Deciphering Type 2 Diabetes-associated SNP Effects on Islet Chromatin Accessibility and Beta Cell Expression Levels Using Functional Genomics

Shubham Khetan

University of Connecticut - Storrs, khetan@uchc.edu

Follow this and additional works at: https://opencommons.uconn.edu/dissertations

Recommended Citation

https://opencommons.uconn.edu/dissertations/2608
Deciphering Type 2 Diabetes-associated SNP Effects on Islet Chromatin Accessibility and Beta Cell Expression Levels Using Functional Genomics.

Shubham Khetan, Ph.D.
University of Connecticut, 2020

Type 2 diabetes (T2D) results when pancreatic beta cells are unable to secrete adequate amounts of insulin in response to elevated blood glucose levels. Genome-wide association studies (GWAS) have linked single nucleotide polymorphisms (SNPs) in >250 regions of the human genome to a higher risk for developing T2D. These T2D-associated GWAS SNPs are predominantly non-coding (~90%) and enriched at islet regulatory elements, implicating altered transcriptional regulation in islets as a molecular mechanism underlying T2D risk. However, the sharp rise in T2D incidence over the past few decades cannot be attributed to changes in the human genome, but our environment, such as rising overnutrition and pollution levels. Therefore, T2D is a multifactorial disorder with both genetic and environmental components risk factors. In order to elucidate the mechanistic link between non-coding SNPs and a complex trait like T2D, we first identified functional SNPs among the 10s-100s that are in linkage disequilibrium (LD) at any given T2D-associated GWAS signal. We then show that endoplasmic reticulum (ER) stress in beta cells is an important environmental context for the genetic risk of T2D. ER folding capacity, therefore, may be a major factor determining how much insulin can be released by beta cells before stress ensues. SNPs at regulatory elements modulating the expression of genes relevant to meet higher demands for insulin synthesis may impair or improve ER capacity, ultimately determining the threshold at which ER stress ensues, leading to beta cell failure and T2D.
Deciphering Type 2 Diabetes-associated SNP Effects on Islet Chromatin Accessibility and Beta Cell Expression Levels Using Functional Genomics.

B.Tech., IIT Roorkee, 2014

A Dissertation

Submitted

in Partial Fulfillment of the

Requirements for the Degree of

Doctor of Philosophy

at the

University of Connecticut

2020
Deciphering Type 2 Diabetes-associated SNP Effects on Islet Chromatin Accessibility and Beta Cell Expression Levels using Functional Genomics.

Presented by
Shubham Khetan, B.Tech.

Major Advisor – 1
Duygu Ucar, Ph.D.

Major Advisor – 2
Michael Stitzel, Ph.D.

Associate Advisor
Justin Cotney, Ph.D.

Associate Advisor
Stormy Chamberlain, Ph.D.

University of Connecticut
2020
PREFACE

Portions of this dissertation have been published:


Acknowledgements

This thesis is dedicated to my elder brother, Aditya, who introduced me to the boundless wonder in books, and to my parents, who loved me enough to send me away to learn.

I thank Nikunj and Neha, who illuminate my life with their presence every day.

Ashley, Kobi, Marc, Brittany, Adam and Anu - I love you guys and thank you so much for your friendship. You have brightened many dark days.

I thank Siddhant, who steered me in the direction of seeking truth for a living.

Michael and Duygu - I thank you for your guidance, support and trust. I am blessed to have found nurturing mentors like you.

I thank all members of the human race, past and present, who have sought truth and taught it. This thesis stands on your shoulders.
Table of Contents

Chapter 1 - Introduction .......................................................................................................... 1
   A. Type 2 diabetes is a multifactorial disorder with both genetic and environmental risk factors ..... 2
   B. Interaction of genetic and environmental factors determine endoplasmic reticulum load in beta cells ................................................................................................................................................. 3
   C. Approaches to study T2D-associated genetic and environmental variations altering transcriptional activation in beta cells ............................................................................................................................. 5

Chapter 2 – Type 2 diabetes-associated genetic variants regulate chromatin accessibility in human islets ........................................................................................................................................................................... 15

Chapter 3 – Functional characterization of thousands of type 2 diabetes-associated and chromatin-modulating variants under steady state and endoplasmic reticulum stress ........ 45

Chapter 4 – Discussion and Future Directions ........................................................................ 95
   A. T2D-associated genetic variants regulate chromatin accessibility in human islets .......... 96
   B. T2D-associated SNPs reside in regulatory elements responsive to changes in the ER load of beta cells .................................................................................................................................................................................. 98
   C. Future Directions ............................................................................................................. 101
<table>
<thead>
<tr>
<th>Abbreviation</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>T2D</td>
<td>Type 2 diabetes</td>
</tr>
<tr>
<td>GWAS</td>
<td>Genome-Wide Association Studies</td>
</tr>
<tr>
<td>SNP</td>
<td>Single Nucleotide Polymorphism</td>
</tr>
<tr>
<td>RE</td>
<td>Regulatory Element</td>
</tr>
<tr>
<td>TF</td>
<td>Transcription Factor</td>
</tr>
<tr>
<td>LD</td>
<td>Linkage Disequilibrium</td>
</tr>
<tr>
<td>ER</td>
<td>Endoplasmic Reticulum</td>
</tr>
<tr>
<td>ERAD</td>
<td>ER-associated degradation</td>
</tr>
<tr>
<td>HFD</td>
<td>High Fat Diet</td>
</tr>
<tr>
<td>EMSA</td>
<td>Electrophoretic Mobility Shift Assay</td>
</tr>
<tr>
<td>QTL</td>
<td>Quantitative Trait Loci</td>
</tr>
<tr>
<td>eQTL</td>
<td>expression Quantitative Trait Loci</td>
</tr>
<tr>
<td>caQTL</td>
<td>chromatin accessibility Quantitative Trait Loci</td>
</tr>
<tr>
<td>MPRA</td>
<td>Massively Parallel Reporter Assay</td>
</tr>
</tbody>
</table>
Chapter 1

Introduction
A. Type 2 diabetes is a multifactorial disorder with both genetic and environmental risk factors

Glucose is the primary source of energy for all bodily functions. Islets of Langerhans in the pancreas play critical roles in maintaining homeostatic levels of glucose in the blood, which is very important for long-term health\(^1\). The various cell types comprising islets release hormones to maintain optimal blood glucose levels\(^1\sim3\). Alpha cells secrete glucagon in response to low blood glucose levels, which stimulates gluconeogenesis in the liver to increase levels of glucose in the blood\(^1\sim2\). If blood glucose levels drop too low, resulting hypoglycemia may impair mental functioning, and cause shaking, twitching and loss of consciousness. On the other hand, beta cells secrete insulin in response to high blood glucose levels, which stimulate glucose uptake in peripheral tissues (adipose, skeletal muscle, liver) to decrease blood glucose levels\(^1\sim3\). Elevated blood glucose levels can cause several short-term complications such as anxiety, sleepiness, headaches, sweating and numbness in fingers, toes and lips.

