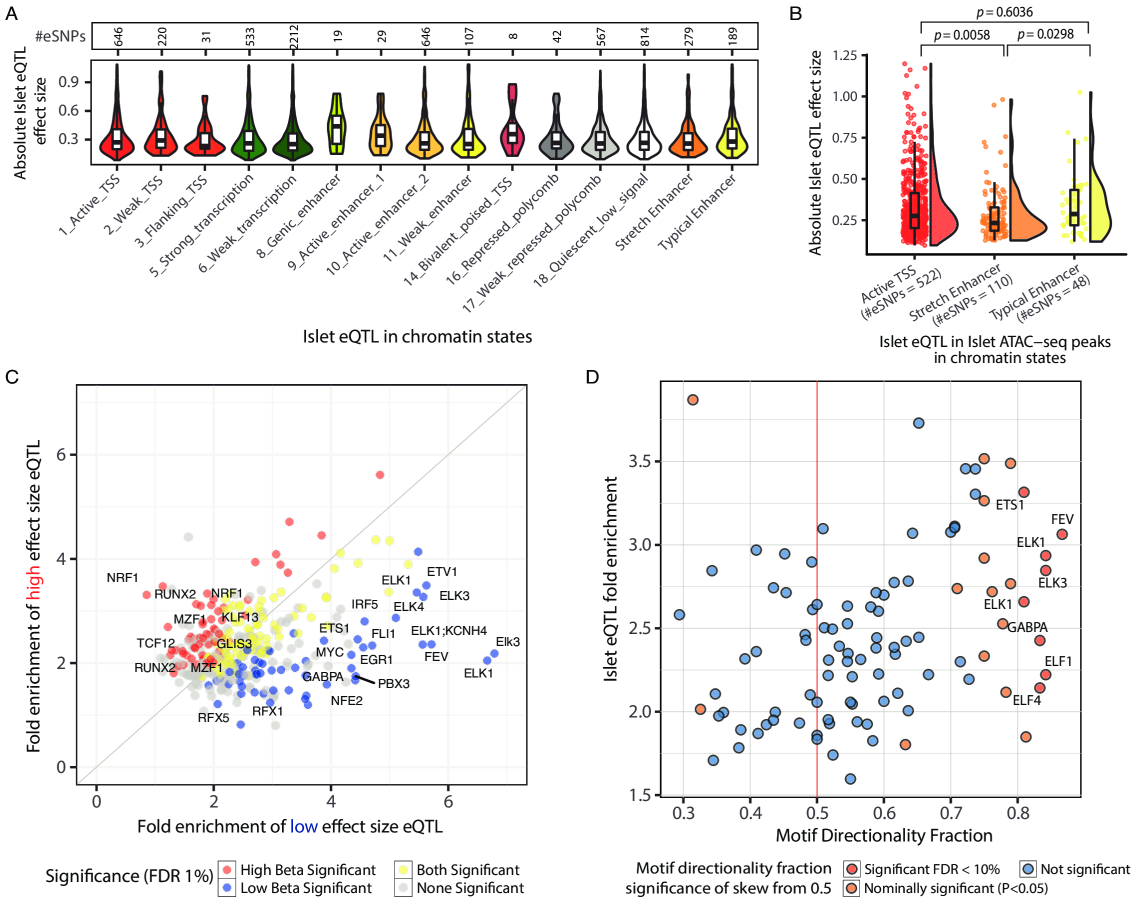


Figure 4.9: Integration of Islet eQTL with epigenomic information reveals characteristics of gene expression regulation. A: Distribution of absolute effect sizes for Islet eQTLs in each Islet chromatin state. B: Distribution of absolute effect sizes for Islet eQTL in ATAC-seq peaks in three Islet chromatin states. eQTL SNPs in ATAC-seq peaks in stretch enhancers have significantly lower effect sizes than SNPs in ATAC-seq peaks in active TSS and typical enhancer states. P values obtained from a Wilcoxon rank sum test. C: Fold Enrichment for transcription factor footprint motifs to overlap low vs high effect size islet eQTL SNPs. D: TF footprint motif directionality fraction vs fold enrichment for the TF footprint motif to overlap islet eQTL. TF footprint motif directionality fraction is calculated as the fraction of eQTL SNPs overlapping a TF footprint motif where the base preferred in the motif is associated with increased expression of the eQTL eGene. Significance of skew of this fraction from a null expectation of 0.5 was calculated using Binomial test.

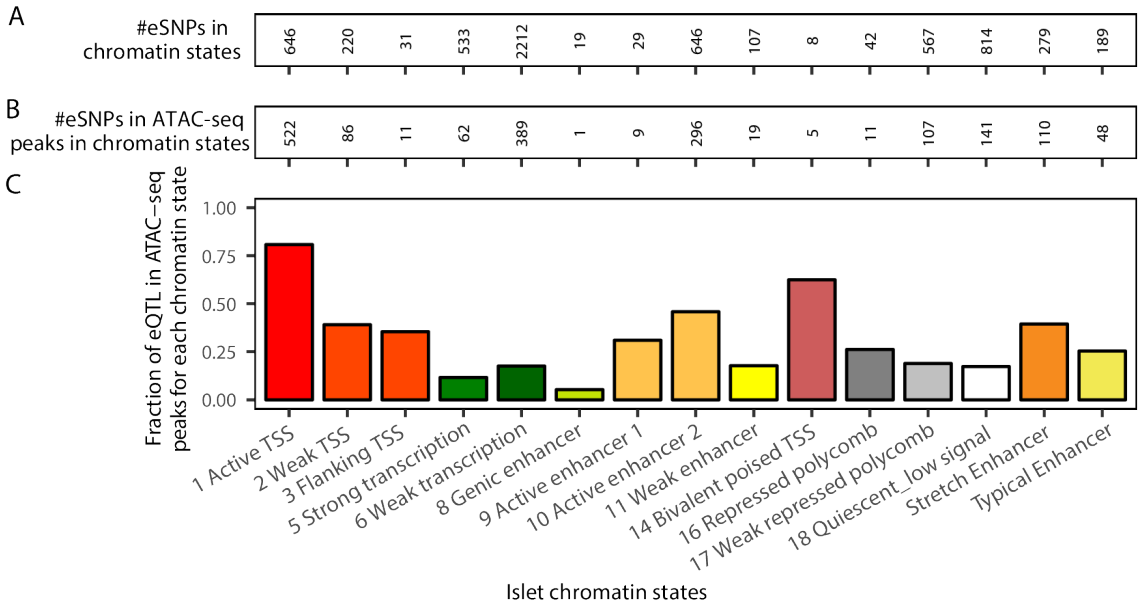


Figure 4.10: Fraction of eQTLs in ATAC-seq peaks in chromatin states. A: Number of eQTL Islet eQTL overlapping with Islet chromatin states and stretch/typical enhancers. B Number of Islet eQTL in Islet ATAC-seq peaks in chromatin states. C: Fraction of Islet eQTL in ATAC-seq peaks in each chromatin states. An eQTL overlap is considered if the eQTL lead eSNP or proxy SNP (LD $r^2 > 0.99$) overlaps the feature.

(80.8%) occur in the (ATAC-defined) open chromatin portion of that chromatin state Fig. 4.10. We note that 49.7% of the islet active TSS chromatin state territory is occupied by islet ATAC-seq broad-peaks.

When we examined the distribution of absolute effect sizes for eSNPs that occur within islet ATAC-seq peaks within active TSS, islet stretch enhancers, which were defined as enhancer chromatin state segments longer than 3kb and were show to be islet-specific [142] and typical enhancer (enhancer chromatin states smaller than median size of 800bp) annotations, we found that eSNPs in stretch enhancers had significantly lower effect sizes than those in either typical enhancer chromatin (Wilcoxon Rank Sum Test P=0.0088) or active TSS chromatin states (P=0.0099) (Fig. 4.9B). One important corollary of this observation that eSNPs in different chromatin contexts have different regulatory effect sizes is that eSNPs in cell-specific

stretch enhancers may be less detectable as eQTLs, and that higher sample sizes are needed to ensure better powered discovery of eQTLs within these critical cell identity regions.

We next sought to use the combination of islet chromatin state annotation and eQTL data to identify TFs driving islet regulatory networks. For these analyses, we used published TF footprint (in vivo predicted occupied TF motif binding sites) results generated from human islet ATAC-seq data [196]. We previously reported enrichment of selected TF footprint motifs at islet eSNPs, using a smaller islet expression dataset [196]: here, the larger eSNP catalog allowed us to determine how eSNP effect size and target gene expression directionality is associated with base-specific TF binding preferences. We partitioned eSNPs into two equally-sized bins representing those with lower (absolute beta from regression <0.254) and higher ($geq 0.254$) effect sizes. Higher-effect size eSNPs were preferentially enriched ($<1\%$ FDR) for a subset of footprint motifs, characteristic of islet-relevant TF families including KLF11 (motif=KLF13_1, $P=5.3\times10^{-6}$) and GLIS3 (motif GLIS3_1, $P=5.2\times10^{-6}$). Other sets of footprint motifs, including those related to the RFX and ETS families of TFs, were significantly enriched for low effect size eSNPs ($P<2\times10^{-4}$) (Fig. 4.9C). Collectively, these results further demonstrate a relationship between local chromatin environment at the level of a TF footprint motif and eSNP effect size.

Since TFs can act as activators, repressors, or both [42], we asked if eSNP alleles that match the base preference at TF footprint motifs have a consistent directional impact on gene expression. We defined a motif directionality fraction score for each TF footprint motif by calculating the fraction of overlapping eSNP where the preferred base in the motif was associated with increased expression of the eGene (Methods 'TF motif directionality'). Directionality fractions indicate if the TF motifs are activating (fraction near 1), repressive (fraction near 0), or show no preference (fraction near 0.5). We found that the motif activity measures generated with this islet eQTL and ATAC-

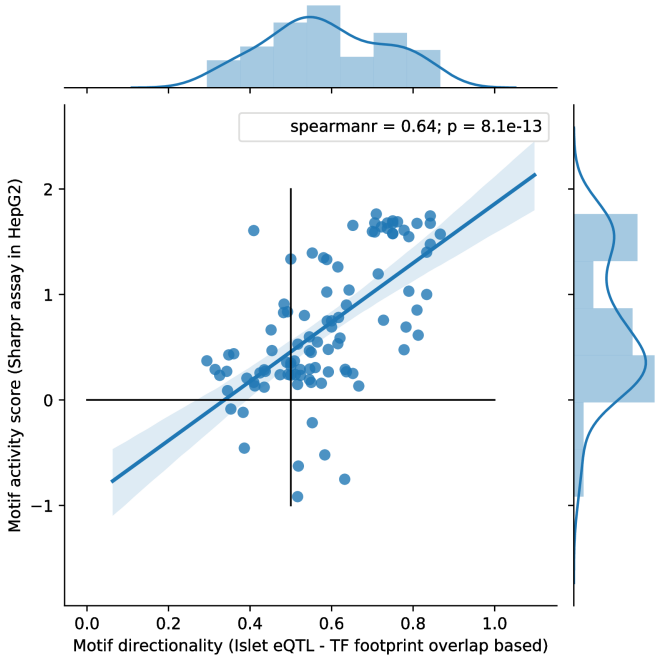


Figure 4.11: TF motif directionality comparison with MPRA activity. Transcription factor motif activity scores from Sharpr MPRA in HepG2 cells [42] vs Motif directionality fractions from Islet eQTL and ATAC-seq TF footprinting data. TF Motifs that were reported to be either activating or repressive (P¡0.01) from the MPRAs in both HepG2 and K562 are shown.

seq footprint-based metric were largely concordant (Spearman's r=0.64, P=8.110-13) with orthogonal motif activity measures derived from massively parallel reporter assays (MPRAs) performed in HepG2 and K562 cell lines [42] (Fig. 4.11). There were 99 motifs reported as consistently activating or repressive across HepG2 and K562 cell lines present in our study: for these, we tested whether the motif directionality fraction deviated from null expectation (no preference for activator/repressor) using a binomial test. We found that only 8% (n=8) of the motifs showed evidence of skewed activator preference (<10% FDR; Fig. 4.9D). The activator motifs we identified include many ETS family members, which have a known preference for transcriptional activation [42].

Our analyses demonstrate how integrating diverse epigenomic information with rich eQTL data can reveal characteristics of gene regulation and its regulators. While

contrasting tissue-specific stretch enhancers with the more ubiquitous TSS states in the context of eQTL effect sizes delineated the role of underlying chromatin on function; integrating eQTL information with ATAC-seq and high-resolution TF footprinting revealed in vivo activities of these upstream regulators.

4.3.5 Islet eQTLs are enriched among T2D and glycemic GWAS variants

Diverse lines of evidence emphasize the contribution of reduced pancreatic islet function to the development of T2D, and there is evidence, based on patterns of association across diabetes-related quantitative traits, that many T2D GWAS loci act primarily through their impact on insulin secretion [110, 111, 36, 188]. To examine the relationships between T2D predisposition alleles and the tissue-specific regulation of gene expression, we combined the human islet eQTL data with equivalent exon-level information for 44 tissues available through GTEx (version 6p) [58]. We examined 122 GWAS lead variants with genome-wide significant associations to T2D (focusing on 78 signals with the most pronounced effects on T2D risk as detected in 3 or 44 continuous glycemic traits relevant to T2D predisposition (including fasting glucose, and beta-cell function (HOMA-B) in non-diabetic individuals) [163, 113, 175]. For each of these GWAS lead variants, we extracted the lead eSNP from the 44 GTEx tissues and the InsPIRE pancreatic islets. To determine the extent to which the lead T2D GWAS variant showed tissue-specific enrichment for islet eQTL associations, we compared these observed effect size estimates from the eQTLs to those derived from a null distribution of 15,000 random eSNPs, matched to the GWAS SNPs with respect to the number of SNPs in LD, distance to TSS, number of nearby genes and minor allele frequency. We were particularly focused on the enrichment in eQTL effect sizes at T2D/glycemic GWAS-associated variants for six tissues implicated in T2D pathogenesis (subcutaneous adipose tissue, skeletal muscle, liver, hypothalamus, islets and whole pancreas), plus whole blood for comparison (Fig. 4.13A, Fig. 4.12).

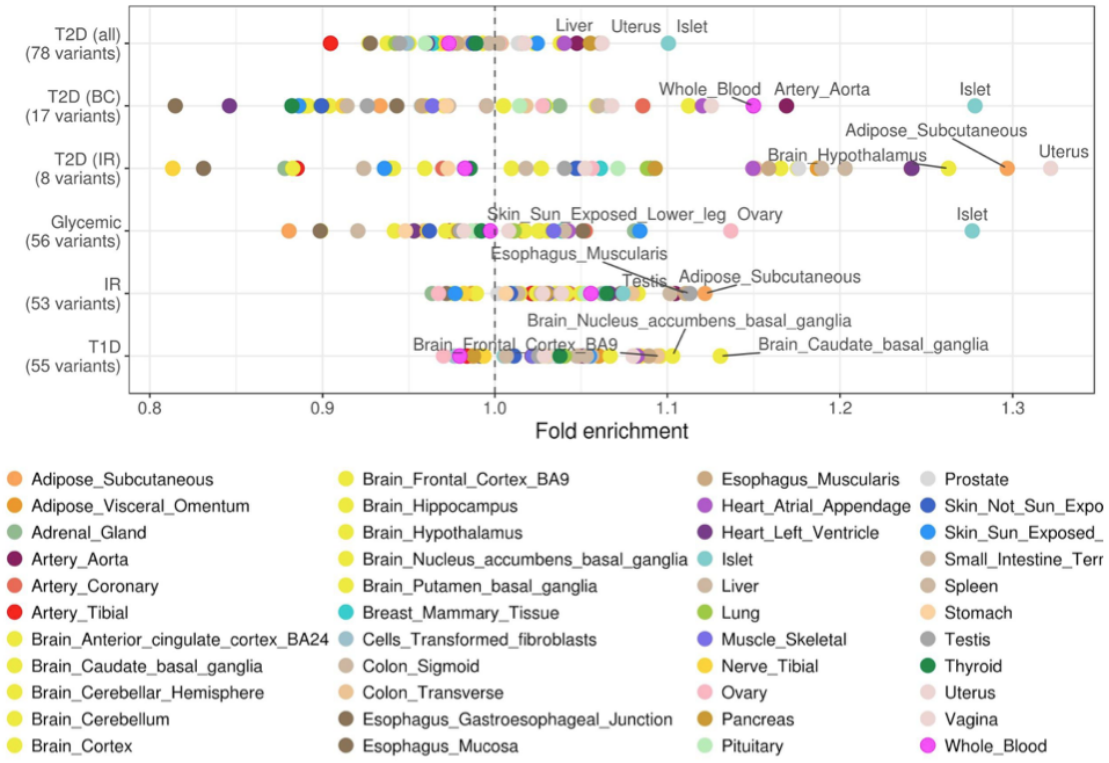


Figure 4.12: Enrichment of GWAS loci in eQTL for GTEx tissues.

We included a set of 55 lead variants implicated by GWAS in predisposition to T1D for comparison [138].

Across the 45 tissues, we detected significant enrichment for islet eQTLs amongst variants associated with continuous glycemic traits (normalize enrichment score (NES)=1.27; P=3.6×10^{-3}). Apart from a modest signal in ovary (NES=1.13, P=0.02), there was no enrichment in any other GTEx tissue. The enrichment for islet eQTLs for the full set of 78 T2D variants was directionally consistent with the results for continuous glycemic traits but did not reach nominal significance (NES=1.10; P=0.07). However, T2D GWAS signals influence disease risk through physiological effects in multiple tissues. In the subset (n=17) of the 78 T2D GWAS signals with the strongest evidence (from patterns of association to other T2D-related traits) of mediation through reduced insulin secretion (implicating islet dysfunction) [111, 36, 206], we observed

more marked enrichment of islet eQTL signals (NES=1.27; p=0.025). For this subset there was no enrichment for eQTL effect sizes in whole pancreas (NES=0.90, P=0.88). There was no enrichment of islet eQTL effects for the set of T1D association signals, consistent with the consensus that most genetic risk for T1D is mediated through immune mechanisms [138]. In the subset of 8 T2D GWAS signals with the strongest evidence of mediation through defects in insulin action (n=8), enrichment was seen in insulin target tissues such as liver (NES=1.10; P=0.03), adipose tissue (NES=1.12; P=0.04) and brain cortex (NES=1.10; P=0.03), but not in islets (NES=1.07, P=0.17). Similar patterns of eQTL enrichment were seen for a broader, partly-overlapping, set of 53 lead variants influencing insulin sensitivity derived from a multivariate GWAS [107]. These data reveal tissue-specific patterns of genetic regulatory impact for variants at T2D- and glycemic-trait loci which mirror the mechanistic inferences generated by physiological analysis of those signals. They also highlight the importance of matching the tissue origin of the transcriptomic data used for mechanistic inference, to the tissue-specific impact of each GWAS signal on disease predisposition.

4.3.6 Identifying effector transcripts for T2D and glycemic traits

This evidence of generalized overlap between islet eQTLs and (selected) T2D and/or glycemic GWAS signals motivates further efforts to characterize these relationships at individual loci. Previous studies have identified GWAS signals displaying apparent overlap between islet eQTLs and the T2D/glycemic GWAS signals [196, 44, 193], but not all of these signals have been evaluated with respect to the statistical evidence for co-localization (i.e. testing whether the eQTL and the GWAS signals are likely to emanate from the same causal variants), and not all coincident signals have replicated despite ostensibly similar designs and power.

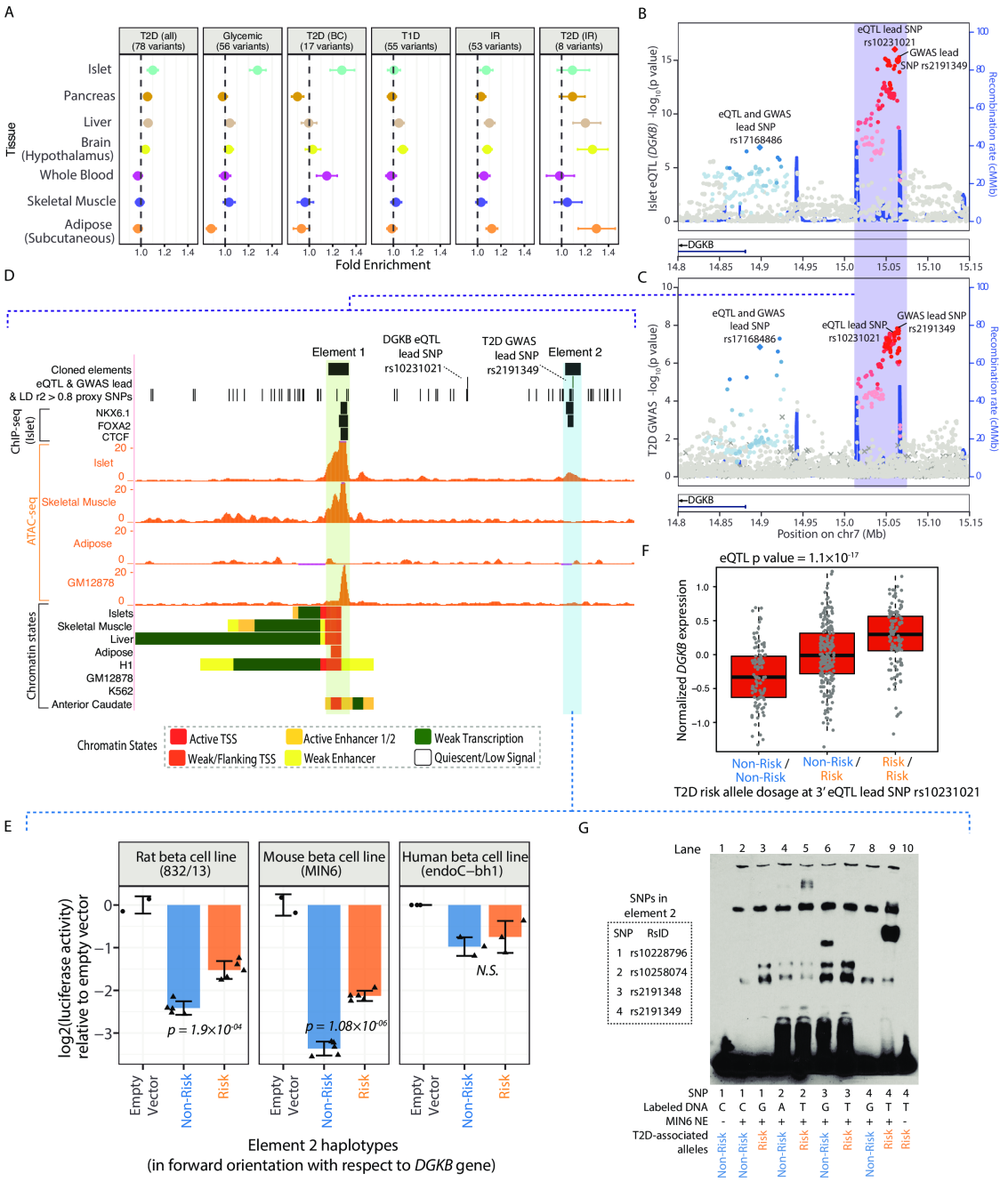


Figure 4.13: Functional validation of DGKB eQTL locus. A: Enrichment of eQTL effect sizes in different GTEx tissues at T2D/glycemic GWAS-associated variants. Numbers within square brackets denote the number of variants implicated for the trait. Also shown are subsets of T2D GWAS associated with reduced insulin secretion or islet beta cell dysfunction (T2D (BC)) or insulin resistance (T2D (IR)), type 1 diabetes (T1D) signals, insulin resistance (IR). B: Two independent islet eQTL signals (lead SNP rs17168486 referred at as the 5' signal and lead SNP rs10231021 referred to as the 3' signal) are identified near the DGKB gene locus. Continued on the next page.

Figure 4.13: continued - These signals co-localize with two independent T2D GWAS signals shown in C: where (rs17168486 referred to as the 5' signal and lead SNP rs2191349 referred to as the 3' signal and. LD information was not available for SNPs denoted by (X). D: Genome browser view of the region highlighted in purple in (B) and (C) that contains the 3' DGKB eQTL and T2D GWAS signals. Two regulatory elements overlapping islet ATAC-seq peaks (element 1 highlighted in green, element 2 highlighted in blue) were cloned into a luciferase reporter assay construct for functional validation. E: Normalized DGKB gene expression levels relative to the T2D risk allele dosage at the 3' islet eQTL for DGKB lead SNP rs10231021. eQTL P value adjusted to the beta distribution is shown. F: Log 2 Luciferase assay activities (normalized to empty vector) in rat (832/13), mouse (MIN6) and human (endoC) beta cell lines for the element 2 highlighted in blue in (D). Risk haplotype shows significantly higher ($P<0.05$) activity than the non-risk haplotype in 832/13 and MIN6, consistent with the eQTL direction shown in (F). P values were determined using unpaired two-sided t-tests. G: Electrophoretic mobility shift assay (EMSA) for probes with risk and non-risk alleles at the four SNPs overlapping the regulatory element validated in (F) using nuclear extract from MIN6 cells.

