Using Integrative Multiomics Approaches to Dissect Type 2 Diabetes Genetic Risk in Pancreatic Islets

by

Vivek Rai

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
(Bioinformatics)
in the University of Michigan
2022

Doctoral Committee:

Associate Professor Stephen CJ Parker, Chair
Professor Veerabhadran Baladandayuthpani
Professor Margit Burmeister
Assistant Professor Jie Liu
Associate Professor Scott Soleimanpour

It is only with the heart that one can see rightly; what is essential is invisible to the eye.

Antoine de Saint-Exupéry
The Little Prince

Vivek Rai
vivekrai@umich.edu
ORCID iD: 0000-0002-2058-7238

To the undying compassion of my partner.

ACKNOWLEDGMENTS

We are the sum of our all experiences — and never have these words been truer than in the past five years of my life. Graduate school provides an opportunity for deep growth but not without its demands. However, I have been very fortunate to be surrounded by the most knowledgeable, kind, and above all, incredibly supportive people I have ever known.

First and foremost, I want to thank my dissertation supervisor Dr. Steve Parker for his support and guidance throughout my graduate training. He's an incredible mentor who strives for scientific rigor and career development of his trainees. I found shared excitement in how we put science and people at the core of our endeavors. While Steve was a foundational pillar for my scientific progress, I could not have made any progress if my colleagues at the Parker Lab would not have been kind to me with their time and words. I want to acknowledge the support and dedication of Drs. Ricardo Albanus, Arushi Varshney, Peter Orchard, and Daniel Quang. I collaborated on numerous projects with Dr. Nandini Manickam and Christa Ventresca who both did excellent work in generating high-quality data and insights — a contribution that spans most of my chapters in this thesis. I also want to acknowledge the support of Cynthia Zajac, Dr. Venkat Elangovan, Dr. Adelaide Tovar, Dr. Cassie Robertson who contributed significantly to my experience in the lab and cheered for me even though I did not get much opportunity to work with them.

Much of the work discussed in my thesis is a direct outcome of many productive collaborations throughout my training. I want to acknowledge my collaborators who have provided feedback, data, revisions, and mentorship that have been critical to my success. While too long to list, I want to acknowledge two key collaborations. First, I want to thank Drs. Narisu Narisu, Michael Erdos, Francis Collins at the National Institutes of Health (NIH) and Drs. Darren Cusanovich and Jay Shendure at the University of Washington who were pivotal for my work discussed in chapter 2, and second I want to thank Drs. Jack Walker, Diane Saunders, Marcela Brissova, and Alvin Powers at the Vanderbilt University (VU) for collaborating with us on a large body of work that I discuss in chapter 3.

During my PhD, one of my pillars for growth has been the continuous feedback and encouragement from my dissertation committee. Drs. Margit Burmeister, Veera Baladandayuthpani, Jie Liu, and Scott Soleimanpour have been invaluable mentors. Not only have they advised me on

the scientific merits of my work but have been forthcoming in helping me plan a career and ensure that I receive appropriate support. I also want to highlight and acknowledge the dual role of Dr. Margit Burmeister, who as the Director of the Bioinformatics department has helped me plan not only plan my courses but also navigate difficult situations during the pandemic due to travel restrictions. I am also thankful to many people in the Bioinformatics department who I have had the opportunity to interact on many occasions and who remain in touch with me beyond graduate school. Specially, I want to acknowledge the help of Julia Eussen and the department staff in helping me navigate the academic environment as an international student.

I also want to acknowledge the mentorship I received during my undergraduate education. Dr. Yannick Wurm for providing me the opportunity to work on Sequenceserver and getting me excited about bioinformatics software development. Dr. Gustavo Glusman at the Institute for Systems Biology in Seattle and Dr. Sandhya Visweswariah at the Indian Institute of Science Bangalore for hosting me in their respective labs and sharing their scientific knowledge and wisdom. Open Bioinformatics Foundation for selecting my Google Summer of Code proposal. And finally, my master's thesis advisor Dr. Amit Kumar Das at the Indian Institute of Technology Kharagpur for allowing me the freedom and flexibility to pursue my scientific curiosities.

Outside research, I want to thank my friends and Tango community who have contributed many memories and moments that I will cherish forever — Sagnik who has shared hundreds of laughs with me and suffered through my jokes, and the Michigan Argentine Tango Club and the tango friends for providing a creative and welcoming space for me to learn and grow.

Lastly and most importantly, I want to thank my partner Tanvi for her unconditional love and support throughout this journey. She prioritized our time together above everything, and I am eternally grateful for these wonderful years together. I look forward to continuing the next phase of our journey together.

I extend my heartfelt gratitude to you all for making this possible.

TABLE OF CONTENTS

LIST OF FIGURES

FIGURE

LIST OF TABLES

TABLE

LIST OF ACRONYMS

ATAC-seq Assay for Transposase Accessible Chromatin Sequencing

BMI Body Mass Index

BH Benjamini-Hochberg

CODEX Co-detection by Indexing

DE Differential Expression

EUR European (ancestry)

FACS/FANS Fluorescence Activated Cell/Nuclear Sorting

FDR False Discovery Rate

GO Gene Ontology

GCG Glucagon

GWAS Genome-wide Association Study

INS Insulin

IGT Impaired Glucose Tolerance

KEGG Kyoto Encyclopedia of Genes and Genomes

LDSC LD-score Regression

LD Linkage Disequilibrium

MODY Maturity-onset Diabetes of the Young

ND Non-Diabetic

QTL Quantitative Trait Loci

RFX6 Regulatory Factor X-6

RNA-seq RNA Sequencing

RUVSeq Remove Unwanted Variation from RNA-seq

sci-ATAC-seq Single-cell-combinatorial-indexing ATAC-seq

SNP Single Nucleotide Polymorphisms

shRNA Short-hairpin RNA

SST Somatostatin

T2D Type 2 Diabetes

T1D Type 1 Diabetes

TSS Transcription Start Site

WGCNA Weighted Gene Correlation Network Analysis

LIST OF SYMBOLS

β Beta cells

α Alpha cells

δ Delta cells

ϵ Epsilon cells

PP Gamma cells

ABSTRACT

Type 2 Diabetes (T2D) is a complex disease characterized by pancreatic β-cell dysfunction and dysregulation of blood glucose levels. Genome-wide association studies for diabetes and related traits suggest a complex genetic architecture of the disease and identify >400 independent signals throughout the genome. However, more than 90 percent of the signals map to the non-protein-coding regions suggesting a strong regulatory component to the disease. It is hypothesized that these non-coding variants affect disease susceptibility by modulating the transcription factor (TF) binding in a tissue- and context-specific manner. As such, understanding the genetic architecture of the disease involves a careful assessment of the complexity across all layers of the genetic organization. While existing studies have used high-throughput sequencing ('omics) approaches to dissect the disease at different layers, they have either been limited to a bulk-sample view or have focused on a specific layer (modality) — thereby limiting our ability to map biological mechanisms and the consequences of their dysregulation comprehensively. In this work, I utilize high-throughput molecular profiling data-driven approaches, "multiomics," in human pancreatic islets to characterize the tissue heterogeneity (complex interplay of cell types and their organization) and gene regulatory interactions (linking genetic variation to target genes and functions) to discover mechanistic insights relevant to the disease pathophysiology.

In chapter 1, I discuss the genetic architecture of T2D and emphasize how multiomic approaches driven by high-throughput sequencing technologies can help us link variants to genes and genes to their function. I emphasize the need of understanding the epigenomic landscape of different constituent cell types within the pancreatic islets, and how we can use that to complement our understanding of gene expression and regulation from transcriptomic and genetic studies.

In chapter 2, I use single-cell ATAC-seq to profile chromatin accessibility in pancreatic islets and identify molecular signatures unique to constituent cell types — one of the first published studies in this domain with a novel dataset. I show that major cell types can be easily identified from their epigenomic profiles and can be used to dissect genetic-risk associations across different cell types. We identify the pancreatic islet β cells to be the most enriched cell type for T2D genetic risk; and within each cell type, we use co-accessibility approaches to link variants to genes.

In chapter 3, I build upon the findings from the previous chapter, where we identify β cell chromatin accessibility peaks to be highly enriched for T2D genetic risk and investigate how β cell function is impacted in early-stage T2D. Using integrative approaches combining data from RNA-seq, ATAC-seq, secretion assays, imaging, and mRNA knockdown experiments, we discover Regulatory Factor X 6 (RFX6) as the key transcription factor implicated in dysregulation of insulin response in β cells.

Finally, in chapter 4, I discuss how my work establishes a framework for investigating complex diseases, where starting from genetic associations, which are non-specific and do not provide any mechanistic insight, we can integrate information across layers of the genetic organization using 'omics-driven approaches to build a mechanistic understanding in a stepwise manner. Applied to T2D, we show the strength of this approach and dissect the genetic heterogeneity to identify context-specific molecular signatures. Identification of such signatures will provide a higher-resolution map of our existing knowledge and enable the discovery of novel targets and approaches to prevent, monitor, and treat T2D.

CHAPTER 1

Introduction

Diabetes is a global epidemic that is rising at an alarming rate. As of 2021, 1 in 10 adults worldwide — over 500 million people — now live with the disease that is one of the top 10 causes of global mortality [115]. In 2021 alone, Diabetes and related health complications were responsible for 6.7 million deaths and nearly $966 billion in health expenditure — a 316% increase over the last 15 years [118]. It is estimated that the total number of cases with diabetes and diabetes-related health complications will exceed 600 billion by 2030, resulting in devastating medical and socioeconomic strain on the individuals, families, and societies [289].

Of all the cases of diabetes, the most prevalent diagnosis is the "common" non-insulin dependent or type 2 diabetes (T2D) accounting for >90% of all the individuals [236]. Unlike the monogenic forms of diabetes such as maturity-onset diabetes of the young (MODY) [85] which are caused by single-gene mutations [209], T2D is caused by a combined combinatorial effect of the genetic, environmental, and lifestyle factors, therefore making it hard to study, prevent, or treat (Figure 1.1).

While numerous studies have dissected the multifactorial pathogenesis of T2D at clinical and basal levels to discover individual risk factors and their contribution to the disease [101], the molecular and genetic characterization of the disease has been limited.

Family and twin-studies have shown T2D to have a strong genetic component with a positive family-history conferring a 2.4x increased risk of T2D [118]. Follow up by genome-wide association studies (GWAS) with over a million of people have further identified >400 independent

signals [166] associated with the disease highlighting the complex genetic etiology of the disease.

However, despite the compelling evidence of the genetic predisposition, identification of target genes and gene-product alternations for clinical management of the disease remains limited. Studies of the more extreme forms of diabetes — including maturity-onset diabetes of the young (MODY), mitochondrial diabetes with deafness, and neonatal diabetes — have identified single-gene Mendelian disorders [175]. However, expansion of our mechanistic insight beyond monogenic forms of the disease to more common form of diabetes has been difficult. This suggests that T2D is both polygenic and heterogeneous — i.e., multiple genes are involved across different tissues, and different combinations of genes play a role in different subsets of individuals. As a result, the exact role of the genes and their relative contributions to the disease risk remain uncertain.

The focus of these studies and the results presented in this work underscore the importance of investigating the cellular and molecular mechanisms that underlie the multifactorial pathogenesis of T2D and how they orchestrate the disease associated outcomes. Identification of such mechanisms not only will provide a higher-resolution map of our existing knowledge but also enable the discovery of novel targets and approaches to prevent, monitor, and treat T2D.

1.1 T2D is a disease of dysregulated glucose homeostasis

1.1.1 Pancreatic islet is a mini-organ critical for blood glucose control

T2D is a complex, heterogeneous disorder involving several key organs such as pancreas, liver, muscle, and fat [246]. However, central to the dysregulation that leads to hyperglycemia, the hallmark of T2D, are a cluster of hormone-producing cells called the *islets of Langerhans* (Figure 1.2) that are scattered throughout the pancreas. The pancreatic islets, together, comprise only about 1-2% (by weight) of the pancreas but perform coordinated functions to maintain glucose homeostasis by sensing blood glucose levels and regulating insulin action, production, and secretion.

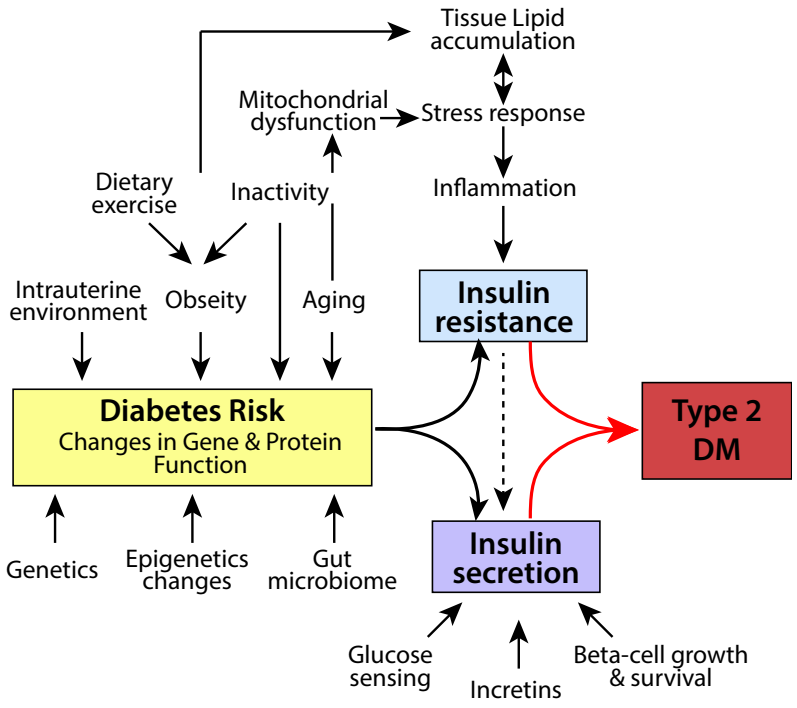


Figure 1.1: Complex Pathogenesis of Type 2 Diabetes. Mechanisms influencing the risk of type 2 diabetes. Reprinted from [72].

Pancreatic islets are a heterogeneous cluster of cells consisting of at least five cell types — α, β, δ, ε, and PP — that have been identified (See Figure 1.3). Moreover, these cell types are not distributed uniformly within the islet and their spatial distribution is known to change through age and development stage [24, 137]. In fact, changes to the islet morphology have been documented much prior to insulin's discovery in 1921 [68, 196].

In T2D, there are two interrelated problems at work that contribute uniquely to the clinical outcome of a patient. First is a secretory defect, where the pancreatic islets — specifically the β cells within the islet — do not produce enough insulin; and second is that the peripheral and target tissues that utilize insulin respond poorly to insulin and take up less sugar and develop resistance to insulin. For an individual, the disease etiology may range from predominantly insulin resistance with relative insulin deficiency to predominantly secretory defect with minimal insulin resistance.

At onset, as blood sugar levels increase, the insulin-producing β cells in the pancreas increase

in mass and release more insulin to compensate for insulin insensitivity. At this stage, most patients do not require insulin to survive. However, as type 2 diabetes manifests within an individual with persistent hyperglycemia, the β cells become impaired and may also reduce in mass resulting in insufficient supply of insulin to meet the body's demands, necessitating the need of clinical intervention to achieve optimal glucose control.

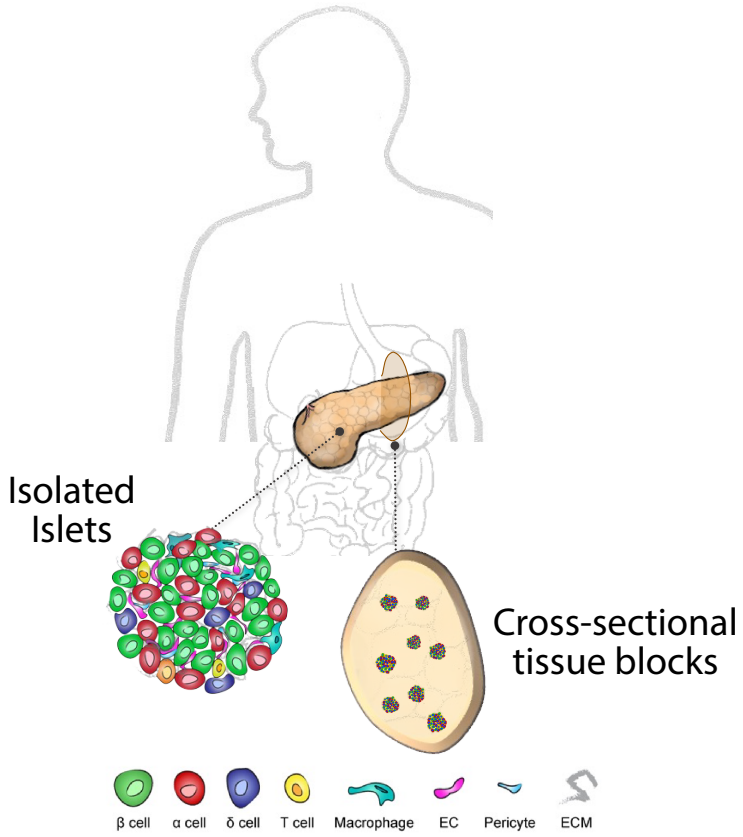


Figure 1.2: Pancreas is a mini-organ. Pancreas is the central organ for the glucose metabolism regulation. The *islets of Langerhans* are a cluster of different cell types scattered throughout the pancreas that produce hormones for glucose absorption and metabolism.

1.1.2 β cell dysfunction is a critical component in the pathogenesis of T2D

While insulin resistance is often the initiating defect in type 2 diabetic individuals, most people who develop the disease have a defect in the pancreatic β cell compensatory mechanism [126,

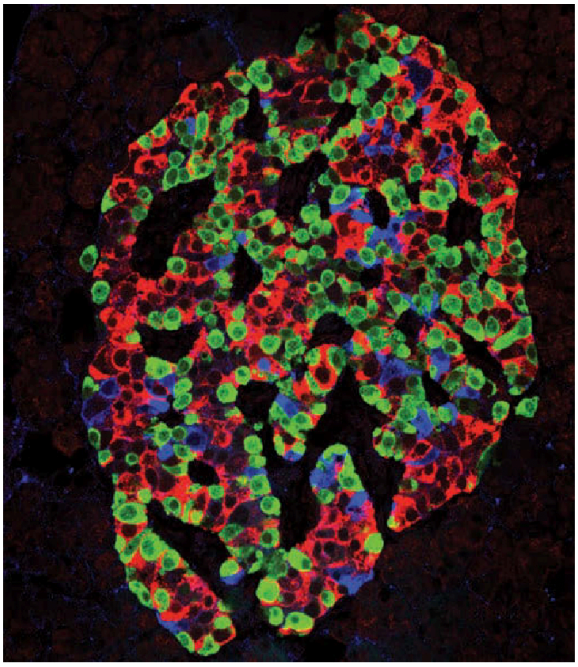


Figure 1.3: Human islet of Langerhans. Pancreatic section stained for insulin (red), glucagon (green), and somatostatin (blue). Reprinted from [34].

169]. Functionally, this defect manifests as reduction in total insulin secretion and maximal insulin response under glucose stimulation. While defects due to loss of β cell mass and its relative contribution has been debated in the literature [172], Recent studies of metabolically profiled donors suggested that β cell loss is not prominent in early T2D [50, 276].

Earlier studies that focused on candidate-gene testing or GWAS approaches to identify genomic loci associated with polygenic type 2 diabetes, found that most of the risk loci are involved in β cell function and turnover [261]. In fact, examples of single-gene defects discovered in patients with MODY have all been shown to produce a defect in glucose-induced insulin release.

These results reinforce the critical role of the β cell in controlling blood glucose and its involvement in the pathophysiology of T2D. Therefore, our hypothesis is that the β cell dysfunction occurs early in T2D and that prevention and/or rapid intervention may be critical to preserve β cell function and treat T2D.

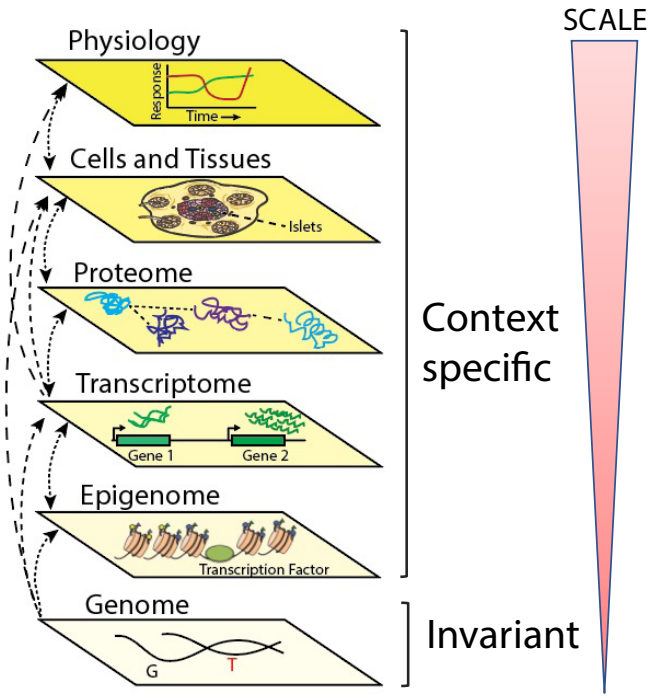


Figure 1.4: Diabetes is a culmination of dysregulation across layers of genetic organization. Genetic architecture of a complex disease such as T2D consists of distinct layers of molecular organization. Information from the bottom most layer, the genome, which is largely invariant within an individual propagates through context-specific regulatory layers to express as the hallmark disease phenotype. Adapted from [48].

1.2 T2D risk spans layers of genetic organization

1.2.1 Genetic architecture of T2D

Genome-wide association studies (GWAS), where researchers look to see if single-nucleotide polymorphisms (SNPs) are associated with a phenotype or not, have identified hundreds of loci associated with T2D and related traits across the genome. A recent such study in populations of European descent identified >400 distinct T2D association signals at 243 loci that explain ~20% of the trait variance and ~50% of the heritability of T2D [166]. Additionally, parallel GWAS in populations of East Asian, South Asian [144], Hispanic/Latino [278], and African ethnicities, and trans-ethnic meta-analyses [165, 267] have identified >100 additional loci.

While GWAS allow us to scan the genome and identify disease-associated regions without guessing where to look first, the untargeted nature also poses significant challenges in interpre-

tation. First, the reported association signals do not implicate the specific DNA sequences that cause the molecular phenotype (*causal variants*) and the precise effector transcripts involved in increasing T2D risk (*causal genes*). Instead, they merely flag regions of the genome that are over-represented in cases (people with our desired trait, T2D for example) versus controls (people without the trait, non-diabetic individuals for example). Second, due to linkage disequilibrium, the lead variants reported for each association signal — the nucleotide changes associated with the trait of the interest — may not necessarily be the nucleotides that alter gene function or regulation in a relevant cell type to induce disease. Finally, most of the reported associations have small effect sizes on the risk and explain only a small fraction of the expected genetic component of risk [168]. As such, identification of causal genetic variants, their effector genes, and the relevant cell types underlying the observed associations is a significant challenge for T2D.

1.2.2 Regulatory complexity of T2D

While genomic DNA is essentially invariant across cells and tissues within an individual, the information from genomic DNA propagating through other molecular domains is highly specific and dynamic. This regulation is at the core of distinct functional identities for all the cells in our bodies — the mechanism for which are encoded within the genome itself. The genomic regions that regulate the gene expression are called *regulatory elements* and consist of specific nucleotide sequences that facilitate downstream action through transcription factors and enzymes. Approaches to identify and characterize such regulatory elements and their downstream effects is a growing area of research in functional genomics studies.

Using functional genomics approaches guided by GWAS, recent studies have shown that T2D associated variants are significantly enriched to overlap chromatin-defined transcriptional regulatory elements like stretch enhancers or enhancer clusters that are highly tissue- and context-specific [114, 201, 264]. For instance, T2D GWAS loci are specifically enriched in pancreatic islet Regulatory Factor X (RFX) footprints [264], and T2D GWAS variants were shown to overlap skeletal-muscle-specific regulatory enhancers [235]. Because a majority of the disease associated

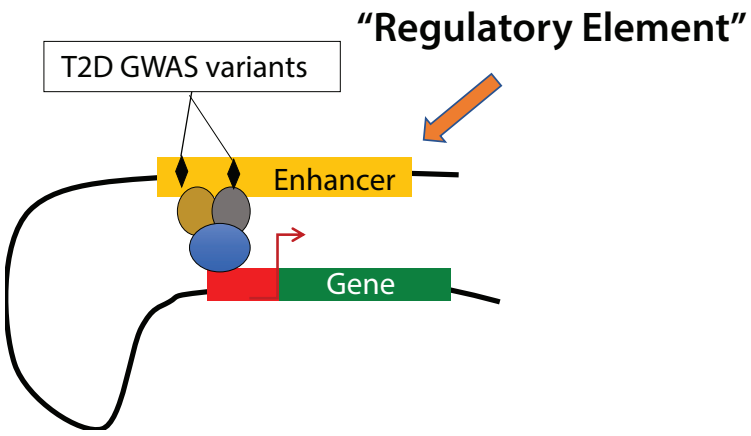


Figure 1.5: Schematic showing likely mechanism of gene regulation. Most variants likely regulate gene expression rather than directly altering protein function as shown in the schematic.

variants do not disrupt protein function directly, these studies reinforce that the non-protein-coding variants play a primary role in modulating genetic risk for T2D through gene expression regulation, many by altering regulatory elements, which also can be tissue- and context-dependent (Figure 1.5).

However, mapping the effects of non-coding variation is less straightforward than identification of coding variants that directly point to an effector transcript. Because of the large number of such variants, non-coding variants are presumed to have smaller effect sizes, and thus their consequences are much harder to map and identify.

Nonetheless, while it may be difficult to directly identify causal non-coding variants and their downstream effector transcripts through GWAS alone, mapping the consequence of this variation through the molecular layers (Figure 1.4) can inform us of underlying mechanisms that propagate this information from genotype to the clinical phenotype.

For example, the epigenomic layer which consists of molecular modifications to the DNA that affect its chromatin structure and DNA accessibility can alter the binding of transcription factors. This altered binding, when performed in a cell- and context-specific manner provides a unique mechanism of gene regulation. Histone modifications and methylation are two common processes used by cells to dynamically alter the expression landscape of a cell and create a unique epigenomic profile.

Similarly, profiling gene expression has clarified the impact of non-coding variation through identification of disease-linked variants that alter RNA-splicing, RNA stability, or interfering with long non-coding RNA (lncRNA) integrity. Further, studies have combined gene expression and genotyping data to identify variants that are statistically associated with the expression of a gene in a population. Such studies have been conducted for many tissues, to discover tissue-specific quantitative trait loci (QTLs) and across many populations as well [84, 158, 235, 262, 264], enabling direct variant to target gene assignment.

Transcriptomic and epigenomic profiling therefore are necessary for dissecting the function of non-coding variation. Further, integrating these datasets together can provide us concordant maps of changes in the two layers and help us identify disease causing variants and their molecular mechanisms.

1.3 Single-cell technologies enable high-resolution maps

Bulk profiling assays are unable to capture the cellular heterogeneity of the disease-relevant tissues and are dominated by the most common cell types within the tissue sample. For example, within pancreas, exocrine cells are the majority and will contribute overwhelmingly to any readout masking the changes in low-abundance cell types. While bulk analysis of sorted cell types is possible, such approaches are necessarily biased due to limitations in our ability to identify cell-specific markers unique to a cell population, and negatively impact the assay quality [198]. Since single-cell approaches provide a largely unbiased, data-driven approach to identify cell types and capture cellular heterogeneity, they have gained huge popularity. While gene-expression measurement continues to be the dominant readout, single-cell resolution versions of many common molecular assays have been developed — sci-ATAC-seq [55], scATAC-seq [286] for chromatin accessibility profiling; scRNA-seq for gene expression; sci-Hi-C for chromatin architecture profiling [187]; and CUT&Tag for protein-DNA interactions [111]. Depending on the assay, a molecular readout may be obtained from the whole-cell fraction or only nuclei — differentiating single-cell

and single-nuclear assays. However, systematic comparison of both types of dataset have established numerous practical advantages of single-nuclear assays [69, 280] and many assays for RNA-seq and ATAC-seq have been adapted to work with isolated nuclei. At the same time, parallel efforts to develop computational tools and methods to tackle specific challenges of single-cell data have also continued to grow [154, 250]. As such, our ability to generate multiomic single-cell resolution maps of relevant cells and tissues in specific contexts could provide unprecedented opportunity to study mechanisms underlying disease development.

1.3.1 Single-cell maps of chromatin accessibility for epigenomic profiling

Single-nucleus Assay for Transposase Accessible Chromatin by Sequencing (snATAC-seq) is a powerful technique to study the chromatin accessibility landscape and gene regulation in single cells, highlighting context-specific biology [27, 49, 54]. By mapping chromatin regulatory landscape at a single-cell resolution, researchers have demonstrated potential to discover complex cell-populations, link regulatory elements to the target genes, and map regulatory dynamics during complex cellular differentiation processes [26, 56, 97, 231]. The snATAC-seq approach enables this innovation by first isolating cells at a limiting dilution and then delivering the Tn5-transposase enzyme with necessary buffers and unique-sequence barcodes to carry out the reaction. The presence of unique barcodes then allows one to uniquely identify each cell or nuclei at the data processing stage. Currently, the most dominant approach to snATAC-seq profiling is based on the partitioning of samples and reagents into droplets, each called a Gel Bead in Emulsion (GEM) [286]. However, alternative approaches such as sci-ATAC-seq [54], as used in chapter 2, do exist which use combinatorial indexing to uniquely identify each cell or nuclei. Both approaches have unique advantages and disadvantages, but overall, benefit from the increased resolution of the data.

1.3.2 Joint single-cell RNA and ATAC profiling

Most single-cell studies focus on profiling one modality or molecular layer at a time and thus provide a limited view of the cell state. Jointly profiling the gene regulation traits, principally by the single-cell RNA and ATAC-seq assays, would provide the first readout of the activity of genetic variants and their activity in a cell and context-specific manner. Therefore, allowing us to map simultaneous changes between the molecular traits and establish the sequence of disease-inducing changes within the cells. Several studies that have profiled two distinct molecular modalities on single cells — methylation and chromatin [158], transcriptome and histone modification [288], and chromatin accessibility and gene expression [37] for example — have demonstrated a powerful utility of single-cell co-assays over assays that solely profile a single molecular trait. Recently, commercial availability of droplet gel-bead based protocol for joint single-nuclei RNA and ATAC from 10X Genomics has a high-quality approach to generate these single-cell resolution datasets. In fact, a recent study utilized the joint profiling approach to investigate cellular heterogeneity, identify causal cell types and regulatory elements in the human and rat skeletal muscle in the context of T2D [198], suggesting the high value of using such an approach to create a high-quality multi-modal map at a single-cell resolution in the pancreatic islets.

1.4 Multiomic integration to dissect genetic architecture

With different type of molecular profiles available for each layer in the genetic architecture of the disease (see Figure 1.4), and often at single-cell resolution thanks to new technologies, integration of the distinct data modalities can yield powerful insights about the mechanisms driving disease associations and enable biomarker discovery. In a recent report [276], authors performed multi-omics analysis of transcriptomic and proteomic profiles from bulk pancreatic islets obtained from metabolically profiled pancreatectomized living human donors along the glycemic continuum of T2D. Thus, by building an integrative model that allows us to assess the relative importance of each molecular layer and consequences of its disruption in the disease pathophysiology, we can

gain a more holistic view of the system and draw conclusions regarding key pathways, targets and biomarkers, and therapies in metabolic and other diseases.

1.5 Thesis outline

The two main chapters described in this thesis present the body of work published in references [217] and [268]. Additional work highlighting applications of ATAC-seq to liver, another tissue central to glucose metabolism and type 2 diabetes, is described in appendix section A.

Broadly, in chapter 2, I discuss the findings from the first single-cell epigenomic profiling study in pancreatic islets. We show that constituent cell types can be inferred using their unique epigenomic molecular signatures and can help us partition the disease risk or heritability across cell types. This increased resolution allows us to build more sophisticated models of disease risk mechanisms and link variants to genes and genes to function. We also demonstrate the novel application of a deep-learning based approach to impute data from sparse or low-abundance cell populations that will be of immense significance in studies where it is difficult to obtain a large amount or a high quality of data.

In chapter 3, we build on understanding from previous chapter and strive to understand the differences in the molecular signatures between normal and early-stage type 2 diabetic individuals. We profile function and physiology and tissue architecture of pancreatic islets along with transcriptomic profiling of sorted α and β cell populations. Using integrative analysis approaches, we define functional modules within β cell that are associated with disease relevant traits and enriched for T2D GWAS genetic variants. We then identify key genes across these modules that are dysregulated in early-stage T2D and perform knockdown experiment in pseudo-islets to validate our hypothesis.

Finally, in chapter 4, I conclude our findings and discuss the broader contributions made to the current scientific understanding of diabetes and how we can use these approaches in different disease context. I then discuss strategies and options to build upon this work that could contribute

further to our understanding of complex disease genetics and improved healthcare outcomes for patients in the future.

CHAPTER 2

Single Cell ATAC-seq in Human Pancreatic Islets and Deep Learning Upscaling of Rare Cells Reveals Cell-specific Type 2 Diabetes Regulatory Signatures

2.1 Abstract

Type 2 diabetes (T2D) is a complex disease characterized by pancreatic islet dysfunction, insulin resistance, and disruption of blood glucose levels. Genome wide association studies (GWAS) have identified 400 independent signals that encode genetic predisposition. More than 90% of the associated single nucleotide polymorphisms (SNPs) localize to non-coding regions and are enriched in chromatin-defined islet enhancer elements, indicating a strong transcriptional regulatory component to disease susceptibility. Pancreatic islets are a mixture of cell types that express distinct hormonal programs, and therefore each cell type may contribute differentially to the underlying regulatory processes that modulate T2D-associated transcriptional circuits. Existing chromatin profiling methods such as ATAC-seq and DNase-seq, applied to islets in bulk, produce aggregate profiles that mask important cellular and regulatory heterogeneity.

We present genome-wide single cell chromatin accessibility profiles in 1,600 cells derived from a human pancreatic islet sample using single-cell-combinatorial-indexing ATAC-seq (sci-ATAC-

seq). We also developed a deep learning model based on the U-Net architecture to accurately predict open chromatin peak calls in rare cell populations.

We show that sci-ATAC-seq profiles allow us to deconvolve α, β, and δ cell populations and identify cell-type-specific regulatory signatures underlying T2D. Particularly, we find that T2D GWAS SNPs are significantly enriched in β cell-specific and cross cell-type shared islet open chromatin, but not in α or δ cell-specific open chromatin. We also demonstrate, using less abundant δ cells, that deep-learning models can improve signal recovery and feature reconstruction of rarer cell populations. Finally, we use co-accessibility measures to nominate the cell-specific target genes at 104 non-coding T2D GWAS signals.

Collectively, we identify the islet cell type of action across genetic signals of T2D predisposition and provide higher-resolution mechanistic insights into genetically encoded risk pathways.

2.2 Introduction

Pancreatic islets consist of a cluster of at least five different endocrine cell-types (α, β, δ, γ, and ϵ), each producing a unique hormone in a distinct but coordinated manner [66]. Collectively, these clusters of cells work together to maintain insulin production and glucose homeostasis. Disruption of the complex interplay between the cell types, their organization, and their underlying regulatory interaction is known to be associated with type-2-diabetes (T2D) pathophysiology [243]. However, the exact cellular mechanisms through which different risk factors contribute to the disease risk are not completely understood. Using GWAS and eQTL mapping approaches, recent studies have discovered >400 independent signals (>240 loci) associated with T2D and T2D-associated traits [166], although remarkably, more than 90% of them localize to non-protein-coding regions of the genome [266]. Growing evidence suggests that many of these variants likely influence the RNA expression and cellular function of human pancreatic islets by altering transcription factor binding, critical components of a cellular regulatory network [31, 201, 202, 257, 264].

High-throughput epigenomic profiling methods such as ATAC-seq [28] and DNase-seq [112] have enabled profiling of chromatin accessibility across samples in a tissue-wide manner, providing the opportunity to identify millions of context-specific regulatory elements. However, these bulk-measurements of chromatin accessibility limit the precise understanding of how tissue heterogeneity and multiple cell-types in the population contribute to overall disease etiology [239]. Recent advances in single-cell transcriptomic and epigenomic profiling methods have enabled an unbiased identification of cell-type populations and regulatory elements in a heterogeneous biological sample. By mapping the chromatin-regulatory landscape at a single-cell resolution, recent single-nuclei studies have demonstrated the potential to discover complex cell populations, link regulatory elements to their target genes, and map regulatory dynamics during complex cellular differentiation processes [54, 56, 210, 231]. The pancreatic islet gene expression landscape has been investigated at single-cell resolution in existing studies [152, 182], but chromatin accessibility studies have been limited to fluorescence-activated cell sorting (FACS) methods for obtaining cell-type populations [2, 8]. FACS based methods will miss identification of unknown or rarer cell-populations and are unable to produce pure cell-type populations due to reliance on the specificity of cell-surface markers [74, 190].

Here, we present a genome-wide map of chromatin accessibility in >1,600 nuclei derived from a human pancreatic islet sample using single-nucleus-combinatorial-indexing ATAC-seq (sci-ATAC-seq) [55]. sci-ATAC-seq enables us to deconvolve cell populations and identify cell-type-specific regulatory signatures underlying T2D. Notably, we find that T2D GWAS SNPs are significantly enriched in β cell-specific and cross cell-type shared islet open chromatin, but not in α or δ cell-specific open chromatin. We also demonstrate, using the less represented δ cell population (< 5% of total islet population), that deep learning can improve signal recovery and feature reconstruction for less abundant cell-populations using concepts borrowed from image upscaling methods. We anticipate that our deep learning method will enable analysis of heterogeneous tissues that may be harder to obtain in large numbers or contain rare sub-populations. Collectively, these results identify the islet cell-type of action across genetic signals of T2D predisposition and

provide higher-resolution mechanistic insights into genetically encoded pathophysiology.

2.3 Results

2.3.1 sci-ATAC-seq captures tissue relevant characteristics similar to bulk ATAC-seq

Pancreatic islets represent approximately 1-2% (by mass) of total pancreatic tissue [66] and therefore requires specialized approaches to isolate in a manner that maintains viability. We obtained a highly pure (>95% purity and >92% viability) sample of human pancreatic islet tissue from one individual (cadaveric donor, female, 43 years old, and non-diabetic) and profiled chromatin-accessibility using sci-ATAC-seq protocol [55] as described previously (Figure 2.1A). In total, we obtained 1,690 single-cell ATAC-seq datasets with depth ranging from 17,667 to 415,237 (median: 79,482) reads per nucleus, and TSS enrichment from 0.77 to 9.80 (median: 3.91) after removing background barcodes (Figure 2.2A). For quality assessment of each single nucleus, we reasoned that total reads and TSS enrichment values are more suitable metrics for identifying nuclei with poor signal-to-noise ratio than using fraction of reads in peaks as the latter may bias counts for under-represented cell-type populations in the analysis (Figure 2.2B-C). Based on these criteria, we obtained high-quality sci-ATAC-seq data for 1,456 single-nuclei. In addition to sci-ATAC-seq data, we generated high-quality bulk ATAC-seq data for ten islet samples with >47 M reads and >4.4 TSS enrichment per sample. Using our approach to identify high-confidence (master) peak calls across samples (see methods), we obtained 106,460 bulk islet accessible chromatin peaks.

We then compared the aggregate islet sci-ATAC-seq data with bulk ATAC-seq samples from islets and other tissues. For this, we called 156,311 peaks on the aggregate sci-ATAC-seq. We found that aggregate sci-ATAC-seq profiles were concordant and clustered together with the other bulk islet samples indicating that aggregate sci-ATAC-seq can capture chromatin accessibility in a manner equivalent to bulk ATAC-seq assays (Figure 2.1B-C, Figure 2.2D-E). Further, to understand if the aggregate sci-ATAC-seq peaks capture islet-specific regulatory features, we

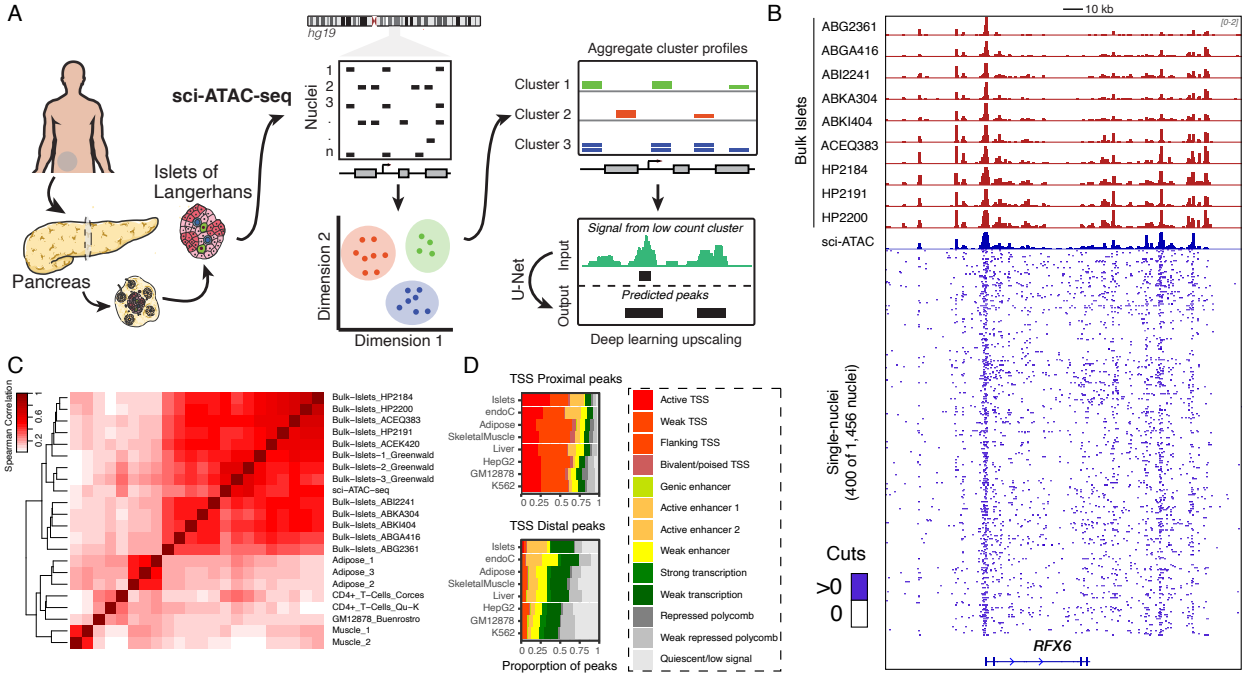


Figure 2.1: Schematic of sci-ATAC-seq study. (A) sci-ATAC-seq protocol for generating single-nuclei ATAC-seq data from a pancreatic islet sample. The data is then used to identify constituent cell-types and use deep-learning model to predict peaks on the clusters with fewer nuclei count. (B) ATAC-seq signal tracks for 10 bulk islet samples and sci-ATAC-seq islet sample. Bottom tracks show the signal across a random subset of up to 400 single-nuclei. Signal tracks are normalized to one million reads and scaled between 0-2. (C) Spearman correlation between aggregate sci-ATAC-seq, 13 bulk islets, 3 adipose, 2 muscle, 2 CD4+ T-cells, and 1 GM12878 sample. (D) Distribution of aggregate sci-ATAC-seq TSS proximal and distal peaks across bulk islet derived ChromHMM segmentations.

compared the distribution of peaks across chromHMM chromatin state maps in eight tissues, including islets and the EndoC human β cell line [264]. We found that islet sci-ATAC-seq peaks overlap active TSS and active enhancer segmentations in islet and EndoC (a β cell line) chromatin state maps to a larger extent compared to other tissues (Figure 2.1D). Because chromHMM enhancer states are driven by H3K27ac marks and are known to be associated with tissue-specific enhancer activity [114, 201], our results indicate that sci-ATAC-seq data capture the underlying islet-specific chromatin architecture similarly to bulk islet ATAC-seq assays. Overall, these results indicate that our aggregate islet sci-ATAC-seq data are of high quality and suggests that the underlying individual nuclei could reveal valuable cell-specific patterns of the constituent cell-types.

2.3.2 sci-ATAC-seq reveals constituent cell-types in pancreatic islets

The aggregate sci-ATAC-seq profile of the islet is constituted of signal from distinct cell-types. For identifying these cell-types, we leveraged the observation that TSS distal regions capture cell-type-specific accessibility patterns and are effective at classifying constituent cell-types [29]. We adopted a multi-step process to robustly detect and identify islet subpopulations (see methods). This approach produced four distinct clusters (Figure 2.3A). In order to assign a cell-type identity to the clusters, we merged nuclei in each cluster to create aggregate chromatin accessibility profiles and systematically examined the patterns of accessibility at multiple cell-type marker loci. We found three clusters to have distinct chromatin-accessibility patterns at *GCG*, *INS-IGF2*, and *SST* genes corresponding to three major islet cell-types: α, β, and δ cells (Figure 2.3B). The fourth cluster (95 nuclei, ~7% of all nuclei) showed a "mixed" cell-type appearance as shown by signal at multiple cell-specific markers. We reasoned that these are likely to be nuclei doublets resulting from barcode collisions inherent to the combinatorial indexing protocol, and thus should have skewed ATAC-seq read coverage. Indeed, we observed that nuclei assigned to the mixed cell cluster were significantly (nominal P-value = 7.3e-7, binomial test) enriched in the high sequencing depth bin relative to nuclei from other clusters (Figure 2.3C). As such, these nuclei were removed

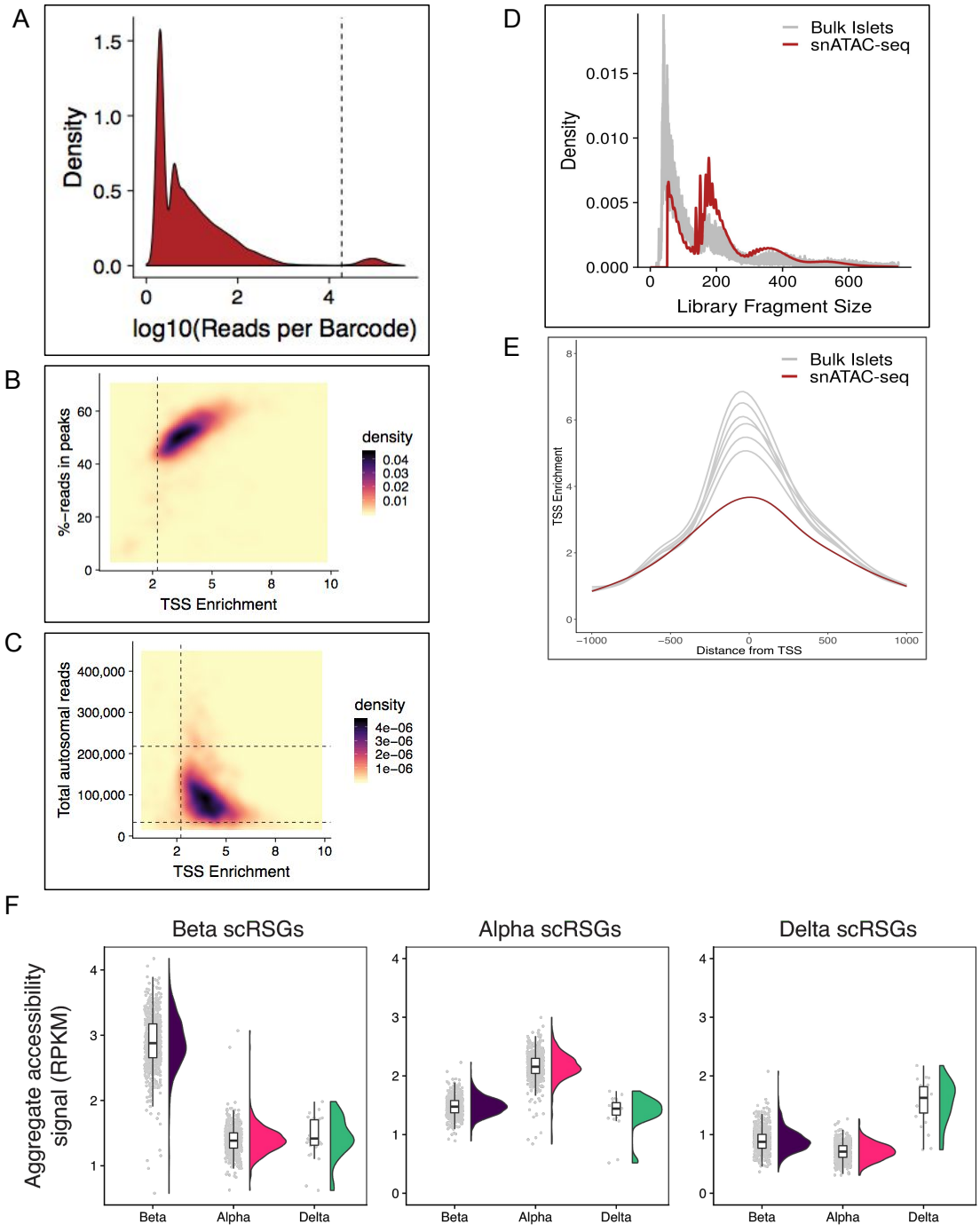


Figure 2.2: ATAC-seq metrics of nuclei from sci-ATAC-seq. (A) Distribution of reads per barcodes shown with the threshold chosen for filtering background barcodes. (B) Fraction of reads in peaks versus TSS Enrichment, and (C) Total autosomal reads versus TSS enrichment for all single-nuclei. Density units are arbitrary. (D) TSS coverage of aggregate sci-ATAC-seq, and (E) Fragment length distribution of aggregate sci-ATAC-seq compared with ten bulk islet ATAC-seq samples. (F) Chromatin accessibility signal in single-cell RNA-seq derived cell-type signature genes (scRSGs; β=83, α=168, δ=53) across three sci-ATAC-seq identified cell clusters. scRSGS obtained from [237].

from further analyses yielding a total of 1,361 nuclei with 51%, 47%, and 2% assigned to β, α, and δ cell-type respectively. These estimates agree with the existing estimates of pancreatic islet cell-type proportions observed in confocal microscopy or single-cell transcriptomics experiments [24, 34, 151, 237]. As additional validation of our cell-type assignments, we used cell-type signature genes from a published islet scRNA-seq study [151, 237] and observed cluster-specific chromatin accessibility consistent with our assigned cell identities (Figure 2.3D-E).

We then analyzed the chromatin accessibility profile for each cell-type cluster. For this, we aggregated nuclei within each cluster and identified peaks using MACS2. We identified 129,046 sites for α and 120,116 sites for β cells. However, because the δ cluster had only 28 cells (corresponding to ~2 M reads), we reasoned that MACS2 would not perform ideally on data with such low depth. Indeed, we only identified 49,293 peaks using MACS2 on the δ cell aggregate reads.

2.3.3 Deep learning enables robust peak calls on less abundant δ cells

To solve the challenge of learning cell-type-specific features from the sparse signal in the low-count δ cell cluster, we developed a novel a deep learning approach based on the U-Net architecture (Figure 2.4A). U-Net was first developed for biomedical image segmentation but has since been applied to many other problems including audio and super resolution images. Its use in the super resolution problem served as the main impetus for our choice of model to upscale genomic signals. We formulated our approach as a classification problem in which we used sparse signal and corresponding peak calls to predict dense and high-quality peak calls. To avoid overfitting and create a robust, generalizable model, we adopted a rigorous training scheme. We divided the chromosomes into training, validation, and testing sets (Figure 2.5A, Figure 2.4B) and tested the performance of the models within the same cell type as well as across different cell types. We reasoned that our islet sci-ATAC-seq data were an ideal fit for this problem as all the nuclei came from the same individual and processing batch and should, therefore, contain no genetic or technical biases that would influence within or across cell-type predictions. Because we had high-quality data from two cell types, we trained two models: one model was trained using 28-

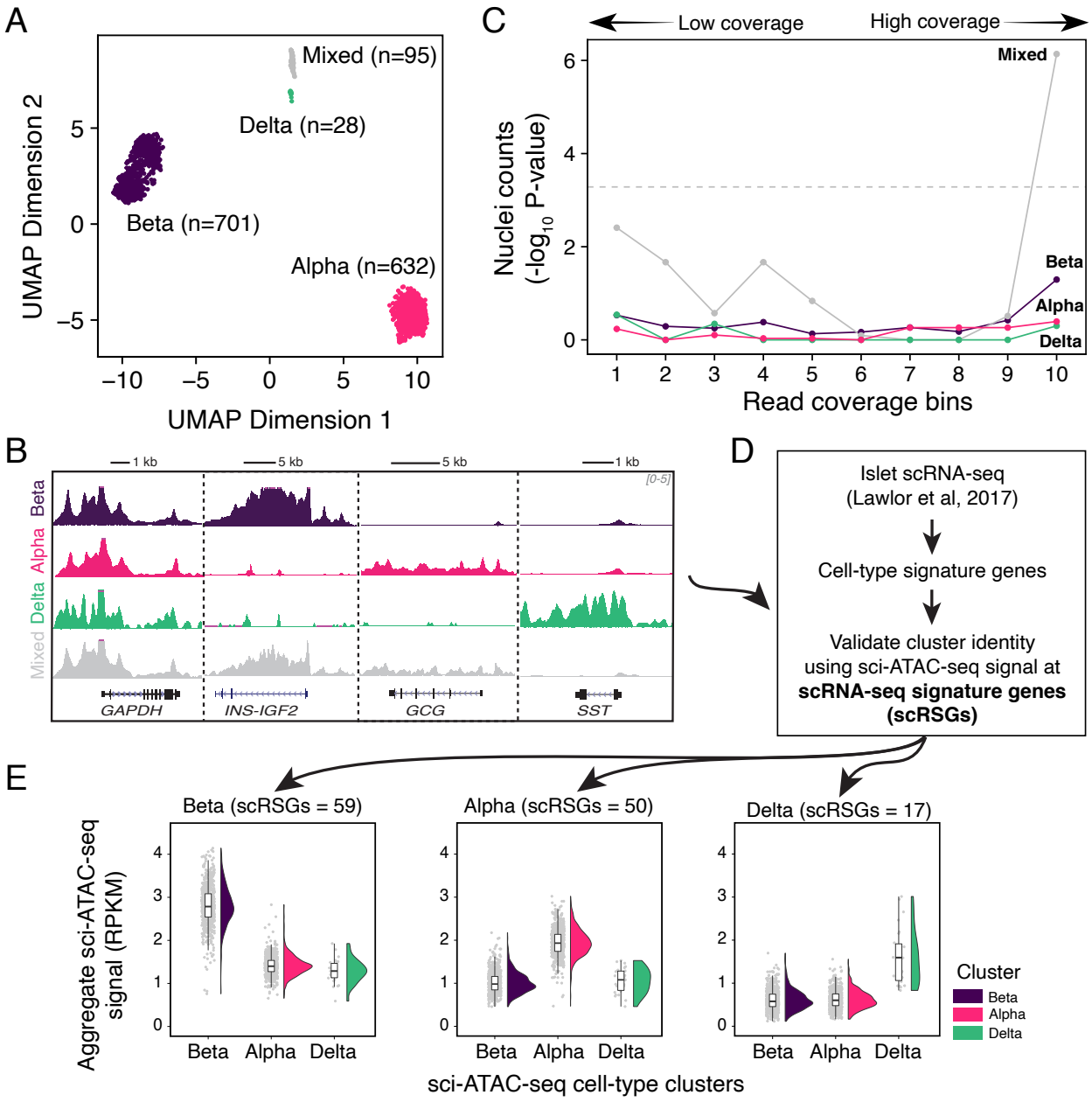


Figure 2.3: Clustering and identification of cell-type clusters in sci-ATAC-seq data. (A) UMAP projection with clustering of 1,456 single-nuclei islets represented by each single point into four clusters as identified by density-based clustering. (B) Enrichment of cells from each cluster relative to their expected population proportion across different read sequencing depth bins. Sequencing depth increases with the bin number. (C) Genome browser tracks showing signal at different cell-type marker loci: α (*GCG*), β (*INS-IGF2*), δ (*SST*), and a housekeeping gene (*GAPDH*). Tracks are normalized to one million reads and scaled between 0-5. (D) Overview of independent cluster verification scheme utilizing cell-type signature genes as identified by an islet scRNA-seq study by [152]. (E) Plot of aggregate ATAC-seq signal (normalized using RPKM) at scRNA-seq derived cell-type signature genes for α, β, and δ cells. Number of signature genes for each cell type indicated in the legend.

cell and 600-cell data from the α cells (α-trained model), while the second model was trained similarly on the data from the β cells (β-trained model).

We then compared peak predictions from both models to corresponding MACS2 peaks from the 600-cell data. We found that the results from across cell-type predictions of both models outperformed the MACS2 peak results as measured by the mean average precision (Figure 2.5B), suggesting that the U-Net model was able to reconstruct peak calls from sparse signals independent of the specific cell type it was trained on. We highlight several examples in which the model was able to successfully predict peaks that were absent in the sparse 28-cell data but present in the 600-cell data of a cell type (Figure 2.5C). Because the training cell type had no signal or peak at the given locus, these predictions could not have been transferred or "copied over" from the training data, indicating a possible use across cell types or tissues. Based on these results, we used the U-Net models to predict peaks for the low-count δ cell cluster. As the U-Net model provides a posterior probability score for each peak call prediction, we sought to create a high-confidence set of predicted peak calls for each cell type. We used a threshold of 0.625 to filter the predicted peaks for each cell type. The choice of threshold was used to control for potential false positives and the final number of predicted cell-type peaks (Figure 2.4C-D). Further, considering that the δ peak predictions from both the α and β models were highly concordant (Jaccard index of 0.85), we used the intersection of the results as the final predicted outcome.

We then validated our peak predictions using an orthogonal strategy, where we computed the enrichment of scRNA-seq derived signature genes (scRSGs) for the α, β, and δ cells across chromatin accessibility peaks. We found that the scRSGs for each cell type consistently had higher enrichment in the predicted peaks than the MACS2 peaks derived from the same 28-cell data (Figure 2.6A), indicating that our predicted peaks captured cell-type specificity. In the next step, we compared them to the bulk islet ATAC-seq master peak calls. We found that master peaks derived from the bulk islets were highly reproducible across samples, with >70% of the peaks occurring in five or more of the 10 samples (Figure 2.6B). Predictably, we also observed that the chromatin states corresponding to "active TSS" and "active enhancer" showed enrichment with increasing

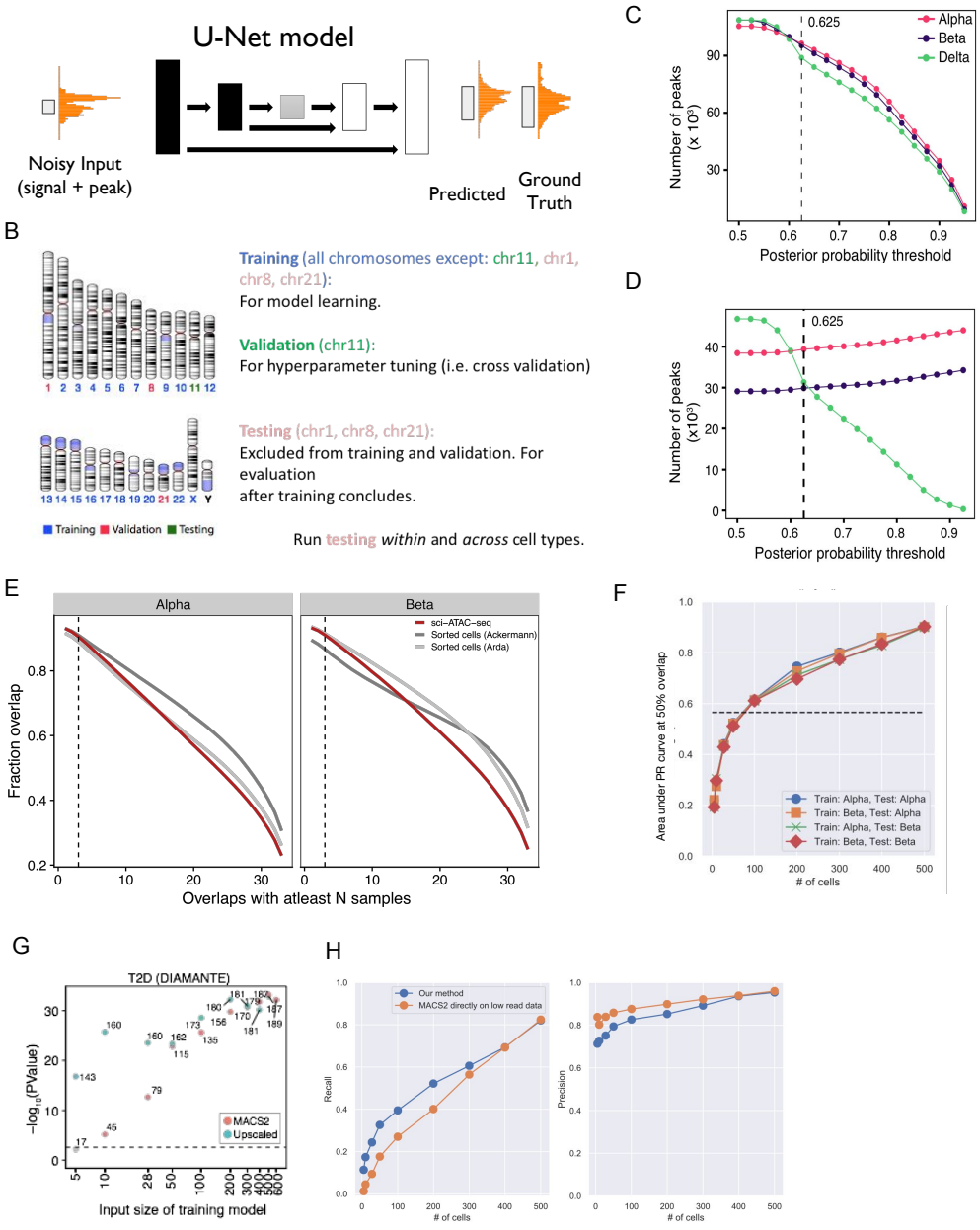


Figure 2.4: Peak calling using deep learning approach. (A) Schematic of U-Net learning strategy. (B) The training, testing, and validation scheme used for training the models delineating which chromosomes were part of what dataset. (C) Number of predicted peaks (from 28-cell trained model) for each cell type with different output posterior probability thresholds. (D) Number of cell-type specific peaks for α, β, and δ after partitioning into mutually exclusive sets (see methods) with different output posterior probability thresholds. (E) Fraction overlap of cell-type peaks (α, β) from our study and sorted cell populations from [2, 8] with different sets of reproducible bulk islet ATAC-seq peaks obtained from 33 bulk islet ATAC-seq samples. (F) Average precision in predicting peaks compared for all four models (two training and two prediction datasets) with different sizes of input training data. (G) Enrichment of T2D GWAS SNPs (N=378) in predicted beta peak calls (from α-trained model) compared with peaks calls from MACS2 on the data with varying size of input training data. (H) Precision and recall curves comparing predicted beta peaks (from α-trained model) for varying size of input training data.