Type 2 diabetes (T2D) is diagnosed when blood glucose levels are persistently higher than normal, causing a significant decline in life expectancy due to long-term complications such as retinopathy, nephropathy and neuropathy\(^4\). T2D is often preceded by insulin resistance in peripheral tissues, which causes pancreatic islets to secrete higher amounts of insulin than typically required to normalize blood glucose levels. Insulin resistance, therefore, is an important predictor of future development of T2D\(^5\). However, T2D results only when pancreatic beta cells are unable to secrete adequate amounts of insulin in response to elevated blood glucose levels\(^3\sim6\).

Genome-wide association studies (GWAS) have linked single nucleotide polymorphisms (SNPs) in >250 regions of the human genome to a higher risk for developing T2D\(^7,8\). The overwhelming majority of these T2D-associated GWAS SNPs are non-coding (~90%), implicating
altered transcriptional regulation as a common molecular mechanism underlying disease risk\textsuperscript{7–10}. However, the past few decades have witnessed a rapid rise in the incidence of T2D. In 2015, 9.4\% of the USA population were diagnosed with T2D, compared to 1\% in 1958. This sharp rise in T2D incidence cannot be attributed to changes in the human genome, but our environment. Sedentary lifestyles\textsuperscript{11}, overnutrition\textsuperscript{12–14}, pollution\textsuperscript{15,16}, and depression\textsuperscript{17,18} are some of the environmental factors that have been linked to a higher risk for developing T2D, and have risen in the past few decades as well. Therefore, T2D is a multifactorial disorder with both genetic and environmental components, manifesting when environmental risk factors expose genetic susceptibilities\textsuperscript{19}. However, while islet dysfunction and failure have been clearly implicated in T2D etiology, it is not immediately apparent how environmental risk factors for T2D, such as overnutrition, affect islet function.

B. Interaction of genetic and environmental factors determine endoplasmic reticulum load in beta cells

Beta cells possess highly developed endoplasmic reticulum (ER) to synthesize approximately 1 million molecules of insulin every minute\textsuperscript{20,21}. Insulin consists of two polypeptide chains, the A- and B- chains, linked together by disulfide bonds\textsuperscript{22}. It is first synthesized as a single polypeptide called in the ER of beta cells, after which the signal sequence is removed to generate pro-insulin\textsuperscript{22}. The folding of pro-insulin into its native shape is aided by ER resident oxido-reductases, which catalyze the formation of three intramolecular disulfides\textsuperscript{20}. Demonstrating the importance of this process, pro-insulin variants lacking a cysteine needed to form one of the intramolecular disulfide bonds rapidly exhaust ER folding capacity, leading to beta cell apoptosis and diabetes in both humans and mice\textsuperscript{20,23–25}.
Despite efforts of the protein-folding machinery in the ER, a subset of pro-insulin molecules normally fails to fold properly\textsuperscript{20}. Since insulin mediates crucial signaling roles, incompletely folded forms of pro-insulin are removed for degradation either though ER-associated degradation (ERAD) or autophagy\textsuperscript{20,26}. Mice deficient for genes participating in either ERAD or autophagy in beta cells have been shown to develop hyperglycemia, highlighting the importance of these processes in maintaining beta cell identity and function\textsuperscript{20,27,28}. However, due to the high demand of insulin production and secretion in beta cells, protein-folding in the ER often exceeds capacity. During such imbalanced states of ER stress, unfolded forms of insulin molecules accumulate in the ER. These are detected by proteins, IRE1a and PERK, combinatorial signals from which up-regulate the expression of genes encoding chaperones and oxidoreductases\textsuperscript{20,21}. This increases ER protein-folding capacity in beta cells to meet increased demands for insulin secretion. Low levels of ER stress can also be a signal for beta cells to proliferate as a mechanism to meet higher insulin secretion demands\textsuperscript{29}. However, if ER stress persists despite attempts at adaptation, apoptosis is initiated, leading to a decline in beta cell mass and diabetes\textsuperscript{20}.

The relevance of ER capacity and stress in beta cells for the future development of T2D is highlighted by genetic ablation studies, in which mice deficient for genes encoding IRE1a or PERK in beta cells develop hyperglycemia due to defective insulin production\textsuperscript{20,30,31}. However, beta cells can also encounter environmental challenges during which protein-folding demand in the ER exceeds capacity. Modelling the rise of T2D incidence in humans due to overnutrition, mice fed a high fat diet (HFD) have an increased risk for developing T2D due to higher levels of ER stress in beta cells\textsuperscript{32}. Importantly, increasing ER folding capacity in beta cells protects against HFD-induced diabetes\textsuperscript{33}. On the other hand, decreasing ER folding capacity in beta cells exacerbates HFD-induced diabetes\textsuperscript{34}. Therefore, an interaction of both genetic and environmental factors determines ER load in beta cells, which in turn governs risk for developing T2D in the future.
C. Approaches to study T2D-associated genetic and environmental variations altering transcriptional activation in beta cells

A host of diverse cell-types comprise a multicellular organism. Distinctions between them do not arise due to genetic variation, but due to differences in the manner in which the genome is transcribed. A host of proteins, which are themselves encoded in the genome, regulate transcription. Transcription factors (TFs) bind to non-coding regulatory elements, and recruit chromatin remodelers, histone modifiers and RNA polymerases, to regulate target gene expression\(^3\). Changes to the cellular environment, such as the presence of a growth factor or ER stress, are signaled to TFs so that transcriptional programs can be enacted for cells to respond appropriately, giving rise to the diversity of cell-types observed in multicellular organisms\(^3,36,37\). SNPs at non-coding regulatory elements in the human population modulate transcriptional programs to alter the response of cells to changes in the cellular environment\(^3,38,39\), such as ER stress. Genetic differences across and environmental variation within individuals, therefore, both can affect the spatial and temporal patterns of gene expression.

GWA studies have linked SNPs in >250 regions of the human genome to a higher risk for developing T2D\(^7,8\). In order to elucidate the mechanistic link between non-coding SNPs and a complex trait like T2D, it is important to know:

1. T2D-associated GWAS SNPs are active in which cell-types or tissues in the body?
   
   Genome-wide mapping of regulatory elements, such as promoters, enhancers, and insulators in >150 human cell types and tissues, have revealed T2D-associated GWAS SNPs to be significantly and specifically enriched in islet-
specific regulatory elements, suggesting alterations in islet regulatory element activity underlie the molecular genetics of T2D \cite{9,10,40}.

2. What are the genes whose expression in islets is altered by T2D-associated SNPs? Identifying target genes is important to know why small changes in their expression lead to a higher risk for T2D. Expression quantitative trait loci (eQTL) approaches have successfully linked SNPs to variation in target gene expression across islet samples \cite{40}. However, despite well-powered large-scale eQTL studies, SNPs at only 23 T2D-associated GWAS signals have been linked to effector transcripts in islets \cite{40,41}. Since such eQTL studies have been restricted so far to islets cultured under steady state conditions, we hypothesize that important gene-environment interactions, characteristic of a complex disorder like T2D, were missed.

3. Under what cellular conditions are T2D-associated SNPs active in islets? Identifying the cellular environment in which DNA regulatory elements are active in beta cells is key to understanding how SNPs in them modulate risk for diseases such as T2D.
At a mock T2D-associated GWAS signal, there are 8 SNPs in LD. This thesis takes two complementary approaches to identify functional SNPs among them.