There are multiple methods for evaluating the evidence for co-localization: these make different assumptions and often lead to discrepant results [80]. In this analysis, we focused on the co-localization evidence provided by two complementary algorithms: COLOC, which assesses the differences in regression coefficients of variants associated to two traits, and RTC (Regulatory Trait Concordance), which assesses the differences in ranking of SNPs associated to one trait after conditioning on the most associated SNP for the other trait [53, 140]. We detected evidence for co-localization (with either method) of islet eQTLs at 23 GWAS loci, comprising 24 independent signals (the DGKB hosts two signals), 16 of which reflect T2D associations and 8 glycemic traits. Evidence for co-localization was most compelling for 11 loci (12 signals) at which both RTC and COLOC provided strong support: including extending confirmation of previously observed co-localizations at ADCY5, TCF7L2, HMG20A, IGF2BP2 and DGKB [188, 20].

At other loci, we observe islet *cis*-eQTL co-localization for the first time. For example, rs7903146, the lead variant at the T2D-risk signal at TCF7L2, co-localizes

with islet expression of TCF7L2 ($P=1.9\times10^{-7}$) (Fig. 4.14), with the T2D-risk allele increasing TCF7L2 expression (eQTL beta = 0.218). The same eQTL signal was also detected in the smaller beta-cell-specific eQTL analysis (n=26; eQTL beta= 0.724; $P=1.0\times10^{-3}$). Previous efforts to characterize the mechanism of action at this signal have demonstrated that the fine-mapped T2D-risk allele at rs7903146 influences chromatin accessibility and enhancer activity in islets37, but evidence linking these events to TCF7L2 expression has been missing. Indeed, recent studies have proposed other nearby genes as possible effectors transcripts, such as ACSL5 [56]: however, we found no evidence in any tissue (from GTEx or InsPIRE) to indicate that rs7903146 influences ACSL5 expression. The association between rs7903146 and TCF7L2 expression was restricted to islets, consistent with evidence that non-diabetic carriers of the TCF7L2 risk-allele display markedly reduced insulin secretion [211].

Several previously-reported co-localizing signals were not observed in our exon-eQTL based analysis. MTNR1B has shown consistent islet *cis*-eQTL signals across multiple previous studies [193, 190], but was excluded from our exon-level analysis due to low exonic-read coverage. However, in gene-level expression analyses, we once again observed strong evidence of co-localization between the lead T2D GWAS variant (rs10830963) and MTNR1B expression ($P=5.3\times10^{-21}$). At ZMIZ1, the previously-reported *cis*-eQTL was nominally significant (rs185040218; $P=3.0\times10^{-5}$) but this particular signal did not reach the 1% FDR threshold for inclusion in co-localization testing.

At other loci, complex, but divergent, patterns of association between the eQTL and T2D GWAS signals (likely reflecting the impact of multiple enhancers active in different tissues to the T2D signal) challenged the assumptions of these co-localization methods. At the ZBED3 locus for example, the association plots highlight two dis-

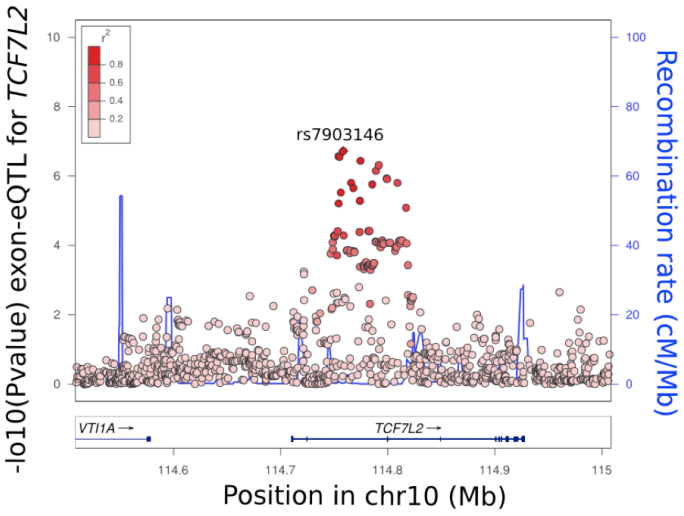


Figure 4.14: *TCF7L2* eQTL locus.

tinct T2D signals (500kb apart), and two islet eQTL signals for the PDE8B gene, but only the signal at rs7708285 appears coincident (Fig. 4.15). COLOC detects this as co-localization, but this configuration cannot easily be tested using RTC as it restricts analysis to variants that lie between a single pair of recombination hotspots.

Finally, we attempted to further characterize eGenes that overlapped signals from T2D and glycemic trait GWAS studies by assessing the impact of acute changes in glycemic status on their expression in islets. We used data from a recent analysis of human islets obtained from a set of T2D, and non-diabetic donors and focused on transcripts that showed acute changes in expression when exposed to altered glucose levels in culture (that is, islets from diabetic individuals cultured at normal glucose, and islets from non-diabetic subjects cultured in high glucose) [141]. This revealed multiple islet eGenes, including STARD10, WARS, SIX3, NKX6-3 and KLHL42 which may be of particular interest given that their expression in islets is regulated both by T2D-associated variation and by acute changes in glucose exposure.

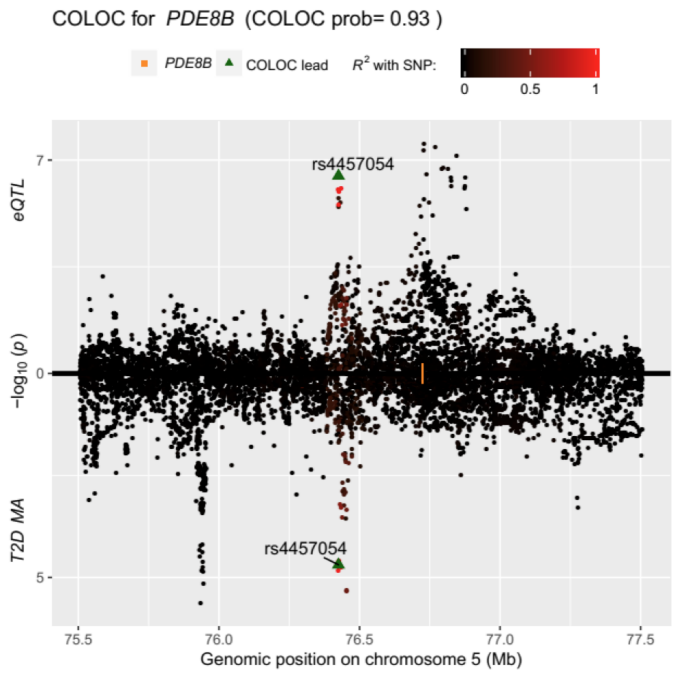


Figure 4.15: *PDF8B* eQTL and T2D GWAS loci. Miami plot of the *PDE8B* eQTL locus is shown on the top and the T2D GWAS locus is on the bottom.

4.3.7 Experimental validation at DGKB

The DGKB locus features two independent co-localizing signals: at both of these, the T2D-risk allele is associated with increased islet expression of DGKB (Fig. 4.13A). At the 5' signal, the lead SNP for both the T2D GWAS and the islet *cis*-eQTL is rs17168486. At the 3' signal, the lead eSNP, rs10231021 (Fig. 4.13B), is in high LD (r^2=1, D'=1) with the lead GWAS variant rs10231021 (Fig. 4.13C). The pattern of GWAS association signals for diabetes-related traits for both signals is consistent with a primary impact on insulin secretion (implying islet dysfunction)5,7. We prioritized, for functional analysis, variants that were in high LD (r^2>0.8) with the lead SNPs and located in islet ATAC-seq peaks (Fig. 4.13D).

At the 3' signal, seven variants met these criteria: three (rs7798124, rs7798360 and rs7781710, Fig. 4.13D, 'element 1') overlap an ATAC-seq peak shared across islets, skeletal muscle and the lymphoblastoid cell line GM12878 [17] cell-line, and

four others (rs10228796, rs10258074, rs2191348 and rs2191349, Fig. 4.13D 'element 2') lie in a smaller but more islet-specific ATAC-seq peak. We cloned these putative regulatory elements (Fig. 4.13D) into luciferase reporter constructs and performed transcriptional reporter assays in three insulin-34secreting beta-cell models, including human EndoC-βH1, rat INS1-derived 823/13 and mouse MIN6. Element 1 demonstrated consistent enhancer activity across all three beta-cell lines but did not show allelic differences consistent with the eQTL direction of effect Fig. 4.16. Element 2, when in forward orientation with respect to DGKB, showed reduced luciferase expression in all three beta-cell lines compared to control. The T2D-risk haplotype showed significantly higher expression than the non-risk haplotype in 832/13 (P = 1.910-4) and MIN6 cell-lines (P = 1.110-6): equivalent experiments in EndoC-βH1 showed a consistent trend, which did not reach significance (Fig. 4.13E). Luciferase assays using element 2 in reverse orientation also showed consistent trends across all three cell lines, reaching significance in 832/13 alone (Supp Figure S14). These data suggest that T2D risk alleles alleviate regulatory element repression and are directionally consistent with the 3' DGKB eQTL (Fig. 4.13F). In electrophoretic mobility assays using MIN6 nuclear extract, three of the four 'element 2' variants (rs10228796, rs2191348, and rs2191349) showed allele-specific binding (Fig. 4.13G), supporting a functional regulatory role.

At the 5' eQTL, we focused attention on rs17168486, which was both the lead *cis*-expression and GWAS SNP at the 5' eQTL, and is located in an islet ATAC-seq peak Fig. 4.18A. We cloned an element including this variant into luciferase reporter constructs but observed no consistent allelic effects on transcriptional activity Fig. 4.18B.

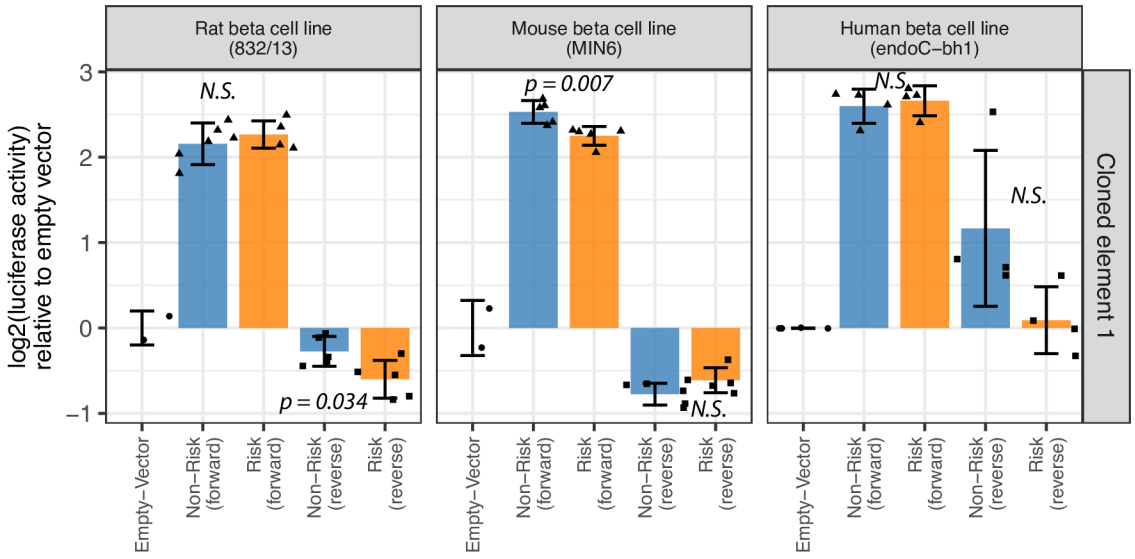


Figure 4.16: Luciferase assay results for *DGKB* 3' eQTL element 1. Log_2 luciferase assay activities (normalized to empty vector) in rat (832/13), mouse (MIN6) and human (endoC) beta cell lines for the element 1 highlighted in green in Fig. 4.13D. The element was cloned in both forward and reverse orientation with respect to the *DGKB* gene. P values were determined using unpaired two-sided t-tests.

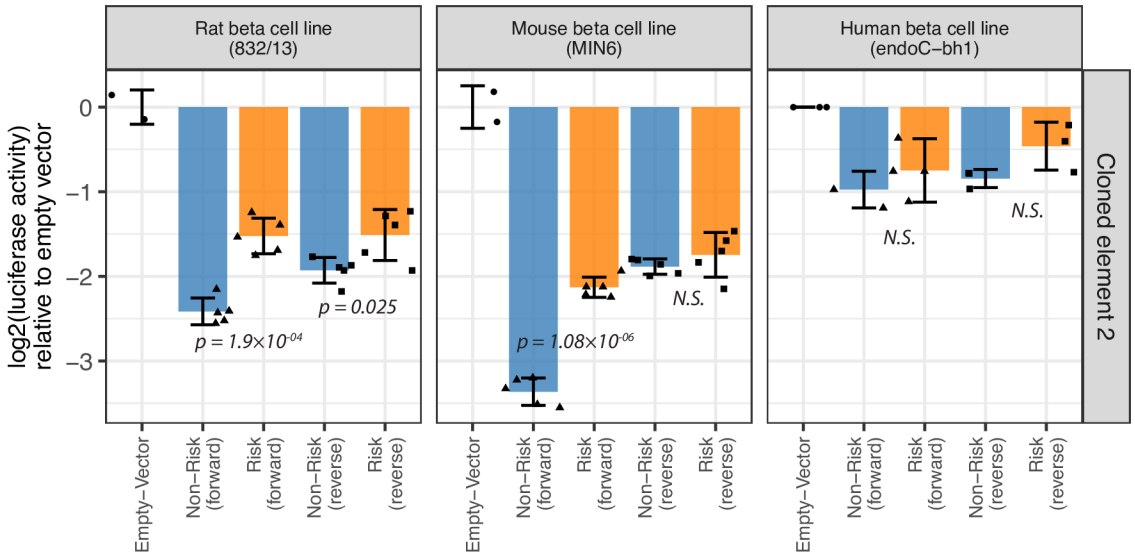


Figure 4.17: Luciferase assay results for *DGKB* 3' eQTL element 2. Log_2 luciferase assay activities (normalized to empty vector) in rat (832/13), mouse (MIN6) and human (endoC) beta cell lines for the element 2 highlighted in blue in Fig. 4.13D, cloned in both forward and reverse orientation with respect to the DGKB gene. P values were determined using unpaired two-sided t-tests.

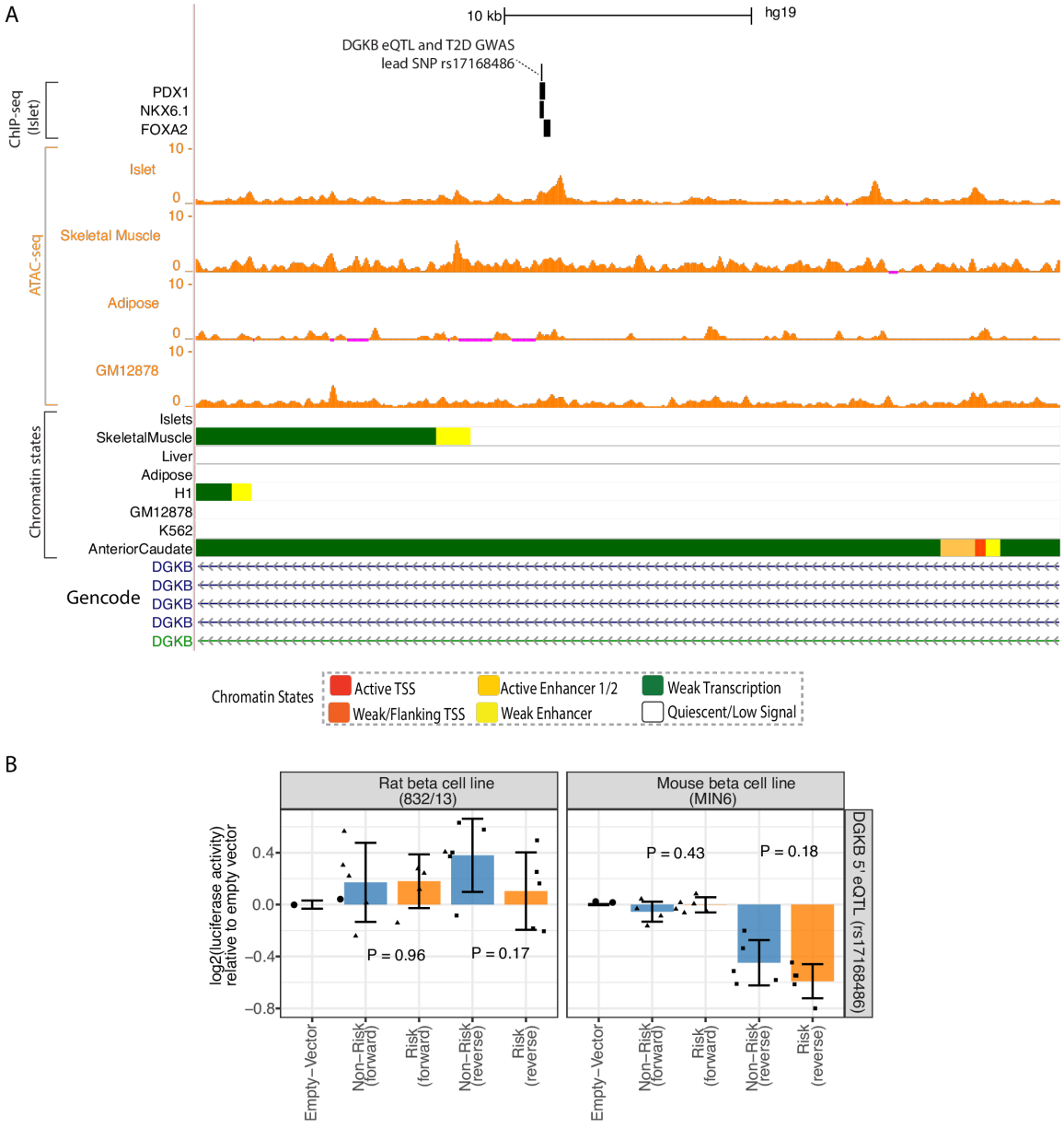


Figure 4.18: 5' *DGKB* eQTL and T2D GWAS lead SNP 17168486 locus. A: Genome browser shot of the 5' *DGKB* eQTL along with ChIP-seq, ATAC-seq and chromatin state profiles in Islets and other tissues. B. Luciferase assay activities (normalized to empty vector) in rat (832/13) and mouse (MIN6) cell lines for the element containing the T2D GWAS and islet eQTL lead SNP (rs17168486), cloned in both forward and reverse orientation with respect to the DGKB gene. Differences between activities of the risk and non-risk allele containing elements were non-significant.

4.4 Discussion

In this manuscript, we used transcriptome sequencing in 420 human islet preparations to address issues that are of general relevance to the mechanistic interpretation of non-coding association signals detected by GWAS, in addition to their specific importance for T2D. We documented the degree to which RNA-sequencing of a disease-relevant tissue missing from a reference data set (e.g. GTEx) provides access to a more complete survey of eQTLs active in islets. We used this information to extend the number of GWAS signals for T2D and related glycemic traits that have been shown to co-localise with islet eQTLs, providing clues to potential effector transcripts at several of these loci. We have demonstrated how tissue heterogeneity (cellular heterogeneity within the tissue of interest, and contamination with cells that are not of direct relevance) can complicate the interpretation of eQTLs co-localizing with GWAS signals. We also integrated our eQTL catalogue with islet epigenomic data to reveal effect size heterogeneity based on local chromatin context and to infer in vivo TF directional activities. Finally, we used our results to nominate and experimentally test causal SNPs at the DGKB locus, which displays coordinated regulatory effects at two statistically independent T2D GWAS signals.

Several lines of evidence including analysis of the physiological association patterns of T2D-associated alleles, and genome-wide enrichment analyses indicate that many, though by no means all, established T2D association signals act through the islet [206, 192, 130, 159]. One of the major motivations behind this study was to bring an enhanced islet eQTL analysis to bear on the challenge of delivering robust mechanistic inference to non-coding GWAS signals, with particular emphasis on the identity of the effector transcripts that may mediate the downstream consequences of the associated variants. At DGKB, evidence that both the T2D signals co-localize with eQTLs with directionally-consistent impacts on DGKB expression in islets lends

support to a causal role for DGKB in T2D predisposition.

However, it is important to emphasize some of the complexities of accurate inference from the coincidence of eQTLs and GWAS signals. First, the RNA-seq data from which these analyses are derived from human islets maintained in culture in basal glycemic conditions. eQTL signals that are restricted to a subset of the cells within those islets would have been hard to detect, and the same would be true for genes whose expression is dependent on stimulation. Genes that mediate T2D-risk through an impact on islet development may be under different transcriptional control in adult islets: in some circumstances, this may incriminate co-localizing eGenes that are not directly responsible for the phenotype. Similarly, given that not all T2D loci act through the islet, some of the eQTLs detected may reflect tissue-specific regulation that is not germane to the development of the diabetic phenotype. Reassuringly, for the co-localizing loci we detected, we were able to perform analyses that are generally supportive of the idea that their T2D effects are mediated through islet dysfunction. For example, the islet eQTLs we detected were enriched in the subset of T2D and glycemic loci for which the patterns of GWAS association indicate a primary effect on insulin secretion.

Second, the confident assignment of co-localization can be difficult. There are a diversity of algorithms to measure the evidence that two association signals (here, a trait GWAS and an eQTL signal) are likely to reflect the same causal variants, but agreement between them is not complete. An additional challenge arises from the complex architecture of many GWAS signals that feature multiple overlapping signals that require conditional decomposition before co-localization can be accurately assigned. This is likely to be especially important when the sets of GWAS and *cis*-eQTL signals at a given locus are not completely overlapping, such that clear

co-localization at one of the contributing signals, can be masked by differences in the overall shape of the association signals that confounds simplistic analysis.

Third, recent studies have shown that functionally constrained genes those that are depleted for missense or loss of function variants are less likely to have eQTLs, suggesting uniform intolerance of both regulatory and coding variation at the same genes [95, 55]. Complementary studies focusing on regulatory elements have shown that large, cell-specific stretch enhancers harbor smaller effect size eQTLs than ubiquitous promoter regions 42 and that genes with more cognate enhancer sequence are depleted for eQTLs43. The results we report are consistent with these observations: we have shown that islet eQTLs that map to the islet stretch enhancers most frequently implicated in GWAS regions have smaller eQTL effect sizes (and therefore may be more difficult to detect). One consequence, for example, is that, at a GWAS variant that has regulatory impact on multiple *cis*-genes, eQTL signals for bystander genes (those not directly implicated in disease pathogenesis) may be easier to detect than those that are actually mediating the signal.

Finally, it is important to emphasize that, even when co-localization has been robustly demonstrated between a GWAS signal and a tissue-appropriate eQTL signal, this does not of itself implicate the eGene concerned as mediating disease predisposition. Causal relationships other than 'variant to gene to disease' are possible, including the possibility the variant has separate (horizontally) pleiotropic effects on both. Growing understanding of the extent of shared local regulatory activity and regulatory pleiotropy makes such an alternative explanation all the more credible. In our view, it is best to regard the genes highlighted by coincident GWAS and eQTL signals as 'candidate' effector transcripts, and to proceed to experimental approaches that enable direct tests of causality. These may involve perturbing the gene across

a range of disease-relevant cell-lines and animal models, and determining the impact on phenotypic readouts that represent reliable surrogates of disease pathophysiology.

4.5 Materials and Methods

4.5.1 Pancreatic Islet sample collection and processing

Geneva samples: Islet sample procurement, mRNA processing and sequencing procedure has been described in [137]. Briefly, Islets isolated from cadaveric pancreas were provided by the Cell Isolation and Transplant Center, Department of Surgery, Geneva University Hospital (Drs. T. Berney and D. Bosco) through the Juvenile Diabetes Research Foundation (JDRF) award 31-2008-416 (ECIT Islet for Basic Research Program). mRNA was extracted using RLT buffer (RNeasy, Qiagen) and total RNA was prepared according to the standard RNeasy protocol. The original RNA libraries were 49-bp paired-end sequenced however, in order to allow joint analysis with the other available datasets for this study, mRNA samples were re-processed using a 100-bp paired-end sequencing protocol. The library preparation and sequencing followed customary Illumina TruSeq protocols for next generation sequencing as described in the original paper [137]. All procedures followed ethical guidelines at the University Hospital in Geneva.