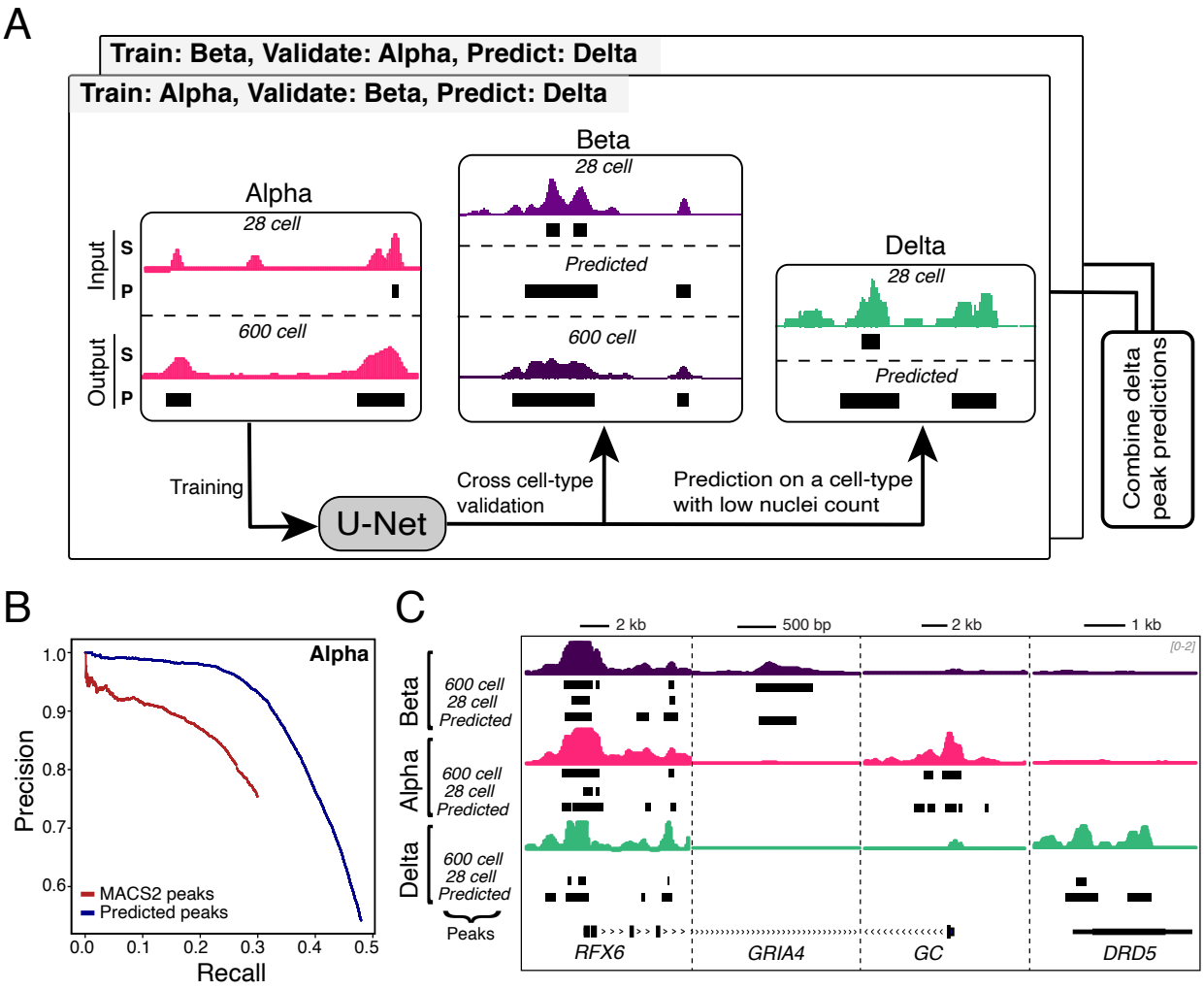


Figure 2.5: Deep learning upscaling from sparse low-count nuclei clusters. (A) Schematic of U-Net training scheme. Two models are depicted in the illustration: one trained on α cells data as input and other trained on β cells as input. δ cell peak predictions from both models are combined to get final predictions (see Methods). (B) Precision-recall curve comparing peak calls from MACS2 on downscaled data (α cell-type) with predicted peak calls from 28 cell U-net model (trained on β, predicted on α). (C) Example loci illustrating peak upscaling with the model. For each cell-type, four tracks are shown: full signal track, peak calls on full data, peak calls on subsampled data, and predicted peak calls. The predicted peak calls are obtained from a model trained on a different cell-type. For δ predicted peak calls, intersection of prediction from both α and β models are shown. Signal tracks normalized to one million reads and scaled between 0-2.

reproducibility of the master peaks. Like- wise, chromatin states such as "repressed polycomb," "weak transcription," and "quiescent/low signal" showed a depletion with the increasing islet ATAC-seq peak reproducibility (Figure 2.6C). Similarly, when we compared the cell-type peaks to the master peaks, we found that the proportion of peaks from each cell type increased with the increasing reproducibility of the bulk peaks (Figure 2.6D), suggesting that highly reproducible peaks were driven by all the constituent cell types while the peaks that occurred in fewer samples might have originated from underlying cell-population variability. For further validation, we also compared our cell-type peaks and sorted cell-type population peaks [2, 8] with master peaks derived from 33 independent bulk islet ATAC-seq samples and observed a high degree of concordance (Figure 2.4E). For example, >90% of the α and β peaks were reproducible across three or more ATAC-seq samples, which was comparable to the 85-92% peak overlap observed for α and β cell-type peaks from sorted cell populations in previous studies [2, 8]. While the primary model of our interest was trained using data from 28 cells to predict 600-cell equivalent peaks, we asked if the model would perform similarly for a varying resolution of input data. To accomplish this, we subsampled cells from α and β cell clusters to sets of different cell counts, starting with as few as five cells to 500 cells. We found that the performance of the model increased with the increasing number of cells used in the input training data (Figure 2.4F). There was up to a five-fold gain in the coverage of the T2D GWAS SNPs in the β predicted peaks compared to the MACS2 peaks (Figure 2.4G) even when fewer cells were used as input training data (Figure 2.4H). These results suggest that the deep learning strategy is applicable to a range of input data typically seen in single-cell sequencing experiments.

Overall, our results show that deep learning driven feature prediction can help recover tissue and cell-type relevant chromatin accessibility patterns from sparse and noisy data. Using this approach can enhance biological discoveries, which is challenging with rare cell populations.

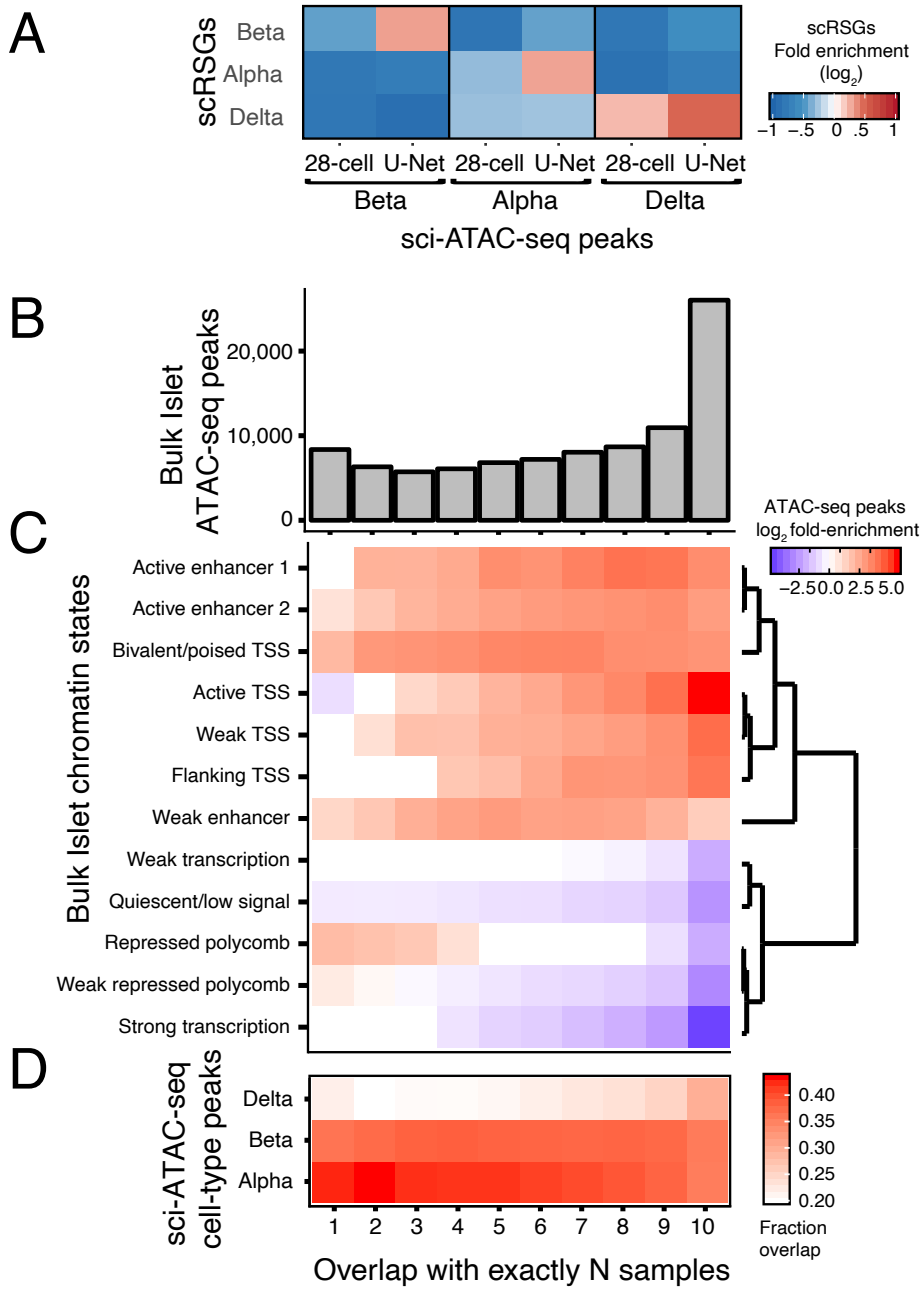


Figure 2.6: Upscaled predicted peaks are enriched for cell type specific signatures. (A) Reproducibility of master peaks from bulk islet ATAC-seq across individual samples. (B) Fold enrichment (log2) of different sets of reproducible peaks from bulk islet ATAC-seq across 13 islet chromatin states. Genic Enhancer is not shown because of no enrichment. (C) Overlap of cell-type peaks (α, β, predicted δ) with different sets of reproducible peaks from bulk islet ATAC-seq data. (D) Overlap of cell-type peaks (α, β, and predicted δ) with different sets of reproducible peaks from bulk islet ATAC-seq data.

2.3.4 T2D GWAS enrichment at cell-type-specific chromatin signatures

We computed the overlap enrichment of T2D GWAS loci in cell-type peak annotations from α, β, and δ cells using a Bayesian hierarchical model, as implemented in fGWAS [207]. fGWAS allows calculation of marginal enrichment associations for one cell type conditioned on another by using not only the subset of genome-wide significant loci but also the full genome-wide association summary statistics. We observed that annotations from all three cell types were highly enriched for T2D GWAS loci, with β-cell annotations having the highest enrichment values (Figure 2.8A). However, when we accounted for marginal associations using a joint model, we found that β cells are the only cell type to remain enriched after adjusting for the other two cell types. This result suggests that shared or β cell-specific chromatin accessibility peaks drive the association with T2D GWAS. More broadly, these findings illustrate how single cell chromatin profiling results, when coupled with conditional statistical enrichment analyses, can dissect specific cell types that drive enrichment in bulk tissue samples.

We next partitioned the peaks into exclusive sets based on the cell-types shared by each peak. Because the δ cell cluster has fewer reads compared to α and β cells, we did not utilize read count-based approaches to determine cell-type-specific peaks. Instead, we used peak level metrics to identify peaks exclusive to a combination of cell types. We found that a majority of peaks (47,209) were shared across all cell-types and that each cell type had a set of unique accessible sites (29,884 β, 39,353 α, 31,330 δ) (Figure 2.8B). Consistent with our expectations, TSS-proximal shared peaks mostly overlapped active TSS chromatin states compared to cell-type-specific peaks which had a larger proportion of peaks in active enhancer states (Figure 2.7A).

To further understand the regulatory logic, we looked for TF motifs enriched in cell-type-specific peaks using GAT [110]. We found enrichment of motifs implicated in islet cell-type-specific functionality consistent with known islet TFs (Figure 2.7B). For example, PDX1 is enriched in beta-specific peaks, while GATA6 and FOXA are enriched in alpha-specific peaks. We also observed enrichment of motifs relevant to endocrine function such as PAX6 and MAF. For delta cell peaks, we found HHEX to be the only TF signature gene (out of 17 scRNA-seq cell-type

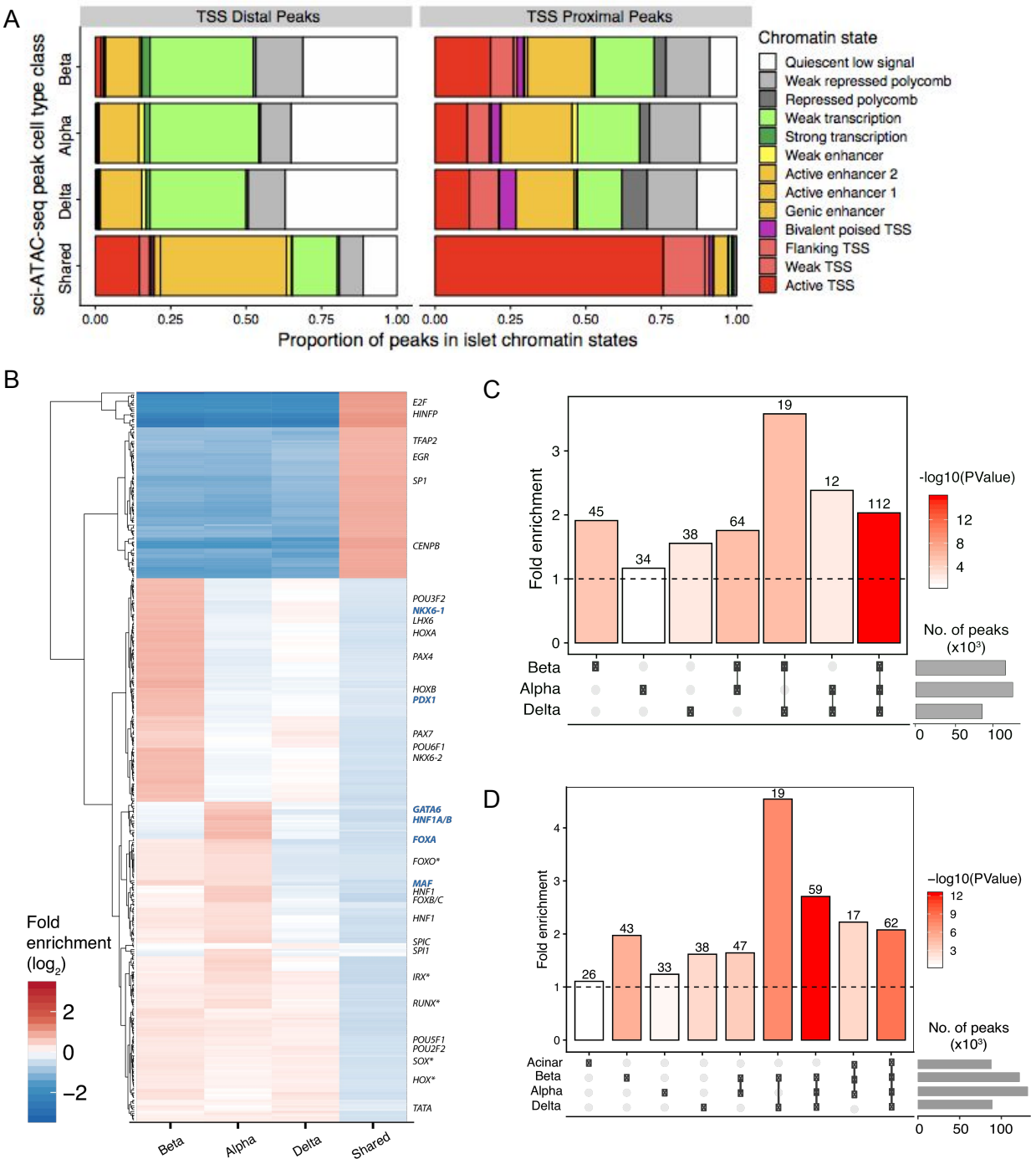


Figure 2.7: sci-ATAC-seq peaks have unique cell type and shared chromatin accessibility signatures. (A) Distribution of TSS proximal and distal peaks (>5kb from nearest Refseq TSS) in shared peaks and peaks assigned only to α, β, and δ cells. (B) Transcription factor (TF) motif enrichment (log2) across cell-type specific and shared peaks. (C) Enrichment of T2D GWAS SNPs (N=378) across all cell-type specific sets of peaks. (D) Enrichment of T2D GWAS SNPs (N=378) across all cell-type specific sets of peaks including peaks from acinar cells.

signature genes) that encodes a transcription factor, but we saw no delta specific enrichment. We think this could be because HHEX is a member of the homeobox family of TFs, and therefore has a highly degenerate motif, which could result in less specific enrichment within delta-specific peaks. Overall, the alpha and beta peak motif enrichments are consistent with known cell-specific TFs. We then used a complementary enrichment approach with the GREGOR tool [233] to determine if T2D GWAS loci are enriched in each subclass of peaks. We found that T2D GWAS loci were highly enriched in shared peaks (P-value=1.64e-16, fold enrichment=2.03) and beta cell-specific peaks (P-value=6.42e-6, fold enrichment=1.91) (Figure 2.7C). We also observed moderate enrichment of T2D GWAS SNPs in other sets of cell-type-specific peaks, but strikingly, there was little enrichment in delta cell-specific peaks (P-value=3.12e-3, fold enrichment=1.55) and no significant enrichment in alpha cell-specific peaks (P-value=1.83e-01, fold enrichment=1.16). This suggests that the role of alpha and delta cells in the mechanisms underlying genetic predisposition to T2D pathophysiology might be limited compared to beta cells. To further tease apart the role of shared peaks into islet endocrine specific peaks and constitutive peaks shared across more broad cell types, we added peaks from a sorted acinar cell population. Reassuringly, we observed that acinar-specific peaks showed no enrichment (P-value=0.31, fold enrichment=1.10) (Figure 2.7D). These independent enrichment findings from the GREGOR tool are consistent with the results from the fGWAS analysis (Figure 2.8A), indicating the robust nature of these results.

2.3.5 Linking cell-type-specific chromatin accessibility to target genes

One of the primary challenges in understanding the underlying biological mechanisms at non-coding T2D GWAS variants is the identification of their target genes. Risk variants occurring in enhancer regions can often interact with their target genes that are not adjacent. Multiple studies have examined the regulatory landscape of pancreatic islets and relevant cell lines using chromosome conformation capture techniques to nominate target genes [99, 150, 179]. However, most of these studies were conducted on bulk islet samples, thereby obscuring any cell-specific signatures of chromatin looping. Additionally, chromatin looping studies tend to have noisy signals

when two regions are close in linear space, which leads to a bias towards detecting longer-range interactions. In order to mitigate these limitations, we adopted a recently published approach, Cicero [208], which leverages profiles of chromatin co-accessibility across single cells to infer pairs of chromatin peaks that are likely to be in close physical proximity. For this analysis, we focused on α and β cell-types as they were the clusters with the most nuclei. In order to filter the Cicero co-accessible scores for those peak pairs that are more likely to represent true looping, we compared our results to experimentally-defined loops from three independent chromatin looping data sets: islet Hi-C [99], islet promoter capture Hi-C (pcHi-C) [179], and EndoC Pol2 ChIA-PET [150] loops. We found that Cicero peak-pairs from our sci-ATAC-seq data with score >0.05 were strongly enriched to be called as loops in each of the three reference data sets (Figure 2.8C, Figure 2.9A). With this threshold, we found 190,176 β cell and 147,716 α cell co-accessible peak-pairs.

Using our new catalog of Cicero-inferred chromatin loops, we next sought to link TSS-distal T2D GWAS variants to target gene promoters. We focused on the latest T2D GWAS results and used SNPs in association signals that were genetically fine-mapped to be in a 99% credible set and had >0.05 posterior probability of association (PPAg) [166]. For this mapping procedure, we required that the credible set SNP was not within 1 kb of an annotated TSS and that the other end of the chromatin loop occurs within 1 kb of an annotated TSS. Using this approach across both α and β cells, we found that of the 265 independent GWAS signals containing SNPs that met our criteria, we were able to nominate target genes at 104 of them (Figure 2.8D). Similarly, we checked if the SNPs within each locus overlapped a cell-type-specific peak. We observed several notable examples. At the C2CD4A/B locus, we found rs7163757 (PPAg 0.095) to be linked to *C2CD4B* in α cells (Figure 2.8E).Using an islet gene expression and genetic integration approach to identify expression quantitative trait loci (eQTL), we previously showed that rs7163757 is associated with *C2CD4B* expression [264], and a subsequent functional study corroborated these findings [146]. At a different locus, we found rs11708067 (PPAg 0.79), located in an islet enhancer within the *ADCY5* gene, to be linked to the TSS of the corresponding gene (Figure 2.8F). The risk allele of rs11708067 has been reported to be associated with reduced expression of *ADCY5* [225] and

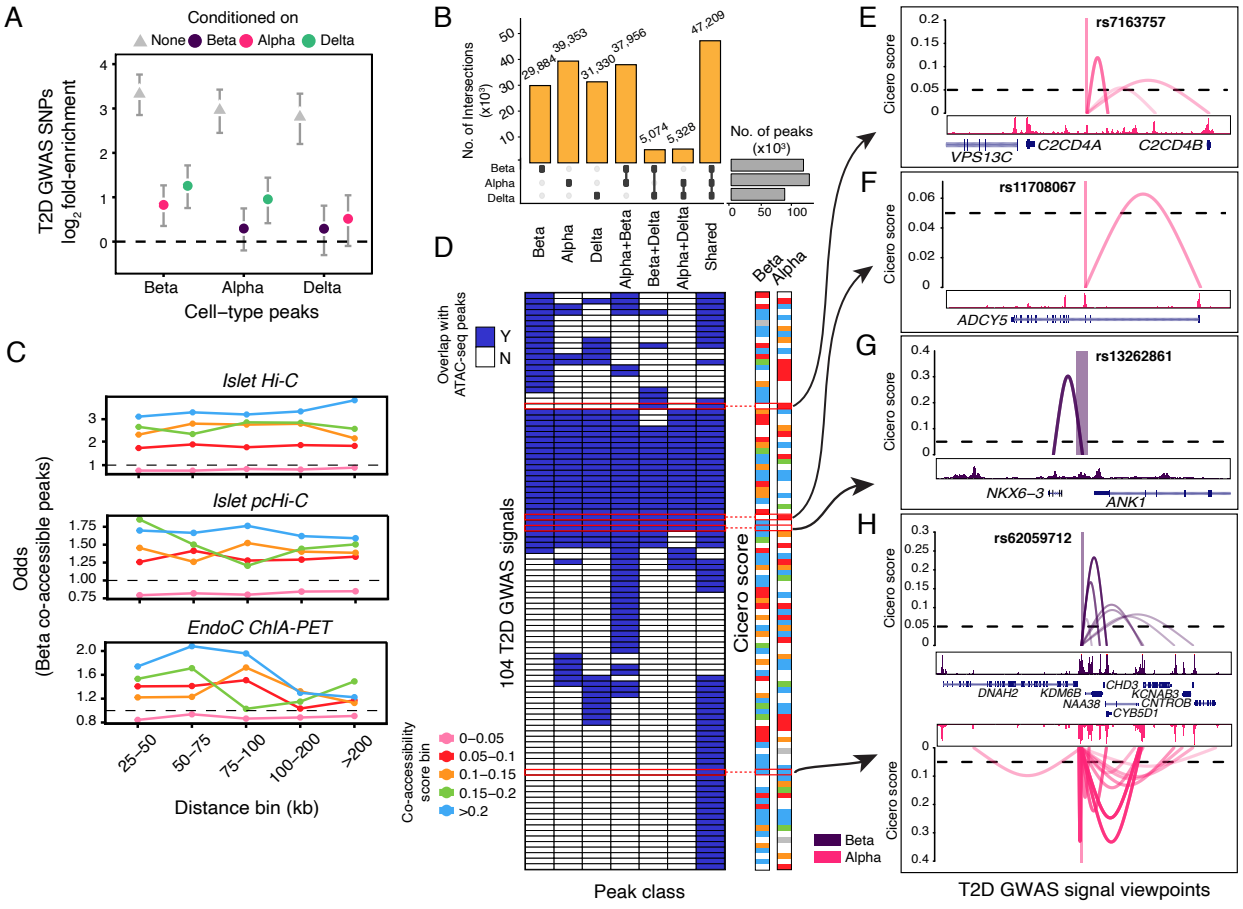


Figure 2.8: Enrichment of T2D GWAS signals in cell-type-specific chromatin and linking them to target genes. (A) Fold enrichment (log2) of T2D GWAS SNPs in cell-type peaks in single and conditional analysis mode using fGWAS tool. For each cell-type, three enrichment values with 95% confidence intervals are shown: None (single annotation mode), α (conditioned on β and δ), β (conditioned on α and δ), and δ (conditioned on α and β). (B) Partitioning of α, β, and predicted δ peaks in mutually exclusive sets of cell-type-specific peaks. The subplot (on right) shows the total number of peaks for each cell-type. (C) Distance-matched Fisher odds that β cell co-accessibility links overlap islet Hi-C, islet pcHi-C, and ChIA-PET chromatin loops across different co-accessibility threshold bins. (D) Overlap of T2D GWAS credible set SNPs with cell-type-specific peaks. Bin is colored if there's at least one SNP (PPAg > 0.05) in the 99-pct genetic credible set of the T2D GWAS signal located within 1 kb of an ATAC-seq peak. Cicero score columns are colored to indicate the score of the highest scoring link to the target gene. (E) Viewpoint plot of α Cicero connections centered at rs7163757 for *C2CD4A/B* locus, (F) α Cicero connections centered at rs11708067 for *ADCY5* locus, (G) β Cicero connections centered at rs13262861 for *ANK1* locus, and (H) Cicero connections for both α and β centered at rs62059712 for *ATP1B2* locus. The viewpoint region is +/− 1kb of the region from the variant.

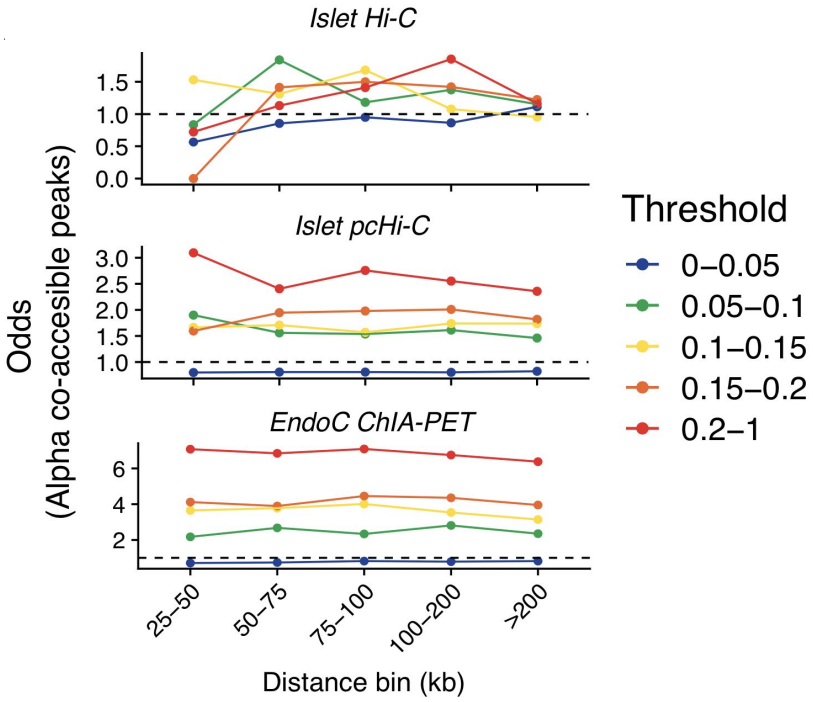


Figure 2.9: Enrichment of α cell co-accessible peaks in chromatin loop anchors. Fisher odds score for enrichment of alpha co-accessible sites in loop anchors from three different datasets: Islet Hi-C, Islet pcHi-C, and EndoC ChIA-PET.

functional validation experiments show association with impaired insulin secretion [31, 265]. As an example of a β cell-specific connection, we found variant rs13262861 (PPAg 0.97) within the ANK1 locus to be linked to nearby *NKX6-3* (Figure 2.8G). We have previously used islet eQTL data to nominate *NKX6-3* as an islet target gene at this locus [264, 265]. The extensive support from previous publications for these three loci serves as positive controls for our results and reinforces the quality of this sci-ATAC-seq data and analyses. Finally, we highlight rs62059712 (PPAg 0.34) within the *ATP1B2* locus as an example of a variant linked to multiple gene promoters across both β and α cell-types (Figure 2.8H). Notably, of the 104 T2D GWAS signals for which we were able to nominate target genes in either cell type, 60 (~58%) had more than one nominated target gene.

2.4 Discussion

Single-nuclei chromatin accessibility profiling provides a unique approach for mapping of cell-type-specific regulatory signatures. Here, we utilized the sci-ATAC-seq protocol to generate and study chromatin accessibility profiles for 1,456 high-quality nuclei from a purified pancreatic islet sample. Our dataset and analyses provide high-quality maps of cell-type accessibility profiles and regulatory architecture using an unbiased approach compared to prior maps from sorted cell-type populations. However, it is essential to emphasize that single-cell data present unique challenges, and that our study, which analyzed only one pancreatic islet sample, may be limited in how it can address some of them.

First, de-novo identification of cell types from the sparse single-cell chromatin accessibility data continues to be a challenge. We adopted several strategies to address potential biases in our analysis. Our logistic regression approach to eliminate read depth as a confounding technical variable, combined with the binomial counting strategy to infer doublet enrichment in clusters, enabled us to identify three major cell-type populations corresponding to α, β, and δ cells. In order to assign these cell identities, we relied not only on classical hormone markers, but we also leveraged findings from an independent islet single-cell RNA-seq study to validate our results. While islets have been reported to contain other rarer cell-type populations (<5% of all islet cells) [34], our ability to observe them was limited due to the size of our dataset.

Second, we faced the challenge of identifying reliable cell-specific accessibility patterns across all cell types due to the relatively low abundance of δ cells. As such, our U-Net-based deep learning approach presents a novel strategy for addressing this particular problem. Our model differs from a related deep learning method, Coda [142], by focusing on single cell ATAC-seq as opposed to bulk histone ChIP-seq data and uses a more complex architecture (U-Net) which has been used before in image processing related tasks [86, 226] but seen little mention in genomics [80]. We demonstrated, using α and β cells as reciprocal training and testing datasets, that our model successfully learns to predict high-quality peak calls from low cell count data. We observed, however, that there are diminishing returns from using deep learning models when 200

or more cells are used as input to the model, an observation consistent with the threshold of experimental reproducibility highlighted in a recent large-scale single nuclei ATAC-seq study [231]. This consistency with an independent study reinforces the value of our deep learning approach but also highlights a limitation of our δ peak predictions which derive from a low cell count input dataset. Nonetheless, we envision that our method will be useful in scenarios where it is challenging or cost-prohibitive to obtain specific cell populations.

Overall, an important implication of our findings comes from our ability to generate cell-specific chromatin accessibility maps and to infer looping connections from accessible regions to target genes of T2D GWAS variants. A recent T2D GWAS [166] reported >400 independent association signals, but the molecular mechanisms underlying these signals is known only for a subset of the variants. Single nuclei resolution cell-specific regulatory signatures provide a unique opportunity to infer target gene links with non-coding elements. Thus, we integrated cell-type co-accessibility links with T2D GWAS SNPs that were genetically fine-mapped to 99% credible sets to create a higher resolution map of the regulatory landscape underlying 104 distinct T2D GWAS signals. Focusing on the cell-specificity of the chromatin accessibility peaks that anchor these target gene associations, we observed seven classes, representing: i. peaks that are unique to a cell type (three classes), ii. peaks that are shared across all three cell types (one class), iii. peaks that occur in a pair of cell types (three classes). Interestingly, the class of peaks shared across all three cell types comprised 26 of the 104 (25%) T2D GWAS to target gene links even though this class is only one of seven. These results paint a complex picture of disease mechanisms where certain risk variants may mediate target effects through cell-type-specific pathways, while others might affect multiple target genes shared across cell-type populations.

We noted specific examples at the *C2CD4A/B* and *ANK1* loci, where we were able to nominate specific variants linked with islet gene expression and their role in T2D pathophysiology as compelling targets for future mechanistic studies. As this manuscript was under preparation, another similar study appeared as a preprint (published as of April 1, 2022) [44], and as such an important future topic will be to combine and meta-analyze multiple islet single-cell ATAC-seq

datasets. Such an endeavor will increase statistical power to detect chromatin features, including loops, at GWAS loci, and eventually enable single-cell resolution chromatin QTL studies, which will help to further narrow in on functional SNPs. Overall, we believe that the data, results, and methodology from this study will be of value to the broader research community.

2.5 Materials and methods

2.5.1 Bulk Islet ATAC-seq

2.5.1.1 Sample processing

The human pancreatic islet samples were procured and processed as described in [264]. Briefly, the islets were obtained from the National Disease Research Interchange (NDRI) and processed according to the NHGRI institutional review board-approved protocols. The islet was shipped overnight from the distribution center. Upon receipt, we pre-warmed the islet to 37 degree in shipping media for 1-2h before harvest. ~50-100 islet equivalents (IEQs) were harvested and transposed in triplicate following the methods in [28]. The ATAC-seq library was barcoded and sequenced 2 × 125bp on a HiSeq 2000.

2.5.1.2 ATAC-seq analysis

Sequencing adapters were trimmed using cta (v0.1.2) [121] and aligned to hg19 reference genome using BWA-MEM (v0.7.15-r1140, options: -I 200,200,5000) [156]. Picard MarkDuplicates (v2.18.27) was used for duplicate removal and samtools [157] was used to filter for autosomal, properly-paired and mapped read pairs with mapping quality >= 30 (v1.9, options: -f 3 -F 3340 -q 30). Replicates across each sample were merged into a single file using samtools merge. For peak calling, each sample was downsampled to 25 million (M) reads and converted to BED file. We then used MACS2 [285] to call broad peaks (v2.1.1.20160309; options: –nomodel –broad –shift -100 – extsize 200 –keep-dup all –SPMR) and removed those with FDR >0.05 and overlapping with ENCODE hg19 blacklists [253]. ATAC-seq coverage tracks were displayed using UCSC Genome

Browser and Integrative Genomics Viewer (IGV). Summary statistics were calculated using Ataqv (v1.0) [199] and are available in interactive and downloadable format online. For comparative purposes, we performed the same read trimming, alignment, filtering, downsampling, and peak calling steps on publicly available ATAC-seq data. Peaks from each sample were merged to create a master peak set and Spearman correlation was computed on the RPKM normalized read-count matrix.

2.5.1.3 Determination of high-confidence peaks

We randomly sampled 2.5 M reads from each sample using samtools view and pooled them into one file so that each sample is equally represented. Peaks were called on the pooled file as discussed in the previous paragraph. We then determined the number of samples overlapping with each master peak using peaks called on individual samples.

2.5.1.4 Overlap of reads with chromHMM states

We tested for enrichment of ATAC-seq peaks across 13 islet-specific chromatin states using Genomic Association Tester (GAT) [110]. We ran GAT (v1.3.5, options: –number-samples 10,000) and filtered chromatin states with no significant enrichment (Bonferroni adjusted p-value < 0.05) of peaks in them. The log2 fold enrichment values across chromatin states were clustered using hierarchical clustering of the correlation matrix.

2.5.2 sci-ATAC-seq analysis

2.5.2.1 Sample processing

We used the combinatorial cellular indexing method to generate single-nuclei chromatin accessibility data as previously described in [55]. Briefly, a suspension of islet cells were obtained and pelleted 5 min at 4℃ 500 x g. The media was aspirated and the cells were washed once in 1 ml PBS. The cells were pelleted again for 5 min at 4 ℃ 500 x g and then resuspended in 1 ml of cold lysis buffer (10 mM Tris-HCl, pH 7.4, 10 mM NaCl, 3 mM MgCl2, 0.1% IGEPAL CA-630, supplemented

with 1X protease inhibitors (Sigma P8340)). Nuclei were maintained on ice whenever possible after this point. 10 µl of 300 µM DAPI stain was added to 1 ml of lysed nuclei for sorting. To prepare for sorting, 19 µl of Freezing Buffer (50 mM Tris at pH 8.0, 25% glycerol, 5 mM MgOAc2, 0.1 mM EDTA, supplemented with 5 mM DTT and 1X protease inhibitors (Sigma P8340)) was aliquot into each well of a 96-well Lo-Bind plate. 2,500 DAPI+ nuclei (single cell sensitivity) were sorted into each well of the plate containing Freezing Buffer. The plate was sealed with a foil plate sealer and then snap frozen by placing in liquid nitrogen. The frozen plate was then transferred directly to a -80 ℃ freezer. Subsequently, the sample was shipped from NIH to UW overnight on dry ice. The plate was then thawed on ice and supplemented with 19 µl of Illumina TD buffer and 1 µl of custom indexed Tn5 (each well received a different Tn5 barcode). The nuclei were tagmented by incubating at 55 ℃ for 30 min. The reaction was then quenched in 20 mM EDTA and 1 mM spermidine for 15 min at 37 ℃. The nuclei were then pooled and stained with DAPI again. 25 DAPI+ nuclei were then sorted into each well of a 96-well Lo-bind plate containing 11.5 µl Qiagen EB buffer, 800 µg/µl BSA, and 0.04% SDS. 2.5 µl of 10 µM P7 primers were added to each sample and the plate was incubated at 55 ℃ for 15 min. 7.5 µl of NPM was then added to each well. Finally, 2.5 µl of 10 µM P5 primers were added to each well and the samples were PCR amplified with following cycles: 72 ℃ 3min, 98 ℃ 30s, then 20 cycles of 98 ℃ for 10 s, 63 ℃ for 30 s, 72 ℃ for 1 min. The exact number of cycles was determined by first doing a test run on 8 samples on a real-time cycler with SYBR green (0.5X final concentration). PCR products were then pooled and cleaned on Zymo Clean&Concentrator-5 columns (the plate was split across 4 columns) eluting in 25 µl Qiagen EB buffer and then all 4 fractions were combined and cleaned using a 1X Ampure bead cleanup before eluting in 25 µl Qiagen EB buffer again. The molar concentration of the library was then quantified on a Bioanalyzer 7500 chip (including only fragments in the 200-1000 bp range) and sequenced on an Illumina NextSeq at 1.5 pM concentration.

2.5.2.2 QC and pre-processing

Step 1. Barcode correction and filtering. Each barcode consists of four 8-bp long indexes (i5, i7, p5, and p7). Reads with barcode combinations containing more than 3 edit distance for any index were removed. If a barcode was within 3 edits of an expected barcode and the next best matching barcode was at least 2 edits further away, we corrected the barcode to its best match. Otherwise, the barcode was classified as ambiguous or unknown.

Step 2. Adapter trimming and alignment. Adapters were removed using Trimmomatic [17] with NexteraPE adapters as input (ILLUMINACLIP:NexteraPE.fa:2:30:10:1:true TRAILING:3 SLIDINGWINDOW:4:10 MINLEN:20) and aligned to hg19 reference using BWA-MEM (v0.7.15-r1140, options: -I 200,200,5000). The final alignment was filtered using samtools to remove unmapped reads and reads mapping with quality < 10 (-f3 -F3340 -q10) as well as reads that were associated with ambiguous or unknown barcodes.

Step 3. Deduplication and nuclei detection. Duplicates from the pruned file were removed using a custom Python script on a per-nucleus basis. Using the distribution of reads per barcode, we applied bi-culstering, as implemented in the mclust [90] R package, to differentiate between background barcodes and barcodes that correspond to a nucleus. Using the list of non-background barcodes, we split the aggregate bam file into constituent bam files corresponding to each barcode representing a single nucleus using a custom Python script.

Step 4. Quality assessment of each single nucleus. For each single nucleus, we computed ATAC-seq quality metrics such as fragment length distribution, transcription start site (TSS) enrichment, short-to-mononucleosomal reads ratio, total autosomal reads, and fraction of reads overlapping peaks. We removed nuclei with a) total reads outside 5% to 95% range [34578, 226755] of all the nuclei, and b) TSS enrichment of <2.7 (5%-tile) from further downstream analysis.

Step 5. Aggregate sci-ATAC-seq peaks. We pooled reads from filtered barcodes from the previous steps to create an aggregate bam file. Peaks were called and filtered as described previously in the Bulk Islet ATAC-seq analysis section.

2.5.3 Cluster analysis

2.5.3.1 Feature selection and clustering

We generated a list of TSS distal peaks (>5 kb away from the nearest TSS based on RefSeq genes [192]) from the aggregate sci-ATAC-seq data. For each nucleus, we counted the number of reads overlapping the peaks using the Rsubread package [34]. We then adopted a logistic regression approach to remove peaks where binarized accessibility across nuclei was significantly associated (Bonferroni corrected p-value < 0.05) with sequencing depth. This approach should help to reduce the bias associated with sequencing depth, as the remaining peaks are no longer associated with this technical factor, a strategy that has been successfully implemented in single cell RNA-seq data analysis [96]. The resulting count matrix was RPKM-normalized and reweighted using the term-frequency and inverse-document-frequency (TF-IDF) method [54]. To do this, we first weighted all the sites for individual nuclei by the total number of sites accessible in that cell ("'term frequency"'). We then multiplied these weighted values by log(1 + the inverse frequency of each site across all cells), the "inverse document frequency." The TF-IDF transformed matrix was then reduced to 30 principal components using Principal Component Analysis (PCA) and used as input to generate a two-dimensional embedding using the Uniform Manifold Approximation Method (UMAP, n_neighbors = 20) [176]. We identified clusters in the two-dimensional embedding in an unsupervised manner using a density-based clustering method (hdbscan, minPts = 20) as implemented in `dbscan` R package [102].

2.5.3.2 Cell identity assignment and validation

The cell identities were assigned based on de-facto cell-type-specific hormone markers: *INS-IGF2* (β),*GCG* (α), *SST* (δ) etc. A marker gene was said to be present in a cell if a read mapped within 5 kb

of the GENCODE (v19) gene body annotation [107]. For additional verification of cell-identity, we computed RPKM normalized aggregate ATAC-seq signal across cell-type marker genes reported in two independent islet scRNA-seq studies [151, 237]. Finally, we evaluated the enrichment of cells from each cell-type cluster relative to their expected population proportion using two-sided binomial test across ten bins of sequencing depth (~145 cells/bin).

2.5.4 Deep learning signal and peak upscaling

2.5.4.1 Model design, training, and validation strategy

The U-Net model [226] takes input sequences and outputs prediction sequences. The goal of model training is to reduce the error between the prediction output and a representation of ground truth. For signal upscaling, the input sequence is base-wise scores of BAM pileups (read-depth) corresponding to a subsample of n cells (randomly sampled from 600 cells) and the output sequence is base-wise scores of BAM pileups using reads from all 600 cells. Peak upscaling not only uses the subsampled BAM pileup scores as inputs, but it also uses the binary base-wise values from calling peaks with MACS2 on the subsampled BAM alignments. Output sequences for peak upscaling are the binary base-wise values from calling peaks with MACS2 on the data. We created two models, each separately on data from α and β cells. Because both had different number of constituent single-nuclei, we matched the size of output dataset by randomly sampling 600 cells from each cell-type cluster. The input datasets were created by sampling n cells from the set of 600 cells such that the total number of reads is similar across both models. We did not set any explicit constraints on the number of peaks to be called by this approach.

2.5.4.2 Model architecture

The network architecture of the U-Net used in this study is illustrated in Figure 2.4A. It consists of a contracting, convolutional path (left side) and an expansive, deconvolutional path (right side). The contracting path consists of repeated applications of two kernel size 11 convolutions (unpadded convolutions) with rectified linear unit (ReLU) activation, and a kernel size 2 max

pooling operation with stride 2 for downsampling. Each downsampling step halves the length of the activation sequence while doubling the number of feature channels. Every step in the expansive path consists of a kernel size 2 deconvolution layer with a linear activation function that halves the number of feature channels, a concatenation with the correspondingly cropped feature map from the contracting path, and two kernel size 11 convolution layers with ReLU activations. The cropping is necessary due to the loss of border sequence steps in non-padded convolution. At the final layer a kernel size 1 convolution with either an ReLU (for signal upscaling) or sigmoid (for peak upscaling) activation function generates the sequence of predictions. Due to the use of unpadded convolutions, the prediction sequence is shorter than the input sequence by a constant border width. Although the U-Net can accept arbitrary length input sequences, we fix all training samples to be of length 6700, which results in output prediction sequences of length 4820. In total, the network has five steps each in the contracting and expansive paths for a total of 27 convolutional layers and 8,998,529 training parameters. The model was implemented using Keras [91] with the Tensorflow [1] backend, and experiments were run using Titan Xp and GTX 1080 Ti GPUs.

To reduce overfitting, we split chromosomes into training, validation, and testing sets. The model was fit using the ADAM optimizer [139] with a learning rate of 1e-5 and a batch size of 128 for 50 epochs. Separate loss functions, and hence models, were used to solve signal and peak upscaling. For signal upscaling, we used the mean squared error base-wise loss function. For peak upscaling, the loss function was the sum of the cross-entropy base-wise loss and the Dice-coefficient loss, also known as F1 score. We used mean average precision, a common evaluator for object detection, and Pearson correlation as the output evaluation metrics for peak and signal upscaling, respectively. This downscaling and model training were repeated for n=5, 10, 28, 50, 100, 200, 300, 400, and 500 cells.

2.5.4.3 Generating upscaled peaks

In order to select a subset of high-confidence peaks from the predicted model output, we adopted a post-hoc approach where we compared the number of cell-type-specific peaks for α, β, and δ cells, and chose a threshold where they had a similar number. For predicted δ cell peaks, we combined the results from α and β models at the same threshold using bedtools [215] intersect (v.2.27.1) after filtering for the chosen threshold.

2.5.5 Cell-type-specific peaks analysis

2.5.5.1 Cell-type-specific peaks

Peaks specific to each cell-type were obtained by comparing peaks in one cell-type with all other cell-types using bedtools.

2.5.5.2 T2D GWAS SNPs enrichment

Enrichment of T2D associated GWAS SNPs from DIAMANTE [166] was tested using GREGOR (v1.3.1) [233]. Specifically, we used the following parameters: r^2 threshold (for inclusion of SNPs in LD with the diabetes associated GWAS SNPs) = 0.80, LD window size = 1 Mb, and minimum neighbor number = 500. P-values were adjusted according to Bonferroni threshold for multiple testing burden.

2.5.5.3 Conditional fGWAS enrichment analysis

We used fGWAS [207] to model shared properties of loci affecting a trait. We ran fGWAS (v0.3.6) with DIAMANTE T2D GWAS summary data and cell-type ATAC-seq peaks from three cell types as input annotations. For each individual annotation, the output model provided maximum likelihood enrichment parameters and annotations were considered as significantly enriched if the parameter estimates and 95% confidence interval (CI) did not overlap zero. We then used fGWAS to run a conditional analysis in a pair-wise manner where enrichment of one model was evaluated

conditional on the output models from other annotations.

2.5.5.4 Validating cell-type peaks using scRNA-seq signature genes

We evaluated the enrichment of scRNA-seq derived signature genes (scRSGs) in 28-cell MACS2 and upscaled peak calls using GAT [110]. We ran GAT (v1.3.5, options: –number-samples 10,000) with union of all peaks as workspace and scRSGs as segments.

2.5.5.5 Transcription factor motif enrichment

We used motif PWN scans from [235]. Briefly, we used biallelic SNPs and short indels from the 1,000 Genomes project (release v5) [252] to generate comprehensive scans with FIMO [49] using the background nucleotide frequencies from hg19 and a p-value < 1e-4. We only kept motif instances that intersected mappable regions and did not intersect blacklisted regions. We then tested for enrichment of motifs across cell-type-specific peaks using GAT (v1.3.5, options: –number-samples 100,000). We used union of top 100 motifs (by log fold enrichment) for each annotation and clustered them using hierarchical clustering.

2.5.6 Linking SNPs to target genes

2.5.6.1 Cicero co-accessibility analysis

In order to link TSS distal ATAC-seq peaks with target genes, we used Cicero [208], which identifies co-accessible pairs of DNA elements using single-cell chromatin accessibility data. We used these results to infer connections between regulatory elements and their target genes. We ran Cicero (v1.0.15, default parameters) with cells from the α and β cell clusters separately. To do this, we first called peaks on each cluster and counted the number of reads per nuclei overlapping the peaks. The resulting count matrix was used as input to Cicero along with the UMAP projection for each cluster. Finally, in order to decide a threshold for filtering co-accessible peak pairs, we computed Fisher odds ratio for enrichment of co-accessible peaks versus distance matched non co-accessible peaks (co-accessibility < 0) with three different three-dimensional chromatin

looping data sets: islet Hi-C [99], islet promoter capture Hi-C (pcHi-C) [179], and EndoC Pol2 ChIA-PET anchors [150]. For overlap, we checked if both the ends of the Cicero loops intersected with both the anchors from the experimental chromatin looping data. Public epigenome browser session links have been included in Table S7 of [217].

2.5.6.2 T2D GWAS SNP overlap analysis.

In order to link T2D GWAS SNPs with target genes, we utilized 380 independent GWAS signals from DIAMANTE that were genetically fine-mapped to 99% credible sets using a Bayesian approach. In this framework, each SNP has a posterior probability for being causal for the association in that region. These posterior probabilities are the ratio of evidence for each variant versus all others which makes it easy to compare the variants directly. A genetic credible set is then defined as the minimum set of SNPs that contains all SNPs with probability greater than or equal to 0.01. We filtered SNPs within each set to have >0.05 posterior probability of association (PPAg). We then checked for each GWAS signal whether SNPs passing the criteria mapped within 1 kb of cell-type-specific ATAC-seq peaks. To obtain Cicero target genes, we checked if an ATAC-seq peak was a) within 1 kb of a variant, b) outside 1 kb range of a RefSeq TSS, and (c) linked to an ATAC-seq peaks that was within 1 kb of a RefSeq TSS. The binary overlap matrix was clustered using hierarchical clustering with binary distance method. Tables containing number of SNPs within each credible set and specific variants overlapping cell-type-specific ATAC-seq peaks are available in Table S8 and Table S9 of [217].

2.6 Data availability

Extended data and tables referenced in the chapter are not included in the dissertation and can be obtained from the online version of the manuscript referenced in [217]. Code for analysis done in this work is openly available on GitHub at https://github.com/ParkerLab/islet_sci-ATAC-seq_2019.

2.7 Acknowledgments

FSC, JS, and SCJP conceived the study; MRE, DC, and RMD generated the data; LSZ, JPD, and NN performed initial analysis; VR, DXQ, and SCJP analyzed the data and performed the research presented here; DXQ and YG implemented the U-Net strategy and wrote the code; FSC, JS, and SCJP jointly supervised the work; and VR, DXQ, and SCJP wrote the manuscript with feedback from all the authors.

This work was supported by American Diabetes Association (ADA) Fellowship to JPD, National Institute of Diabetes and Digestive and Kidney Diseases grant R01 DK117960, National Heart, Lung, and Blood Institute grant U01 HL137182, and the ADA Pathway to Stop Diabetes Grant 1-14-INI-07 to SCJP, and National Institutes of Health Grants 1-ZIA-HG000024 to FSC. The medical art images were obtained from Servier Medical Art by Servier under CC-BY-3.0 license. We thank F. Steemers and L. Christiansen at Illumina for providing indexed Tn5 transposase. We also thank NVIDIA Corporation for donating the Titan Xp GPUs to SCJP.

2.8 My contributions

The results presented in this chapter are the outcome of an outstanding collaboration between teams across three different institutions. I want to thank John Didion, Narisu Narisu, Mike Erdos, and Francis Collins at the National Institutes of Health for sharing the data and initial analysis; Darren Cusanovich and Jay Shendure at the University of Washington for generating the sci-ATAC-seq data; and Daniel Quang and Steve Parker at the University of Michigan for co-leading the analyses presented here. I also want to thank all the team members across the three labs for contributing to this project. The study (as published in [217]) reflects the success of the team's effort.

Of the results presented in this chapter, I contributed to bulk islet ATAC-seq analysis, sci-ATAC-seq data analysis, cluster analysis, cell-type-specific peak analysis, and the efforts to link variants to target genes. I also compiled the manuscript figures corresponding to these analyses,

wrote the first draft, and led the manuscript writing and revision with input from all the authors.

CHAPTER 3

RFX6-mediated Dysregulation Defines Human β Cell Dysfunction in Early Type 2 Diabetes

3.1 Abstract

A hallmark of type 2 diabetes (T2D), a major cause of world-wide morbidity and mortality, is dysfunction of insulin-producing pancreatic islet β cells [3, 220, 228]. T2D genome-wide association studies (GWAS) have identified hundreds of signals, mostly in the non-coding genome and overlapping β cell regulatory elements, but translating these into biological mechanisms has been challenging [45, 166, 217]. To identify early disease-driving events, we performed single cell spatial proteomics, sorted cell transcriptomics, and assessed islet physiology on pancreatic tissue from short-duration T2D and control donors. Here, through integrative analyses of these diverse modalities, we show that multiple gene regulatory modules are associated with early-stage T2D β cell-intrinsic defects. One notable example is the transcription factor RFX6, which we show is a highly connected β cell hub gene that is reduced in T2D and governs a gene regulatory network associated with insulin secretion defects and T2D GWAS variants. We validated the critical role of RFX6 in β cells through direct perturbation in primary human islets followed by physiological and single nucleus multiome profiling, which showed reduced dynamic insulin secretion and large-scale changes in the β cell transcriptome and chromatin accessibility landscape. Understanding the molecular mechanisms of complex, systemic diseases necessitates integration of signals from multiple molecules, cells, organs, and individuals, and thus we anticipate this ap-

proach will be a useful template to identify and validate key regulatory networks and master hub genes for other diseases or traits with GWAS data.

3.2 Introduction

Type 2 diabetes mellitus (T2D), a metabolic disease defined by hyperglycemia, is a major cause of macro and microvascular morbidity and mortality for more than 460 million individuals worldwide [228]. Clinically heterogenous, T2D involves genetic, environmental, and physiologic components that impact multiple molecular pathways and tissues [3, 220]. Initial management frequently involves diet and lifestyle alterations but often escalates to require multiple oral or injectable medications and ultimately exogenous insulin to lower blood glucose [9, 128]. T2D is associated with obesity and age, both of which reduce peripheral tissue sensitivity to insulin; however, most individuals with insulin resistance do not develop T2D. Instead, the key defining feature of those who develop T2D is impaired insulin secretion [103, 128]. Insulin is secreted endogenously by the β cell within the pancreatic islet. In addition to β cells, the islet also contains other endocrine cells (α, δ, γ, and ε), vascular structures (endothelial cells and pericytes), and immune cells, which collectively function as a mini-organ to control glucose homeostasis in a coordinated fashion [191, 273]. While islet dysfunction is a hallmark of T2D, it remains unclear whether this is the result of an intrinsic β cell defect, a reduction in β cell number, systemic signals from altered levels of fatty acids, glucose, or lipids, or some combination of these.

T2D has a strong genetic component with more than 400 signals identified through genome-wide association studies (GWAS) [166]. Loci linked to T2D through GWAS are enriched in β cell-specific open chromatin regions, suggesting impaired β cell processes as a key determinant for whether T2D develops and how quickly it progresses [43, 217]. Further, 90% of GWAS-identified single nucleotide polymorphisms (SNPs) are located in non-coding parts of the genome, and they are enriched in predicted islet enhancer regions where many likely modulate cell-specific gene expression regulatory networks by altering transcription factor binding [201, 202, 257, 259, 264].

How personalized genetic variation causes changes in cell-specific gene and protein expression, tissue architecture, and cellular physiology in T2D islets is not well understood.

Postulated T2D disease processes include β cell loss and/or dedifferentiation, endoplasmic reticulum (ER) stress, amyloid deposition, oxidative stress, glucotoxicity, lipotoxicity, and islet inflammation [77, 94, 249, 274]. These processes have been primarily studied in rodent models of T2D due to difficulty in obtaining and studying human pancreatic tissue and islets. Importantly, human islets show several key differences from mouse islets, including endocrine and non-endocrine cell composition and arrangement, basal and stimulated insulin secretion, response to dyslipidemia and hyperglycemia, and expression of key islet-enriched transcription factors [24, 60, 75, 188], highlighting the need for studies to define initiating and sustaining mechanisms of islet dysfunction in primary human islets.

Recent advances in cadaveric pancreas procurement and processing have increased availability of human tissue for histological analysis as well as ex vivo molecular and functional profiling of islets isolated from individuals with diabetes. However, many studies utilize only tissue or islets, and further, do not differentiate study outcomes based on T2D duration. Since different stages of T2D may involve different processes, studies that combine cases with different T2D duration make it difficult to discern cellular and molecular causes from disease consequences. The association of physiological measurements with transcriptomic profiles of islet cells have begun to identify key pathways for β cell function [36, 276], but integration of these studies with disease stage, tissue-based analyses, and genetic risk remains a challenge.

Here, we used an integrated approach to study the pancreas and isolated islets from donors with short-duration T2D and nondiabetic controls to identify disease-driving molecular defects early in the course of T2D. We analyzed islet function both ex vivo and in vivo using a transplant system and performed comprehensive transcriptional analysis by bulk RNA-sequencing (RNA-seq) of whole islets and purified β cells and α cells, correlating these profiles to functional parameters and GWAS variants using weighted gene co-expression network analysis (WGCNA). Concurrently, we described changes in the pancreatic islet microenvironment via traditional and

multiplexed imaging approaches, including assessing spatial cell relationships. We found that dysfunction in short-duration T2D is defined primarily by β cell-intrinsic defects, including an RFX6-governed and GWAS-enriched transcriptional regulatory network.

3.3 Results

3.3.1 Identification, collection, and processing of short-duration T2D donor pancreata

To identify early, disease-driving mechanisms in islets, we focused on short-duration T2D as defined by a combination of disease duration and treatment approach (Figure 3.1a). Using a national network, we identified high-quality organs to ensure minimal ischemic time and consistently applied multiple tissue processing and fixation methods, including simultaneous collection of isolated islets and tissue from the same pancreas when possible. Twenty pancreata were obtained from individuals with T2D aged 37-66y (mean 52y) with T2D duration of 0-10y (mean 3.5y). Of these donors, 25% were without pharmaceutical treatment (HbA1c range 6.2-9.9; mean 7.6) and 75% were on diabetes medication, mostly oral agents (HbA1c range 6.3-11.2; mean 8.0) (Figure 3.1a). Pancreata from nondiabetic (ND) donors (n=17) were also collected and processed for multi-modality study. Partnerships with the Integrated Islet Distribution Program (IIDP) and the Alberta Diabetes IsletCore provided access to additional islets from ND donors (n=19) to assist with matching of donor characteristics. Detailed information, including sample types and experimental usage for each case, is available in Extended Data Table 1 of [268]. Application of multiple modalities allowed for integrative analysis of ex vivo and in vivo islet function, tissue architecture and microenvironment including spatial relationships, and cell type-specific gene expression (Figure 3.1b).

3.3.2 Short-duration T2D islets show reduced stimulated insulin secretion

To investigate islet function, we assessed dynamic hormone secretion in isolated islets from age- and body mass index (BMI)-matched T2D and ND donors (Figure 3.2a-b) by a standardized perifusion approach that interrogates multiple steps of the insulin secretory pathway and has been adopted by the Human Islet Phenotyping Program of the IIDP to assess over 400 human islet preparations [22]. When normalized by islet volume, stimulated insulin secretion was substantially reduced in response to high glucose, cyclic AMP (cAMP)-evoked potentiation, and potassium chloride (KCl)-mediated depolarization (Figure 3.1c-f and Figure 3.2c). Both first and second phases of insulin secretion were reduced, with the first phase showing a more significant reduction (Figure 3.2d-e). Inhibition of insulin secretion by low glucose and epinephrine was similar between ND and T2D islets, as was insulin content (Figure 3.1g and Figure 3.2f); as such, normalization of response by islet insulin content showed similar reductions in stimulated insulin secretion but also showed reduced basal insulin secretion (Figure 3.2g-l). Together, these data suggest that short-duration T2D islets ex vivo maintain insulin production and storage but have defects at multiple steps of the insulin secretory pathway, including those distal to glucose metabolism, which persist after islet isolation from the in vivo environment.

In contrast to insulin secretion, neither basal nor stimulated glucagon secretion was different in T2D islets when normalized by islet volume (Figure 3.11h-1k and Figure 3.2m), and both ND and T2D islets showed glucose-mediated suppression of glucagon secretion (Figure 3.2n). Glucagon content was similar between islets from ND and T2D individuals and normalization by glucagon content showed similar secretion dynamics (Figure 3.1l and Figure 3.2o-t). While there is substantial evidence of dysregulated glucagon secretion in T2D [5, 260], these data suggest that either α cell dysfunction is not present in the early stages of T2D or defects are present in vivo but not maintained after islet isolation.

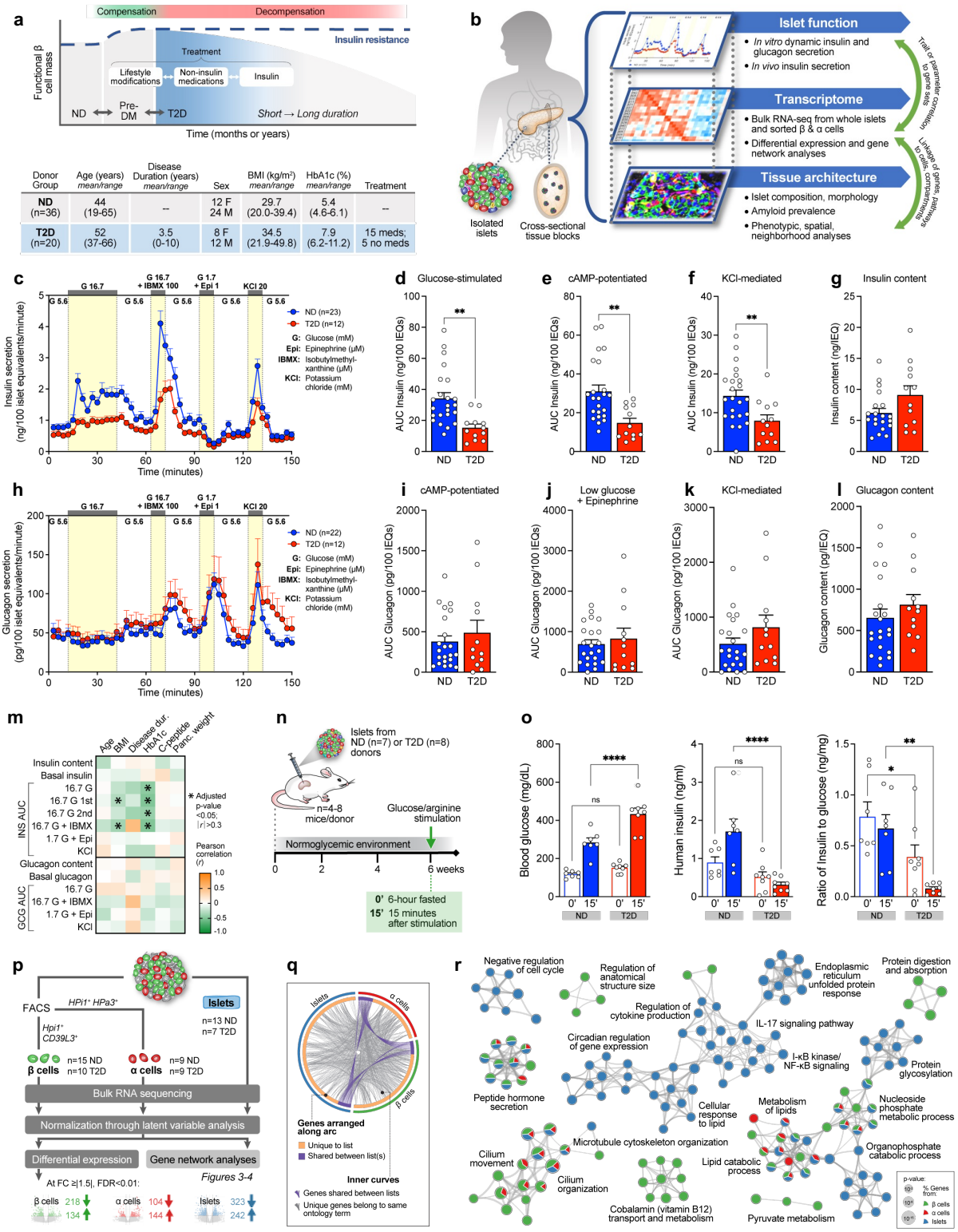
Correlation of donor attributes to functional metrics highlighted a significant negative correlation between donor HbA1c and stimulated insulin secretion ($r<-0.40$, $p<0.05$; Figure 3.1m).