A. The Islet caQTL approach identifies significant correlations between genotypes at SNPs and in vivo islet chromatin accessibility at overlapping ATAC-seq peaks. While islet caQTL approaches are useful in narrowing down T2D-associated SNPs affecting islet RE activity, multiple SNPs are often located within a given open chromatin region, making it difficult to identify functional SNPs using caQTL approaches alone.

B. The Massively Parallel Reporter Assay (MPRA) can test allelic effects of thousands of SNPs under multiple experimental conditions. A mock SNP, rs12345, is used as an example to show how activity is measured. Briefly, 200 bp test sequences are cloned upstream of a minimal promoter driving expression of associated 20 bp barcodes in the 3’ UTR of GFP. The MPRA plasmid pool was transfected into beta cells, and harvested thirty hours later for RNA isolation, GFP mRNA capture, and Illumina sequencing of the regulatory sequence associated barcodes.
At any given T2D-associated GWAS, there are 10s-100s of SNPs that are in linkage disequilibrium (LD) with each other, not all of which are functional. In order to elucidate the mechanistic link between non-coding SNPs and a complex trait like T2D, it is therefore crucial to first identify functional SNPs among them. Under what cellular conditions do they become active? Are there multiple, functional SNPs among them, or just one? This thesis takes two complementary approaches to answer these questions:

1. In Chapter 2, we employed a chromatin accessibility quantitative trait loci (caQTL) approach (see Introductory Figure 1) to link SNPs at 13 T2D-associated GWAS signals to changes in islet chromatin accessibility\(^42\). However, QTL approaches identify correlative associations, and experimental testing of variants is required to establish causality. Moreover, while islet caQTL approaches are useful in narrowing down T2D-associated SNPs affecting islet RE activity, multiple SNPs are often located within a given open chromatin region, making it difficult to identify functional SNPs using caQTL approaches alone.

2. In Chapter 3, to overcome these limitations we used a massively parallel reporter assay\(^43\) (MPRA; see Introductory Figure 1) to test allelic effects of >6,500 SNPs under multiple experimental conditions. By testing all SNPs in LD (\(r^2>0.8\)) with 259 T2D association signals, we identified 220 functional SNPs in 104 T2D-associated GWAS signals. For 54 T2D-associated signals, only one functional SNP was identified, implicating them as the putative causal SNP within their respective locus. More than 1 functional SNPs were detected at 50 T2D association signals, suggesting T2D risk from individual GWAS signals may result from a combined effect of multiple functional SNPs.
Finally, since caQTL approaches require islets from ≥100 individuals, wherein power is still highly dependent on minor allele frequency, we used MPRA to also show that ER stress is an important cellular context under which regulatory elements harboring T2D-associated SNPs become active (see Introductory Figure 2). A significant fraction of functional T2D-associated SNPs resided in regulatory elements responsive to ER stress, clearly linking ER stress in beta cells to the genetic risk of common T2D.

References:


Chapter 2

Type 2 Diabetes–Associated Genetic Variants Regulate Chromatin Accessibility in Human Islets.

Abstract:
Genetic and environmental factors both contribute to islet dysfunction and failure, resulting in type 2 diabetes (T2D). The islet epigenome integrates these cues and can be remodeled by genetic and environmental variation. However, our knowledge of how genetic variants and T2D disease state alter human islet chromatin landscape and cis-regulatory element (RE) use is lacking. To fill this gap, we profiled and analyzed human islet chromatin accessibility maps from 19 genotyped individuals (5 with T2D) using ATAC-seq technology. Chromatin accessibility quantitative trait locus (caQTL) analyses identified 3001 sequence variants (FDR<10%) altering putative cis-RE use/activity. Islet caQTL were significantly and specifically enriched in islet stretch enhancers and islet-specific transcription factor binding motifs, such as FOXA2, NKX6.1, RFX5/6 and PDX1. Importantly, these analyses identified putative functional single nucleotide variants (SNVs) in 13 T2D-associated GWAS loci, including those previously associated with altered ZMIZ1, MTNR1B, RNF6, and ADCY5 islet expression, and linked the risk alleles to increased (n=8) or decreased (n=5) islet chromatin accessibility. Luciferase reporter assays confirmed allelic differences in cis-RE activity for 5/9 caQTL sequences tested, including a T2D-associated SNV in the IL20RA locus. Comparison of T2D and non-diabetic islets revealed 1882 open chromatin sites exhibiting T2D-associated chromatin accessibility changes (FDR<10%). Together, this study provides new insights into genetic variant and T2D disease state effects on islet cis-RE use and serves as an important resource to identify putative functional variants in T2D- and islet dysfunction-associated GWAS loci and link their risk allele to in vivo loss or gain of chromatin accessibility.
Introduction

Pancreatic islet function is central to maintaining glucose homeostasis. T2D is a complex disease resulting from the combined effects of genetic susceptibility and environmental exposures. Genome-wide association studies (GWAS) have associated single nucleotide variants (SNVs) in >100 loci with increased susceptibility to type 2 diabetes (T2D) and related quantitative measures of islet dysfunction (Fuchsberger et al. 2016; Mohlke and Boehnke 2015). The majority of these variants overlap islet-specific enhancer elements (Parker et al. 2013; Fuchsberger et al. 2016; Pasquali et al. 2014). This enrichment establishes perturbed islet transcriptional regulation in the genetic etiology of islet dysfunction and T2D (Lawlor et al. 2017). In addition to individual genetic variation, environmental insults to islet functions such as oxidative, endoplasmic reticulum (ER), and inflammatory stresses, have also been linked to T2D.

Genetic and environmental factors shape the epigenome to modulate the transcriptional programs governing steady state and stress responsive-factors. Common genetic variants in the human population, contributing to complex phenotypes and disease susceptibility, have been linked to alterations in regulatory element use, as monitored by changes in chromatin accessibility (Degner et al. 2012; Pique-Regi et al. 2011; Kumasaka et al. 2016; McDaniell et al. 2010; Alasoo et al. 2017) and histone modifications (McVicker et al. 2013; Ng et al. 2017) in diverse cell types. Moreover, cis-regulatory element use in a given cell type can be modified by its local environment (Lavin et al. 2014) and cellular responses to stimuli (Ostuni et al. 2013) and stressors (Brown et al. 2014). Currently, our understanding of how individual genetic variation and the type 2 diabetic disease state alter cis-regulatory element use in human pancreatic islets is limited.

In this study, we sought to understand how (1) genetic variants, particularly those associated with T2D susceptibility and quantitative measures of islet dysfunction; and (2)
T2D state alter chromatin accessibility and cis-regulatory element use in human islets. Using the assay for transposase-accessible chromatin with sequencing (ATAC-seq), we studied the chromatin accessibility patterns in human islets obtained from 19 individuals, five of whom were type 2 diabetic. Using RASQUAL (Kumasaka et al. 2016), we integrated genotypes generated in each individual with their corresponding open chromatin profiles to identify chromatin accessibility quantitative trait loci (caQTLs), i.e., genetic variants that alter chromatin accessibility and cis-RE use in islets. Finally, by comparing ATAC-seq profiles between diabetic and normal donors, we identified changes in chromatin accessibility and putative cis-RE use associated with the T2D disease state.