Lund Samples: Islet sample procurement, mRNA processing and sequencing procedure has been described in [44]. Along with the 89 islet samples previously published in [44], we included 102 islet samples and processed these uniformly following the same protocol. These islet samples were obtained from 191 cadaver donors of European ancestry from the Nordic Islet Transplantation Programme (http://www.nordicislets.org). Purity of islets was assessed by dithizone staining, while measurement of DNA content and estimate of the contribution of exocrine and endocrine tissue were assessed

as previously described [44]. Total RNA was isolated with the AllPrep DNA/RNA Mini Kit following the manufacturer's instructions (Qiagen), sample preparation was performed using Illumina's TruSeq RNA Sample Preparation Kit according to manufacturer's recommendations. The target insert size of 300 bp was sequenced using a paired end 101 bp protocol on the HiSeq2000 platform (Illumina). Illumina Casava v.1.8.2 software was used for base calling. All procedures were approved by the ethics committee at Lund University. Oxford samples: Samples collected in Oxford and Edmonton that were jointly sequenced in Oxford are included in this set of samples. Islet sample procurement, mRNA processing and sequencing procedure has been described in 16. To the 117 samples previously published (78 from Edmonton and 39 from Oxford), 57 samples were added and processed following similar protocols as before (27 from Edmonton and 30 from Oxford). Briefly, freshly isolated human islets were collected at the Oxford Centre for Islet Transplantation (OXCIT) in Oxford, or the Alberta Diabetes Institute IsletCore (www.isletcore.ca) in Edmonton, Canada. Additional islets were obtained from the Alberta Diabetes Institute IsletCore's long-term cryopreserved biobank. Freshly isolated islets were processed for RNA and DNA extraction after 13 days in culture in CMRL media. Cryopreserved samples were thawed as described 45 [Lyon, et al., Endocrinology, 2016]. RNA was extracted from human islets using Trizol (Ambion, UK or Sigma Aldrich, Canada). To clean remaining media from the islets, samples were washed three times with phosphate buffered saline (Sigma Aldrich, UK). After the final cleaning step 1 mL Trizol was added to the cells. The cells were lysed by pipetting immediately to ensure rapid inhibition of RNase activity and incubated at room temperature for ten minutes. Lysates were then transferred to clean 1.5 mL RNase-free centrifuge tubes (Applied Biosystems, UK). RNA quality (RIN score) was determined using an Agilent 2100 Bioanalyser (Agilent, UK), with a RIN score > 6 deemed acceptable for inclusion in the study. Samples were stored at -80C prior to sequencing. Poly-A selected libraries

were prepared from total RNA at the Oxford Genomics Centre using NEBNext ultra directional RNA library prep kit for Illumina with custom 8bp indexes 46. Libraries were multiplexed (3 samples per lane), clustered using TruSeq PE Cluster Kit v3, and paired-end sequenced (100nt) using Illumina TruSeq v3 chemistry on the Illumina HiSeq2000 platform. All procedures were approved by the Human Research Ethics Board at the University of Alberta (Pro00013094), the University of Oxford's Oxford Tropical Research Ethics Committee (OxTREC Reference: 215), or the Oxfordshire Regional Ethics Committee B (REC reference: 09/H0605/2). All organ donors provided informed consent for use of pancreatic tissue in research.

USA samples: Islet sample procurement, mRNA processing and sequencing has been described in [196]. Briefly, 39 Islet samples from organ donors were received from the Integrated Islet Distribution Program, the National Disease Research Interchange (NDRI), and Prodo- Labs. Total RNA from 2000-3000 islet equivalents (IEQ) was extracted and purified using Trizol (Life Technologies). RNA quality was confirmed with Bioanalyzer 2100 (Agilent); samples with RNA integrity number (RIN) > 6.5 were prepared for mRNA sequencing. We added the ERCC spike-in controls (Life Technologies) to one microgram of total RNA. PolyA+, stranded mRNA RNA-sequencing libraries were generated for each islet using the TruSeq stranded mRNA kit according to manufacturer's protocol (Illumina). Each islet RNA-seq library was barcoded, pooled into 12-sample batches, and sequenced over multiple lanes of HiSeq 2000 to obtain an average depth of 100 million 2 x 101 bp sequences. All procedures followed ethical guidelines at the National Institutes of Health (NIH.)

4.5.2 Beta-cell sample collection and processing

Sample collection, mRNA processing and sequencing procedure has been described in [137]. To the 11 FAC sorted beta-cells population samples previously published,

we added 15 more samples that were processed following the same protocols. Briefly, islets were dispersed into single cells, stained with Newport Green, and sorted into 'beta' and 'non-beta' populations as described previously [143]. The proportion of beta (insulin), alpha (glucagon), and delta (somatostatin) cells in each population (as percentage of total cells) was determined by immunofluorescence. mRNA extractions as well as sequencing followed the same details described for islets samples processing for the Geneva samples.

4.5.3 Read-mapping and exon quantification

The 100-bp sequenced paired-end reads were mapped to the GRCh37 reference genome with GEM [116]. Exon quantifications were calculated for all elements annotated in GENCODE v19 [61], removing genes with more than 20% zero read count. All overlapping exons of a gene were merged into meta-exons with identifier of type ENSG000001.1_exon.start.pos_exon.end.pos, as described in [94]. Read counts over these elements were calculated without using read pair information, except for excluding reads where the pairs mapped to two different genes. We counted a read in an exon if either its start or end coordinates overlapped an exon. For split reads, we counted the exon overlap of each split fragment, and added counts per read as 1/(number of overlapping exons per gene). Gene level quantifications used the sum of all reads mapped to exons from the gene. Genes with more than 20% zero read counts were removed.

4.5.4 Genotype imputation

Genotypes for all islet samples, including 19 beta-cell samples, were available from omniexpress and omni2.5 genotype arrays. Quality of genotyping from the shared SNPs in both arrays was assessed before imputation separately by removing SNPs as follows: 1) SNPs with minor allele frequency (MAF) $< 5\%$; 2) SNP geno-

type success rate <95%; 3) Palindromic SNPs with MAF > 40%; 4) HWE < 1e-6; 5) Absence from 1000G reference panel; 6) Allele inconsistencies with 1000G reference panel; 7) Probes for same rsID mapping to multiple genomic locations (1000G reference-consistent probe kept). Finally, samples were excluded if they had an overlap genotype success rate lower than 90%; and MAF differences larger than 20% compared to the 1000G reported european MAF.

The two panels were separately pre-phased with SHAPEIT v2 [34] using the IMPUTE2-supplied genetic maps. After pre-phasing the panels were imputed with IMPUTE2 v2.3.1 [72] using the 1000 Genomes Phase I integrated variant set (March 2012) as the reference panel. SNPs with INFO score > 0.4 and HWE p > 1e-6 (for chrX this was calculated from female individuals only) from each panel were kept. A combined vcf for each chromosome was generated from the intersection of the checked variants in each panel. Directly genotyped SNPs with a MAF < 1% (including the exome-components of the chips not shared between all centres) were merged into the combined vcfs: i) If SNPs were not imputed they were added and ii) If SNPs had been imputed, the imputed calls for the individual were replaced by the typed genotype. Dosages were calculated from the imputation probabilities (genotyped samples) or genotype calls (WGS samples). For the 22 autosomes the dosage calculation was: 2x((0.5*heterozygous call) + homozygous alt call). For chromosome X (where every individual should be functionally hemizygous), the calculation was: (0.5*heterozygous call) + homozygous alt call). Genotype calls for males can only be '0/0' and '1/1'. The total number of variants available for analysis after quality assessment was 8,056,952.

For the 26 beta cell samples, 19 had genotypes available from omniexpress arrays, whereas 7 had the DNA sequence available. Variant calling from DNA sequence has

been previously described in [137]. Briefly, the Genome Analysis Toolkit (GATK) 1.5.31 [121] was used following the Best Practice Variant Detection v3 to call variants. Reads were aligned to the hg19 reference genome with BWA [97]. We used a confidence score threshold of 30 for variant detection and a minimum base quality of 17 for base calling. Good confidence (1% FDR) SNP calls were then imputed on the 1000 Genomes reference panel and phased with BEAGLE 3.3.2 56. Imputation of variants from samples with arrays genotyping were imputed together with genotypes from individuals with islets samples as described before and then merged with genotypes from DNA sequences. SNPs with INFO score > 0.4, HWE p $>$ 1e-6 and MAF $> 5\%$, were kept for further analysis. The total number of variants available for analysis after quality assessment was 6,847,993.

4.5.5 RNAseq quality assessment and data normalization

Heterozygous sites per sample were matched with genotype information to confirm the ID of the samples [177]. 11 samples did not match with their genotypes, 6 of which would be corrected by identifying a good match. For the remaining samples, no matches were found on the genotypes and they were removed from the dataset, giving a total of 420 samples with genotypes. Raw read counts from exons and genes were scaled to 10 million to allow comparison between samples with different libraries. Scaled raw counts were then quantile normalized. We used principal component analysis (PCA) to evaluate the effects of unwanted technical variation and the expected batch effects due to fact that the islet sample processing mRNA sequencing was performed across four labs. We evaluated a) the optimal number of principal components (PCs) for the discovery of eQTLs and b) the minimum number of PCs necessary to control for laboratories of origin batch effects (Fig. 4.2). We performed eQTL discovery controlling for 1,5,10, 20 30 40 and 50 PCs for expression, as well as gender, 4 PCs derived from genotype data, and a variable defining the laboratory of origin (coded

as: OXF, LUND, GEN and USA). After evaluation of the results, we conclude that controlling for 20 PCs was optimal. To ensure that we controlled for batch effects with these variables, we used a permutation scheme as follows: expression sample labels and expression covariates were permuted within each of the 4 laboratories before performing a standard eQTL analysis against non-permuted genotypes (and matched PCs for genotypes) using different numbers of PCs for expression. Significant eQTLs in any of these analyses are considered a false positive due to technical differences across laboratories of origin of the samples. Our results indicate that 10PCs were sufficient to minimize the number of false positives due to batch effects originating from differences in processing of the islet samples (Fig. 4.2).

4.5.6 eQTL analysis

eQTL analysis for islets and beta-cells were performed using fastQTL [139] on 420 islets samples and 26 beta-cells samples with available genotypes. *Cis*-eQTL analysis was restricted to SNPs in a 1MB window upstream and downstream the transcription start site (TSS) for each gene and SNPs with MAF>1%. For the analysis of beta-cell samples, we used a filter of MAF>5%. Exon-level eQTLs identified best exons-SNP association per gene (using the group flag), while gene level eQTLs used gene quantifications and identified the best gene-SNP association. Variables included in the linear models were the first 4 PCs for genotypes, the first 25 PCs for expression, gender and a variable identifying the laboratory of origin of the samples. Significance for the SNP-gene association was assessed using 1000 permutations per gene, correcting P values with a beta approximation distribution 18. Genome-wide multiple testing correction was performed using the q-value correction [173] implemented in largeQvalue [14].

To discover multiple independent eQTLs, we applied a stepwise regression proce-

dure as described in [15]. Briefly, we started from the set of eGenes discovered in the first pass of association analysis (FDR $< 1\%$). Then, the maximum beta-adjusted P value (correcting for multiple testing across the SNPs and exons) over these genes was taken as the gene-level threshold. The next stage proceeded iteratively for each gene and threshold. A *cis*-scan of the window was performed in each iteration, using 1,000 permutations and correcting for all previously discovered SNPs. If the beta adjusted P value for the most significant exon-SNP or gene-SNP (best association) was not significant at the gene-level threshold, the forward stage was complete and the procedure moved on to the backward step. If this P value was significant, the best association was added to the list of discovered eQTLs as an independent signal and the forward step proceeded to the next iteration. The exon level *cis*-eQTL scan identified eQTLs from different exons, but reported only the best exon-SNP in each iteration. Once the forward stage was complete for a given gene, a list of associated SNPs was produced which we refer to as forward signals. The backward stage consisted of testing each forward signal separately, controlling for all other discovered signals. To do this, for each forward signal we ran a *cis* scan over all variants in the window using fastQTL, fitting all other discovered signals as covariates. If no SNP was significant at the gene-level threshold the signal being tested was dropped, otherwise the best association from the scan was chosen as the variant that represented the signal best in the full model.

4.5.7 GTEx eQTLs

We identified exon level eQTLs for 44 GTEx tissues using fastQTL 18 following the same procedure as for the islet eQTLs. Covariates included followed the previously published number of PCs for expression [58] and included 15 PCs for expression for tissues with less than 154 samples; 30 PCs for samples between 155 and 254 samples; and 35 PCs for samples with more than 254 samples. Independent eQTLs from

exons were identified as described for islets eQTLs. The proportion of shared eQTLs between islet and beta-cell eQTLs and the eQTLs from GTEx tissues were identified using Π1 [173].

4.5.8 Tissue de-convolution

To identify the contribution of the beta-cells, non-beta cells and exocrine components (non-islets cell) expression to the total gene expression measure in islets we performed an expression deconvolution analysis. Expression profiles from GTEx whole pancreas was used as a model for the exocrine component of expression [58], while FAC-sorted expression profiles from beta-cell and non-beta-cells from Nica et al [137] were used to identify the fraction of expression derived from islets cells. First, we performed differential expression analysis of a) exocrine versus whole islet samples; b) beta-cell versus whole islet samples; c) non-beta-cell versus whole islet samples. The top 500 genes from each analysis were combined, and a deconvolution matrix of log2-transformed median expression values was prepared for each cell type. Next, deconvolution was performed using the Bioconductor package DeconRNASeq 61.Deconvolution values per sample are included in the covariates file, together with the expression values in the EGA submission.

4.5.9 Enrichment of eQTLs in T2D and glycemic GWAS

To perform an enrichment analysis of T2D and glycemic traits GWAS associations among eQTLs across tissues, we examined 78 T2D associated signals [49], and 44 variants from associations with continuous glycemic traits relevant to T2D predisposition (including fasting glucose, and beta-cell function (HOMA-B) in non-diabetic individuals) [163, 175, 114]. For each GWAS lead variant, we extracted the eQTL with the greatest absolute effect size estimate from the results for all GTEx tissues and the InsPIRE pancreatic islets. We then compared their observed effect size es-

timates to those derived from a null distribution of 15,000 random variants matched in terms of the number of SNPs in LD, distance to TSS, number of nearby genes and minor allele frequency. For comparison with results observed for T2D loci, we also included the set of 50 lead variants implicated by GWAS in predisposition to T1D [138].

4.5.10 Co-localization of islet eQTL with T2D GWAS

Co-localization of GWAS variants and eQTLs were performed using both COLOC [PMID: 24830394] and RTC [ref]. For the analysis using COLOC, all variants within 250 kilobase flanking regions around the index variants were tested for co-localization using default parameters from the software were used on summary statistics from T2D GWAS from [164] and fasting glucose [113]. GWAS variants and eSNPs pairs were considered to co-localize if the COLOC score for shared signal was larger than 0.9. RTC analysis was also performed using defaults parameters from the software with a list of 86 lead GWAS variants for T2D and fasting glucose. Associations between GWAS and gene expression were considered as co-localizing if RTC score was larger than 0.9.

4.5.11 Chromatin states, Islet ATAC-seq and Transcription factor (TF) footprints

We used a previously published 13 chromatin state model that included Pancreatic Islets along with 30 other diverse tissues [196]. Briefly, these chromatin states were generated from cell/tissue ChIP-seq data for H3K27ac, H3K27me3, H3K36me3, H3K4me1, and H3K4me3, and input control from a diverse set of publically available data [142, 186, 43, 125] using the ChromHMM program 65. Chromatin states were learned jointly from 33 cell/tissues that passed QC by applying the ChromHMM (version 1.10) hidden Markov model algorithm at 200-bp resolution to five chromatin

marks and input 12. We ran ChromHMM with a range of possible states and selected a 13-state model, because it most accurately captured information from higher-state models and provided sufficient resolution to identify biologically meaningful patterns in a reproducible way. As reported previously [196], Stretch Enhancers were defined as contiguous enhancer chromatin state (Active Enhancer 1 and 2, Genic Enhancer and Weak Enhancer) segments longer than 3kb, whereas Typical Enhancers were enhancer state segments smaller than the median length of 800bp [142].

We used the union of ATAC-seq peaks previously identified from two human islet samples called using MACS2 v2.1.0 [196]. We also used previously published TF footprints that were generated in a haplotype-aware manner using ATAC-seq and genotyping data from the phased, imputed genotypes for each of two islet samples using vcf2diploid v0.2.6a [196].

4.5.12 Filtering eQTL SNPs for epigenomic analyses

Since low MAF SNPs, due to low power, can only be identified as significant eQTL SNP (eSNPs) with high eQTL effect sizes (slope or the beta from the linear regression), we observed that absolute effect size varies inversely with MAF Fig. 4.19. To conduct eQTL effect size based analyses in an unbiased manner, we selected significant (FDR 1%) eSNPs with MAF >= 0.2. We then pruned this list to retain the most significant SNPs with pairwise LD(r^2)<0.8 for the EUR population using PLINK 66 and 1000 genomes variant call format (vcf) files (downloaded from `ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/release/20130502/`) for reference (European population). This filtering process resulted in N=3832 islet eSNPs.

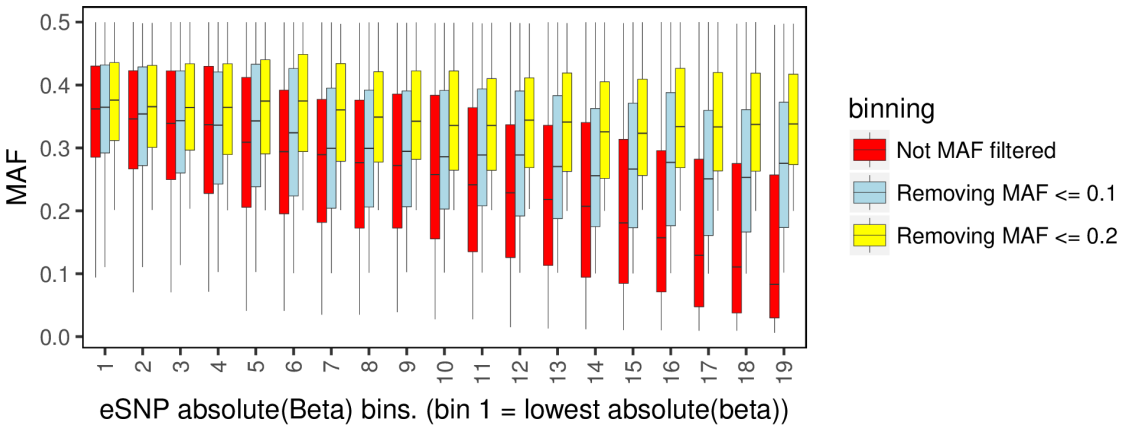


Figure 4.19: MAF filtering for eSNPs. MAF for islet eQTL eSNPs binned by absolute effect size into equal sized, 50% overlapping bins. Bin 1 contains eSNPs with lowest absolute effect sizes, bin 19 contains eSNPs with highest absolute effect sizes.

4.5.13 Enrichment of genetic variants in genomic features

To calculate the enrichment of islet eSNPs to overlap with genomic features such as chromatin states and transcription factor (TF) footprint motifs, we used the GREGOR tool [160]. For each input SNP, GREGOR selects 500 control SNPs matched for MAF, distance to the gene, and number of SNPs in LD(r^2) *geq* 0.99. A unique overlap is reported if the feature overlaps any input lead SNP or its LD(r^2)>0.99 LD SNPs. Fold enrichment is calculated as the number unique overlaps over the mean number of loci at which the matched control SNPs (or their LD(r^2>)0.99 SNPs) overlap the same feature. This process accounts for the length of the features, as longer features will have more overlap by chance with control SNP sets. We used the following parameters in GREGOR for eQTL enrichment: r^2 threshold (for inclusion of SNPs in linkage disequilibrium (LD) with the lead eSNP) = 0.99, LD window size = 1Mb, and minimum neighbor number = 500.

For enrichment of T2D GWAS SNPs in islet chromatin states, we downloaded the list of T2D GWAS SNPs from [110]. We pruned this list to retain the most sig-

nificant SNPs with pairwise LD(r^2)<0.2 for the EUR population using PLINK [22] and 1000 genomes vcf files (downloaded from `ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/release/20130502/`) for reference (European population). This filtering process resulted in N=378 T2D GWAS SNPs. We used GREGOR to calculate enrichment using the following specific parameters: r^2 threshold (for inclusion of SNPs in linkage disequilibrium (LD) with the lead eSNP) = 0.8, LD window size = 1Mb, and minimum neighbor number = 500

We investigated if footprint motifs were more enriched to overlap eQTL of high vs low effect sizes. We sorted the filtered (as described above) eQTL list by absolute effect size values and partitioned into two equally sized bins (N eSNPs = 1,916). Since TF footprints were available for a large number of motifs (N motifs = 1,995), the enrichment analysis had a large multiple testing burden and limited power with 1,916 eSNPs in each bin. Therefore, we only considered footprint motifs that were significantly enriched (FDR <1%, Benjamini & Yekutieli method from R p.adjust function, N motifs = 283) to overlap the bulk set of eSNPs (LD r^2<0.8 pruned but not MAF filtered, N eSNPs = 6,468) for enrichment to overlap the binned set of eSNPs. This helped reduce the multiple testing burden. We then calculated enrichment for the selected footprints to overlap SNPs in each bin using GREGOR with same parameters as described above.

4.5.14 eSNP effect size distribution in chromatin states and ATAC-seq peaks within chromatin states

We identified the islet eQTL eSNPs (after LD pruning and MAF filtering as described above) occurring in chromatin states or ATAC-seq peaks within chromatin states using BEDtools intersect [154]. Similar to the enrichment calculation procedure, we considered a unique eQTL overlap if the lead eSNP or a proxy SNP with

LD(r^2)>0.99 occurred in these regions. We considered the effect size as the slope or the beta from the linear regression for the eQTL overlapping each region. P values were calculated using the Wilcoxon Rank Sum Test in R.

4.5.15 TF motif directionality analysis

For TF footprint motifs that were significantly enriched to overlap the full set of islet eQTLs (after LD pruning to r^2<0.8) with (FDR 1%, Benjamini & Yekutieli method from R p.adjust function, N motifs = 283), we determined the overlap position of the eSNP (pruned LD r^2<0.8 lead eSNPs and their LD r^2>0.99 proxy SNPs) with each TF footprint motif. We considered instances where the eSNP overlapped the TF footprint motif at a position with information content >= 0.7 and either the eSNP effect or the non-effect allele was the most preferred base in the motif. We selected TF footprint motifs that had 10 or more such eSNP overlap instances (N=278). For each TF footprint motif and eSNP overlap, we re-keyed the direction of effect on the target gene being positive or negative with respect to the most preferred base in the motif. For each TF motif, we compiled the fraction of instances where the SNP allele that was most preferred in the TF footprint motif (i.e. base with highest probability in the motif) associated with increased expression of the associated gene. We refer to this metric as the motif directionality fraction where fraction near 1 suggests activating and fraction near 0 suggests repressive preferences towards the target gene expression. Motif directionality fraction near 0.5 suggests no activity preference or context dependence. We compared our results to a previously published study that quantified transcription activating or repressive activities based on massively parallel reported assays in HepG2 and K562 cells 27. We then considered 99 motifs from our analyses that were reported to have significant (P<0.01) activating or repressive scores from MPRAs in both HepG2 and K562. With the null expectation of the motif directionality fraction being equal to 0.5, i.e. TF binding equally likely to increase or

decrease target gene expression, we used a binomial test to calculate TF that show significant deviation from the null (N = 8 at FDR < 10%).

4.5.16 Cell culture

MIN6 mouse insulinoma beta cells [128] were grown in Dulbecco's modified Eagle's Medium (Sigma-Aldrich, St. Louis, Missouri/USA) with 10% fetal bovine serum, 1 mM sodium pyruvate, and 0.1 mM beta-mercaptoethanol. INS-1-derived 832/13 rat insulinoma beta cells (a gift from C. Newgard, Duke University, Durham, North Carolina/USA) were grown in RPMI-1640 medium (Corning, New York/USA) supplemented with 10% fetal bovine serum, 10 mM HEPES, 2 mM L-glutamine, 1 mM sodium pyruvate, and 0.05 mM beta-mercaptoethanol. EndoC-βH1 cells (Endocell) were grown according to (Ravassard et al., 2011) in Dulbecco's modified Eagle's medium (DMEM; Sigma-Aldrich), 5.6mmol/L glucose with 2% BSA fraction V fatty acid free (Roche Diagnostics), 50mol/L 2-mercaptoethanol, 10mmol/L nicotinamide (Calbiochem), 5.5g/ml transferrin (Sigma-Aldrich), 6.7ng/ml selenite (Sigma-Aldrich), 100U/ml penicillin, and 100g/ml streptomycin. Cells were grown on coating consisting of 1% matrigel and 2g/mL fibronectin (Sigma). We maintained cell lines at 37 C and 5% CO2.

4.5.17 Transcriptional reporter assays

To test haplotypes for allele-specific effects on transcriptional activity, we PCR-amplified a 765-bp genomic region (haplotype A) containing variants: rs7798124, rs7798360, and rs7781710, and a second 592-bp genomic region (haplotype B) containing variants: rs10228796, rs10258074, rs2191348, and rs2191349 from DNA of individuals homozygous for each haplotype. We cloned the PCR amplicons into the multiple cloning site of the Firefly luciferase reporter vector pGL4.23 (Promega, Fitchburg, Wisconsin/USA) in both orientations, as described previously [45]. Vectors are

designated as 'forward' or 'reverse' based on the PCR-amplicon orientation with respect to DGKB gene. We isolated and verified the sequence of five independent clones for each haplotype in each orientation. For the 5' eQTL a 250 bp construct containing the rs17168486 SNP (Origene) was subcloned into the Firefly luciferase reporter vector pGL4.23 (Promega) in both orientations.