To test whether the systemic environment contributed to β cell dysfunction in T2D islets, we transplanted T2D or ND islets from a subset of donors into normoglycemic, non-insulin resistant immunodeficient NOD-*scid-IL2rγ*^{null} (NSG) mice (Figure 3.1n). After six weeks in this environment, T2D islets secreted less human insulin than ND islets, especially after stimulation with glucose/arginine (Figure 3.1o, average per donor and Figure 3.2u, individual mice), consistent with ex vivo findings of impaired stimulated insulin secretion. In sum, these experiments highlight that β cell dysfunction in early T2D persists in a normoglycemic, non-insulin resistant environment and suggest that intrinsic β cell dysregulation and/or cellular and molecular alterations within the islet microenvironment are key features driving reduced insulin secretion.

3.3.3 Broad transcriptional dysregulation revealed through integrated transcriptome analysis of islets and purified α and β cells

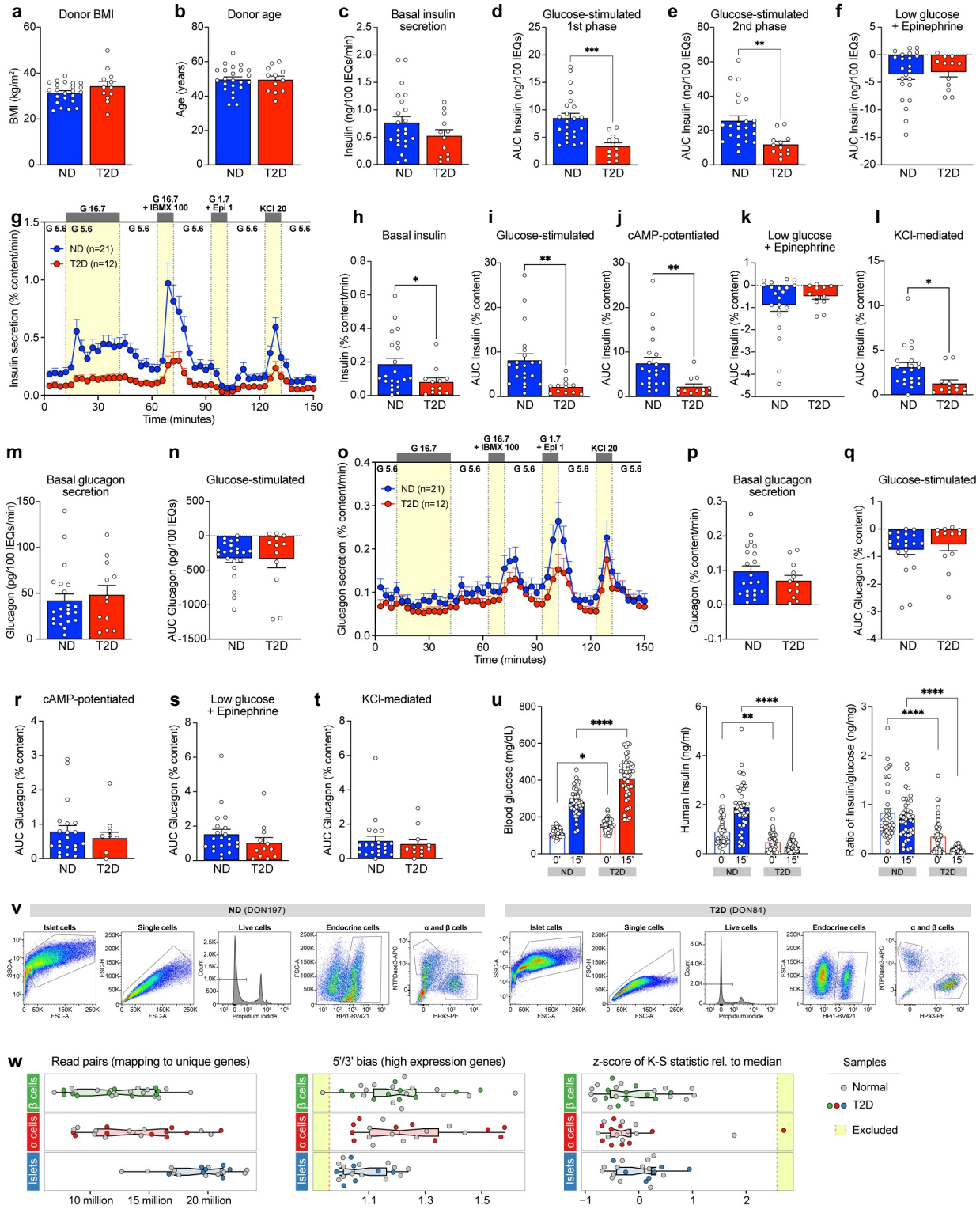
To assess both the β and α cell-specific transcriptional landscapes as well as global islet dysregulation in the short-duration T2D cohort, we purified β and α cells by fluorescence-activated cell sorting (FACS) using well-characterized cell surface antibodies and hand-picked isolated islets

Figure 3.1 *(following page)*: Integrated analysis of islet function, gene expression, and histology in a cohort of donors with short-duration type 2 diabetes (T2D) reveals substantially reduced stimulated insulin secretion ex vivo and in vivo despite similar insulin content and highlights dysregulated pathways in purified β and α cells as well as whole islets. (a) Schematic of functional β cell mass during disease progression from nondiabetic (ND) to pre-diabetes (Pre-DM) and T2D, highlighting the divergence of insulin supply and demand and escalation of treatment mirroring progressive loss of functional β cell mass. Shaded blue represents targeted disease stage in this cohort with clinical profile shown below in table. (b) Schematic of multimodal study of islet function, transcriptome, and tissue architecture. Coordinated study on islets and tissue from same donor allowed integration between analyses (green arrows). (c-l) Dynamic insulin and glucagon secretory responses measured by islet perifusion. Panels d-f and i-k: secretagogue response as area under the curve (AUC); g, l: hormone content normalized to islet volume. (m) Pearson correlation of perifusion metrics to clinical traits. (n) Schematic of human islet transplantation and in vivo assessment of function. (o) Blood glucose, human insulin levels, and human insulin:blood glucose ratio measured before and after glucose and arginine stimulation of mice with human islet grafts. Symbols show donor average. (p) Schematic of RNA sample collection and analysis. (q) Overlap of differentially expressed (DE) genes in T2D β cell (green) α cell (red), and islet (blue) samples at the level of genes (purple curves) or ontology terms (grey curves). (r) Metascape network showing a subset of enriched terms from DE genes. Edges denote similarity and node colors reflect contribution of sample(s). * p<0.05, ** p<0.01, *** p<0.001, **** p<0.0001 (d-g, i-l: two-tailed t-test; o: two-way ANOVA).


a
Compensation
Decompensation
Insulin resistance
Functional β cell mass
Treatment
Lifestyle modifications
Non-insulin medications
Insulin
ND
Pre-DM
T2D
Short → Long duration
Time (months or years)
b
Islet function
• In vitro dynamic insulin and glucagon secretion
• In vivo insulin secretion
Transcriptome
• Bulk RNA-seq from whole islets and sorted β & α cells
• Differential expression and gene network analyses
Tissue architecture
• Islet composition, morphology
• Amyloid prevalence
• Phenotypic, spatial, neighborhood analyses
Trait or parameter correlation to gene sets
Linkage of genes, pathways to cells, compartments, structures
Isolated islets
Cross-sectional tissue blocks
c
G 16.7
G 16.7 + IBMX 100
G 1.7 + Epi 1
KCl 20
G 5.6
ND (n=23)
T2D (n=12)
G: Glucose (mM)
Epi: Epinephrine (μM)
IBMX: Isobutylmethyl-xanthine (μM)
KCl: Potassium chloride (mM)
Insulin secretion (ng/100 islet equivalents/minute)
Time (minutes)
d Glucose-stimulated
e cAMP-potentiated
f KCl-mediated
g Insulin content
AUC Insulin (ng/100 IEQs)
Insulin content (ng/IEQ)
**
ND
T2D
h
Glucagon secretion (pg/100 islet equivalents/minute)
ND (n=22)
T2D (n=12)
i cAMP-potentiated
j Low glucose + Epinephrine
k KCl-mediated
l Glucagon content
AUC Glucagon (pg/100 IEQs)
Glucagon content (pg/IEQ)
m
Age
BMI
Disease dur.
HbA1c
C-peptide
Panc. weight
Insulin content
Basal insulin
16.7 G
16.7 G 1st
16.7 G 2nd
16.7 G + IBMX
1.7 G + Epi
KCl
Glucagon content
Basal glucagon
16.7 G
16.7 G + IBMX
1.7 G + Epi
KCl
INS AUC
GCG AUC
* Adjusted p-value <0.05; |r|>0.3
Pearson correlation (r)
n
Islets from ND (n=7) or T2D (n=8) donors
n=4-8 mice/donor
Glucose/arginine stimulation
Normoglycemic environment
0 2 4 6 weeks
0' 6-hour fasted
15' 15 minutes after stimulation
o
Blood glucose (mg/dL)
Human insulin (ng/ml)
Ratio of Insulin to glucose (ng/mg)
****
**
*
ns
0' 15' 0' 15'
ND
T2D
p
FACS
HPi1+ HPa3+
Hpi1+ CD39L3+
Islets
n=13 ND n=7 T2D
β cells n=15 ND n=10 T2D
α cells n=9 ND n=9 T2D
Bulk RNA sequencing
Normalization through latent variable analysis
Differential expression
Gene network analyses
Figures 3-4
At FC ≥|1.5|, FDR<0.01:
β cells 218 134
α cells 104 144
Islets 323 242
q
Islets
α cells
β cells
Genes arranged along arc
Unique to list
Shared between list(s)
Inner curves
Genes shared between lists
Unique genes belong to same ontology term
r
Negative regulation of cell cycle
Regulation of anatomical structure size
Endoplasmic reticulum unfolded protein response
Protein digestion and absorption
Regulation of cytokine production
IL-17 signaling pathway
Circadian regulation of gene expression
I-κB kinase/ NF-κB signaling
Protein glycosylation
Peptide hormone secretion
Cellular response to lipid
Metabolism of lipids
Nucleoside phosphate metabolic process
Microtubule cytoskeleton organization
Organophosphate catabolic process
Cilium movement
Lipid catabolic process
Cilium organization
Cobalamin (vitamin B12) transport and metabolism
Pyruvate metabolism
p-value:
% Genes from:
β cells
α cells
Islets

for RNA-sequencing (Figure 3.1p and Figure 3.2v). Studying sorted β and α cells together with whole islets, which has not been done in prior studies, allowed detailed appreciation of both cell type-specific and islet-wide transcriptional changes in T2D. As collection of these rare tissues spanned more than 3.5 years, we used a latent variable analysis to discern biological variation from technical variation and then examined the datasets by both differential gene expression (Figure 3.1p and Figure 3.2w, Figure 3.3a-f) and gene network analyses. Differential expression analysis yielded 352, 248, and 564 differentially expressed genes in β cells, α cells, and whole islets, respectively (Figure 3.3g-i), highlighted by genes involved in stimulated insulin secretion in β cells (*G6PC2, GLP1R*) and changes in non-endocrine components in islets (*CXCL8, ADAMTS4*). Numerous metabolic and mitochondrial, exocytosis, ion transport and protein secretion pathways were enriched in T2D β cells (Figure 3.3j), while α cell gene changes were in amino acid and steroid signaling pathways and regulation of blood vessel morphology (Figure 3.2k). In T2D islets, cytokine signaling and immune terms were enriched, as were pathways related to ER processing and unfolded proteins (Figure 3.3l). These were less prominent in isolated α or β cells (Figure 3.3j-k). Despite diverse differentially expressed genes across sample types (Figure 3.1q),

Figure 3.2 *(following page)*: Additional metrics from functional and transcriptional profiling of islets from donors with short-duration T2D (related to Figure 3.1). (a-b) Matching of ND and T2D donor BMI (a) and age (b) for perifusion experiments. (c) Basal insulin secretion calculated as the average of the first three points of perifusion trace. (d-e) Integrated area under the curve (AUC) breaking down the total 16.7 mM glucose response into the first phase (d; through minute 24) and second phase (e; remainder of stimulation). (f) Area "under" the curve calculated from trace baseline for inhibition with low glucose and epinephrine. (g-l) Dynamic insulin secretion and metrics equivalent to Figure 3.1 but normalized by total insulin content. (m) Basal glucagon secretion calculated as average of first three points of perifusion trace. (n) Area "under" the curve calculated from trace baseline for inhibition with high glucose. (o-t) Dynamic glucagon secretion and metrics equivalent to Figure 3.1 but normalized by total glucagon content. (u) Blood glucose, human insulin levels, and human insulin:blood glucose ratio measured at 0' (six-hour fasted) and 15' after glucose and arginine stimulation of mice with human islet grafts. Symbols represent individual mice. * $p<0.05$, ** $p<0.01$, *** $p<0.001$, **** $p<0.0001$ (two-tailed t-test, panels a-f, h-n, and p-t; two-way ANOVA, panel u); error bars are SEM. (v) Gating strategy for sorted α and β cells identified by cell surface markers. Cell debris were excluded by forward scatter (FSC) and side scatter (SSC), single cells were identified by voltage pulse geometry (FSC-A v. FSC-H), and non-viable cells were excluded using propidium iodide (PI). Endocrine cell subpopulations were then gated based on positivity for HPi1 (pan-endocrine marker) and additional positivity for HPa3 (α cells) or NTPDase3 (β cells). (w) Select metrics used to assess library quality, organized by sample type. Outlier samples are highlighted in yellow and were excluded from downstream analyses.

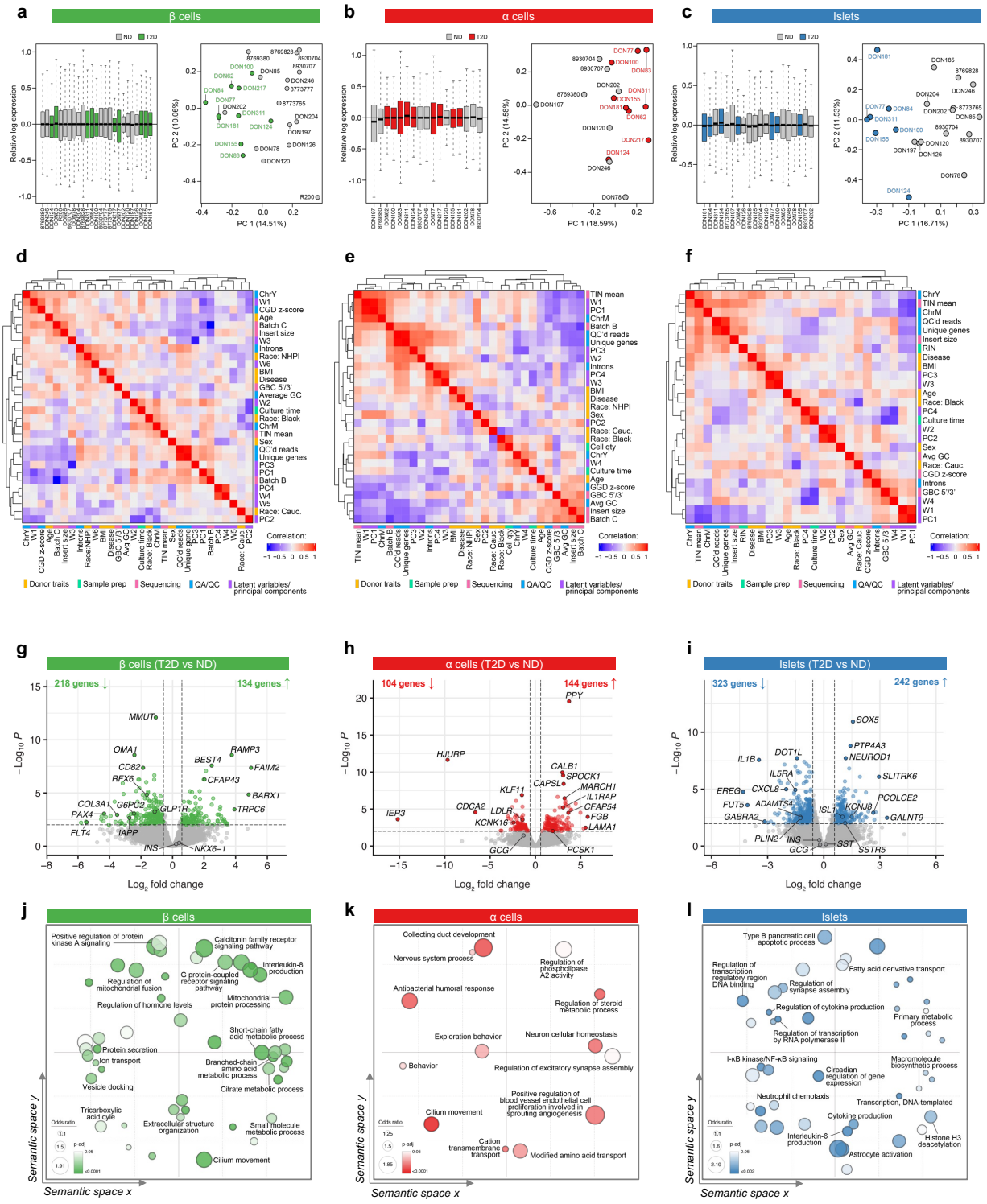

a
Donor BMI
BMI (kg/m²)
ND
T2D
b
Donor age
Age (years)
c
Basal insulin secretion
Insulin (ng/100 IEQs/min)
d
Glucose-stimulated 1st phase
AUC Insulin (ng/100 IEQs)
e
Glucose-stimulated 2nd phase
f
Low glucose + Epinephrine
g
G 16.7
G 16.7 + IBMX 100
G 1.7 + Epi 1
KCl 20
G 5.6
ND (n=21)
T2D (n=12)
Insulin secretion (% content/min)
Time (minutes)
h
Basal insulin
Insulin (% content/min)
i
Glucose-stimulated
AUC Insulin (% content)
j
cAMP-potentiated
k
Low glucose + Epinephrine
l
KCl-mediated
m
Basal glucagon secretion
Glucagon (pg/100 IEQs/min)
n
Glucose-stimulated
AUC Glucagon (pg/100 IEQs)
o
Glucagon secretion (% content/min)
p
Basal glucagon secretion
Glucagon (% content/min)
q
Glucose-stimulated
AUC Glucagon (% content)
r
cAMP-potentiated
s
Low glucose + Epinephrine
t
KCl-mediated
u
Blood glucose (mg/dL)
Human Insulin (ng/ml)
Ratio of Insulin/glucose (ng/mg)
0'
15'
v
ND (DON197)
T2D (DON84)
Islet cells
Single cells
Live cells
Endocrine cells
α and β cells
FSC-A
SSC-A
Propidium iodide
HPi1-BV421
HPa3-PE
w
Read pairs (mapping to unique genes)
5'/3' bias (high expression genes)
z-score of K-S statistic rel. to median
β cells
α cells
Islets
10 million
15 million
20 million
Samples
Normal
T2D
Excluded

there was considerable overlap at the level of biological pathways in which these genes are involved – among the most enriched across samples were hormone secretion, lipid metabolism, and cilia organization (Figure 3.1r). In sum, analysis of differential gene expression of sorted β and α cells and whole islets emphasizes common dysregulated pathways among sample types as well as cell-specific transcriptomic changes.

3.3.4 Short-duration T2D donors do not show significant changes in endocrine cell mass

To understand the context in which these functional and transcriptomic changes occur, we comprehensively evaluated the islet architecture in pancreatic tissue from T2D donors. High-throughput traditional immunohistochemistry (IHC) was applied across pancreas head, body, and tail regions for the entire donor cohort, and in parallel, a subset of samples was analyzed

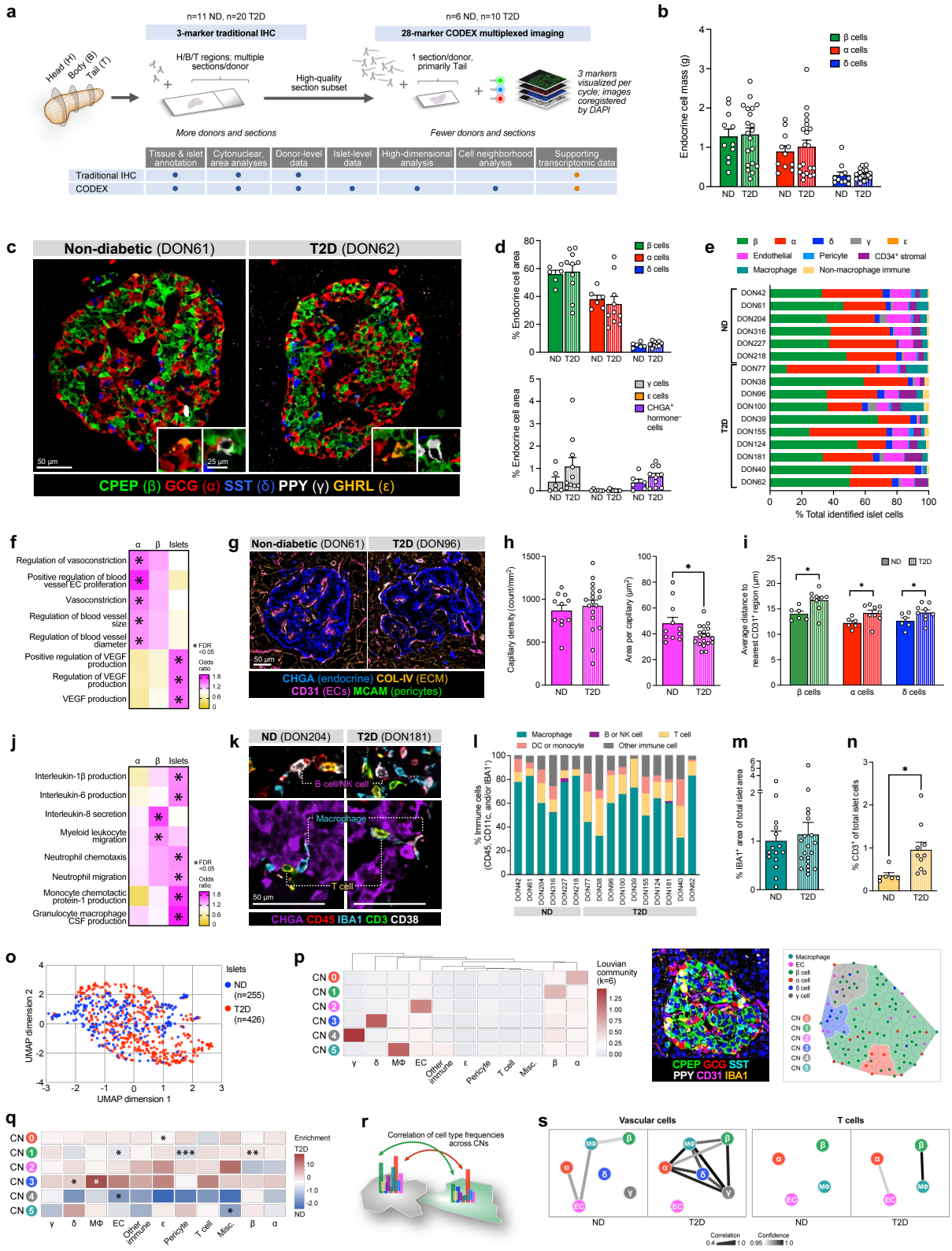
Figure 3.3 *(following page)*: Transcriptional analysis of islets and sorted α and β cells reveals dysregulation of metabolic pathways in T2D β cells and immune signaling in T2D islets (related to Figure 3.1). (a-c) Relative expression of individual libraries post-correction and principal component (PC) analysis of each sample type. RRIDs (donor labels beginning with '8') are abbreviated; see Extended Data Table 1 for complete alphanumeric RRIDs. Nondiabetic (ND) samples, grey; T2D samples, colored according to sample type. (d-f) Pearson correlation between sample covariates and PCs using the DEseq model. Colored bands next to row/column labels indicate whether variable is a donor trait (yellow), sample preparation variable (mint green), sequencing metric (pink), quality assurance/quality control (QA/QC) metric (blue), or latent variable or PC (purple). Culture time, duration of time (hours) between islet isolation and cell dispersion/sorting; Cell qty, number of sorted cells from which RNA was isolated (β and α cells only); RIN, RNA integrity number; Batch x, sequencing batch; TIN mean, mean transcript integrity number; Insert size, median length of sequenced RNA fragments; GBC 5'/3', ratio of gene body coverage at 5' and 3' end, describing reads distribution along a gene; QC'd reads, number of read pairs that pass initial filters; unique reads, number of read pairs that map to genomic area covering exactly one gene; Introns, reads mapping to intronic regions of genes; Avg GC, average GC content of all reads; CGD z-score, z-score quantifying cumulative gene diversity of libraries from median based on Kolmogorov Smirnov test; ChrM, reads mapping to MT chromosome; ChrY, reads mapping to Y chromosome; PCx, principal components; Wx, RUV-seq latent variables. (g-i) Volcano plots illustrating differentially expressed genes between ND and T2D β cells (g), α cells (h), and islets (i). Lines denote cutoffs for fold-change (±1.5) and significance (<0.01); genes passing both thresholds are colored and select genes are labeled. (j-l) Enriched gene ontology terms (FDR<0.05) obtained from RNA Enrich were condensed using the RelSim function of Revigo (similarity=0.5) and plotted in semantic space to emphasize relatedness. Dot size represents odds ratio and color represents p-value. Select terms are labeled.


a
β cells
b
α cells
c
Islets
ND
T2D
Relative log expression
PC 1 (14.51%)
PC 2 (10.06%)
PC 1 (18.59%)
PC 2 (14.58%)
PC 1 (16.71%)
PC 2 (11.53%)
d
e
f
Correlation:
Donor traits
Sample prep
Sequencing
QA/QC
Latent variables/ principal components
g
β cells (T2D vs ND)
218 genes ↓
134 genes ↑
h
α cells (T2D vs ND)
104 genes ↓
144 genes ↑
i
Islets (T2D vs ND)
323 genes ↓
242 genes ↑
Log2 fold change
j
β cells
k
α cells
l
Islets
Semantic space x
Semantic space y

with a 28-marker panel using co-detection by indexing (CODEX) (Figure 3.4a). This multiplexed technique for fluorescence-based imaging of large tissue sections without tissue destruction provided simultaneous visualization of multiple tissue compartments as well as spatially resolved cellular phenotypes defined by combined expression/exclusion of multiple markers (Figure 3.5a-b). Images are available in Pancreatlas (https://pancreatlas.org/datasets/904/explore) for reader exploration.

Because changes in endocrine cell number or ratio could explain the reduced insulin secretion in T2D islets, we first evaluated β, α, and δ cell populations. Multiple analyses across pancreas head, body, and tail, including evaluation of islet cell area and islet cell count within entire cross-sections, revealed that β and α cell mass in short-duration T2D were similar to controls (Figure 3.4b and Figure 3.5c-h), supporting the similar insulin content in the two groups of islets. We additionally assessed cell death and found apoptotic and/or necrotic cells to be exceedingly rare in both ND and T2D islets (data not shown). Donor-to-donor variability in β and α cell ratio was notable underscoring the challenge in working with heterogeneous human tissues. CODEX permitted simultaneous assessment of rarer γ and ε cell populations as well as identification of cells positive for chromogranin A (CHGA) but negative for all hormones, previously suggested to define "dedifferentiated" β cells [6, 47]. These cells were rare but present in both ND and T2D

Figure 3.4 *(following page)*: Integrated tissue analysis reveals no change to endocrine cell mass or number, but alteration in intraislet capillaries, T cells, and cellular neighborhoods in short-duration T2D cohort. (a) Schematic illustrating parallel analysis by traditional and multiplexed immunohistochemistry (IHC). (b) Mass of β, α, and δ cells in ND and T2D donors. (c) Representative images of islets from co-detection by indexing (CODEX) imaging; insets show γ and ε cells. (d) Cross-sectional area of endocrine cell types. (e) Relative proportions of islet endocrine, vascular, stromal, and immune cells. (f) Enrichment of vascular-related ontology terms in T2D transcriptome. (g) Representative images of islet capillaries, pericytes, and extracellular matrix (ECM). (h) Islet capillary density and area per capillary. (i) Spatial analysis of endocrine cells and islet capillaries. (j) Enrichment of immune-related ontology terms in T2D transcriptome. (k-l) Islet immune cell phenotypes and composition. (m-n) Islet macrophage (m) and T cell (n) abundance. (o) High-dimensional component analysis of islet cell composition per islet (n=255 ND, n=426 T2D). (p-s) Cellular neighborhood assignment (p) and corresponding cell composition and correlation changes in T2D vs. ND islets (q-s). Traditional IHC data: panels b, h, m; CODEX data: panels c-e, g, i, k-l, n-s. Symbols in bar graphs represent donors; * $p<0.05$ (two-tailed t-test, ND vs. T2D). RNA data: panels f, j; * FDR<0.05.


a
n=11 ND, n=20 T2D
3-marker traditional IHC
n=6 ND, n=10 T2D
28-marker CODEX multiplexed imaging
Head (H)
Body (B)
Tail (T)
H/B/T regions; multiple sections/donor
High-quality section subset
1 section/donor, primarily Tail
3 markers visualized per cycle; images coregistered by DAPI
More donors and sections
Fewer donors and sections
Tissue & islet annotation
Cytonuclear, area analyses
Donor-level data
Islet-level data
High-dimensional analysis
Cell neighborhood analysis
Supporting transcriptomic data
Traditional IHC
CODEX
b
Endocrine cell mass (g)
β cells
α cells
δ cells
ND T2D
c
Non-diabetic (DON61)
T2D (DON62)
50 µm
25 µm
CPEP (β) GCG (α) SST (δ) PPY (γ) GHRL (ε)
d
% Endocrine cell area
β cells
α cells
δ cells
γ cells
ε cells
CHGA+ hormone− cells
e
β
α
δ
γ
ε
Endothelial
Pericyte
CD34+ stromal
Macrophage
Non-macrophage immune
ND
T2D
DON42
DON61
DON204
DON316
DON227
DON218
DON77
DON38
DON96
DON100
DON39
DON155
DON124
DON181
DON40
DON62
% Total identified islet cells
f
α
β
Islets
Regulation of vasoconstriction
Positive regulation of blood vessel EC proliferation
Vasoconstriction
Regulation of blood vessel size
Regulation of blood vessel diameter
Positive regulation of VEGF production
Regulation of VEGF production
VEGF production
* FDR <0.05
Odds ratio
g
Non-diabetic (DON61)
T2D (DON96)
50 µm
CHGA (endocrine) COL-IV (ECM)
CD31 (ECs) MCAM (pericytes)
h
Capillary density (count/mm²)
Area per capillary (µm²)
ND
T2D
i
ND
T2D
Average distance to nearest CD31+ region (µm)
β cells
α cells
δ cells
j
α
β
Islets
Interleukin-1β production
Interleukin-6 production
Interleukin-8 secretion
Myeloid leukocyte migration
Neutrophil chemotaxis
Neutrophil migration
Monocyte chemotactic protein-1 production
Granulocyte macrophage CSF production
* FDR <0.05
Odds ratio
k
ND (DON204)
T2D (DON181)
B cell/NK cell
Macrophage
T cell
50 µm
CHGA CD45 IBA1 CD3 CD38
l
Macrophage
B or NK cell
T cell
DC or monocyte
Other immune cell
% Immune cells (CD45, CD11c, and/or IBA1+)
DON42
DON61
DON204
DON316
DON227
DON218
DON77
DON38
DON96
DON100
DON39
DON155
DON124
DON181
DON40
DON62
ND
T2D
m
% IBA1+ area of total islet area
ND
T2D
n
% CD3+ of total islet cells
ND
T2D
o
UMAP dimension 2
UMAP dimension 1
Islets
ND (n=255)
T2D (n=426)
p
CN 0
CN 1
CN 2
CN 3
CN 4
CN 5
Louvian community (k=6)
γ
δ
MΦ
EC
Other immune
ε
Pericyte
T cell
Misc.
β
α
CPEP GCG SST
PPY CD31 IBA1
Macrophage
EC
β cell
α cell
δ cell
γ cell
q
Enrichment
T2D
ND
r
Correlation of cell type frequencies across CNs
s
Vascular cells
T cells
ND
T2D
Correlation
Confidence

at similar proportions (Figure 3.4c-d and Figure 3.5i). Evidence of amyloid deposits, the abnormal buildup of β cell-produced islet amyloid polypeptide (IAPP) that manifests in T2D, was detectable in 75% of donors in this cohort but with variable prevalence and did not correlate to endocrine cell abundance or area (Figure 3.6a-b). Thus, tissue analysis suggests that changes in endocrine cell numbers, including β cell mass, are not a substantial component of short-duration T2D. Instead, these data point to reduction in β cell function as the predominant feature of this disease stage.

3.3.5 Reduced capillary size, increased T cell populations, and altered cellular neighborhoods highlight alterations in T2D islet microenvironment

Adequate islet vascularization and blood flow are critical for sensing and delivery of hormones to systemic circulation, so we next investigated islet capillary endothelial cells (ECs), the most abundant non-endocrine islet cell population (Figure 3.4e and Figure 3.5j). Pathway analysis from RNA-seq highlighted enrichment in T2D samples for processes controlling blood vessel size, particularly in α cells, as well as regulation of growth factors critical to islet capillary maintenance (Figure 3.4f and Figure 3.6c). Morphometric analysis demonstrated that capillary size, but not density, was reduced in T2D islets (Figure 3.4h-i), resulting in a greater distance of endocrine cells to the nearest capillary in T2D islets (Figure 3.4i). Interestingly, α and δ cells were closer to capillaries than β cells in both ND and T2D islets (Figure 3.6d), aligning with α cells expressing more angiogenic ligands and receptors than β cells (Figure 3.6e). Phenotypic markers CD34, a cell adhesion molecule that is prevalent in progenitor capillary ECs [87], and HLA-DR, a major histocompatibility class II (MHCII) receptor, were unchanged in T2D ECs (Figure 3.6f).

In addition to vasculature-related processes, transcriptional profiling also revealed enrichment in T2D β cells and islets for cytokine signaling and immune cell recruitment pathways (Figure 3.4j and Figure 3.6g). Macrophages, the largest population of intraislet immune cells, did not differ between ND and T2D based on either abundance or phenotypic classification by proinflamma-

tory (HLA-DR$^+$) or anti-inflammatory (CD163 and/or CD206$^+$) markers (Figure 3.4k-m and Figure 3.6h). T cells were rarer in the islet than macrophages but elevated in T2D islets across CD4$^+$ (helper), CD8$^+$ (cytotoxic), and CD4$^-$ CD8$^-$ (double negative) phenotypes (Figure 3.4n and Figure 3.6i). HLA-DR$^+$ T cells, previously observed in T2D islets [281], were not increased, though they were more abundant in a subset of T2D donors (Figure 3.6j). High dimensional data analysis using Uniform Manifold Approximation and Projection (UMAP) of all identified cell types within individually annotated islets revealed a high degree of overlap between islets from ND and T2D donors, emphasizing that although there are subtle differences, the overall islet composition is similar (Figure 3.4o).

Because analyses of islet composition did not consider the spatial organization of islet cells, we next applied two neighborhood analyses in parallel to annotated islet regions in an effort to identify differential cell architecture. A community detection algorithm tailored to islet cell frequencies, termed CF-IDF, categorized six different cellular neighborhoods (CNs), clusters of cells with distinct cell type compositions that were defined by the most enriched cell type (CN0-CN5; Figure 3.4p). A modified k-means clustering algorithm previously developed for CODEX data corroborated CN classifications (Figure 3.6k), and both approaches found similar CN distribution between ND and T2D islets (Figure 3.6l). ECs and pericytes were depleted in β CNs (CN1) of T2D islets (Figure 3.4q and Figure 3.6m), consistent with our findings of decreased proximity between β cells and ECs in T2D. In contrast, T2D β CNs had higher β cell enrichment than ND (Figure 3.4q).

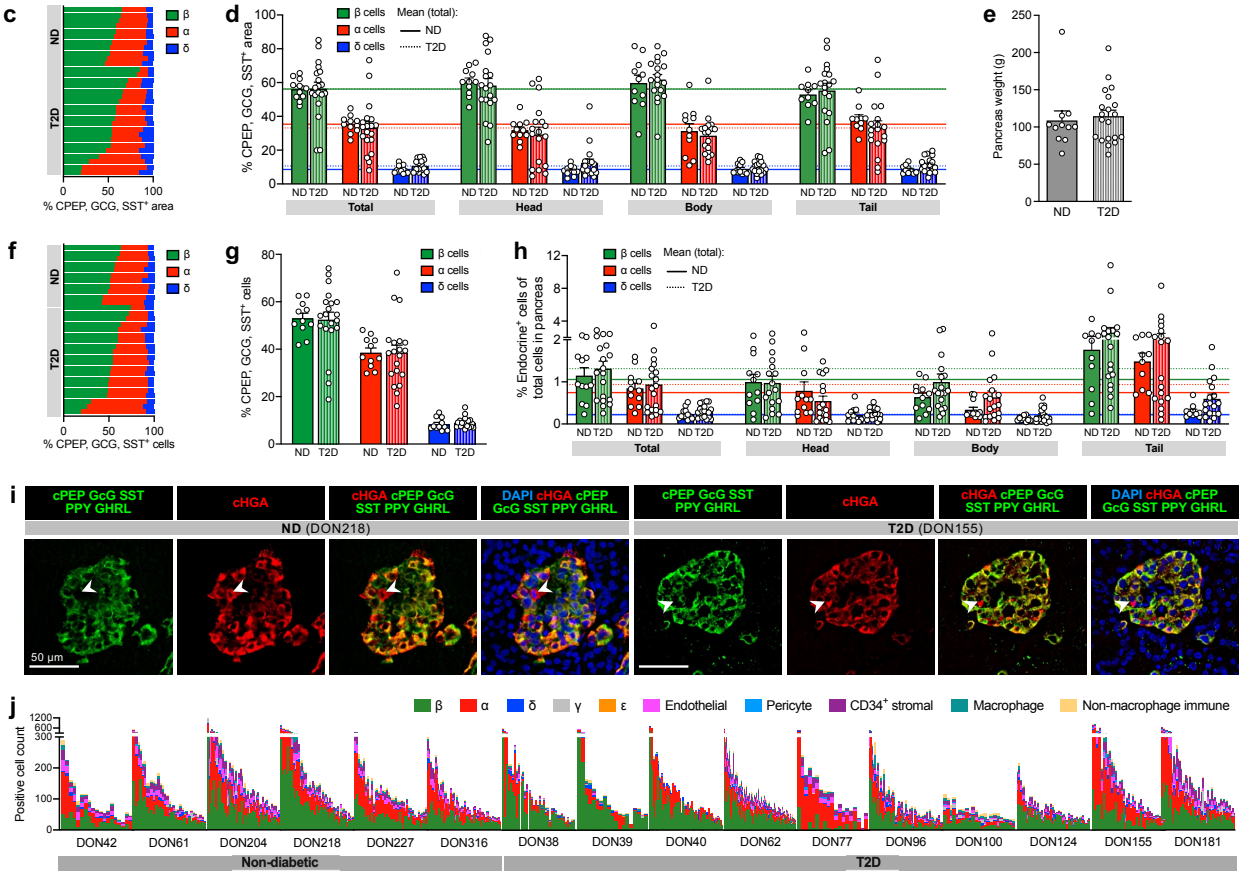
Figure 3.5 *(following page)*: Parallel approaches of multiplexed imaging and high-throughput traditional immunohistochemistry enable profiling of endocrine cells in addition to intraislet vascular and immune cells (related to Figure 3.4). Markers, cell populations, and specific phenotypes distinguished by the CODEX antibody panel. (c-h) Cross-sectional area (c-d) and cytonuclear quantification (f-h) of β cells (CPEP; green), α cells (GCG; red), and δ cells (SST; blue). Individual donor data shown in stacked bar graphs (c, f); bar graphs (d-e, g-h) show mean + SEM, one symbol per donor. Stratification by pancreas region (d, g) includes horizontal lines (solid, ND; dotted, T2D) for mean values from combined analysis ('Total'). (e) Pancreas weight measured during organ procurement; used to calculate endocrine cell mass in Figure 3.4b. (i) Representative images depicting rare cells positive for chromogranin A (CHGA; red) but negative for all hormones (green). Scale bars, 50 μm; arrowheads denote CHGA+ hormone− cells. (j) Abundance of endocrine and non-endocrine cells in ND and T2D islets; one vertical bar per islet and colored by cell type. Islets are grouped by donor and ordered from largest (highest total cell number) to smallest. See also Figure 3.4e.

a

Marker	Abbreviation	cell type(s)
α-Amylase	AMY2A	Acinar
β-Tubulin	TUBB3	Neuronal
Arginase i	ARG1	Macrophage phenotypic marker
c-peptide	cPEP	Endocrine (β)
cD11c	cD11c	Macrophages, monocytes, Dcs
cD14	cD14	Macrophages
cD163	cD163	Macrophage phenotypic marker
cD206	cD206	Macrophage phenotypic marker
cD3	cD3	t cells
cD31	cD31	Endothelial
cD34	cD34	HScs, vascular endothelial, fibroblasts
cD38	cD38	t cells, b cells, NK cells, myeloid cells, plasma cells, Dcs
cD4	cD4	t cell subset
cD45	cD45	Hematopoeitic (pan-leukocyte)
cD8	cD8	t cell subset
chromogranin a	cHGA	Endocrine
collagen IV	cOL-IV	EcM
E-cadherin	EcAD	Membrane (mostly ductal and acinar cells)
Ghrelin	GHRL	Endocrine (ε)
Glucagon	GcG	Endocrine (α)
HLA-DR	HLA-DR	Antigen-presenting cells
IBA1	IBA1	Macrophages
Ki67	Ki67	Proliferation marker
McAM (cD146)	McAM	Pericytes, other stromal
Pan-cytokeratin	KRT	Epithelial
Pancreatic polypeptide	PPY	Endocrine (γ)
Somatostatin	SST	Endocrine (δ)
Thioflavin s	ThioS	Amyloid deposits

b

	cell type	Phenotype	Inclusive markers	Exclusionary markers
Endocrine	**All endocrine**		**cHGA**	**cD31, cD45, IBA1, KRT**
	β cell		cPEP	GcG, SST, PPY, GHRL
	α cell		GcG	cPEP, SST, PPY, GHRL
	δ cell		SST	cPEP, GcG, PPY, GHRL
	γ cell		PPY	cPEP, GcG, SST, GHRL
	ε cell		GHRL	cPEP, GcG, SST, PPY
	cHGA$^+$ Hormone$^-$		cHGA	cPEP, GcG, SST, PPY, GHRL
Vascular	Endothelial cell (Ec)	**All Ec**	**cD31**	**cD45, IBA1**
		HLA-DR$^+$ Ec	cD31, HLA-DR	cD34
		cD34$^+$ Ec	cD31, cD34	HLA-DR
		HLA-DR$^+$ cD34$^+$ Ec	cD31, HLA-DR, cD34	
		HLA-DR$^-$ cD34$^-$ Ec	cD31	HLA-DR, cD34
	Pericyte		McAM	cD31, KRT
	cD34$^+$ stromal		cD34	cD31
Immune	**All immune**		**cD45/IBA1/cD11c**	**cD31**
	Non-macrophage immune		cD45/cD11c	IBA1
	b or natural killer (NK) cell		cD45, cD38	cD3, IBA1
	Dendritic cell (Dc) or monocyte		cD11c	IBA1, cD3, cD38
	t cell	**All t cell**	**cD45, cD3**	cD38, IBA1
		cD4$^+$ t cell	cD45, cD3, cD4	cD38, cD8, IBA1
		cD8$^+$ t cell	cD45, cD3, cD8	cD38, cD4, IBA1
		cD4$^-$ cD8$^-$ t cell	cD45, cD3	cD38, cD4, cD8, IBA1
	Macrophage (mac)	**All macrophage**	**IBA1**	**cD38, cD3**
		HLA-DR$^+$ mac	IBA1, HLA-DR	cD163, cD206
		cD163/cD206$^+$ mac	IBA1; cD163 or cD206	HLA-DR
		HLA-DR$^+$ cD163/206$^+$ mac	IBA1, HLA-DR; cD163 or cD206	
		HLA-DR$^-$ cD163/206$^-$ mac	IBA1	HLA-DR, cD163, cD206
	Other immune		cD45	cD3, cD38, IBA1, cD11c

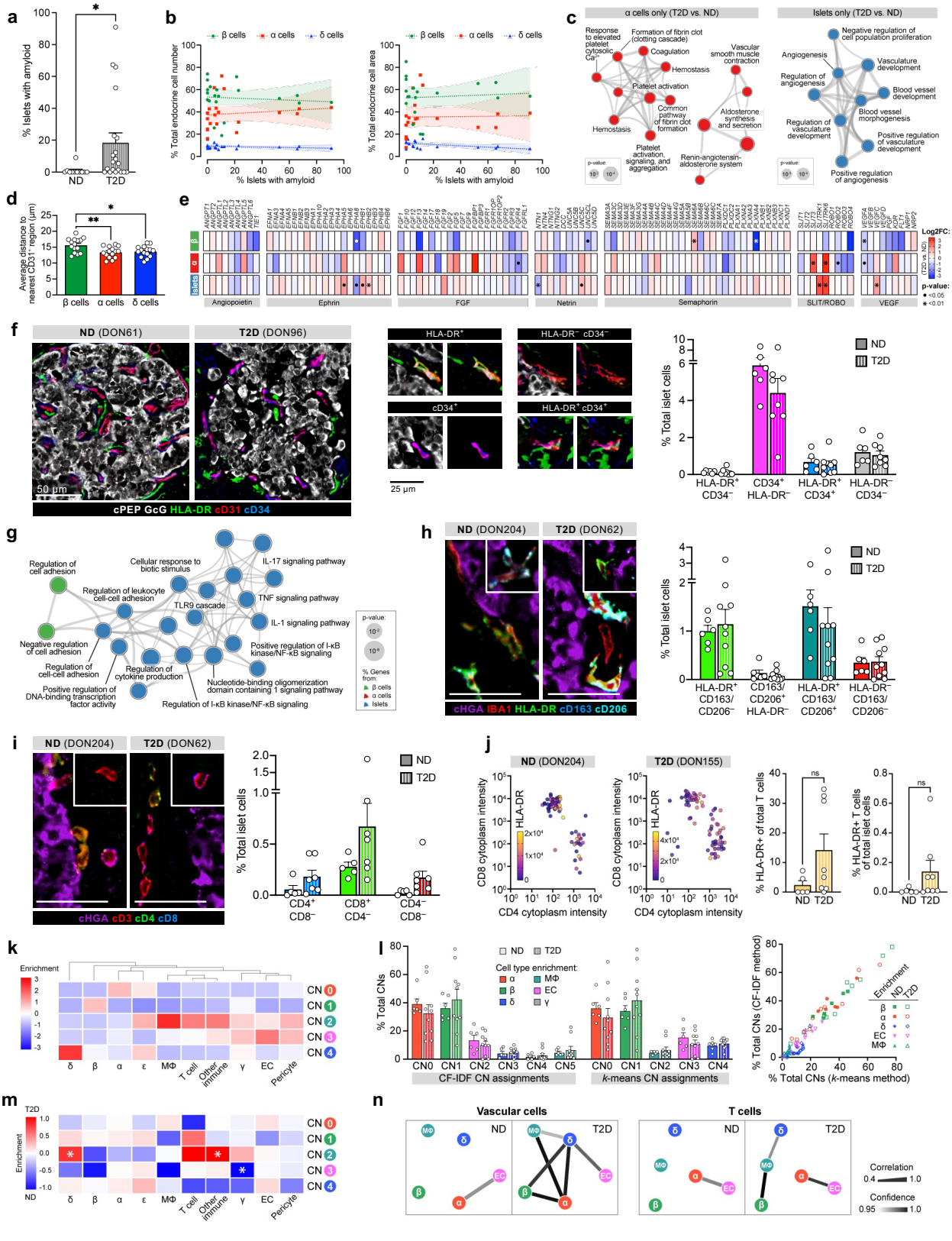


We also asked whether cell type frequencies correlated between CNs, i.e., if there was evidence for connectivity between spatially distinct regions (Figure 3.4r). Vascular cell frequencies were correlated between more CNs in T2D compared to ND islets, while T cell frequencies were specifically correlated between EC and α CNs as well as β and macrophage CNs in T2D (Figure 3.4s and Figure 3.6n), congruent with findings by islet RNA-seq that EC-specific and immune signals were upregulated in T2D. Together, these results demonstrate modest disruptions of islet organization by vascular and immune cells in early-stage T2D.

3.3.6 Co-expression network analyses identified gene modules related to donor and islet traits and revealed disrupted metabolism and cilia homeostasis in T2D

To understand the key gene networks that were contributing to β cell dysfunction in short-duration T2D, we performed weighted gene co-expression network analysis (WGCNA) on α cell (Figure 3.8d-f), β cell(Figure 3.8a-c), and islet samples (Figure 3.8g-i). This approach created mod-

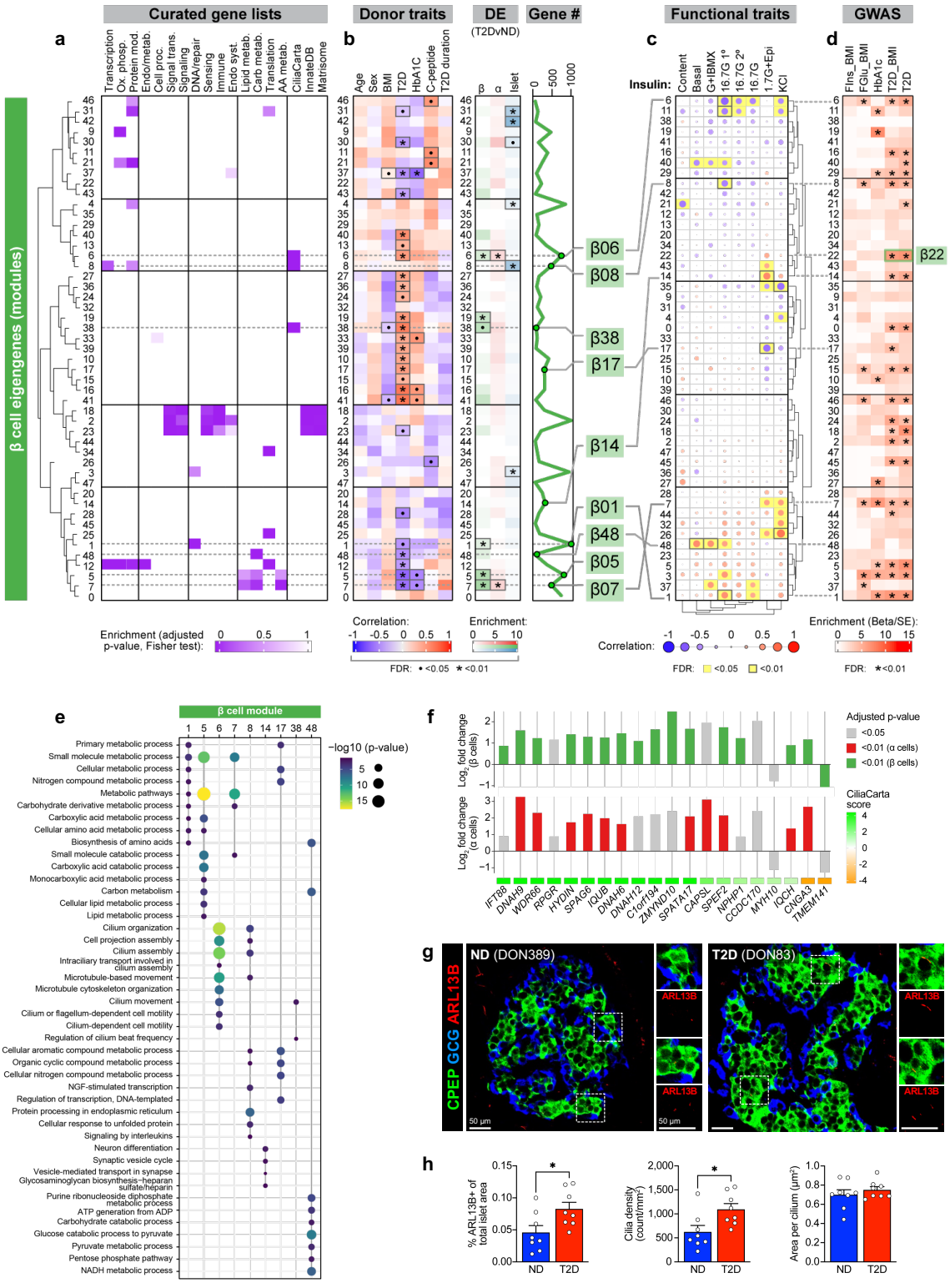
Figure 3.6 ***(following page)*****: Integration of multiplexed imaging and transcriptional profiling highlight disrupted capillaries and immune cells within T2D islets (related to Figure 3.4).** (a) Amyloid prevalence (% total islets with amyloid, averaged over multiple regions); * p<0.05 (two-tailed t-test). (b) Correlation of amyloid prevalence with β, α, and δ cell populations as percentage of total endocrine cell number or cross-sectional area; one symbol per donor with 95% confidence interval of linear regression (shading). No slopes were significantly nonzero at p<0.01 threshold. (c) Metascape visualization of select terms enriched for differentially expressed genes in T2D α cells (left) and islets (right). (d) Average distance of each endocrine cell type to nearest capillary; one symbol per donor (both ND and T2D); asterisks signify results of one-way ANOVA with Tukey's multiple comparisons test (** p<0.01; * p<0.05). (e) Gene expression fold-change of selected vascular and neuronal ligands and their receptors in β cells, α cells, and islets; • FDR<0.05; * FDR<0.01. (f) Phenotypes of endothelial cells (CD31; red) defined by single or dual positivity for HLA-DR (green) and CD34 (blue). Examples of each combination (HLA-DR+ CD34−, CD34+ HLA-DR−, HLA-DR+ CD34+, and HLA-DR− CD34−) are shown to right. (g) Magnification of select clusters depicted in Figure 3.1r (terms enriched across β, α, and islet samples). (h-i) Macrophages (IBA1+) and T cells (CD3+) phenotyped by various cell surface markers; insets show additional cells to illustrate phenotypic variety. Scale bars, 50 μm. (j) Expression of HLA-DR in CD4+ and CD8+ T cell populations. (k-l) Cellular neighborhood assignment and corresponding cell composition changes in T2D vs. ND islets. Panels k and m-n show results from the k-means method and panel l compares these results to CF-IDF method shown in Figure 3.4o-s. Traditional IHC data: panels a-b; CODEX data: panels d, f, h-n. RNA data: panels c, e, g.


a
% Islets with amyloid
ND
T2D
b
β cells
α cells
δ cells
% Total endocrine cell number
% Islets with amyloid
% Total endocrine cell area
c
α cells only (T2D vs. ND)
Islets only (T2D vs. ND)
Response to elevated platelet cytosolic Ca2+
Formation of fibrin clot (clotting cascade)
Coagulation
Hemostasis
Platelet activation
Common pathway of fibrin clot formation
Vascular smooth muscle contraction
Aldosterone synthesis and secretion
Renin-angiotensin-aldosterone system
Platelet activation, signaling, and aggregation
Negative regulation of cell population proliferation
Angiogenesis
Vasculature development
Regulation of angiogenesis
Blood vessel development
Blood vessel morphogenesis
Regulation of vasculature development
Positive regulation of vasculature development
Positive regulation of angiogenesis
p-value:
d
Average distance to nearest CD31+ region (μm)
β cells
α cells
δ cells
e
Angiopoietin
Ephrin
FGF
Netrin
Semaphorin
SLIT/ROBO
VEGF
Log2FC:
p-value:
f
ND (DON61)
T2D (DON96)
50 μm
cPEP GcG HLA-DR cD31 cD34
HLA-DR+
HLA-DR− cD34−
cD34+
HLA-DR+ cD34+
25 μm
% Total islet cells
ND
T2D
HLA-DR+ CD34−
CD34+ HLA-DR−
HLA-DR+ CD34+
HLA-DR− CD34−
g
Regulation of cell adhesion
Cellular response to biotic stimulus
IL-17 signaling pathway
Regulation of leukocyte cell-cell adhesion
TNF signaling pathway
TLR9 cascade
IL-1 signaling pathway
Negative regulation of cell adhesion
Positive regulation of I-κB kinase/NF-κB signaling
Regulation of cell-cell adhesion
Regulation of cytokine production
Nucleotide-binding oligomerization domain containing 1 signaling pathway
Positive regulation of DNA-binding transcription factor activity
Regulation of I-κB kinase/NF-κB signaling
p-value:
% Genes from:
β cells
α cells
Islets
h
ND (DON204)
T2D (DON62)
cHGA IBA1 HLA-DR cD163 cD206
% Total islet cells
ND
T2D
HLA-DR+ CD163/ CD206−
CD163/ CD206+ HLA-DR−
HLA-DR+ CD163/ CD206+
HLA-DR− CD163/ CD206−
i
ND (DON204)
T2D (DON62)
cHGA cD3 cD4 cD8
% Total islet cells
ND
T2D
CD4+ CD8−
CD8+ CD4−
CD4− CD8−
j
ND (DON204)
T2D (DON155)
CD8 cytoplasm intensity
HLA-DR
CD4 cytoplasm intensity
% HLA-DR+ of total T cells
% HLA-DR+ T cells of total islet cells
ns
ND T2D
k
Enrichment
CN 0
CN 1
CN 2
CN 3
CN 4
δ β α ε MΦ T cell Other immune γ EC Pericyte
l
% Total CNs
ND
T2D
Cell type enrichment:
α
β
δ
MΦ
EC
γ
CN0 CN1 CN2 CN3 CN4 CN5
CF-IDF CN assignments
CN0 CN1 CN2 CN3 CN4
k-means CN assignments
% Total CNs (CF-IDF method)
% Total CNs (k-means method)
Enrichment
ND T2D
m
T2D
ND
Enrichment
CN 0
CN 1
CN 2
CN 3
CN 4
δ β α ε MΦ T cell Other immune γ EC Pericyte
n
Vascular cells
T cells
ND
T2D
MΦ
δ
EC
β
α
Correlation
0.4 1.0
Confidence
0.95 1.0

ules ("eigengenes") of up to 2,000 genes each, labeled by sample type and numbered consecutively (β cells, modules β00-β48; α cells, α00-α54; islets, i00-i67). Collapsing the expression patterns across >14,000 genes into a smaller number of modules reduced gene-level multiple testing burden and enabled association of transcriptomic profiles with sample features including donor traits, islet functional parameters from the same donors defined by dynamic islet perifusion, and enrichment of open chromatin peaks to overlap GWAS variants (β cells: Figure 3.7a-e; α cells: Figure 3.9a-e; islets: Figure 3.9f-i). Modules with significant correlations were then queried, based on their member genes, for ontology terms to determine biological processes related to significant associations. Noteworthy observations are highlighted below, and results are available for further exploration online (https://theparkerlab.shinyapps.io/Islet-RNAseq-WGCNA/).

Several β cell modules were significantly (FDR < 5%) associated with whole-body glucose homeostasis (HbA1c), and some of these, such as β05 and β07, were also significantly enriched for genes differentially expressed in T2D β cells (Figure 3.7b). Both β05 and β07 contained genes related to carbohydrate, lipid, and amino acid metabolism (Figure 3.7a,e), with β07 significantly correlating with KCl-mediated insulin secretion (r=0.49, p=0.027; Figure 3.7c). Modules significantly positively correlated with glucose-stimulated insulin secretion (GSIS) included β01, β03, and β48, all enriched for metabolism-related processes, while β06 and β08, both enriched for cilium movement and motility, were significantly negatively correlated to GSIS (Figure 3.7c,e).

Figure 3.7 *(following page)*: Weighted Gene Co-expression Network Analysis (WGCNA) distinguishes β cell gene modules associated with donor and islet traits as well as those enriched in GWAS loci and identifies disruption in cilia processes as a conserved feature across sample types. (a) Relative enrichment of β cell module eigengenes for curated gene lists, based on genes present in each module. (b) Module correlation to donor characteristics, enrichment of differentially expressed (DE) genes, and total number of genes per module. • p<0.05; * p<0.01. Modules of interest highlighted (green). (c) Module correlation to β cell function described in Figure 3.1; significant associations highlighted (yellow). G+IBMX, 16.7 mM glucose with 100 μM isobutylmethylxanthine; 16.7G, 16.7 mM glucose; 16.7G 1°, first phase; 16.7G 2°, second phase; 1.7G+Epi, 1.7 mM glucose and 1 μM epinephrine; KCl, 20 mM potassium chloride. (d) Module enrichment for GWAS traits. FIns, fasting insulin; FGlu, fasting glucose. * FDR<0.01. (e) Enrichment of select gene ontology terms in β cell modules with notable correlations and/or enrichment. (f) Cilia-related genes with fold change ≥ |1.5| in both α and β cells in T2D. (g-h) Visualization by immunohistochemistry of cilia (ARL13B; red) and quantification of abundance, density, and size in ND and T2D tissue. * p<0.05 (two-tailed t-test).