Results

Chromatin accessibility maps in human pancreatic islets

To determine the genome-wide location of cis-regulatory elements in human islets, we transposed the nuclei of islet samples obtained from 23 cadaveric organ donors (Table 1; 18 non-diabetic (ND) and 5 T2D) and measured chromatin accessibility using ATAC-seq. 19/23 (n=14 ND, 5 T2D) donor islet ATAC-seq datasets passed quality control filters (Methods) and were used in subsequent analyses (Figures 1A, S1A; Supplemental Table 1). As shown in Figure 1B, the genome-wide chromatin accessibility profiles of these islets were highly correlated and all islet profiles clustered separately from other cell types, including T2D-relevant skeletal muscle and adipose tissues. Notably, T2D donor islet profiles (n=5) did not cluster distinctly from those of ND donor islets (n=14), suggesting that the T2D disease state itself does not lead to global restructuring of islet chromatin accessibility. ATAC-seq profiles of representative islets for the GCK locus are shown in Figure 1C, revealing both common (gray) and islet-specific (orange) ATAC-seq peaks.
A. Study design

B. ATAC-seq libraries

C. Examples of ATAC-seq peaks in GCK locus

D. Functional annotations of islet ATAC-seq peaks

E. Stretch Enhancer enrichment

F. TF motif enrichment -log(p-value)
Collectively, 154,438 ATAC-seq peaks were identified across the 19 islet donors (see Methods), representing putative cis-REs (i.e., promoter, enhancer, repressor, and insulator elements). 40% of these putative islet REs were detected in only 1 out of 19 individuals in the cohort (Supplementary Figure S1B). Not surprisingly, 45% of these individual-specific peaks were un-annotated in reference islet ChromHMM states (i.e., low signal state) (Supplementary Figure S1C). In contrast, ATAC-seq peaks at gene promoters were consistently accessible across the cohort (Supplementary Figure S1C), suggesting that promoter elements are less variable across individuals compared to other cis-REs.

Islet ATAC-seq peaks were compared against previously reported (Varshney et al. 2017) ChromHMM-defined functional states in islets from our previous studies (Stitzel et al. 2010; Parker et al. 2013) and 30 other tissues from the NIH Epigenome Roadmap project (Roadmap Epigenomics Consortium et al. 2015), including adipose, skeletal muscle, liver, and brain. As anticipated, islet ATAC-seq peaks were most enriched in ChromHMM-defined islet enhancers (Figure 1D). Similarly, we compared the islet ATAC-seq peaks with stretch enhancers (SEs) in 31 tissues. SEs are long (>3kb) stretches of contiguous cell-specific enhancer chromatin states that are enriched for disease-associated SNVs relevant to the cognate cell type (Parker et al. 2013). 90% of islet SEs overlapped islet ATAC-seq peaks (Figure 1E); this overlap was significantly greater.

Figure 1. Chromatin accessibility maps in human islets. (A) Schematic of study design. (B) Spearman correlation between genome-wide ATAC-seq read distributions of islets and other cell types. Islets from type 2 diabetic donors are marked with asterisk. (PBMC->peripheral blood mononuclear cells). (C) An example locus in and around the GCK gene representing chromatin accessibility landscapes in 3 ND and 3 T2D islets, and other tissues. The regions marked in orange are specifically accessible in islet cells, whereas regions marked in gray are ubiquitously accessible. All chromatin accessibility maps are normalized to the same depth and have the same scale. (D) Overlap of islet ATAC-seq peaks with chromatin states in islets and 30 other cell types. Tissues are sorted from highest to lowest overlap between ATAC-seq peaks and enhancer states. (TSS = Transcription Start Site). (E) Percent of tissue-specific stretch enhancers (SEs) overlapping islet ATAC-seq peaks. Fisher’s exact test p-values are shown to represent enrichment. (F) Enrichment of transcription factor (TF) motifs in islet-specific ATAC-seq peaks. TFs are clustered with respect to the similarity of their position weight matrices (PWMs).
(Fisher’s Exact p<2.2e-16 for islet SEs) than that observed for other tissue SEs. Moreover, islet-specific peaks (i.e., those that do not overlap open chromatin sites in Skeletal Muscle, Adipose, GM12878, CD4+ T cells, PBMCs), were enriched in motifs of islet cell-specific transcription factors, such as PDX1 and NKX6.1, when compared to ATAC-seq peaks that were common across tissues (Figure 1F). These data thus represent high quality chromatin accessibility maps of human islets and captures islet-specific regulatory elements.

**Identification of genetic variants affecting islet chromatin accessibility**

Since genetic variation in cis-RE use/activity has been implicated in diverse phenotypes and complex diseases, including T2D, we sought to identify genetic variants that alter chromatin accessibility in human islets (Figure 2A). Chromatin accessibility QTL (caQTL) analysis (Kumasaka et al. 2016) was used to identify the genetic variants within each ATAC-seq peak that correlated with changes in its accessibility (see Methods). In total, we uncovered 3001 SNVs significantly associated (FDR<0.10; Fig S2A) with chromatin accessibility changes in this cohort. Figure 2B highlights an example caQTL overlapping an intronic islet SE in the CELF4 gene, which exhibits islet-selective expression (Varshney et al. 2017). Islet chromatin accessibility was reduced in rs488797 CC homozygotes, potentially by disrupting a FOXA2 binding motif (Ward and Kellis 2016). In agreement, almost all ATAC-seq reads overlapping this variant in CT heterozygous islets contain the T allele (Figure 2B, inset).
Figure 2. Identification of genetic variants affecting islet chromatin accessibility. (A) Schematic of chromatin accessibility quantitative trait locus (caQTL) analyses linking genotype to chromatin accessibility changes. (B) An example caQTL in an islet-SE within an intron of CELF4. Aggregate ATAC-seq profiles of individuals with C/C (red), C/T (blue), and T/T (black) genotypes at rs488797 are displayed. The inset boxplot shows the fraction of ATAC-seq reads containing the T allele at rs488797 in each C/T heterozygous islet (n=11). Sequence logo for FOXA2 TF is displayed, which is predicted by HaploReg to be altered by this variant. (C) Distance between caQTLs and the transcription start site (TSS) of the nearest expressed gene versus the significance of the caQTL association. The majority of caQTLs are within 200 kb of the TSS of the nearest expressed gene. (D) Percent of tissue-specific SEs overlapping caQTLs. Fisher’s exact test p-values are shown for enrichment. (E) TF motifs enriched in caQTLs. TFs are clustered based on their PWM similarity using hierarchical clustering, resulting in four major TF groups. Bar plots of p-values are color coded according to this clustering. A representative PWM logo is represented for each cluster, where the corresponding TF is marked with an asterisk.
The majority of islet caQTLs (97%) were within 200 kilobases (kb) of the transcription start site (TSS) of an islet-expressed gene. (Figure 2C). Approximately 30% of islet caQTLs overlapped islet enhancer chromatin states (Supplementary Figure S2C). When compared to tissue SEs, the islet caQTLs were specifically enriched in islet SEs only (Figure 2E). Peaks containing caQTLs were also significantly enriched in islet SEs when compared to all islet ATAC-seq peaks (p=0.0024; OR=1.18; Fisher’s exact test), suggesting that islet caQTLs alter cis-REs encoding important islet-specific functions. Consistently, sequence motifs for islet-specific TFs, such as NKX6.1, PDX1, and MAFA, and not for general TFs, were enriched in caQTL-containing ATAC-seq peaks (Figure 2F). Surprisingly, sequence motifs of oxidative stress-responsive TFs (Ma 2013; Dhakshinamoorthy et al. 2005), such as BACH1, BACH2, and NRF2, were also enriched in caQTL peaks. Together, these data and analyses enumerate sequence variants that alter chromatin accessibility of islet cis-REs and suggest that these changes may be associated with altered binding of TFs governing islet cell identity, function, and stress response.