We plated the MIN6 (200,000 cells) or 832/13 (300,000 cells) in 24-well plates 24 hrs before transfections and the EndoC-βH1 cells (140.000 cells) plated 48H prior to transfection. We co-transfected the pGL4.23 constructs with phRL-TK Renilla luciferase reporter vector (Promega) in duplicate into MIN6 or 832/13 cells and in triplicate for EndoC-βH1 cells. For the transfections we used Lipofectamine LTX (ThermoFisher Scientific, Waltham, Massachusetts/USA) with 250 ng of plasmid DNA and 80 ng Renilla for MIN6 cells, Fugene6 (Promega) with 720 ng of plasmid and 80 ng Renilla for 832/13 cells per each welll and Fugene6 with 700 ng plasmid and 10 ng renilla for EndoC-βH1 cells. We incubated the transfected cells at 37 C with 5% CO2 for 48 hours. We measured the luciferase activity with cell lysates using the Dual-Luciferase Reporter Assay System (Promega). We normalized Firefly luciferase activity to the Renilla luciferase activity. We compared differences between the haplotypes using unpaired two-sided t-tests. All experiments were independently repeated on a second day and yielded comparable results.

4.5.18 Electrophoretic Mobility Shift Assays

Electrophoretic mobility shift assays were performed as previously described [45]. We annealed 17-nucleotide biotinylated complementary oligonucleotides (Integrated DNA Technologies) centered on variants: rs10228796, rs10258074, rs2191348, and rs2191349. MIN6 nuclear protein extract was prepared using the NE-PER kit (Thermo Scientific). To conduct the EMSA binding reactions, we used the LightShift Chemi-

luminescent EMSA kit (Thermo Scientific) following the manufacturer's protocol. Each reaction consisted of 1 g poly(dI-dC), 1x binding buffer, 10 g MIN6 nuclear extract, 400 fmol biotinylated oligonucleotide. We resolved DNA-protein complexes on nondenaturing DNA retardation gels (Invitrogen) in 0.5x TBE. We transferred the complexes to Biodyne B Nylon membranes (Pall Corporation), and UV cross-linked (Stratagene) to the membrane. We used chemiluminescence to detect the DNA-protein complexes. EMSAs were repeated on a second day with comparable results.

4.6 Acknowledgements

This work was a huge team effort, I thank co-authors Dr. Ana Viñuela, Dr. Martijn van de Bunt, Prof. Mark McCarthy and Prof. Steve Parker and all the team members for their contributions towards this project. I specifically contributed towards the analyses of epigenomic data and integration with eQTL data and manuscript preparation.

CHAPTER V

Integrating Enhancer RNA Signatures with Diverse Omics Data to Identify Characteristics of Transcription Initiation in Pancreatic Islets

5.1 Abstract

Identifying active regulatory elements and their characteristics is critical to understand gene regulatory mechanisms and subsequently better delineating biological mechanisms of complex disease/trait predisposition. Studies have shown that many active enhancers are transcribed into enhancer RNA (eRNA). Here, we identify actively transcribed regulatory elements in human pancreatic islets in high genomic resolution by generating eRNA profiles using cap analysis of gene expression (CAGE) across 70 islet samples. We identify >10,000 clusters of CAGE tag transcription start sites (TSS) or tag clusters (TCs) in islets, 20% of which are islet specific when compared to CAGE TCs in other publicly available tissues. Islet TCs are most enriched to overlap GWAS loci for islet-relevant traits such as fasting glucose. We integrated islet CAGE profiles with diverse epigenomic information such as chromatin immunoprecipitation followed by sequencing (ChIP-seq) profiles of five histone modifications

and accessible chromatin profiles from the assay for transposase accessible chromatin followed by sequencing (ATAC-seq), to understand how the underlying chromatin landscape affects transcription. As expected, we observe that transcription largely initiates downstream of ATAC-seq peak summits. We identify that ATAC-seq informed transcription factor binding sites (TF footprint motifs) for the RFX TF family are highly enriched in transcribed regions occurring in enhancer associated chromatin, whereas footprint motifs for the ETS family TFs are highly enriched in transcribed regions with promoter associated chromatin. Using massively parallel reporter assays in a rat pancreatic islet beta cell line, we tested the activity of 3,240 islet TC elements, 70% (2,206) of which show significant regulatory activity (5% FDR). This work provides a high-resolution transcriptional regulatory map of human pancreatic islets.

5.2 Introduction

T2D is a complex disease that is caused due to an interplay of factors such as pancreatic islet dysfunction and insulin resistance in peripheral tissues such as fat and muscle. GWASs to date have identified >240 loci that modulate risk for T2D [110]. However, these SNPs mostly occur in non protein-coding regions and are highly enriched to overlap islet-specific enhancer regions [182, 120, 189, 142, 144, 153]. This suggests the variants likely affect gene expression rather than directly altering protein structure or function. Moreover, due to the correlated structure of common genetic variations across genome, GWAS signals are usually marked by numerous SNPs in high linkage disequilibrium (LD). Therefore, identifying causal SNP(s) is extremely difficult using genetic information alone. These factors have impeded our understanding of the molecular mechanisms by which genetic variants modulate gene expression in orchestrating disease.

In order to understand gene regulatory mechanisms, it is essential to identify regu-

latory elements at high genomic resolution, since these are most likely to house causal variant(s). Active regulatory elements can be delineated by profiling covalent modifications of the histone H3 subunit such as H3 lysine 27 acetylation (H3K27ac) which is associated with enhancer activity, H3 lysine 4 trimethylation (H3K4me3) which is associated with promoter activity, among others. However, such chromatin modification based methods identify regions of the genome that typically span hundreds of base pairs. Since TF binding can affect gene expression, TF accessible regions of the chromatin within these active enhancer and promoter elements can enable identifying the regulatory element at a higher resolution. Numerous studies have utilized this diverse information in islets to nominate causal gene regulatory mechanisms [196, 193, 44, 188, 155, 188].

In the light of identifying active regulatory elements, studies have shown that a subset of enhancers are also transcribed into enhancer RNA (eRNA), and that transcription is a robust predictor of enhancer activity [7, 124]. eRNAs are nuclear, short, mostly-unspliced, 5 capped and non-polyadenylated [7]. eRNAs have generally shown to be bidirectionally transcribed with respect to the regulatory element [85, 123, 7], however, unidirectional transcription at enhancers has also been reported. Previous studies have indicated that these transcripts could be stochastic output of Pol2 and TF machinery at active regions, whereas in some cases, the transcripts could serve important functions such as sequestering TFs or potentially assisting in chromatin looping [79, 73, 100, 209]. Therefore, identifying active sites where transcription initiates can pinpoint active regulatory elements at a higher genomic resolution.

Genome-wide sequencing of 5' capped RNAs using Cap Analysis of Gene Expression (CAGE) can detect transcription start sites (TSSs) and thereby profile transcribed promoter and enhancer regions [85, 7]. CAGE-identified enhancers are two to three times more likely to validate in vitro than non-transcribed enhancers de-

tected by chromatin-based methods [7]. An advantage of CAGE is that it can be applied on RNA samples from hard to acquire biological tissue such as islets and does not require live cells that are imperative for other TSS profiling techniques such as GRO-cap seq [28, 27, 105]. The functional annotation of the mammalian genome (FANTOM5) project [183] has generated an exhaustive CAGE expression atlas across 573 primary cell types and tissues, including the pancreas. However, islets, that secrete insulin and are relevant for T2D and related traits, constitute only 1% of the pancreas tissue. Therefore, pancreas transcriptome cannot accurately represent the islet enhancer transcription landscape. Motivated by these reasons, we profiled the islet transcriptomes using (CAGE). Here, we present the islet CAGE TSS atlas in pancreatic islets and complement the omics compendium for the tissue.

5.3 Results

5.3.1 The CAGE landscape in human pancreatic islets

We analyzed transcriptomes in 70 human pancreatic islet samples obtained from unrelated organ donors by employing CAGE on total RNA from each sample. To enrich for the non poly-adenylated and short in size (<1kb) eRNA transcripts [7], we performed polyA depletion and small fragment size selection (<1kb, methods CAGE library preparation) to enrich for the eRNA transcript fraction. CAGE libraries were prepared according to the non-Amplified non-Tagging Illumina Cap Analysis of Gene Expression (MAF) protocol [135], and an 8 bp unique molecular identifier was added to identify PCR duplicates. We sequenced CAGE libraries, performed quality control (QC) and mapped to the hg19 genome and identified transcription start sites (TSSs) or CAGE tags. We selected 51 samples that passed our QC measures (see methods) for all further analyses. To identify regions with high density of transcription initiation events, we called clusters of CAGE tags or tag clusters (TCs) using the paraclu [46]

method in each islet sample. We then identified a consensus set of aggregated islets TCs by merging TCs across samples in a strand-specific manner and retaining TC segments that were supported by at least 10 individual samples (Methods, Fig. 5.1). We identified 10,373 tag clusters with median length of 191 bp (Fig. 5.2), spanning a total genomic territory of 2.5 Mb. To analyze characteristics of islet TCs and explore the chromatin landscape underlying these regions, we utilized publicly available ChIP-seq data for five histone modifications along with ATAC-seq data in islets [196]. We integrated the datasets for histone modifications, namely, promoter associated H3K4me3, enhancer associated H3K4me1, active promoter and enhancer associated H3K27ac, transcribed gene-associated H3K36me3 and repressed chromatin associated H3K27me3 across islets and analyzed the data jointly with corresponding publicly available ChIP-seq datasets for Skeletal Muscle, Adipose and Liver (included for other ongoing projects) using ChromHMM [40, 43, 41]. This analysis produced 11 unique and recurrent chromatin states (Fig. 5.3), including promoter, enhancer, transcribed, and repressed states. Fig. 5.4A shows an example locus in the intronic region of the *ST18* gene where a TC identified in islets overlaps an active TSS chromatin state and an ATAC-seq peak. The regulatory activity of this element was validated by the VISTA project in an *in vivo* reporter assay in mouse embryos [198].

We next compared the islet TCs with CAGE peaks identified across across diverse cell and tissue types by FANTOM project. Using CAGE profiles across hundreds of cell/tissues, the FANTOM project identified peaks using a decomposition-based peak identification (DPI) method [183], following which a set of robust peaks were defined that included a CAGE TSS with more than 10 read counts and 1 TPM (tags per million) in at least one sample. We observed that 77.8% of Islet TCs segments overlapped (at least 1bp) with FANTOM robust peaks, and the total overlapping region comprised 24% of the total Islet TC territory (Fig. 5.4B). To compare islet TCs with individual FANTOM tissues, we identified TCs in each FANTOM human

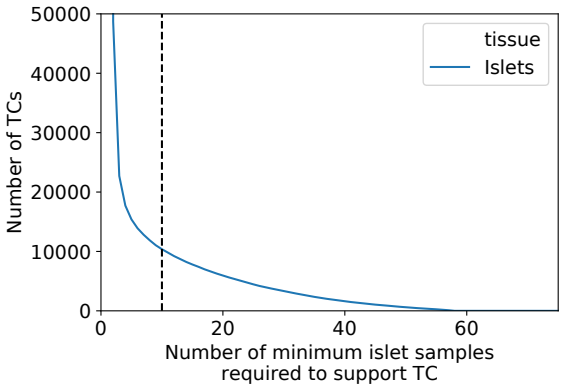


Figure 5.1: TC segments called using the paraclu in each of the 51 islet samples were merged in a strand specific manner. Shown here is the number of merged TC segments that overlap TC territory in x or more islet samples. We required support in a minimum of 10 islet samples to include a TC segment in the aggregate list of islet TCs.

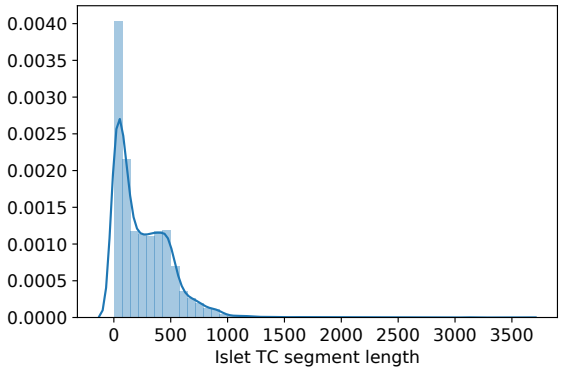


Figure 5.2: Distribution of islet TC lengths.

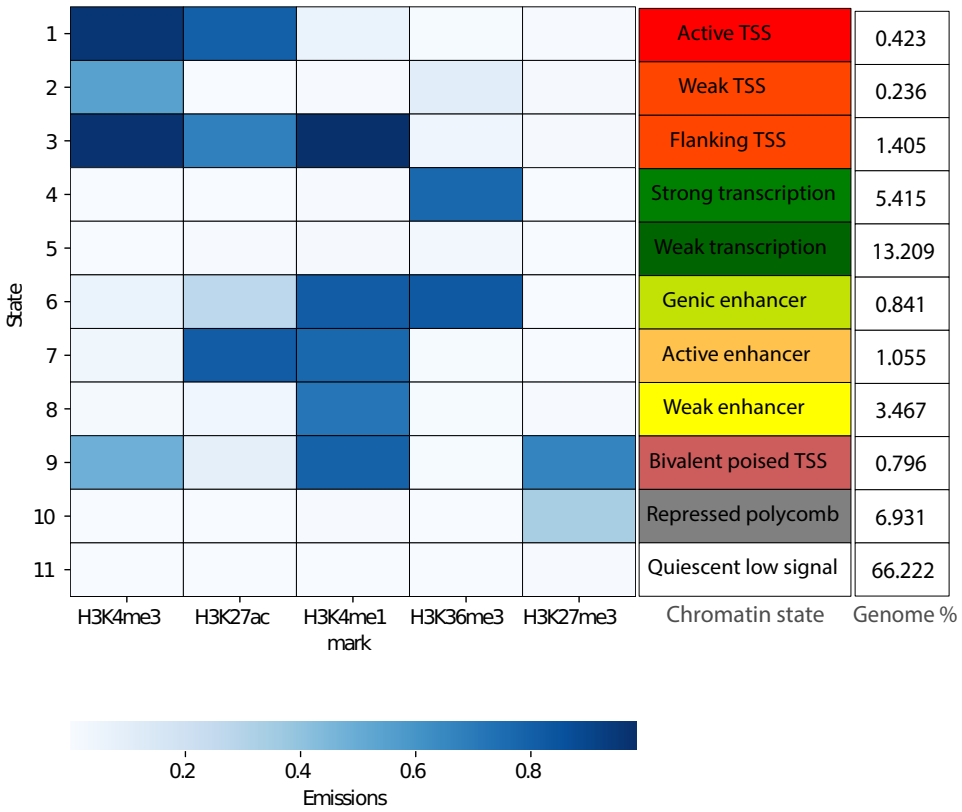


Figure 5.3: 11 chromatin state model. Shown are the emission probabilities of each of the five histone marks, chromatin state annotation and the percent genome coverage of each state

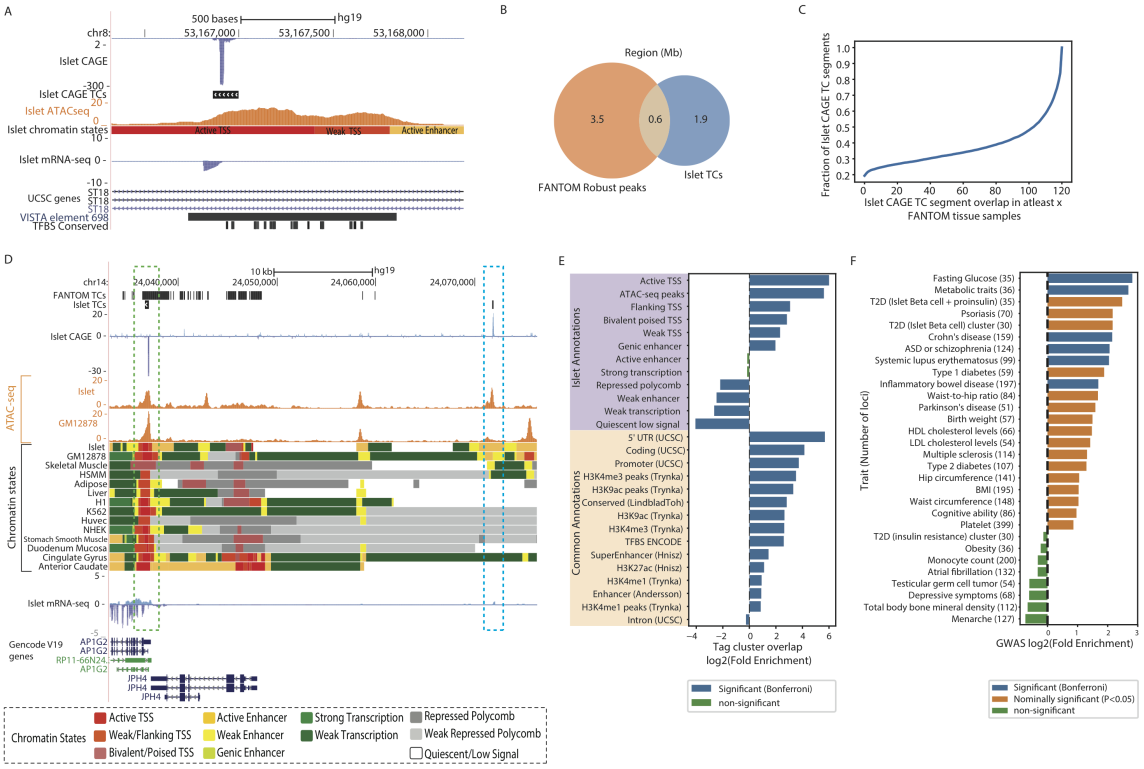


Figure 5.4: Islet CAGE tag cluster identification. A: Genome browser view of the intronic region of the *ST18* gene as an example locus where a TC is identified in islets that overlaps an islet ATAC-seq peak and an active TSS chromatin state. This TC also overlaps an enhancer element validated by the VISTA project [198]. B: Base-pair level overlap between islet CAGE TC territory and FANTOM robust CAGE peaks. C: Distribution of the number of tissues in which TCs identified by the FANTOM project overlap each islet TC segment. D: Genome browser view of an example locus near the *AP1G2* gene that highlights an islet TC that is also identified in FANTOM tissues (FANTOM TCs track is a dense depiction of TCs called across >120 tissues) (green box), occurs in a ATAC-seq peak region in both islets and lymphoblastoid cell line GM12878 (ATAC-seq track) and overlap active TSS chromatin states across numerous other tissues. Another islet TC 34 kb distal to the *AP1G2* gene is not identified as a TC in other FANTOM tissues, occurs in an islet ATAC-seq peak and a more islet-specific active enhancer chromatin state region (blue box). E: Enrichment of islet TCs to overlap islet chromatin state annotations and other common annotations. F: Enrichment of islet TCs to overlap GWAS loci of various disease/traits. Number of loci for each trait are noted in parentheses.

tissue using the paraclu method similarly as islets. For each Islet TC segment, we then calculated the number of FANTOM tissues in which TCs overlapped the islet segment. We observed that 20% of Islet TCs were unique, whereas about 60% of segments were shared across 60 or more tissues (Fig. 5.4C). We highlight an example locus where an islet TC in the *AP1G2* gene occurs in active TSS chromatin states across multiple tissues, and overlaps shared ATAC-seq peaks in islet and the lymphoblastoid cell line GM12878 [17] (Fig. 5.4D). This region was also identified as a TC in FANTOM tissues (Fig. 5.4D, green box). Another islet TC 34kb away however occurs in a region lacking gene annotations, and overlaps a more islet-specific active enhancer chromatin state and ATAC-seq peak (Fig. 5.4D, blue box). This region was not identified as a TC in the 120 FANTOM tissues that were analyzed. These data highlight that islet TCs comprise both shared and also islet-specific sites of active transcription initiation.

We next asked if islet TCs preferentially occurred in certain genomic annotations. We computed the enrichment of islet TCs to overlap islet annotations such as active TSS and other chromatin states and islet ATAC-seq peaks. We also included static annotations such as known gene promoters, coding, untranslated (UTR) regions, or annotations such as super enhancers, or histone ChIP-seq peaks that were aggregated across multiple cell types. We observed that Islet TCs were highly enriched to overlap Islet active TSS states (> 60 fold, P=0.0001, Fig. 5.4E). This result is largely expected since CAGE profiles transcription start sites where the underlying chromatin is more likely to look like active TSS. TCs were also enriched to overlap islet ATAC-seq peaks, which signifies that these regions of transcription initiation are bound by TFs, and 5 untranslated regions (UTRs).

To gauge if these transcribed elements could be relevant for diverse disease/traits, we computed enrichment for TCs to overlap GWAS loci for >100 traits from the NHGRI catalog [18]. We observed that traits such as Fasting Glucose (FGlu) (fold

enrichment = 7.05, P value = 3.30×10^{-4}), metabolic traits (fold enrichment = 6.44, P value = 2.09×10^{-4}) were the among the most highly enriched, highlighting the relevance and of these transcribed elements for islets (Fig. 5.4E). GWAS loci for T2D were also enriched in islet TCs (fold enrichment = 2.45, P value = 0.02). Since T2D is orchestrated through a complex interplay between islet beta cell dysfunction and insulin resistance in peripheral tissues, we reasoned that some underlying pathways in T2D might be more relevant to islets than others. To explore this rationale, we utilized results from a previous study that analyzed GWAS data for T2D along with 47 other diabetes related traits and identified clusters of related loci at the T2D GWAS signals [191]. Interestingly, we observe that signals in the islet beta cell and proinsulin cluster were highly enriched to overlap islet TCs (fold enrichment = 5.59, P value = 0.004), whereas signals in the insulin resistance cluster were depleted (fold enrichment = 0.91). These results suggest that islet TCs comprise active regulatory elements relevant for traits specifically related to islet function.

5.3.2 Integrating CAGE TCs with epigenomic information

We further explored CAGE profiles relative to the underlying chromatin landscape to identify characteristics of transcription initiation. We first overlayed CAGE profiles over accessible chromatin (ATAC-seq) profiles. Aggregated CAGE signal over ATAC-seq narrow peak summits highlighted the characteristic divergent pattern of transcription (Fig. 5.5A). Conversely, we anchored ATAC-seq signal over islet TC centers and observed that the summit of the ATAC-seq signal lies upstream of the TC center (Fig. 5.5B). We next asked if TF binding sites were more enriched to occur upstream of TCs vs downstream. We utilized TF footprint motifs previously identified using islet ATAC-seq data and TF DNA binding position weight matrices (PWMs) [196]. These footprint motifs represent putative TF binding sites that are also supported by accessible chromatin profiles, as opposed to TF motif matches

that are only informed by DNA sequence. We observe that most TFs were more enriched to occur in TCs and the 500bp upstream region as compared to TCs and 500 bp downstream region (Fig. 5.5C). These observations highlight that the region upstream of the TC is most accessible and more TF binding events occur.

We next explored the characteristics of TCs that occurred in the two main regulatory classes - promoter and enhancers, relative to each other. We focussed on transcribed, accessible regions in promoter and in enhancer states (TCs overlapping ATAC-seq peaks within promoter or enhancer segments). We considered the proximity of these elements to know gene TSSs and further classified the segments as TSS proximal or distal using a 5kb distance threshold from the nearest protein coding genes (Gencode V19). We then interrogated the chromatin landscape at these regions across 98 Roadmap Epigenomics cell types for which chromatin state annotations are publicly available (18 state extended model, see methods) [186]. We observed that TSS proximal islet TCs in accessible islet TSS chromatin states were nearly ubiquitously identified as TSS chromatin states across roadmap cell types (Fig. 5.5D, left). A fraction of TSS distal islet TCs in accessible islet TSS chromatin states however were more specific for pancreatic islets (Fig. 5.5D, right). In contrast, we observed that islet TCs in accessible islet enhancer chromatin states, both proximal and distal to known gene TSSs were more specifically identified as enhancer states in pancreatic islets (Fig. 5.5E). This pattern was more clear for pancreatic islets than whole pancreas (Fig. 5.5D and E, labelled) which further emphasizes the differences between epigenomic profiles for islets vs pancreas tissue.