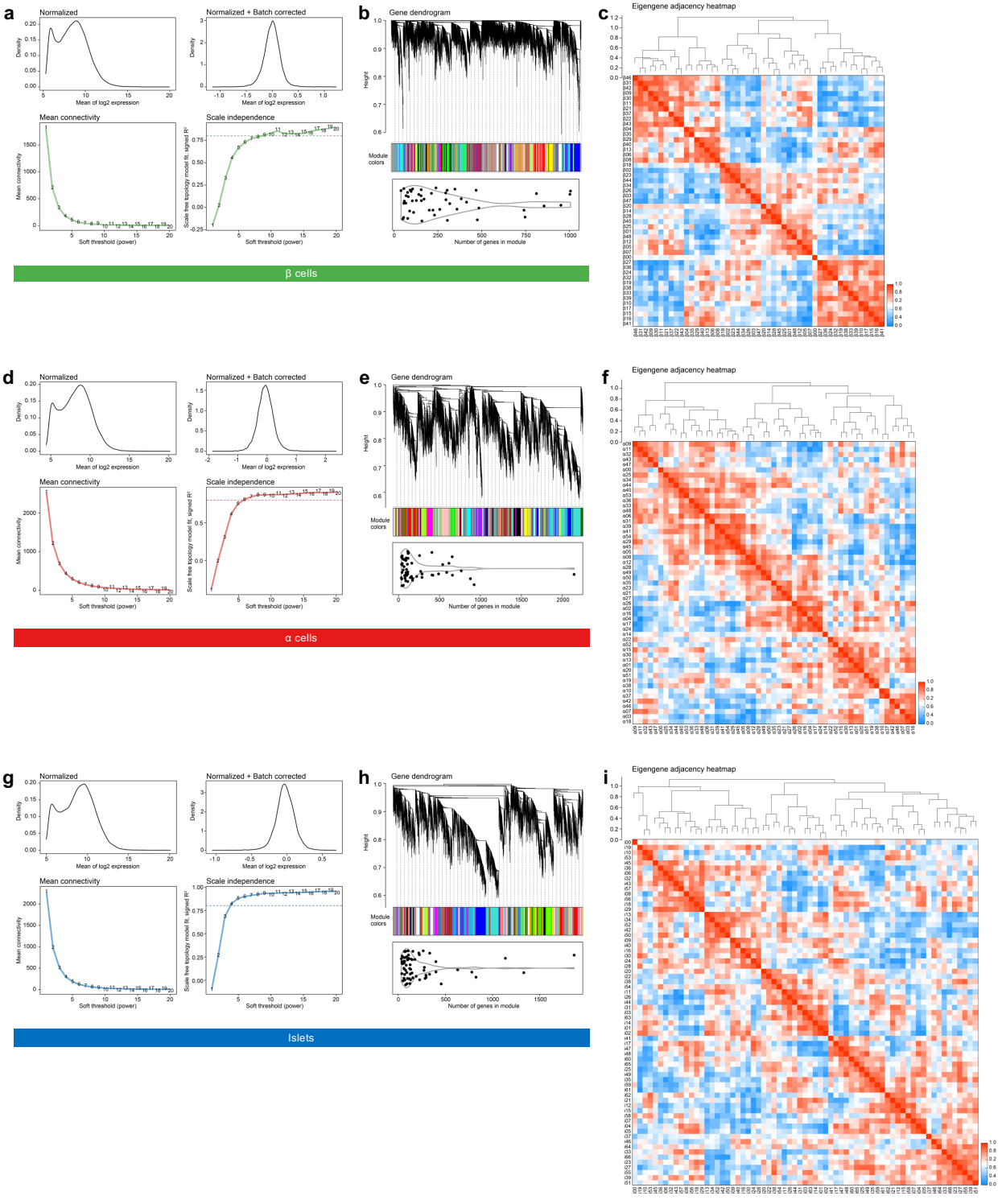

Curated gene lists
Donor traits
DE
(T2DvND)
Gene #
Functional traits
GWAS
a
b
c
d
β cell eigengenes (modules)
Transcription
Ox. phosp.
Protein mod.
Endo/metab.
Cell proc.
Signal trans.
Signaling
DNA/repair
Sensing
Immune
Endo syst.
Lipid metab.
Carb metab.
Translation
AA metab.
CiliaCarta
InnateDB
Matrisome
Age
Sex
BMI
T2D
HbA1C
C-peptide
T2D duration
β
α
Islet
Insulin:
Content
Basal
G+IBMX
16.7G 1°
16.7G 2°
16.7G
1.7G+Epi
KCl
FIns_BMI
FGlu_BMI
HbA1c
T2D_BMI
T2D
β06
β08
β38
β17
β14
β01
β48
β05
β07
β22
Enrichment (adjusted p-value, Fisher test):
Correlation:
Enrichment:
FDR: • <0.05 ★ <0.01
Enrichment (Beta/SE):
e
β cell module
−log10 (p-value)
Primary metabolic process
Small molecule metabolic process
Cellular metabolic process
Nitrogen compound metabolic process
Metabolic pathways
Carbohydrate derivative metabolic process
Carboxylic acid metabolic process
Cellular amino acid metabolic process
Biosynthesis of amino acids
Small molecule catabolic process
Carboxylic acid catabolic process
Monocarboxylic acid metabolic process
Carbon metabolism
Cellular lipid metabolic process
Lipid metabolic process
Cilium organization
Cell projection assembly
Cilium assembly
Intraciliary transport involved in cilium assembly
Microtubule-based movement
Microtubule cytoskeleton organization
Cilium movement
Cilium or flagellum-dependent cell motility
Cilium-dependent cell motility
Regulation of cilium beat frequency
Cellular aromatic compound metabolic process
Organic cyclic compound metabolic process
Cellular nitrogen compound metabolic process
NGF-stimulated transcription
Regulation of transcription, DNA-templated
Protein processing in endoplasmic reticulum
Cellular response to unfolded protein
Signaling by interleukins
Neuron differentiation
Synaptic vesicle cycle
Vesicle-mediated transport in synapse
Glycosaminoglycan biosynthesis−heparan sulfate/heparin
Purine ribonucleoside diphosphate metabolic process
ATP generation from ADP
Carbohydrate catabolic process
Glucose catabolic process to pyruvate
Pyruvate metabolic process
Pentose phosphate pathway
NADH metabolic process
f
Log2 fold change (β cells)
Log2 fold change (α cells)
Adjusted p-value
<0.05
<0.01 (α cells)
<0.01 (β cells)
CiliaCarta score
IFT88
DNAH9
WDR66
RPGR
HYDIN
SPAG6
IQUB
DNAH6
DNAH12
C1orf194
ZMYND10
SPATA17
CAPSL
SPEF2
NPHP1
CCDC170
MYH10
IQCH
CNGA3
TMEM141
g
CPEP GCG ARL13B
ND (DON389)
T2D (DON83)
ARL13B
50 µm
h
% ARL13B+ of total islet area
Cilia density (count/mm²)
Area per cilium (µm²)
ND
T2D

Importantly, aligning functional correlations with enrichment for GWAS loci (Figure 3.7d) enabled of modules that are more likely to be disease-causing (e.g., β01, β03) as opposed to those without GWAS enrichment (e.g., β48) that may instead represent disease-induced transcriptional changes. Thus, this approach allows linking of transcriptional profiles to islet physiological parameters and facilitates prioritization of signatures based on T2D genetic risk.

Though α cell modules showed weaker correlations to donor and functional traits than did β cells, several modules were significantly enriched for cilia-related genes and α08 was also enriched for α cell genes differentially expressed in T2D α cells (Figure 3.9a-b). Both α08 and α16 significantly inversely correlated with epinephrine-mediated glucagon secretion and were closely related across functional parameters (Figure 3.9c), with α08 showing significant enrichment for T2D GWAS variants (Figure 3.9d). In addition to genes enriched for cilia processes, α08 also included genes related to cytokine signaling and immune response (Figure 3.9e). Similarly, several islet modules showed notable enrichment for immune- and matrisome-related genes (Figure 3.9f); of these, i25 correlated positively with T2D status and inversely with basal insulin secretion and GSIS, while i26 correlated inversely with KCl-mediated insulin secretion (Figure 3.9g-h). Genes in both modules corresponded to cell-cell communication, including response to stimulus (i26) and leukocyte activation and migration (i25) (Figure 3.9i). Overall, these patterns suggest that β cell function may be influenced by α and other non-endocrine cells residing within the islet.

Interestingly, cilia-related processes not only defined key functionally correlated modules in every sample type, but they were also some of the most enriched pathways across all samples based on differential gene expression (Figure 3.9j). Further β06, β08, and α08 were enriched for T2D and related trait GWAS loci, suggesting a potential casual role (Figure 3.7d and Figure 3.9d). We compared fold change of validated cilia-related genes [61] and determined that the majority were expressed at higher levels in T2D compared to ND for both β and α cells (Figure 3.7f).

Figure 3.8 *(following page)*: Quality assessment of Weighted Gene Co- Expression Network Analysis (related to Figure 3.7). Analyses for β cell (a-c), α cell (d-f), and islet (g-i) datasets were conducted in parallel. Metrics are shown for batch correction and network parameter selection (a, d, g), module size and assignment (b, e, h), and module relatedness (c, f, i).

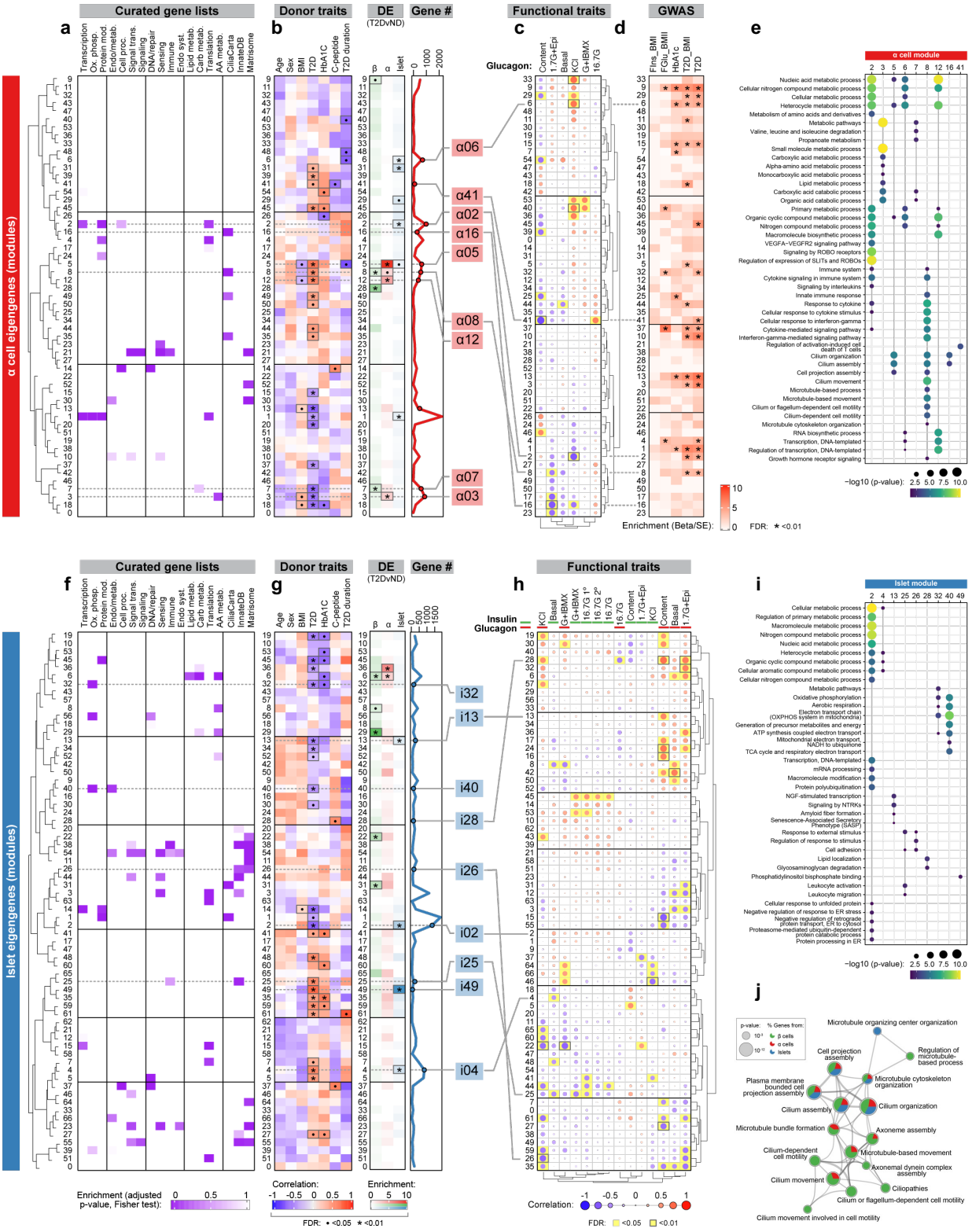

a
Normalized
Normalized + Batch corrected
Density
Mean of log2 expression
Mean connectivity
Mean connectivity
Soft threshold (power)
Scale independence
Scale free topology model fit, signed R²
Soft threshold (power)
b
Gene dendrogram
Height
Module colors
Number of genes in module
c
Eigengene adjacency heatmap
β cells
d
Normalized
Normalized + Batch corrected
Density
Mean of log2 expression
Mean connectivity
Mean connectivity
Soft threshold (power)
Scale independence
Scale free topology model fit, signed R²
Soft threshold (power)
e
Gene dendrogram
Height
Module colors
Number of genes in module
f
Eigengene adjacency heatmap
α cells
g
Normalized
Normalized + Batch corrected
Density
Mean of log2 expression
Mean connectivity
Mean connectivity
Soft threshold (power)
Scale independence
Scale free topology model fit, signed R²
Soft threshold (power)
h
Gene dendrogram
Height
Module colors
Number of genes in module
i
Eigengene adjacency heatmap
Islets

To investigate whether these changes translated to cellular alterations, we stained tissue sections from the same donors with cilia marker ARL13B (Figure 3.7g). Total cilia area within the islet was greater in T2D tissue, attributable to a higher cilia density with unchanged cilia size (Figure 3.7h), consistent with elevations in cilia transcripts. Thus, integration of functional, transcriptional, genetic, and tissue-based analyses highlights cilia-related processes as playing a key role in early T2D.

3.3.7 β cell hub gene RFX6 is reduced in T2D and controls glucose-stimulated insulin secretion

The network approach of WGCNA enables identification of "hub" genes that are highly connected, i.e., whose expression highly correlates with many other genes, both within and across modules, making it a powerful analysis to understand central transcriptional regulators that may be driving β cell dysfunction in short-duration T2D (Figure 3.10a). Of the highly connected β cell genes, *RFX6* stood out as a key islet-enriched transcription factor that has been linked to both monogenic and polygenic forms of diabetes [203, 241, 264] and thus is in prime position to exert disproportionate influence on the β cell transcriptional state. *RFX6* was more highly connected than other islet-enriched transcription factors specifically in β cells (Figure 3.10a-b and Figure 3.11a-d) and was one of the most reduced islet-enriched transcription factors at the tran-

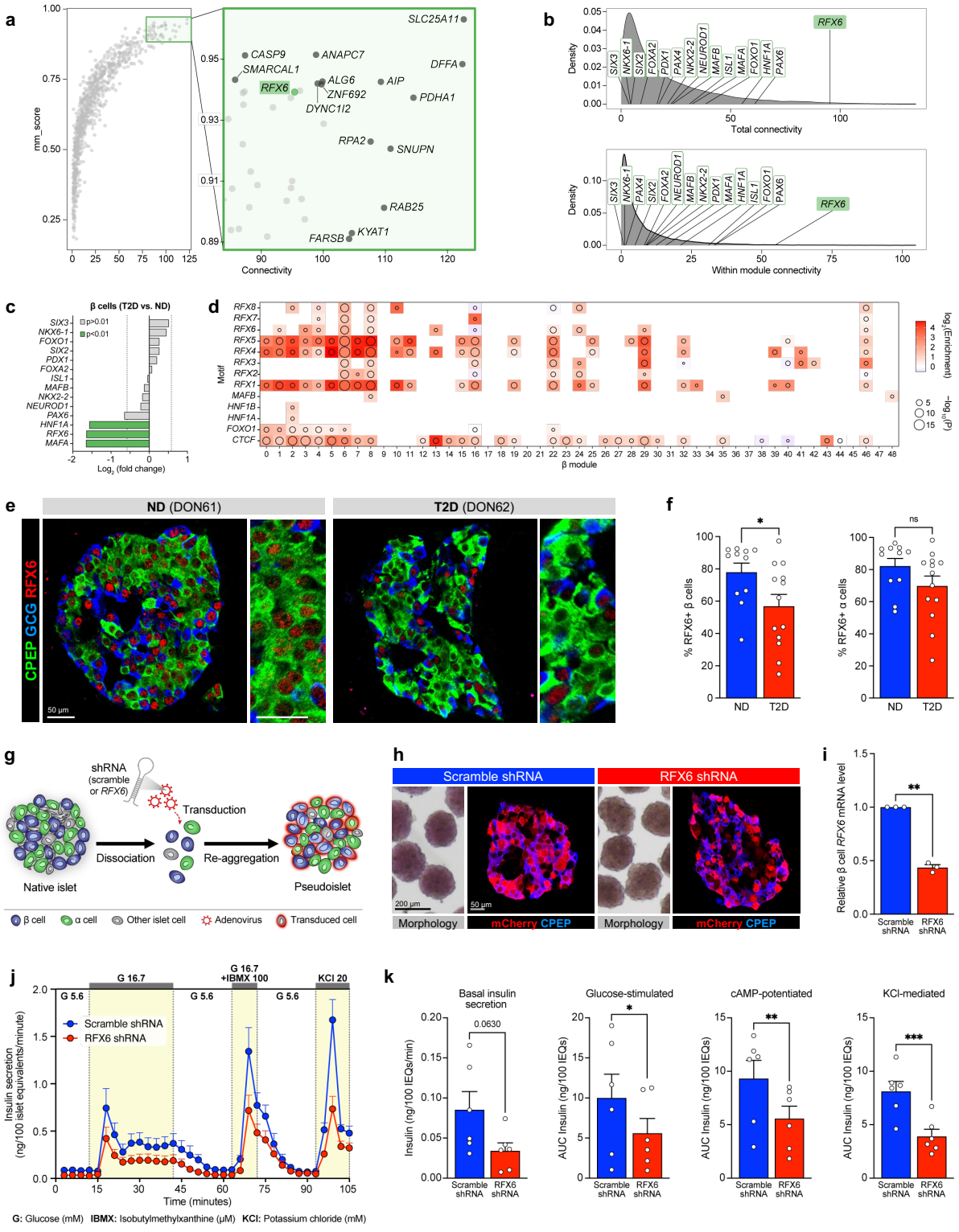
Figure 3.9 *(following page)*: WGCNA emphasizes α and islet cell gene modules associated with donor and islet traits as well as those enriched in GWAS loci (related to Figure 3.7). Module eigengenes for α cells (a-e) and islets (f-i) shown in parallel to β cells (Figure 3.7 a-e). (a, f) Modules clustered by similarity and showing relative enrichment of curated gene lists. (b, g) Module correlation to donor characteristics, enrichment of differentially expressed (DE) genes, and total number of genes per module. • p<0.05; * p<0.01. Modules of interest highlighted (b: red, g: blue). (c, h) Module correlation to α and β cell function (Figure 3.1); significant associations highlighted (yellow). For islets (g), modules were correlated to both insulin and glucagon secretion. G+IBMX, 16.7 mM glucose with 100 μM isobutylmethylxanthine; 16.7G, 16.7 mM glucose; 16.7G 1°, first phase; 16.7G 2°, second phase; 1.7G+Epi, 1.7 mM glucose and 1μM epinephrine; KCl, 20 mM potassium chloride. (d) Module enrichment for GWAS traits. FIns, fasting insulin; FGlu, fasting glucose. * FDR<0.01. (e, i) Enrichment of select gene ontology terms in β cell modules with notable correlations and/or enrichment. (j) Magnification of select clusters depicted in Figure 3.1r (terms enriched across β, α, and islet samples).


a
b
c
d
e
Curated gene lists
Donor traits
DE
(T2DvND)
Gene #
Functional traits
GWAS
α cell eigengenes (modules)
Transcription
Ox. phosp.
Protein mod.
Endo/metab.
Cell proc.
Signal trans.
Signaling
DNA/repair
Sensing
Immune
Endo syst.
Lipid metab.
Carb metab.
Translation
AA metab.
CiliaCarta
InnateDB
Matrisome
Age
Sex
BMI
T2D
HbA1C
C-peptide
T2D duration
β
α
Islet
Glucagon:
Content
1.7G+Epi
Basal
KCl
G+IBMX
16.7G
Fins_BMI
FGlu_BMI
HbA1c
T2D_BMI
T2D
α06
α41
α02
α16
α05
α08
α12
α07
α03
Enrichment (Beta/SE):
FDR: ★ <0.01
α cell module
Nucleic acid metabolic process
Cellular nitrogen compound metabolic process
Cellular metabolic process
Heterocycle metabolic process
Metabolism of amino acids and derivatives
Metabolic pathways
Valine, leucine and isoleucine degradation
Propanoate metabolism
Small molecule metabolic process
Carboxylic acid metabolic process
Alpha-amino acid metabolic process
Monocarboxylic acid metabolic process
Lipid metabolic process
Carboxylic acid catabolic process
Organic acid catabolic process
Primary metabolic process
Organic cyclic compound metabolic process
Nitrogen compound metabolic process
Macromolecule biosynthetic process
VEGFA−VEGFR2 signaling pathway
Signaling by ROBO receptors
Regulation of expression of SLITs and ROBOs
Immune system
Cytokine signaling in immune system
Signaling by interleukins
Innate immune response
Response to cytokine
Cellular response to cytokine stimulus
Cellular response to interferon-gamma
Cytokine-mediated signaling pathway
Interferon-gamma-mediated signaling pathway
Regulation of activation-induced cell death of T cells
Cilium organization
Cilium assembly
Cell projection assembly
Cilium movement
Microtubule-based process
Microtubule-based movement
Cilium or flagellum-dependent cell motility
Cilium-dependent cell motility
Microtubule cytoskeleton organization
RNA biosynthetic process
Transcription, DNA-templated
Regulation of transcription, DNA-templated
Growth hormone receptor signaling
−log10 (p-value):
f
g
h
i
j
Islet eigengenes (modules)
Insulin
Glucagon
i32
i13
i40
i28
i26
i02
i25
i49
i04
Enrichment (adjusted p-value, Fisher test):
Correlation:
FDR: • <0.05 ★ <0.01
Islet module
Cellular metabolic process
Regulation of primary metabolic process
Macromolecule metabolic process
Nitrogen compound metabolic process
Nucleic acid metabolic process
Heterocycle metabolic process
Organic cyclic compound metabolic process
Cellular aromatic compound metabolic process
Cellular nitrogen compound metabolic process
Metabolic pathways
Oxidative phosphorylation
Aerobic respiration
Electron transport chain (OXPHOS system in mitochondria)
Generation of precursor metabolites and energy
ATP synthesis coupled electron transport
Mitochondrial electron transport, NADH to ubiquinone
TCA cycle and respiratory electron transport
Transcription, DNA-templated
mRNA processing
Macromolecule modification
Protein polyubiquitination
NGF-stimulated transcription
Signaling by NTRKs
Amyloid fiber formation
Senescence-Associated Secretory Phenotype (SASP)
Response to external stimulus
Regulation of response to stimulus
Cell adhesion
Lipid localization
Glycosaminoglycan degradation
Phosphatidylinositol bisphosphate binding
Leukocyte activation
Leukocyte migration
Cellular response to unfolded protein
Negative regulation of response to ER stress
Negative regulation of retrograde protein transport, ER to cytosol
Proteasome-mediated ubiquitin-dependent protein catabolic process
Protein processing in ER
p-value:
% Genes from:
β cells
α cells
Islets
Microtubule organizing center organization
Cell projection assembly
Regulation of microtubule-based process
Plasma membrane bounded cell projection assembly
Microtubule cytoskeleton organization
Cilium assembly
Cilium organization
Microtubule bundle formation
Axoneme assembly
Cilium-dependent cell motility
Microtubule-based movement
Axonemal dynein complex assembly
Cilium movement
Ciliopathies
Cilium or flagellum-dependent cell motility
Cilium movement involved in cell motility

script level in T2D β cells (Figure 3.10c). Importantly, RFX6 is a member of module β01, which had the strongest positive association with high glucose-stimulated insulin secretion and was among the most significantly enriched for both GWAS variants and RFX binding motifs (Figure 3.7c-d and Figure 3.10d). Immunohistochemistry analysis revealed a reduction in number of β cells expressing RFX6 in T2D (Figure 3.10e-f). Together, these data support RFX6 as a critical hub gene in β cells that may contribute to the functional deficits observed in short-duration T2D.

To determine the role of RFX6 in adult human β cell function in an islet-like context, we used shRNA knockdown in a primary human pseudoislet system that allows for functional and transcriptomic assessment (Figure 3.10g). Scramble shRNA ('control') and RFX6 shRNA ('shRFX6') pseudoislets exhibited similar size and morphology, and preferential β cell transduction resulted in β cell RFX6 knockdown that did not change β or α cell proportion (Figure 3.10h-i and Figure 3.11e-g), suggesting that acute (6-day) reduction of RFX6 expression does not lead to β cell loss. Following RFX6 knockdown, dynamic insulin secretion in the presence of three secretagogues (high glucose, high glucose + IBMX, and KCl) was significantly blunted, similar to that seen in T2D islets (Figure 3.10j-k). Normalization to insulin content, which was greater in shRFX6 pseudoislets, made this secretory response even more prominent (Figure 3.11h-j). In sum, not only is RFX6 decreased in T2D β cells, but the results of targeted knockdown are consistent with the *RFX6*-containing module β01 association with glucose-stimulated insulin secretion (Figure 3.7d) and strongly implicate RFX6 as a major regulator of stimulated insulin secretion.

Figure 3.10 *(following page)*: RFX6, a central regulator of transcript changes in short-duration T2D, is reduced in T2D β cells and controls stimulated insulin secretion. (a-b) Overall connectivity (a) and cross- and within-module connectivity (b) of individual genes based on β cell WGCNA. Select genes with high connectivity scores (a) and select transcription factors (b) are labeled. (c) RNA fold change in T2D β cells of transcription factors highlighted in panel b. Vertical lines denote fold change = |1.5|. (d) Enrichment of transcription factor motifs in β cell modules. (e-f) Expression of RFX6 in β and α cells of ND and T2D donors. (g) Schematic of adenoviral shRNA delivery and formation of pseudoislets. (h) Morphology and immunofluorescent staining of transduced pseudoislets. (i) Relative RFX6 mRNA expression in β cells treated with scramble or RFX6 shRNA. (j) Pseudoislet insulin secretion assessed by perifusion; n=6 donors per group. (k) Area under the curve (AUC) for secretory response to each of the stimuli shown in panel j. Panels f, i, k: * p<0.05, ** p<0.01, *** p<0.001 (two-tailed t-test).

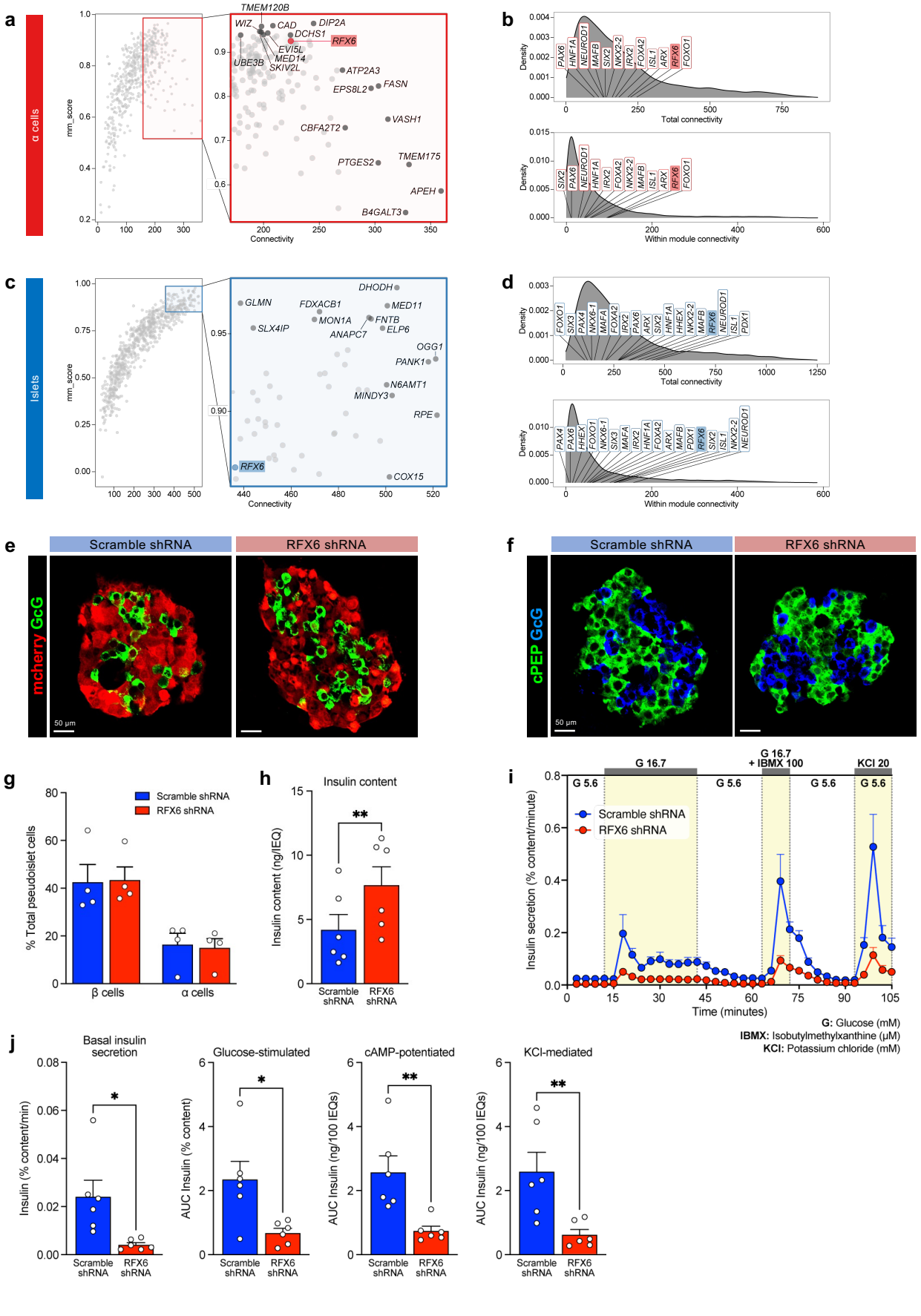

a
mm_score
Connectivity
SLC25A11
CASP9
ANAPC7
DFFA
SMARCAL1
RFX6
ALG6
AIP
ZNF692
PDHA1
DYNC1I2
RPA2
SNUPN
RAB25
FARSB
KYAT1
b
Density
Total connectivity
Within module connectivity
SIX3
NKX6-1
SIX2
FOXA2
PDX1
PAX4
NKX2-2
NEUROD1
MAFB
ISL1
MAFA
FOXO1
HNF1A
PAX6
RFX6
c
β cells (T2D vs. ND)
p>0.01
p<0.01
Log2 (fold change)
d
Motif
β module
log2(Enrichment)
−log10(P)
e
ND (DON61)
T2D (DON62)
CPEP GCG RFX6
50 µm
f
% RFX6+ β cells
% RFX6+ α cells
ND
T2D
*
ns
g
shRNA
(scramble
or RFX6)
Transduction
Dissociation
Re-aggregation
Native islet
Pseudoislet
β cell
α cell
Other islet cell
Adenovirus
Transduced cell
h
Scramble shRNA
RFX6 shRNA
200 µm
50 µm
Morphology
mCherry CPEP
i
Relative β cell RFX6 mRNA level
Scramble shRNA
RFX6 shRNA
**
j
G 16.7
G 16.7
+IBMX 100
KCl 20
G 5.6
Scramble shRNA
RFX6 shRNA
Insulin secretion
(ng/100 islet equivalents/minute)
Time (minutes)
G: Glucose (mM) IBMX: Isobutylmethylxanthine (µM) KCl: Potassium chloride (mM)
k
Basal insulin
secretion
Glucose-stimulated
cAMP-potentiated
KCl-mediated
Insulin (ng/100 IEQs/min)
AUC Insulin (ng/100 IEQs)
0.0630
*
**
***

3.3.8 RFX6 knockdown alters the β cell chromatin and transcriptional landscape and downregulates secretory vesicle components

To determine the molecular mechanism by which RFX6 knockdown impacted insulin secretion, shRFX6 and control pseudoislets (n=7 matched donors) were multiplexed using a blocked study design and processed for single nucleus multiome profiling (Figure 3.12a). Single nucleus (sn)RNA and snATAC reads were collected and filtered to yield 15,825 (RNA) and 5,706 (ATAC) high-quality nuclei for downstream analysis (Figure 3.13a). Islet cell types were resolved by clustering (Figure 3.12b-c and Figure 3.13b) where we found representation of all major cell types across all donors (Figure 3.13c) and equal distribution between shRFX6 and control constructs (Figure 3.12d). Consistent with the previously observed preferential adenoviral targeting of β relative to α cells, fluorescent reporter expression was much higher in β cell nuclei than in α cell nuclei (Figure 3.12e). Data are available via the UCSC Cell Browser at https://theparkerlab.med.umich.edu/data/public/cellbrowser/?ds=Pseudoislet10XMultiome for further exploration.

Supporting the role of RFX6 as a major β cell regulator, 13% of total detected genes were differentially expressed in β cell nuclei compared with <3% in other cell types (Figure 3.12f). Nuclear *RFX6* was not among those reduced, consistent with shRNA silencing occurring in the cytoplasm. Differentially expressed genes included those encoding cytoskeletal and scaffold/adaptor proteins (11% of those classified), membrane traffic proteins (4%), and gene-specific transcriptional regulator or chromatin/chromatin-binding or -regulatory proteins (13%) (Figure 3.12g). Upreg-

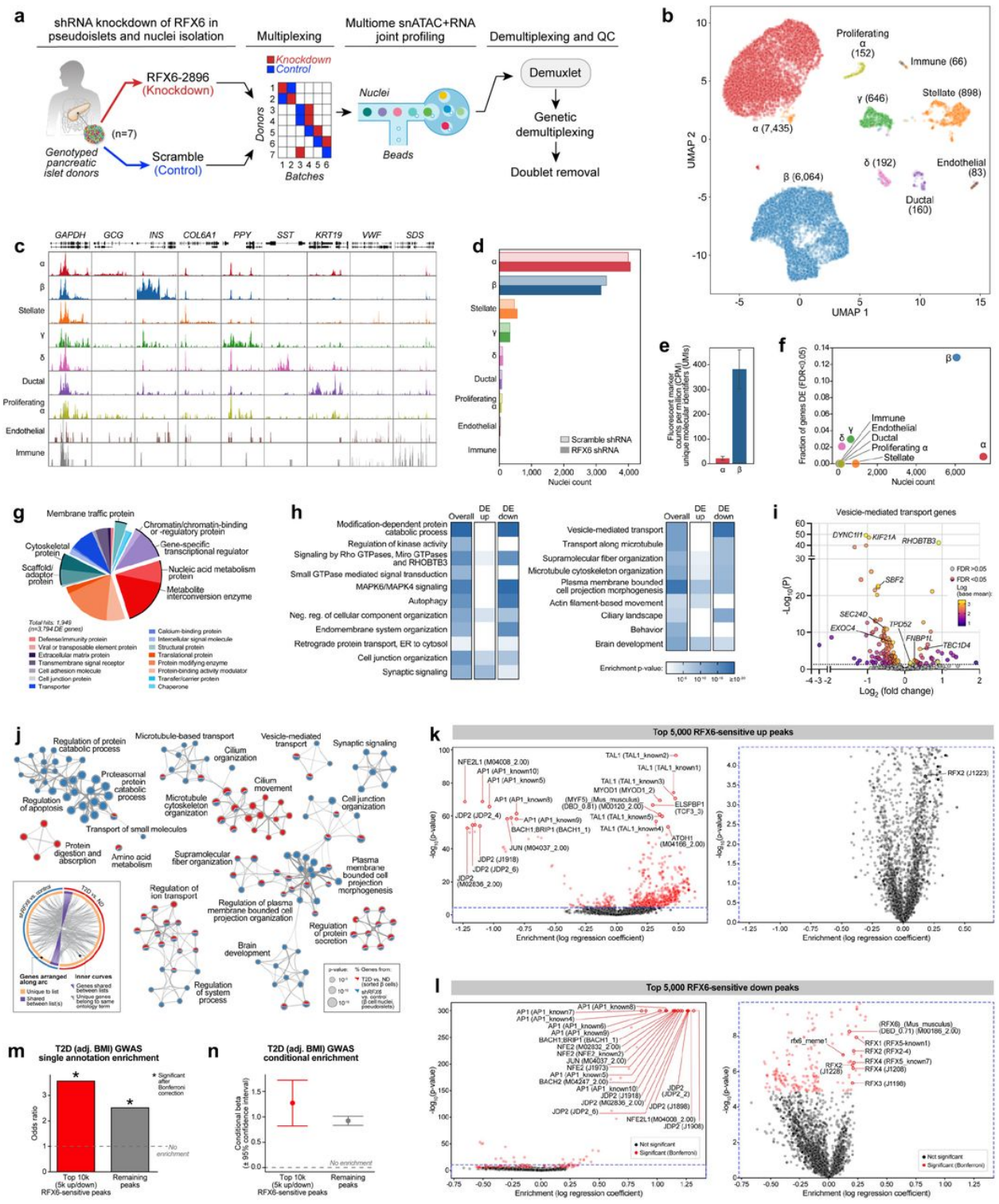
Figure 3.11 *(following page)*: Connectivity of *RFX6* by WGCNA is β cell-specific and RFX6 reduction impairs insulin secretion (related to Figure 3.10). (a-d) Connectivity of genes in α cell (a-b) and islet (c-d) modules, in parallel to data for β cell modules in Figure 3.10a-b. (a, c) Overall connectivity of individual genes; select genes with high connectivity scores are labeled. (b, d) Cross- and within-module connectivity; select transcription factors are labeled. (e-f) Immunofluorescent staining of pseudoislets embedded in type I collagen. (e) Transduced α cells marked by mCherry; see Figure 3.10h for β cells. (f) Distribution of β cells (CPEP; green) and α cells (GCG; blue). (g) Quantification of % β and % α cells in control (scramble) and shRFX6 pseudoislets. (h) Insulin content in control and shRFX6 pseudoislets (** p<0.01, two-tailed t-test). (i-j) Dynamic insulin secretion and metrics equivalent to Figure 3.10j-k but normalized by total insulin content.


a
α cells
mm_score
Connectivity
TMEM120B
WIZ
CAD
DIP2A
DCHS1
RFX6
EVI5L
MED14
SKIV2L
UBE3B
ATP2A3
EPS8L2
FASN
VASH1
CBFA2T2
PTGES2
TMEM175
APEH
B4GALT3
b
Density
Total connectivity
Within module connectivity
PAX6
HNF1A
NEUROD1
MAFB
SIX2
NKX2-2
IRX2
FOXA2
ISL1
ARX
RFX6
FOXO1
c
Islets
GLMN
FDXACB1
DHODH
MED11
MON1A
FNTB
SLX4IP
ANAPC7
ELP6
OGG1
PANK1
N6AMT1
MINDY3
RPE
COX15
d
FOXO1
SIX3
PAX4
NKX6-1
MAFA
FOXA2
IRX2
PAX6
ARX
SIX2
HNF1A
HHEX
NKX2-2
MAFB
RFX6
NEUROD1
ISL1
PDX1
e
Scramble shRNA
RFX6 shRNA
mcherry GcG
50 μm
f
cPEP GcG
g
Scramble shRNA
RFX6 shRNA
% Total pseudoislet cells
β cells
α cells
h
Insulin content
Insulin content (ng/IEQ)
i
G 16.7
G 16.7 + IBMX 100
KCl 20
G 5.6
Insulin secretion (% content/minute)
Time (minutes)
G: Glucose (mM)
IBMX: Isobutylmethylxanthine (μM)
KCl: Potassium chloride (mM)
j
Basal insulin secretion
Glucose-stimulated
cAMP-potentiated
KCl-mediated
Insulin (% content/min)
AUC Insulin (% content)
AUC Insulin (ng/100 IEQs)

ulated genes were enriched for actin filament-based movement and synaptic signaling, while downregulated genes were enriched for membrane trafficking, autophagy, and ciliary pathways (Figure 3.12h-i). To investigate overlap in differentially expressed genes between sh*RFX6* β cell nuclei and sorted T2D β cells, we compared the top 1,000 most significantly differential genes in each group and observed common pathway enrichment related to microtubule cytoskeleton organization, ion transport, and regulation of protein secretion (Figure 3.12j). Also of note, sh*RFX6* β cell nuclei differentially expressed genes were overrepresented in WGCNA module β22 (Figure 3.13d) that was enriched for T2D GWAS variants and RFX binding motifs. Genes in this module corresponded to cellular membrane and vesicle components, mirroring pathways dysregulated in sh*RFX6* β cell nuclei (Figure 3.13e) and further implicating exocytosis as a target of RFX6-mediated dysfunction in T2D β cells.

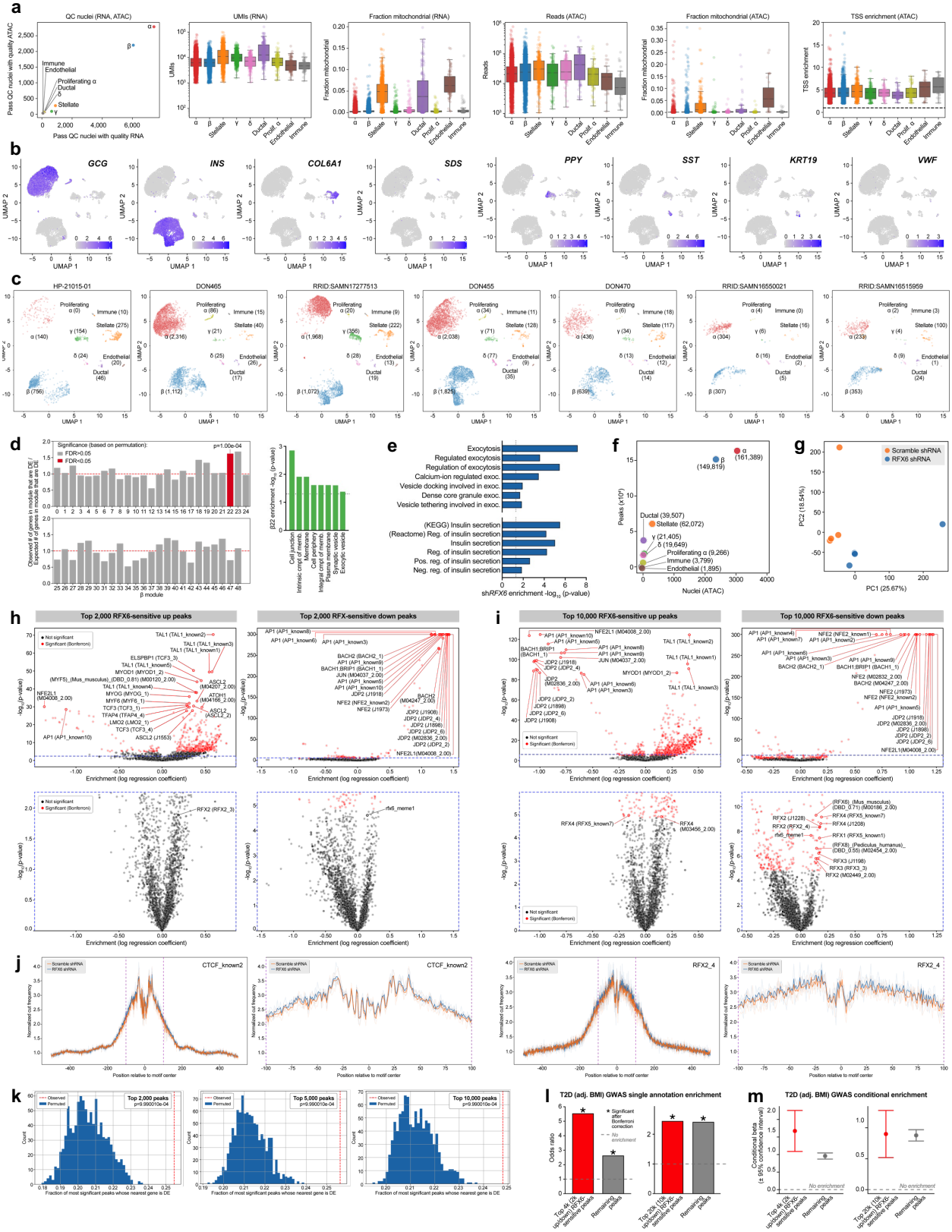
We next sought to identify the landscape of chromatin alterations in sh*RFX6* β cells and observed global changes compared to matched controls (Figure 3.13f-g). We took n=2,000-10,000 peaks with smallest p-values in either direction ('top RFX6-sensitive peaks') for use in downstream analyses. These peaks were significantly enriched for motifs corresponding to the known chromatin modifier activator protein 1 (AP1), as well as RFX6 and related family member motifs (Figure 3.12k-l and Figure 3.13h-i). CCCTC-binding factor (CTCF) and RFX motif footprint sig-

Figure 3.12 ***(following page)*****: RFX6 controls glucose-stimulated insulin secretion in human β cells through chromatin modifications and vesicle trafficking pathways.** (a) Schematic depicting randomized study design to mitigate batch effects in single nuclear (sn) RNA- and ATAC-sequencing of scramble shRNA (control) and RFX6 shRNA (shRFX6) pseudoislets. (b) Cell type assignment by clustering on RNA. (c). Pseudobulk ATAC signal at marker genes. (d) Post-QC nuclei counts from control and shRFX6 pseudoislets. (e) Abundance of fluorescent marker gene expression (mCherry/mKate2) in α and β cell nuclei. (f) Proportion of differentially expressed (DE) genes per cell type. (g) Classification of protein-coding DE genes in shRFX6 β cells by PANTHER. (h) Pathway enrichment for DE genes (FDR<0.01); second two columns separate genes up- or downregulated in shRFX6. (i) DE genes in Reactome pathway R-HSA-5653656. (j) Overlap of 1,000 most significant DE genes in shRFX6 vs. control β cell nuclei (blue) and T2D vs. ND sorted β cells (red), analyzed by Metascape. Circos plot illustrates overlap at the level of genes (purple) or ontology terms (grey). Network displays a subset of enriched terms, where edges denote term similarity and node colors represent contribution of each gene list. (k-l) Motif enrichment for top 5,000 RFX6-sensitive up-(k) and downregulated (l) ATAC peaks in shRFX6 β cell nuclei. Right panels show enlarged views of plots on left. (m-n) Odds ratio of T2D GWAS enrichment (m) and model estimate from conditional analysis (n) of RFX6-sensitive peaks.


a
shRNA knockdown of RFX6 in pseudoislets and nuclei isolation
Multiplexing
Multiome snATAC+RNA joint profiling
Demultiplexing and QC
Genotyped pancreatic islet donors
(n=7)
RFX6-2896 (Knockdown)
Scramble (Control)
Knockdown
Control
Donors
Batches
Nuclei
Beads
Demuxlet
Genetic demultiplexing
Doublet removal
b
UMAP 1
UMAP 2
α (7,435)
β (6,064)
γ (646)
δ (192)
Proliferating α (152)
Immune (66)
Stellate (898)
Endothelial (83)
Ductal (160)
c
GAPDH
GCG
INS
COL6A1
PPY
SST
KRT19
VWF
SDS
α
β
Stellate
γ
δ
Ductal
Proliferating α
Endothelial
Immune
d
Scramble shRNA
RFX6 shRNA
Nuclei count
e
Fluorescent marker counts per million (CPM) unique molecular identifiers (UMIs)
f
Fraction of genes DE (FDR<0.05)
Nuclei count
g
Membrane traffic protein
Chromatin/chromatin-binding or -regulatory protein
Gene-specific transcriptional regulator
Nucleic acid metabolism protein
Metabolite interconversion enzyme
Cytoskeletal protein
Scaffold/ adaptor protein
h
Overall
DE up
DE down
Modification-dependent protein catabolic process
Regulation of kinase activity
Signaling by Rho GTPases, Miro GTPases and RHOBTB3
Small GTPase mediated signal transduction
MAPK6/MAPK4 signaling
Autophagy
Neg. reg. of cellular component organization
Endomembrane system organization
Retrograde protein transport, ER to cytosol
Cell junction organization
Synaptic signaling
Vesicle-mediated transport
Transport along microtubule
Supramolecular fiber organization
Microtubule cytoskeleton organization
Plasma membrane bounded cell projection morphogenesis
Actin filament-based movement
Ciliary landscape
Behavior
Brain development
Enrichment p-value:
i
Vesicle-mediated transport genes
DYNC1I1
KIF21A
RHOBTB3
SBF2
SEC24D
EXOC4
TPD52
FNBP1L
TBC1D4
FDR >0.05
FDR <0.05
Log (base mean)
-Log10(P)
Log2 (fold change)
j
Regulation of protein catabolic process
Proteasomal protein catabolic process
Regulation of apoptosis
Transport of small molecules
Protein digestion and absorption
Amino acid metabolism
Microtubule-based transport
Cilium organization
Cilium movement
Microtubule cytoskeleton organization
Vesicle-mediated transport
Synaptic signaling
Cell junction organization
Supramolecular fiber organization
Plasma membrane bounded cell projection morphogenesis
Regulation of ion transport
Regulation of plasma membrane bounded cell projection organization
Regulation of protein secretion
Brain development
Regulation of system process
T2D vs. ND
shRFX6 vs. control
Genes arranged along arc
Inner curves
k
Top 5,000 RFX6-sensitive up peaks
NFE2L1 (M04008_2.00)
AP1 (AP1_known10)
AP1 (AP1_known5)
JDP2 (JDP2_4)
AP1 (AP1_known9)
BACH1:BRIP1 (BACH1_1)
JUN (M04037_2.00)
JDP2 (J1918)
JDP2 (JDP2_6)
JDP2 (M02836_2.00)
TAL1 (TAL1_known2)
TAL1 (TAL1_known1)
TAL1 (TAL1_known3)
MYOD1 (MYOD1_2)
AP1 (AP1_known8)
(MYF5)_(Mus_musculus) (DBD_0.81) (M00120_2.00)
ELSPBP1 (TCF3_3)
TAL1 (TAL1_known5)
TAL1 (TAL1_known4)
ATOH1 (M04166_2.00)
RFX2 (J1223)
-log10(p-value)
Enrichment (log regression coefficient)
l
Top 5,000 RFX6-sensitive down peaks
AP1 (AP1_known8)
AP1 (AP1_known7)
AP1 (AP1_known4)
AP1 (AP1_known6)
AP1 (AP1_known9)
BACH1:BRIP1 (BACH1_1)
NFE2 (M02932_2.00)
NFE2 (NFE2_known2)
JUN (M04037_2.00)
NFE2 (J1973)
AP1 (AP1_known3)
BACH2 (M04247_2.00)
AP1 (AP1_known10)
JDP2 (J1918)
JDP2 (M02836_2.00)
JDP2 (JDP2_6)
JDP2 (JDP2_2)
JDP2 (J1898)
NFE2L1(M04008_2.00)
JDP2 (J1908)
Not significant
Significant (Bonferroni)
(RFX6)_(Mus_musculus) (DBD_0.71) (M00186_2.00)
rfx6_meme1
RFX1 (RFX5-known1)
RFX2 (RFX2-4)
RFX4 (RFX5_known7)
RFX4 (J1208)
RFX3 (J1198)
RFX3 (J1228)
m
T2D (adj. BMI) GWAS
single annotation enrichment
Significant after Bonferroni correction
Odds ratio
No enrichment
Top 10k (5k up/down) RFX6-sensitive peaks
Remaining peaks
n
T2D (adj. BMI) GWAS
conditional enrichment
Conditional beta (± 95% confidence interval)
No enrichment
Top 10k (5k up/down) RFX6-sensitive peaks
Remaining peaks

natures like those previously observed in bulk islet ATAC data15 confirmed the high quality of the snATAC data (Figure 3.13j). Further, top RFX6-sensitive peaks were significantly enriched to occur near differentially expressed genes (Figure 3.13k), indicating concordance between the snATAC and snRNA modalities. We and others have shown that β cell ATAC peaks are enriched for T2D GWAS variants5,6, and indeed, top RFX6-sensitive peaks were also significantly enriched to overlap with these variants (Figure 3.12m and Figure 3.13l). Importantly, enrichment remained significant after conditional analysis controlled for remaining (not RFX6-sensitive) peaks (Figure 3.12n and Figure 3.13m), which emphasizes the importance of β cell *RFX6*-sensitive peaks in the genetic predisposition to T2D. Overall, these results show that knockdown of *RFX6* in β cells results in widespread transcriptional and chromatin changes that are associated with downregulated vesicle transport and coordinated disruption of regulatory elements that overlap T2D GWAS variants, consistent with the role of *RFX6* as a master regulator of β cell identity.

Figure 3.13 *(following page)*: Application of dual RNA and ATAC-sequencing to single nuclei from *RFX6* shRNA pseudoislets (related to Figure 3.12). (a) Quality control of nuclei for RNA and ATAC modalities. UMI, unique molecular identifier; TSS, transcription start site. (b) Expression of marker genes in cell type clusters. (c) Per-donor cell type counts. See also Figure 3.12b-c. (d) Enrichment of shRFX6 β cell nuclei differentially expressed genes within each β cell module derived from transcriptomes of sorted ND and T2D β cells (see Figure 3.7a-d). Right panel: cell component (cmpt) terms enriched for genes in β module 22. Memb., membrane. (e) Membership enrichment for exocytosis (exoc.) and insulin secretory pathways based on shRFX6 β cell nuclei differentially expressed (p<0.01) genes. All pathways are GO terms unless otherwise indicated. Neg., negative; pos., reg., regulation. (f) Per-cluster ATAC peaks (exact number listed in parentheses next to cell type). (g) PCA of pseudobulk β cell ATAC peak signal, each marker representing nuclei from a single donor/construct combination. (h-i) Motif enrichment for top 2,000 (h) or 10,000 (i) RFX6-sensitive up- and downregulated ATAC peaks in shRFX6 β cell nuclei. Motifs with highest significance are labeled in top panels; significant RFX motifs (or the single RFX motif closest to significance, in the case that no RFX motifs reach significance) are labeled in bottom panels. (j) ATAC footprints for CTCF_known2 and RFX2_4 motifs in β cell ATAC peaks. Light lines represent per-donor footprints; bold lines represent the average across donors. (k) Enrichment of top RFX6-sensitive up- and downregulated ATAC peaks (n=2,000, 5,000, or 10,000) in shRFX6 β cell nuclei near shRFX6 β cell differentially expressed genes. (l-m) Odds ratio of T2D GWAS enrichment (l) and model estimate from conditional analysis (m) of top 2,000 or 10,000 RFX6-sensitive peaks.


a
QC nuclei (RNA, ATAC)
UMIs (RNA)
Fraction mitochondrial (RNA)
Reads (ATAC)
Fraction mitochondrial (ATAC)
TSS enrichment (ATAC)
Pass QC nuclei with quality ATAC
Pass QC nuclei with quality RNA
UMIs
Fraction mitochondrial
Reads
TSS enrichment
Immune
Endothelial
Proliferating α
Ductal
Stellate
b
GCG
INS
COL6A1
SDS
PPY
SST
KRT19
VWF
UMAP 1
UMAP 2
c
HP-21015-01
DON465
RRID:SAMN17277513
DON455
DON470
RRID:SAMN16550021
RRID:SAMN16515959
d
Significance (based on permutation):
FDR>0.05
FDR<0.05
p=1.00e-04
Observed # of genes in module that are DE / Expected # of genes in module that are DE
β module
β22 enrichment -log10 (p-value)
Cell junction
Intrinsic cmpt of memb.
Membrane
Cell projection
Integral cmpt of memb.
Plasma membrane
Synaptic vesicle
Endocytic vesicle
e
Exocytosis
Regulated exocytosis
Regulation of exocytosis
Calcium-ion regulated exoc.
Vesicle docking involved in exoc.
Dense core granule exoc.
Vesicle tethering involved in exoc.
(KEGG) Insulin secretion
(Reactome) Reg. of insulin secretion
Insulin secretion
Reg. of insulin secretion
Pos. reg. of insulin secretion
Neg. reg. of insulin secretion
shRFX6 enrichment -log10 (p-value)
f
α (161,389)
β (149,819)
Ductal (39,507)
Stellate (62,072)
γ (21,405)
δ (19,549)
Proliferating α (9,266)
Immune (3,799)
Endothelial (1,895)
Peaks (x10⁴)
Nuclei (ATAC)
g
Scramble shRNA
RFX6 shRNA
PC2 (18.54%)
PC1 (25.67%)
h
Top 2,000 RFX6-sensitive up peaks
Top 2,000 RFX-sensitive down peaks
Not significant
Significant (Bonferroni)
Enrichment (log regression coefficient)
-log10(p-value)
i
Top 10,000 RFX6-sensitive up peaks
Top 10,000 RFX6-sensitive down peaks
j
Scramble shRNA
RFX6 shRNA
CTCF_known2
RFX2_4
Normalized cut frequency
Position relative to motif center
k
Observed
Permuted
Top 2,000 peaks
p=9.990010e-04
Top 5,000 peaks
p=9.990010e-04
Top 10,000 peaks
p=9.990010e-04
Count
Fraction of most significant peaks whose nearest gene is DE
l
T2D (adj. BMI) GWAS single annotation enrichment
Odds ratio
★ Significant after Bonferroni correction
No enrichment
Top 4k (2k up/down) RFX6-sensitive peaks
Remaining peaks
Top 20k (10k up/down) RFX6-sensitive peaks
m
T2D (adj. BMI) GWAS conditional enrichment
Conditional beta (± 95% confidence interval)
No enrichment

3.4 Discussion

The pancreatic β cell, a major focus in diabetes, exists within the multicellular pancreatic islet mini-organ, where interactions between various cell types are increasingly recognized. In T2D, like in other chronic, complex, multi-organ diseases, teasing apart the causes, correlates, and consequences of cellular and tissue dysfunction is challenging due to limited availability of primary tissue, constraints of sample processing at different disease stages, and in many cases, removal of cells from their native environment. To address these challenges and identify early disease-driving events, we applied a comprehensive, multimodal, integrated approach to isolated islets and pancreatic tissue from a unique cohort of short-duration T2D and control donors that included analyses of islet physiology, transcriptome, and pancreas tissue cellular architecture. Furthermore, we integrated donor and islet functional traits with gene network analysis and GWAS to understand central transcriptional regulators driving β cell dysfunction in short-duration T2D. Co-registration of multimodal data and clinical information yielded several important findings (Figure 4.1a): (1) impaired β cell function, a hallmark of early-stage T2D, persisted ex vivo and in nondiabetic environments; in contrast, α cell function was not changed; (2) islet endocrine composition was unchanged though there were modest alterations to the islet microenvironment in endothelial and immune cells; (3) transcriptional network analysis proportioned genetic risk into gene modules with specific functional properties, and (4) *RFX6* emerged as a highly connected hub transcription factor that was reduced in T2D β cells and associated with reduced glucose-stimulated insulin secretion. We validated a critical role for RFX6 by performing dynamic functional analyses and integrated snRNA and snATAC-seq on primary human pseudoislets with knockdown of *RFX6* in β cells. Reduction of RFX6 led to reduced insulin secretion defined by transcriptional dysregulation of vesicle trafficking, exocytosis, and ion transport pathways that was mediated by chromatin architectural changes overlapping with T2D GWAS variants (Figure 4.1b). Thus, our integrated, multimodal studies identify β cell dysfunction that results from cell-intrinsic defects, including an RFX6-mediated, T2D GWAS-enriched transcriptional network, as a key event in early T2D pathogenesis.

3.4.1 Dysfunction of β cells, and not β cell loss, is primary defect in early-stage T2D

This study demonstrates β cell functional defects ex vivo – which persist in culture and following transplantation into a normoglycemic environment – but no change to insulin content or β cell mass. The relative contributions of impaired β cell function and/or reduced β cell mass have long been debated in T2D [18, 127, 178]. Though postmortem studies suggest mild β cell loss [33, 216, 229, 230], most studies mixed short- and long-term disease duration together and noted that defects were more severe with longer duration and/or insulin treatment. Recent studies of metabolically profiled donors suggested that β cell loss is not prominent in early T2D [50, 276]. By integrating studies of both pancreatic tissue and isolated islets from the same donors, our data indicate that β cell loss is not a major component in disease pathogenesis at early-stage T2D. Further, the continued dysfunction of islets in a transplant setting also underscores the persistence of initial β cell defect. In sum, this study illustrates that β cell dysfunction occurs early in T2D and that prevention and/or rapid intervention may be critical to preserve β cell function.

3.4.2 Changes to islet microenvironment emphasize additional disease processes that may become more prominent in later disease stages

Our transcriptional analyses in isolated islets identified altered vascular and immune signaling as features in sorted α and β cells as well as in whole islets. Although isolated islets do not provide a physiologic context, particularly for endothelial cells without their connection to systemic circulation, similar transcriptional changes were found in laser capture microdissected T2D islets [276]. Further, our comprehensive tissue analyses of the same donors allowed in situ characterization of non-endocrine islet cell abundance, phenotype, and localization. We demonstrated that T2D islets had subtle reductions in islet capillary size, increased intraislet T cells, and altered communication between cellular neighborhoods, but overall the microenvironment was largely

similar to ND islets. While most donors showed some evidence of amyloid deposits as a unique feature of the T2D islet microenvironment, only a minority of islets demonstrated detectable amyloid at this stage of disease. Together, these observations are unlikely to explain the degree of β cell dysfunction in this cohort but, given that they are present without any associated changes in endocrine cell composition, may represent early consequences of β cell dysfunction or may act to exacerbate initial β cell-intrinsic defects. Indeed, inflammatory signals and other trophic factors have been shown to influence β cell function, especially in the presence of amyloid, and may become a more prominent feature of the disease at later stages [77, 171, 181, 275]. Further study is needed to determine whether changes to the microenvironment are truly an independent disease process or whether there is bidirectional signaling between dysfunctional β cells, α cells, and/or other islet cell types.

3.4.3 Integrated co-expression network analysis reveals gene modules of genetic risk in T2D

The transcriptomic profiles of sorted α and β cells in addition to islets provided new insight into cell-specific contributions to T2D pathogenesis. Co-expression network analysis and association with GWAS variants and physiological parameters, similar to a recent approach [227], allowed us to prioritize processes with physiological relevance that were more likely to be disease-causing rather than disease-induced. For instance, both β01 (metabolism-enriched) and β06 (cilia-enriched) modules are associated with T2D GWAS variants, indicating that regulatory circuitry related to metabolism and cilia function may have causative roles in development of T2D. Notably, insulin secretion was positively correlated to β01, whose genes were decreased in T2D β cells, but negatively correlated to β06, whose genes were increased in T2D β cells. These results suggest that β01 genes enhance insulin secretion while β06 genes decrease it, thus one expects that T2D risk alleles likely decrease β01 gene expression and activate β06 genes, both of which would negatively influence β cell function. Future work directly testing key candidate genes from this dataset, analogous to the studies of RFX6 described here, will be important to validate these

processes and how they contribute to T2D pathogenesis.

Genetic risk for complex metabolic diseases such as T2D results from the combined influence of many small-effect variants, with at-risk individuals likely having multiple parallel processes affected. This concept has been described as a "palette" model [174], and our work aids in deciphering components of the palette by proportioning genetic risk into cell-specific functional modules derived from transcriptome signatures across early stages of disease. Thus, this opens the opportunity to assess downstream consequences of an individual's innate genetic risk by identifying specific molecular and functional processes that would be most affected and hopefully allowing for precise targeting of those to achieve personalized medicine in diabetes.

3.4.4 RFX6 plays a central role in dysregulation of β cell function early in T2D

By identifying an RFX6 regulatory network that strongly correlates with insulin secretion and T2D genetic risk, this study provides new insight into previous work which has linked RFX6 to both monogenic and polygenic forms of diabetes [203, 241, 264]. Our results suggest that RFX6 exerts a disproportionate transcriptional influence on β cell state and that its dysregulation is a key molecular event in early T2D pathogenesis. We pursued this finding by directly testing the role of RFX6 in pseudoislets and demonstrated a clear function for RFX6 in governing stimulated insulin secretion in primary human β cells. Previous studies with direct perturbation of RFX6 in adult β cells, performed in cell lines and mouse models, highlighted downstream effects on Ca^{2+} and K_{ATP} channels [41, 205]. Our work confirms defective ion transport processes but identifies vesicle trafficking and exocytosis pathways as major drivers of defective insulin secretion in primary human β cells with impaired release likely responsible for the buildup of insulin content. Additionally, we show that these transcriptional changes are mediated by changes in β cell chromatin regions significantly overlapping with T2D GWAS loci, emphasizing the central role of RFX6 in mediating genetic risk to functional defects that define early T2D. Further, cilia-related genes were also significantly dysregulated following RFX6 reduction, in line with

evidence that the RFX family of transcription factors control ciliogenesis [46, 204]. Given their role in environment sensing, cell-to-cell communication, and signal transduction, cilia represent a potential link between β cell-intrinsic, RFX6-mediated dysregulation and changes within the islet microenvironment seen in early T2D and warrant future study.