**T2D-associated GWAS SNVs alter chromatin accessibility in islets**

The large majority (>90%) of common variants associated with T2D and quantitative measures of islet dysfunction, such as fasting plasma glucose and insulin levels, reside in non-coding loci and significantly and specifically overlap islet SEs. However, in vivo effects of these GWAS SNV alleles on cis-regulatory element use, as assessed by chromatin accessibility, islet TF ChIP, or active enhancer histone modifications such as H3K27ac, have been assessed and reported at only a handful of loci (Gaulton et al. 2010; Roman et al. 2017) to date. We hypothesized that T2D- and islet (dys)function-associated GWAS SNVs alter chromatin accessibility in islets, and exhibit significant and specific overlap with islet caQTLs. To test this, we assessed the
enrichment (Schmidt et al. 2015) of islet caQTLs for SNVs exhibiting genome-wide significant ($p<5\times10^{-8}$) associations with 184 diverse traits and diseases retrieved from the NHGRI/EBI GWAS Catalog (Methods). Among all GWAS traits and diseases assessed, islet caQTLs only exhibited significant enrichment of GWAS SNVs associated with T2D (2.97-fold), fasting glucose (13.46-fold), and BMI-adjusted fasting glucose (7.43-fold) (Figure 3A, $p<5.43\times10^{-4}$, < 0.10 after Bonferroni correction).

These analyses highlighted 13 T2D-associated variants overlapping islet caQTLs (Figure 3B). They included 4 loci ($ADCY5$, $ZMIZ1$, $MTNR1B$, and $RNF6$) in which the caQTL SNV has been previously linked to altered *in vitro* enhancer activity or *in vivo* steady state gene expression in islets (van de Bunt et al. 2015; Roman et al. 2017; Lyssenko et al. 2009, 1; Varshney et al. 2017; Fadista et al. 2014). Importantly, for all loci that harbor both islet caQTL and eQTL variants (i.e., $ADCY5$, $ZMIZ1$, $MTNR1B$, and $RNF6$), the risk alleles have a concordant effect both on chromatin accessibility and gene expression levels (Figure 3B).
Figure 3. T2D associated caQTLs. (A) Enrichment of caQTLs for disease associated GWAS SNVs. The dashed red line represents $p=5.43 \times 10^{-4}$, equivalent to 10% cutoff after Bonferroni correction for the number of diseases tested (n=184). (B) Table enumerating the 13 caQTLs linked to T2D-associated GWAS SNVs. Asterisks mark those loci that were tested for differential luciferase activity in panel E. (C) Average chromatin accessibility profiles at the ADCY5 locus (loss-of-function T2D caQTL). The inset boxplot shows the fraction of ATAC-seq reads containing the G allele in each of the heterozygous islet samples (n=5). (D) Average chromatin accessibility profiles at the IL20RA locus (gain-of-function T2D caQTL). The inset boxplot shows the fraction of ATAC-seq reads containing the C allele in each of the heterozygous islet samples (n=11). (E) Luciferase activity of 9 tested caQTLs with reference and alternate alleles (normalized to empty construct) **** and *** indicate $p<0.0001$ and $p<0.001$, respectively; two-sided Mann-Whitney test $p$-values are shown on boxplots. ns = not significant.
For 5 out of these 13 GWAS loci, the risk allele decreased chromatin accessibility (Figure 3B). This includes the T2D-associated index SNV rs11708067, which resides in the third intron of the ADCY5 gene and overlaps an islet SE. The risk allele for this variant (A) is associated with reduced chromatin accessibility (Figures 3B, 3C). This is consistent with recent reports by us and others linking the rs11708067 risk allele (A) to decreased enhancer activity in luciferase reporter assays in vitro (Roman et al. 2017), to reduced histone H3 lysine 27 acetylation (H3K27ac) (Roman et al. 2017) and to decreased ADCY5 expression in human islets in vivo (van de Bunt et al. 2015; Varshney et al. 2017; Roman et al. 2017). In the remaining 8 T2D-associated caQTL loci (Figures 3B, D), the T2D risk allele was associated with higher chromatin accessibility than the non-risk allele, suggesting that the risk allele is associated with a gain-of-function.

To validate a subset (n=9) of the islet caQTLs, we tested whether the human caQTL alleles altered enhancer activity of the sequences overlapping these putative cis-REs using luciferase reporter assays in MIN6 beta cells. Comparison of sequences containing either the reference or alternate allele for each caQTL site (Table 4) confirmed differential enhancer activity for 5 out of 9 loci (Figure 3E). For example, the rs6937795 “A” allele in the IL20RA locus, which is associated with increased T2D susceptibility and increased islet chromatin accessibility (Figure 3D), showed 2.5-fold higher enhancer activity than the non-risk “C” allele (Figures 3B, 3D, 3E).

**Chromatin accessibility changes in T2D versus ND islets**

To uncover T2D associated changes in chromatin accessibility, we compared chromatin accessibility maps from 5 ND and 5 T2D donors (Figure 4A). Out of 52,387 ATAC-seq peaks tested, 1882 differentially accessible peaks between T2D and ND islets were identified (FDR 10%). Of these, 980 showed an increase and 902 showed a decrease in accessibility with the T2D state, hereafter referred to as “opening” and
“closing” peaks, respectively (Figure 4B). There was a remarkable difference in the functional annotation of differential peaks, where closing peaks were mostly found at enhancers (48%), and opening peaks were mainly at promoters (70%) (Figure 4C). Figure 4D and 4E represents examples of closing and opening peaks respectively.

However, when differential peaks were categorized with respect to the presence or absence of ATAC-seq peaks in normal and diabetic donors (Fig S3E), we found that the majority of T2D-associated changes in chromatin accessibility were gradient in nature, i.e., peaks do not completely appear/disappear in T2D islets, with a few exceptions (<1%) (Supplementary Figure S3E). Additionally, we note that a subset of the 1882 differentially accessible peaks (42 opening peaks, 51 closing peaks) overlapped caQTLs (Supplementary Figure S3F), suggesting that these T2D associated accessibility changes are driven by genetic factors.