Having observed differences in cell-type specificities in islet TCs in TSS vs enhancer states, we next asked if transcription factors displayed preferences to bind in these regions. We observed that footprint motifs for regulatory factor X (RFX) TF family were highly enriched (>3 fold, P value = 0.0001) in TCs in accessible enhancers. On the other hand, TCs in accessible TSS regions were highly enriched

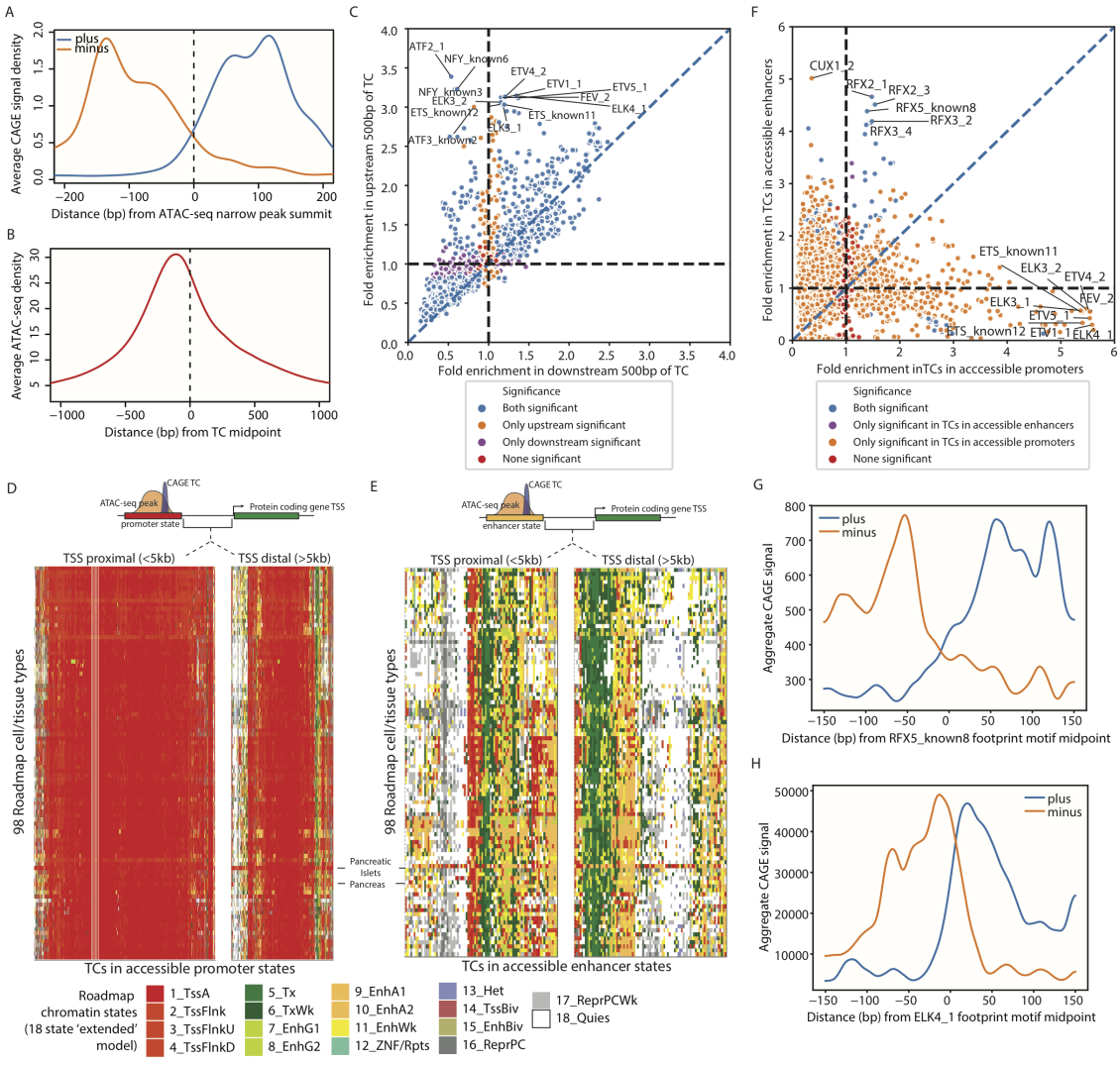


Figure 5.5: Integrating Islet CAGE TCs with other epigenomic information reveals characteristics of transcription initiation. A: Aggregate CAGE profiles over ATAC-seq peak summits. B: Aggregate ATAC-seq profile over TC midpoints. C: Enrichment of TF footprint motifs to overlap TC and 500 bp upstream region (y axis) vs TC and 500 bp downstream region (x axis). D: Chromatin state annotations across 98 Roadmap Epigenomics cell types (using the 18 state extended model, [186]) for TC segments that occur in islet active TSS chromatin state and overlap ATAC-seq peaks. These segments were segregated into those occurring 5kb proximal (left) and distal (right) to known protein coding gene TSS (Gencode V19). E: Chromatin state annotations across 98 Roadmap Epigenomics cell types for TC segments that occur in islet active enhancer chromatin state and overlap ATAC-seq peaks, segregated into those occurring 5kb proximal (left) and distal (right) to known gene TSS. F: Aggregate CAGE profiles centered and oriented relative to RFX5_known8 footprint motifs occurring in 5kb TSS distal regions. G: Aggregate CAGE profiles centered and oriented relative to ELK4_1 footprint motifs.

to overlap footprint motifs of the E26 transformation-specific (ETS) TF family.

We observe divergent aggregate CAGE profiles over TF footprint motifs enriched in enhancers for example RFX5_known8 footprint motifs in 5kb TSS distal regions and ELK4_1 motif (Fig. 5.5G and H).

5.3.3 Experimental validation of transcribed regions

We next sought to experimentally validate the regulatory activity of islet TCs. We utilized the massively parallel reporter assay platform wherein thousands of elements can be simultaneously tested by including unique barcode sequences for each element and determining the transcriptional regulatory activity using sequencing based barcode quantification. This approach is also known as the self-transcribing active regulatory region sequencing (STARR-seq) assay. We generated libraries of TC sequences (198 bp elements) and cloned these along with unique 16 bp barcode sequences into the STARR-seq vector, downstream of the GFP gene which was in control with the SCP1 promoter. We transfected the STARR-seq libraries into rat beta cell insulinoma (INS1 832/13) cell line, extracted RNA and sequenced the barcodes. We added 8 bp unique molecular identifier (UMI) sequences before the PCR amplification of the RNA libraries to enable accounting for and removing PCR duplicates while quantifying true biological RNA copies. We then modeled the RNA and DNA barcode counts in generalized linear models (GLMs) to model RNA and DNA counts for each barcode using MPRAnalyze [8] to quantify transcriptional activity of the TC element inserts. Our STARR-seq library included 6,798 insert elements (198 bp each) that overlapped 5,898 TCs. We selected barcodes that each had at least 10 DNA counts and non zero RNA counts in at least one technical replicate, and selected insert elements that had at least two of such qualifying barcodes. We had 3,240 such insert elements which we then tested for significant activity in the STARR-seq assay. We observed that 68.1% (N = 2,260) of these elements showed significant regulatory activity (5% FDR) (Fig.

5.6A (top)). On classifying insert elements based on active TSS, enhancer or other chromatin state overlap in islets, we observe that a larger fraction of TC elements overlapping the active TSS state had significant transcriptional activity compared to elements overlapping enhancer states, which was in turn higher than TCs in other chromatin states (Fig. 5.6A (bottom)). We also observed that the STARR-seq activity Z scores for TC elements in active TSS states were higher than those in enhancer states (Wilcoxon rank sum test $P = 2.99\times10^{-9}$) (Fig. 5.6B). Z scores for TCs that overlapped ATAC-seq peaks were also significantly higher than those that did not occur in peaks (Wilcoxon rank sum test $P = 4.01\times10^{-15}$) (Fig. 5.6C). Also, Z scores for TCs proximal to gene TSSs were higher than TCs that were distal to gene TSS locations (Wilcoxon rank sum test $P = 4.21\times10^{-9}$) (Fig. 5.6D). In Fig. 5.6E, we highlight an islet TC for which we tested three insert elements (Fig. 5.6E, STARR-seq elements track), which occurred in active TSS and enhancer states and overlapped ATAC-seq peak. All three of the elements showed significant transcriptional activity in our assay (Z score > 2.94, Z score P values < 0.001). Interestingly, while there are no known gene TSS annotations in this region, clear islet polyA+ mRNA-seq profiles that overlap the CAGE signal can be observed here. Another example TC locus that occurred in the intronic region of the *ABCC8* gene marked a region of islet-specific enhancer chromatin state and overlapped an ATAC-seq peak (Fig. 5.6F). The regulatory activity of this region was previously validated in the pancreatic bud in mouse embryos from a LacZ assay [142]. We included 39 insert elements that tiled this region (50 bp offset) (Fig. 5.6F, STARR-seq element track, which is a dense depiction of these tiles) in the STARR-seq assay and observed significant activity in multiple tiles within and neighboring the TC and the ATAC-seq peak (Fig. 5.6F, STARR-seq Z scores track). Through these analyses we could experimentally validate a considerable proportion of TCs for transcriptional regulatory activity in a STARR-seq assay in a rodent beta cell model system.

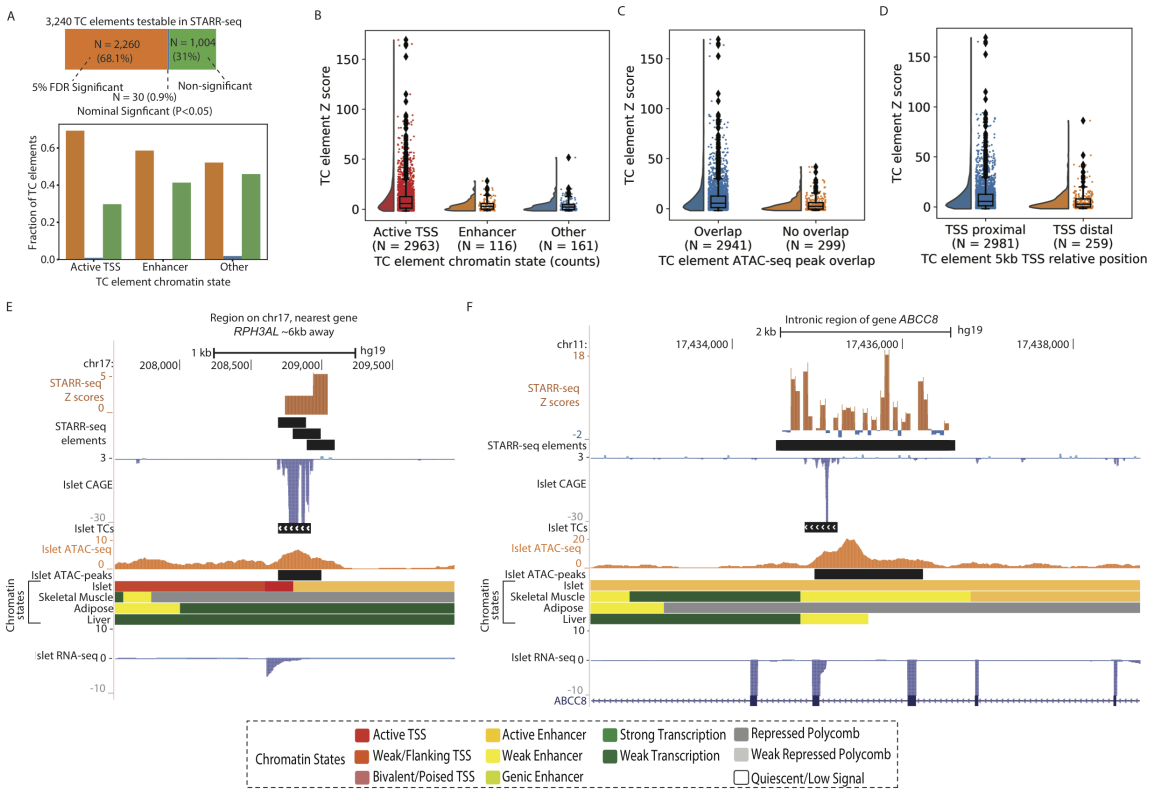


Figure 5.6: Experimental validation of TCs using STARR-seq MPRA: A: (Top) Number and fraction of TCs identified as significantly active (5% FDR), nominally active ($P < 0.05$) or non-significant in STARR-seq assay performed in rat beta cell insulinoma (INS1 832/13) cell line model. (Bottom) panel indicates the proportion of TCs that overlapped active TSS, enhancer or other chromatin states that were identified as active in STARR-seq assay. B. STARR-seq activity Z scores for TCs occurring in active TSS, enhancer or other chromatins states. C: STARR-seq activity Z scores for TCs that overlap ATAC-seq peak vs those that do not overlap peaks. D: STARR-seq activity Z scores for TCs based on relative position (5kb TSS proximal or distal) to known protein coding gene TSSs (Gencode V19). E: Example locus were TC elements that occur in active TSS and enhancer chromatin state and overlap ATAC-seq peak that were tested in the STARR-seq assay. The CAGE profile coincides with islet mRNA profile that is detected despite no known gene annotation in the region and the nearest protein coding gene is 6kb away. F: The intronic locus of the *ABCC8* gene, where an islet TC overlaps an ATAC-seq peak and active enhancer chromatin states. 198 bp tiles spanning the region shown in the STARR-seq elements track were included in the assay.

5.3.4 CAGE profiles augment functional genomic annotations to better understand GWAS and eQTL associations

Observing characteristics of TCs in islets in different epigenomic contexts and validating the activities of these elements, we next asked if islet TCs taken as an additional layer of functional genomic information could supplement fine mapping efforts in understanding GWAS or eQTL associations. We classified genomic annotations utilizing different layers of epigenomic data such as histone modification based chromatin states, accessible regions within these states and transcribed accessible regions within these states. We then computed enrichment for T2D GWAS loci to overlap these annotations using full GWAS summary statistics [110] using a Bayesian hierarchical model implemented in the fGWAS tool [146]. This method allows using not only the genome wide significant loci, instead, leveraging genome wide association statistics such that marginal associations can also be accounted for. The prior probabilities of a region of the genome containing an association and a SNP being causal are then allowed to vary based on overlap with annotations. We observed that TCs in accessible enhancer regions were the most highly enriched for T2D GWAS loci (Fig. 5.7A, left). We performed similar analysis using islet eQTL summary data [196], where we observed that TCs in accessible regions in both enhancers and promoters were most highly enriched (Fig. 5.7A, right). These data suggest that including TC information with other functional genomics data help delineate more relevant regions for gene expression and trait association signals.

We also asked if TCs or ATAC-seq data can be more informative in pinpointing active elements within enhancers or promoters. We performed GWAS and islet eQTL enrichment analyses for TCs and ATAC-seq peaks while conditioning on active TSS or active enhancer chromatin states. We observed that TCs had a higher conditional enrichment over enhancer states for T2D (5.7B) and ATAC-seq peaks. TCs also had a higher conditional enrichment over enhancer and promoter states for islet eQTL loci

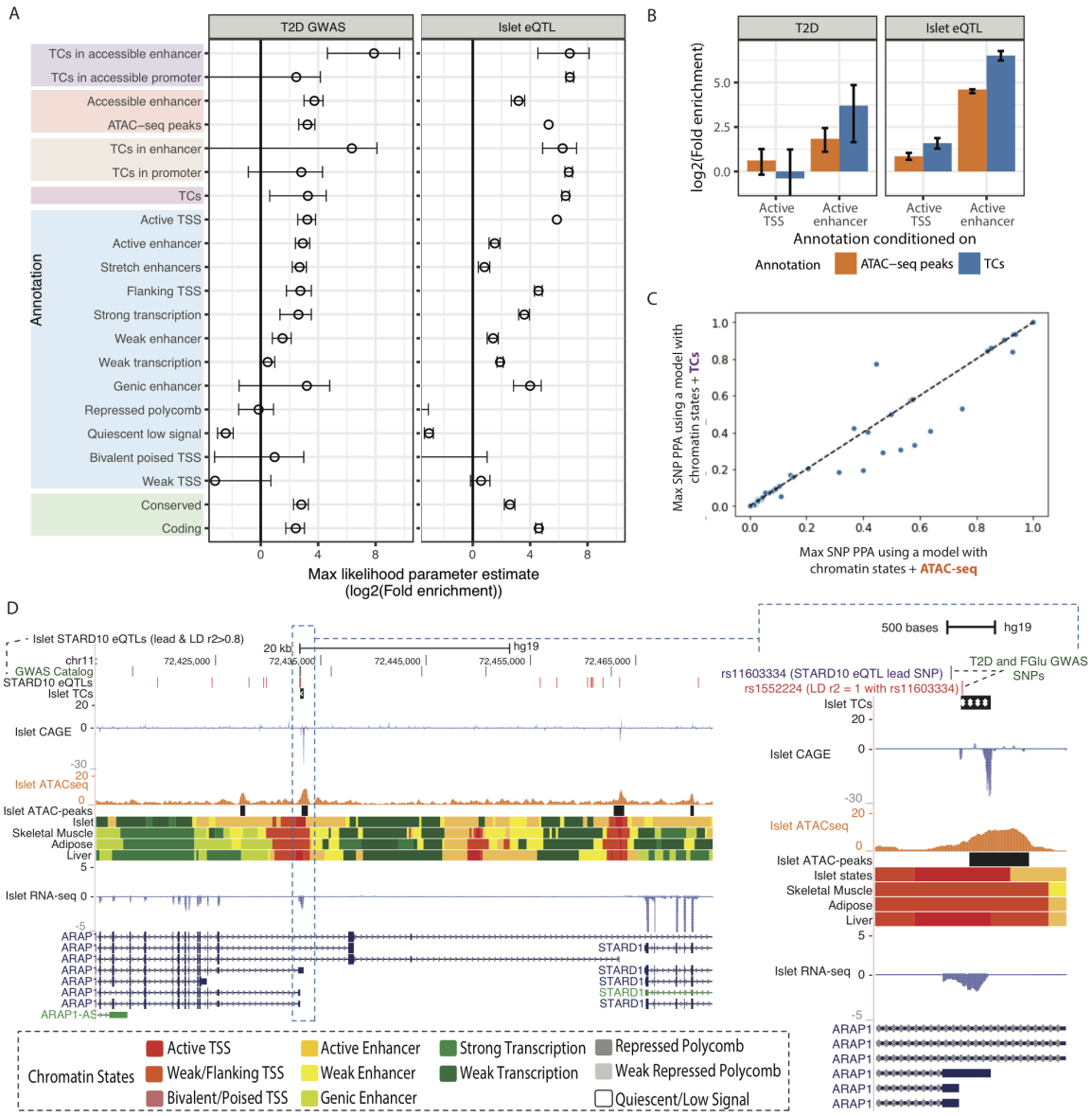


Figure 5.7: Islet TCs along with ATAC-seq and chromatin state information supplement GWAS finemapping efforts: Enrichment of T2D GWAS (A) or islet eQTL (B) loci in annotations that comprise different levels of epigenomic information, including including chromatin state, ATAC-seq and TCs. Annotations were defined using combinations of these datasets such as accessible enhancers (ATAC-seq peaks in enhancer states) transcribed accessible enhancers (TCs that overlap ATAC-seq peaks in enhancer states) etc. Enrichment was calculated using fGWAS [146] using summary statistics from GWAS (in A) [110] or eQTL (in B) [196]. C: fGWAS conditional enrichment analysis testing the contribution of annotations such as islet TCs or ATAC-seq peaks after conditioning on histone-only based annotations such as promoter and enhancer chromatin states in islets. D: Maximum SNP PPA per FGlu GWAS locus when using a model including ATAC-seq (x axis) or TCs (y axis) another annotation such a STARR-seq activity Z scores for TCs occurring in active TSS, enhancer or other chromatins states. Continued on the next page.

Figure 5.7: Continued - E: Genome browser view of the *STARD10* gene locus where T2D and Fasting Glucose GWAS SNPs and eQTL SNPs for the *STARD10* gene occur (left). *STARD10* eQTL Lead and LD r^2>0.8 proxy SNPs are shown in the eQTL SNP track. Genome browser view on the right shows the region zooming in on the lead eQTL SNP rs11603334 and another SNP rs1552225 (LD r^2=1 with the lead SNP) which overlaps an islet TC. Functional reweighting of Fasting Glucose GWAS data using chromatin state, ATAC-seq and TC data resulted in the PPA of the SNP rs1552225 = 0.772.

as compared to ATAC-seq peaks (5.7B). We then sought to reweight the GWAS posterior probabilities of association (PPAs) by including these functional annotations in order to fine-map Fasting glucose GWAS loci and compared the maximal SNP PPA at each locus (Fig. 5.7C). We highlight one such region within the *ARAP1* gene that includes many variants in high LD. Variants at this T2D and FGlu GWAS locus are identified as eQTL for the *STARD10* gene [199] but not for ARAP1. The GWAS and eQTL index SNP rs11603334 is in high LD (r^2=1) with rs1552224. Including TC information results in increased PPA of rs1552224 to 0.772. Without TC data the PPA for both rs11603334 and rs1552224 and were 0.446. We observed significant activity of the TC element that overlaps rs1552224 in our STARR-seq assay (Z score = 4.90, Z score P value = 4.78×10^{-7}). A previous study showed stronger evidence for rs11603334 to be the causal variant [89], whereas another study pointed towards another variant (rs140130268) as more likely causal [20] which highlights the complexity at this locus. These analyses demonstrate the utility of transcription initiation information to demarcate active regulatory elements at higher genomic resolution.

5.4 Discussion

We profiled transcription start sites in human pancreatic islets using CAGE. We observe high enrichment of CAGE TCs in TSS chromatin states and ATAC-seq peaks in islets, which expectedly reflects the chromatin landscape at regions where transcription initiation occurs. Comparison of islet CAGE TCs with those identified across

multiple tissues revealed that 20% of islet TCs were islet specific. Furthermore, comparing the chromatin states underlying these TCs across multiple cell types and tissues indicated that TCs that occur distal to known TSSs of protein coding genes comprised of more islet specific promoter and enhancer chromatin states. These analyses also highlighted the differences in TCs and their underlying chromatin contexts between islets and pancreas tissues, which further demonstrate the need for molecular profiling in the islet tissue to better understand islet mechanisms. Islet TCs were also enriched to overlap GWAS loci of fasting glucose and specifically the islet beta cell related components of T2D loci, while being depleted for the insulin resistance related components of T2D GWAS loci. These analyses demonstrate that islet TCs mark active, specific and relevant Islet regulatory elements.

Our work revealed that transcribed and accessible enhancer regions were most enriched to overlap TF footprint motifs for the RFX family of TFs. We previously showed that RFX footprint motifs are confluently disrupted by T2D GWAS risk alleles [196], which are enriched to occur in islet specific enhancer regions. These observations together highlight the role of islet specific enhancer regions, and the potential of ATAC-seq and CAGE profiling to identify the active regulatory elements within these large enhancer elements at high genomic resolution.

Utilizing the STARR-seq enhancer MPRA approach, we observed that 68% of Islet TCs induce significant transcriptional activity which highlights how CAGE identifies active regulatory elements. A larger fraction of TCs that occurred in active TSS chromatin states were significantly active than those occurring in active enhancer or other chromatin states. The transcriptional activities of these active TSS state overlapping TCs were also higher than the latter. We note here that only a small fraction of TCs (0.4%) were identified to overlap the active enhancer state. Studies have shown that gene distal transcripts are more unstable, which would therefore be difficult to profile from a total RNA sample. Of course, given the relative instability of

enhancer RNAs some chromatin-defined sites may be active but fall below the limits of detection of CAGE. Therefore, it is understandable that islet CAGE profiling from total RNA samples would comprise more stable promoter-associated RNA transcripts and have lesser representation of weaker transcripts originating from enhancer regions. In our previous work [195], we showed that genetic variants in more cell type-specific enhancer regions have lower effects on gene expression (measured as eQTL effect sizes) than the variants occurring in more ubiquitous promoter regions, in the un-stimulated or basal cell state. This is consistent with our observation of lower transcriptional activities and even low representation of transcription initiation identified in enhancer state regions.

To better understand the mechanisms underlying GWAS loci, we interrogated the potential of TC information in identifying the causal SNP(s) using functional fine mapping approach (fGWAS). While most islet TCs overlapped islet ATAC-seq peaks (>70%), we observed that regions supported by TCs, ATAC-seq peaks and enhancer chromatin states (transcribed, accessible enhancer regions) were most enriched to overlap T2D GWAS loci. This enrichment was higher than in regions only informed by ATAC-seq peaks and enhancer chromatin states, indicating that the small set of TCs in enhancer regions actually delineate highly relevant elements. Our work demonstrates that transcription start site information profiled using CAGE in islets can be used in addition to other relevant epigenomic information such as histone mark informed chromatin states and chromatin accessibility in nominating relevant variants and biological mechanisms.

5.5 Materials and Methods

5.5.1 Islet Procurement and Processing

Islet samples from organ donors were received from the Integrated Islet Distribution Program, the National Disease Research Interchange (NDRI), and Prodo- Labs. Islets were shipped overnight from the distribution centers. On receipt, we prewarmed islets to 37 C in shipping media for 12 h before harvest; 2,5005,000 islet equivalents (IEQs) from each organ donor were harvested for RNA isolation. We transferred 5001,000 IEQs to tissue culture-treated flasks and cultured them as in the work in [52].

5.5.2 RNA isolation, CAGE-seq library preparation and sequencing

Total RNA from 2000-3000 islet equivalents (IEQ) was extracted and purified using Trizol (Life Technologies). RNA quality was confirmed with Bioanalyzer 2100 (Agilent); samples with RNA integrity number (RIN) > 6.5 were prepared for CAGE sequencing. 1ug Total RNA samples were sent to DNAFORM, Japan, where polyA negative selection and size selection (<1000bp) was performed. Stranded CAGE-sequencing libraries were generated for each islet sample using the () kit according to manufacturers protocol (Illumina). Each islet CAGE-seq library was barcoded, pooled into 24-sample batches, and sequenced over multiple lanes of HiSeq 2000 to obtain an average depth of 100 million 2 x 101 bp sequences. All procedures followed ethical guidelines at the National Institutes of Health (NIH).

5.5.3 CAGE data mapping and processing

Because read lengths differed across libraries, we trimmed all reads to 51 bp using fastx_trimmer (FASTX Toolkit v. 0.0.14). Adapters and technical sequences were trimmed using trimmomatic (v. 0.38; paired-end mode, with options ILLUMI-

NACLIP:adapters.fa:1:30:7:1:true). To remove potential E. coli contamination, we mapped to the E. coli chromosome (genome assembly GCA_000005845.2) with bwa mem (v. 0.7.15; options: -M). We then removed read pairs that mapped in a proper pair (with mapq $\geq$ 10) to E. coli. We mapped the remaining reads to hg19 using STAR (v. 2.5.4b; default parameters). We pruned the mapped reads to high quality autosomal read pairs (using samtools view v. 1.3.1; options -f 3 -F 4 -F 8 -F 256 -F 2048 -q 255). We then performed UMI-based deduplication using umitools dedup (v. 0.5.5; –method directional).