This work raises important questions about what factors or events initially dysregulate RFX6 to start this cascade. Given the coordinating role RFX6 plays in islet cell development35, it may be that early defects driven by RFX6 dysfunction only become apparent after superimposed environmental, nutritional, and/or age-related stressors. Alternatively, the strong enrichment of T2D GWAS variants in β01 (the *RFX6*-containing module) and position of RFX6 as a hub gene may point to cumulative genetic effects compounding over time in an irreversible cascade that disrupts β cell homeostasis. Thus, precisely what underlies the initial dysregulation of RFX6, and whether it can be targeted to prevent or reverse early molecular defects in the β cell, should be an active area of investigation.

3.5 Materials and methods

3.5.1 Human subjects

Pancreata from nondiabetic (ND) (n=19) and T2D (n=20) donors were obtained through partnerships with the International Institute for Advancement of Medicine (IIAM), National Disease Research Interchange (NDRI), and local organ procurement organizations. Pancreata were processed in Pittsburgh by Dr. Rita Bottino for both islet isolation and histological analysis as previously described [15, 23, 57]. Additional ND human islet preparations (n=27) were obtained through partnerships with the Integrated Islet Distribution Program (IIDP) and Alberta Diabetes Institute (ADI) Isletcore and served as assay-specific controls or were used for pseudoislet studies. Deidentified medical histories provided both information for T2D staging as well as clinical characteristics to correlate with generated data. The Vanderbilt University Institutional Review Board declared that studies on de-identified human pancreatic specimens do not qualify as human

subject research.

Some human islets used in this research study were provided by the ADI IsletCore at the University of Alberta in Edmonton (http://www.bcell.org/adi-isletcore.html) with the assistance of the Human Organ Procurement and Exchange (HOPE) program, Trillium Gift of Life Network (TGLN), and other Canadian organ procurement organizations. Islet isolation was approved by the Human Research Ethics Board at the University of Alberta (Pro00013094). All donors' families gave informed consent for the use of pancreatic tissue in research. This study also used data from the Organ Procurement and Transplantation Network (OPTN) that was in part compiled from the Data Hub accessible to IIDP-affiliated investigators through IIDP portal (https://iidp.coh.org/secure/isletavail). The OPTN data system includes data on all donors, waitlisted candidates, and transplant recipients in the US, submitted by the members of the OPTN. The Health Resources and Services Administration (HRSA), U.S. Department of Health and Human Services provides oversight to the activities of the OPTN contractor. The data reported here have been supplied by UNOS as the contractor for the Organ Procurement and Transplantation Network (OPTN). The interpretation and reporting of these data are the responsibility of the authors and in no way should be seen as an official policy of or interpretation by the OPTN or the U.S. Government.

3.5.2 Pancreas procurement and processing

Pancreata from ND and T2D donors were received within 18 hours from cross clamp and maintained in cold preservation solution on ice until processing, as described previously [23]. Pancreas was then cleaned from connective tissue and fat, measured and weighed. Prior to islet isolation, multiple cross-sectional slices of pancreas with 2-3 mm thickness were obtained from the head, body and distal tail, further divided into quadrants, and processed into paraformaldehyde-fixed cryosections as described previously [220]. Islet isolation was performed via ductal collagenase infusion and purification by density gradient as described previously [15, 23], then shipped to Vanderbilt for further analysis following shipping protocols developed by the IIDP. Islets were

cultured in CMRL 1066 media (5.5 mM glucose, 10% FBS, 1% Pen/Strep, and 2 mM L-glutamine) in 5% CO2 at 37℃ for 24–48 hours prior as reported in previous studies [21, 23, 58]. Pseudoislets were cultured in Vanderbilt Pseudoislet media [269]. Limitations of tissue availability and processing dictated that not all assays could be performed on each donor.

3.5.3 Assessment of native pancreatic islet and pseudoislet function by macroperifusion

Function of islets from ND and T2D donors and pseudoislets were studied in a dynamic cell perifusion system at a perifusate flow rate of 1 mL/min [132, 269]. The effluent was collected at 3-minute intervals using an automatic fraction collector, then islets were retrieved and lysed with acid-ethanol solution to extract. Insulin and/or glucagon concentrations in each perifusion fraction, as well as total hormone content, were measured by radioimmunoassay (RIA) (human insulin, RI-13K, Millipore; glucagon, GL-32K, Millipore), enzyme-linked immunosorbent assay (ELISA) (Human insulin, 10-1132-01, Mercodia; glucagon, 10-1281-01, Mercodia), or Homogeneous Time Resolved Fluorescence (HTRF) assay (glucagon, 62CGLPEH, Cisbio). Area under the curve (AUC) above baseline hormone release was calculated with the trapezoidal method in GraphPad Prism 8.0 as previously described [269].

3.5.4 Human islet transplantation

Immunodeficient *NOD.Cg-Prkdc^{scid}Il2rg^{tm1Wjl}/Sz* (NSG) [9] 10-12-week old male mice were maintained by Vanderbilt Division of Animal Care in group housing in sterile containers within a pathogen-free barrier facility housed with a 12 hour light/12 hour dark cycle and access to free water and standard rodent chow. All animal procedures were approved by the Vanderbilt Institutional Animal Care and Use Committees. Between 1000-2000 islet equivalents per mouse (n=4-8 mice per islet preparation) were transplanted beneath the kidney capsule. After 6 weeks, mice were fasted for 6 hours and then injected with glucose + arginine (2g/kg body weight) intraperi-

toneally as previously described [21, 23, 58, 59]. Blood samples were obtained before (0') and after (15') injection and human-specific insulin was analyzed by ELISA (Alpco, 80-ISNHU-E01.1) or radioimmunoassay (Millipore, RI-13K).

3.5.5 Purification of α and β cells by FACS

Human islets from ND and T2D donors were dispersed and sorted for collection of RNA from α and β cells as described previously [23, 232]. Briefly, 0.025% trypsin was used to disperse islet cells by manual pipetting and subsequently quenched with RPMI containing 10% FBS. Cells were washed in the same medium and counted on a hemocytometer, then transferred to FACS buffer (2 mM EDTA, 2% FBS, 1X PBS). Indirect antibody labeling was completed via two sequential incubation periods at 4C, with one wash in the FACS buffer following each incubation. Primary and secondary antibodies have been characterized previously and used to isolate high-quality RNA [23, 73, 74, 232]. Appropriate single color compensation controls were run alongside samples. For sorting of β cells for use in pseudoislets, quenching step post-dispersion was performed with 100% FBS at 1/3 volume trypsin. Cells then underwent an additional filtration step using a 40 μl strainer prior to staining. For all preparations, propidium iodide (0.05 ug/100,000 cells; BD Biosciences, San Jose, CA) was added to samples prior to sorting for non-viable cell exclusion. Flow analysis was performed using an LSRFortessa cell analyzer (BD Biosciences, San Jose, CA), and a FACSAria III cell sorter (BD Biosciences, San Jose, CA) was used for FACS. Cells for RNA were collected into FACS buffer, washed once in 1X PBS, and stored in RNA lysis buffer for RNA extraction. Cells for pseudoislets were washed once in 1X PBS, resuspended in Vanderbilt pseudoislet media, and processed as described in Pseudoislet section below. Analysis of flow cytometry data was completed using FlowJo 10.1.5 (Tree Star, Ashland, OR).

3.5.6 Traditional and multiplexed immunohistochemical imaging and analysis

3.5.6.1 Traditional Immunohistochemistry

Multiple sections from pancreatic head, body, and tail regions of 20 T2D and 11 age-matched ND donors were lightly paraformaldehyde (PFA)-fixed and prepared for immunohistochemistry and stained as described previously [23, 104, 232]. Primary and secondary antibodies and their dilutions are listed in Extended Data Table 2 of [268]. Amyloid was visualized using a 2-minute incubation in Thioflavin S (0.5% w/v; #T-1892, Sigma, St. Louis, MO) followed by a brief wash in 70% ethanol as described previously [58, 59, 108]. Images were acquired at 20X with 2X digital zoom using a FV3000 confocal laser scanning microscope (Olympus) or a ScanScope FL (Aperio) and processed using cytonuclear algorithms (HighPlex FL v3.2.1) or tissue classifiers via HALO software (Indica Labs) or morphometric measurement via Metamorph software v7.10 (Molecular Devices, LLC). Analyses were run on the entire tissue section or manually annotated islets as indicated in figure legends. Endocrine cell mass was quantified by using pancreas weight and the ratio of hormone positive cells as identified by cytonuclear algorithm within the entire pancreatic section from multiple blocks representing the head, body, and tail regions. To obtain islet capillary measurements, caveolin-1 channel was isolated and color thresholding was used on a per-image basis to gather object data using the Integrated Morphometry Analysis (IMA) function (Metamorph). The following analysis metrics represent mean ± standard error: endocrine cells (Figure 3.4b, Figure 3.5f-h) 16,151 ± 1,715 islet cells/donor and 570,508 ± 51,866 total cells/donor; endocrine cell area (Figure 3.5c-d) 2.34 ± 0.24 mm2/donor; capillary morphology (Figure 3.4h) 48 ± 4 islets/donor; macrophage area (Figure 3.4m) 0.64 ± 0.07 mm2/donor; amyloid (Figure 3.6a) 108 ± 19 islets/donor; cilia (Figure 3.7h) 0.32 ± 0.05 mm2/donor; RFX6 (Figure 3.10f) 1,863 ± 362 cells/donor; pseudoislets (Figure 3.11g) 2,797 ± 508 cells/sample.

3.5.6.2 CODEX multiplexed imaging

Antibodies were purchased preconjugated from Akoya Biosciences or sourced from other vendors and conjugated in-house using the CODEX Conjugation Kit (Akoya Biosciences) or by Leinco Technologies, Inc. (St. Louis, MO, USA). 10-µm lightly fixed [23] pancreas sections were mounted onto 22x22 mm glass coverslips (Electron Microscopy Sciences) coated in 0.1% Poly-L-lysine (Sigma) and stained with the CODEX Staining Kit (Akoya Biosciences) in uncoated 6-well tissue culture plates (VWR) per manufacturer instructions. Fluorescent oligonucleotide-conjugated reporters were combined with Nuclear Stain and CODEX Assay Reagent (Akoya Biosciences) in light-protected 96-well plates sealed with foil (Akoya Biosciences) and automated image acquisition and fluidics exchange were performed using the Akoya CODEX instrument and CODEX Instrument Manager (CIM) v1.29 driver software (Akoya Biosciences) integrated with a BZ-X800 epifluorescent microscope (Keyence). Tissue was hydrated in 1X CODEX buffer (10X CODEX Buffer diluted in Milli-Q water) and hybridization/stripping of the fluorescent oligonucleotides was performed using dimethyl sulfoxide (Sigma). After loading of coverslip into stage insert, tissue was visualized with Nuclear Stain diluted 1:1000 in PBS and imaging area was set by center point and tile number using BZ-X800 viewing software (Keyence). All images were acquired using a CFI plan Apo I 20x/0.75 objective (Nikon) with 30% tile overlap and 5 z-planes (1.5 µm/z).

3.5.6.3 Processing and annotation of CODEX images

A total of 16 tissue regions were captured from 6 ND and 10 T2D donors (mean 50 mm2 tissue/donor). Image alignment, stitching, background subtraction, and deconvolution were performed using the CODEX Processor v1.7.0.6 (Akoya Biosciences; see https://help.codex.bio/codex/processor/technical-notes for details). Individual channel images (TIFF files) were imported into HALO software v3.1 (Indica Labs) for all analyses as described below. Tissue and islet areas were annotated by hand to exclude out-of-focus regions and poor tissue quality. Islets (estimated diameter ≥50 µm; mean 42 islets/donor) were annotated based on DAPI and CHGA channels. Cell segmentation and cell type annotations were performed using

the HALO HighPlex FL v3.2.1 module with consistent cytonuclear parameters (nuclear contrast threshold 0.456, maximum cytoplasm radius 0.48). Due to marker intensity variability among samples, thresholds were manually set for each marker and donor. Unless otherwise noted, cells were counted positive for a given marker if minimum intensity was reached in 50% of cytoplasm area (see Figure 3.5a-b for complete list of markers, abbreviations, and cell types). For cells with more variable morphology, positivity was also counted for nuclear area (30%: ARG1, CD11c, CD14, CD163, CD206, CD31, CD34, CD45, HLA-DR, IBA1, KRT, MCAM). Proliferating cells were counted only if minimum 60% of nuclear area met Ki67 intensity threshold. Vascular structures (CD31) were also measured by random forest classification algorithm (HALO Tissue Classifier module). The following analysis metrics represent mean ± standard error: endocrine cell area (Figure 3.4d) 0.88 ± 0.10 mm2/donor; islet cell composition (Figure 3.4e, Figure 3.5j) 7,322 ± 852 cells/donor; immune cells (Figure 3.4l,n) 309 ± 43 cells/donor; endothelial cell phenotypes (Figure 3.6f) 460 ± 92 cells/donor; macrophage phenotypes (Extended Figure 3.6h) 191 ± 29 cells/donor; T cell phenotypes (Extended Figure 3.6i-j) 40 ± 17 cells/donor.

3.5.6.4 High-dimensional, spatial, and neighborhood analyses

The R implementation of the UMAP algorithm (https://CRAN.R-project.org/package=umap) was used for dimensionality reduction. Cell marker percentages obtained through HALO were standardized across islets (n=255 ND islets and 426 T2D islets; mean 172 cells/islet), and default parameters were used for UMAP reduction (Figure 3.4o) except for nearest neighbors (80) and minimum distance (0.05). For spatial analyses, CD31 area classifications were converted to an annotation layer. A nearest neighbor algorithm (HALO Spatial Analysis module) was applied to obtain average distance of endocrine cells (n=4,830 ± 692 cells/donor) to islet capillaries (CD31+ region) (Figure 3.4i, Extended Figure 3.6d).

For cell neighborhood (CN) analysis, two methods were applied in parallel to CODEX data from annotated islets. In the community detection method, termed Dynamic CF-IDF (Figure 3.4p-q, Figure 3.4s), a weighted undirected heterogeneous graph for each islet was constructed based

on the cell types and normalized distance between cells. A greedy-based graph community detection method [16] was applied to segment the graph into a set of cell communities, then cell communities were stratified into 6 CNs (n=5,582 total CNs with median 11 cells/CN). Cell type enrichment was determined by a new proposed scoring function CF-IDF, which is a modification of the widely used text sequence analysis method term frequency (TF)–inverse document frequency (IDF) scoring [163]. Our cell frequency (CF)-inverse dataset frequency (IDF) score emphasizes the cell type that is not only prevailing, but also uniquely representative in a group of target islets. Therefore, it will deemphasize the most dominant cell types (e.g., α and β) throughout all the islets while paying more attention to the relative enrichment of less abundant cell types (e.g., vascular and immune cells) in the local regions. The downstream analysis not only introduces insightful results on T2D feature analysis but also shows a robust performance across different resolution levels.

The second CN analysis method, a k-means approach (Extended Figure 3.6k-n), built on a previously published algorithm used to identify CNs in the tumor microenvironment [234]. For each cell, we first found its 10 nearest neighbors in the islet and assigned the i-th nearest neighbor which was an α cell, β cell, macrophage, EC cell, or ɣ cell, a score $\cos(i\pi/20)$. Then we calculated the total score for each cell type, applied L1 normalization to the scores, and standardized them across all cells. The resulting representations of cells were finally used for k-means clustering to form 5 CNs (n=5,021 total CNs with median 5 cells/CN).

3.5.7 Transcriptional analysis of α and β cells and islets from ND and T2D donors

3.5.7.1 RNA isolation and bulk RNA-sequencing

RNA was extracted from sorted α and β cells (see above, Purification of α and β cells by FACS) or from pelleted whole islets using the Invitrogen RNAqueous-Micro Total RNA Isolation kit (Thermo Fisher #AM1931). TURBO DNA-free (Ambion) was used to treat any trace DNA con-

tamination. RNA was quantified by Qubit Fluorometer 2.0 and RNA integrity was confirmed (RIN >7) by 2100 Bioanalyzer (Agilent). Amplified cDNA libraries were constructed using SMART-seq v4 Ultra low Input RNA-kit (Takara) and sequencing was performed on an NovaSeq platform (Illumina) using paired-end reads (100 bp) and 25 million reads per sample.

We processed the raw RNA-seq reads using FastQC (v0.11.8) for broad quality assessment. Briefly, we examined the following parameters: (1) base quality score distribution, (2) sequence quality score distribution, (3) average base content per read, (4) GC distribution in thereads, (5) PCR amplification issue, (6) overrepresented sequences, (7) adapter content. Based on the quality report of fastq files, we trimmed sequence reads using fastq-mcf (v1.05) and cutadapt (v2.5) to only retain high quality sequence for further analysis. The paired-end reads were aligned to the GRCh37/hg19 human reference with GENCODE v19 gene annotation using STAR splice-aware aligner (v2.5.4b; –outSAMUnmapped Within KeepPairs) [71].

We counted fragments mapping to features type in GENCODE v19 gene annotation using featureCounts from Subread package [159]. The gene list was pruned to contain only protein-coding genes mapping to autosome and chrX, resulting in a total of 20,260 genes. We assessed libraries using comprehensive quality metrics generated by QoRTs [109] as well as computed derived metrics. Briefly, on the top of QoRTs reported metrics, we computed (1) 5'-3' gene coverage bias (as the ratio of coverage values at the 90%-ile and 10%-ile of the coverage distribution), (2) Kolmogorov-Smirov test statistic between cumulative gene diversity of each library relative to median distribution of all libraries within each cell type and standardized to a mean of 0 and standard deviation of 1 to yield a z-score, (3) number of reads mapped mapped to Xist and SRY genes, (4) average number of reads mapped to chrM, and (5) transcript integrity number (TIN) [270] for each library. The labeled sex of donors was matched against the gene expression quantified for sex genes to rule out any sample swaps or mislabeling. We also computed principal components for TPM normalized count matrix for each cell type in order to detect potential outliers.

3.5.7.2 Differential gene expression analysis

We performed differential gene expression analysis between T2D and ND samples for each cell type individually using DESeq2 [162]. In order to minimize potential effects of known and unknown confounding factors, we included known covariates in the DESeq2 model as well accounted for unknown covariates using RUVseq latent variable approach [222]. Briefly, we used the following multi-step process: (1) We first removed genes from the raw count matrix which had less than 10 reads in fewer than 25% of the samples for that cell type. (2) We then ran a first-pass differential expression analysis using DESeq2 with Age, Sex, BMI, and Batch as known covariates. The output result was filtered for genes that were non-significant i.e., not differentially expressed between T2D and ND samples and had p-value > 0.5. These genes were used as "control" or "empirical" genes for RUVSeq::RUVg function to estimate latent variables accounting for variation in the data not attributed to disease status. (3) The latent variables estimated from the RUVseq run were then used as additional covariates (on the top of Age, Sex, BMI, and Batch where applicable) for the second run of DESeq2. We selected the number of latent variables to provide the most reasonable separation between T2D and ND samples and minimal deviation from mean in the relative log expression plots. The output results from DESeq2 were filtered for 1% FDR to generate the final list of genes differentially expressed between T2D and ND for each cell type. We performed functional enrichment analysis using RNA-enrich [153] and retained terms with an FDR threshold of 5%. Terms were condensed using the RelSim function in REVIGO [247] with similarity parameter set to 0.5 and visualized in semantic space using an.xgmml file imported into Cytoscape software26 v3.8.2. Combined analysis of differentially expressed genes (fold change ≥1.5 or ≤-1.5; p<0.01) was performed using Metascape [287]. Metascape's heuristic algorithm samples the 20 top-score clusters, selects up to the 10 best scoring terms (lowest p-values) within each cluster, and connects terms pairs with Kappa similarity above 0.3. The resulting network was exported as a .cys file and visualized using Cytoscape, with the most representative term name in each cluster selected manually.

3.5.7.3 Gene network analysis

We adopted Weighted Gene Co-expression Network Analysis (WGCNA) [148] approach to create networks from the gene expression data. Briefly, we first filtered genes following the same rule established in Differential Gene Expression analysis where we only kept genes that had at least 10 reads in at least 25% of the samples for each cell type. We then processed raw counts using the varianceStabilizedTransformation function in DESeq2 package and used removeBatchEffect from the limma R package [223] to adjust for effects of age, sex, and BMI while protecting for disease status in the design matrix. The normalized and batch corrected count matrix was then used as input to blockwiseModules to create a "signed hybrid" network with "bicor" as the correlation function. The power (k) parameter was selected such that the scale free topology fit reached at least 80% fit. To examine cell type modules associated with quantitative traits of interest, we utilized a linear regression-based framework. We (1) inverse normalized the raw quantitative trait, (2) adjusted for Age, Sex, and BMI by linear regression, and (3) computed the spearman rank correlation between residuals and eigengene of all modules. Within each network, we also computed the module membership score and network connectivity for each gene. Estimated enrichment of curated gene lists [20, 61, 185] was calculated using Fisher's exact test. Functional enrichment of genes in each module was performed using gprofiler2 [143], and the results were visualized as a dotplot.

3.5.7.4 Integration of network analysis with chromatin accessibility

We integrated chromatin accessibility information with gene network analysis using sci-ATAC-seq data for α and β cells derived from our previously published study [217]. For each module within each cell type, we selected (a) accessible sites that were present within a specified distance of the transcription start site (TSS) of the genes within that module, and (b) the distal chromatin peaks that were linked to the peaks within this set based on the Cicero peak interaction results from the same study. This set of TSS proximal and distal peaks for all the genes within each module and for each cell type were then used for downstream enrichment analyses.

For variant enrichment analysis in the module linked peaks, we collected the latest published summary statistics for selected traits [42, 166]. Using a threshold of +- 10kb to define our gene TSS boundary for linking peaks with modules, we created a set of accessible sites for each module. The union of peaks across all modules was used as a "bulk" positive enrichment control. We then tested the enrichment of trait-associated variants from multiple GWAS across module peaks using GARFIELD [119] and used a p-value threshold of 5e-08 as input parameter for selecting trait-associated variants.

Next, we considered whether specific Transcription Factor Binding Motifs (TFBMs) are enriched to occur in certain modules. To test this, we defined module linked peaks for each module as described before but using a threshold of +- 1kb from gene TSS. For each peak within a module, we then identified the peak summit and extended the summit by 50 bp in each direction. Using genomic sequence in this region as our "test sequence", we used Analysis of Motif Enrichment (AME, v5.3.2) tool from MEME-Suite [14] (using default parameters) to identify enriched TFBMs represented in cisBP (v 2.0) [271]. The control set of sequence was generated using –shuffle– parameter in AME which generates a control sequence by shuffling the test sequence but preserving the 2-mer frequency. The enrichment score was computed as scaled log2 transformed (TP+1)/(FP+1) for each TFBM.

3.5.8 Pseudoislet formation and assessment of RFX6 knockdown

Pseudoislets were formed as previously described [269]. Briefly, nondiabetic human islets were handpicked to purity and then dispersed with 0.025% HyClone trypsin (Thermo Scientific) for 7 minutes at room temperature before counting with an automated Countess II cell counter or manually by hemacytometer. Dispersed human islets or purified β cells (see above, Purification of α and β cells by FACS) were incubated in adenovirus at a multiplicity of infection of 500 for 2 hours in Vanderbilt pseudoislet media before being spun and washed. Adenovirus containing U6 driven scramble or RFX6 targeted shRNA as well as CMV driven mCherry or mKate2 red fluorescent tag were prepared, amplified and purified by Welgen, Inc (Worcester, MA). Cells were

then resuspended in appropriate volume of Vanderbilt pseudoislet media to allow for seeding into wells at 2000 cells per 200 μL each well of CellCarrier Spheroid Ultra-low attachment microplates (PerkinElmer). Pseudoislets were allowed to reaggregate for 6 days before being harvested and studied. To assess knockdown, RNA was extracted from pseudoislets containing only β cells using an RNAqueous RNA isolation kit (Ambion). cDNA synthesis and quantitative reverse transcriptase PCR were performed as previously described [57]; briefly, cDNA was synthesized using a High-Capacity cDNA Reverse Transcription Kit (Applied Biosystems #4368814) according to the manufacturer's instructions. Quantitative PCR (qPCR) was performed using TaqMan probes for ACTB (Hs99999903_m1) as endogenous control and RFX6 (Hs00941591_m1). Relative changes in mRNA expression were calculated by the comparative ΔCt method.

3.5.9 Multiome single nuclear RNA/ATAC-sequencing

3.5.9.1 Nuclear isolation

Pseudoislet samples treated with RFX6 shRNA or scramble RNA were pooled together using a randomized study design, so the targeting and scramble conditions were not confounded by batch (Figure 3.12a). To accomplish this, samples were allocated into six groups (batches) of n=490-494 pseudoislets for nuclei isolation. A customized protocol was developed based on recommendations by 10x Genomics (https://www.10xgenomics.com/resources/demonstrated-protocols/) which included optimization steps described below. Briefly, the samples were suspended in 1X PBS and pelleted at 2000 x g for 3 minutes at 4℃. The pellet was resuspended in lysis buffer (10mM Tris-HCl 7.4 pH, 10mM NaCl, 3mM MgCl2, 0.1% Tween-20, 0.1% NP40, 0.01% Digitonin, 1% BSA, 1mM DTT, and 2U/μl RNase Inhibitor) and rocked in an Eppendorf thermomixer C (EP #5382000015) at 300 x g for 5 minutes at 4℃. Keeping the samples on ice as much as possible, tubes were then transferred to a prechilled 2 mL glass dounce homogenizer and homogenized with 15 strokes of tight pestle B before being transferred to a 1.5 mL tube and centrifuged at 500 x g for 5 minutes at 4℃. The resulting pellet was then resuspended in 1 mL of wash buffer (10mM Tris-HCL 7.4 pH, 10mM NaCl, 3mM MgCl2, 1% BSA, 0.1% Tween-20, 1Mm

DTT, and 2U/µl RNase Inhibitor) and centrifuged at 100 x g for 1 minute at 4℃. The supernatant was collected, filtered through a pre-wetted 30 µm filter, and centrifuged at 500 x g for 5 minutes at 4℃. Nuclei were resuspended in 300 µl of wash buffer, then 300 µl of sucrose cushion (0.88M sucrose, 1mM DTT, 1mM RNAse Inhibitor, and 10% wash buffer) was added to the bottom of the tube and the resulting layered solution was centrifuged at 1000 x g for 10 minutes at 4℃. Both layers of supernatant were removed, and pellet was resuspended in 1 mL of wash buffer and centrifuged at 500 x g for 5 minutes at 4℃. Nuclei were then resuspended in 30 µl of nuclei resuspension buffer before counting and quality assessment. The desired concentration of nuclei was achieved by resuspending the appropriate number of nuclei in 1X diluted nuclei buffer for joint (on the same nucleus) snATAC-seq and snRNA-seq multiome profiling. Nuclei were processed by the University of Michigan Advanced Genomics Core using the 10x Genomics Chromium platform at 20K nuclei per well.

3.5.9.2 Multiome sample genotyping and imputation

Samples were genotyped with the Infinium Multi-Ethnic Global-8 v1.0 kit using 50 ng/uL DNA samples in two batches. Probes were mapped to Build 37. We merged the .ped files for the two batches along with samples from other projects that were genotyped on the same chip (resulting in a combined 68 samples). We removed variants with multi mapping probes and updated the variant rsIDs using Illumina support files "Multi-EthnicGlobal_D1_MappingComment.txt and Multi-EthnicGlobal_D1.annotated.txt"[1]. We performed pre-imputation QC using the "HRC-1000G-check-bim.pl" script (version 4.2.9) obtained from the Mark McCarthy lab website[2] to check for strand, alleles, position, Ref/Alt assignments and update the same based on the 1000G reference[3]. We did not conduct allele frequency checks at this step (i.e. used the "–noexclude" flag) since we had 68 samples from mixed ancestries. These filters resulted in 958,427 variants. We performed pre-phasing and imputation using the Michigan Imputation Server [63]. The stan-

[1]https://support.illumina.com/downloads/infinium-multi-ethnic-global-8-v1-support-files.html
[2]https://www.well.ox.ac.uk/w̃rayner/tools/
[3]https://www.well.ox.ac.uk/w̃rayner/tools/1000GP_Phase3_combined.legend.gz

dard pipeline[4] included pre-phasing using Eagle2 [160] and genotype dosage imputation using Minimac4 (https://github.com/statgen/Minimac4) and the 1000g phase 3 v5 (build GRCh37/hg19) reference panel [10]. Post-imputation, we selected biallelic variants with estimated imputation accuracy (r^2>0.3), variants not significantly deviating from Hardy Weinberg Equilibrium (P>1e-6), MAF in 1000G European individuals > 0.05 and minor allele count (MAC) >1 in our 12 samples, resulting in 6,665,607 variants.

3.5.9.3 Data processing (RNA component)

The RNA component of the multiome data was processed using starSOLO (STAR v. 2.7.3a, with GENCODE v19 annotation; options –soloUMIfiltering MultiGeneUMI –soloCBmatchWLtype 1MM_multi_pseudocounts –soloCellFilter None), which outputs the count matrices needed for most of the analyses [71]. Quality control metrics were gathered on a per-nucleus basis using a custom Python script on the corrected gene counts and aligned BAM file.

Following processing with STAR, we constructed a custom count matrix by combining information from the GeneFull and Gene matrices output by STAR. The GeneFull matrix contains per-gene counts based on intronic and exonic reads, while the Gene matrix contains per-gene counts based on exonic reads only. As nuclear RNA may contain introns, the GeneFull matrix should be preferred. However, due to overlapping transcript annotations that render some read gene assignments ambiguous, some genes may receive fewer counts in the GeneFull matrix than in the Gene matrix. The INS gene was an extreme example of this, receiving very low counts in the GeneFull matrix but high counts in the Gene matrix. To salvage counts for such genes, our custom matrix utilized the GeneFull counts for most genes but utilized the Gene counts for the subset of genes that had greater counts in the Gene matrix than in the GeneFull matrix.

[4] https://imputationserver.readthedocs.io/en/latest/pipeline/

3.5.9.4 Data processing (ATAC component)

Adapters were trimmed using cta (https://github.com/ParkerLab/cta). We used a custom Python script, available in the Parker lab Github repository, for barcode correction. Barcodes were corrected in a similar manner as in the 10x Genomics Cell Ranger ATAC v. 1.0 software. In brief, barcodes were checked against the 10x Genomics whitelist. If a barcode was not on the whitelist, then we found all whitelisted barcodes within a hamming distance of two from the bad barcode. For each of these whitelisted barcodes, we calculated the probability that the bad barcode should be assigned to the whitelisted barcode using the Phred scores of the mismatched base(s) and the prior probability of a read coming from the whitelisted barcode (based on the whitelisted barcode's abundance in the rest of the data). If there was at least a 97.5% probability that the bad barcode was derived from one specific whitelisted barcode, it was corrected to the whitelisted barcode.

Reads were mapped using BWA-MEM [155] with flags '-I 200,200,5000 -M' (v. 0.7.15-r1140). We used Picard MarkDuplicates (v. 2.25.1; https://broadinstitute.github.io/picard/) to mark duplicates, and filtered to high-quality, non-duplicate autosomal read pairs using SAMtools view44 with flags '-f 3 -F 4 -F 8 -F 256 -F 1024 -F 2048 -q 30' (v. 1.10). Quality control metrics were gathered on a per-nucleus basis using ataqv45 (v. 1.2.1) on the BAM file with duplicates marked.

3.5.9.5 Selection of quality nuclei (barcodes) for downstream analysis

We performed rigorous QC of all RNA nuclei and only included those deemed as high-quality based on the following four definitions: 1) nUMI > 1000, 2) mitochondrial fraction < 0.2, 3) nuclei where the RNA profile was statistically different from the background/ambient RNA signal, and 4) nuclei identifiable as a singlet and assignable to a sample using genotypes. We considered droplets with UMIs < 10 to be "empty" and therefore representative of the background/ambient RNA profile. Top genes in the ambient RNA included highly expressed genes across prominent islet cell types such as INS, GCG, and SST, along with several mitochondrial genes. We used the testEmptyDrops function from DropletUtils (v 1.6.1) [164], specifying the 'lower' parame-

ter as 10 and selecting droplets with P<0.05 as droplets significantly different from the ambient RNA profile. To identify singlets and assign to samples, we ran Demuxlet [130] using using the BAM files and the genotype VCF file considering all post-QC variants in gene bodies with minor allele count (MAC) >1. We used the command `demuxlet --sam $bam --tag-group CB --tag-UMI UB --vcf $vcf --alpha 0 --alpha 0.5 --field GT`, and selected singlets. To account for ambient RNA contamination while identifying singlets, we also masked the top 1% genes expressed in the ambient RNA and re-ran Demuxlet with the same parameters; nuclei were considered singlets and kept for downstream analysis if they were called as singlets in either Demuxlet run.

We also performed QC of the ATAC component of the multiome data. For ATAC, we required nuclei to have a minimum TSS enrichment (as calculated by ataqv) of 2, minimum filtered read count of 1000 (ataqv 'HQAA' metric), and maximum mitochondrial fraction of 0.5. We also ran Demuxlet on the ATAC component (command: `demuxlet --sam $bam --tag-group CB --vcf $vcf --field GT`) and required that a prospective nucleus be called as a singlet. The ATAC component of nuclei in two wells showed low TSS enrichment and all nuclei from these two wells were therefore excluded from analysis.

If the RNA and the ATAC component of a barcode both passed QC and the Demuxlet sample assignment was the same, both modalities were utilized for downstream analysis. If only the RNA component passed QC, only the RNA component was used in downstream analysis. As we performed clustering on the RNA component, we excluded the few (twelve) barcodes that passed ATAC QC and failed RNA QC.

3.5.9.6 Removal of ambient RNA counts from single nucleus gene expression UMI matrices

Prior to clustering and downstream analysis, we used DecontX [282] (celda v. 1.8.1, in R v. 4.1.1) to adjust the nucleus x gene expression count matrices for ambient RNA. DecontX was run on a per-batch basis, as the amount of ambient contamination may vary across batches. Decontaminated

counts were generated via the decontX() function, passing barcodes with total UMI count <= 10 to the background argument. Rounded decontaminated counts were used for clustering and all downstream analyses. Nuclei with estimated contamination level > 0.2 were excluded from downstream analysis.

3.5.10 Clustering of multiome data

Nuclei were clustered on the RNA component using Seurat [32, 106, 245] (v. 3.9.9.9010, in R v. 3.6.3). After normalizing counts with the NormalizeData function, we identified the top 2000 variable features (FindVariableFeatures function, with selection.method='vst') and scaled with the ScaleData function. We identified neighbors using the top 20 PCs and k.param = 20, and called clusters using resolution = 0.1 with n.start = 100. We used the top 20 PCs for generating the UMAP.

This clustering protocol identified 10 clusters. One of the smaller clusters shows expression of both INS and GCG, suggesting it may consist of doublets that were not caught by demuxlet. To verify this was a doublet cluster, we ran a different, genotype-independent, ATAC-based doublet detection method (AMULET; v. 1.0-beta, run with default parameters separately on data from each multiome well) [256] on the ATAC nuclei that otherwise passed QC. This method tagged ~40% of the nuclei in the suspected doublet cluster as doublets, while only ~5% of nuclei in any other cluster were tagged as doublets. We therefore removed the small doublet cluster from the clustering and downstream analysis. Data are available via the UCSC Cell Browser [242] at https://theparkerlab.med.umich.edu/data/public/cellbrowser/?ds=Pseudoislet10XMultiome for further exploration.

3.5.10.1 Differential gene expression analysis

Differential gene expression was performed within each cluster using DESeq2 (v. 1.28.0) [162] on pseudobulk counts. UMI counts were summed across nuclei within a donor + construct + cluster. Only donors with paired data (RFX6-2896 and scrambled-mCherry constructs) were used, and

the analysis was performed in a paired fashion (DESeq2 model: ~donor + construct). We used an FDR threshold of 5% for considering genes differentially expressed.

3.5.10.2 Per-cluster processing of ATAC component

All ATAC reads from pass-QC, clustered nuclei were merged within each cluster. To generate per-cluster peaks, these BAM files were converted to single-ended BED format using bedtools bamtobed [214] before calling ATAC-seq peak summits with MACS2 [285] (flags -g hs –nomodel –shift -37 –extsize 73 -B –keep-dup all –call-summits). We removed summits in blacklist regions, filtered to FDR 0.1% summits, and then generated a peak list from the summits by extending the ATAC-seq peak summits for each cluster +/- 150 bps to get 300bp peaks (within each cluster, if two 300bp peaks overlapped the one with the greater MACS2 score was kept). We then removed peaks in blacklist regions. To get the ATAC peak counts used in the ATAC PCA and differential chromatin accessibility analyses, we determined the number of ATAC fragments overlapping each of these peaks in each of the per-cluster, per-donor, per-construct pseudobulk samples.

For visualization of ATAC signal, we generated a normalized bedGraph file using MACS2 on the single-end BED file (macs2 callpeak command, with options –SPMR –nomodel –shift -100 –extsize 200 -B –broad –keep-dup all) and then converted to bigWig format using the UCSC bedGraphToBigWig [133]. For PCA on the pseudobulk ATAC counts, we first removed any peaks on the mCherry or mKate2 contigs. We then converted peak counts to counts per million and removed the bottom 10% of features with the lowest average CPM across samples. For each peak, we filled any 0s with a value equal to half of the minimum non-zero CPM for that peak across samples. We then log transformed prior to performing the PCA.

3.5.10.3 Differential chromatin accessibility analysis

Differential chromatin accessibility was performed within each cluster using DESeq2 (v. 1.28.0) on pseudobulk ATAC peak counts. Only donors with paired ATAC data (RFX6-2896 and scrambled-mCherry constructs) were used, and we additionally excluded donor 17277513 due

to very low read counts. The DESeq2 analysis was performed in a paired fashion, with model: " donor + tss_enrichment + construct". To compute TSS enrichment for each pseudobulk sample, we merged all ATAC nuclei (regardless of cluster) from each donor and computed TSS enrichment with ataqv.

3.5.11 Testing for enrichment of peak subsets near differential genes

We used a permutation test to determine whether the most significant peaks ('top peaks') from the beta cell differential peak analysis were enriched near beta cell differentially expressed (DE) genes. First, we assigned each peak to the gene with the nearest TSS (if multiple TSS were equally close, we took the TSS with the smallest chromosomal coordinate). We then calculated the fraction of top peaks whose nearest gene was DE. To get the null expectation for this value, we permuted the 'DE/not DE' gene labels, such that the same number of genes were always labeled as 'DE' but the identity of these DE genes changed in each permutation. While permuting, we split genes into deciles based on the expression of each gene and permuted the labels only within each decile (this controls for the fact that highly expressed genes are more likely to be DE than lowly expressed genes due to statistical power in the DE analysis). We performed 10,000 permutations, in each permutation re-calculating the fraction of top peaks whose nearest gene was DE to build up the null distribution. We then calculated an empirical p-value based on our observed value and the null distribution, adding a pseudocount to avoid a p-value of 0 (p = [1 + # of permutations where the test statistic was greater than or equal to our observed value] / 10,001).

3.5.11.1 Motif scanning for multiome motif enrichment analyses

The motif scans were performed using FIMO (v. 5.0.4) [98] with a background model calculated from the hg19 reference genome and otherwise default parameters. We used the motifs from [134], excluding "*_disc" motifs; motifs from cisBP (v. 2.0); motifs from [122]; and custom RFX6 motifs generated using mouse RFX6 ChIP-seq data from [206].

The custom RFX6 motifs were generated during a previous project [264]. Sequencing reads

from [206] were mapped to the mouse mm9 genome63 using bwa (v. 0.7.12-r1039) and peaks were called using MACS2 (flags: "-t MIN6_Rfx6-HA_IP.bam -c MIN6_Control-HA.bam -B –nomodel -g mm –keep-dup 1 -q 1.00e-4"). The MEME (v. 4.11.0) [12] and DREME (v. 4.9.1) [11] tools from the MEME suite [13] were used to discover novel motifs in the resulting peaks. One non-repetitive motif from the MEME tool and two motifs from the DREME tool, bearing similarity to known RFX family motifs, were selected for use in downstream analysis.

3.5.11.2 Motif enrichment in most significant peaks

We used logistic regression to measure enrichment of motifs in subsets of ATAC-seq peaks. We ran one model per peak category and motif. For testing for enrichment in the peaks that had the smallest p-values and leaned towards higher signal in shRFX6 samples, we modeled:

```
peak_leans_higher_in_shRFX6 ~ peak_gc_content + peak_size +
                              n_motif_hits_in_peak
```

Where "peak_leans_higher_in_shRFX6" is 1 if the peak was one of the most significant peaks in the "up in RFX6 KD condition" direction and 0 otherwise; "peak_gc_content" was the GC content of the sequence within the peak; "peak_size" was the mean DESeq2-normalized count for the peak across the samples in the DESeq2 analysis; and "n_motif_hits_in_peak" was the number of motif hits in the peak as determined by the FIMO motif scans. The coefficient of the "n_motif_hits_in_peak" term was taken as the measure of motif enrichment. For testing for enrichment in the peaks that had the smallest p-values and leaned towards lower signal in shRFX6 samples, we used the same model except the outcome variable was "peak_leans_lower_in_shRFX6."

3.5.11.3 Generation of ATAC footprint plots

To generate the ATAC footprint plots, we first separated the motif occurrences into those within the beta cells ATAC peaks and those outside of peaks. For each of these two groups, we computed an aggregate Tn5 cut matrix for the 500 bps on either side of the motifs, using beta cell

ATAC reads from each individual donor+construct (using the "make_cut_matrix" script within the atactk package (https://github.com/ParkerLab/atactk); options: -a -r 500). The cut matrices were generated separately for each donor+construct, utilizing only donors with paired ATAC data (RFX6-2896 and scrambled-mCherry constructs) and additionally excluding donor 17277513 due to very low ATAC read counts. To reduce the impact of Tn5 insertion sequence bias, we normalized the Tn5 cut frequency at each position for the motifs in peaks by the corresponding frequencies for the motifs outside of peaks. To adjust for technical differences (e.g., TSS enrichment) between the donors+constructs, we then divided these normalized cut frequencies by the average normalized cut frequency between the -500 and -400 bp positions.

3.5.11.4 GWAS enrichment in most significant peaks

We considered if β cell ATAC-seq peaks that score highly for differential accessibility, as measured by p-value, are specifically enriched to overlap T2D-GWAS variants. We compared the enrichment of T2D (adj. BMI) GWAS variants to overlap top 5000 ATAC-seq differential peaks leaning up and down with the remaining peaks for β cell using GARFIELD [119]. Using a p-value threshold of 1e-05, we also performed a conditional analysis where GARFIELD evaluates if both annotations are conditionally independent of each other in the enrichment model. The coefficients corresponding to each annotation from the conditional enrichment model were shown along with the 95%-CI. To ensure robustness of our results, we repeated the analysis for top 2000 (up and down each) and top 10000 (up and down each) differential peaks.

3.6 Data availability

Extended data and tables referenced in the chapter are not included in the dissertation and can be obtained from the online version of the manuscript referenced in [268]. Code for analysis done in this work is openly available on GitHub at https://github.com/ParkerLab/2021_islet-rfx6, and a summary web-page of the project is available at http://theparkerlab.org/manuscripts/2021_islet-

rfx6/ with link to all resources.

3.7 Acknowledgments

Conceptualization, JTW, DCS, VR, CD, JJW, SCJP, ACP, MB; Software, VR, PO, YT, SF, AV; Formal Analysis, VR, PO, ALH, YT, SF, SS, AV, SA; Investigation, JTW, DCS, CD, ALH, CVR, JJW, YDP, CV, RA, GP, RJ, NJH; Resources, DLG, LDS, RB; Data Curation, JTW, DCS, VR, PO; Writing — Original Draft, JTW, DCS, VR, SCJP, ACP, MB; Writing — Review & Editing, all authors; Visualization, JTW, DCS, VR, PO, SS, AV; Supervision, JL, SCJP, ACP, MB; Funding Acquisition, SCJP, ACP, MB.

We thank the organ donors and their families for their invaluable donations and the International Institute for Advancement of Medicine (IIAM), Organ Procurement Organizations, National Disease Research Exchange (NDRI), and the Alberta Diabetes Institute IsletCore together with the Human Organ Procurement and Exchange (HOPE) program and Trillium Gift of Life Network (TGLN) for their partnership in studies of human pancreatic tissue for research. We thank Drs. Jing Hughes, Jeong Hun Jo, Seung Kim, Yan Hang, Joana Almaça, Roland Stein, and Rachana Haliyur for their valuable scientific insight regarding experimental design and methods. This study used human pancreatic islets that were provided by the NIDDK-funded Integrated Islet Distribution Program at the City of Hope (DK098085). This work was supported by the Human Islet Research Network (RRID:SCR_014393), the Human Pancreas Analysis Program (RRID:SCR_016202), DK106755, DK123716, DK123743, DK120456, DK104211, DK108120, DK104218, DK112232, DK117960, DK126185, DK117147, EY032442, T32GM007347, F30DK118830, DK20593 (Vanderbilt Diabetes Research and Training Center), The Leona M. and Harry B. Helmsley Charitable Trust, JDRF, Doris Duke Charitable Foundation, and the Department of Veterans Affairs (BX000666). Cell sorting was performed in the Vanderbilt Flow Cytometry Shared Resource (P30 CA68485, DK058404) and whole-slide imaging was performed in the Islet and Pancreas Analysis Core of the Vanderbilt DRTC (DK20593).

3.8 My contributions

The results presented in this chapter are an outcome of a huge team effort. I thank my co-authors Jack Walker, Diane Saunders, Marcela Brissova, Steve Parker, Alvin Powers, all the team members at the Vanderbilt University and the University of Michigan for their contributions to the project.

Of the results presented in this chapter, I specifically contributed to the analysis of RNA-seq data (differential gene expression, gene-network construction, module-level analysis, integration with ATAC-seq data), RFX6 knockdown experiment design, and single-nuclear multiome data processing and QC for cluster identification. I also implemented the interactive and static web resources (see section 3.6) available with the manuscript, published the source code, and deposited the data in EGA. Finally, I contributed to figure design, manuscript writing, and reviewing of the work presented here. Because the manuscript corresponding to this chapter is currently in review, the final published version may differ.

CHAPTER 4

Conclusion

In this thesis, I have sought to understand and dissect the genetic complexity of T2D by profiling molecular characteristics from healthy and diabetic individuals across different layers of organization. By focusing on different data modalities, such as epigenome and transcriptome, and their integration within the pancreatic islets, I have identified mechanisms of gene regulation that underlie the associations between genetic variation and disease traits such as T2D. In section 4.1, I summarize the broader themes emerging from this thesis, and in section 4.2 I highlight the specific contributions made by this thesis to the current field of research. Finally, in section 4.3, I conclude by providing current and future directions to build upon this body of work in the extant literature.

4.1 Summary

4.1.1 High-resolution chromatin accessibility map of pancreatic islets

In chapter 2, I present our work on profiling chromatin accessibility in pancreatic islets from a single donor using sci-ATAC-seq protocol. While our ability to collect multiple replicates was limited, our work clearly demonstrates the utility of single-cell resolution epigenomic profiling for study of cell-type specific regulatory elements and their role in disease pathophysiology. The results and data generated from our study contributed significantly to our existing understanding of islet heterogeneity in several ways. First, while many earlier studies had focused on profiling

heterogeneity through cell-surface markers and fluorescence-activated cell-sorting to obtain distinct cell populations, we employed an unbiased data driven approach that allowed us to discover low abundance cell types with as few as 28 cells in the sample. Second, earlier single-cell studies have focused mainly on profiling the gene expression of the islets and constituent cell types. However, transcriptomic state of a cell is more dynamic than the epigenetic state and is generally a consequence of the regulatory processes within the cell. As a result, creating a map of chromatin accessibility provides us a unique opportunity to identify and characterize mechanisms of regulation that may contribute to the functional changes relevant to the disease. Here, we collected single-cell chromatin accessibility data on ~1500 cells from the islets and identified three major cell types — α, β, and δ cells. Surprisingly, even though delta cells constitute only 2-3% of the cells in the islet, we were able to identify 28 cells (out of ~1500 cells) cluster as δ using their accessibility profile at the *SST* marker gene. However, because we only had few δ cells and much lower read coverage, conventional approach for identifying chromatin accessibility *peaks* was not successful. We then developed a novel deep-learning based approach, inspired from image upscaling algorithms in the field of computer vision, to impute the sparse and noisy signal from low abundance δ cell cluster to an "upscaled" signal equivalent to a high abundance cluster. Using the high-abundance data for the α and β cell types and their downsampled versions, we trained the model to predict chromatin accessibility (or regulatory) peaks from the sparse data. This allowed us to identify thousands of new δ cell peaks that were enriched for cell-type specific signatures genes previously defined by single-cell RNA-seq studies. Our comparison to ATAC-seq peak data derived from bulk islet samples further reinforced the high-quality of upscaled peaks.

In fact, since our publication of the deep-learning approach, another study [147] developed a toolkit based on similar principles to denoise and upscale sequencing coverage data from low abundance, low coverage cell types. Based on our work and subsequent findings by AtacWorks [147], we establish that novel deep-learning based methods can enhance the sensitivity of single-cell experiments by "upscaling" or denoising data from low-abundance and low coverage cell types. Further, these approaches can be generalized to different modalities such as ChIP-seq for

inference of protein-DNA interaction, critical for identifying regulatory regions and mechanisms of regulation.

Once we had cell-type chromatin accessibility maps, we asked if T2D GWAS variants are differentially enriched in chromatin maps of a particular cell type. β cells due to their critical role in insulin synthesis and secretion are known to harbor mutations for monogenic diabetes (MODY). Here, we find that while all three cell types are enriched for T2D GWAS variants, only β cells are enriched to overlap the variants after conditional analysis reinforcing the idea that majority of disease genetic risk burden is carried by the β cells. With this information, we sought to identify and nominate putative target genes for associated T2D GWAS variants. Using co-accessibility approach implemented in Cicero, we identified distinct examples of regulatory interactions in α and β cells. We highlighted specific examples at the *C2CD4A/B*, *ADCY5*, *ANK1* and *NKX6-3* GWAS loci where specific variants were linked to target gene promoters in a cell-type specific manner.

Overall, we contributed the first single-cell resolution chromatin accessibility data for pancreatic islets resolved for three major constituent cell types.

4.1.2 Multiomic integration for variant prioritization and target discovery

Joint profiling of molecular domains across layers of the genetic organization is essential for a holistic understanding of development and disease. Our understanding of how genetic variation constitutes a cascading effect through the layers before culminating in the disease can be combined with the clinical knowledge of the phenotype to identify precise therapeutic targets [19, 36, 43]. In chapter 3, I expanded our study to focus on concerted changes across these layers in type 2 diabetic individuals. However, a significant challenge of integrative multiomic analyses is the differences between data types and their properties. For example, ATAC-seq is performed genome-wide and generates a continuous measure of signal across, RNA-seq assays yield discrete transcript counts, whereas secretion assays or physiological profiles may provide a time-series

measure or a single quantitative summary. Our integrative analysis approach focused not only on molecular signatures associated with the disease in each data modality but also used innovative approaches to link them together and follow the cascade of functional consequences from genome to physiology. Using the gene-network approach for RNA-seq data from sorted cell types, I created functional modules that could be correlated with donor traits and linked to ATAC-seq data to proportion disease risk across cell-specific modules. Our identification of insulin secretion, ion signaling, and cilia-associated modules that are dysregulated in β cells corroborate previous findings on signature pathways attributed to islet dysfunction and early-stage T2D. Overall, these results are a natural consequence of our funnel-like investigation scheme where we begin from broader, bulk molecular profiles, and trait-GWAS, to cell-specific signatures, and eventually to specific groups of genes and variants within those cell types. Prioritization of these functional modules and disease-associated variants for downstream experimental validation can be an effective starting point for effector transcript discovery and opportunities for targeted intervention.

4.1.3 Identification of RFX6 as critical for β cell function

Efforts to identify regulators of β cell function in type 2 diabetes have been underway for a long time — initially through candidate gene testing approaches and later through GWAS — with limited success. For example, studies of monogenic disorders (MODY) of type 2 diabetes that are characterized by single-gene defects and early onset, have identified many genes controlling β cell function and formation and relevant to the disease onset and progression. However, such disorders are driven by rare mutations in the coding sequences of the genes, resulting in direct change in the activity of corresponding transcription factor or enzyme. For example, MODY 2 results from an abnormal glucokinase enzyme encoded by the *GCK* gene on chromosome 7, while MODY 1, MODY 3-7 are caused by mutations in the transcription factor genes such as *HNF4α*, *HNF1α*, *PDX1*, *HNF1β*, and *NEUROD1*. Mutations in many of these genes, and thus their dysregulation, can cause diabetes, highlighting the pathway's importance in human β-cell formation and insulin production.

However, in non-monogenic forms of diabetes, which constitutes the majority of diabetes cases or individuals at the risk of developing diabetes, most disease pre-disposing variants are common and non-coding, as evidenced by recent large scale GWAS for type 2 diabetes and related traits, and make very small contribution to the disease risk (see subsection 1.2.1). Combined with our understanding of the severe forms of the disease, non-coding mutations are hypothesized to contribute to the disease risk through context-specific modulation of β-cell gene expression (see subsection 1.2.2).

In chapter 3, my work integrating multiomic data from islet function, transcriptome, and tissue architecture identified the dysregulation of *RFX6* and its regulatory network as a key molecular event in early T2D pathogenesis. *RFX6* encodes for a transcription factor that functions downstream of *NEUROG3* and upstream of *PDX1* in β cell development, and coding-defects in the gene are known to cause neonatal and potentially fatal form of diabetes called Mitchell-Riley syndrome.

Here I show that broad transcriptional changes in β cell implicated pathways related to ion-transport, vesicle trafficking, exocytosis, and insulin secretion. Further, co-expression network analysis and association with GWAS variants and physiological parameters allowed us to prioritize processes with physiological relevance that were more likely to be disease-causing rather than disease-induced. I identified *RFX6*-containing β01 module to be (a) significantly associated with insulin secretion, (b) enriched for dysregulated genes, and (c) enriched to overlap T2D GWAS risk variants, emphasizing the central role of RFX6 in β-cell defect driven disease development. In fact, a previous study which looked at bulk chromatin accessibility in pancreatic islets found that T2D GWAS loci were strikingly and specifically enriched in islet Regulatory Factor X-binding (RFX) footprints and uniformly disrupt the RFX motifs at high-information content positions.

Overall, this integrative work identifies RFX6 as critical for β-cell function in early-stage T2D and show that common genetic variation linked to the type 2 diabetes, unlike monogenic defects, disrupts the regulatory context of *RFX6* gene resulting in a cascading effect leading to irreversible changes in β-cell homeostasis.

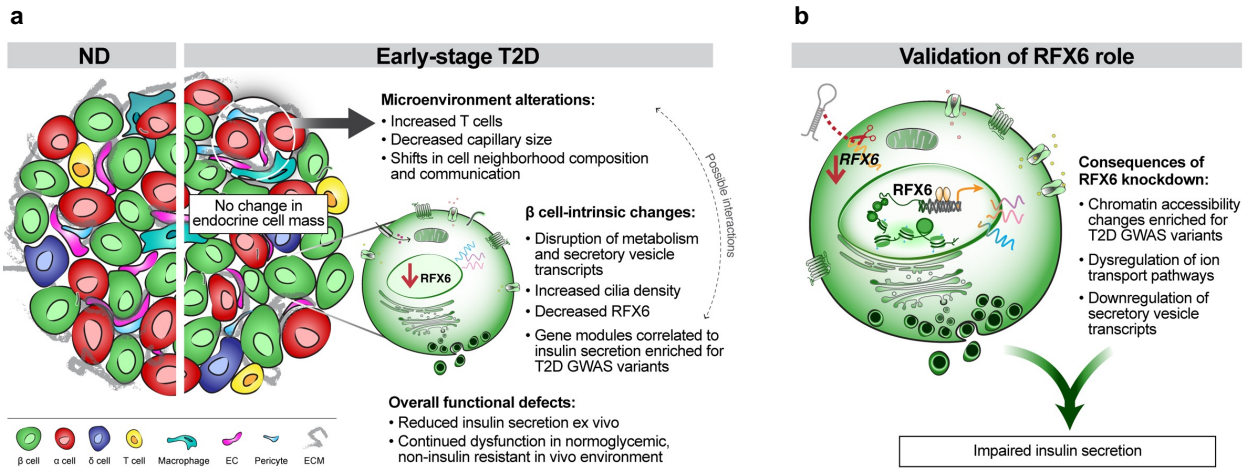


Figure 4.1: RFX6-mediated chromatin, transcriptome, and insulin secretion dysregulation in human β cells. (a) Major β cell-intrinsic and islet microenvironment alterations that define islet dysfunction in early-stage T2D. Observations from transcriptomic and histologic studies revealed no change to endocrine cell composition but evidence of dysregulated β cell processes and modest changes to intraislet vascular and immune cell populations. Insulin secretion was reduced and persisted in a nondiabetic environment. (b) RFX6 knockdown using a primary human pseudoislet system resulted in dysregulated vesicle trafficking and ion transport pathways, mediated by chromatin architectural changes overlapping with T2D GWAS variants. This led to reduced insulin secretion, confirming the critical role of RFX6 in human β cell function. Figure created by Diane C. Saunders for [268].

4.2 Contributions

In summary, the contributions of my thesis are as follows:

1. The first to be published single-cell resolution chromatin accessibility map of human pancreatic islets, followed closely by a similar study from a different group [44]. We identified constituent cell types within the islets based on their unique epigenomic signatures and using those cell-type-specific signatures, we showed that:

   (a) β cell chromatin accessibility peaks are enriched to overlap T2D GWAS signals and are conditionally independent of α and δ cell chromatin peaks;

   (b) A deep-learning based approach can be a valuable tool in the single-cell data analysis where the sparse and noisy signal from low abundance cell types, δ cells for example, can be imputed ("upscaled") to fill the gaps using data from high-abundance cell types such as α and β cells; and

   (c) Co-accessibility between chromatin accessibility peaks using single-cell resolution data can be used to link genetic signals to target genes can provide a higher resolution variant-to-gene interaction map than and help us precisely identify targets that are otherwise not identified by GWAS.

2. A multimodal, integrative analysis framework to investigate pancreatic islet cell types (α, β, and whole islets) between early-stage T2D and healthy individuals for identification of central transcriptional regulators driving β cell dysfunction in short-duration T2D. We found that:

   (a) Impaired β cell function, a hallmark of early-stage T2D, persisted ex vivo and in non-diabetic environments; in contrast, α cell function was not changed;

   (b) Islet endocrine composition was unchanged though there were modest alterations to the islet microenvironment in endothelial and immune cells;

(c) Transcriptional network analysis proportioned T2D genetic risk (disease predisposition) into cell- and context-specific gene modules with specific functional properties, and

(d) RFX6 emerged as a highly connected hub transcription factor that was reduced in T2D β cells and associated with reduced glucose-stimulated insulin secretion.

4.3 Limitations and future work

There are many new directions that emerge from our findings with exciting new questions to pursue. In this section, I consider several possibilities, including addressing limitations of many approaches considered, that would further improve our understanding of T2D and also allow us to extend it to other complex, polygenic diseases.

4.3.1 Mapping islet heterogeneity across disease and donor development stages

In chapter 2, I presented results from pancreatic islet tissue obtained from a single healthy donor. However, since then, a similar study containing >15,000 nuclei from three donors has been published [43], and many more are underway. For example, the recently published *Tabula Sapiens* study analyzed single-cell transcriptomic data across multiple human organs including pancreas [255]. The increased number of donors and nuclei provides a unique opportunity to combine and meta-analyze multiple islet single-cell RNA and ATAC-seq datasets. This will allow us to study donor variation, heterogeneity in cell populations such as distinct cell-states, and an increased statistical power to detect chromatin features, including loops, at GWAS loci [81]. Finally, with enough donors and nuclei, one may also do single-cell resolution QTL studies using the joint gene expression and chromatin accessibility data collected from the same cell. This will help with variant prioritization and identification of functional SNPs. Finally, while our analysis was limited to early-stage T2D and healthy individuals, T2D is a progressive disease, and future studies can

benefit from analyses that aim to discover multiomic dynamic changes across diverse conditions and across different stages of development and disease state.

4.3.2 Integrating population data to explore genotype-phenotype landscape

Determining the phenotypic consequences of genetic variation affecting a gene and its regulatory network is further made difficult by the heterogeneity of the trait. In chapter 3, our integrative analysis identified *RFX6* gene and its extended regulatory networks as critical for β cell function — a finding validated through knockdown experiments in pancreatic pseudoislets. Although we showed that T2D GWAS variants are enriched to overlap β cell regulatory regions for the genes linked to the *RFX6*-containing module, we did not show a causal relationship between the genetic variation affecting *RFX6* expression and the risk of T2D and related phenotypes. A future approach to establish such a causal relationship can use islet eQTL results for RFX6 and population genetics data from biobanks in a mendelian randomization (MR) framework [78]. MR methods are increasingly being used to account for unobserved confounding in observational studies and have been successfully deployed for studying many cardiovascular diseases [83] as well as T2D [248]. Further discussion on using MR approach can be found in previous work [7, 83].

4.3.3 Network approaches to discover higher order interactions

Networks provide a powerful way of visualizing interactions in a complex system and infer relationships. For example, within a cell, interactions between different proteins or proteins with the DNA can be assayed to create protein-protein interaction or protein-DNA interaction networks. In this dissertation, we utilized networks to model two types of interaction. In the first instance in chapter 2, we used Cicero to model interaction between genomic elements using co-accessibility and estimate enhancer-gene interactions for gene regulation. This allowed us to link genetic

variation in the regulatory regions to putative target genes. In chapter 3, we used weighted gene co-expression networks to identify genes that perform in a coordinated manner in islet cell types across disease conditions. This allowed us to identify module level drivers of islet function and predict function of genes and the consequence of their disruption in the disease condition. Using these approaches in unison with other data modalities, we were able to derive insights into key disease-driving cellular mechanisms. However, biological networks and their topological analysis constitute a complex sub-field in itself and our approaches present a limited scope of what is possible. Below I highlight a few limitations of the approaches considered, and how we can address them in a future study.

First, these network approaches we considered rely on the guilt-by-association, a heuristic that has been broadly applied in genomics to characterize gene function [279]. However, the input data to these approaches is typically processed through several steps that may influence the final topology of the network. For example, coverage and quality of the data, normalization method used, the biological and experimental context in which the data was collected, confounding variables, the number of genes or features used to create the network, and finally how many samples we have. While WGCNA is quite robust to the choice of data normalization scheme and has been shown to work well for the sample sizes considered in our study, a future study may benefit from additional number of samples and potentially using alternative network creation approach discussed later in this section.

Second, creating networks or inferring biological insights from the network relationships often requires selecting arbitrary cut-off thresholds based on an a-priori requirement, but these assumptions might not be correct. For example, Cicero, used in chapter 2, generates a list of pairwise co-accessible genomic regions but which correlations are significant for further exploration requires selecting a cutoff. In our work, we selected cutoff using orthogonal validation of co-accessible peak pairs using experimental chromatin looping data. Clearly, such a validation approach will not be possible in absence of such data and would limit the usability of the approach. In chapter 3, we used WGCNA which constructs a scale-free network topology of genes.