The differential peaks were annotated to the closest active genes in islets (Methods), resulting in 898 genes associated with opening and 665 genes associated with closing peaks (Supplemental Table S5). Differential gene expression analyses between ND and T2D samples revealed small changes in gene expression levels, where only 90 (38) genes were significantly up (down) regulated in T2D islets (FDR 10%). We observed a modest yet positive correlation between T2D-associated chromatin accessibility changes at gene promoters and the changes in the expression levels of these genes (p=0.038, Wilcoxon, Figure 4F). Opening and closing peaks were enriched in different TF motifs (Figure 4G). Interestingly, TFs that regulate stress responses such as ATF3, AP-1 were enriched in closing peaks.
Using ATAC-seq, we profiled the chromatin accessibility of human islets from 19 individuals. Integration of these open chromatin maps with each individual’s genotypes identified 3001 sequence variants (caQTL) that modulate in vivo islet regulatory element
These caQTLs were enriched in islet-specific enhancers and TF motifs. Importantly, a subset of these was significantly and specifically enriched for T2D and fasting glucose GWAS index or linked ($r^2>0.8$) sequence variants. Comparison of ATAC-seq profiles from T2D and ND individuals revealed quantitative changes in chromatin accessibility at 1882 putative islet cis-REs. Together, these data and analyses contribute significantly to (1) enumerating the genetic variants that alter islet cis-RE use; (2) delineating putative functional variants among the T2D- and islet dysfunction-associated GWAS SNVs; (3) linking the risk allele to in vivo loss or gain of cis-RE use in islets; and (4) assessing the relative chromatin accessibility effects of genetic variation and T2D state on human pancreatic islets.

ATAC-seq profiling in islets obtained from multiple cadaveric organ donors identified genetic effects on accessibility of 3.5% (3001/84499) of putative islet cis-REs genome-wide, and linked the alternate allele to increased chromatin accessibility at 43.5% (1307/3001) of sites and decreased accessibility in 56.5% (n=1694/3001) of sites compared to the reference allele. 6.2% (187/3001) of caQTLs variants identified in this study were in linkage disequilibrium ($r^2>0.8$) with previously described islet eQTL (Varshney et al. 2017). Reports studying other cell types from larger cohorts have observed overlaps between chromatin-based QTLs (such as DNase-sensitivity and histone acetylation) and eQTLs ranging from 16%-45% (Degner et al. 2012; Li et al. 2016; del Rosario et al. 2015; Ng et al. 2017). Lower overlap observed in our study could be explained, at least in part, by the differences in sample size and resulting disparities in power between these two studies. It may also reflect the effect of the mixed cellular composition of islets, which might be resolved by studies measuring these features in sorted cell types. Finally, this overlap may reflect different inherent features measured by RNA-seq and chromatin-based assays that may contribute to these modest overlaps. For example, features reflected in RNA-seq data such as mRNA stability,
polyadenylation, and splicing are not captured by chromatin profiling assays. This warrants future studies examining the impact of genetic variation on the ability of islets to respond to environmental changes (response QTLs), alongside islet caQTL studies with higher sample sizes, perhaps in sorted cells.

Using luciferase assays, we assessed allelic effects of the sequences overlapping nine of the islet caQTL sites on enhancer activity. Only 3/9 sequences tested exhibited transcriptional enhancer activity compared to the minimal promoter sequence alone, reinforcing the concept that caQTLs capture both enhancer and repressor cis-REs (Petrykowska et al. 2008). Importantly, five of these sites exhibited significant allelic differences in cis-RE activity. In each case, the direction of allelic effect on enhancer activity matched the allelic changes in chromatin accessibility, including for the rs6937795 variant in the T2D-associated IL20RA locus. For the remaining four sequences, lack of allelic differences in in vitro enhancer activity may be due to the human enhancers not being active in the mouse MIN6 beta cell line used for luciferase assays or to REs displaying enhancer activity only under certain conditions, such as oxidative stress and not in baseline conditions. Indeed, studies in other cell types suggest that regulatory elements can be primed for and activated by specific environmental stimuli or stressors (Ostuni et al. 2013; Alasoo et al. 2017; Brown et al. 2014).

In this study, we identified SNVs in 13 T2D-associated loci that alter chromatin accessibility. These include four loci (ZMIZ1, MTNR1B, RNF6, and ADCY5) in which the same or linked (r2>0.8) genetic variant has been identified as an islet eQTL. Importantly, the caQTL and eQTL studies identified a consistent direction-of-effect (e.g., gain- or loss-of-function) for the risk allele in each of these loci. T2D risk alleles in 5/13 loci were associated with reduced chromatin accessibility. For the remaining loci, the risk alleles were associated with increased chromatin accessibility, representing potential gain-of-
function variants. Unfortunately, sequence motif analysis of these caQTL variants did not reveal obvious trans-factors that may be responsible for these accessibility differences or whose binding is affected by this sequence variant. This could be due in part to incomplete information on position weight matrices for TFs, including ARX, which is an islet alpha cell transcription factor. Together, these data and analyses have identified novel SNV effects on islet cis-REs, including their direction-of-effect, that can be further dissected in a site-specific and hypothesis-driven manner.

By comparing ATAC-seq profiles from T2D and ND donors, we identified 980 and 902 regulatory elements that exhibit quantitative T2D-associated increases or decreases in chromatin accessibility, respectively. These data suggest that T2D state by itself may not lead to widespread changes in chromatin accessibility. However, we acknowledge that T2D-associated epigenomic changes may be masked by multiple factors, including: 1) the small and genetically heterogeneous islet cohort analyzed; (2) cell type-specific changes that are hidden by other islet constituent cells; and (3) steady-state, normoglycemic culture conditions that may mask changes elicited by the diabetic milieu. Moving forward, studies that account for these potential confounders in larger, genetically-stratified islet cohorts will be necessary to further confirm these T2D-associated changes and identify novel ones.

Methods:

Study subjects and primary islet culture:

Fresh human cadaveric pancreatic islets were procured from ProdoLabs or the Integrated Islet Distribution Program (IIDP). Upon arrival, cells were transferred into PIM(S) media (ProdoLabs) supplemented with PIM(ABS) (ProdoLabs) and PIM(G) (ProdoLabs) and kept in a T-150 non-tissue culture treated flask (VWR) for recovery at 37 C and 5% CO₂ overnight. ATAC-seq and RNA-seq were performed the following day
as described below. For genotyping, genomic DNA was collected from islets cultured in CMRL + 10% FBS + Pen/Strep + Glutamax (Life Technologies) on tissue treated T175 until confluent and then prepped with the Qiagen Blood and tissue kit.

**Islet genotyping and imputation:**

Islets were genotyped using the Illumina Omni2.5Exome (n=11) or the Omni5Exome chips (n=8) (See Table S1). Genotype calls were made using the Genome Studio software (Illumina). The resulting vcf files were merged using the vcf-merge command in the vcftools/0.1.12a suite, and subsequently filtered for sites with any missing data (--max-missing 1). 2.38 million genotyped SNVs passed QC and were used for imputation (1000G Phase 3 v5) (1000 Genomes Project Consortium et al. 2015) and phasing (Eagle v2.3) (Loh et al. 2016) using the Michigan Imputation Server (Das et al. 2016), to get a total of 47 million SNVs. After removing SNVs that were either monomorphic or outside islet ATAC-seq peaks, 1.21 million SNVs were kept for downstream analysis (caQTL/eQTL).