We selected Islet samples with strandedness measures >0.85 calculated from QoRTS [62] for all downstream analyses. 50 Islet samples passed this threshold.

5.5.4 Tag cluster calling

We used the paralu method to identify clusters of CAGE start sites (CAGE tag clusters) [46]. The algorithm uses a density parameter d and identifies segments that maximize the value of $Number\ of\ events - d * size\ of\ the\ segment$ (bp). Here, large values of d would favor small, dense clusters and small values of d would favor larger more rarefied clusters. The method identifies segments over all values of d beginning at the largest scale, where d = 0, where all of the events are merged into one big cluster. It then calculates the density (events per nucleotide) of every prefix and suffix of the big cluster. The lowest value among all of these densities is the maximum value of d for the big cluster because at higher values of d the big cluster will no longer be a maximal-scoring segment (because zero-scoring prefixes or suffixes are not allowed).

We called TCs in each individual sample using raw tag counts, requiring at least 2 tags at each included start site and allowing single base-pair tag clusters (singletons) if supported by >2 tags. We then merged the tag clusters on each strand across samples. For each resulting segment, we calculated the number of islet samples in

which TCs overlapped the segment. We included the segment in the consensus TCs set if it was supported by independent TCs in at least 10 individual islet samples. This threshold was selected based on comparing the number of tag clusters with the number of samples across which support was required to consider the segment (Fig. 5.1).

5.5.5 Chromatin state analysis

We collected publicly available cell/tissue ChIP-seq data for H3K27ac, H3K27me3, H3K36me3, H3K4me1, and H3K4me3 and input for Islets, Adipose and Skeletal Muscle and Liver. Data for Adipose, Skeletal Muscle and Liver tissues were included in the joint model for other ongoing projects. We performed read mapping and integrative chromatin-state analyses in a manner similar to that of our previous reports and followed quality control procedures reported by the Roadmap Epigenomics Study [186]. Briefly, we trimmed reads across datasets to 36bp and overrepresented adapter sequences as shown by FASTQC (version v0.11.5, `https://www.bioinformatics.babraham.ac.uk/projects/fastqc/`) using cutadapt (version 1.12) [117]. We then mapped reads using BWA (version 0.5.8c), removed duplicates using samtools [98], and filtered for mapping quality score of at least 30. To assess the quality of each dataset, we performed strand cross-correlation analysis using phantompeakqualtools (v2.0; `code.google.com/p/phantompeakqualtools`) [93]. We converted bam files for each dataset to bed using the bamToBed tool. To more uniformly represent datasets with different sequencing depths across histone marks and tissues, we randomly subsampled each dataset bed file to the mean depth for that mark across the four included tissues. This allowed comparable chromatin state territories across tissues and ensured that chromatin state territories were not heavily driven by high sequencing depth. Chromatin states were learned jointly for the three cell types using the ChromHMM (version 1.10) hidden Markov model algo-

rithm at 200-bp resolution to five chromatin marks and input [40, 43, 41]. We ran ChromHMM with a range of possible states and selected a 11-state model, because it most accurately captured information from higher-state models and provided sufficient resolution to identify biologically meaningful patterns in a reproducible way. We have used this state selection procedure in previous analyses [162, 196]. To assign biological function names to our states that are consistent with previously published states, we performed enrichment analyses in ChromHMM comparing our states with the states reported previously [196] for the four matched tissues. We assigned the name of the state that was most strongly enriched in each of our states (Fig. 5.3).

5.5.6 ATAC-seq data analysis

We used previously published data for chromatin accessibility profiled using ATAC-seq in islets from two human organ donor samples [196]. For each sample, we trimmed reads to 36 bp (to uniformly process ATAC-seq from other tissues for ongoing projects) and removed adapter sequences using Cutadapt (version 1.12) [117], mapped to hg19 used bwa-mem (version 0.7.15-r1140) [96], removed duplicates using Picard (`http://broadinstitute.github.io/picard`) and filtered out regions blacklisted by the ENCODE consortium due to poor mappability (wgEncodeDacMapabilityConsensusExcludable.bed and wgEncodeDukeMapabilityRegionsExcludable.bed). For each tissue we subsampled both samples to the same depth so that each tissue had overall similar genomic region called as peaks. We used MACS2 (`https://github.com/taoliu/MACS`), version 2.1.0, with flags -g hsnomodelshift -100extsize 200 -Bbroadkeep-dup all, to call peaks and retained all broad-peaks that satisfied a 1% FDR.

5.5.7 Overlap enrichment between TCs and annotations

Enrichment for overlap between each pair of regulatory annotations in Figure S1 was calculated using the Genomic Association Tester (GAT) tool [65]. To ask if two sets of regulatory annotations overlap more than that expected by chance, GAT randomly samples segments of one regulatory annotation set from the genomic workspace (hg19 chromosomes) and computes the expected overlaps with the second regulatory annotation set. We used 10,000 GAT samplings for each enrichment run. GAT outputs the observed overlap between segments and annotation along with the expected overlap and an empirical p-value.

5.5.8 GWAS data collection and LD pruning

We downloaded the GWAS data for various traits from the NHGRI website on June 12, 2018 (file gwas_catalog_v1.0.2-associations_e92_r2018-05-29.tsv from `https://www.ebi.ac.uk/gwas/docs/file-downloads`). We selected genome-wide significant GWAS SNPs (P $<5\times10^{-8}$) for traits for which the study included European samples. To retain independent signals, we linkage diequilibrium (LD) pruned the list of SNPs to retain SNPs with the most significant P values that had LD r^2 <0.2 between each pair. This procedure was performed using the PLINK (v1.9) tool [152, 22] clump option and 1000 genomes phase 3 vcf files (downloaded from ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/release/20130502), subsetted to the European samples as reference. We selected traits that had >30 independent signals for following analyses.

5.5.9 Enrichment of genetic variants in genomic features

Enrichment for genome wide association study (GWAS) variants for different traits in Islet TCs was calculated using the Genomic Regulatory Elements and Gwas Overlap algoRithm (GREGOR) tool (version 1.2.1) [160]. Since the causal SNP(s) for the

traits are not known, GREGOR allows considering the input lead SNP along with SNPs in high linkage disequilibrium (LD) (based on the provided R2THRESHOLD parameter) while computing overlaps with genomic features (such as islet TCs). Therefore, as input to GREGOR, we supplied SNPs that were not in high linkage disequilibrium with each other. For each input SNP, GREGOR selects 500 control SNPs that match the input SNP for minor allele frequency (MAF), distance to the nearest gene, and number of SNPs in LD. Fold enrichment is calculated as the number of loci at which an input SNP (either lead SNP or SNP in high LD) overlaps the feature over the mean number of loci at which the matched control SNPs (or SNPs in high LD) overlap the same features. This process accounts for the length of the features, as longer features will have more overlap by chance with control SNP sets.

Specific parameters for the GWAS enrichment were: GREGOR: r2 threshold = 0.8. LD window size = 1Mb; minimum neighbor number = 500, population = European.

5.5.10 Comparison of features with Roadmap chromatin states

We downloaded the chromatin state annotations identified in 127 human cell types and tissues by the Roadmap epigenomics project [186] after integrating ChIP-seq data for five histone 3 lysine modifications (H3K4me3, H3K4me1, H3K36me3, H3K9me3 and H3K27me3) that are associated with promoter, enhancer, transcribed and repressed activities, across each cell type. For each TC feature, for example, TCs in ATAC-seq peaks within islet enhancer chromatin states, we identified segments occurring proximal to (within 5kb) and distal from (further than 5kb) known protein coding gene TSS (gencode V19 [61]). For each such segment, we identified the maximally overlapping chromatin state across 98 cell types publicly available from the Roadmap Epigenomics project in their 18 state extended model using BEDtools intersect. We then ordered the segments using clustering (hclust function in R) based

on the gower distance metric (daisy function in R) for the roadmap state assignments across 127 cell types.

5.5.11 Aggregate signal

We generated the ATAC-seq density plot over islet TC midpoints using the Agplus tool (version 1.0) [109]. We used the ATAC-seq signal track for reads per 10 Million to aggregate over stranded TCs.

To obtain CAGE tracks, we merged CAGE bam files for islet samples that passed QC (see CAGE data processing section) and obtained the read 1 start sites or TSSs. To better visualise the CAGE signal, we then flanked each TSS 10bp upstream and downstream and normalized the TSS counts to 10M mapped reads. We generated CAGE density plots over ATAC-seq narrow peak summits by using the agplus tool.

To obtain aggregate CAGE signal over TF footprint motifs, we oriented the CAGE signal with respect to the footprint taken on the plus strand. We used HTSeq GenomicPosition method [6] to obtain the sum of CAGE signal at each base pair relative to the footprint motif mid point.

5.5.12 fGWAS analyses and finemapping

fGWAS (version 0.3.6) [146] employs a Bayesian hierarchical model to determine shared properties of loci affecting a trait. The model uses association summary level data, divides the genome into windows generally larger than the expected LD patterns in the population. The model estimates the probabilities that an association lies in a window and that a SNP is causal. These probabilities are then allowed to depend on genomic annotations, and are estimated based on enrichment patterns of annotations across the genome using a Bayes approach. We used fGWAS with default parameters for enrichment analyses for individual annotations in Fig. 5.7 A and B. For each individual annotation, the model provided maximum likelihood enrichment

parameters and annotations were considered as significantly enriched if the parameter estimate and 95% CI was above zero.

We performed conditional analyses using the -cond option.

To reweight GWAS summary data based on functional annotation overlap, we used the -print option in an fGWAS model run after including multiple annotations that were individually significantly enriched. We included Active TSS, active enhancer, stretch enhancer, quiescent and polycomb repressed annotations along with ATAC-seq or TCs in a model to derive enrichment priors which can then be used to evaluate both the significance and functional impact of associated variants in GWAS regions; such that variants overlapping more enriched annotations carry extra weight.

5.6 Acknowledgements

I'd like to thank several people who contributed to this project. I thank Dr. Mike Erdos, Narisu Narisu Dr. Nandini Manickam for preparing the RNA samples, coordinating shipping to DNAFORM, Japan for the CAGE library preparation and subsequent sequencing performed at the National Institutes of Health (NIH), USA. I especially thank Dr. Yasuhiro Kyono, Dr. Venkateswaran Ramamoorthy Elangovan and Prof. Jacob Kitzman for all their help in designing, performing and the subsequent analysis of the STARR-seq assay. I thank Prof. Steve Parker conceptualizing and organizing this project and for his insight and support all throughout.

CHAPTER VI

Implications and Future Work

Through my dissertation work, I have analyzed large omics datasets to supplement our understanding of gene regulatory mechanisms that underlie the associations between genetic variation and disease traits such as T2D. Several important themes have emerged from my work which are exciting avenues to further augment our understanding of how predisposition to complex disease is encoded in our non-coding genome.

6.0.1 Regulatory buffering and the need for molecular context specific studies

The regulatory nature of a majority of common disease associated variants motivated our investigation of regulatory regions across the genome. We asked asked how the genetics of gene regulation differed across regulatory annotations that had been defined using different sets of epigenomic marks - all of which were shown to marked active regulatory regions. We observed that eQTL occurring in HOT regions that represent mostly promoter-like regions had significantly higher effect sizes than those occurring in more cell-specific stretch enhancers. This effect remained after controlling for distance to the target gene. However, chromatin accessibility QTL in stretch enhancers have significantly larger effect sizes compared to those in HOT

regions. These observations were quite robust in that we observed these trends in multiple cell/tissue types such as blood/K652 cell lines, LCL/GM12878 cell lines and islet tissue.

These seemingly conflicting results indicated that the chromatin in the cell-specific stretch enhancers was genetically primed for larger effects on accessibility, however, genetic effects on modulation of gene expression were lower. We noted here that both chromatin and expression QTLs analyzed were identified in the basal state for the cells/tissues. A recent study showed that 60% of eQTL identified in stimulated condition in macrophages were identified as chromatin QTL in the basal state [3]. Our observations and other supportive evidence suggest that lower effects on gene expression in the basal state despite higher propensity for chromatin effects could be a mechanism to ensure stable expression of critical genes in the basal state while priming these for quick response to patho-physiologic stimuli. Similar inferences about robust gene expression and enhancer redundancy have been reported recently. One such study showed that dosage-sensitive genes have evolved robustness to the disruptive effects of genetic variation by expanding their regulatory domains [201]. Others have questioned if stretch/super enhancers are really different from other enhancers that just happen to occur close-by [148]. Numerous studies perturbed elements within these enhancers but showed conflicting results - some, where the perturbation affected gene expression and others where there was no observable effect on gene expression, again implying redundancy in gene regulation by individual components of super enhancers [63, 166, 133, 207].

Considering regulatory annotations from a disease genetics perspective, I and others have observed high enrichment of GWAS loci in trait-relevant stretch enhancers (eg. T2D GWAS loci enriched in islet stretch enhancers, Rheumatoid Arthritis GWAS loci enriched in lymphoblastoid cell line GM12878 stretch enhancers, Fasting Insulin GWAS loci enriched in Adipose and Skeletal Muscle stretch enhancers among sev-

eral such observations). However, eQTL loci, which have long been touted to help identify the target genes of the GWAS signals through co-localization methods, are highly enriched in promoter regions. The differences in the genetics of gene regulation between these annotations that I have demonstrated could help reconcile why many cis-eQTLs are shared across cell types and infrequently co-localize with GWAS signals [103, 74, 58].

We hypothesize that cell-specific stretch enhancers drive critical responses to external stimuli and therefore some genetic associations with gene expression might be evident only under relevant conditions. Stimuli such as nutrient conditions, stress or hormone signaling could modulate TF abundance and localization which could drive context-specific mechanisms. Therefore, one critical and exciting direction is to understand the genetics of gene regulation under carefully selected stimulatory conditions in relevant cell types. Studying the impact of genetic variants under such contexts may be crucial for revealing functional convergence of disease-associated variants. More specifically, we advocate conducting molecular QTL screens under stimulus/treatment induced conditions with carefully considered sample sizes and time points (Fig. 6.1 A and B).

Response studies have been limited in the T2D genomics literature. The potential of such studies however is demonstrated by other work that highlights context specific effects. One such study explored a T2D GWAS variant (rs508419) that lies in a skeletal muscle-specific active promoter region at the ANK1 locus [162, 208]. Human skeletal muscle eQTL data indicated that the T2D risk allele was associated with higher ANK1 expression [162]. Interestingly, however, increased ANK1 protein only affected glucose uptake when treated with insulin; there was no detectable effect of increased ANK1 protein under basal conditions. The same variant, however, is associated with reduced expression of the transcription factor NKX6-3 in the islet tissue [162, 196], representing a tissue-dependent effect of regulatory variants, and po-

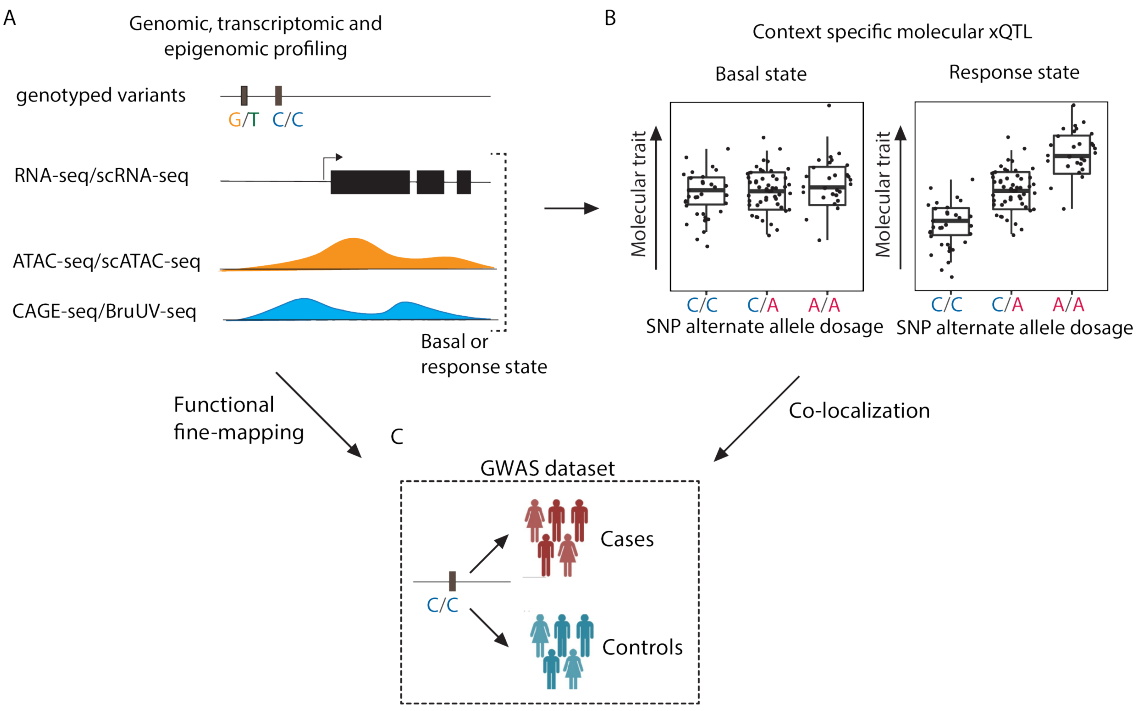


Figure 6.1: Context-specific xQTL mapping to better understand GWAS

tentially more complicated genetic architecture. Therefore, modelling environmental stimuli in functional T2D genomic studies is both important and challenging.

The increased glucose levels in the body resulting from insulin resistance are critical environmental factors for the pancreatic islets. Therefore, glucose specific effects on islet function can be assessed by culturing islets in low vs high glucose conditions and profiling RNA (mRNA-seq) or open chromatin (ATAC-seq) followed by QTL analysis to identify context specific molecular QTL (xQTL) in islets. Patient-derived induced pluripotent stem cell (iPSC) lines differentiated towards the islet lineage could also be useful for this purpose and better suited to culture in the laboratory.

6.0.2 Single-cell molecular profiling approaches to dissect islet heterogeneity

It is important to note that islets are heterogeneously composed of multiple subtypes (including alpha, beta, gamma and delta cells) that have diverse functions. Recent developments in single cell assays followed by sequencing technologies, both for measuring RNA (scRNA-seq) and open chromatin (scATAC-seq) therefore present exciting avenues. For example, it is now possible to identify regulatory sites that are specific to, say insulin secreting beta cells. scRNA-seq or scATAC-seq can be used to define cell-type proportions, patterns of which when analyzed across individuals can would identify cell-type proportion QTLs, where genetic variation influences proportion of different subtypes. Performing these studies while considering basal or glucose response contexts could again be invaluable in identifying context specific genetic effects on islet cell biology. Such studies could lead to a better understanding of islet biology and potentially at T2D related trait GWAS loci.

6.0.3 Linking molecular profiling and xQTL information with GWAS and identifying causal relationships

Genome-wide molecular profiling data can be jointly analyzed with GWAS summary data to help fine-map causal variants(s). For example, a hierarchical modeling approach implemented in the fGWAS package [146] statistically models the prior probabilities that a genomic region contains an association with the trait and that a SNP in that region is causal, allowing these probabilities to vary based on the underlying functional molecular profiles such as open chromatin. Therefore, statistical integration of molecular profiling data, especially under specific contexts can further enable identification of causal variants (Fig. 6.1 C). Next, to establish mechanistic links between genetic effects on these molecular profiles and on the relevant trait such as T2D, it should be determined if the same genetic variant drives the xQTL as well

as the GWAS signal. Approaches that test for such co-localization between two signals have been established [53, 136, 70], (Fig. 6.1 C). Furthermore, potential causal relationships between the xQTL and GWAS signals can be assessed using Mendelian randomization [70]. For loci with 2 molecular profile QTLs, potential causal direction between each pair can be assessed using the Causal Inference Test [127] and MR-Steiger [67]; these tests provide complementary information. While molecular xQTL data elucidates the genetics of the epigenomic or chromatin landscape, recently developed bayesian strategies [90] can further help determine causal interactions and the relationships between multiple identified regulatory elements .

6.0.4 Functional follow up of prioritized variants

While mapping the epigenomic landscape and identifying genetic associations can inform candidate regulatory regions and potentially causal variants, complementary approaches are needed to functionally validate these effects. Massively parallel reporter assays (MPRAs) are one such tool that can be invaluable in functionally screening regulatory elements and identifying allelic effects. Since enhancers integrate and transduce environmental signals to execute gene expression programs, studying the impact of genetic variants under diverse conditions will be crucial for furthering our understanding of disease-associated variants. Efforts towards more robust, accurate and efficient MPRAs have been ongoing in the lab that enable correcting for PCR duplicates by introducing unique molecular identifier sequences (UMIs), consider effects of different promoter sequences in the assay and also their placement relative to the tested sequence among other advantages.

6.0.5 Concluding remarks

The results presented in this dissertation demonstrate that integrating information from the epigenome, transcriptome and other diverse molecular domains can

help understand how complex disease predisposition is encoded in the (mostly non-coding) part of our genome. Moving forward, obtaining molecular profiles in different environmental contexts and probing genetic associations, followed by computational integration with emerging large-scale GWAS data could help partition swathes of GWAS signals into coherent, tissue-specific subsets to shed light on underlying pathophysiologies. Technological advancements will propel the field forward, for example, further reduction in experimental and sequencing cost could enable increased sample sizes in study design; development of more efficient and robust single cell experimental and analysis tools could supplement biological understanding at higher resolution; more exhaustive trans-ethnic studies will enable higher power for signal discovery and delineate a more complete picture of the genetic underpinnings of disease traits. Collectively, such approaches could help reveal additional convergent functional contexts, which could eventually enable higher-resolution patient stratification and determination of individualised risk.

BIBLIOGRAPHY

BIBLIOGRAPHY

[1] Adli, M., J. Zhu, and B. E. Bernstein (2010), Genome-wide chromatin maps derived from limited numbers of hematopoietic progenitors, *Nature Methods*, *7*(8), 615–618, doi:10.1038/nmeth.1478.

[2] Aftab, S., L. Semenec, J. S.-C. Chu, and N. Chen (2008), Identification and characterization of novel human tissue-specific RFX transcription factors, *BMC Evolutionary Biology*, *8*(1), 226.

[3] Alasoo, K., J. Rodrigues, S. Mukhopadhyay, A. J. Knights, A. L. Mann, K. Kundu, C. Hale, G. Dougan, and D. J. Gaffney (2018), Shared genetic effects on chromatin and gene expression indicate a role for enhancer priming in immune response, *Nature Genetics*, p. 1, doi:10.1038/s41588-018-0046-7.

[4] Allum, F., et al. (2015), Characterization of functional methylomes by next-generation capture sequencing identifies novel disease-associated variants, *Nature Communications*, *6*, 7211.

[5] Almgren, P., M. Lehtovirta, B. Isomaa, L. Sarelin, M. R. Taskinen, V. Lyssenko, T. Tuomi, L. Groop, and for the Botnia Study Group (2011), Heritability and familiality of type 2 diabetes and related quantitative traits in the Botnia Study, *Diabetologia*, *54*(11), 2811, doi:10.1007/s00125-011-2267-5.

[6] Anders, S., P. T. Pyl, and W. Huber (2015), HTSeq–a Python framework to work with high-throughput sequencing data, *Bioinformatics*, *31*(2), 166–169, doi:10.1093/bioinformatics/btu638.

[7] Andersson, R., et al. (2014), An atlas of active enhancers across human cell types and tissues, *Nature*, *507*(7493), 455, doi:10.1038/nature12787.

[8] Ashuach, T., D. S. Fischer, A. Kreimer, N. Ahituv, F. J. Theis, and N. Yosef (2019), MPRAnalyze - A statistical framework for Massively Parallel Reporter Assays: Supplemental Methods - model, *bioRxiv*, doi:10.1101/527887.

[9] Battle, A., et al. (2013), Characterizing the genetic basis of transcriptome diversity through RNA-sequencing of 922 individuals, *Genome Res.*, p. gr.155192.113, doi:10.1101/gr.155192.113.

[10] Benayoun, B. A., et al. (2014), H3K4me3 Breadth Is Linked to Cell Identity and Transcriptional Consistency, *Cell*, *158*(3), 673–688, doi:10.1016/j.cell.2014.06.027.

[11] Bernstein, B. E., et al. (2006), A Bivalent Chromatin Structure Marks Key Developmental Genes in Embryonic Stem Cells, *Cell*, *125*(2), 315–326, doi:10.1016/j.cell.2006.02.041.

[12] Blinka, S., M. H. Reimer, K. Pulakanti, and S. Rao (2016), Super-Enhancers at the Nanog Locus Differentially Regulate Neighboring Pluripotency-Associated Genes, *Cell Reports*, *17*(1), 19–28, doi:10.1016/j.celrep.2016.09.002.

[13] Boyle, A. P., et al. (2014), Comparative analysis of regulatory information and circuits across distant species, *Nature*, *512*(7515), 453–456, doi:10.1038/nature13668.

[14] Brown, A. A. (2015), LargeQvalue: A Program for Calculating FDR Estimates with Large Datasets, *bioRxiv*, p. 010074, doi:10.1101/010074.