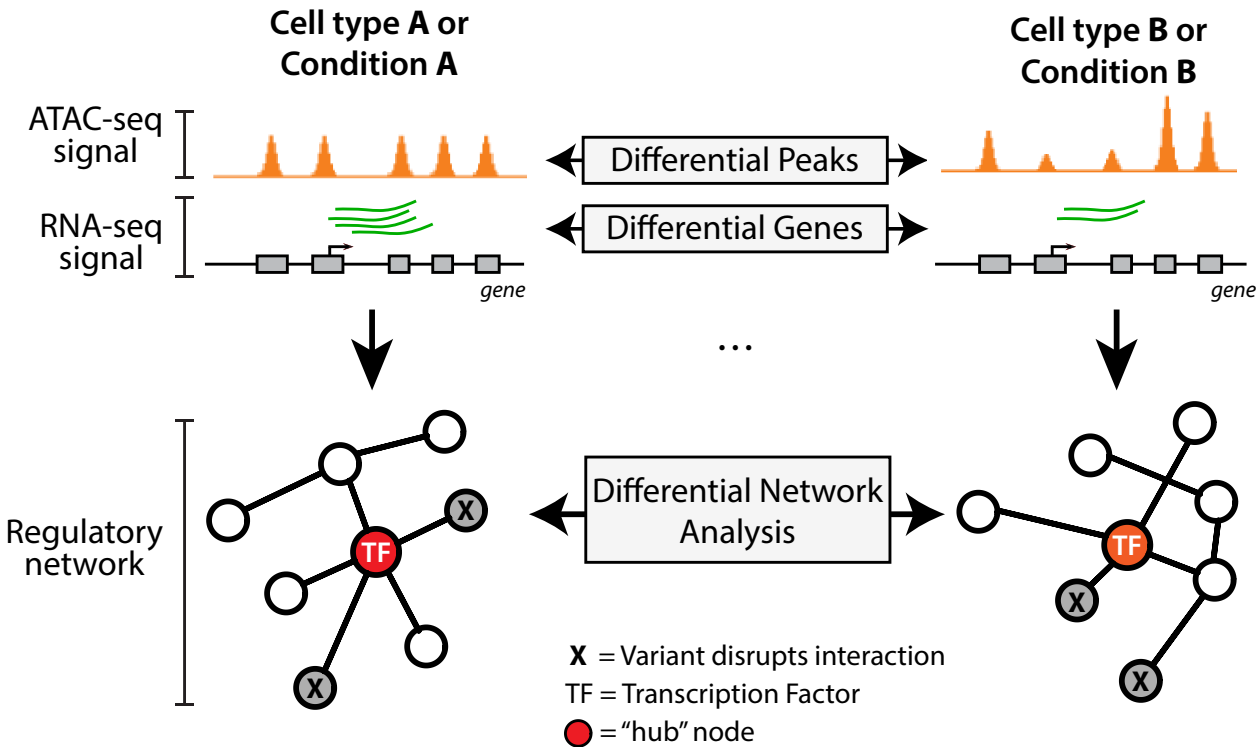


Figure 4.2: Future approaches to create and study networks from multiomics data. Using cell type or condition specific multiomic data such as gene expression (from RNA-seq) and chromatin accessibility (from ATAC-seq) can be used with robust partial correlation approaches to create variety of networks and infer hub genes and transcription factors, their interactions, and compare differences in the network wiring between the two conditions.

Although it tries to reduce the influence of arbitrary cut-off thresholds, a key parameter β (R^2 fit) must be chosen to meet the requirement of the scale-free topology — an assumption that has been shown to be applicable only to a small fraction of biological networks [189]. Further, one may choose different parameters that influence the size of the modules and how genes are clustered. In the absence of a statistical measure, designing experiments for sample size or evaluating the robustness and stability of modules becomes challenging.

Third, the networks created in our work primarily relied on pairwise correlation to compute similarity between different genes or genomic elements. Causal relationships, however, may not be directly related in a one-to-one fashion (first-order interaction) and may instead operate through a cascade of interactions. Because correlation networks represent correlation coefficients in a pairwise manner, they do not capture the higher-order interactions. For example, a

genetic variant in a regulatory element may influence the expression of a transcription factor which in turn modulates the expression of a target gene. One way to address this limitation is to use partial correlation networks [79, 221]. Partial correlation coefficients that form the basis of these networks are calculated for pairs of genes or genomic elements when all other variables are considered [195]. Consequently, partial correlations represent direct associations, whereas correlation analyses do not distinguish between indirect and direct associations. The partial correlation approach has been successfully applied to transcriptome data from yeast, cell lines, tumors, and case-control human disease tissues [120] and can be extended to our study as well.

Finally, creating networks for different cell types or disease conditions provides a unique opportunity to compare the networks and identify differences between the two conditions from a network biology perspective. For example, the WGCNA approach implements a module presevtation metric that compares the gene modules between two different networks allowing one to discover modules that capture the unique properties of the cellular state in a condition [100, 200]. Further, comparing networks can reveal context-specificity of genes and their gene regulatory pathways. Using data from our work, differential network approaches can be applied to transcriptome data from T2D and ND β and α cells to infer gene functions and pathways that might have been rewired.

In summary, network analysis approaches, including a few discussed in this work, present many possibilities for discovering biological insights. Inferring the regulatory network of RFX6 and its interacting partners, changes in the RFX6 network across cell types and disease conditions, as well as identification of other hub genes and module drivers of islet function are exciting questions to pursue.

4.3.4 Translating our findings into clinical knowledge and practice

The ultimate goal of any research finding is to translate it into practice and policy for better patient and healthcare outcomes. Complex diseases, such as T2D, present significant challenge to the translation pipeline due to the combination of risk factors that are difficult to study and

quantify and hence develop therapies for. For T2D, the first line of treatment involves environment control and lifestyle changes such as diet and exercise, followed by use of fasting-glucose lowering drugs as a metformin [224]. However, achieving a good metabolic control of T2D over long term requires a holistic understanding of the multifactorial pathogenesis of the disease and development of personalized medicine to deliver unique care and treatment options to each patient. Therefore, identifying biomarkers to quickly assay and establish the risk of the disease at early-stage or pre-diabetic stage and determine effective treatment options are critical goals for current and future studies focusing on T2D.

In this work, we identified RFX6 transcription factor as an important regulator ("biomarker") for β cell function and showed that disruption of *RFX6* gene expression leads to reduced insulin secretion in β cells. However, a future study might wish to build upon this finding and try to identify the exact genetic cause of RFX6 dysregulation and the cascade of pathways that are affected due to its disruption. Having this knowledge will allow one to develop specific therapeutic interventions or drug targets or screen patients to determine the best course of treatment.

Overall, moving forward in answering these questions would enable precision therapeutics and develop pharmacological therapies for T2D as we better understand how genetic variants affect islet cell functions in orchestrating disease mechanisms.

APPENDIX A

Genetic Effects on Liver Chromatin Accessibility Identify Disease Regulatory Variants

A.1 Foreword

This appendix is included to showcase applications of ATAC-seq to tissue from human liver – an example of a non-pancreatic organ that is critical to glucose homeostasis and other metabolic disorders. As a middle-author on this publication, I contributed to the project discussion, ATAC-seq data analysis (QC and processing), allelic bias mapping, figure design (Figure A.1A, and Figure A.3A), and manuscript methods writing and reviewing.

The full list of acknowledgements with the declaration of interest, funding agencies, and data and code availability can be found in online version of the manuscript referenced in [53]. The supplementary data and figures referenced in this chapter are not included in the dissertation but can be obtained online.

A.2 Abstract

Identifying the molecular mechanisms by which genome-wide association study (GWAS) loci influence traits remains challenging. Chromatin accessibility quantitative trait loci (caQTLs) help identify GWAS loci that may alter GWAS traits by modulating chromatin structure, but caQTLs have been identified in a limited set of human tissues. Here we mapped caQTLs in human liver

tissue in 20 liver samples and identified 3,123 caQTLs. The caQTL variants are enriched in liver tissue promoter and enhancer states and frequently disrupt binding motifs of transcription factors expressed in liver. We predicted target genes for 861 caQTL peaks using proximity, chromatin interactions, correlation with promoter accessibility or gene expression, and colocalization with expression QTLs. Using GWAS signals for 19 liver function and/or cardiometabolic traits, we identified 110 colocalized caQTLs and GWAS signals, 56 of which contained a predicted caPeak target gene. At the *LITAF* LDL-cholesterol GWAS locus, we validated that a caQTL variant showed allelic differences in protein binding and transcriptional activity. These caQTLs contribute to the epigenomic characterization of human liver and help identify molecular mechanisms and genes at GWAS loci.

A.3 Introduction

Genome-wide association studies (GWASs) have identified thousands of loci associated with complex traits, but the vast majority of variants fall outside the coding region. As a consequence, the causal variants, molecular mechanisms, target genes, and tissues of action for most loci have not been characterized. Studies of gene expression quantitative trait loci (eQTLs) have been instrumental in identifying plausible target genes and tissues for GWAS loci [93]. Chromatin conformation capture techniques, such as Hi-C, have identified variants at GWAS loci that physically interact with gene promoters [125]. However, additional approaches are needed to further pinpoint functional variants and to identify how these variants alter gene expression.

Variants at GWAS loci are enriched in transcriptional regulatory elements, which are typically marked by chromatin accessibility, in trait-relevant tissues [52]. Recent studies have identified chromatin accessibility QTLs (caQTLs), many of which overlap transcription factor (TF) binding sites and motifs [4, 25, 67, 92, 135, 145]. A subset of caQTLs are colocalized with eQTLs and GWAS loci, suggesting that variants at these loci influence gene expression and GWAS traits by altering chromatin accessibility [4, 25, 67, 92, 135, 145]. However, caQTLs have been mapped in

a limited set of human tissues. Mapping caQTLs in additional tissues and cell types is valuable to characterize the transcriptional regulatory mechanisms for a larger set of GWAS loci.

Liver is involved in numerous processes, including lipid metabolism, glucose storage, drug metabolism, and immune response [258]. Several studies have mapped eQTLs in liver tissue, and liver eQTLs are colocalized with GWAS loci for lipid, drug response, and other traits [35, 82, 244]. Lipid GWAS loci are enriched in regulatory chromatin states, including enhancers and promoters, in HepG2 hepatocytes [277]. QTLs for the active regulatory element histone marks H3K27ac and H3K4me3 have been identified in liver tissue, including a subset colocalized with liver eQTLs and GWAS loci [35]. Chromatin accessibility marks active regions containing H3K4me3 and H3K27ac, as well as poised promoters and enhancers that often do not display these histone marks [51, 141]. Consequently, mapping caQTLs in liver tissue can help functionally characterize GWAS loci that act by altering gene expression in liver.

In this study, we jointly mapped genotypes, gene expression, and chromatin accessibility in liver tissue from 20 organ donors and identified caQTLs in liver tissue. We predicted the impact of caQTL variants on TF binding and predicted caQTL target genes using four approaches. Finally, we used caQTLs, TF binding motifs, and target gene links to predict mechanisms at GWAS loci for multiple traits.

A.4 Results

A.4.1 Joint profiling of gene expression and chromatin accessibility in human liver tissue

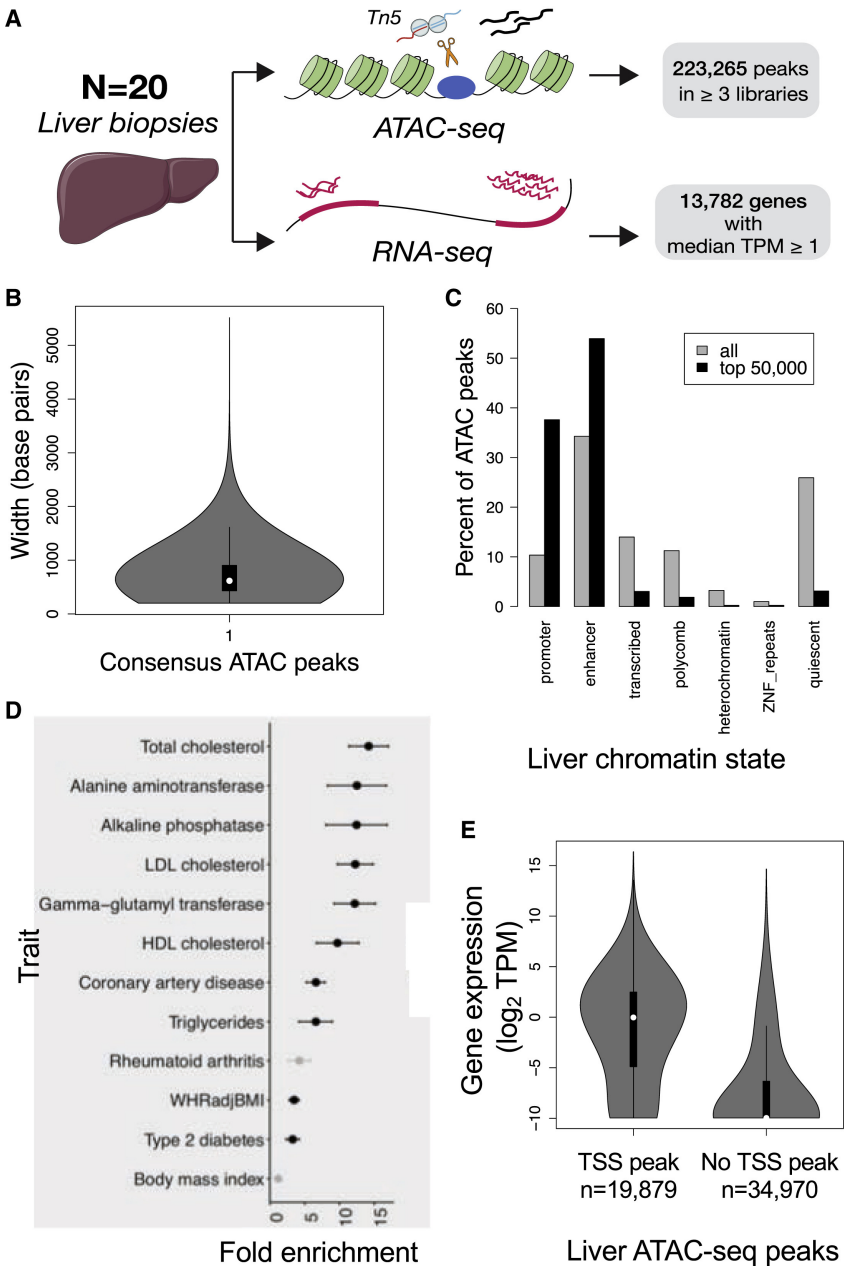
We obtained liver tissue from 20 deceased donors from the St. Jude liver bank and profiled gene expression using RNA-seq and chromatin accessibility using ATAC-seq [28] (Figure A.1A). All RNA libraries had RNA integrity number (RIN) of at least 6.5, and the median RIN value was 8, indicating that we extracted high-quality RNA. We identified 13,782 expressed genes, 13,317 of which are on autosomal chromosomes. By generating triplicate ATAC-seq libraries, we obtained

an average of 204 million high-quality autosomal ATAC-seq alignments (HQAAs) per sample and all libraries had >13% of HQAAs within peaks and TSS enrichment > 4, indicating that we generated libraries from tissue with high signal-to-noise. We identified 223,265 consensus accessible chromatin regions (peaks) with median peak width of 617 base pairs (Figure A.1B).

To predict the regulatory function of ATAC-seq peaks, we assigned peaks to liver tissue chromatin states from the Roadmap Epigenomics Project [52] and tested for enrichment of transcription factor (TF) binding sites and motifs in peaks. Among all 223,265 peaks, 34% were located in enhancers and 10% in promoters, and among the 50,000 most accessible peaks, ranked by median DESeq2 normalized count across individuals, 54% were located in enhancers and 38% in promoters (Figure A.1C). These results indicate that the strongest peaks were mostly located in promoters and enhancers, as expected, but that weaker peaks observed in at least three individuals were located in less well-characterized regions. We found 90 TF motifs enriched in peaks (E-value < 1×10^{-100}), including motifs for HNF4G (MIM: 605966), FOXA family members (HNF3), CEBPB [186] (MIM: 189965), the multifaceted protein CTCF [138] (MIM: 604167), and KLF family members, which regulate numerous processes in liver [193]. Of 17 TFs with ChIP-seq data in liver tissue [218], binding sites for all TFs were significantly enriched (permutation $p < 1 \times 10^{-3}$) in ATAC peaks, and 11 TFs had over 90% of their binding sites within ATAC peaks, similar to previous findings [51]. Taken together, ATAC peaks marked previously annotated transcriptional regulatory elements and TF binding sites in liver tissue.

We tested whether liver ATAC peaks were enriched for heritability of liver-relevant traits us-

Figure A.1 *(following page)*: Joint profiling of gene expression and chromatin accessibility in human liver tissue. (A) RNA-seq and ATAC-seq was performed in liver samples from 20 donors. (B) Distribution of consensus ATAC peak widths in base pairs. (C) Percent of consensus ATAC peaks by chromatin state in liver tissue from the Roadmap Epigenomics Project. All peaks, gray; 50,000 most accessible consensus peaks, black; quiescent represents unannotated regions. (D) Heritability enrichment of GWAS variants for multiple traits in all 223,265 liver ATAC peaks using stratified LD score regression. Points represent fold enrichment (proportion of heritability divided by proportion of SNPs within ATAC peaks) and error bars represent standard error. Significant enrichment (enrichment_p < 0.05), black; non-significant enrichment (enrichment_p > 0.05), gray. (E) Comparison of the distribution of expression between genes with and without an ATAC peak overlapping the transcription start site (TSS).


A
N=20
Liver biopsies
Tn5
ATAC-seq
RNA-seq
223,265 peaks in ≥ 3 libraries
13,782 genes with median TPM ≥ 1
B
Width (base pairs)
0
1000
2000
3000
4000
5000
1
Consensus ATAC peaks
C
Percent of ATAC peaks
0
10
20
30
40
50
60
all
top 50,000
promoter
enhancer
transcribed
polycomb
heterochromatin
ZNF_repeats
quiescent
Liver chromatin state
D
Trait
Total cholesterol
Alanine aminotransferase
Alkaline phosphatase
LDL cholesterol
Gamma−glutamyl transferase
HDL cholesterol
Coronary artery disease
Triglycerides
Rheumatoid arthritis
WHRadjBMI
Type 2 diabetes
Body mass index
0
5
10
15
Fold enrichment
E
Gene expression (log2 TPM)
−10
−5
0
5
10
15
TSS peak n=19,879
No TSS peak n=34,970
Liver ATAC-seq peaks

ing stratified LD score regression [88]. We observed significant heritability enrichment ($p < 0.05$) for 11 of 13 tested traits (Figure A.1D), and total cholesterol displayed the strongest enrichment (enrichment = 14.2, $p = 7.2 \times 10^{-5}$). We also observed strong enrichments (fold enrichment > 10) for LDL cholesterol and the liver enzymes. Heritability enrichment for cholesterol traits in liver regulatory elements marked by H3K4me1 has been previously identified [88], consistent with our results. As expected, we did not observe significant enrichment for rheumatoid arthritis and body mass index. These results indicate that liver ATAC peaks are enriched for genetic variants associated with liver-relevant traits.

We next determined whether genes with ATAC peaks at their transcription start site (TSS) were more likely to be expressed compared to genes without TSS peaks. A larger proportion of expressed genes had an ATAC peak directly overlapping the TSS (9,904 of 13,317, 74%) compared to non-expressed genes (9,975 of 41,532, 24%). Similarly, genes with a peak at the TSS tended to have higher expression than genes without a peak at the TSS (Figure A.1E; Kolmogorov-Smirnov test, $p < 2.2 \times 10^{-16}$). Together, the data provide high-quality gene expression and chromatin accessibility profiles in human liver tissue.

A.4.2 Identification of genetic variants associated with liver chromatin accessibility

We identified chromatin accessibility quantitative trait loci (caQTLs) using RASQUAL [145] and two distance thresholds: variants within 100 kilobases (kb) and within 1 kb of peak centers (Figure A.2A). Testing variants within 100 kb of peak centers, we identified significant caQTLs for 1,770 peaks (caPeaks), corresponding to 1,740 unique lead caQTL variants (Figure A.2A). For a substantial portion of caPeaks, the lead caQTL variant was within 1 kb of the caPeak center (n = 692, 39%, Figures S4B and S4C), and 654 of these 692 variants were within the caPeak. Testing variants within 1 kb of peak centers, we identified a significant caQTL for 3,123 peaks (Figure A.2A). We likely identified more caQTLs using a smaller window size because of a reduced multiple testing burden. We used this set of 3,123 caQTLs for all subsequent analyses unless noted

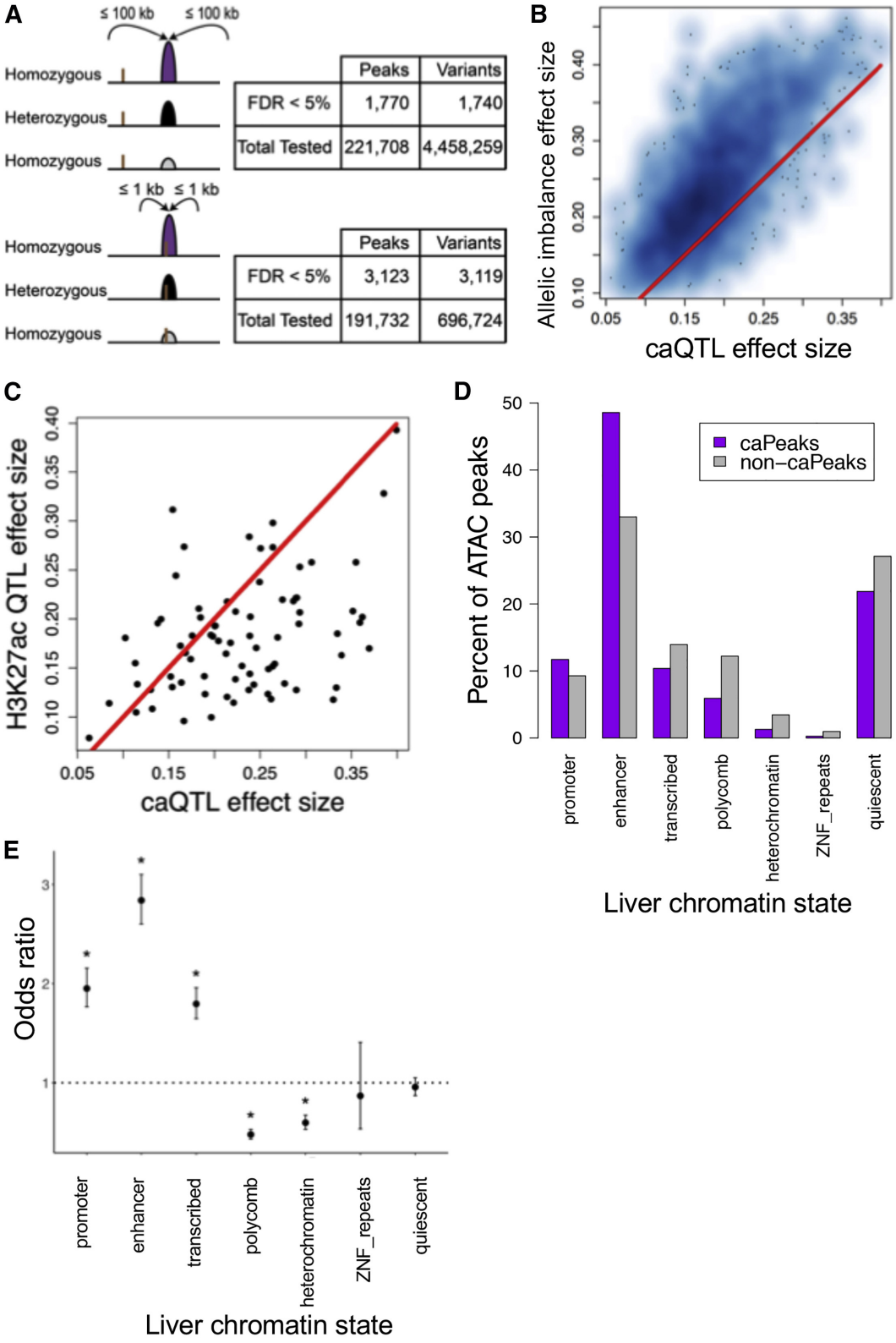
otherwise.

We next tested whether any caQTLs were strongly influenced by a single sample. Of the 3,123 caQTLs, 355 were no longer significant when one specific sample was removed, but remained significant when any other sample was removed. However, all but 6 remained nominally significant ($p < 0.05$). The most common influential sample (sample 459) accounted for only 48 of the 355 caQTLs (14%) and had the highest percent of HQAA within peaks, indicating that this sample has high quality. Taken together, the vast majority of the caQTLs are not strongly influenced by one sample.

To compare the RASQUAL model to another method that accounts for allelic mapping bias, we used WASP to remove reads exhibiting allelic mapping bias [262] and then calculated allelic imbalance (AI). 1,912 (81%) caQTLs identified by RASQUAL exhibited nominal (beta-binomial $p < 0.05$) and 1,112 (47%) exhibited genome-wide AI (FDR < 5%), all with the same direction of effect as the caQTL. Lead caQTL variants and representative AI variants exhibiting nominal AI showed strongly correlated effect sizes (Pearson's R = 0.75, Figure A.2B). AI effect sizes tended to be larger than caQTL effect sizes (Figure A.2B), possibly because AI was calculated using individual variants whereas caQTLs were identified using entire peaks. Therefore, we conclude that allelic mapping bias has no systematic effect on the caQTL results.

To determine the extent of shared genetic effects across different markers of transcriptional regulatory elements, we compared the 3,123 caQTLs to 921 H3K27ac QTLs from a recent report [35]. Of the 921 H3K27ac QTL peaks, 77 (8%) overlap a caPeak and have a lead variant in strong LD ($r^2 > 0.8$) with the caQTL lead. The 77 colocalized caQTL-H3K27ac QTL signals all showed

Figure A.2 *(following page)*: Identification and characterization of caQTLs. (A) caQTLs identified using variants within 100 kb or 1 kb of peak centers. (B) Comparison of effect sizes between caQTLs and simple allelic imbalance (Pearson's R = 0.75). The red line is the one-to-one line for caQTL effect sizes. (C) Comparison of effect sizes between caQTLs and H3K27ac QTLs (Pearson's R = 0.40). The red line is the one-to-one line for caQTL effect sizes. (D) Comparison of the number of caPeaks and non-caPeaks assigned to each chromatin state in liver tissue from the Roadmap Epigenomics Project. caPeaks, purple; non-caPeaks, gray; quiescent represents unannotated regions. (E) Enrichment of caQTL variants in liver chromatin states. Error bars represent 95% confidence intervals. * indicates significant enrichment ($p < 0.0071$).


A
≤ 100 kb
≤ 100 kb
Homozygous
Heterozygous
Homozygous
Peaks
Variants
FDR < 5%
1,770
1,740
Total Tested
221,708
4,458,259
≤ 1 kb
≤ 1 kb
Homozygous
Heterozygous
Homozygous
Peaks
Variants
FDR < 5%
3,123
3,119
Total Tested
191,732
696,724
B
Allelic imbalance effect size
caQTL effect size
C
H3K27ac QTL effect size
caQTL effect size
D
Percent of ATAC peaks
caPeaks
non−caPeaks
promoter
enhancer
transcribed
polycomb
heterochromatin
ZNF_repeats
quiescent
Liver chromatin state
E
Odds ratio
promoter
enhancer
transcribed
polycomb
heterochromatin
ZNF_repeats
quiescent
Liver chromatin state

the same direction of effect, and their effect sizes were moderately correlated (Pearson's R = 0.40, Figure A.2C). The largely distinct results may be due to the small sample sizes, analysis differences, and different genetic effects on the two epigenetic marks.

To predict the regulatory function of caPeaks, we compared caPeaks to liver tissue chromatin states from the Roadmap Epigenomics Consortium [52]. Relative to non-caPeaks (eigenMT-adjusted p > 0.5), caPeaks were more frequently located in enhancers (48.6% versus 33.0%) and promoters (11.7% versus 9.3%) (Figure A.2D). caQTL variants were significantly enriched in enhancers (OR = 2.9), promoters (OR = 2.0), and transcribed regions (OR = 1.8) and depleted in polycomb (OR = 0.5) and heterochromatin (OR = 0.6) states, which are associated with gene repression and presumably inaccessible chromatin (Figure A.2E). Taken together, caQTLs showed strong overlap with active transcriptional regulatory elements, with particularly strong enrichment in enhancers.

To identify liver caQTLs that would not be identified in blood, we counted liver caPeaks that overlapped macrophage ATAC peaks [4], using all macrophage ATAC peaks, not just caPeaks, due to limited sample sizes. Of the liver caPeaks, 1,268 (41%) overlapped a macrophage ATAC peak, suggesting that 59% of liver caQTLs mark regulatory elements not present in macrophages. This estimate is likely conservative because we included macrophage ATAC peaks that do not have caQTLs and demonstrates the importance of mapping caQTLs in a diverse set of tissues.

A.4.3 Disruption of transcription factor binding motifs by caQTLs

One way genetic variants may alter chromatin accessibility is by disrupting TF binding sites [67, 135, 145]. Among 4,585 variants within a caPeak and in strong LD with the caQTL lead, 3,132 (68%) variants altered the binding affinity of a TF motif (Figure A.3A). Of the 2,793 caPeaks containing a variant, 2,249 (81%) contained at least one variant predicted to disrupt a motif, and 602 of these contained 2 or more predicted motif-disrupting variants. Motifs for many TFs were disrupted by multiple caQTL variants, with 109 TF motifs disrupted by 20 or more variants. Disruption of motifs for 29 of these 109 TFs was significantly associated with caQTL status (logn

OR > 0, $p < 4.6 \times 10^{-4}$) (Figure A.3B), including TFs from the HNF, FOXA, and CEBP families [186], CTCF, and ATF2 (MIM: 123811). FOXA and CEBP factors can act as pioneer factors by binding to inaccessible chromatin and initiating the establishment of accessible chromatin [173] and ATF2 can alter chromatin structure to activate or repress transcription [149], suggesting that this approach identifies TFs that may influence chromatin accessibility.

To investigate how often TFs bind the more accessible allele, we compared alleles associated with higher chromatin accessibility to the motifs. Among 7,629 motifs for all TFs, the more accessible allele matched the motif better for 4,770 motifs (63%, binomial $p < 4.1 \times 10^{-107}$). Similarly, among 3,132 motifs for the highest expressed TF at each variant, the more accessible allele matched the motif better for 1,953 motifs (62%, binomial $p < 8.0 \times 10^{-44}$). When restricting analysis to 993 observations of the 29 TFs for which motif disruption is associated with caQTL status, the more accessible allele matched the motif better for 834 motifs (84%, binomial $p < 5.1 \times 10^{-111}$). TFs exhibited variation in the percent of motifs that matched better to the more accessible allele (Figure A.3C). For 11 TFs, including HNF4A (MIM: 600281), ATF4 (MIM: 604064), ERF (MIM: 611888), and FOXA2 (MIM: 600288), more than 90% of stronger motif matches corresponded to the more accessible allele, while for SPI1 (MIM: 165170) only 56% of stronger motif matches corresponded to the more accessible allele. These results suggest that TFs typically, but not always, bind to the more accessible allele.

A.4.4 Identifying putative target genes for caPeaks

Connecting caPeaks to their target genes is challenging, particularly when the caPeaks are distal to transcription start sites (TSSs). Individual approaches for identifying target genes have limitations and may not always show a direct regulatory relationship between a caPeak and gene. To address these challenges, we used four approaches to connect caPeaks to genes (Figure A.4A).

First, we identified caPeaks proximal (−2 kb/+1 kb) to TSSs of genes expressed in liver. Of 3,123 total caPeaks, 114 (4%) were proximal to the TSS of at least 1 gene. Among these 114 caPeaks, 15 were proximal to the TSS of two or three genes. This approach identified 131 unique caPeak-gene

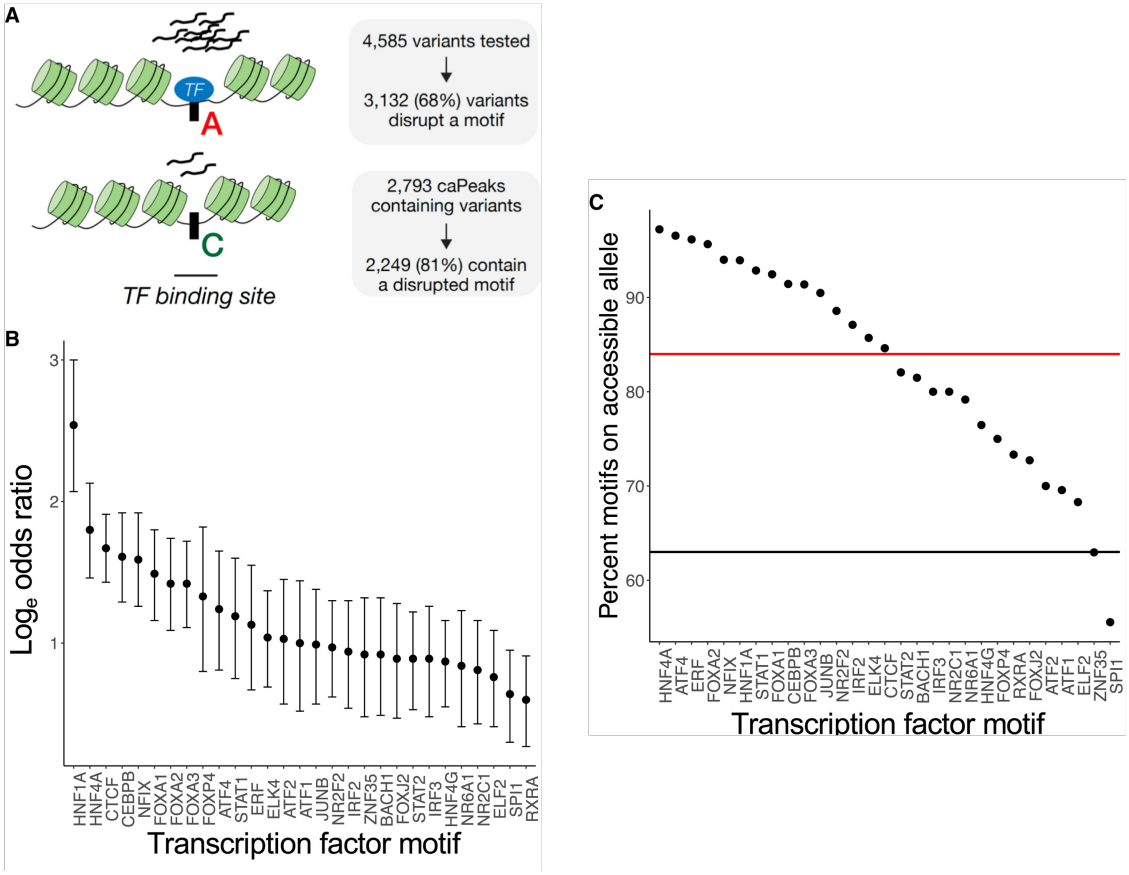


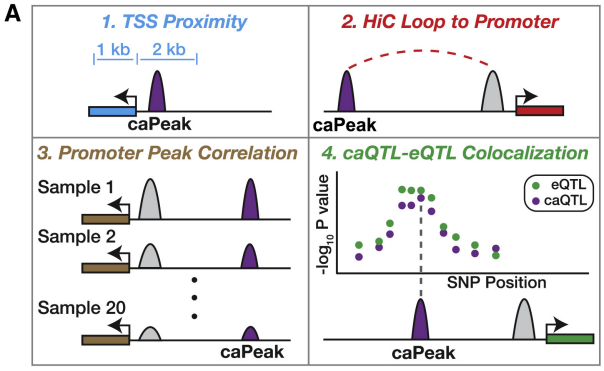
Figure A.3: Disruption of TF binding motifs by caQTL variants. (A) Allele affinities for TF binding and chromatin accessibility for variants within caPeaks and in strong LD with the caQTL lead variant (r^2 > 0.8). (B) Association of caQTL status with motif disruption status. Only the 109 TFs with at least 20 motifs disrupted by caQTL variants were included in the analysis, and only the 29 significant associations ($p < 4.6 \times 10^{-4}$) are shown. Error bars indicate 95% confidence intervals. (C) Percent of disrupted motifs for which the allele with higher chromatin accessibility matched the motif better. Percents are shown for the 29 TFs that had at least 20 motifs disrupted by caQTL variants. Black line, percent for all disrupted motifs across all tested TFs; red line, average percent across the 29 TFs.

connections (Figure A.4A).

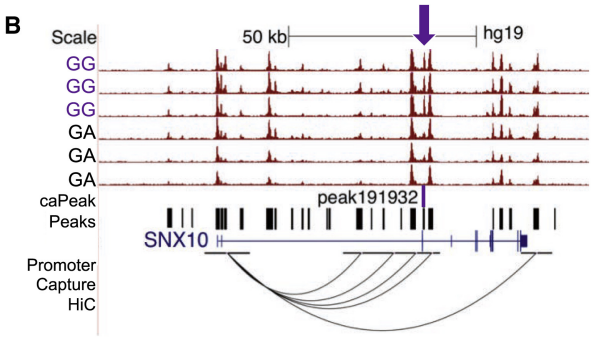
Second, we used liver tissue promoter capture Hi-C2 to identify caPeaks that physically interact with gene promoters. We identified 329 distal caPeaks (>15 kb from any promoter as defined in the Hi-C analysis) that interact with promoters for 451 genes, including a caPeak that interacts with the promoter of SNX10 (MIM: 614780; Figure A.4B and S6A). The caPeak near SNX10 was identified even though only two genotypes were observed in these samples, demonstrating that caQTL effect sizes can be large. Among caPeaks that overlapped the promoter of one gene and interact with the promoter of another gene, we identified an additional 104 caPeaks that interact with promoters of 190 genes. Combining promoter-distal and promoter-promoter interactions, we identified 697 caPeak-gene connections (Figure A.4A).

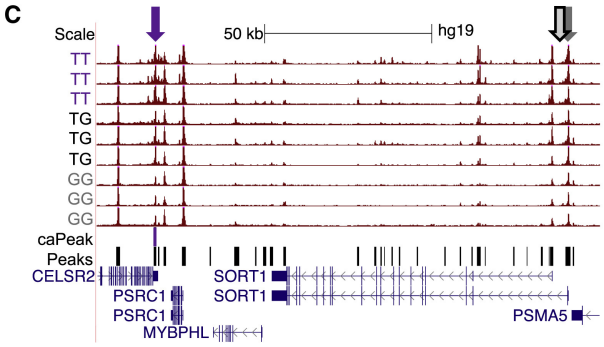
Third, we identified caPeak sizes that either correlated with expression level of nearby genes or with the size of ATAC peaks at promoters. More caPeaks were correlated with promoter ATAC peaks than with gene expression level; 120 caPeaks were significantly correlated (FDR < 5%) with promoter ATAC peaks while only 2 caPeaks were correlated with gene expression (FDR < 5%), resulting in 121 unique caPeaks because gene RP11-101E14.2 had both types of correlations (Figure A.4A). When using the same p value threshold for both analyses ($p < 2.9 \times 10^{-4}$), 5 additional caPeaks were correlated with gene expression. As an example at a regulatory element previ-

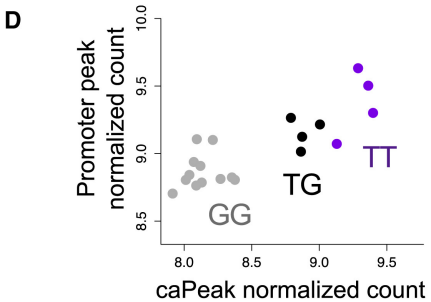
Figure A.4 *(following page)*: Prediction of target genes for caPeaks using four approaches. (A) Illustrations of four approaches to predict caPeak target genes. (B) Hi-C chromatin contact shown as an arc between caPeak191932 and the SNX10 promoter. Selected ATAC-seq signal tracks are shown for each caQTL genotype of rs12534816. More accessible homozygotes, purple; heterozygotes, black. (C) Genome browser image showing the correlation across rs12740374 genotypes of caPeak9372 and a peak at a SORT1 promoter. The purple arrow indicates the caPeak and the gray arrow indicates the promoter peak. (D) The same peak correlation with points representing normalized peak counts of individual samples colored by rs12740374 genotype. (E and F) SORT1 eQTL associations at the signal colocalized with the caQTL for caPeak9372 (E) and caQTL associations with caPeak9372 (F). In both plots, the caQTL lead variant within 1 kb of the peak center is indicated by a purple diamond and LD is based on 1000G phase 3 Europeans. (G) Comparison of directions of effect among all colocalized caQTL and eQTL signals. The A allele represents the more accessible allele than C, and more red marks indicate higher gene expression. (H) UpSet plot comparing the number of shared and unique caPeak-gene links identified by the four approaches. It is not possible for a caPeak-gene pair to be predicted using all four methods because if a caPeak is TSS proximal, it cannot form a Hi-C loop with the same gene and it cannot be a distal caPeak correlated with the promoter peak for the same gene.

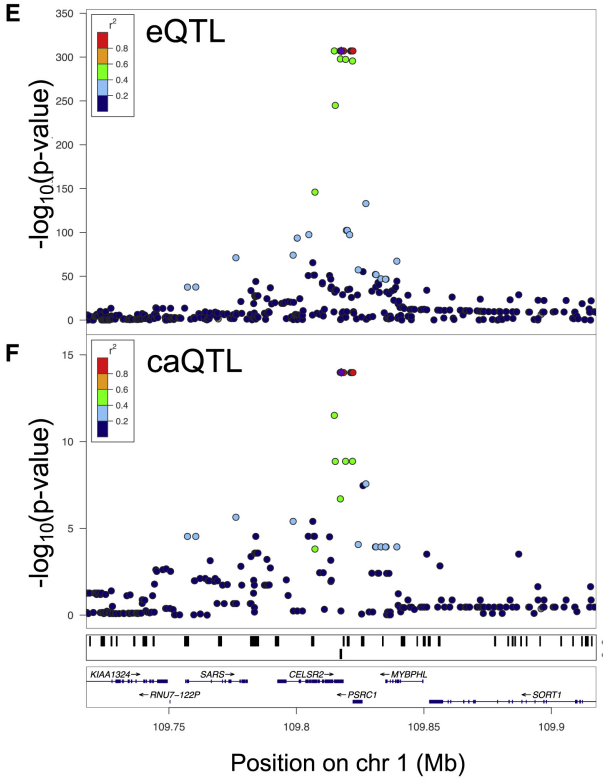


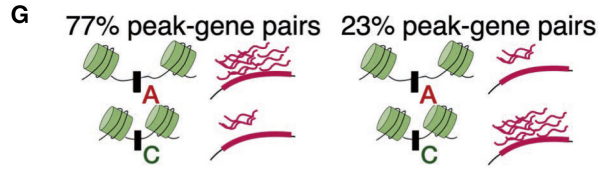
Method	caPeaks	Genes	caPeak-Gene Pairs
1	114	131	131
2	432	629	697
3	121	185	196
4	415	463	506

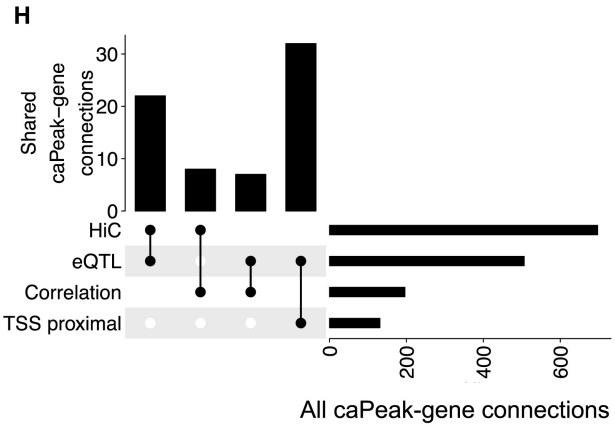












ously shown to regulate SORT177 (MIM: 602458), caPeak9372 is positively correlated with a peak proximal to a SORT1 TSS (peak9400, Spearman rho = 0.76, $p < 1.6 \times 10^{-4}$; Figure A.4C-D) and nominally correlated with SORT1 expression (Spearman rho = 0.69, $p < 1.2 \times 10^{-3}$). The vast majority of peak-peak correlations (167 of 173, 97%) are positive, suggesting that higher caPeak accessibility is usually associated with higher accessibility of connected promoter peaks. Using either caPeak-promoter peak or caPeak-gene correlations, we identified 196 caPeak-gene connections (Figure A.4A).

Finally, we identified caQTLs for which the lead variant exhibited high LD ($r^2 > 0.8$) with an eQTL lead variant for 15,418 autosomal genes from a liver tissue eQTL meta-analysis of 1,183 individuals [82]. Of 3,119 unique caQTL lead variants, 414 (13%) were in strong LD with at least 1 eQTL lead variant, which is similar to the percentage reported in a previous caQTL study [67]. Among caQTL lead variants, 71 were in strong LD with more than one eQTL lead variant, suggesting that some caPeaks may affect expression of multiple genes. In total, we identified 463 target genes for 415 caPeaks, representing 506 unique caPeak-gene connections (Figure A.4A). For example, we identified a caQTL signal with the same variants as an eQTL signal for SORT1 (Figure A.4E-F). At connected loci, the allele associated with higher chromatin accessibility was usually associated with higher gene expression (390 of 506 loci, 77%; Figure A.4G), suggesting caPeaks frequently act as promoters or enhancers to gene expression. We obtained a similar result when restricting to caQTL variants associated with only one peak and colocalized with eQTL variants associated with only one gene (273 of 337 loci, 81%). Of the 506 caQTL-eQTL signals colocalized based on LD, 28 showed strong evidence of colocalization using coloc (PP4 > 0.8) [95], and an additional 48 showed suggestive evidence of colocalization (PP4 > 0.5 but < 0.8). Of the 430 signals that did not show suggestive evidence of colocalization, 409 (95%) did not have sufficient power to detect colocalization (PP0+PP1+PP2 > 0.5) and no signals showed evidence of separate, but not colocalized signals (PP3 > 0.5). Therefore, we conclude that the study is underpowered to detect colocalizations using coloc.

Together the four methods identified a total of 1,461 caPeak-gene connections, although the

approaches showed low overlap. Only 69 caPeak-gene connections were predicted by two methods, and no connections by three methods, likely due to the low power of many of the approaches (Figure A.4H). The 69 caPeak-gene associations consist of 67 unique caPeaks and 67 unique genes; two caPeaks had two target genes. It is not possible for a caPeak-gene pair to be predicted using all four methods because if a caPeak is TSS proximal, it cannot be found within the distal end of a Hi-C loop >15 kb from the same TSS and it cannot be a distal caPeak correlated with the promoter peak for the same gene. Thus, the only method that can corroborate TSS proximity is caQTL-eQTL colocalization. Of the 131 caPeak-gene connections identified by TSS proximity, 32 (24%) were supported by caQTL-eQTL colocalization. In addition, when considering peak correlations for which the distal caPeak was tested by other approaches, 98 of 307 (32%) caQTL-eQTL colocalizations and 108 of 436 (25%) Hi-C loops showed at least nominal ($p < 0.05$) evidence. These methods are limited by power and technical factors, suggesting that the 69 caPeak-gene connections identified by two methods may be a conservative estimate. This integrated approach predicted a target gene for 861 of 3,123 caPeaks (28%), suggesting that caPeaks frequently interact with genes.

A.4.5 Prediction of regulatory mechanisms at GWAS loci

To identify genetic variants that may influence disease by altering chromatin accessibility, we identified colocalized caQTL and GWAS signals, based on strong LD ($r^2 > 0.8$) between lead caQTLs and lead GWAS variants. Using GWAS variants for 19 traits relevant to liver function and cardiometabolic traits from the NHGRI-EBI GWAS catalog [30], we identified 110 potentially colocalized caQTL and GWAS signals, corresponding to 111 caPeaks, because one caQTL signal was associated with two caPeaks. We identified at least one colocalized caQTL for 15 of the 19 traits, and of the GWAS signals for these traits, liver enzymes showed the highest percentage of potentially colocalized caQTLs (14 signals, 18%) (Table A.1). For traits with at least 5 GWAS-caQTL signals, we identified a relatively high percentage of colocalized signals (>5%) for total cholesterol and LDL cholesterol, consistent with the involvement of liver in lipid metabolism

[258]. As a negative control, we observed a relatively low percentage (<2%) of GWAS signals colocalized with liver caQTLs for height and rheumatoid arthritis.

Table A.1: Colocalized GWAS-caQTL signals by trait

Trait	Number of GWAS signals [a]	Number of colocalized caQTL-GWAS signals [b]	Percent of colocalized caQTL-GWAS signals [c]
Liver enzymes	77	14	18.2
Total cholesterol	292	18	6.2
Glucose	54	3	5.6
Insulin	18	1	5.6
LDL cholesterol	240	13	5.4
Bilirubin	20	1	5.0
HDL cholesterol	314	12	3.8
C-reactive protein	81	3	3.7
Triglycerides	279	10	3.6
Cardiovascular disease	454	14	3.1
Body mass index	986	29	2.9
Blood pressure	1,540	38	2.5
Type 2 diabetes	268	5	1.9
HbA1c	66	1	1.5
WHRadjBMI	209	3	1.4
Glycated albumin	2	0	0.0
Liver injury	17	0	0.0
NAFLD	9	0	0.0
Serum albumin	15	0	0.0

LDL, low-density lipoprotein; HDL, high-density lipoprotein; WHRadjBMI, waist-hip ratio adjusted for BMI; NAFLD, non-alcoholic fatty liver disease.

[a] Counted as lead GWAS variants not in high LD ($r^2 < 0.8$) with another.
[b] Colocalized if the caQTL lead variant was in strong LD ($r^2 > 0.8$) with the GWAS lead.
[c] Percent of all GWAS signals that are colocalized with a caQTL.

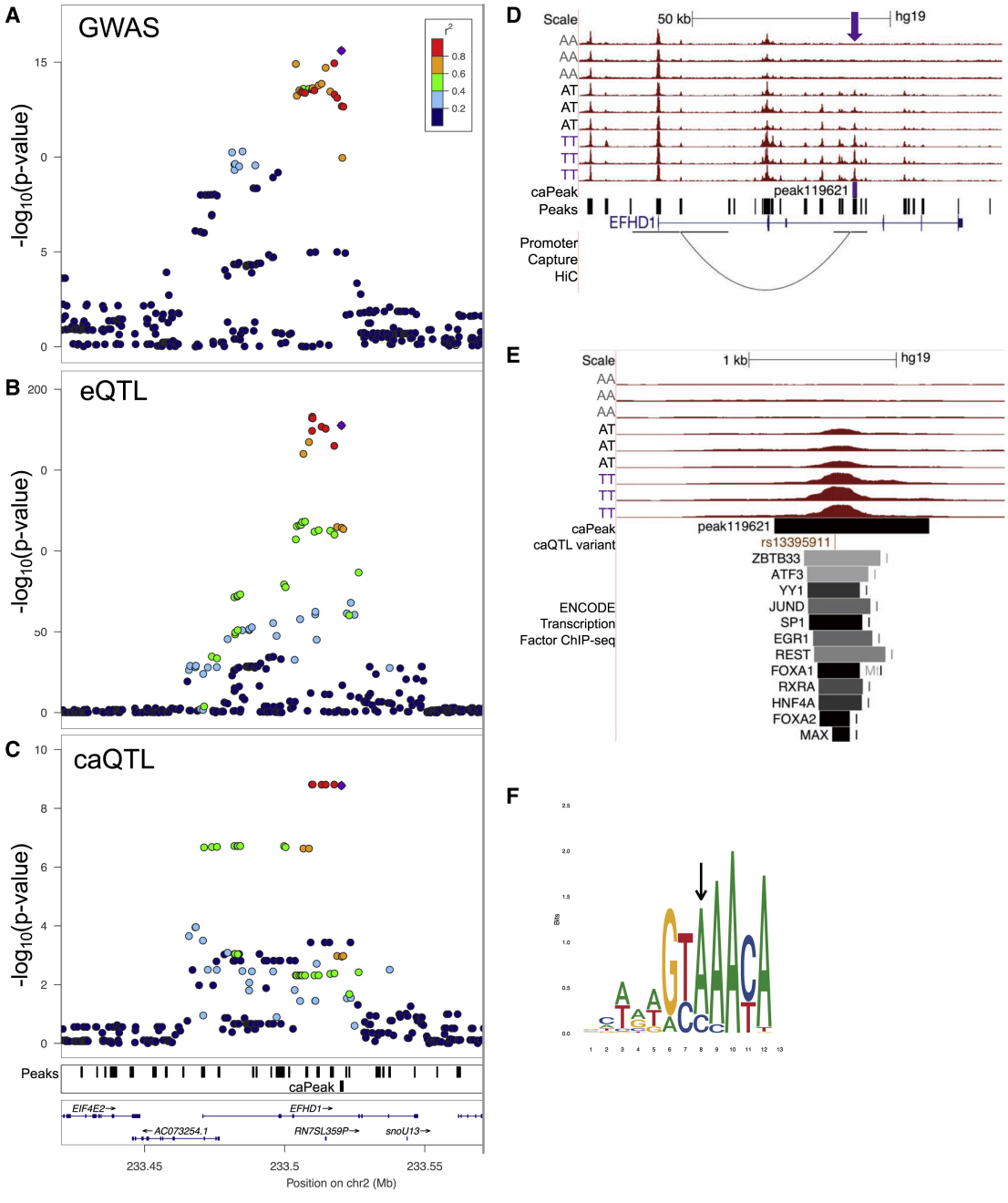
Only 26 of the 143 (18%) liver caQTL-GWAS colocalizations were observed using blood caQTL datasets. For liver enzymes, total cholesterol, and LDL cholesterol, respectively, only 3 of 14, 3 of 18, and 2 of 13 liver caQTL-GWAS colocalizations were observed in blood. GWAS signals for liver enzymes were colocalized with a higher percentage of liver caQTLs (0.51%) than each of the blood cell type caQTLs (0.06%–0.12%), whereas GWAS signals for rheumatoid arthritis were colocalized with a higher percentage of blood caQTLs (0.09%–0.21%) than liver caQTLs (0.06%). However, many of these colocalization differences between liver and blood may be due to limited caQTL

sample sizes. Larger studies using identical caQTL pipelines are needed to robustly identify cell type-specific caQTL-GWAS colocalizations.

To identify plausible regulatory mechanisms at GWAS loci, we integrated our GWAS-colocalized caQTLs with TF motif-disrupting variants and predicted caPeak target genes. Of the 111 caPeaks at potentially colocalized caQTL-GWAS signals for liver function or cardiometabolic traits, 85 harbored a TF motif-disrupting variant, 56 had a predicted target gene, and 45 of these overlapped with both types of data. The gene with a TSS closest to the GWAS lead variant was predicted to be a target gene for 25 of 56 caPeaks (45%).

We identified seven liver function or cardiometabolic GWAS-caQTL colocalized signals with strong evidence of regulatory mechanisms. At these GWAS loci, the caPeak had a target gene identified by two approaches and harbored TF motif-disrupting variants (Table A.2). We identified colocalized caQTL, eQTL, and GWAS signals and a correlated caPeak-promoter peak pair (Table A.2; Figure A.4C–F) at the *SORT1* locus associated with LDL cholesterol for which the alternate allele (rs12740374-T) has been shown to create a *CEBP* binding site and increase hepatic SORT1 expression [183]. At a less well-characterized locus, the caQTL signal with lead variant rs13395911 associated with caPeak119621 is colocalized with GWAS signals for plasma liver enzyme levels in European [40] and Asian [129] individuals and an eQTL for *EFHD111* (MIM: 611617; Figure A.5A-C and S7). Increased accessibility corresponds to higher *EFHD1* expression level and higher liver enzyme levels. caPeak119621 physically interacts with the promoter of *EFHD1* in liver tissue promoter capture Hi-C data [125] (Figure A.5D), further suggesting that caPeak119621 may

Figure A.5 *(following page)*: A plausible regulatory mechanism at the EFHD1 locus for plasma liver enzyme levels. (A–C) GWAS association with plasma levels of the liver enzyme alanine transaminase in Japanese individuals (A), eQTL association for EFHD1 (B), and caQTL associations for caPeak119621 (C). For all three plots, the caQTL lead variant within 1 kb of the peak center is indicated by a purple diamond and LD is based on 1000G phase 3 East Asians (A) or Europeans (B and C). Additional plots are shown in Figure S7. (D) Hi-C chromatin contact shown as an arc between caPeak119621 and the *EFHD1* promoter. Selected ATAC-seq signal tracks are shown for each rs13395911 genotype. More accessible homozygotes, purple; heterozygotes, black; less accessible homozygote, gray. (E) Transcription factor ChIP-seq peaks in liver tissue from ENCODE that overlap caPeak119621. (F) Sequence logo plot for the FOXA2 motif s disrupted by caQTL variant rs13395911 (arrow). The motif match is shown on the negative strand, and variant alleles in (D) and (E) are shown on the positive strand.


A
GWAS
r^2
0.8
0.6
0.4
0.2
-log10(p-value)
B
eQTL
C
caQTL
Peaks
caPeak
EIF4E2
EFHD1
AC073254.1
RN7SL359P
snoU13
233.45
233.5
233.55
Position on chr2 (Mb)
D
Scale
50 kb
hg19
AA
AT
TT
caPeak
Peaks
peak119621
EFHD1
Promoter Capture HiC
E
1 kb
caQTL variant
rs13395911
ENCODE Transcription Factor ChIP-seq
ZBTB33
ATF3
YY1
JUND
SP1
EGR1
REST
FOXA1
RXRA
HNF4A
FOXA2
MAX
F
Bits

affect *EFHD1* expression. CaPeak119621 does not overlap an ATAC peak in macrophages [4]. The peak overlaps ChIP-seq peaks for 12 TFs in liver (Figure A.5E), and rs13395911 disrupts motifs for eight TFs expressed in liver. The motif with the largest difference between rs13395911 alleles is for *FOXA2*, and the allele with higher chromatin accessibility matches the motif better (Figure A.5F). These and other connections provide potential regulatory mechanisms linking variants to regulatory element, transcription factors and genes that may influence the GWAS traits.

A.4.6 Identification of a putative functional variant at the *LITAF* locus

Near *LITAF* (MIM: 603795), which encodes lipopolysaccharide (LPS)-induced TNF factor, we identified a caQTL signal for caPeak75869 and tested variants for allelic differences in transcriptional activity and protein binding. This caQTL signal is potentially colocalized with a GWAS signal for LDL cholesterol [140] and an eQTL signal for *LITAF* [82] (Figure A.6A-B and S8). caPeak75869 loops to the promoter of LITAF in liver tissue promoter capture Hi-C [125] (Figure A.6C). caPeak75869 contains the lead caQTL variant rs57792815 (caQTL $p < 5.0 \times 10^{-17}$) and two additional variants in strong LD with the caQTL lead, rs3784924 ($r^2 = 0.95$) and rs11644920 ($r^2 = 0.98$). The haplotype associated with higher accessibility consists of the rs57792815-T, rs3784924-A, and rs11644920-A alleles. We tested a 666-bp DNA construct spanning the three variants for haplotype differences in transcriptional activity using luciferase reporter assays, testing the construct in two orientations relative to a minimal promoter. Given that *LITAF* is involved in lipopolysaccharide (LPS)-stimulated immune response [184], we tested transcriptional activity in four cell types: HepG2 hepatocytes, THP-1 monocytes, THP-1 differentiated macrophages, and LPS-stimulated THP-1 macrophages. In all four cell types, the forward orientation construct containing the alleles associated with higher accessibility showed significantly higher transcriptional activity than the construct containing the other alleles, with the strongest differences observed in hepatocytes (fold change = 2.49, p = 2 × 10−4) and LPS-stimulated macrophages (fold change = 1.39, p = 7 × 10−4; Figure A.6D). The same haplotype showed significantly higher transcriptional activity in the reverse orientation for hepatocytes (p = 1 × 10−4) and unstimulated macrophages

Table A.2: Selected caQTLs at GWAS loci

caQTL variant	caPeak	GWAS variant	GWAS trait	LD r^2 [a]	Gene	Methods [b]	caQTL, eQTL directions [c]
rs12740374	peak9372	rs12740374	LDL cholesterol	1.00	SORT1	eQTL, Corr	D, D
rs17276527	peak13768	rs4077194	HDL cholesterol	1.00	RALGPS2	eQTL, HiC	D, D
rs13395911	peak119621	rs13395911	ALT	1.00	EFHD1	eQTL, HiC	I, I
rs2232015	peak9185	rs1730859	LDL cholesterol	0.97	PRMT6	TSS, eQTL	D, D
rs2037517	peak71475	rs832890	Pulse pressure	0.90	PLEKHO2	eQTL, HiC	D, D
rs12677006	peak205272	rs1906672	Sys. blood pressure	0.89	DDHD2	eQTL, HiC	I, I
rs57792815	peak75869	rs34318965	LDL cholesterol	0.81	LITAF	eQTL, HiC	I, I

Loci are shown for shared caQTL-GWAS signals if the caPeak was linked to a target gene by two methods and if the caPeak harbored motif-disrupting variants. ALT, alanine aminotransferase levels; Sys, systolic.

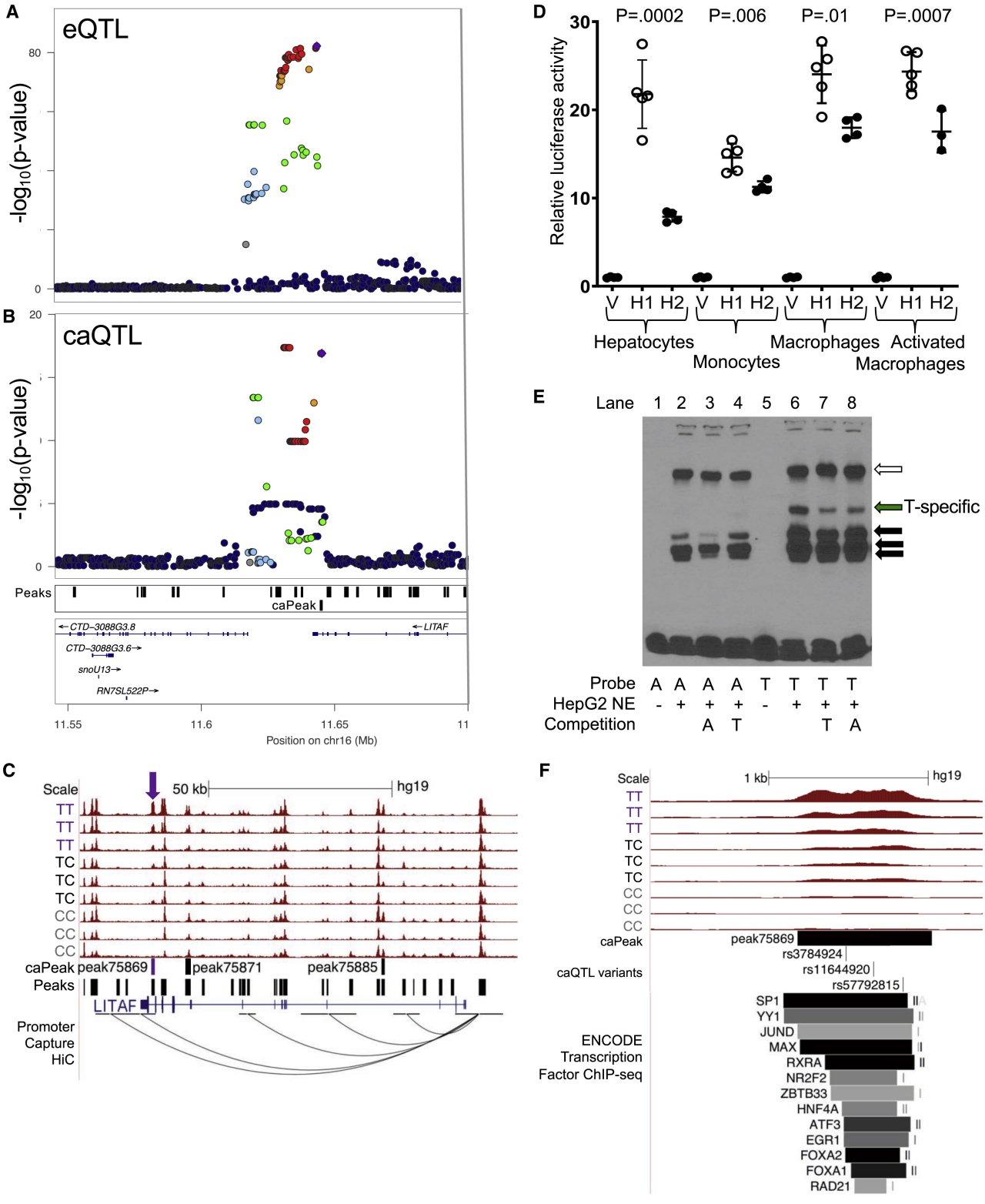
[a] LD r^2 between the caQTL and GWAS lead variants.