**Chromatin accessibility analysis (ATAC-seq):**

Human islet ATAC-seq libraries were prepared as described (Varshney et al. 2017). Approximately 50-100 islet equivalents (50,000-100,000 cells) per sample were transposed in triplicate. Libraries were sequenced on an Illumina NextSeq500 (see Table S2). Paired-end 75-bp ATAC-seq reads were trimmed to remove low quality base calls using trimmomatic, and aligned to the hg19 human genome assembly with the Burrows Wheeler Aligner-MEM (Li and Durbin 2009). For each sample, duplicates were removed and the residual reads were shifted as previously described (Ucar et al. 2017). For each sample, technical replicates were merged using samtools, and peaks were called from the resulting merged bam file for each individual using MACS2 (Zhang et al. 2008) (with parameters -callpeak --nomodel -f BAMPE). Islets with less than 30,000 peak calls were removed, resulting in 19 islets for downstream analyses. An average
sequencing depth of 62.6 million (SD=18.6 million) reads was obtained for each of the remaining 19 islets, after merging the 3 technical replicates. ATAC-seq peaks on sex chromosomes and those overlapping regions with low mappability (http://hgdownload.cse.ucsc.edu/goldenpath/hg19/encodeDCC/wgEncodeMapability/) were removed. The remaining autosomal peaks with q-values < 0.01 were selected for downstream analysis. The R Diffbind package (Stark and Brown) was used to define 154,437 ATAC-seq peaks for the 19 islets and to obtain read counts for each ATAC-seq peak for all the samples.

**Islet chromatin accessibility quantitative trait locus (caQTL) analyses:** VerifyBamID (Jun et al. 2012) was used to match ATAC-seq bam files for each sample to each individual’s genotypes and ensure no samples were switched. We removed 69,939/154,438 islet ATAC-seq peaks containing monomorphic SNVs from the analyses. For the 1.21 million SNVs that were non-monomorphic and found within the remaining 84,499 islet peaks, allele-specific counts were obtained. Along with the read count information for the islet peaks for each sample local caQTLs were mapped using the RASQUAL statistical approach. The first 5 principal components were used as covariates to minimize confounding factors. The Bonferroni method was used to correct for the number of SNVs tested for each ATAC-seq peak. 10 random permutations were generated for each feature, and used to correct for the number of features tested, with an FDR cutoff of 10%.

**Differential ATAC-seq peak analyses (T2D vs. ND):**

Islets from 5 T2D individuals and from 5 ND individuals with the best demographic match (e.g., age, sex and race) were selected for comparative/differential analysis (see Table 1). The R Diffbind package (Stark and Brown) was used to define 117599 consensus ATAC-seq peaks among these 10 islets and to determine read counts in each ATAC-seq peak for each of the 10 samples. Peaks were excluded from differential analysis if they
met the following criteria: 1) the peak is present in fewer than three islet donors; 2) it is present in only 1 T2D and 2 ND (or 1ND and 2 T2D) islet donor. Surrogate Variable Analysis (SVA) (Leek et al. 2012) was used to summarize sources of unwanted variability in the read count table for the remaining 52,387 consensus peaks. The two significant surrogate variables were used as covariates in the design matrix to minimize confounding factors. The edgeR package (Robinson et al. 2010) was used to identify 1882 differential peaks at FDR 10%.

**Enrichment of genome-wide association study (GWAS) SNVs in differential (T2D vs. ND) and caQTL open chromatin sites:**

Lists of reference SNV identifiers were obtained from the NHGRI-EBI Catalog of GWAS SNVs (https://www.ebi.ac.uk/gwas/; accessed on January 19th, 2017) for 642 disease categories. For each disease category, GWAS SNVs were pruned using PLINK (Purcell et al. 2007) version 1.9 and parameters “--maf 0.05 --clump --clump-p1 0.0001 --clump-p2 0.01 --clump-r2 0.2 --clump-kb 1000” to ensure that each variant haplotype was tested only once during the enrichment analysis. For each SNV pair in linkage disequilibrium (LD) ($R^2 > 0.2$) the SNV with the less significant p-value was discarded. GREGOR (Schmidt et al. 2015) was used to determine if the LD-pruned GWAS SNVs were enriched ($r^2 > 0.8$) in (1) differential or (2) caQTL ATAC-seq peaks. Diseases for which there weren’t any GWAS SNVs in LD ($r^2 > 0.8$) with the tested genomic regions were excluded from downstream analysis.

**Transcription factor (TF) motif enrichment:**

The findMotifsGenome.pl (with parameters hg19 and -size given) script in the Homer suite (Heinz et al. 2010) was used to identify significantly enriched transcription factor (TF) motifs in islet ATAC-seq data. For Figure 1F, motifs enriched in the islet-specific ATAC-seq peaks were identified using the common ATAC-seq peaks as the background set. Common (background) peaks were defined as those that overlapped any given
ATAC-seq peaks for adipose, CD4T, GM12878 and 2 PBMC samples. Islet-specific peaks were those that did not overlap an open chromatin loci in the other tissues. For Figure 4G, motifs enriched in the 1882 differential peaks were identified using the 52387 peaks as the background set. For Figure 2E, motifs enriched in the 3001 caQTL peaks were identified using the 154437 peaks as the background set. TFs are clustered based on the similarity of their Position Weight Matrices (PWMs) using Kullback Leibler divergence method as implemented in the TFBSTools R package (Tan and Lenhard 2016).

ChromHMM annotation:
Harmonized ChromHMM files (13 state) for islets, the ENCODE cell lines and the Roadmap tissues were used as previously determined (Varshney et al. 2017). The ggplot2 package was used to plot the overlap of peak sets to the harmonized ChromHMM states. For cases when a peak overlapped two or more ChromHMM states, the order of preference for overlaps were as follows: Active TSS, Bivalent TSS, Weak TSS, Flanking TSS, Active Enhancer-1, Active Enhancer-2, Weak Enhancer, Genic Enhancer, Strong Transcription, Weak Transcription, Repressed Polycomb, Weak Repressed Polycomb, and Quiescent.

Stretch enhancer annotation:
Stretch enhancers were defined using the harmonized ChromHMM definitions. Briefly, stretch enhancers are defined as > 3kb consecutive segments that overlap and enhancer state including Active Enhancer 1 and 2, Weak Enhancer and Genic Enhancer ChromHMM states. To test whether a peak set is enriched in a given tissue stretch enhancers, fisher’s exact test was performed. The background set was the union of the stretch enhancers of all 31 tissues, except the one being tested.

RNA-seq profiling:
Libraries for the 19 islets exhibiting high-quality ATAC-seq profiles were prepared using the stranded TruSeq kit (Illumina), and had either ERCC Mix 1 or Mix 2 randomly spiked-in (ThermoFisher, catalog #4456740; see Supplemental Table S3). The 10 islets used for the T2D vs. ND differential analysis were sequenced on an Illumina NextSeq500 sequencer. The remaining 9 islets were sequenced separately on Illumina HiSeq 2500. The paired-end RNA-seq reads for each islet was trimmed for low quality base calls using trimmomatic (Bolger et al. 2014). Bowtie2 (Langmead and Salzberg 2012), in conjunction with RSEM (Li and Dewey 2011) (rsem-calculate-expression), was used to obtain the FPKM and Expected read counts for all genes across the 19 samples. An average depth of 87.2 ± 27.8 million reads was obtained for the 19 islets.