[15] Brown, A. A., A. Viñuela, O. Delaneau, T. D. Spector, K. S. Small, and E. T. Dermitzakis (2017), Predicting causal variants affecting expression by using whole-genome sequencing and RNA-seq from multiple human tissues, *Nature Genetics*, *49*(12), 1747–1751, doi:10.1038/ng.3979.

[16] Buenrostro, J., B. Wu, H. Chang, and W. Greenleaf (2015), ATAC-seq: A Method for Assaying Chromatin Accessibility Genome-Wide, *Curr Protoc Mol Biol*, *109*, 21.29.1–21.29.9, doi:10.1002/0471142727.mb2129s109.

[17] Buenrostro, J. D., P. G. Giresi, L. C. Zaba, H. Y. Chang, and W. J. Greenleaf (2013), Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position, *Nature Methods*, *10*(12), 1213–1218, doi:10.1038/nmeth.2688.

[18] Buniello, A., et al. (2019), The NHGRI-EBI GWAS Catalog of published genome-wide association studies, targeted arrays and summary statistics 2019, *Nucleic Acids Res*, *47*(D1), D1005–D1012, doi:10.1093/nar/gky1120.

[19] Butler, P. C., M. Elashoff, R. Elashoff, and E. A. M. Gale (2013), A critical analysis of the clinical use of incretin-based therapies: Are the GLP-1 therapies safe?, *Diabetes Care*, *36*(7), 2118–2125, doi:10.2337/dc12-2713.

[20] Carrat, G. R., et al. (2017), Decreased STARD10 Expression Is Associated with Defective Insulin Secretion in Humans and Mice, *The American Journal of Human Genetics*, *100*(2), 238–256, doi:10.1016/j.ajhg.2017.01.011.

[21] Chandra, V., et al. (2014), RFX6 regulates insulin secretion by modulating Ca2+ homeostasis in human β cells, *Cell Rep*, *9*(6), 2206–2218, doi:10.1016/j.celrep.2014.11.010.

[22] Chang, C. C., C. C. Chow, L. C. Tellier, S. Vattikuti, S. M. Purcell, and J. J. Lee (2015), Second-generation PLINK: Rising to the challenge of larger and richer datasets, *Gigascience*, *4*(1), 1–16, doi:10.1186/s13742-015-0047-8.

[23] Chen, K., et al. (2015), Broad H3K4me3 is associated with increased transcription elongation and enhancer activity at tumor-suppressor genes, *Nature Genetics*, *47*(10), 1149–1157, doi:10.1038/ng.3385.

[24] Civelek, M., and A. J. Lusis (2014), Systems genetics approaches to understand complex traits, *Nat. Rev. Genet.*, *15*(1), 34–48, doi:10.1038/nrg3575.

[25] Civelek, M., et al. (2017), Genetic Regulation of Adipose Gene Expression and Cardio-Metabolic Traits, *The American Journal of Human Genetics*, *100*(3), 428–443, doi:10.1016/j.ajhg.2017.01.027.

[26] Claussnitzer, M., et al. (2015), FTO Obesity Variant Circuitry and Adipocyte Browning in Humans, *New England Journal of Medicine*, *373*(10), 895–907, doi:10.1056/NEJMoa1502214.

[27] Core, L. J., J. J. Waterfall, and J. T. Lis (2008), Nascent RNA Sequencing Reveals Widespread Pausing and Divergent Initiation at Human Promoters, *Science*, *322*(5909), 1845–1848, doi:10.1126/science.1162228.

[28] Core, L. J., A. L. Martins, C. G. Danko, C. Waters, A. Siepel, and J. T. Lis (2014), Analysis of nascent RNA identifies a unified architecture of initiation regions at mammalian promoters and enhancers, *Nat Genet*, *46*(12), 1311–1320, doi:10.1038/ng.3142.

[29] Corradin, O., A. Saiakhova, B. Akhtar-Zaidi, L. Myeroff, J. Willis, R. Cowper-Sal·lari, M. Lupien, S. Markowitz, and P. C. Scacheri (2014), Combinatorial effects of multiple enhancer variants in linkage disequilibrium dictate levels of gene expression to confer susceptibility to common traits, *Genome Res.*, *24*(1), 1–13, doi:10.1101/gr.164079.113.

[30] Danecek, P., et al. (2011), The variant call format and VCFtools, *Bioinformatics*, *27*(15), 2156–2158, doi:10.1093/bioinformatics/btr330.

[31] Dave, K., et al. (2017), Mice deficient of Myc super-enhancer region reveal differential control mechanism between normal and pathological growth, *eLife Sciences*, *6*, e23,382, doi:10.7554/eLife.23382.

[32] Degner, J. F., et al. (2012), DNase I sensitivity QTLs are a major determinant of human expression variation, *Nature*, *482*(7385), 390, doi:10.1038/nature10808.

[33] Delaneau, O., J.-F. Zagury, and J. Marchini (2013), Improved whole-chromosome phasing for disease and population genetic studies, *Nat. Methods*, *10*(1), 5–6, doi:10.1038/nmeth.2307.

[34] Delaneau, O., et al. (2014), Integrating sequence and array data to create an improved 1000 Genomes Project haplotype reference panel, *Nature Communications*, *5*, 3934, doi:10.1038/ncomms4934.

[35] Dickson, S. P., K. Wang, I. Krantz, H. Hakonarson, and D. B. Goldstein (2010), Rare Variants Create Synthetic Genome-Wide Associations, *PLOS Biology*, *8*(1), e1000,294, doi:10.1371/journal.pbio.1000294.

[36] Dimas, A. S., et al. (2014), Impact of Type 2 Diabetes Susceptibility Variants on Quantitative Glycemic Traits Reveals Mechanistic Heterogeneity, *Diabetes*, *63*(6), 2158–2171, doi:10.2337/db13-0949.

[37] Dobin, A., C. A. Davis, F. Schlesinger, J. Drenkow, C. Zaleski, S. Jha, P. Batut, M. Chaisson, and T. R. Gingeras (2013), STAR: Ultrafast universal RNA-seq aligner, *Bioinformatics*, *29*(1), 15–21, doi:10.1093/bioinformatics/bts635.

[38] Dong, X., T.-W. Chang, S. T. Weiss, and W. Qiu (2017), powerEQTL: Power and Sample Size Calculation for eQTL Analysis.

[39] Duggirala, R., J. Blangero, L. Almasy, T. D. Dyer, K. L. Williams, R. J. Leach, P. O'Connell, and M. P. Stern (1999), Linkage of type 2 diabetes mellitus and of age at onset to a genetic location on chromosome 10q in Mexican Americans., *Am J Hum Genet*, *64*(4), 1127–1140.

[40] Ernst, J., and M. Kellis (2010), Discovery and characterization of chromatin states for systematic annotation of the human genome, *Nature Biotechnology*, *28*(8), 817–825, doi:10.1038/nbt.1662.

[41] Ernst, J., and M. Kellis (2012), ChromHMM: Automating chromatin state discovery and characterization, *Nat Methods*, *9*(3), 215–216, doi:10.1038/nmeth.1906.

[42] Ernst, J., A. Melnikov, X. Zhang, L. Wang, P. Rogov, T. S. Mikkelsen, and M. Kellis (2016), Genome-scale high-resolution mapping of activating and repressive nucleotides in regulatory regions, *Nature Biotechnology*, *34*(11), 1180–1190, doi:10.1038/nbt.3678.

[43] Ernst, J., et al. (2011), Mapping and analysis of chromatin state dynamics in nine human cell types, *Nature*, *473*(7345), 43–49, doi:10.1038/nature09906.

[44] Fadista, J., et al. (2014), Global genomic and transcriptomic analysis of human pancreatic islets reveals novel genes influencing glucose metabolism., *Proceedings of the National Academy of Sciences of the United States of America*, *111*(38), 13,924–13,929.

[45] Fogarty, M. P., M. E. Cannon, S. Vadlamudi, K. J. Gaulton, and K. L. Mohlke (2014), Identification of a Regulatory Variant That Binds FOXA1 and FOXA2 at the CDC123/CAMK1D Type 2 Diabetes GWAS Locus, *PLOS Genetics*, *10*(9), e1004,633, doi:10.1371/journal.pgen.1004633.

[46] Frith, M. C., E. Valen, A. Krogh, Y. Hayashizaki, P. Carninci, and A. Sandelin (2008), A code for transcription initiation in mammalian genomes, *Genome Res.*, *18*(1), 1–12, doi:10.1101/gr.6831208.

[47] Fu, Y., M. Sinha, C. L. Peterson, and Z. Weng (2008), The Insulator Binding Protein CTCF Positions 20 Nucleosomes around Its Binding Sites across the Human Genome, *PLOS Genetics*, *4*(7), e1000,138, doi:10.1371/journal.pgen.1000138.

[48] Fuchsberger, C., G. R. Abecasis, and D. A. Hinds (2015), Minimac2: Faster genotype imputation, *Bioinformatics*, *31*(5), 782–784, doi:10.1093/bioinformatics/btu704.

[49] Fuchsberger, C., et al. (2016), The genetic architecture of type 2 diabetes, *Nature*, *536*(7614), nature18,642, doi:10.1038/nature18642.

[50] Gamazon, E. R., et al. (2018), Using an atlas of gene regulation across 44 human tissues to inform complex disease- and trait-associated variation, *Nature Genetics*, *50*(7), 956, doi:10.1038/s41588-018-0154-4.

[51] Gaulton, K. J., et al. (2010), A map of open chromatin in human pancreatic islets, *Nature Genetics*, *42*(3), 255–259, doi:10.1038/ng.530.

[52] Gershengorn, M. C., A. A. Hardikar, C. Wei, E. Geras-Raaka, B. Marcus-Samuels, and B. M. Raaka (2004), Epithelial-to-Mesenchymal Transition Generates Proliferative Human Islet Precursor Cells, *Science*, *306*(5705), 2261–2264.

[53] Giambartolomei, C., D. Vukcevic, E. E. Schadt, L. Franke, A. D. Hingorani, C. Wallace, and V. Plagnol (2014), Bayesian Test for Colocalisation between Pairs of Genetic Association Studies Using Summary Statistics, *PLOS Genetics*, *10*(5), e1004,383, doi:10.1371/journal.pgen.1004383.

[54] Girard, C., F. Duprat, C. Terrenoire, N. Tinel, M. Fosset, G. Romey, M. Lazdunski, and F. Lesage (2001), Genomic and Functional Characteristics of Novel Human Pancreatic 2P Domain K+ Channels, *Biochemical and Biophysical Research Communications*, *282*(1), 249–256, doi:10.1006/bbrc.2001.4562.

[55] Glassberg, E. C., Z. Gao, A. Harpak, X. Lan, and J. K. Pritchard (2019), Evidence for Weak Selective Constraint on Human Gene Expression, *Genetics*, *211*(2), 757–772, doi:10.1534/genetics.118.301833.

[56] Grant, S. F. A., et al. (2006), Variant of transcription factor 7-like 2 (*TCF7L2*) gene confers risk of type 2 diabetes, *Nature Genetics*, *38*(3), 320–323, doi:10.1038/ng1732.

[57] Gross, D. S., and W. T. Garrard (1988), Nuclease Hypersensitive Sites in Chromatin, *Annu. Rev. Biochem.*, *57*(1), 159–197, doi:10.1146/annurev.bi.57.070188.001111.

[58] GTEx Consortium (2017), Genetic effects on gene expression across human tissues, *Nature*, *550*(7675), 204, doi:10.1038/nature24277.

[59] Guo, C., et al. (2015), Coordinated regulatory variation associated with gestational hyperglycaemia regulates expression of the novel hexokinase HKDC1, *Nat Commun*, *6*, 6069, doi:10.1038/ncomms7069.

[60] Gupta, S., J. A. Stamatoyannopoulos, T. L. Bailey, and W. S. Noble (2007), Quantifying similarity between motifs, *Genome Biol.*, *8*(2), R24, doi:10.1186/gb-2007-8-2-r24.

[61] Harrow, J., et al. (2012), GENCODE: The reference human genome annotation for The ENCODE Project, *Genome Res.*, *22*(9), 1760–1774, doi:10.1101/gr.135350.111.

[62] Hartley, S. W., and J. C. Mullikin (2015), QoRTs: A comprehensive toolset for quality control and data processing of RNA-Seq experiments, *BMC Bioinformatics*, *16*, 224, doi:10.1186/s12859-015-0670-5.

[63] Hay, D., et al. (2016), Genetic dissection of the α-globin super-enhancer in vivo, *Nat Genet*, *48*(8), 895–903, doi:10.1038/ng.3605.

[64] He, B., C. Chen, L. Teng, and K. Tan (2014), Global view of enhancer–promoter interactome in human cells, *PNAS*, *111*(21), E2191–E2199, doi:10.1073/pnas.1320308111.

[65] Heger, A., C. Webber, M. Goodson, C. P. Ponting, and G. Lunter (2013), GAT: A simulation framework for testing the association of genomic intervals, *Bioinformatics*, *29*(16), 2046–2048, doi:10.1093/bioinformatics/btt343.

[66] Heintzman, N. D., et al. (2007), Distinct and predictive chromatin signatures of transcriptional promoters and enhancers in the human genome, *Nature Genetics*, *39*(3), 311–318, doi:10.1038/ng1966.

[67] Hemani, G., K. Tilling, and G. D. Smith (2017), Orienting the causal relationship between imprecisely measured traits using GWAS summary data, *PLOS Genetics*, *13*(11), e1007,081, doi:10.1371/journal.pgen.1007081.

[68] Hindorff, L. A., P. Sethupathy, H. A. Junkins, E. M. Ramos, J. P. Mehta, F. S. Collins, and T. A. Manolio (2009), Potential etiologic and functional implications of genome-wide association loci for human diseases and traits, *Proc. Natl. Acad. Sci. U.S.A.*, *106*(23), 9362–9367, doi:10.1073/pnas.0903103106.

[69] Hnisz, D., B. J. Abraham, T. I. Lee, A. Lau, V. Saint-André, A. a Sigova, H. a Hoke, and R. a Young (2013), Super-Enhancers in the Control of Cell Identity and Disease, *Cell*, *155*(4), 934–947, doi:10.1016/j.cell.2013.09.053.

[70] Hormozdiari, F., et al. (2016), Colocalization of GWAS and eQTL Signals Detects Target Genes, *The American Journal of Human Genetics*, *99*(6), 1245–1260, doi:10.1016/j.ajhg.2016.10.003.

[71] Howie, B., C. Fuchsberger, M. Stephens, J. Marchini, and G. R. Abecasis (2012), Fast and accurate genotype imputation in genome-wide association studies through pre-phasing, *Nat. Genet.*, *44*(8), 955–959, doi:10.1038/ng.2354.

[72] Howie, B. N., P. Donnelly, and J. Marchini (2009), A flexible and accurate genotype imputation method for the next generation of genome-wide association studies, *PLoS Genet.*, *5*(6), e1000,529, doi:10.1371/journal.pgen.1000529.

[73] Hsieh, C.-L., et al. (2014), Enhancer RNAs participate in androgen receptor-driven looping that selectively enhances gene activation, *PNAS*, *111*(20), 7319–7324, doi:10.1073/pnas.1324151111.

[74] Huang, H., et al. (2017), Fine-mapping inflammatory bowel disease loci to single-variant resolution, *Nature*, *547*(7662), 173, doi:10.1038/nature22969.

[75] Huopio, H., et al. (2016), Clinical, Genetic, and Biochemical Characteristics of Early-Onset Diabetes in the Finnish Population, *J. Clin. Endocrinol. Metab.*, *101*(8), 3018–3026, doi:10.1210/jc.2015-4296.

[76] Jiang, L., F. Schlesinger, C. A. Davis, Y. Zhang, R. Li, M. Salit, T. R. Gingeras, and B. Oliver (2011), Synthetic spike-in standards for RNA-seq experiments, *Genome Res.*, *21*(9), 1543–1551, doi:10.1101/gr.121095.111.

[77] Jolma, A., et al. (2013), DNA-Binding Specificities of Human Transcription Factors, *Cell*, *152*(1), 327–339, doi:10.1016/j.cell.2012.12.009.

[78] Jun, G., M. Flickinger, K. N. Hetrick, J. M. Romm, K. F. Doheny, G. R. Abecasis, M. Boehnke, and H. M. Kang (2012), Detecting and estimating contamination of human DNA samples in sequencing and array-based genotype data, *Am. J. Hum. Genet.*, *91*(5), 839–848, doi:10.1016/j.ajhg.2012.09.004.

[79] Kaikkonen, M. U., et al. (2013), Remodeling of the Enhancer Landscape during Macrophage Activation Is Coupled to Enhancer Transcription, *Molecular Cell*, *51*(3), 310–325, doi:10.1016/j.molcel.2013.07.010.

[80] Kanduri, C., C. Bock, S. Gundersen, E. Hovig, and G. K. Sandve (), Colocalization analyses of genomic elements: Approaches, recommendations and challenges, *Bioinformatics*, doi:10.1093/bioinformatics/bty835.

[81] Keller, M. P., et al. (2008), A gene expression network model of type 2 diabetes links cell cycle regulation in islets with diabetes susceptibility, *Genome Res*, *18*(5), 706–716, doi:10.1101/gr.074914.107.

[82] Khera, A. V., et al. (2018), Genome-wide polygenic scores for common diseases identify individuals with risk equivalent to monogenic mutations, *Nature Genetics*, *50*(9), 1219–1224, doi:10.1038/s41588-018-0183-z.

[83] Kheradpour, P., and M. Kellis (2014), Systematic discovery and characterization of regulatory motifs in ENCODE TF binding experiments, *Nucleic Acids Res.*, *42*(5), 2976–2987, doi:10.1093/nar/gkt1249.

[84] Kim, A., K. Miller, J. Jo, G. Kilimnik, P. Wojcik, and M. Hara (2009), Islet architecture: A comparative study, *Islets*, *1*(2), 129–136, doi:10.4161/isl.1.2.9480.

[85] Kim, T.-K., et al. (2010), Widespread transcription at neuronal activity-regulated enhancers, *Nature*, *465*(7295), 182–187, doi:10.1038/nature09033.

[86] Köster, J., and S. Rahmann (2012), Snakemake—a scalable bioinformatics workflow engine, *Bioinformatics*, *28*(19), 2520–2522, doi:10.1093/bioinformatics/bts480.

[87] Kriegel, M. A., C. Rathinam, and R. A. Flavell (2012), Pancreatic islet expression of chemokine CCL2 suppresses autoimmune diabetes via tolerogenic CD11c+ CD11b+ dendritic cells, *Proc. Natl. Acad. Sci. U.S.A.*, *109*(9), 3457–3462, doi:10.1073/pnas.1115308109.

[88] Kudron, M. M., et al. (2017), The modERN Resource: Genome-Wide Binding Profiles for Hundreds of Drosophila and Caenorhabditis elegans Transcription Factors, *Genetics*, doi:10.1534/genetics.117.300657.

[89] Kulzer, J. R., et al. (2014), A common functional regulatory variant at a type 2 diabetes locus upregulates ARAP1 expression in the pancreatic beta cell, *Am. J. Hum. Genet.*, *94*(2), 186–197, doi:10.1016/j.ajhg.2013.12.011.

[90] Kumasaka, N., A. J. Knights, and D. J. Gaffney (2019), High-resolution genetic mapping of putative causal interactions between regions of open chromatin, *Nature Genetics*, *51*(1), 128, doi:10.1038/s41588-018-0278-6.

[91] Kvon, E. Z., G. Stampfel, J. O. Yáñez-Cuna, B. J. Dickson, and A. Stark (2012), HOT regions function as patterned developmental enhancers and have a distinct cis-regulatory signature, *Genes Dev.*, *26*(9), 908–913, doi:10.1101/gad.188052.112.

[92] Kyono, Y., J. O. Kitzman, and S. C. J. Parker (2019), Genomic annotation of disease-associated variants reveals shared functional contexts, *Diabetologia*, doi:10.1007/s00125-019-4823-3.

[93] Landt, S. G., et al. (2012), ChIP-seq guidelines and practices of the ENCODE and modENCODE consortia, *Genome Res.*, *22*(9), 1813–1831, doi:10.1101/gr.136184.111.

[94] Lappalainen, T., et al. (2013), Transcriptome and genome sequencing uncovers functional variation in humans, *Nature*, *501*(7468), 506–511, doi:https://doi.org/10.1038/nature12531.

[95] Lek, M., et al. (2016), Analysis of protein-coding genetic variation in 60,706 humans, *Nature*, *536*(7616), 285–291, doi:10.1038/nature19057.

[96] Li, H. (2013), Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM, *arXiv:1303.3997 [q-bio]*.

[97] Li, H., and R. Durbin (2009), Fast and accurate short read alignment with Burrows-Wheeler transform, *Bioinformatics*, *25*(14), 1754–1760, doi:10.1093/bioinformatics/btp324.

[98] Li, H., B. Handsaker, A. Wysoker, T. Fennell, J. Ruan, N. Homer, G. Marth, G. Abecasis, and R. Durbin (2009), The Sequence Alignment/Map format and SAMtools, *Bioinformatics*, *25*(16), 2078–2079, doi:10.1093/bioinformatics/btp352.

[99] Li, H., F. Liu, C. Ren, X. Bo, and W. Shu (2016), Genome-wide identification and characterisation of HOT regions in the human genome, *BMC Genomics*, *17*, doi:10.1186/s12864-016-3077-4.

[100] Li, W., et al. (2013), Functional roles of enhancer RNAs for oestrogen-dependent transcriptional activation, *Nature*, *498*(7455), 516–520, doi:10.1038/nature12210.

[101] Li, Y. R., and B. J. Keating (2014), Trans-ethnic genome-wide association studies: Advantages and challenges of mapping in diverse populations, *Genome Med*, *6*(10), 91, doi:10.1186/s13073-014-0091-5.

[102] Lin, C. Y., et al. (2016), Active medulloblastoma enhancers reveal subgroup-specific cellular origins, *Nature*, *530*(7588), 57–62, doi:10.1038/nature16546.

[103] Liu, X., et al. (2017), Functional Architectures of Local and Distal Regulation of Gene Expression in Multiple Human Tissues, *The American Journal of Human Genetics*, *100*(4), 605–616, doi:10.1016/j.ajhg.2017.03.002.

[104] Lizio, M., et al. (2015), Mapping Mammalian Cell-type-specific Transcriptional Regulatory Networks Using KD-CAGE and ChIP-seq Data in the TC-YIK Cell Line, *Front Genet*, *6*, 331, doi:10.3389/fgene.2015.00331.

[105] Lopes, R., R. Agami, and G. Korkmaz (2017), GRO-seq, A Tool for Identification of Transcripts Regulating Gene Expression, *Methods Mol. Biol.*, *1543*, 45–55, doi:10.1007/978-1-4939-6716-2_3.

[106] Lotshaw, D. P. (2007), Biophysical, pharmacological, and functional characteristics of cloned and native mammalian two-pore domain K+ channels, *Cell Biochem Biophys*, *47*(2), 209–256, doi:10.1007/s12013-007-0007-8.

[107] Lotta, L. A., et al. (2017), Integrative genomic analysis implicates limited peripheral adipose storage capacity in the pathogenesis of human insulin resistance, *Nature Genetics*, *49*(1), 17–26, doi:10.1038/ng.3714.

[108] Lovén, J., H. a Hoke, C. Y. Lin, A. Lau, D. a Orlando, C. R. Vakoc, J. E. Bradner, T. I. Lee, and R. a Young (2013), Selective Inhibition of Tumor Oncogenes by Disruption of Super-Enhancers, *Cell*, *153*(2), 320–334, doi:10.1016/j.cell.2013.03.036.

[109] Maehara, K., and Y. Ohkawa (2015), Agplus: A rapid and flexible tool for aggregation plots, *Bioinformatics*, *31*(18), 3046–3047, doi:10.1093/bioinformatics/btv322.

[110] Mahajan, A., et al. (2018), Fine-mapping type 2 diabetes loci to single-variant resolution using high-density imputation and islet-specific epigenome maps, *Nature Genetics*, *50*(11), 1505, doi:10.1038/s41588-018-0241-6.

[111] Mahajan, A., et al. (2018), Refining the accuracy of validated target identification through coding variant fine-mapping in type 2 diabetes, *Nature Genetics*, *50*(4), 559, doi:10.1038/s41588-018-0084-1.

[112] Manichaikul, A., J. C. Mychaleckyj, S. S. Rich, K. Daly, M. Sale, and W.-M. Chen (2010), Robust relationship inference in genome-wide association studies, *Bioinformatics*, *26*(22), 2867–2873, doi:10.1093/bioinformatics/btq559.

[113] Manning, A. K., et al. (2012), A genome-wide approach accounting for body mass index identifies genetic variants influencing fasting glycemic traits and insulin resistance, *Nat. Genet.*, *44*(6), 659–669, doi:10.1038/ng.2274.

[114] Manning Fox, J. E., et al. (2015), Human islet function following 20 years of cryogenic biobanking, *Diabetologia*, *58*(7), 1503–1512, doi:10.1007/s00125-015-3598-4.

[115] Manolio, T. A., et al. (2009), Finding the missing heritability of complex diseases, *Nature*, *461*(7265), 747–753, doi:10.1038/nature08494.