[b] Methods that linked the caPeak to a gene. Corr, correlation between caPeak and promoter peak accessibility.

[c] Direction of chromatin accessibility and gene expression relative to the allele associated with an increase in the GWAS trait, where "I" indicates increased and "D" indicates decreased accessibility or expression.

(p = 0.02) and a trend toward higher transcriptional activity in the other cell types (Figure S8G). Although allelic differences were observed in all four cell types, caPeak75869 does not overlap an ATAC peak in macrophages4. We next tested each of the three haplotype variants for allelic differences in protein binding using nuclear extract from HepG2 cells. Only rs11644920 showed allele-specific binding, with the T allele showing increased binding (Figure A.6E). caPeak75869 contained liver ChIP-seq binding sites for numerous TFs and all three variants within the peak disrupted motifs (Figure A.6F). We focused on the motif disrupted by rs11644920 because it was the only variant that showed allelic differences in binding in the EMSA results. Variant rs11644920 disrupted a motif for ATF2, and the A allele matched the motif better, which is also the allele associated with higher chromatin accessibility. This result contrasts the EMSA results, which showed greater binding for the T allele. Together, these results suggest that altered transcription factor binding at rs11644920 and increased chromatin accessibility of the regulatory element marked by caPeak75869 may lead to increased transcriptional activity and higher *LITAF* expression.

Figure A.6 *(following page)*: Identification of a putative functional variant at the LITAF locus for LDL cholesterol. (A and B) eQTL association for LITAF (A) and caQTL associations for caPeak75869 (B) at an LDL cholesterol GWAS signal. In both plots, the caQTL lead variant within 1 kb of the peak center is indicated by a purple diamond, and LD is based on 1000G phase 3 Europeans. Additional plots are shown in Figure S8. (C) Hi-C chromatin contact between caPeak75869 and the LITAF promoter. Selected ATAC signal tracks are shown for each rs57792815 genotype. More accessible homozygotes, purple; heterozygotes, black; less accessible homozygotes, gray. (D) Transcriptional activity of a 666-bp DNA element spanning caPeak75869 and containing rs3784924, rs11644920, and rs57792815 in HepG2 hepatocytes, THP-1 monocytes, THP-1 differentiated macrophages, and LPS-stimulated THP-1 macrophages. The DNA element was tested in the forward orientation relative to the genome (reverse orientation in Figure S8G). V, empty vector; H1, haplotype 1 of more accessible alleles rs3784924-A, rs11644920-A, and rs57792815-T; H2, haplotype 2 of less accessible alleles rs3784924-G, rs11644920-T, and rs57792815-C. Symbols represent 4–5 independent clones for each haplotype tested in duplicate wells; bars indicate mean ± standard deviation; p values from t tests of allelic differences. (E) EMSA using HepG2 nuclear extract (NE) shows allelic differences in protein binding for rs11644920. rs3784924 and rs57792815 are shown in Figure S8H. Green arrow, band represents T-allele-specific binding; black arrows, T-allele-preferential binding; white arrow, non-specific binding. Competition probes were unlabeled and included in 10-fold excess. (F) TF ChIP-seq peaks in liver tissue from ENCODE that overlap caPeak75869.


A
eQTL
-log10(p-value)
B
caQTL
-log10(p-value)
Peaks
caPeak
CTD-3088G3.8
LITAF
CTD-3088G3.6
snoU13
RN7SL522P
11.55
11.6
11.65
Position on chr16 (Mb)
C
Scale
50 kb
hg19
TT
TC
CC
caPeak
peak75869
peak75871
peak75885
Peaks
LITAF
Promoter Capture HiC
D
P=.0002
P=.006
P=.01
P=.0007
Relative luciferase activity
V H1 H2
Hepatocytes
Monocytes
Macrophages
Activated Macrophages
E
Lane 1 2 3 4 5 6 7 8
T-specific
Probe A A A A T T T T
HepG2 NE - + + + - + + +
Competition A T T A
F
Scale
1 kb
hg19
caPeak
peak75869
caQTL variants
rs3784924
rs11644920
rs57792815
ENCODE Transcription Factor ChIP-seq
SP1
YY1
JUND
MAX
RXRA
NR2F2
ZBTB33
HNF4A
ATF3
EGR1
FOXA2
FOXA1
RAD21

A.5 Discussion

We profiled chromatin accessibility in 20 individuals and identified caQTLs in human liver tissue. caQTL variants frequently disrupt TF binding motifs, and alleles that better match a motif often have higher chromatin accessibility, consistent with TFs stabilizing chromatin in an accessible state. We identified 1,461 putative caPeak-gene links using four approaches, suggesting that caPeaks frequently regulate gene expression. We identified 110 caQTLs at GWAS signals, including 56 with a predicted caPeak target gene, identifying regulatory mechanisms that may be responsible for trait variation. Among variants at a colocalized caQTL, eQTL, and LDL cholesterol GWAS signal near LITAF, one variant showed allelic differences in transcriptional activity and in vitro TF binding. This study contributes to the epigenomic characterization of human liver tissue and will aid in functional characterization of GWAS loci that act in liver.

Combining caQTLs, caPeak-gene links, and disrupted TF motifs helps identify mechanisms at GWAS loci. At the well-characterized *SORT1* GWAS locus for lipid and cardiovascular traits [183], we showed that the previously described functional variant rs12740374 is associated with chromatin accessibility and that the caPeak containing this variant is correlated with a peak at the *SORT1* promoter. We also identified plausible regulatory mechanisms at less well-characterized loci. At a GWAS signal for BMI [136] and LDL cholesterol [140], we identified a caQTL potentially colocalized with a *PRMT6* (MIM: 608274) eQTL signal and observed that the caPeak overlapped the *PRMT6* TSS. *PRMT6* has been shown to regulate hepatic glucose metabolism in mice [105]. Our data suggest that a variant at this locus may increase chromatin accessibility and alter TF binding at the *PRMT6* TSS, leading to higher P*PRMT6* expression and decreased LDL cholesterol. At a GWAS locus for plasma liver enzyme levels [40, 129], we predicted *EFHD1* as a target gene based on both caQTL-eQTL colocalization and a promoter capture Hi-C link. While *EFHD1* is expressed in liver tissue, the GTEx portal shows that expression is much higher in other tissues [93], and the gene's roles in liver have not been characterized [76]. The caPeak at this locus does not overlap an ATAC peak in macrophages [4], but additional experiments, such as single nucleus ATAC-seq, are needed to determine the relevant cell type within liver tissue. Our data suggest

that *EFHD1* may be a target gene at this locus and act through one or more of the cell types in liver tissue. These and other results highlight the utility of caQTLs to identify mechanisms at GWAS loci.

At the *LITAF* locus, we provided direct evidence that variant rs11644920 can alter transcriptional regulation. Here, the caQTL, liver eQTL, and LDL cholesterol GWAS signals are colocalized, and the variant, mechanism, and cell type responsible for these associations were unknown. LITAF encodes a transcription factor that can mediate effects on inflammation [184], suggesting a potential role in hepatocytes and/or macrophages in an inflammatory environment. We showed that variants in the caPeak alter transcriptional reporter activity in hepatocytes, monocytes, macrophages, and lipopolysaccharide-stimulated macrophages. In all cell types, the caPeak showed a similar magnitude of enhancer activity and alleles showed differences in transcriptional activity, suggesting that the variant may act in any or all of these cell types. The caPeak at this locus does not overlap an ATAC peak in macrophages [4], but additional experiments, such as single nucleus ATAC-seq, are needed to determine the relevant cell type within liver tissue. We further provided evidence that rs11644920 alters protein binding, at least in vitro. Further study is needed to provide direct evidence that these variants alter transcription of *LITAF* and how altered levels of *LITAF* may affect cholesterol levels.

The maximum distance threshold between peaks and tested variants had a substantial impact on caQTL detection. Analyzing variants within a narrow region around a peak reduced the multiple testing burden for nearby variants, whereas testing variants in a broader region allowed identification of variants within one peak that may also influence another peak. A wide range of distance thresholds have been applied to caQTL discovery, including variants within 1 kb and 20 kb of peak centers [67], 50 kb from peak ends [4], and 1 Mb from peak ends [92]. We found many more significant results when using variants within 1 kb of peak centers compared to variants within 100 kb of peak centers, potentially due to reduced multiple testing burden and low power to detect long-range caQTL effects due to small sample size. Future caQTL studies with larger sample sizes will be more powered to detect longer-range caQTLs.

Due to the modest sample size of this study, we only tested for caQTLs using common variants (MAF ≥ 0.1) and did not predict regulatory variants at low-frequency GWAS signals. Based on three large GWASs for height [284], body mass index [284], and blood lipids [140] (see web resources), 77%–91% of signals had lead variant MAF ≥ 0.1, suggesting that we could test the majority of GWAS signals for caQTLs. However, allele frequencies in small sample sizes may differ from population allele frequencies, and larger caQTL studies will have more power to detect caQTLs at low frequency variants.

We used four approaches to suggest genes that may be regulated by caPeaks. However, several factors limit how many caPeak-gene connections can be identified and how many are shared by two or more approaches. TSS proximity is useful to detect variation in promoter accessibility, although our results showed that only 4% of caPeaks are TSS proximal, and caQTL-eQTL colocalization is the only method we tested that can corroborate TSS proximity. Promoter capture Hi-C data [125] identifies distal regions that physically interact with promoters, although additional Hi-C loops may be identified in additional samples and with higher sequencing depth. Hi-C loops < 15 kb were removed [125], indicating that the Hi-C data cannot corroborate caQTL-eQTL colocalizations or caPeak-promoter peak/gene expression correlations located < 15 kb from the promoter. The identification of caPeaks correlated with promoter peaks [238] or with gene expression is limited by sample size, and gene expression is affected by many other factors. The LD-based method we used to predict colocalized caQTL and eQTL signals helps identify peaks and genes with a shared genetic basis, although this method is influenced by low-resolution fine-mapping of the lead caQTL variant, use of an LD threshold, and choice of LD reference panel. Due to the modest sample size of this study, we were underpowered to detect colocalizations using coloc [95], and we recommend that future caQTL studies consider larger sample sizes for more robust colocalizations. Identification of conditional liver eQTLs, which tend to be further from gene TSSs compared to primary eQTLs [70, 219], could lead to additional caQTL-eQTL colocalizations. While each of these approaches was useful to predict links between caPeaks and genes, additional experiments are needed to identify causal relationships.

The caQTLs presented here are a resource for studying liver regulatory elements and will help identify mechanisms at GWAS loci for multiple traits that act through liver. The 56 caQTLs at GWAS loci with predicted target genes are strong candidates for future functional studies. While caQTLs can pinpoint functional regulatory variants, the modest sample size and analyses restricted to common variants limit fine-mapping potential and highlight the importance of considering LD proxies. The promising regulatory mechanisms identified here motivate identification of liver caQTLs in larger sample sizes.

A.6 Materials and Methods

A.6.1 Liver tissue samples

Healthy human liver tissue was collected from 20 deceased organ donors through the National Institutes of Health Liver Tissue Cell Distribution System (LTCDS). Tissue was obtained from LTCDS and approved for use in this study as non-human subjects research by the Institutional Review Boards (IRBs) at St Jude Children's Research Hospital (Memphis, TN) and the University of North Carolina (Chapel Hill, NC).

A.6.2 Genotyping and imputation

We genotyped more than 2.5 million variants using the Infinium Omni2.5Exome-8 BeadChip array v1.3 (Illumina) at the NHGRI Genomics Core facility. Overall genotyping call rates ranged from 99.0%–99.6%. We mapped the Illumina array probe sequences to the hg19 genome assembly17 using novoalign (see web resources), excluding variants with ambiguous probe alignments and variants with 1000 Genomes (1000G) phase 3 minor allele frequency (MAF) > .01 within 7 bp of the 3′ end of probes. No individuals were related at a 3rd-degree relationship threshold using KING v.1.4 [167]. Prior to performing genotype principal component analysis (PCA), we removed variants with minor allele count < 4 and that were found within regions of unusually high linkage disequilibrium (LD, see web resources) using VCFtools v.0.1 [62] and selected distinct (r^2 < 0.2)

variants using PLINK v.1.9 [213] We performed PCA of 59,674 genotypes using PLINK v.1.920 and found that each principal component (PC) explained essentially the same amount of variation (5%), and no PC explained a disproportionate amount of variation. Therefore, we did not include any genotype PCs as covariates when identifying caQTLs.

Prior to genotype imputation, we combined the genotypes of the samples in this study with genotypes from 177 samples from a separate study genotyped on similar chips and removed variants that met the following criteria: allele frequency difference > 20% with 1000G phase 3 Europeans, palindromic variants with MAF > .2, genotype missingness > 2.5%, and deviation from Hardy-Weinberg equilibrium ($p < 1 \times 10^{-4}$). Using the Michigan Imputation Server [63], we phased 1,789,889 autosomal variants using Eagle v.2.3 [161] and imputed missing genotypes using minimac3 [63] with the Haplotype Reference Consortium (hrc.r1.1.2016) panel [254]. We retained variants with imputation $r^2 > .3$ for downstream analyses.

A.6.3 RNA-seq library preparation, read alignment, and selection of expressed genes

We extracted and purified total RNA from 20 frozen liver tissue samples using Trizol as previously described [264]. Paired-end, strand-specific, poly(A) RNA sequencing (RNA-seq) was performed on an Illumina NovaSeq 6000 with 2× 151 bp cycles. RNA-seq reads were trimmed using Trimmomatic25 and aligned to the hg19 genome assembly using STAR v.2.53 with default parameters [71]. Using verifyBamID v.1.1.1 [124], we found no evidence of library contamination or sample swaps. Expression levels of GENCODE v.19 genes were quantified using QoRTs v.1.2.42 [109] We classified genes as expressed if the median transcripts per million (TPM) across the 20 individuals was at least 1. We performed principal component analysis on gene counts normalized by library size and variance-stabilized using DESeq2 [162]. Principal components (PCs) were correlated against technical factors to identify covariates in downstream analyses (see section A.6.17).

A.6.4 ATAC-seq library preparation

Nuclei were isolated as previously described [235] with the following modifications. We pulverized 50-mg pieces of frozen human liver tissue in liquid nitrogen using a Cell Crusher (Cell-Crusher), homogenized the tissue powder in ice-cold nuclei isolation buffer (NIB: 20 mM Tris-HCl, 50 mM EDTA, 5 mM spermidine, 0.15 mM spermine, 0.1% mercaptoethanol, 40% glycerol [pH 7.5]) using a 1-mL dounce for 40 strokes, and rotated for 5 min at 4℃. We filtered the solution through a Miracloth (Calbiochem), centrifuged at 1,100 × g for 10 min at 4℃, washed the pellet with 250 µL NIB containing 0.5% Triton-X, centrifuged at 500 × g for 5 min at 4℃, and resuspended the pellet in 250 µL of resuspension buffer (10 mM Tris-HCl, 10 mM NaCl, 3 mM MgCl2 [pH 7.4]). After counting isolated nuclei, we pelleted 50,000 nuclei at 500 × g for 5 min at 4℃ for each of three replicate ATAC-seq libraries per sample. Libraries were prepared using Nextera kits (Illumina) as previously described [28].

A.6.5 ATAC-seq read alignment and identification of consensus peaks

We trimmed ATAC-seq reads to a uniform length of 126 bp using cutadapt [170] and aligned reads as previously described [217]. Briefly, we trimmed sequencing adapters using CTA (see web resources) and aligned reads to the hg19 human genome17 using BWA-MEM (see web resources). We selected properly paired autosomal alignments with high mapping quality (mapq > 30) with samtools35 and removed duplicate alignments using Picard (see web resources). We used ataqv [197] to generate ATAC-seq quality metrics and confirmed ATAC-seq libraries corresponded to the correct genotypes using verifyBamID.

To assess reproducibility of libraries from the same individual, we called narrow peaks separately for each library using MACS2 [285] with parameters –nomodel –shift -100 –extsize 200, then merged peaks across all individuals and replicates using BEDTools merge [215], and selected peaks present in at least 3 libraries. We counted the number of reads overlapping each peak using featureCounts [159] and performed library size normalization and variance-stabilization using DESeq2 [162]. We computed pairwise Pearson correlations of normalized counts for all

peaks and for the 10,000 most variable peaks between libraries and visualized the results using the heatmap.2 function in the gplots R package [251] (see web resources). Libraries from the same individual were highly correlated, so we merged the alignment .bam files across libraries for each individual using SAMtools [157].

To identify consensus peaks, we converted the merged .bam files for each individual to .bed files using BEDTools, called narrow peaks for each individual using MACS2 with parameters –nomodel –shift -100 –extsize 200 –keep-dup all, and removed peaks overlapping blacklisted regions [131]. We then merged peaks across individuals using BEDTools and defined consensus peaks as merged peaks that shared at least 1 base with a peak present in samples from at least 3 individuals.

A.6.6 Overlap of consensus peaks with roadmap chromatin states

We computed overlap of ATAC-seq consensus peaks with chromatin states in adult liver tissue from the Roadmap Epigenomics Consortium [52]. We defined the following states: promoter (1_TssA, 2_TssFlnk, 3_TssFlnkU, 4_TssFlnkD, 14_TssBiv), transcribed (5_Tx, 6_TxWk), enhancer (7_EnhG1, 8_EnhG2, 9_EnhA1, 10_EnhA2, 11_EnhWk, 15_EnhBiv), polycomb (16_ReprPC, 17_ReprPCWk), heterochromatin (13_Het), ZNF repeats (12_ZNF/Rpts), and quiescent (18_Quies). For each consensus ATAC peak, we computed the fraction of bases that overlapped each chromatin state in liver tissue (Roadmap epigenome ID E066) using BEDTools coverage [214]. We assigned each peak to the chromatin state with which it shared the most bases, except for the quiescent state; we only assigned a peak to a quiescent state if all bases of a peak were found within a quiescent state. If a peak shared most, but not all, of its bases with a quiescent state, we assigned the peak to the state with the second highest coverage.

A.6.7 Selection of transcription factor motifs

We obtained transcription factor (TF) binding motifs from Cis-BP v.1.02 [272], selected all directly determined motifs per TF or the best inferred motif when a TF did not have a directly determined

motif (TF_Information.txt dataset from Cis-BP), and restricted to motifs for TFs expressed in liver tissue from GTEx v.8 (median transcripts per million ≥ 1). We performed clustering to remove redundant motifs using RSAT matrix-clustering [39] with parameters -hclust_method average -calc sum -metric_build_tree Ncor -lth w 5 -lth cor 0.8 -lth Ncor 0.8 -quick, resulting in 516 motif clusters. For each motif cluster, we defined the representative TF as the TF with the highest expression in liver tissue from GTEx v.8 (measured in median TPM) and the representative motif as the motif assigned to the representative TF. If multiple motifs existed for the representative TF in a given cluster, we selected the motif with the highest information content. Although we often use the representative TF name to refer to motif clusters for convenience, any TF in the cluster may bind at a given locus. Therefore, we listed all expressed TFs in the cluster in supplemental tables. Some TFs were assigned as the representative TF for multiple clusters, potentially representing distinct binding profiles for the same TF. We retained all of these clusters unless otherwise noted.

A.6.8 Enrichment of TF motifs and ChIP-seq binding sites in ATAC peaks

We tested for enrichment of 286 non-redundant transcription factor (TF) motifs in consensus ATAC peaks using Analysis of Motif Enrichment (AME) [177] with parameters –control –shuffle– –kmer 2 –scoring max –hit-lo-fraction 0.75. We classified motifs with E-value < 1×10^{-100} as significantly enriched. We derived the 286 motifs from the set of 516 non-redundant motifs (see "Selection of transcription factor motifs") by selecting the motif with the highest information content per TF.

We downloaded liver tissue ChIP-seq peaks for 17 TFs [218] from the ENCODE portal [123] (sample accession ENCDO882MMZ) and defined binding sites as the summit of the ChIP-seq peaks. We computed the number of binding sites overlapping consensus ATAC-seq peaks for each TF using BEDTools intersect. To determine whether the number of binding sites overlapping ATAC peaks was more than expected given their genomic frequency, we permuted binding

sites across the genome 1,000 times excluding blacklisted regions using BEDTools shuffle and computed the number of overlaps for each permutation. We calculated an enrichment p value by determining the fraction of permuted overlaps that were equal to or greater than the observed number of overlaps.

A.6.9 Enrichment of heritability in ATAC peaks

Using stratified LD score regression as implemented in LDSC v.1.0.1 [88], we tested whether liver ATAC peaks were enriched for heritability of 13 GWAS traits: liver enzymes traits alanine aminotransferase (ALT) [40], alkaline phosphatase (ALP) [40], and gamma-glutamyl transferase (GGT) [40]; cardiometabolic traits body mass index [212], high-density lipoprotein cholesterol (HDL) [277], low-density lipoprotein cholesterol (LDL) [277], triglycerides [277], total cholesterol [277], coronary artery disease [263], waist-hip ratio adjusted for body mass index (WHRadjBMI),49 and type 2 diabetes [166]; and two negative control traits likely less relevant to liver, height [284] and rheumatoid arthritis [194] (see web resources). We computed LD scores for liver ATAC peaks using LDSC with 1000G phase 3 European LD and restricting to HapMap3 SNPs. We computed partitioned heritability of the ATAC peaks using LDSC correcting for the baseline v.1.2 model, which consists of 53 annotations [88]. We report heritability enrichment as the proportion of heritability explained by SNPs within ATAC peaks divided by the proportion of SNPs within ATAC peaks and classify enrichments with enrichment p value (enrichment_p) <0.05 as significant.

A.6.10 Chromatin accessibility QTL identification

We identified caQTLs using RASQUAL,5 which jointly tests for association of genotype with peak accessibility across individuals and allelic imbalance in read counts at heterozygous variants within the same individual. We selected 4 million genetic variants with MAF > 0.1 in the 20 individuals and within 100 kb of consensus peak centers and then restricted to variants present in 1000G phase 3 Europeans. To quantify peak accessibility across samples, we extended alignments 100 bp from either end of the 5′-most base using BEDTools and counted the number of

alignments overlapping each peak using featureCounts. We did not use WASP54 to remove reads exhibiting allelic mapping bias because RASQUAL models and accounts for allelic mapping bias.5 We used DESeq2 size factors [162] to adjust for library size and the gcCor.R script provided with RASQUAL [145] to adjust for GC bias. To identify global variation between samples that may confound caQTL detection, we performed PCA on peak counts adjusted for library size and variance-stabilized by DESeq2. We ran RASQUAL using differing numbers of PCs as covariates ranging from 0 to 10 in increments of 1 and selected 2 PCs to maximize the number of peaks with a caQTL at false discovery rate (FDR) of 5%. We performed multiple testing correction using the two-step eigenMT-BH procedure [117]. First, we used eigenMT [64] with the 1000G phase 3 European reference panel to adjust for the differing variant density around each peak, taking into account the LD between variants. Second, we selected the most significant eigenMT-adjusted p value for each peak and calculated FDR using the Benjamini-Hochberg (BH) procedure. We selected significant caQTLs with FDR < 5% and correlation r^2 between prior and posterior genotypes > 0.8. We refer to peaks with a significant caQTL as caPeaks. We repeated the caQTL analysis using 0.6 million variants within 1 kb of peak centers. Unless otherwise noted, all downstream analyses were performed using caQTLs identified using variants within 1 kb of peak centers.

A.6.11 Identification of caQTLs strongly influenced by one sample

To identify caQTLs strongly influenced by one sample, we separately removed each sample from the analysis and re-identified caQTLs in the 20 sets of 19 samples. We used the same caQTL parameters as for all 20 samples, except that we reduced the minimum MAF threshold to 0.05 to retain variants with MAF of 0.1 in the 20 samples. We restricted analyses to the lead variant-peak pairs detected in the 20-sample analysis. Given our small sample size, we would expect some caQTLs to no longer be significant when one sample is removed due to power even if no influential samples are present. Therefore, we defined caQTLs that are strongly influenced by one specific sample as caQTLs that no longer meet the FDR < 5% threshold (eigenMT-adjusted p < 8.4×10^{-4}) only when one specific sample is removed, but remain significant when any other

sample is removed.

A.6.12 ATAC-seq allelic imbalance and comparison to caQTL effect sizes

Instead of removing reads that exhibit allelic mapping bias, RASQUAL estimates and accounts for allelic mapping bias during QTL mapping.5 To compare the RASQUAL results to another strategy, we used an alternative method to remove reads exhibiting allelic mapping bias and calculate allelic imbalance (AI). We removed ATAC-seq reads exhibiting allelic mapping bias using the WASP mapping pipeline [262] and counted the number of ATAC-seq reads mapping to each allele at heterozygous variants using ASEReadCounter [38] with the option –min-base-quality 30. We removed variants that had aligned bases other than the two genotyped alleles and included variants with >10 total reads, >3 reads per allele, and that were heterozygous in >3 individuals. After pooling reads across individuals, each variant had a minimum of 30 total reads and 9 reads per allele. The average reference allele fraction across all heterozygous sites for each sample ranged from 0.502 to 0.505, and the average reference allele fraction after combining samples was 0.503, indicating that little to no systematic allelic mapping bias remains. We fit allele counts to a beta-binomial distribution using the VGAM R package [283], tested for AI using a two-tailed beta-binomial test, and adjusted for multiple testing using the BH procedure.

To compare effect sizes of AI variants and caQTL signals, we selected caQTLs that had at least one AI variant in strong LD (r^2 > 0.8, 1000G phase 3 Europeans) with the caQTL lead variant and that resided within the caPeak; LD was calculated using PLINK v.1.9. For each caQTL with a linked AI variant, we selected the AI variant with the strongest evidence of imbalance (smallest beta-binomial p value). For both methods, we calculated an effect size by subtracting 0.5 from the estimated fraction of reads containing the alternate allele, which is the RASQUAL PI value for caQTLs. An alternate allele fraction of 0.5 corresponds to an equal number of reads for each allele, which is an effect size of 0. We then computed the Pearson correlation between the absolute value of effect sizes between the caQTLs and AI variants.

A.6.13 Colocalization of caQTL and H3K27ac QTL signals

We retrieved QTLs for 921 histone 3 lysine 27 acetylation (H3K27ac) peaks (termed H3K27ac QTLs, FDR < 5%, n = 18) from a recent report [35]. We only tested for colocalization between QTL signals where the caPeak and H3K27ac peak overlapped (defined as sharing at least one base). We calculated LD and haplotype phase between H3K27ac QTL and caQTL lead variants using PLINK v.1.9 and classified signals as colocalized if these lead variants exhibited strong pairwise LD (r^2 > 0.8, 1000G phase 3 Europeans). We calculated effect sizes for caQTLs and H3K27ac QTLs by subtracting 0.5 from the RASQUAL PI values. We then computed the Pearson correlation between the absolute value of caQTL and H3K27ac QTL effect sizes.

A.6.14 caQTL enrichment in chromatin states

To identify which regulatory elements preferentially contain caPeaks, we compared the number of caPeaks (FDR < 5%) and non-caPeaks (eigenMT-adjusted p > 0.5) assigned to various liver tissue chromatin states from Roadmap [52]. We tested whether caQTL variants were enriched in specific liver tissue chromatin states relative to variants matched for MAF, number of LD proxies, and distance to nearest gene using the logistic regression model implemented in GARFIELD [119]. We defined caQTL variants as significantly enriched in a chromatin state if the p value for the logistic regression beta was less than the Bonferroni-corrected threshold (alpha of 0.05 for 7 chromatin states) of 7.1 × 10−3 and the odds ratio was greater than 1. We defined caQTL variants as significantly depleted in a chromatin state if $p < 7.1 \times 10^{-3}$ and odds ratio < 1.

A.6.15 Overlap of caPeaks with macrophage ATAC peaks

We retrieved a set of 296,220 ATAC peaks mapped across macrophages exposed to four experimental conditions: naive, IFNγ stimulation, *Salmonella* infection, and both exposures [4] (see web resources). To compare peak positions, we used liftOver61 with the option -minMatch = 0.75 to convert the 3,123 liver caPeaks from GRCh37 (hg19) to GRCh38 coordinates. We identi-

fied liver caPeaks that overlapped (defined as sharing at least 1 base) with a macrophage peak using BEDTools intersect. We also applied liftOver to the macrophage peaks and obtained the same results.

A.6.16 Transcription factor motif disruption by caQTL variants

We selected 5,378 caQTL variants that resided within a caPeak using BEDTools intersect and that were in strong LD ($r^2 > 0.8$, calculated with PLINK) with the caQTL lead variant. To ensure that each motif occurrence was disrupted by only one variant, we removed 793 variants located within 30 bp of another caQTL variant, resulting in 4,585 variants. For both alleles of each caQTL variant, we extracted the nucleotide sequence for the region containing the variant and the 30 nucleotides on either side of the variant using the BEDTools slop and getfasta tools [215]. We scanned these sequences for occurrences of 516 non-redundant TF motifs using Find Individual Motif Occurrences (FIMO) [98] with parameters –thresh 0.01–max-stored-scores 1000000–no-qvalue–skip-matched-sequence –text and only retained motif occurrences that overlapped caQTL variant positions. For each motif-variant pair, we selected the strongest motif match (smallest p value) per allele and only retained motif occurrences that matched strongly to at least one allele ($p < 1 \times 10^{-4}$). If different motifs for the same representative TF overlapped the same variant, we selected the motif with the strongest match.

Similar to a recent study [180], we quantified the difference in motif match between alleles of a variant using the log ratio of FIMO p values. The FIMO p value for a given motif occurrence is the probability of observing a motif occurrence with the same or greater score, which inherently accounts for differences in score distributions between different motifs. For a given variant-motif pair, we define motif disruption as $\log10(p_{aw}) - \log10(p_{as})$, where p_{aw} and p_{as} are the FIMO p values for the alleles with the weaker and stronger motif match, respectively. As motif disruption is always positive, we classified a motif as disrupted if motif disruption was > 1, corresponding to a 10-fold difference in the FIMO p values between alleles.

We identified motifs whose disruption was associated with caQTL status using logistic regres-

sion. To generate a set of non-caQTL variants, we first selected peaks with no evidence of genetic regulation (caQTL eigenMT-adjusted p > 0.5), that overlapped at least one variant tested in the caQTL analysis and that were similar to caPeaks in GC content (±5%), peak width (±20%), and distance to nearest transcription start site (TSS) of a protein-coding gene in GENCODE (±20%). We identified 10 non-caPeaks for >99% of the caPeaks used in the motif disruption analysis and defined non-caQTL variants as the 50,054 variants that were within non-caPeaks and were located more than 30 bp from the nearest variant. We tested these non-caQTL variants for TF motif disruption using the same procedure as for caQTL variants and restricted analysis to the 109 motifs with at least 20 disruptions by caQTL variants. For each representative TF, we selected the motif with the most disruptions by caQTL variants to ensure that we used only one motif per representative TF. We then regressed caQTL status (1 = caQTL, 0 = non-caQTL) against motif disruption status (1 = disrupted, 0 = not disrupted) for each motif-variant pair using logistic regression. We classified motif disruption as associated with caQTL status if the p value for the logistic regression beta was less than the Bonferroni-corrected threshold (alpha of 0.05 for 109 motifs) of $4.6 \times 10{-}4$. Because residual differences may exist in peak GC content, width, and distance to nearest protein coding TSS, we performed logistic regression with and without these features as covariates and obtained the same set of significantly enriched motifs.

A.6.17 caPeak target gene identification

We used four methods to identify target genes for caPeaks: proximity to a gene's TSS, overlap of caPeaks with promoter-centered chromatin contacts, correlation of caPeaks with peaks at gene promoters or with gene expression, and colocalization of caQTLs and eQTLs. We excluded genes from the analysis if their Entrez ID did not map to exactly one Ensembl ID (eQTL data only) or if their symbol (common name) didn't map to exactly one Ensembl ID. When combining results across the four methods, we matched genes based on Ensembl ID.

TSS proximity

We classified a caPeak as TSS proximal if it was located within 2 kb upstream and 1 kb downstream of the TSS of any of the 13,782 expressed genes (median TPM > 1) in our 20 liver samples using BEDTools closest.

Promoter-centered chromatin contacts

We obtained promoter-distal and promoter-promoter contacts mapped in liver tissue using promoter capture Hi-C from a recent study [125] (see web resources). Using described filtering criteria [125], we selected contacts with p value < 0.01 and interaction frequency ≥ 5. We identified caPeaks overlapping distal ends of promoter-distal contacts or either end of promoter-promoter contacts using BEDTools intersect.

Correlation of caPeaks with promoter peaks and gene expression

We classified an ATAC-seq peak as the promoter peak for an expressed gene if it was the closest peak to the TSS of the gene and it was within 2 kb upstream and 1 kb downstream of the TSS [65]. A promoter peak may or may not be a caPeak. We identified promoter peaks for 10,074 of 13,782 expressed genes. For each gene with a promoter peak, we identified caPeaks for correlation that were within 1 Mb of the gene's TSS but that were not TSS proximal. For peak and gene counts, we performed library size normalization and variance-stabilization using DESeq2 and GC bias-correction using RASQUAL [145]. We additionally adjusted peak counts by the percent of high-quality autosomal alignments (HQAA) in peaks (a measure of ATAC signal-to-noise), which was strongly correlated with the first ATAC-seq PC, and gene counts by the percent of reads mapping to the most expressed gene and the percent of reads mapping to the top 10 most expressed genes (geneDiversityProfile_top1pct and geneDiversityProfile_top10pct metrics from QoRTs), which were strongly correlated with RNA-seq PCs 1 and 2, respectively, using the limma removeBatchEffects [223]. We then computed the Spearman correlation between (1) gene expression and caPeaks and (2) promoter peaks and caPeaks using the cor.test function in R. We

adjusted for multiple testing using the BH procedure and classified correlations with FDR < 5% as significant.

Colocalization of caQTLs and eQTLs

We obtained liver tissue expression quantitative trait loci (eQTLs) for 15,668 genes (FDR < 5%) from a meta-analysis of 1,183 individuals [82] and restricted to the 15,418 eQTLs on autosomes. We calculated LD and haplotype phase between eQTL and caQTL lead variants using PLINK v.1.9 and classified signals as colocalized if these lead variants exhibited strong pairwise LD (r^2 > 0.8, 1000G phase 3 Europeans). To compare the direction of effect for colocalized caQTLs and eQTLs, we compared the sign of the caQTL effect size (RASQUAL pi statistic - 0.5) and the eQTL effect size (meta T statistic).

For the caQTL-eQTL colocalizations identified based on LD, we also assessed colocalization using the Bayesian approach implemented in coloc [95]. We ran coloc using p values and minor allele frequencies because regression coefficients and variances are not available from the RASQUAL model. coloc estimates five posterior probabilities (PP): no variant in the tested region affects either trait (PP0), a variant affects one trait but not the other (PP1 for caQTL and PP2 for eQTL), different variants affect each trait (PP3, no colocalization), and the same variant affects both traits (PP4, colocalization). We considered signals to show strong evidence of colocalization if PP4 > 0.8, suggestive evidence of colocalization if PP4 > 0.5, and evidence against colocalization if PP3 > 0.5. If the sum of PP0, PP1, and PP2 was > 0.5, we concluded that power was too low to assess colocalization. We note that coloc was designed to operate on results from linear regression or logistic regression [95] and may not be appropriate for the caQTL results generated from RASQUAL, which combines results from a negative binomial generalized linear model and tests of allelic imbalance [145].

A.6.18 Colocalization of caQTL and GWAS signals

We downloaded the NHGRI-EBI GWAS catalog [30] on October 28, 2019, extracted only single variant associations, and converted variant genomic coordinates from GRCh38 to GRCh37 (hg19) using liftOver. We extracted variants associated with 19 trait groups ($p < 5 \times 10^{-8}$) relevant to liver function and cardiometabolic diseases: liver enzymes, high-density lipoprotein cholesterol (HDL), low-density lipoprotein cholesterol (LDL), total cholesterol (TC), triglycerides (TG), cardiovascular disease (CVD), hypertension/blood pressure (HTBP), type 2 diabetes (T2D), insulin, glucose, glycated albumin, serum albumin, glycated hemoglobin (HbA1c), C-reactive protein (CRP), bilirubin, body mass index (BMI), waist-hip ratio adjusted for BMI (WHRadjBMI), liver injury, and non-alcoholic fatty liver disease (NAFLD). We also included two negative control traits, height and rheumatoid arthritis, which presumably have less relevance to the liver. We extracted alleles for each variant from the dbSNP [240] build 151 common variant set (see web resources), restricting to bi-allelic variants. To select one variant per association signal, we performed LD clumping separately for each trait using swiss (see web resources); variants in strong LD ($r^2 > 0.8$, 1000G phase 3 Europeans) and within 1 Mb of a variant with a more significant p value were removed. We calculated LD between lead caQTL and GWAS variants using PLINK v.1.9 and classified signals in high LD ($r^2 > 0.8$) as colocalized. We made LocusZoom plots for specific loci using LocusZoom v.1.4 [211].

To identify liver caQTL-GWAS colocalizations also observed in blood, we retrieved caQTLs mapped in macrophages exposed to four experimental conditions [4] and activated T cells [92] (see web resources). For macrophages, we downloaded the caQTL lead variant summary statistics and selected significant caQTLs at FDR < 10% using the same procedure described in the previous report [4], and we converted the genomic coordinates from GRCh38 to hg19 using liftOver [113]. T cells, we used the set of publicly available caQTLs at FDR < 5% mapped to hg19 coordinates [92]. For both datasets, we identified caQTL signals colocalized with GWAS signals using the procedure described above. We considered a liver caQTL-GWAS colocalization to be present in a blood cell type if the liver and blood caPeaks shared at least one base and if the lead variant

of the blood caQTL was in strong LD ($r^2 > 0.8$) with the same GWAS variant as the liver caQTL. Blood caQTL lead variants were not tested for colocalization if variants were not in the 1000G LD reference panel.

A.6.19 Transcriptional activity reporter assays

HepG2 hepatocyte cells were cultured in MEM-alpha supplemented with 10% FBS and 1 mM sodium pyruvate, THP-1 monocyte cells were cultured in RPMI-1640 supplemented with 10% FBS, and both cell types were maintained at 37℃ with 5% CO2. To test haplotypic differences in transcriptional activity, we designed PCR primers (5′-TATGTTGCACAGGCTGGTCT-3′ and 5′-GGCAATAACGCCCACCTC-3′) to amplify a 666-bp DNA element (chr16:11,644,551–11,645,216) spanning the ATAC-seq peak and containing variants rs3784924, rs11644920, and rs57792815, and we generated PCR products using DNA from individuals homozygous for both haplotypes. We cloned the derived PCR products into luciferase reporter vector pGL4.23 (Promega) as described previously [89]. The day before transfection, we plated 120,000 HepG2 cells, and on the day of transfection, we plated 300,000 THP-1 cells. We transfected duplicate wells with four to five sequence-verified independent constructs for each haplotype. We co-transfected wells with phRL-TK Renilla reporter vector using lipofectamine 3000 (Life Technologies) following the manufacturer's protocol. To induce differentiation into macrophages [116], we added 100 nM 1α,25-Dihydroxyvitamin D3 (Sigma) to the THP-1 cells at the time of transfection. To obtain activated macrophages, we added 100 ng/mL lipopolysaccharides (Sigma) to vitamin D3-treated cells 24 h after transfection and incubated cells for an additional 24 h. Firefly luciferase activity was measured 48 h post-transfection and normalized to Renilla activity to adjust for differences in transfection efficiency. Fold-changes in luciferase activity were calculated relative to an empty pGL4.23 vector, and statistical differences in activity were determined using two-tailed Student's t tests. We repeated transcriptional activity experiments on a separate day and obtained equivalent results.

A.6.20 Electrophoretic mobility shift assays (EMSAs)

We designed and annealed 3 biotin-labeled and unlabeled 17-bp complementary oligonucleotide probes centered on each of variants rs3784924, rs11644920, and rs57792815. We conducted EMSAs using the LightShift Chemiluminescent EMSA kit (Thermo Scientific) following the manufacturer's protocol. The binding reactions consisted of 6 μg HepG2 nuclear extract (NE-PER Kit, Thermo Fisher Scientific), 1 μg poly(dI-dC), 1x binding buffer, and 400 fmol biotinylated oligonucleotide as described previously [89]. To test the specificity of the protein complexes to each allele, we added 10-fold excess unlabeled probes. Protein-DNA complexes were resolved by gel electrophoresis and transferred and detected by chemiluminescence as described previously [89]. We repeated EMSA experiments on a separate day and obtained equivalent results.

Bibliography

[1] Martin Abadi et al. "TensorFlow: A System for Large-Scale Machine Learning". In: *12th USENIX Symposium on Operating Systems Design and Implementation (OSDI 16)*. 2016, pp. 265–283. URL: https://www.usenix.org/system/files/conference/osdi16/osdi16-abadi.pdf (visited on 08/30/2019) (page 42).

[2] Amanda M. Ackermann et al. "Integration of ATAC-seq and RNA-seq Identifies Human Alpha Cell and Beta Cell Signature Genes". In: *Molecular Metabolism* 5.3 (Mar. 1, 2016), pp. 233–244. ISSN: 2212-8778. DOI: 10.1016/j.molmet.2016.01.002 (pages 16, 24, 26).

[3] E. Ahlqvist, R.B. Prasad, and L. Groop. "Subtypes of Type 2 Diabetes Determined From Clinical Parameters". In: *Diabetes* 69 (2020), pp. 2086–2093. DOI: 10.2337/dbi20-0001 (pages 48, 49).

[4] Kaur Alasoo et al. "Shared Genetic Effects on Chromatin and Gene Expression Indicate a Role for Enhancer Priming in Immune Response". In: *Nature Genetics* (Jan. 29, 2018), p. 1. ISSN: 1546-1718. DOI: 10.1038/s41588-018-0046-7 (pages 122, 129, 139, 143, 144, 154, 159).

[5] N.J.W. Albrechtsen. "The Liver-α-Cell Axis and Type 2 Diabetes". In: *Endocrine reviews* 40 (2019), pp. 1353–1366. DOI: 10.1210/er.2018-00251 (page 52).

[6] K. Amo-Shiinoki. "Islet Cell Dedifferentiation Is a Pathologic Mechanism of Long-Standing Progression of Type 2 Diabetes". In: *JCI Insight* 6 (2021), p. 143791. DOI: 10.1172/jci.insight.143791 (page 59).

[7] Stylianos E. Antonarakis et al. "Mendelian Disorders and Multifactorial Traits: The Big Divide or One for All?" In: *Nature Reviews Genetics* 11.5 (May 2010), pp. 380–384. ISSN: 1471-0056, 1471-0064. DOI: 10.1038/nrg2793 (page 116).

[8] H. Efsun Arda et al. "A Chromatin Basis for Cell Lineage and Disease Risk in the Human Pancreas". In: *Cell Systems* 7.3 (Sept. 26, 2018), 310–322.e4. ISSN: 2405-4712. DOI: 10.1016/j.cels.2018.07.007. pmid: 30145115 (pages 16, 24, 26).

[9] A.D. Association. "Pharmacologic Approaches to Glycemic Treatment: Standards of Medical Care in Diabetes—2020". In: *Diabetes Care* 43 (Suppl. 1 2020), pp. 98–110 (pages 49, 86).

[10] A. Auton. "A Global Reference for Human Genetic Variation". In: *Nature* 526 (2015), pp. 68–74. DOI: 10.1038/nature15393 (page 98).

[11] T.L. Bailey. "DREME: Motif Discovery in Transcription Factor ChIP-seq Data". In: *Bioinformatics (Oxford, England)* 27 (2011), pp. 1653–1659. DOI: 10.1093/bioinformatics/btr261 (page 104).

[12] T.L. Bailey and C. Elkan. "Fitting a Mixture Model by Expectation Maximization to Discover Motifs in Biopolymers". In: *Proceedings. International Conference on Intelligent Systems for Molecular Biology* 2 (1994), pp. 28–36 (page 104).

[13] T.L. Bailey et al. "The MEME Suite". In: *Nucleic acids research* 43 (2015), pp. 39–49. DOI: 10.1093/nar/gkv416 (page 104).

[14] Timothy L. Bailey et al. "MEME Suite: Tools for Motif Discovery and Searching". In: *Nucleic Acids Research* 37 (suppl_2 July 1, 2009), W202–W208. ISSN: 0305-1048. DOI: 10.1093/nar/gkp335 (page 95).

[15] A.N. Balamurugan et al. "Flexible Management of Enzymatic Digestion Improves Human Islet Isolation Outcome from Sub-Optimal Donor Pancreata". In: *American journal of transplantation : official journal of the American Society of Transplantation and the American Society of Transplant Surgeons* 3 (2003), pp. 1135–1142. DOI: 10.1046/j.1600-6143.2003.00184.x (pages 84, 85).

[16] Vincent D. Blondel et al. "Fast Unfolding of Communities in Large Networks". In: *Journal of Statistical Mechanics: Theory and Experiment* 2008.10 (Oct. 2008), P10008. ISSN: 1742-5468. DOI: 10.1088/1742-5468/2008/10/P10008 (page 91).

[17] Anthony M. Bolger, Marc Lohse, and Bjoern Usadel. "Trimmomatic: A Flexible Trimmer for Illumina Sequence Data". In: *Bioinformatics* 30.15 (Aug. 1, 2014), pp. 2114–2120. ISSN: 1367-4803. DOI: 10.1093/bioinformatics/btu170 (page 39).

[18] S. Bonner-Weir and T.D. O'Brien. "Islets in Type 2 Diabetes: In Honor of Dr". In: *Robert C. Turner. Diabetes* 57 (2008), pp. 2899–2904 (page 81).

[19] Lori L Bonnycastle et al. "Single-Cell Transcriptomics from Human Pancreatic Islets: Sample Preparation Matters". In: *Biology Methods and Protocols* 4.1 (Jan. 1, 2019), bpz019. ISSN: 2396-8923. DOI: 10.1093/biomethods/bpz019 (page 110).

[20] K. Breuer. "InnateDB: Systems Biology of Innate Immunity and beyond—Recent Updates and Continuing Curation". In: *Nucleic acids research* 41 (2013), pp. 1228–1233. DOI: 10.1093/nar/gks1147 (page 94).

[21] M. Brissova. "Islet Microenvironment, Modulated by Vascular Endothelial Growth Factor-A Signaling, Promotes β Cell Regeneration". In: *Cell metabolism* 19 (2014), pp. 498–511. DOI: 10.1016/j.cmet.2014.02.001 (pages 86, 87).

[22] M. Brissova. "The Integrated Islet Distribution Program Answers the Call for Improved Human Islet Phenotyping and Reporting of Human Islet Characteristics in Research Articles". In: *Diabetologia* 62 (2019), pp. 1312–1314. DOI: 10.1007/s00125-019-4876-3 (page 52).

[23] Marcela Brissova et al. "α Cell Function and Gene Expression Are Compromised in Type 1 Diabetes". In: *Cell Reports* 22.10 (Mar. 6, 2018), pp. 2667–2676. ISSN: 2211-1247. DOI: 10.1016/j.celrep.2018.02.032. pmid: 29514095 (pages 84–89).

[24] Marcela Brissova et al. “Assessment of Human Pancreatic Islet Architecture and Composition by Laser Scanning Confocal Microscopy”. In: *The Journal of Histochemistry and Cytochemistry: Official Journal of the Histochemistry Society* 53.9 (Sept. 2005), pp. 1087–1097. ISSN: 0022-1554. DOI: 10.1369/jhc.5C6684.2005. pmid: 15923354 (pages 3, 21, 50).

[25] J. Bryois et al. “Evaluation of Chromatin Accessibility in Prefrontal Cortex of Individuals with Schizophrenia”. In: *Nature communications* 9 (2018), p. 3121. DOI: 10.1038/s41467-018-05379-y (page 122).

[26] Jason D. Buenrostro et al. “Integrated Single-Cell Analysis Maps the Continuous Regulatory Landscape of Human Hematopoietic Differentiation”. In: *Cell* 0.0 (Apr. 26, 2018). ISSN: 0092-8674, 1097-4172. DOI: 10.1016/j.cell.2018.03.074. pmid: 29706549, 29706549 (page 10).

[27] Jason D. Buenrostro et al. “Single-Cell Chromatin Accessibility Reveals Principles of Regulatory Variation”. In: *Nature* 523.7561 (July 2015), pp. 486–490. ISSN: 1476-4687. DOI: 10.1038/nature14590 (page 10).

[28] Jason D. Buenrostro et al. “Transposition of Native Chromatin for Fast and Sensitive Epigenomic Profiling of Open Chromatin, DNA-binding Proteins and Nucleosome Position”. In: *Nature Methods* 10.12 (Dec. 2013), pp. 1213–1218. ISSN: 1548-7091. DOI: 10.1038/nmeth.2688 (pages 16, 36, 123, 148).

[29] Michael Bulger and Mark Groudine. “Functional and Mechanistic Diversity of Distal Transcription Enhancers”. In: *Cell* 144.3 (Feb. 4, 2011), pp. 327–339. ISSN: 0092-8674. DOI: 10.1016/j.cell.2011.01.024 (page 19).

[30] Annalisa Buniello et al. “The NHGRI-EBI GWAS Catalog of Published Genome-Wide Association Studies, Targeted Arrays and Summary Statistics 2019”. In: *Nucleic Acids Research* 47.D1 (Jan. 8, 2019), pp. D1005–D1012. ISSN: 0305-1048. DOI: 10.1093/nar/gky1120 (pages 135, 159).

[31] Martijn van de Bunt et al. “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”. In: *PLOS Genetics* 11.12 (Dec. 1, 2015), e1005694. ISSN: 1553-7404. DOI: 10.1371/journal.pgen.1005694 (pages 15, 33).

[32] A. Butler et al. “Integrating Single-Cell Transcriptomic Data across Different Conditions, Technologies, and Species”. In: *Nature biotechnology* 36 (2018), p. 411. DOI: 10.1038/nbt.4096 (page 101).

[33] A.E. Butler. “β-Cell Deficit and Increased β-Cell Apoptosis in Humans with Type 2 Diabetes”. In: *Diabetes* 52 (2003), pp. 102–110 (page 81).

[34] Over Cabrera et al. “The Unique Cytoarchitecture of Human Pancreatic Islets Has Implications for Islet Cell Function”. In: *Proceedings of the National Academy of Sciences of the United States of America* 103.7 (Feb. 14, 2006), pp. 2334–2339. ISSN: 0027-8424. DOI: 10.1073/pnas.0510790103. pmid: 16461897 (pages 5, 21, 34).

[35] Minal Çalışkan et al. “Genetic and Epigenetic Fine Mapping of Complex Trait Associated Loci in the Human Liver”. In: *The American Journal of Human Genetics* 105.1 (July 3, 2019), pp. 89–107. ISSN: 0002-9297. DOI: 10.1016/j.ajhg.2019.05.010 (pages 123, 127, 154).

[36] Joan Camunas-Soler et al. “Patch-Seq Links Single-Cell Transcriptomes to Human Islet Dysfunction in Diabetes”. In: *Cell Metabolism* 31.5 (May 5, 2020), 1017–1031.e4. ISSN: 1550-4131. DOI: 10.1016/j.cmet.2020.04.005 (pages 50, 110).

[37] Junyue Cao et al. “Joint Profiling of Chromatin Accessibility and Gene Expression in Thousands of Single Cells”. In: *Science* 361.6409 (Sept. 28, 2018), pp. 1380–1385. ISSN: 0036-8075, 1095-9203. DOI: 10.1126/science.aau0730 (page 11).

[38] Stephane E. Castel et al. “Tools and Best Practices for Data Processing in Allelic Expression Analysis”. In: *Genome Biology* 16.1 (Sept. 17, 2015), p. 195. ISSN: 1474-760X. DOI: 10.1186/s13059-015-0762-6 (page 153).

[39] Jaime Abraham Castro-Mondragon et al. “RSAT Matrix-Clustering: Dynamic Exploration and Redundancy Reduction of Transcription Factor Binding Motif Collections”. In: *Nucleic Acids Research* 45.13 (July 27, 2017), e119. ISSN: 0305-1048. DOI: 10.1093/nar/gkx314 (page 150).

[40] John C. Chambers et al. “Genome-Wide Association Study Identifies Loci Influencing Concentrations of Liver Enzymes in Plasma”. In: *Nature Genetics* 43.11 (11 Nov. 2011), pp. 1131–1138. ISSN: 1546-1718. DOI: 10.1038/ng.970 (pages 137, 143, 151).

[41] V. Chandra. “RFX6 Regulates Insulin Secretion by Modulating Ca2+ Homeostasis in Human β Cells”. In: *Cell Reports* 9 (2014), pp. 2206–2218. DOI: 10.1016/j.celrep.2014.11.010 (page 83).

[42] Ji Chen et al. “The Trans-Ancestral Genomic Architecture of Glycemic Traits”. In: *Nature Genetics* 53.6 (6 June 2021), pp. 840–860. ISSN: 1546-1718. DOI: 10.1038/s41588-021-00852-9 (page 95).

[43] Joshua Chiou et al. “Large-Scale Genetic Association and Single Cell Accessible Chromatin Mapping Defines Cell Type-Specific Mechanisms of Type 1 Diabetes Risk”. In: *bioRxiv* (Jan. 15, 2021), p. 2021.01.13.426472. DOI: 10.1101/2021.01.13.426472 (pages 49, 110, 115).

[44] Joshua Chiou et al. “Single Cell Chromatin Accessibility Reveals Pancreatic Islet Cell Type- and State-Specific Regulatory Programs of Diabetes Risk”. In: *bioRxiv* (July 9, 2019), p. 693671. DOI: 10.1101/693671 (pages 35, 114).

[45] Joshua Chiou et al. “Single-Cell Chromatin Accessibility Identifies Pancreatic Islet Cell Type– and State-Specific Regulatory Programs of Diabetes Risk”. In: *Nature Genetics* (Apr. 1, 2021), pp. 1–12. ISSN: 1546-1718. DOI: 10.1038/s41588-021-00823-0 (page 48).

[46] S.P. Choksi et al. “Switching on Cilia: Transcriptional Networks Regulating Ciliogenesis”. In: *Development (Cambridge, England)* 141 (2014), pp. 1427–1441. DOI: 10.1242/dev.074666 (page 84).

[47] F. Cinti. "Evidence of β-Cell Dedifferentiation in Human Type 2 Diabetes". In: *J Clin Endocrinol Metabolism* 101 (2016), pp. 1044–1054. DOI: 10.1210/jc.2015-2860 (page 59).

[48] Mete Civelek et al. "Genetic Regulation of Adipose Gene Expression and Cardio-Metabolic Traits". In: *The American Journal of Human Genetics* 100.3 (Mar. 2017), pp. 428–443. ISSN: 00029297. DOI: 10.1016/j.ajhg.2017.01.027 (page 6).

[49] Stephen J. Clark et al. "Single-Cell Epigenomics: Powerful New Methods for Understanding Gene Regulation and Cell Identity". In: *Genome Biology* 17.1 (Apr. 18, 2016), p. 72. ISSN: 1474-760X. DOI: 10.1186/s13059-016-0944-x (page 10).

[50] Christian M. Cohrs et al. "Dysfunction of Persisting β Cells Is a Key Feature of Early Type 2 Diabetes Pathogenesis". In: *Cell Reports* 31.1 (Apr. 7, 2020), p. 107469. ISSN: 2211-1247. DOI: 10.1016/j.celrep.2020.03.033 (pages 5, 81).

[51] ENCODE Project Consortium et al. "An Integrated Encyclopedia of DNA Elements in the Human Genome". In: *Nature* 489.7414 (2012), p. 57 (pages 123, 124).

[52] Roadmap Epigenomics Consortium et al. "Integrative Analysis of 111 Reference Human Epigenomes". In: *Nature* 518.7539 (Feb. 2015), pp. 317–330. ISSN: 1476-4687. DOI: 10.1038/nature14248 (pages 122, 124, 129, 149, 154).

[53] Kevin W. Currin et al. "Genetic Effects on Liver Chromatin Accessibility Identify Disease Regulatory Variants". In: *The American Journal of Human Genetics* (May 2021), S0002929721001853. ISSN: 00029297. DOI: 10.1016/j.ajhg.2021.05.001 (page 121).

[54] Darren A. Cusanovich et al. "A Single-Cell Atlas of In Vivo Mammalian Chromatin Accessibility". In: *Cell* (Aug. 2, 2018). ISSN: 0092-8674. DOI: 10.1016/j.cell.2018.06.052 (pages 10, 16, 40).

[55] Darren A. Cusanovich et al. "Multiplex Single-Cell Profiling of Chromatin Accessibility by Combinatorial Cellular Indexing". In: *Science* 348.6237 (May 22, 2015), pp. 910–914. ISSN: 0036-8075, 1095-9203. DOI: 10.1126/science.aab1601. pmid: 25953818 (pages 9, 16, 17, 37).

[56] Darren A. Cusanovich et al. "The *Cis*-Regulatory Dynamics of Embryonic Development at Single-Cell Resolution". In: *Nature* 555.7697 (Mar. 2018), pp. 538–542. ISSN: 1476-4687. DOI: 10.1038/nature25981 (pages 10, 16).

[57] C. Dai. "Age-Dependent Human β Cell Proliferation Induced by Glucagon-like Peptide 1 and Calcineurin Signaling". In: *The Journal of clinical investigation* 127 (2017), pp. 3835–3844. DOI: 10.1172/JCI91761 (pages 84, 96).

[58] C. Dai. "Stress-Impaired Transcription Factor Expression and Insulin Secretion in Transplanted Human Islets". In: *The Journal of clinical investigation* 126 (2016), pp. 1857–1870. DOI: 10.1172/JCI83657 (pages 86–88).

[59] C. Dai. "Tacrolimus- and Sirolimus-Induced Human β Cell Dysfunction Is Reversible and Preventable". In: *JCI Insight* 5 (2020), p. 130770. DOI: 10.1172/jci.insight.130770 (pages 87, 88).

[60] Chao Dai et al. “Mining 3D Genome Structure Populations Identifies Major Factors Governing the Stability of Regulatory Communities”. In: *Nature Communications* 7 (May 31, 2016), p. 11549. ISSN: 2041-1723. DOI: 10.1038/ncomms11549. pmid: 27240697 (page 50).

[61] T.J.P.van Dam. “CiliaCarta: An integrated and validated compendium of ciliary genes”. In: *PLoS One* 14 (2019), p. 0216705 (pages 68, 94).

[62] Petr Danecek et al. “The Variant Call Format and VCFtools”. In: *Bioinformatics* 27.15 (Aug. 1, 2011), pp. 2156–2158. ISSN: 1367-4803. DOI: 10.1093/bioinformatics/btr330 (page 146).

[63] Sayantan Das et al. “Next-Generation Genotype Imputation Service and Methods”. In: *Nature Genetics* 48.10 (10 Oct. 2016), pp. 1284–1287. ISSN: 1546-1718. DOI: 10.1038/ng.3656 (pages 97, 147).

[64] Joe R. Davis et al. “An Efficient Multiple-Testing Adjustment for eQTL Studies That Accounts for Linkage Disequilibrium between Variants”. In: *The American Journal of Human Genetics* 98.1 (Jan. 2016), pp. 216–224. ISSN: 00029297. DOI: 10.1016/j.ajhg.2015.11.021 (page 152).

[65] Luis de la Torre-Ubieta et al. “The Dynamic Landscape of Open Chromatin during Human Cortical Neurogenesis”. In: *Cell* 172.1-2 (Jan. 2018), 289–304.e18. ISSN: 00928674. DOI: 10.1016/j.cell.2017.12.014 (page 157).

[66] Ralph A. DeFronzo et al. “Type 2 Diabetes Mellitus”. In: *Nature Reviews Disease Primers* 1 (July 23, 2015), p. 15019. ISSN: 2056-676X. DOI: 10.1038/nrdp.2015.19 (pages 15, 17).

[67] Jacob F. Degner et al. “DNase I Sensitivity QTLs Are a Major Determinant of Human Expression Variation”. In: *Nature* 482.7385 (Feb. 2012), pp. 390–394. ISSN: 1476-4687. DOI: 10.1038/nature10808 (pages 122, 129, 134, 144).

[68] Lydia M. Dewitt. “MORPHOLOGY AND PHYSIOLOGY OF AREAS OF LANGERHANS IN SOME VERTEBRATES”. In: *Journal of Experimental Medicine* 8.2 (Mar. 26, 1906), pp. 193–239. ISSN: 1540-9538, 0022-1007. DOI: 10.1084/jem.8.2.193 (page 3).

[69] Jiarui Ding et al. *Systematic Comparative Analysis of Single Cell RNA-sequencing Methods*. preprint. Genomics, May 9, 2019. DOI: 10.1101/632216 (page 10).

[70] Amanda Dobbyn et al. “Landscape of Conditional eQTL in Dorsolateral Prefrontal Cortex and Co-localization with Schizophrenia GWAS”. In: *The American Journal of Human Genetics* 102.6 (June 2018), pp. 1169–1184. ISSN: 00029297. DOI: 10.1016/j.ajhg.2018.04.011 (page 145).

[71] Alexander Dobin et al. “STAR: Ultrafast Universal RNA-seq Aligner”. In: *Bioinformatics* 29.1 (Jan. 1, 2013), pp. 15–21. ISSN: 1367-4803. DOI: 10/f4h523 (pages 92, 98, 147).

[72] Alessandro Doria, Mary-Elizabeth Patti, and C. Ronald Kahn. “The Emerging Genetic Architecture of Type 2 Diabetes”. In: *Cell metabolism* 8.3 (Sept. 2008), pp. 186–200. ISSN: 1550-4131. DOI: 10.1016/j.cmet.2008.08.006. pmid: 18762020 (page 3).

[73] Craig Dorrell et al. “Human Islets Contain Four Distinct Subtypes of β Cells”. In: *Nature Communications* 7.1 (July 11, 2016), p. 11756. ISSN: 2041-1723. DOI: 10.1038/ncomms11756 (page 87).

[74] Craig Dorrell et al. “Isolation of Mouse Pancreatic Alpha, Beta, Duct and Acinar Populations with Cell Surface Markers”. In: *Molecular and cellular endocrinology* 339.1-2 (June 6, 2011), pp. 144–150. ISSN: 0303-7207. DOI: 10.1016/j.mce.2011.04.008. pmid: 21539888 (pages 16, 87).

[75] R.A.e Drigo. “New Insights into the Architecture of the Islet of Langerhans: A Focused Cross-Species Assessment”. In: *Diabetologia* 58 (2015), pp. 2218–2228. DOI: 10.1007/s00125-015-3699-0 (page 50).

[76] Sebastian Dütting, Sebastian Brachs, and Dirk Mielenz. “Fraternal Twins: Swiprosin-1/EFhd2 and Swiprosin-2/EFhd1, Two Homologous EF-hand Containing Calcium Binding Adaptor Proteins with Distinct Functions”. In: *Cell Communication and Signaling* 9.1 (Jan. 18, 2011), p. 2. ISSN: 1478-811X. DOI: 10.1186/1478-811X-9-2 (page 143).

[77] K. Eguchi and R. Nagai. “Islet inflammation in type 2 diabetes and physiology”. In: *The Journal of clinical investigation* 127 (2017), pp. 14–23. DOI: 10.1172/JCI88877 (pages 50, 82).

[78] Eleonora Porcu et al. “Mendelian Randomization Integrating GWAS and eQTL Data Reveals Genetic Determinants of Complex and Clinical Traits”. In: (2019). DOI: 10.1038/s41467-019-10936-0 (page 116).