[116] Marco-Sola, S., M. Sammeth, R. Guigó, and P. Ribeca (2012), The GEM mapper: Fast, accurate and versatile alignment by filtration, *Nature Methods*, *9*(12), 1185–1188, doi:10.1038/nmeth.2221.

[117] Martin, M. (2011), Cutadapt removes adapter sequences from high-throughput sequencing reads, *EMBnet.journal*, *17*(1), 10–12, doi:10.14806/ej.17.1.200.

[118] Mathelier, A., et al. (2014), JASPAR 2014: An extensively expanded and updated open-access database of transcription factor binding profiles, *Nucleic Acids Res.*, *42*(Database issue), D142–147, doi:10.1093/nar/gkt997.

[119] Maurano, M. T., E. Haugen, R. Sandstrom, J. Vierstra, A. Shafer, R. Kaul, and J. A. Stamatoyannopoulos (2015), Large-scale identification of sequence variants influencing human transcription factor occupancy *in vivo*, *Nature Genetics*, *47*(12), 1393–1401, doi:10.1038/ng.3432.

[120] Maurano, M. T., et al. (2012), Systematic Localization of Common Disease-Associated Variation in Regulatory DNA, *Science*, *337*(6099), 1190–1195, doi: 10.1126/science.1222794.

[121] McKenna, A., et al. (2010), The Genome Analysis Toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data, *Genome Res.*, *20*(9), 1297–1303, doi:10.1101/gr.107524.110.

[122] Meier, J. J., and R. C. Bonadonna (2013), Role of reduced β-cell mass versus impaired β-cell function in the pathogenesis of type 2 diabetes, *Diabetes Care*, *36 Suppl 2*, S113–119, doi:10.2337/dcS13-2008.

[123] Melgar, M. F., F. S. Collins, and P. Sethupathy (2011), Discovery of active enhancers through bidirectional expression of short transcripts, *Genome Biology*, *12*(11), R113, doi:10.1186/gb-2011-12-11-r113.

[124] Mikhaylichenko, O., V. Bondarenko, D. Harnett, I. E. Schor, M. Males, R. R. Viales, and E. E. M. Furlong (2018), The degree of enhancer or promoter activity is reflected by the levels and directionality of eRNA transcription, *Genes Dev.*, *32*(1), 42–57, doi:10.1101/gad.308619.117.

[125] Mikkelsen, T. S., Z. Xu, X. Zhang, L. Wang, J. M. Gimble, E. S. Lander, and E. D. Rosen (2010), Comparative Epigenomic Analysis of Murine and Human Adipogenesis, *Cell*, *143*(1), 156–169, doi:10.1016/j.cell.2010.09.006.

[126] Mikkelsen, T. S., et al. (2007), Genome-wide maps of chromatin state in pluripotent and lineage-committed cells, *Nature*, *448*(7153), 553–560, doi: 10.1038/nature06008.

[127] Millstein, J., B. Zhang, J. Zhu, and E. E. Schadt (2009), Disentangling molecular relationships with a causal inference test, *BMC Genetics*, *10*(1), 23, doi: 10.1186/1471-2156-10-23.

[128] Miyazaki, J., K. Araki, E. Yamato, H. Ikegami, T. Asano, Y. Shibasaki, Y. Oka, and K. Yamamura (1990), Establishment of a pancreatic beta cell line that retains glucose-inducible insulin secretion: Special reference to expression of glucose transporter isoforms, *Endocrinology*, *127*(1), 126–132, doi: 10.1210/endo-127-1-126.

[129] Mohlke, K. L., and M. Boehnke (2015), Recent advances in understanding the genetic architecture of type 2 diabetes., *Human Molecular Genetics*, *24*(R1), R85–92.

[130] Moltke, I., et al. (2014), A common Greenlandic *TBC1D4* variant confers muscle insulin resistance and type 2 diabetes, *Nature*, *512*(7513), 190–193, doi: 10.1038/nature13425.

[131] Montgomery, S. B., M. Sammeth, M. Gutierrez-Arcelus, R. P. Lach, C. Ingle, J. Nisbett, R. Guigo, and E. T. Dermitzakis (2010), Transcriptome genetics using second generation sequencing in a Caucasian population, *Nature*, *464*(7289), 773–777, doi:10.1038/nature08903.

[132] Moorman, C., et al. (2006), Hotspots of transcription factor colocalization in the genome of Drosophila melanogaster, *PNAS*, *103*(32), 12,027–12,032, doi:10.1073/pnas.0605003103.

[133] Moorthy, S. D., et al. (2017), Enhancers and super-enhancers have an equivalent regulatory role in embryonic stem cells through regulation of single or multiple genes, *Genome Res.*, *27*(2), 246–258, doi:10.1101/gr.210930.116.

[134] Moyerbrailean, G. A., C. A. Kalita, C. T. Harvey, X. Wen, F. Luca, and R. Pique-Regi (2016), Which Genetics Variants in DNase-Seq Footprints Are More Likely to Alter Binding?, *PLOS Genetics*, *12*(2), e1005,875, doi:10.1371/journal.pgen.1005875.

[135] Murata, M., H. Nishiyori-Sueki, M. Kojima-Ishiyama, P. Carninci, Y. Hayashizaki, and M. Itoh (2014), Detecting Expressed Genes Using CAGE, in *Transcription Factor Regulatory Networks: Methods and Protocols*, edited by E. Miyamoto-Sato, H. Ohashi, H. Sasaki, J.-i. Nishikawa, and H. Yanagawa, Methods in Molecular Biology, pp. 67–85, Springer New York, New York, NY, doi:10.1007/978-1-4939-0805-9_7.

[136] Nica, A. C., S. B. Montgomery, A. S. Dimas, B. E. Stranger, C. Beazley, I. Barroso, and E. T. Dermitzakis (2010), Candidate Causal Regulatory Effects by Integration of Expression QTLs with Complex Trait Genetic Associations, *PLOS Genetics*, *6*(4), e1000,895, doi:10.1371/journal.pgen.1000895.

[137] Nica, A. C., H. Ongen, J.-C. Irminger, D. Bosco, T. Berney, S. E. Antonarakis, P. A. Halban, and E. T. Dermitzakis (2013), Cell-type, allelic, and genetic signatures in the human pancreatic beta cell transcriptome, *Genome Res.*, *23*(9), 1554–1562, doi:10.1101/gr.150706.112.

[138] Onengut-Gumuscu, S., et al. (2015), Fine mapping of type 1 diabetes susceptibility loci and evidence for colocalization of causal variants with lymphoid gene enhancers, *Nature Genetics*, *47*(4), 381–386, doi:10.1038/ng.3245.

[139] Ongen, H., A. Buil, A. A. Brown, E. T. Dermitzakis, and O. Delaneau (2016), Fast and efficient QTL mapper for thousands of molecular phenotypes, *Bioinformatics*, *32*(10), 1479–1485, doi:10.1093/bioinformatics/btv722.

[140] Ongen, H., A. A. Brown, O. Delaneau, N. I. Panousis, A. C. Nica, G. Consortium, and E. T. Dermitzakis (2017), Estimating the causal tissues for complex traits and diseases, *Nature Genetics*, p. ng.3981, doi:10.1038/ng.3981.

[141] Ottosson-Laakso, E., et al. (2017), Glucose-induced Changes in Gene Expression in Human Pancreatic Islets – Causes or Consequences of Chronic Hyperglycemia, *Diabetes*, p. db170311, doi:10.2337/db17-0311.

[142] Parker, S. C. J., et al. (2013), Chromatin stretch enhancer states drive cell-specific gene regulation and harbor human disease risk variants, *PNAS*, *110*(44), 17,921–17,926, doi:10.1073/pnas.1317023110.

[143] Parnaud, G., et al. (2008), Proliferation of sorted human and rat beta cells, *Diabetologia*, *51*(1), 91–100, doi:10.1007/s00125-007-0855-1.

[144] Pasquali, L., et al. (2014), Pancreatic islet enhancer clusters enriched in type 2 diabetes risk-associated variants, *Nature Genetics*, *46*(2), 136–143, doi:10.1038/ng.2870.

[145] Piccand, J., P. Strasser, D. J. Hodson, A. Meunier, T. Ye, C. Keime, M.-C. Birling, G. A. Rutter, and G. Gradwohl (2014), Rfx6 Maintains the Functional Identity of Adult Pancreatic β Cells, *Cell Rep*, *9*(6), 2219–2232, doi:10.1016/j.celrep.2014.11.033.

[146] Pickrell, J. K. (2014), Joint Analysis of Functional Genomic Data and Genome-wide Association Studies of 18 Human Traits, *Am J Hum Genet*, *94*(4), 559–573, doi:10.1016/j.ajhg.2014.03.004.

[147] Pique-Regi, R., J. F. Degner, A. A. Pai, D. J. Gaffney, Y. Gilad, and J. K. Pritchard (2011), Accurate inference of transcription factor binding from DNA sequence and chromatin accessibility data, *Genome Res.*, *21*(3), 447–455, doi:10.1101/gr.112623.110.

[148] Pott, S., and J. D. Lieb (2015), What are super-enhancers?, *Nat Genet*, *47*(1), 8–12, doi:10.1038/ng.3167.

[149] Poulsen, P., K. O. Kyvik, A. Vaag, and H. Beck-Nielsen (1999), Heritability of type II (non-insulin-dependent) diabetes mellitus and abnormal glucose tolerance–a population-based twin study, *Diabetologia*, *42*(2), 139–145.

[150] Prasad, R. B., and L. Groop (2015), Genetics of Type 2 Diabetes—Pitfalls and Possibilities, *Genes*, *6*(1), 87–123, doi:10.3390/genes6010087.

[151] Price, A. L., et al. (2008), Long-Range LD Can Confound Genome Scans in Admixed Populations, *Am J Hum Genet*, *83*(1), 132–135, doi:10.1016/j.ajhg.2008.06.005.

[152] Purcell, S., et al. (2007), PLINK: A Tool Set for Whole-Genome Association and Population-Based Linkage Analyses, *The American Journal of Human Genetics*, *81*(3), 559–575, doi:10.1086/519795.

[153] Quang, D. X., M. R. Erdos, S. C. J. Parker, and F. S. Collins (2015), Motif signatures in stretch enhancers are enriched for disease-associated genetic variants, *Epigenetics & Chromatin*, *8*, 23, doi:10.1186/s13072-015-0015-7.

[154] Quinlan, A. R., and I. M. Hall (2010), BEDTools: A flexible suite of utilities for comparing genomic features, *Bioinformatics*, *26*(6), 841–842, doi:10.1093/bioinformatics/btq033.

[155] Roman, T. S., et al. (2017), A Type 2 Diabetes–Associated Functional Regulatory Variant in a Pancreatic Islet Enhancer at the ADCY5 Locus, *Diabetes*, *66*(9), 2521–2530, doi:10.2337/db17-0464.

[156] Rozowsky, J., et al. (2011), AlleleSeq: Analysis of allele-specific expression and binding in a network framework, *Mol. Syst. Biol.*, *7*, 522, doi:10.1038/msb.2011.54.

[157] Saint-André, V., A. J. Federation, C. Y. Lin, B. J. Abraham, J. Reddy, T. I. Lee, J. E. Bradner, and R. A. Young (2016), Models of human core transcriptional regulatory circuitries, *Genome Res.*, *26*(3), 385–396, doi:10.1101/gr.197590.115.

[158] Sandelin, A., W. Alkema, P. Engström, W. W. Wasserman, and B. Lenhard (2004), JASPAR: An open-access database for eukaryotic transcription factor binding profiles, *Nucleic Acids Res.*, *32*(Database issue), D91–94, doi:10.1093/nar/gkh012.

[159] Saxena, R., et al. (2007), Genome-Wide Association Analysis Identifies Loci for Type 2 Diabetes and Triglyceride Levels, *Science*, *316*(5829), 1331–1336, doi:10.1126/science.1142358.

[160] Schmidt, E. M., J. Zhang, W. Zhou, J. Chen, K. L. Mohlke, Y. E. Chen, and C. J. Willer (2015), GREGOR: Evaluating global enrichment of trait-associated variants in epigenomic features using a systematic, data-driven approach, *Bioinformatics*, *31*(16), 2601–2606, doi:10.1093/bioinformatics/btv201.

[161] Schug, J., W.-P. Schuller, C. Kappen, J. M. Salbaum, M. Bucan, and C. J. Stoeckert (2005), Promoter features related to tissue specificity as measured by Shannon entropy, *Genome Biology*, *6*, R33, doi:10.1186/gb-2005-6-4-r33.

[162] Scott, L. J., et al. (2016), The genetic regulatory signature of type 2 diabetes in human skeletal muscle, *Nature Communications*, *7*, ncomms11,764, doi:10.1038/ncomms11764.

[163] Scott, R. A., et al. (2012), Large-scale association analyses identify new loci influencing glycemic traits and provide insight into the underlying biological pathways, *Nature Genetics*, *44*(9), 991–1005, doi:10.1038/ng.2385.

[164] Scott, R. A., et al. (2017), An Expanded Genome-Wide Association Study of Type 2 Diabetes in Europeans, *Diabetes*, p. db161253, doi:10.2337/db16-1253.

[165] Shabalin, A. A. (2012), Matrix eQTL: Ultra fast eQTL analysis via large matrix operations, *Bioinformatics*, *28*(10), 1353–1358, doi:10.1093/bioinformatics/bts163.

[166] Shin, H. Y., M. Willi, K. H. Yoo, X. Zeng, C. Wang, G. Metser, and L. Hennighausen (2016), Hierarchy within the mammary STAT5-driven Wap super-enhancer, *Nat Genet*, *48*(8), 904–911, doi:10.1038/ng.3606.

[167] Smith, S. B., et al. (2010), Rfx6 directs islet formation and insulin production in mice and humans, *Nature*, *463*(7282), 775–780, doi:10.1038/nature08748.

[168] Solimena, M., et al. (2018), Systems biology of the IMIDIA biobank from organ donors and pancreatectomised patients defines a novel transcriptomic signature of islets from individuals with type 2 diabetes, *Diabetologia*, *61*(3), 641–657, doi:10.1007/s00125-017-4500-3.

[169] Soyer, J., et al. (2010), Rfx6 is an Ngn3-dependent winged helix transcription factor required for pancreatic islet cell development, *Development*, *137*(2), 203–212, doi:10.1242/dev.041673.

[170] Stegle, O., L. Parts, R. Durbin, and J. Winn (2010), A Bayesian framework to account for complex non-genetic factors in gene expression levels greatly increases power in eQTL studies, *PLoS Comput. Biol.*, *6*(5), e1000,770, doi:10.1371/journal.pcbi.1000770.

[171] Stegle, O., L. Parts, M. Piipari, J. Winn, and R. Durbin (2012), Using probabilistic estimation of expression residuals (PEER) to obtain increased power and interpretability of gene expression analyses, *Nature Protocols*, *7*(3), 500, doi:10.1038/nprot.2011.457.

[172] Stitzel, M. L., et al. (2010), Global epigenomic analysis of primary human pancreatic islets provides insights into type 2 diabetes susceptibility loci, *Cell Metab.*, *12*(5), 443–455, doi:10.1016/j.cmet.2010.09.012.

[173] Storey, J. D. (2002), A direct approach to false discovery rates, *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *64*(3), 479–498, doi:10.1111/1467-9868.00346.

[174] Storey, J. D., and R. Tibshirani (2003), Statistical significance for genomewide studies, *Proc. Natl. Acad. Sci. U.S.A.*, *100*(16), 9440–9445, doi:10.1073/pnas.1530509100.

[175] Strawbridge, R. J., et al. (2011), Genome-Wide Association Identifies Nine Common Variants Associated With Fasting Proinsulin Levels and Provides New Insights Into the Pathophysiology of Type 2 Diabetes, *Diabetes*, *60*(10), 2624–2634, doi:10.2337/db11-0415.

[176] Stumvoll, M., B. J. Goldstein, and T. W. van Haeften (2005), Type 2 diabetes: Principles of pathogenesis and therapy, *The Lancet*, *365*(9467), 1333–1346, doi:10.1016/S0140-6736(05)61032-X.

[177] 't Hoen, P. A. C., et al. (2013), Reproducibility of high-throughput mRNA and small RNA sequencing across laboratories, *Nat. Biotechnol.*, *31*(11), 1015–1022, doi:10.1038/nbt.2702.

[178] T2D Knowledge Portal (2019), Type 2 Diabetes Knowledge Portal. Year Month Date of access, http://www.type2diabetesgenetics.org/home/portalHome.

[179] The 1000 Genomes Project Consortium (2015), A global reference for human genetic variation, *Nature*, *526*(7571), 68–74, doi:10.1038/nature15393.

[180] The DIAbetes Genetics Replication And Meta-analysis (DIAGRAM) Consortium, et al. (2012), Large-scale association analysis provides insights into the genetic architecture and pathophysiology of type 2 diabetes, *Nature Genetics*, *44*(9), 981–990, doi:10.1038/ng.2383.

[181] The DIAbetes Genetics Replication And Meta-analysis (DIAGRAM) Consortium, et al. (2014), Genome-wide trans-ancestry meta-analysis provides insight into the genetic architecture of type 2 diabetes susceptibility, *Nature Genetics*, *46*(3), 234–244, doi:10.1038/ng.2897.

[182] The ENCODE project Consortium (2012), An Integrated Encyclopedia of DNA Elements in the Human Genome, *Nature*, *489*(7414), 57–74, doi:10.1038/nature11247.

[183] The FANTOM Consortium, et al. (2014), A promoter-level mammalian expression atlas, *Nature*, *507*(7493), 462, doi:10.1038/nature13182.

[184] The Interanational HapMap Consortium (2003), The International HapMap Project, *Nature*, *426*(6968), 789, doi:10.1038/nature02168.

[185] The modENCODE Consortium, et al. (2010), Identification of Functional Elements and Regulatory Circuits by Drosophila modENCODE, *Science*, *330*(6012), 1787–1797, doi:10.1126/science.1198374.

[186] The Roadmap Epigenomics Consortium, et al. (2015), Integrative analysis of 111 reference human epigenomes, *Nature*, *518*(7539), 317–330, doi:10.1038/nature14248.

[187] Thibodeau, A., E. J. Márquez, D.-G. Shin, P. Vera-Licona, and D. Ucar (2017), Chromatin interaction networks revealed unique connectivity patterns of broad H3K4me3 domains and super enhancers in 3D chromatin, *Scientific Reports*, *7*(1), 14,466, doi:10.1038/s41598-017-14389-7.

[188] Thurner, M., et al. (2018), Integration of human pancreatic islet genomic data refines regulatory mechanisms at Type 2 Diabetes susceptibility loci, *eLife Sciences*, *7*, e31,977, doi:10.7554/eLife.31977.

[189] Trynka, G., C. Sandor, B. Han, H. Xu, B. E. Stranger, X. S. Liu, and S. Raychaudhuri (2013), Chromatin marks identify critical cell types for fine mapping complex trait variants, *Nat. Genet.*, *45*(2), 124–130, doi:10.1038/ng.2504.

[190] Tuomi, T., et al. (2016), Increased Melatonin Signaling Is a Risk Factor for Type 2 Diabetes, *Cell Metab.*, *23*(6), 1067–1077, doi:10.1016/j.cmet.2016.04.009.

[191] Udler, M. S., et al. (2018), Type 2 diabetes genetic loci informed by multi-trait associations point to disease mechanisms and subtypes: A soft clustering analysis, *PLOS Medicine*, *15*(9), e1002,654, doi:10.1371/journal.pmed.1002654.

[192] Vaisse, C., K. Clement, E. Durand, S. Hercberg, B. Guy-Grand, and P. Froguel (2000), Melanocortin-4 receptor mutations are a frequent and heterogeneous cause of morbid obesity, *J. Clin. Invest.*, *106*(2), 253–262, doi:10.1172/JCI9238.

[193] van de Bunt, M., et al. (2015), Transcript Expression Data from Human Islets Links Regulatory Signals from Genome-Wide Association Studies for Type 2 Diabetes and Glycemic Traits to Their Downstream Effectors, *PLOS Genetics*, *11*(12), e1005,694, doi:10.1371/journal.pgen.1005694.

[194] van de Geijn, B., G. McVicker, Y. Gilad, and J. K. Pritchard (2015), WASP: Allele-specific software for robust molecular quantitative trait locus discovery, *Nature Methods*, *12*(11), 1061, doi:10.1038/nmeth.3582.

[195] Varshney, A., H. VanRenterghem, P. Orchard, A. P. Boyle, M. L. Stitzel, D. Ucar, and S. C. J. Parker (2018), Cell Specificity of Human Regulatory Annotations and Their Genetic Effects on Gene Expression, *Genetics*, p. genetics.301525.2018, doi:10.1534/genetics.118.301525.

[196] Varshney, A., et al. (2017), Genetic regulatory signatures underlying islet gene expression and type 2 diabetes, *PNAS*, *114*(9), 2301–2306, doi:10.1073/pnas.1621192114.

[197] Vierra, N. C., P. K. Dadi, I. Jeong, M. Dickerson, D. R. Powell, and D. A. Jacobson (2015), Type 2 Diabetes–Associated K+ Channel TALK-1 Modulates β-Cell Electrical Excitability, Second-Phase Insulin Secretion, and Glucose Homeostasis, *Diabetes*, *64*(11), 3818–3828, doi:10.2337/db15-0280.

[198] Visel, A., S. Minovitsky, I. Dubchak, and L. A. Pennacchio (2007), VISTA Enhancer Browser—a database of tissue-specific human enhancers, *Nucleic Acids Res*, *35*(suppl_1), D88–D92, doi:10.1093/nar/gkl822.

[199] Voight, B. F., et al. (2010), Twelve type 2 diabetes susceptibility loci identified through large-scale association analysis, *Nat. Genet.*, *42*(7), 579–589, doi:10.1038/ng.609.

[200] Wang, J., et al. (2012), Sequence features and chromatin structure around the genomic regions bound by 119 human transcription factors, *Genome Res.*, *22*(9), 1798–1812, doi:10.1101/gr.139105.112.

[201] Wang, X., and D. B. Goldstein (2018), Enhancer redundancy predicts gene pathogenicity and informs complex disease gene discovery, *bioRxiv*, p. 459123, doi:10.1101/459123.

[202] Wang, Y. J., J. Schug, K.-J. Won, C. Liu, A. Naji, D. Avrahami, M. L. Golson, and K. H. Kaestner (2016), Single cell transcriptomics of the human endocrine pancreas, *Diabetes*, p. db160405.

[203] Welter, D., et al. (2014), The NHGRI GWAS Catalog, a curated resource of SNP-trait associations, *Nucleic Acids Res.*, *42*(Database issue), D1001–1006, doi:10.1093/nar/gkt1229.

[204] Whyte, W., D. Orlando, D. Hnisz, B. J. Abraham, C. Y. Lin, M. H. Kagey, P. B. Rahl, T. I. Lee, and R. a Young (2013), Master Transcription Factors and Mediator Establish Super-Enhancers at Key Cell Identity Genes, *Cell*, *153*(2), 307–319, doi:10.1016/j.cell.2013.03.035.

[205] Willer, C. J., Y. Li, and G. R. Abecasis (2010), METAL: Fast and efficient meta-analysis of genomewide association scans, *Bioinformatics*, *26*(17), 2190–2191, doi:10.1093/bioinformatics/btq340.

[206] Wood, A. R., et al. (2017), A Genome-Wide Association Study of IVGTT-Based Measures of First-Phase Insulin Secretion Refines the Underlying Physiology of Type 2 Diabetes Variants, *Diabetes*, *66*(8), 2296–2309, doi:10.2337/db16-1452.

[207] Xie, S., J. Duan, B. Li, P. Zhou, and G. C. Hon (2017), Multiplexed Engineering and Analysis of Combinatorial Enhancer Activity in Single Cells, *Molecular Cell*, *66*(2), 285–299.e5, doi:10.1016/j.molcel.2017.03.007.

[208] Yan, R., S. Lai, Y. Yang, H. Shi, Z. Cai, V. Sorrentino, H. Du, and H. Chen (2016), A novel type 2 diabetes risk allele increases the promoter activity of the muscle-specific small ankyrin 1 gene, *Sci Rep*, *6*, doi:10.1038/srep25105.

[209] Yang, Y., Z. Su, X. Song, B. Liang, F. Zeng, X. Chang, and D. Huang (2016), Enhancer RNA-driven looping enhances the transcription of the long noncoding RNA DHRS4-AS1, a controller of the DHRS4 gene cluster, *Sci Rep*, *6*, 20,961, doi:10.1038/srep20961.

[210] Zhou, V. W., A. Goren, and B. E. Bernstein (2011), Charting histone modifications and the functional organization of mammalian genomes, *Nature Reviews Genetics*, *12*(1), 7–18, doi:10.1038/nrg2905.

[211] Zhou, Y., et al. (2014), TCF7L2 is a master regulator of insulin production and processing, *Hum Mol Genet*, *23*(24), 6419–6431, doi:10.1093/hmg/ddu359.

[212] Zhu, Z., Q. V. Li, K. Lee, B. P. Rosen, F. González, C.-L. Soh, and D. Huangfu (2016), Genome Editing of Lineage Determinants in Human Pluripotent Stem Cells Reveals Mechanisms of Pancreatic Development and Diabetes, *Cell Stem Cell*, *18*(6), 755–768, doi:10.1016/j.stem.2016.03.015.