[79] Sacha Epskamp and Eiko I. Fried. “A Tutorial on Regularized Partial Correlation Networks”. In: *Psychological Methods* 23.4 (Dec. 2018), pp. 617–634. ISSN: 1939-1463. DOI: 10.1037/met0000167. pmid: 29595293 (page 119).

[80] Gökcen Eraslan et al. “Deep Learning: New Computational Modelling Techniques for Genomics”. In: *Nature Reviews Genetics* 20.7 (7 July 2019), pp. 389–403. ISSN: 1471-0064. DOI: 10.1038/s41576-019-0122-6 (page 34).

[81] Gökcen Eraslan et al. “Single-Nucleus Cross-Tissue Molecular Reference Maps toward Understanding Disease Gene Function”. In: *Science* 376.6594 (May 13, 2022), eabl4290. DOI: 10.1126/science.abl4290 (page 115).

[82] A.S. Etheridge et al. *A New Liver eQTL Map from 1,183 Individuals Provides Evidence for Novel eQTLs of Drug Response, Metabolic and Sex-Biased Phenotypes* (pages 123, 134, 139, 158).

[83] David M. Evans and George Davey Smith. “Mendelian Randomization: New Applications in the Coming Age of Hypothesis-Free Causality”. In: *Annual Review of Genomics and Human Genetics* 16.1 (2015), pp. 327–350. DOI: 10.1146/annurev-genom-090314-050016. pmid: 25939054 (page 116).

[84] João Fadista et al. “Global Genomic and Transcriptomic Analysis of Human Pancreatic Islets Reveals Novel Genes Influencing Glucose Metabolism”. In: *Proceedings of the National Academy of Sciences* 111.38 (Sept. 23, 2014), pp. 13924–13929. ISSN: 0027-8424, 1091-6490. DOI: 10.1073/pnas.1402665111. pmid: 25201977;http://web.archive.org/web/20200421195244/https://www.pnas.org/content/111/38/13924 (page 9).

[85] Stefan S. Fajans, Graeme I. Bell, and Kenneth S. Polonsky. "Molecular Mechanisms and Clinical Pathophysiology of Maturity-Onset Diabetes of the Young". In: *New England Journal of Medicine* 345.13 (Sept. 27, 2001), pp. 971–980. ISSN: 0028-4793. DOI: 10.1056/NEJMra002168. pmid: 11575290 (page 1).

[86] Thorsten Falk et al. "U-Net: Deep Learning for Cell Counting, Detection, and Morphometry". In: *Nature Methods* 16.1 (Jan. 2019), pp. 67–70. ISSN: 1548-7105. DOI: 10.1038/s41592-018-0261-2 (page 34).

[87] L. Fina. "Expression of the CD34 Gene in Vascular Endothelial Cells". In: *Blood* 75 (1990), pp. 2417–26 (page 61).

[88] Hilary K. Finucane et al. "Partitioning Heritability by Functional Annotation Using Genome-Wide Association Summary Statistics". In: *Nature Genetics* 47.11 (11 Nov. 2015), pp. 1228–1235. ISSN: 1546-1718. DOI: 10.1038/ng.3404 (pages 126, 151).

[89] Marie P. Fogarty et al. "Identification of a Regulatory Variant That Binds FOXA1 and FOXA2 at the CDC123/CAMK1D Type 2 Diabetes GWAS Locus". In: *PLOS Genetics* 10.9 (Sept. 11, 2014), e1004633. ISSN: 1553-7404. DOI: 10.1371/journal.pgen.1004633 (pages 160, 161).

[90] Chris Fraley et al. *Mclust: Gaussian Mixture Modelling for Model-Based Clustering, Classification, and Density Estimation*. Version 5.4.5. July 8, 2019. URL: https://CRAN.R-project.org/package=mclust (visited on 08/27/2019) (page 39).

[91] François Chollet. *Keras*. GitHub, 2015. URL: https://github.com/keras-team/keras (visited on 08/30/2019) (page 42).

[92] Rachel E. Gate et al. "Genetic Determinants of Co-Accessible Chromatin Regions in Activated T Cells across Humans". In: *Nature Genetics* (July 9, 2018), p. 1. ISSN: 1546-1718. DOI: 10.1038/s41588-018-0156-2 (pages 122, 144, 159).

[93] "Genetic Effects on Gene Expression across Human Tissues". In: *Nature* 550.7675 (7675 Oct. 2017), pp. 204–213. ISSN: 1476-4687. DOI: 10.1038/nature24277 (pages 122, 143).

[94] P.A. Gerber and G.A. Rutter. "The Role of Oxidative Stress and Hypoxia in Pancreatic Beta-Cell Dysfunction in Diabetes Mellitus". In: *Antioxidants & Redox Signaling* 26 (2017), pp. 501–518. DOI: 10.1089/ars.2016.6755 (page 50).

[95] Claudia Giambartolomei et al. "Bayesian Test for Colocalisation between Pairs of Genetic Association Studies Using Summary Statistics". In: *PLOS Genetics* 10.5 (May 15, 2014), e1004383. ISSN: 1553-7404. DOI: 10.1371/journal.pgen.1004383 (pages 134, 145, 158).

[96] Todd M. Gierahn et al. "Seq-Well: Portable, Low-Cost RNA Sequencing of Single Cells at High Throughput". In: *Nature Methods* 14.4 (Apr. 2017), pp. 395–398. ISSN: 1548-7105. DOI: 10.1038/nmeth.4179 (page 40).

[97] Jeffrey M. Granja et al. "Single-Cell Multiomic Analysis Identifies Regulatory Programs in Mixed-Phenotype Acute Leukemia". In: *Nature Biotechnology* 37.12 (Dec. 2019), pp. 1458–1465. ISSN: 1546-1696. DOI: 10/ggfft8 (page 10).

[98] Charles E. Grant, Timothy L. Bailey, and William Stafford Noble. "FIMO: Scanning for Occurrences of a given Motif". In: *Bioinformatics* 27.7 (Apr. 1, 2011), pp. 1017–1018. ISSN: 1367-4803. DOI: 10/fcp52k (pages 103, 155).

[99] William W. Greenwald et al. "Pancreatic Islet Chromatin Accessibility and Conformation Reveals Distal Enhancer Networks of Type 2 Diabetes Risk". In: *Nature Communications* 10.1 (May 7, 2019), p. 2078. ISSN: 2041-1723. DOI: 10.1038/s41467-019-09975-4 (pages 30, 31, 45).

[100] Tyler Grimes, S. Steven Potter, and Somnath Datta. "Integrating Gene Regulatory Pathways into Differential Network Analysis of Gene Expression Data". In: *Scientific Reports* 9.1 (1 Apr. 2, 2019), p. 5479. ISSN: 2045-2322. DOI: 10.1038/s41598-019-41918-3 (page 119).

[101] Leif C. Groop. "The Molecular Genetics of Non-Insulin-Dependent Diabetes Mellitus". In: *Journal of Internal Medicine* 241.2 (1997), pp. 95–101. ISSN: 1365-2796. DOI: 10.1046/j.1365-2796.1997.99897000.x (page 1).

[102] Michael Hahsler et al. *Dbscan: Density Based Clustering of Applications with Noise (DBSCAN) and Related Algorithms*. Version 1.1-4. Aug. 5, 2019. URL: https://CRAN.R-project.org/package=dbscan (visited on 08/07/2019) (page 40).

[103] P.A. Halban. "β-Cell Failure in Type 2 Diabetes: Postulated Mechanisms and Prospects for Prevention and Treatment". In: *Diabetes Care* 37 (2014), pp. 1751–1758. DOI: 10.2337/dc14-0396 (page 49).

[104] R. Haliyur. "Human Islets Expressing HNF1A Variant Have Defective β Cell Transcriptional Regulatory Networks". In: *The Journal of clinical investigation* 129 (2018), pp. 246–251. DOI: 10.1172/JCI121994 (page 88).

[105] Hye-Sook Han et al. "Arginine Methylation of CRTC2 Is Critical in the Transcriptional Control of Hepatic Glucose Metabolism". In: *Science Signaling* 7.314 (Feb. 25, 2014). ISSN: 1945-0877, 1937-9145. DOI: 10.1126/scisignal.2004479 (page 143).

[106] Y. Hao. "Integrated analysis of multimodal single-cell data". In: *Cell* 184 (2021), pp. 3573–3587 29. DOI: 10.1016/j.cell.2021.04.048 (page 101).

[107] Jennifer Harrow et al. "GENCODE: The Reference Human Genome Annotation for The ENCODE Project". In: *Genome Research* 22.9 (Sept. 1, 2012), pp. 1760–1774. ISSN: 1088-9051, 1549-5469. DOI: 10.1101/gr.135350.111. pmid: 22955987 (page 41).

[108] Nathaniel J. Hart et al. "Cystic Fibrosis-Related Diabetes Is Caused by Islet Loss and Inflammation". In: *JCI insight* 3.8 (Apr. 19, 2018). ISSN: 2379-3708. DOI: 10.1172/jci.insight.98240. pmid: 29669939 (page 88).

[109] Stephen W. Hartley and James C. Mullikin. "QoRTs: A Comprehensive Toolset for Quality Control and Data Processing of RNA-Seq Experiments". In: *BMC Bioinformatics* 16.1 (July 19, 2015). ISSN: 1471-2105. DOI: 10.1186/s12859-015-0670-5. pmid: 26187896 (pages 92, 147).

[110] Andreas Heger et al. "GAT: A Simulation Framework for Testing the Association of Genomic Intervals". In: *Bioinformatics* 29.16 (Aug. 15, 2013), pp. 2046–2048. ISSN: 1367-4803. DOI: 10.1093/bioinformatics/btt343 (pages 28, 37, 44).

[111] Steven Henikoff et al. “Efficient Chromatin Accessibility Mapping in Situ by Nucleosome-Tethered Tagmentation”. In: *eLife* 9 (Nov. 16, 2020), e63274. ISSN: 2050-084X. DOI: 10.7554/eLife.63274 (page 9).

[112] Jay R. Hesselberth et al. “Global Mapping of Protein-DNA Interactions *in Vivo* by Digital Genomic Footprinting”. In: *Nature Methods* 6.4 (Apr. 2009), pp. 283–289. ISSN: 1548-7105. DOI: 10.1038/nmeth.1313 (page 16).

[113] A. S. Hinrichs et al. “The UCSC Genome Browser Database: Update 2006”. In: *Nucleic Acids Research* 34 (suppl_1 Jan. 1, 2006), pp. D590–D598. ISSN: 0305-1048. DOI: 10.1093/nar/gkj144 (page 159).

[114] Denes Hnisz et al. “Super-Enhancers in the Control of Cell Identity and Disease”. In: *Cell* 155.4 (Nov. 7, 2013), pp. 934–947. ISSN: 0092-8674, 1097-4172. DOI: 10.1016/j.cell.2013.09.053. pmid: 24119843 (pages 7, 19).

[115] Home et al. *IDF Diabetes Atlas | Tenth Edition*. URL: https://diabetesatlas.org/ (visited on 03/20/2022) (page 1).

[116] Hiroshi Hosoda, Hiroshi Tamura, and Isao Nagaoka. “Evaluation of the Lipopolysaccharide-Induced Transcription of the Human TREM-1 Gene in Vitamin D3-matured THP-1 Macrophage-like Cells”. In: *International Journal of Molecular Medicine* 36.5 (Nov. 2015), pp. 1300–1310. ISSN: 1107-3756, 1791-244X. DOI: 10.3892/ijmm.2015.2349 (page 160).

[117] Qin Qin Huang et al. “Power, False Discovery Rate and Winner’s Curse in eQTL Studies”. In: *Nucleic Acids Research* 46.22 (Dec. 14, 2018), e133. ISSN: 0305-1048. DOI: 10.1093/nar/gky780 (page 152).

[118] *International Diabetes Federation: Facts & Figures*. URL: https://idf.org/aboutdiabetes/what-is-diabetes/facts-figures.html (visited on 03/20/2022) (page 1).

[119] Valentina Iotchkova et al. “GARFIELD Classifies Disease-Relevant Genomic Features through Integration of Functional Annotations with Association Signals”. In: *Nature Genetics* 51.2 (2 Feb. 2019), pp. 343–353. ISSN: 1546-1718. DOI: 10.1038/s41588-018-0322-6 (pages 95, 105, 154).

[120] Åsa Johansson et al. “Partial Correlation Network Analyses to Detect Altered Gene Interactions in Human Disease: Using Preeclampsia as a Model”. In: *Human Genetics* 129.1 (Jan. 2011), pp. 25–34. ISSN: 0340-6717. DOI: 10.1007/s00439-010-0893-5. pmid: 20931231 (page 119).

[121] John Hensley. *Cta: C++ Implementation of Buenrostro Adapter Trimming*. The Parker Lab at the University of Michigan, May 12, 2017. URL: https://github.com/ParkerLab/cta (visited on 08/30/2019) (page 36).

[122] Arttu Jolma et al. “DNA-Binding Specificities of Human Transcription Factors”. In: *Cell* 152.1 (Jan. 17, 2013), pp. 327–339. ISSN: 0092-8674. DOI: 10.1016/j.cell.2012.12.009 (page 103).

[123] Jennifer Jou et al. “The ENCODE Portal as an Epigenomics Resource”. In: *Current Protocols in Bioinformatics* 68.1 (2019), e89. ISSN: 1934-340X. DOI: 10.1002/cpbi.89 (page 150).

[124] Goo Jun et al. “Detecting and Estimating Contamination of Human DNA Samples in Sequencing and Array-Based Genotype Data”. In: *The American Journal of Human Genetics* 91.5 (Nov. 2012), pp. 839–848. ISSN: 00029297. DOI: 10.1016/j.ajhg.2012.09.004 (page 147).

[125] Inkyung Jung et al. “A Compendium of Promoter-Centered Long-Range Chromatin Interactions in the Human Genome”. In: *Nature Genetics* (Sept. 9, 2019), pp. 1–8. ISSN: 1546-1718. DOI: 10.1038/s41588-019-0494-8 (pages 122, 137, 139, 145, 157).

[126] Klaus H. Kaestner et al. “What Is a β Cell? – Chapter I in the Human Islet Research Network (HIRN) Review Series”. In: *Molecular Metabolism* 53 (Nov. 1, 2021), p. 101323. ISSN: 2212-8778. DOI: 10.1016/j.molmet.2021.101323 (page 4).

[127] S.E. Kahn et al. “The Beta Cell Lesion in Type 2 Diabetes: There Has to Be a Primary Functional Abnormality”. In: *Diabetologia* 52 (2009), pp. 1003–1012. DOI: 10.1007/s00125-009-1321-z (page 81).

[128] Steven E. Kahn, Rebecca L. Hull, and Kristina M. Utzschneider. “Mechanisms Linking Obesity to Insulin Resistance and Type 2 Diabetes”. In: *Nature* 444.7121 (Dec. 2006), p. 840. ISSN: 1476-4687. DOI: 10.1038/nature05482 (page 49).

[129] Masahiro Kanai et al. “Genetic Analysis of Quantitative Traits in the Japanese Population Links Cell Types to Complex Human Diseases”. In: *Nature Genetics* (Feb. 5, 2018), p. 1. ISSN: 1546-1718. DOI: 10.1038/s41588-018-0047-6 (pages 137, 143).

[130] H.M. Kang. “Multiplexed droplet single-cell RNA-sequencing using natural genetic variation”. In: *Nature biotechnology* 36 (2018), pp. 89–94. DOI: 10.1038/nbt.4042 (page 100).

[131] Donna Karolchik et al. “The UCSC Table Browser Data Retrieval Tool”. In: *Nucleic Acids Research* 32 (suppl_1 Jan. 1, 2004), pp. D493–D496. ISSN: 0305-1048. DOI: 10.1093/nar/gkh103 (page 149).

[132] N.S. Kayton. “Human Islet Preparations Distributed for Research Exhibit a Variety of Insulin-Secretory Profiles”. In: *American Journal of Physiology-endocrinology and Metabolism* 308 (2015), pp. 592–602. DOI: 10.1152/ajpendo.00437.2014 (page 86).

[133] W.J. Kent et al. “BigWig and BigBed: Enabling Browsing of Large Distributed Datasets”. In: *Bioinformatics (Oxford, England)* 26 (2010), pp. 2204–2207. DOI: 10.1093/bioinformatics/btq351 (page 102).

[134] P. Kheradpour and M. Kellis. “Systematic Discovery and Characterization of Regulatory Motifs in ENCODE TF Binding Experiments”. In: *Nucleic acids research* 42 (2014), pp. 2976–2987. DOI: 10.1093/nar/gkt1249 (page 103).

[135] Shubham Khetan et al. “Type 2 Diabetes–Associated Genetic Variants Regulate Chromatin Accessibility in Human Islets”. In: *Diabetes* 67.11 (Nov. 2018), pp. 2466–2477. ISSN: 0012-1797, 1939-327X. DOI: 10.2337/db18-0393 (pages 122, 129).

[136] Gleb Kichaev et al. “Leveraging Polygenic Functional Enrichment to Improve GWAS Power”. In: *The American Journal of Human Genetics* 104.1 (Jan. 2019), pp. 65–75. ISSN: 00029297. DOI: 10.1016/j.ajhg.2018.11.008 (page 143).

[137] German Kilimnik et al. “Altered Islet Composition and Disproportionate Loss of Large Islets in Patients with Type 2 Diabetes”. In: *PLOS ONE* 6.11 (Nov. 15, 2011), e27445. ISSN: 1932-6203. DOI: 10.1371/journal.pone.0027445 (page 3).

[138] Somi Kim, Nam-Kyung Yu, and Bong-Kiun Kaang. “CTCF as a Multifunctional Protein in Genome Regulation and Gene Expression”. In: *Experimental & Molecular Medicine* 47.6 (6 June 2015), e166–e166. ISSN: 2092-6413. DOI: 10.1038/emm.2015.33 (page 124).

[139] Diederik P. Kingma and Jimmy Ba. “Adam: A Method for Stochastic Optimization”. Dec. 22, 2014. arXiv: 1412.6980 [cs]. URL: http://arxiv.org/abs/1412.6980 (visited on 08/30/2019) (page 42).

[140] Derek Klarin et al. “Genetics of Blood Lipids among ~300,000 Multi-Ethnic Participants of the Million Veteran Program”. In: *Nature Genetics* 50.11 (11 Nov. 2018), pp. 1514–1523. ISSN: 1546-1718. DOI: 10.1038/s41588-018-0222-9 (pages 139, 143, 145).

[141] Sandy L. Klemm, Zohar Shipony, and William J. Greenleaf. “Chromatin Accessibility and the Regulatory Epigenome”. In: *Nature Reviews Genetics* 20.4 (Apr. 2019), p. 207. ISSN: 1471-0064. DOI: 10.1038/s41576-018-0089-8 (page 123).

[142] Pang Wei Koh, Emma Pierson, and Anshul Kundaje. “Denoising Genome-Wide Histone ChIP-seq with Convolutional Neural Networks”. In: *Bioinformatics* 33.14 (July 15, 2017), pp. i225–i233. ISSN: 1367-4803. DOI: 10.1093/bioinformatics/btx243 (page 34).

[143] Liis Kolberg et al. “Co-Expression Analysis Reveals Interpretable Gene Modules Controlled by Trans-Acting Genetic Variants”. In: *bioRxiv* (Apr. 24, 2020), p. 2020.04.22.055335. DOI: 10.1101/2020.04.22.055335 (page 94).

[144] J.S. Kooner et al. *Froguel*. Ed. by Kato P. et al. In collab. with Zimmet M. et al. 2011 (page 6).

[145] Natsuhiko Kumasaka, Andrew J. Knights, and Daniel J. Gaffney. “Fine-Mapping Cellular QTLs with RASQUAL and ATAC-seq”. In: *Nature Genetics* 48.2 (Feb. 2016), pp. 206–213. ISSN: 1061-4036. DOI: 10.1038/ng.3467 (pages 122, 126, 129, 152, 157, 158).

[146] Ina Kycia et al. “A Common Type 2 Diabetes Risk Variant Potentiates Activity of an Evolutionarily Conserved Islet Stretch Enhancer and Increases C2CD4A and C2CD4B Expression”. In: *American Journal of Human Genetics* 102.4 (Apr. 5, 2018), pp. 620–635. ISSN: 1537-6605. DOI: 10.1016/j.ajhg.2018.02.020. pmid: 29625024 (page 31).

[147] Avantika Lal et al. “Deep Learning-Based Enhancement of Epigenomics Data with AtacWorks”. In: *Nature Communications* 12.1 (1 Mar. 8, 2021), p. 1507. ISSN: 2041-1723. DOI: 10.1038/s41467-021-21765-5 (page 109).

[148] Peter Langfelder and Steve Horvath. “WGCNA: An R Package for Weighted Correlation Network Analysis”. In: *BMC Bioinformatics* 9.1 (1 Dec. 2008), pp. 1–13. ISSN: 1471-2105. DOI: 10.1186/1471-2105-9-559 (page 94).

[149] Eric Lau and Ze’ev A. Ronai. “ATF2 – at the Crossroad of Nuclear and Cytosolic Functions”. In: *Journal of Cell Science* 125.12 (June 15, 2012), pp. 2815–2824. ISSN: 0021-9533. DOI: 10.1242/jcs.095000 (page 130).

[150] Nathan Lawlor and Michael L. Stitzel. "(Epi)Genomic Heterogeneity of Pancreatic Islet Function and Failure in Type 2 Diabetes". In: *Molecular Metabolism*. Biomarkers of Beta-Cell Health and Dysfunction: Towards Personalised Diabetes Care 27 (Sept. 1, 2019), S15–S24. ISSN: 2212-8778. DOI: 10.1016/j.molmet.2019.06.002 (pages 30, 31, 45).

[151] Nathan Lawlor et al. "Alpha TC1 and Beta-TC-6 Genomic Profiling Uncovers Both Shared and Distinct Transcriptional Regulatory Features with Their Primary Islet Counterparts". In: *Scientific Reports* 7.1 (Sept. 20, 2017), p. 11959. ISSN: 2045-2322. DOI: 10.1038/s41598-017-12335-1. pmid: 28931935 (pages 21, 41).

[152] Nathan Lawlor et al. "Single-Cell Transcriptomes Identify Human Islet Cell Signatures and Reveal Cell-Type–Specific Expression Changes in Type 2 Diabetes". In: *Genome Research* 27.2 (Feb. 1, 2017), pp. 208–222. ISSN: 1088-9051, 1549-5469. DOI: 10.1101/gr.212720.116. pmid: 27864352 (pages 16, 22).

[153] Chee Lee, Snehal Patil, and Maureen A. Sartor. "RNA-Enrich: A Cut-off Free Functional Enrichment Testing Method for RNA-seq with Improved Detection Power". In: *Bioinformatics (Oxford, England)* 32.7 (Apr. 1, 2016), pp. 1100–1102. ISSN: 1367-4811. DOI: 10.1093/bioinformatics/btv694. pmid: 26607492 (page 93).

[154] Jeongwoo Lee, Do Young Hyeon, and Daehee Hwang. "Single-Cell Multiomics: Technologies and Data Analysis Methods". In: *Experimental & Molecular Medicine* 52.9 (9 Sept. 2020), pp. 1428–1442. ISSN: 2092-6413. DOI: 10.1038/s12276-020-0420-2 (page 10).

[155] H. Li and R. Durbin. "Fast and Accurate Short Read Alignment with Burrows–Wheeler Transform". In: *Bioinformatics (Oxford, England)* 25 (2009), pp. 1754–1760. DOI: 10.1093/bioinformatics/btp324 (page 99).

[156] Heng Li. "Aligning Sequence Reads, Clone Sequences and Assembly Contigs with BWA-MEM". Mar. 16, 2013. arXiv: 1303.3997 [q-bio]. URL: http://arxiv.org/abs/1303.3997 (visited on 08/26/2019) (page 36).

[157] Heng Li et al. "The Sequence Alignment/Map Format and SAMtools". In: *Bioinformatics* 25.16 (Aug. 15, 2009), pp. 2078–2079. ISSN: 1367-4803. DOI: 10.1093/bioinformatics/btp352 (pages 36, 149).

[158] Xin Li et al. "The Impact of Rare Variation on Gene Expression across Tissues". In: *Nature* 550.7675 (Oct. 2017), pp. 239–243. ISSN: 1476-4687. DOI: 10.1038/nature24267 (pages 9, 11).

[159] Y. Liao, G.K. Smyth, and W. Shi. "featureCounts: An Efficient General Purpose Program for Assigning Sequence Reads to Genomic Features". In: *Bioinformatics (Oxford, England)* 30 (2014), pp. 923–930. DOI: 10.1093/bioinformatics/btt656 (pages 92, 148).

[160] P.-R. Loh. "Reference-Based Phasing Using the Haplotype Reference Consortium Panel". In: *Nature genetics* 48 (2016), pp. 1443–1448. DOI: 10.1038/ng.3679 (page 98).

[161] Po-Ru Loh, Pier Francesco Palamara, and Alkes L. Price. "Fast and Accurate Long-Range Phasing in a UK Biobank Cohort". In: *Nature Genetics* 48.7 (7 July 2016), pp. 811–816. ISSN: 1546-1718. DOI: 10.1038/ng.3571 (page 147).

[162] Michael I. Love, Wolfgang Huber, and Simon Anders. "Moderated Estimation of Fold Change and Dispersion for RNA-seq Data with DESeq2". In: *Genome Biology* 15.12 (Dec. 2014), p. 550. ISSN: 1474-760X. DOI: 10.1186/s13059-014-0550-8 (pages 93, 101, 147, 148, 152).

[163] H.P. Luhn. "The Automatic Creation of Literature Abstracts". In: *IBM journal of research and development* 2 (1958), pp. 159–165. DOI: 10.1147/rd.22.0159 (page 91).

[164] Aaron T. L. Lun et al. "EmptyDrops: Distinguishing Cells from Empty Droplets in Droplet-Based Single-Cell RNA Sequencing Data". In: *Genome Biology* 20.1 (1 Dec. 2019), pp. 1–9. ISSN: 1474-760X. DOI: 10.1186/s13059-019-1662-y (page 99).

[165] A. Mahajan et al. "Thorand". In: *Nature genetics* 50 (2018). Ed. by Thorleifsson B. et al. In collab. with H. Grallert et al., pp. 1505–1513. DOI: 10.1038/s41588-018-0241-6 (page 6).

[166] Anubha Mahajan et al. "Fine-Mapping Type 2 Diabetes Loci to Single-Variant Resolution Using High-Density Imputation and Islet-Specific Epigenome Maps". In: *Nature Genetics* (Oct. 8, 2018), p. 1. ISSN: 1546-1718. DOI: 10.1038/s41588-018-0241-6. pmid: 30297969 (pages 2, 6, 15, 31, 35, 43, 48, 49, 95, 151).

[167] Ani Manichaikul et al. "Robust Relationship Inference in Genome-Wide Association Studies". In: *Bioinformatics* 26.22 (Nov. 15, 2010), pp. 2867–2873. ISSN: 1367-4803. DOI: 10.1093/bioinformatics/btq559 (page 146).

[168] Teri A. Manolio et al. "Finding the Missing Heritability of Complex Diseases". In: *Nature* 461.7265 (Oct. 2009), pp. 747–753. ISSN: 1476-4687. DOI: 10.1038/nature08494 (page 7).

[169] Piero Marchetti et al. "An Overview of Pancreatic Beta-Cell Defects in Human Type 2 Diabetes: Implications for Treatment". In: *Regulatory Peptides* 146.1 (Feb. 7, 2008), pp. 4–11. ISSN: 0167-0115. DOI: 10.1016/j.regpep.2007.08.017 (page 4).

[170] Marcel Martin. "Cutadapt Removes Adapter Sequences from High-Throughput Sequencing Reads". In: *EMBnet.journal* 17.1 (1 May 2, 2011), pp. 10–12. ISSN: 2226-6089. DOI: 10.14806/ej.17.1.200 (page 148).

[171] S.L. Masters. "Activation of the NLRP3 Inflammasome by Islet Amyloid Polypeptide Provides a Mechanism for Enhanced IL-1β in Type 2 Diabetes". In: *Nature immunology* 11 (2010), pp. 897–904. DOI: 10.1038/ni.1935 (page 82).

[172] A. V. Matveyenko and P. C. Butler. "Relationship between β-Cell Mass and Diabetes Onset". In: *Diabetes, Obesity and Metabolism* 10.s4 (Nov. 2008), pp. 23–31. ISSN: 1462-8902. DOI: 10.1111/j.1463-1326.2008.00939.x (page 5).

[173] Alexandre Mayran and Jacques Drouin. "Pioneer Transcription Factors Shape the Epigenetic Landscape". In: *Journal of Biological Chemistry* 293.36 (Sept. 2018), pp. 13795–13804. ISSN: 00219258. DOI: 10.1074/jbc.R117.001232 (page 130).

[174] M.I. McCarthy. "Painting a new picture of personalised medicine for diabetes". In: *Diabetologia* 60 (2017), pp. 793–799 (page 83).

[175] Mark I. McCarthy. "Genomics, Type 2 Diabetes, and Obesity". In: *New England Journal of Medicine* 363.24 (Dec. 9, 2010), pp. 2339–2350. ISSN: 0028-4793. DOI: 10.1056/NEJMra0906948. pmid: 21142536 (page 2).

[176] Leland McInnes et al. "UMAP: Uniform Manifold Approximation and Projection". In: *Journal of Open Source Software* 3.29 (Sept. 2, 2018), p. 861. ISSN: 2475-9066. DOI: 10.21105/joss.00861 (page 40).

[177] Robert C. McLeay and Timothy L. Bailey. "Motif Enrichment Analysis: A Unified Framework and an Evaluation on ChIP Data". In: *BMC Bioinformatics* 11.1 (1 Dec. 2010), pp. 1–11. ISSN: 1471-2105. DOI: 10.1186/1471-2105-11-165 (page 150).

[178] J.J. Meier and R.C. Bonadonna. "Role of Reduced β-Cell Mass versus Impaired β-Cell Function in the Pathogenesis of Type 2 Diabetes". In: *Diabetes Care* 36 (2013), pp. 113–119. DOI: 10.2337/dcS13-2008 (page 81).

[179] Irene Miguel-Escalada et al. "Human Pancreatic Islet Three-Dimensional Chromatin Architecture Provides Insights into the Genetics of Type 2 Diabetes". In: *Nature Genetics* 51.7 (July 2019), p. 1137. ISSN: 1546-1718. DOI: 10.1038/s41588-019-0457-0 (pages 30, 31, 45).

[180] Joanna Mitchelmore et al. "Functional Effects of Variation in Transcription Factor Binding Highlight Long-Range Gene Regulation by Epromoters". In: *Nucleic Acids Research* 48.6 (Apr. 6, 2020), pp. 2866–2879. ISSN: 0305-1048. DOI: 10.1093/nar/gkaa123 (page 155).

[181] D.L.Minireview Morris. "Emerging Concepts in Islet Macrophage Biology in Type 2 Diabetes". In: *Molecular endocrinology (Baltimore, Md.)* 29 (2015), pp. 946–962. DOI: 10.1210/me.2014-1393 (page 82).

[182] Mauro J. Muraro et al. "A Single-Cell Transcriptome Atlas of the Human Pancreas". In: *Cell Systems* 3.4 (Oct. 26, 2016), 385–394.e3. ISSN: 2405-4712. DOI: 10.1016/j.cels.2016.09.002 (page 16).

[183] K. Musunuru et al. "From Noncoding Variant to Phenotype via SORT1 at the 1p13 Cholesterol Locus". In: *Nature* 466 (2010), pp. 714–719. DOI: 10.1038/nature09266 (pages 137, 143).

[184] F. Myokai et al. "A Novel Lipopolysaccharide-Induced Transcription Factor Regulating Tumor Necrosis Factor α Gene Expression: Molecular Cloning, Sequencing, Characterization, and Chromosomal Assignment". In: *Proceedings of the National Academy of Sciences of the United States of America* 96.8 (1999), pp. 4518–4523. ISSN: 0027-8424. DOI: 10.1073/pnas.96.8.4518 (pages 139, 144).

[185] A. Naba. "The Matrisome: In Silico Definition and in Vivo Characterization by Proteomics of Normal and Tumor Extracellular Matrices". In: *Molecular & cellular proteomics : MCP* M111.014647 (2012) (page 94).

[186] Masahito Nagaki and Hisataka Moriwaki. "Transcription Factor HNF and Hepatocyte Differentiation". In: *Hepatology Research* 38.10 (2008), pp. 961–969. ISSN: 1872-034X. DOI: 10.1111/j.1872-034X.2008.00367.x (pages 124, 130).

[187] Takashi Nagano et al. "Single-Cell Hi-C Reveals Cell-to-Cell Variability in Chromosome Structure". In: *Nature* 502.7469 (2013), p. 59. DOI: 10.1038/nature12593 (page 9).

[188] G. Nair and M. Hebrok. "Islet Formation in Mice and Men: Lessons for the Generation of Functional Insulin-Producing β-Cells from Human Pluripotent Stem Cells". In: *Current opinion in genetics & development* 32 (2015), pp. 171–180. DOI: 10.1016/j.gde.2015.03.004 (page 50).

[189] Michael Neidlin, Smaragda Dimitrakopoulou, and Leonidas G. Alexopoulos. "Multi-Tissue Network Analysis for Drug Prioritization in Knee Osteoarthritis". In: *Scientific Reports* 9.1 (1 Oct. 23, 2019), p. 15176. ISSN: 2045-2322. DOI: 10.1038/s41598-019-51627-6 (page 118).

[190] Alexandra C. Nica et al. "Cell-Type, Allelic, and Genetic Signatures in the Human Pancreatic Beta Cell Transcriptome". In: *Genome Research* 23.9 (Sept. 2013), pp. 1554–1562. ISSN: 1088-9051. DOI: 10.1101/gr.150706.112. pmid: 23716500 (page 16).

[191] Glyn M. Noguchi and Mark O. Huising. "Integrating the Inputs That Shape Pancreatic Islet Hormone Release". In: *Nature Metabolism* 1.12 (12 Dec. 2019), pp. 1189–1201. ISSN: 2522-5812. DOI: 10.1038/s42255-019-0148-2 (page 49).

[192] Nuala A. O'Leary et al. "Reference Sequence (RefSeq) Database at NCBI: Current Status, Taxonomic Expansion, and Functional Annotation". In: *Nucleic Acids Research* 44.D1 (Jan. 4, 2016), pp. D733–745. ISSN: 1362-4962. DOI: 10.1093/nar/gkv1189. pmid: 26553804 (page 40).

[193] Yumiko Oishi and Ichiro Manabe. "Krüppel-Like Factors in Metabolic Homeostasis and Cardiometabolic Disease". In: *Frontiers in Cardiovascular Medicine* 5 (2018). ISSN: 2297-055X. DOI: 10.3389/fcvm.2018.00069 (page 124).

[194] Yukinori Okada et al. "Genetics of Rheumatoid Arthritis Contributes to Biology and Drug Discovery". In: *Nature* 506.7488 (7488 Feb. 2014), pp. 376–381. ISSN: 1476-4687. DOI: 10.1038/nature12873 (page 151).

[195] Rainer Opgen-Rhein and Korbinian Strimmer. "From Correlation to Causation Networks: A Simple Approximate Learning Algorithm and Its Application to High-Dimensional Plant Gene Expression Data". In: *BMC Systems Biology* 1.1 (Aug. 6, 2007), p. 37. ISSN: 1752-0509. DOI: 10.1186/1752-0509-1-37 (page 119).

[196] Eugene L. Opie. "ON THE RELATION OF CHRONIC INTERSTITIAL PANCREATITIS TO THE ISLANDS OF LANGERHANS AND TO DIABETES MELUTUS". In: *Journal of Experimental Medicine* 5.4 (Jan. 15, 1901), pp. 397–428. ISSN: 1540-9538, 0022-1007. DOI: 10.1084/jem.5.4.397 (page 3).

[197] P. Orchard et al. "Dynamic Visualization, and Validation of Bias in ATAC-Seq Data with Ataqv". In: *Cell systems* 10 (2020), pp. 298–306 4. DOI: 10.1016/j.cels.2020.02.009 (page 148).

[198] Peter Orchard et al. "Human and Rat Skeletal Muscle Single-Nuclei Multi-Omic Integrative Analyses Nominate Causal Cell Types, Regulatory Elements, and SNPs for Complex Traits". In: *Genome Research* (Nov. 23, 2021). ISSN: 1088-9051, 1549-5469. DOI: 10.1101/gr.268482.120 (pages 9, 11).

[199] Peter Orchard et al. “Quantification, Dynamic Visualization, and Validation of Bias in ATAC-Seq Data with Ataqv”. In: *Cell Systems* 10.3 (Mar. 25, 2020), 298–306.e4. ISSN: 2405-4712. DOI: 10.1016/j.cels.2020.02.009 (page 37).

[200] Eva Pampouille et al. “Differential Expression and Co-Expression Gene Network Analyses Reveal Molecular Mechanisms and Candidate Biomarkers Involved in Breast Muscle Myopathies in Chicken”. In: *Scientific Reports* 9.1 (1 Oct. 17, 2019), p. 14905. ISSN: 2045-2322. DOI: 10.1038/s41598-019-51521-1 (page 119).

[201] Stephen C. J. Parker et al. “Chromatin Stretch Enhancer States Drive Cell-Specific Gene Regulation and Harbor Human Disease Risk Variants”. In: *Proceedings of the National Academy of Sciences* 110.44 (Oct. 29, 2013), pp. 17921–17926. ISSN: 0027-8424, 1091-6490. DOI: 10.1073/pnas.1317023110. pmid: 24127591 (pages 7, 15, 19, 49).

[202] Lorenzo Pasquali et al. “Pancreatic Islet Enhancer Clusters Enriched in Type 2 Diabetes Risk-Associated Variants”. In: *Nature Genetics* 46.2 (Feb. 2014), pp. 136–143. ISSN: 1546-1718. DOI: 10.1038/ng.2870 (pages 15, 49).

[203] Kashyap A. Patel et al. “Heterozygous RFX6 Protein Truncating Variants Are Associated with MODY with Reduced Penetrance”. In: *Nature Communications* 8.1 (Oct. 12, 2017), p. 888. ISSN: 2041-1723. DOI: 10.1038/s41467-017-00895-9 (pages 70, 83).

[204] B.P. Piasecki, J. Burghoorn, and P. Swoboda. “Regulatory Factor X (RFX)-Mediated Transcriptional Rewiring of Ciliary Genes in Animals”. In: *Proc National Acad Sci* 107 (2010), pp. 12969–12974. DOI: 10.1073/pnas.0914241107 (page 84).

[205] J. Piccand. “Rfx6 Maintains the Functional Identity of Adult Pancreatic β Cells”. In: *Cell Reports* 9 (2014), pp. 2219–2232. DOI: 10.1016/j.celrep.2014.11.033 (page 83).

[206] Julie Piccand et al. “Rfx6 Maintains the Functional Identity of Adult Pancreatic β Cells”. In: *Cell Reports* (Dec. 17, 2014). DOI: 10.1016/j.celrep.2014.11.033 (pages 103, 104).

[207] Joseph K. Pickrell. “Joint Analysis of Functional Genomic Data and Genome-wide Association Studies of 18 Human Traits”. In: *The American Journal of Human Genetics* 94.4 (Apr. 3, 2014), pp. 559–573. ISSN: 0002-9297, 1537-6605. DOI: 10.1016/j.ajhg.2014.03.004. pmid: 24702953; http://web.archive.org/web/20200604122030/https://www.cell.com/ (pages 28, 43).

[208] Hannah A. Pliner et al. “Cicero Predicts Cis-Regulatory DNA Interactions from Single-Cell Chromatin Accessibility Data”. In: *Molecular Cell* (Aug. 2, 2018). ISSN: 1097-2765. DOI: 10.1016/j.molcel.2018.06.044 (pages 31, 44).

[209] Paolo Pozzilli and Umberto Di Mario. “Autoimmune Diabetes Not Requiring Insulin at Diagnosis (Latent Autoimmune Diabetes of the Adult): Definition, Characterization, and Potential Prevention”. In: *Diabetes Care* 24.8 (Aug. 1, 2001), pp. 1460–1467. ISSN: 0149-5992. DOI: 10.2337/diacare.24.8.1460 (page 1).

[210] Sebastian Preissl et al. “Single-Nucleus Analysis of Accessible Chromatin in Developing Mouse Forebrain Reveals Cell-Type-Specific Transcriptional Regulation”. In: *Nature Neuroscience* 21.3 (Mar. 2018), pp. 432–439. ISSN: 1546-1726. DOI: 10.1038/s41593-018-0079-3 (page 16).

[211] Randall J. Pruim et al. "LocusZoom: Regional Visualization of Genome-Wide Association Scan Results". In: *Bioinformatics* 26.18 (Sept. 15, 2010), pp. 2336–2337. ISSN: 1367-4803. DOI: 10.1093/bioinformatics/btq419 (page 159).

[212] Sara L Pulit et al. "Meta-Analysis of Genome-Wide Association Studies for Body Fat Distribution in 694 649 Individuals of European Ancestry". In: *Human Molecular Genetics* 28.1 (Jan. 1, 2019), pp. 166–174. ISSN: 0964-6906. DOI: 10.1093/hmg/ddy327 (page 151).

[213] Shaun Purcell et al. "PLINK: A Tool Set for Whole-Genome Association and Population-Based Linkage Analyses". In: *The American Journal of Human Genetics* 81.3 (Sept. 2007), pp. 559–575. ISSN: 00029297. DOI: 10.1086/519795 (page 147).

[214] A.R. Quinlan. "BEDTools: The Swiss-Army Tool for Genome Feature Analysis". In: *Curr Protoc Bioinform* 47 (2014), pp. 11 12 1–11 12 34. DOI: 10.1002/0471250953.bi1112s47 (pages 102, 149).

[215] Aaron R. Quinlan and Ira M. Hall. "BEDTools: A Flexible Suite of Utilities for Comparing Genomic Features". In: *Bioinformatics* 26.6 (Mar. 15, 2010), pp. 841–842. ISSN: 1367-4803. DOI: 10.1093/bioinformatics/btq033 (pages 43, 148, 155).

[216] J. Rahier et al. "Pancreatic B-cell Mass in European Subjects with Type 2 Diabetes". In: *Diabetes, obesity & metabolism* 10 (2008), pp. 32–42. DOI: 10.1111/j.1463-1326.2008.00969.x (page 81).

[217] Vivek Rai et al. "Single-Cell ATAC-Seq in Human Pancreatic Islets and Deep Learning Upscaling of Rare Cells Reveals Cell-Specific Type 2 Diabetes Regulatory Signatures". In: *Molecular Metabolism* 32 (Feb. 1, 2020), pp. 109–121. ISSN: 2212-8778. DOI: 10.1016/j.molmet.2019.12.006 (pages 12, 45, 46, 48, 49, 94, 148).

[218] Ryne C. Ramaker et al. "A Genome-Wide Interactome of DNA-associated Proteins in the Human Liver". In: *Genome Research* 27.11 (Nov. 1, 2017), pp. 1950–1960. ISSN: 1088-9051, 1549-5469. DOI: 10.1101/gr.222083.117. pmid: 29021291 (pages 124, 150).

[219] C.K. Raulerson et al. In: *Adipose Tissue Gene Expression Associations Reveal Hundreds of Candidate Genes for Cardiometabolic Traits. Am J Hum Genet* 105 (2019), pp. 773–787 (page 145).

[220] M.J. Redondo. "The Clinical Consequences of Heterogeneity within and between Different Diabetes Types". In: *Diabetologia* 63 (2020), pp. 2040–2048. DOI: 10.1007/s00125-020-05211-7 (pages 48, 49, 85).

[221] Antonio Reverter and Eva K. F. Chan. "Combining Partial Correlation and an Information Theory Approach to the Reversed Engineering of Gene Co-Expression Networks". In: *Bioinformatics* 24.21 (Nov. 1, 2008), pp. 2491–2497. ISSN: 1367-4803. DOI: 10.1093/bioinformatics/btn482 (page 119).

[222] Davide Risso et al. "Normalization of RNA-seq Data Using Factor Analysis of Control Genes or Samples". In: *Nature Biotechnology* 32.9 (9 Sept. 2014), pp. 896–902. ISSN: 1546-1696. DOI: 10.1038/nbt.2931 (page 93).

[223] Matthew E. Ritchie et al. "Limma Powers Differential Expression Analyses for RNA-sequencing and Microarray Studies". In: *Nucleic Acids Research* 43.7 (Apr. 20, 2015), e47–e47. ISSN: 0305-1048. DOI: 10.1093/nar/gkv007 (pages 94, 157).

[224] Lilian Beatriz Aguayo Rojas and Marilia Brito Gomes. “Metformin: An Old but Still the Best Treatment for Type 2 Diabetes”. In: *Diabetology & Metabolic Syndrome* 5.1 (Feb. 15, 2013), p. 6. ISSN: 1758-5996. DOI: 10.1186/1758-5996-5-6 (page 120).

[225] Tamara S. Roman et al. “A Type 2 Diabetes-Associated Functional Regulatory Variant in a Pancreatic Islet Enhancer at the ADCY5 Locus”. In: *Diabetes* 66.9 (Sept. 2017), pp. 2521–2530. ISSN: 1939-327X. DOI: 10.2337/db17-0464. pmid: 28684635 (page 31).

[226] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. “U-Net: Convolutional Networks for Biomedical Image Segmentation”. May 18, 2015. arXiv: 1505.04597 [cs]. URL: http://arxiv.org/abs/1505.04597 (visited on 08/26/2019) (pages 34, 41).

[227] O.L. Sabik et al. “Identification of a Core Module for Bone Mineral Density through the Integration of a Co-Expression Network and GWAS Data”. In: *Cell Reports* 32 (2020), p. 108145. DOI: 10.1016/j.celrep.2020.108145 (page 82).

[228] P. Saeedi. “Global and Regional Diabetes Prevalence Estimates for 2019 and Projections for 2030 and 2045: Results from the International Diabetes Federation Diabetes Atlas”. In: *Diabetes Research and Clinical Practice* 157 (2019), p. 107843. DOI: 10.1016/j.diabres.2019.107843 (pages 48, 49).

[229] H. Sakuraba. “Reduced Beta-Cell Mass and Expression of Oxidative Stress-Related DNA Damage in the Islet of Japanese Type II Diabetic Patients”. In: *Diabetologia* 45 (2002), pp. 85–96. DOI: 10.1007/s125-002-8248-z (page 81).

[230] H. Sasaki. “Reduced Beta Cell Number Rather than Size Is a Major Contributor to Beta Cell Loss in Type 2 Diabetes”. In: *Diabetologia* 64 (2021), pp. 1816–1821. DOI: 10.1007/s00125-021-05467-7 (page 81).

[231] Ansuman T. Satpathy et al. “Massively Parallel Single-Cell Chromatin Landscapes of Human Immune Cell Development and Intratumoral T Cell Exhaustion”. In: *Nature Biotechnology* 37.8 (Aug. 2019), pp. 925–936. ISSN: 1546-1696. DOI: 10.1038/s41587-019-0206-z (pages 10, 16, 35).

[232] D.C. Saunders. “Ectonucleoside Triphosphate Diphosphohydrolase-3 Antibody Targets Adult Human Pancreatic β Cells for in Vitro and in Vivo Analysis”. In: *Cell metabolism* 29 (2019), pp. 745–754 4. DOI: 10.1016/j.cmet.2018.10.007 (pages 87, 88).

[233] Ellen M. Schmidt et al. “GREGOR: Evaluating Global Enrichment of Trait-Associated Variants in Epigenomic Features Using a Systematic, Data-Driven Approach”. In: *Bioinformatics* 31.16 (Aug. 15, 2015), pp. 2601–2606. ISSN: 1367-4803. DOI: 10.1093/bioinformatics/btv201 (pages 30, 43).

[234] C.M. Schürch. “Coordinated Cellular Neighborhoods Orchestrate Antitumoral Immunity at the Colorectal Cancer Invasive Front”. In: *Cell* 182 (2020), pp. 1341–1359 19. DOI: 10.1016/j.cell.2020.07.005 (page 91).

[235] Laura J. Scott et al. “The Genetic Regulatory Signature of Type 2 Diabetes in Human Skeletal Muscle”. In: *Nature Communications* 7 (June 29, 2016), ncomms11764. ISSN: 2041-1723. DOI: 10.1038/ncomms11764 (pages 7, 9, 44, 148).

[236] Tony Scully. “Diabetes in Numbers”. In: *Nature* 485 (May 16, 2012), S2–S3. ISSN: 1476-4687. DOI: 10.1038/485S2a (page 1).

[237] Åsa Segerstolpe et al. “Single-Cell Transcriptome Profiling of Human Pancreatic Islets in Health and Type 2 Diabetes”. In: *Cell Metabolism* 24.4 (Oct. 11, 2016), pp. 593–607. ISSN: 1550-4131. DOI: 10.1016/j.cmet.2016.08.020 (pages 20, 21, 41).

[238] Nathan C. Sheffield et al. “Patterns of Regulatory Activity across Diverse Human Cell Types Predict Tissue Identity, Transcription Factor Binding, and Long-Range Interactions”. In: *Genome Research* 23.5 (May 1, 2013), pp. 777–788. ISSN: 1088-9051, 1549-5469. DOI: 10.1101/gr.152140.112. pmid: 23482648 (page 145).

[239] Efrat Shema, Bradley E. Bernstein, and Jason D. Buenrostro. “Single-Cell and Single-Molecule Epigenomics to Uncover Genome Regulation at Unprecedented Resolution”. In: *Nature Genetics* (Dec. 17, 2018), p. 1. ISSN: 1546-1718. DOI: 10.1038/s41588-018-0290-x (page 16).

[240] S. T. Sherry et al. “dbSNP: The NCBI Database of Genetic Variation”. In: *Nucleic Acids Research* 29.1 (Jan. 1, 2001), pp. 308–311. ISSN: 0305-1048. DOI: 10.1093/nar/29.1.308 (page 159).

[241] S.B. Smith. “Rfx6 Directs Islet Formation and Insulin Production in Mice and Humans”. In: *Nature* 463 (2010), pp. 775–780. DOI: 10.1038/nature08748 (pages 70, 83).

[242] M.L. Speir. “UCSC Cell Browser: Visualize Your Single-Cell Data”. In: *Bioinformatics (Oxford, England)* btab503 (2021). DOI: 10.1093/bioinformatics/btab503 (page 101).

[243] Craig W. Spellman. “Pathophysiology of Type 2 Diabetes: Targeting Islet Cell Dysfunction”. In: *The Journal of the American Osteopathic Association* 110 (3_suppl_2 Mar. 1, 2010), S2–S7. ISSN: 0098-6151. URL: https://jaoa.org/article.aspx?articleid=2093868 (visited on 09/06/2019) (page 15).

[244] Tobias Strunz et al. “A Mega-Analysis of Expression Quantitative Trait Loci (eQTL) Provides Insight into the Regulatory Architecture of Gene Expression Variation in Liver”. In: *Scientific Reports* 8.1 (1 Apr. 12, 2018), p. 5865. ISSN: 2045-2322. DOI: 10.1038/s41598-018-24219-z (page 123).

[245] T. Stuart. “Comprehensive Integration of Single-Cell Data”. In: *Cell* 177 (2019), pp. 1888–1902 21. DOI: 10.1016/j.cell.2019.05.031 (page 101).

[246] Michael Stumvoll, Barry J Goldstein, and Timon W van Haeften. “Type 2 Diabetes: Principles of Pathogenesis and Therapy”. In: *The Lancet* 365.9467 (Apr. 9, 2005), pp. 1333–1346. ISSN: 0140-6736. DOI: 10.1016/S0140-6736(05)61032-X (page 2).

[247] Fran Supek et al. “REVIGO Summarizes and Visualizes Long Lists of Gene Ontology Terms”. In: *PLOS ONE* 6.7 (July 18, 2011), e21800. ISSN: 1932-6203. DOI: 10.1371/journal.pone.0021800 (page 93).

[248] Daniel I. Swerdlow. “Mendelian Randomization and Type 2 Diabetes”. In: *Cardiovascular Drugs and Therapy* 30.1 (Feb. 1, 2016), pp. 51–57. ISSN: 1573-7241. DOI: 10.1007/s10557-016-6638-5 (page 116).

[249] C. Talchai et al. “Pancreatic β Cell Dedifferentiation as a Mechanism of Diabetic β Cell Failure”. In: *Cell* 150 (2012), pp. 1223–1234. DOI: 10.1016/j.cell.2012.07.029 (page 50).

[250] Xiaoning Tang et al. “The Single-Cell Sequencing: New Developments and Medical Applications”. In: *Cell & Bioscience* 9.1 (June 26, 2019), p. 53. ISSN: 2045-3701. DOI: 10.1186/s13578-019-0314-y (page 10).

[251] R.C. Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, 2020 (page 149).

[252] The 1000 Genomes Project Consortium. “A Global Reference for Human Genetic Variation”. In: *Nature* 526.7571 (Oct. 2015), pp. 68–74. ISSN: 1476-4687. DOI: 10/73d (page 44).

[253] The ENCODE Project Consortium. “An Integrated Encyclopedia of DNA Elements in the Human Genome”. In: *Nature* 489.7414 (Sept. 2012), pp. 57–74. ISSN: 1476-4687. DOI: 10.1038/nature11247 (page 36).

[254] the Haplotype Reference Consortium. “A Reference Panel of 64,976 Haplotypes for Genotype Imputation”. In: *Nature Genetics* 48.10 (Oct. 2016), pp. 1279–1283. ISSN: 1061-4036, 1546-1718. DOI: 10.1038/ng.3643 (page 147).

[255] THE TABULA SAPIENS CONSORTIUM. “The Tabula Sapiens: A Multiple-Organ, Single-Cell Transcriptomic Atlas of Humans”. In: *Science* 376.6594 (May 13, 2022), eabl4896. DOI: 10.1126/science.abl4896 (page 115).

[256] Asa Thibodeau et al. “AMULET: A Novel Read Count-Based Method for Effective Multiplet Detection from Single Nucleus ATAC-seq Data”. In: *Genome Biology* 22.1 (Sept. 1, 2021), p. 252. ISSN: 1474-760X. DOI: 10.1186/s13059-021-02469-x (page 101).

[257] Matthias Thurner et al. “Integration of Human Pancreatic Islet Genomic Data Refines Regulatory Mechanisms at Type 2 Diabetes Susceptibility Loci”. In: *eLife* 7 (Feb. 7, 2018), e31977. ISSN: 2050-084X. DOI: 10.7554/eLife.31977 (pages 15, 49).

[258] E. Trefts, M. Gannon, and D.H. Wasserman. “The Liver”. In: *Current biology : CB* 27 (2017), pp. 1147–1151. DOI: 10.1016/j.cub.2017.09.019 (pages 123, 136).

[259] Gosia Trynka et al. “Chromatin Marks Identify Critical Cell Types for Fine Mapping Complex Trait Variants”. In: *Nature Genetics* 45.2 (Feb. 2013), pp. 124–130. ISSN: 1546-1718. DOI: 10.1038/ng.2504 (page 49).

[260] R.H. Unger and A.D. Cherrington. “Glucagonocentric Restructuring of Diabetes: A Pathophysiologic and Therapeutic Makeover”. In: *The Journal of clinical investigation* 122 (2012), pp. 4–12. DOI: 10.1172/JCI60016 (page 52).

[261] Tatsuhiko Urakami. “Maturity-Onset Diabetes of the Young (MODY): Current Perspectives on Diagnosis and Treatment”. In: *Diabetes, Metabolic Syndrome and Obesity: Targets and Therapy* Volume 12 (July 2019), pp. 1047–1056. ISSN: 1178-7007. DOI: 10.2147/DMSO.S179793 (page 5).

[262] Bryce van de Geijn et al. “WASP: Allele-Specific Software for Robust Molecular Quantitative Trait Locus Discovery”. In: *Nature Methods* 12.11 (Nov. 2015), pp. 1061–1063. ISSN: 1548-7091. DOI: 10.1038/nmeth.3582 (pages 9, 127, 153).

[263] Pim van der Harst and Niek Verweij. “Identification of 64 Novel Genetic Loci Provides an Expanded View on the Genetic Architecture of Coronary Artery Disease”. In: *Circulation Research* 122.3 (Feb. 2, 2018), pp. 433–443. ISSN: 0009-7330, 1524-4571. DOI: 10.1161/CIRCRESAHA.117.312086 (page 151).

[264] Arushi Varshney et al. “Genetic Regulatory Signatures Underlying Islet Gene Expression and Type 2 Diabetes”. In: *Proceedings of the National Academy of Sciences* 114.9 (Feb. 28, 2017), pp. 2301–2306. ISSN: 0027-8424, 1091-6490. DOI: 10.1073/pnas.1621192114. pmid: 28193859 (pages 7, 9, 15, 19, 31, 33, 36, 49, 70, 83, 103, 147).

[265] Ana Viñuela et al. “Genetic Variant Effects on Gene Expression in Human Pancreatic Islets and Their Implications for T2D”. In: *Nature Communications* 11.1 (1 Sept. 30, 2020), p. 4912. ISSN: 2041-1723. DOI: 10.1038/s41467-020-18581-8 (page 33).

[266] Ana Viñuela et al. “Influence of Genetic Variants on Gene Expression in Human Pancreatic Islets – Implications for Type 2 Diabetes”. In: *bioRxiv* (May 31, 2019), p. 655670. DOI: 10.1101/655670 (page 15).

[267] M. Vujkovic et al. Wilson, P.W., Edwards, T.L., Rader, D.J., Damrauer, S.M., O’Donnell, C.J., Tsao, P.S., Chang, K.-M., Voight, B.F, 2019 (page 6).

[268] John T Walker et al. *RFX6-mediated Dysregulation Defines Human β Cell Dysfunction in Early Type 2 Diabetes*. Dec. 17, 2021. DOI: 10.1101/2021.12.16.466282 (pages 12, 51, 88, 105, 113).

[269] John T. Walker et al. “Integrated Human Pseudoislet System and Microfluidic Platform Demonstrate Differences in GPCR Signaling in Islet Cells”. In: *JCI Insight* 5.10 (May 21, 2020). ISSN: 0021-9738. DOI: 10.1172/jci.insight.137017 (pages 86, 95).

[270] Liguo Wang et al. “Measure Transcript Integrity Using RNA-seq Data”. In: *BMC Bioinformatics* 17.1 (Feb. 3, 2016), p. 58. ISSN: 1471-2105. DOI: 10.1186/s12859-016-0922-z (page 92).

[271] M.T. Weirauch. “Determination and Inference of Eukaryotic Transcription Factor Sequence Specificity”. In: *Cell* 158 (2014), pp. 1431–1443. DOI: 10.1016/j.cell.2014.08.009 (page 95).

[272] Matthew T. Weirauch et al. “Determination and Inference of Eukaryotic Transcription Factor Sequence Specificity”. In: *Cell* 158.6 (Sept. 11, 2014), pp. 1431–1443. ISSN: 0092-8674. DOI: 10.1016/j.cell.2014.08.009 (page 149).

[273] J. Weitz, D. Menegaz, and A. Caicedo. “Deciphering the Complex Communication Networks That Orchestrate Pancreatic Islet Function”. In: *Diabetes* 70 (2020), pp. 17–26. DOI: 10.2337/dbi19-0033 (page 49).

[274] P. Westermark, A. Andersson, and G.T. Westermark. “Islet Amyloid Polypeptide, Islet Amyloid, and Diabetes Mellitus”. In: *Physiological reviews* 91 (2011), pp. 795–826. DOI: 10.1152/physrev.00042.2009 (page 50).

[275] C.Y. Westwell-Roper, J.A. Ehses, and C.B. Verchere. “Resident Macrophages Mediate Islet Amyloid Polypeptide–Induced Islet IL-1β Production and β-Cell Dysfunction”. In: *Diabetes* 63 (2014), pp. 1698–1711. DOI: 10.2337/db13-0863 (page 82).

[276] L. Wigger. “Multi-Omics Profiling of Living Human Pancreatic Islet Donors Reveals Heterogeneous Beta Cell Trajectories towards Type 2 Diabetes”. In: *Nat Metabolism* 3 (2021), pp. 1017–1031. DOI: 10.1038/s42255-021-00420-9 (pages 5, 11, 50, 81).

[277] Cristen J. Willer et al. “Discovery and Refinement of Loci Associated with Lipid Levels”. In: *Nature genetics* 45.11 (Nov. 2013), pp. 1274–1283. ISSN: 1061-4036. DOI: 10.1038/ng.2797. pmid: 24097068 (pages 123, 151).

[278] A.L. Williams et al. “Sequence Variants in SLC16A11 Are a Common Risk Factor for Type 2 Diabetes in Mexico”. In: *Nature* 506 (2014), pp. 97–101. DOI: 10.1038/nature12828 (page 6).

[279] Cecily J. Wolfe, Isaac S. Kohane, and Atul J. Butte. “Systematic Survey Reveals General Applicability of ”Guilt-by-Association” within Gene Coexpression Networks”. In: *BMC Bioinformatics* 6.1 (Sept. 14, 2005), p. 227. ISSN: 1471-2105. DOI: 10.1186/1471-2105-6-227 (page 117).

[280] Haojia Wu et al. “Advantages of Single-Nucleus over Single-Cell RNA Sequencing of Adult Kidney: Rare Cell Types and Novel Cell States Revealed in Fibrosis”. In: *Journal of the American Society of Nephrology* 30.1 (Jan. 2019), pp. 23–32. ISSN: 1046-6673, 1533-3450. DOI: 10.1681/ASN.2018090912 (page 10).

[281] M. Wu. “Single-Cell Analysis of the Human Pancreas in Type 2 Diabetes Using Multi-Spectral Imaging Mass Cytometry”. In: *Cell Reports* 37 (2021), p. 109919. DOI: 10.1016/j.celrep.2021.109919 (page 62).

[282] S. Yang. “Decontamination of Ambient RNA in Single-Cell RNA-seq with DecontX”. In: *Genome biology* 21 (2020), p. 57. DOI: 10.1186/s13059-020-1950-6 (page 100).

[283] Thomas W. Yee. *Vector Generalized Linear and Additive Models: With an Implementation in R*. Springer Series in Statistics. New York, NY: Springer, 2015. ISBN: 978-1-4939-2818-7 (page 153).

[284] Loic Yengo et al. “Meta-Analysis of Genome-Wide Association Studies for Height and Body Mass Index in ∼700000 Individuals of European Ancestry”. In: *Human Molecular Genetics* 27.20 (Oct. 15, 2018), pp. 3641–3649. ISSN: 0964-6906. DOI: 10.1093/hmg/ddy271 (pages 145, 151).

[285] Yong Zhang et al. “Model-Based Analysis of ChIP-Seq (MACS)”. In: *Genome Biology* 9 (Sept. 17, 2008), R137. ISSN: 1474-760X. DOI: 10.1186/gb-2008-9-9-r137 (pages 36, 102, 148).

[286] Grace X. Y. Zheng et al. “Massively Parallel Digital Transcriptional Profiling of Single Cells”. In: *Nature Communications* 8 (Jan. 16, 2017), p. 14049. ISSN: 2041-1723. DOI: 10.1038/ncomms14049 (pages 9, 10).

[287] Yingyao Zhou et al. “Metascape Provides a Biologist-Oriented Resource for the Analysis of Systems-Level Datasets”. In: *Nature Communications* 10.1 (Apr. 3, 2019), p. 1523. ISSN: 2041-1723. DOI: 10.1038/s41467-019-09234-6. pmid: 30944313 (page 93).

[288] Chenxu Zhu et al. “Joint Profiling of Histone Modifications and Transcriptome in Single Cells from Mouse Brain”. In: *Nature Methods* 3 (Feb. 15, 2021), pp. 1–10. ISSN: 1548-7105. DOI: 10.1038/s41592-021-01060-3 (page 11).

[289] Paul Zimmet, K. G. M. M. Alberti, and Jonathan Shaw. “Global and Societal Implications of the Diabetes Epidemic”. In: *Nature* 414.6865 (Dec. 2001), p. 782. ISSN: 1476-4687. DOI: 10.1038/414782a (page 1).