**Differential gene expression:**

Only autosomal genes with FPKM>5 in more than 3 islets were included in the analysis. SVA was used to summarize sources of unwanted variability in the expected read count matrix for the remaining 10,116 genes, and minimize confounding factors. Differential gene expression analyses between ND and T2D samples were completed using edgeR at FDR 10%.

**Expression QTLs analysis:** Expected counts from RSEM for 9656 genes expressed (FPKM>5) in the 19 islets were used as input to RASQUAL. Only SNVs within the genes or those flanking 50 kilobases (kb) on either side of the gene body were tested for eQTL activity. The first four principal components and race were used as covariates to minimize confounding factors. Bonferroni correction was used to correct for the number of SNVs tested for each gene. 10 random permutations was generated for each gene, and used to correct for the number of genes tested, with an FDR cutoff of 10%.

**Luciferase reporter assays:** Genomic DNA from individuals homozygous for the reference and alternate alleles was used to amplify the 9 loci (see Supplemental Table S4). The 18 total constructs were cloned into the pDONR vector with BP Clonase.
(Invitrogen), which was then used to transfer the constructs into the Gateway-modified pGL4.23-FWD vector (Stitzel et al. 2010) with LR Clonase. Renilla luciferase (pRL-TK) was co-transfected with equimolar amounts of each pGL4.23 vector into MIN6 was used to normalize differences in transfection efficiencies as previously described (Stitzel et al. 2010). Cells were lysed in 1x Passive Lysis Buffer (PLB) 36 hours after transfection and luciferase activity was measured using the Dual Luciferase Reporter (DLR) Assay system (Promega) according to the manufacturer’s instructions. DLR activity was measured using a Synergy2 Microplate Reader (BioTek). The DLR ratio (Firefly/Renilla) for each construct was normalized to the empty pGL4.23 vector. The assay was performed 3 times. Each run included 3 separate mini-preps for each construct, and 3 technical replicates for each mini-prep.

**Supplementary Figure Legends**

**Figure S1. Chromatin accessibility maps in human islets.** (A) Insert size distributions of six representative islets (3 non-diabetic (ND), 3 T2D). ATAC-seq libraries capture nucleosome free and mono-, di-nucleosomal regions. (B) Number of ATAC-seq peaks called across the cohort, ranging from individual-specific peaks (n=1) to common peaks (n=19). (C) Islet ChromHMM annotations for ATAC-seq peaks categorized with respect to their frequency in the cohort. Note that common peaks are mostly promoters, whereas individual-specific or rare peaks include more quiescent or low signal regions. (D) TF motifs enriched in islet-specific peaks (from Figure 1F).

**Figure S2. Chromatin accessibility QTLs in islets.** (A) QQ-plot for expected and observed caQTL p-values. (B) Location of caQTLs (marked in green) across the genome. Note that caQTLs are widely distributed across each chromosome. (C) Functional annotation of caQTLs using ChromHMM states in islets and other tissues.
Tissues are sorted from highest to lowest overlap between ATAC-seq peaks and the ‘Quiescent/Low Signal’ state. Note the enrichment of caQTLs in islet enhancers. (D) Overlap of caQTLs with islet eQTLs from the same cohort. (E) Overlap of caQTLs with previously published islet eQTLs from 112 individuals (Varshney et al. 2017). (F) Table enumerating the 9 caQTLs that were tested for luciferase activity. Note that higher accessibility is associated with higher enhancer activity for all the 5 caQTLs displaying differential luciferase activity.

**Figure S3.** T2D-associated chromatin accessibility changes. (A) Principal Components 1 and 2 for the 10 islets. Note that the ND and T2D islets do not cluster together using all ATAC-seq peaks. (B) The weighted average proportion variance explained for the meta-variables associated with the 10 islets. Note that SVA reduces the variance attributed to all meta-variables, except for the one of interest (Condition) (C) The overlap of differential peaks detected with and without SVA is significant. (D) MA plot of all ATAC-seq peaks used for differential analyses (n=52,387). Positive logFC means the peaks are opening in T2D, and Negative logFC means the peaks are closing in T2D (CPM=counts per million). (E) Each cell in the heat map shows the number of differential peaks that are called as ATAC-seq peaks among ND and T2D islets. 918/1882 differential peaks are found in all 10 islets. (F) Venn diagram showing the number of opening or closing peaks overlapping caQTLs.

**Supplementary Table Legends**

**Table S1.** Meta data associated with the 19 islets.

**Table S2.** ATAC-seq quality control metrics for the 19 islets.

**Table S3.** RNA-seq quality control metrics for the 19 islets.

**Table S4.** Constructs for Luciferase Assay
Table S5. Differentially accessible ATAC-seq peaks in Islets

Acknowledgements

We are indebted to the anonymous pancreatic organ donors and their families, without whom this study would not be possible. We thank Jane Cha for her help in generating artwork for the figures and members of the Stitzel and Ucar labs for helpful discussion and critiques during study design, completion, and write-up. This study was supported in part by the National Institute of Diabetes and Digestive and Kidney Diseases (NIDDK) under award number DK092251 (to MLS) and by The Jackson Laboratory Startup and Director’s Innovation Funds (to M.L.S. and D.U.). Opinions, interpretations, conclusions, and recommendations are solely the responsibility of the authors and do not necessarily represent the official views of the National Institutes of Health.
References:

1000 Genomes Project Consortium, Auton A, Brooks LD, Durbin RM, Garrison EP,

Alasoo K, Rodrigues J, Mukhopadhyay S, Knights AJ, Mann AL, Kundu K, Hale C,
Dougan G, Gaffney DJ. 2017. Shared genetic effects on chromatin and gene
expression reveal widespread enhancer priming in immune response.
http://biorxiv.org/lookup/doi/10.1101/102392 (Accessed September 15,
2017).

Bolger AM, Lohse M, Usadel B. 2014. Trimmomatic: a flexible trimmer for Illumina

Brown JD, Lin CY, Duan Q, Griffin G, Federation AJ, Paranal RM, Bair S, Newton G,
Lichtman AH, Kung AL, et al. 2014. NF-κB directs dynamic super enhancer

Das S, Forer L, Schönherr S, Sidore C, Locke AE, Kwong A, Vrieze SI, Chew EY, Levy S,
McGuie M, et al. 2016. Next-generation genotype imputation service and

Degner JF, Pai AA, Pique-Regi R, Veyrieras J-B, Gaffney DJ, Pickrell JK, De Leon S,
are a major determinant of human expression variation. Nature 482: 390–
394.

2015. Sensitive detection of chromatin-altering polymorphisms reveals

Dhakshinamoorthy S, Jain AK, Bloom DA, Jaiswal AK. 2005. Bach1 competes with
Nrf2 leading to negative regulation of the antioxidant response element
(ARE)-mediated NAD(P)H:quinone oxidoreductase 1 gene expression and

Fadista J, Vikman P, Laakso EO, Mollet IG, Esguerra JL, Taneera J, Storm P, Osmark P,
Ladenvall C, Prasad RB, et al. 2014. Global genomic and transcriptomic
analysis of human pancreatic islets reveals novel genes influencing glucose

Fuchsberger C, Flannick J, Teslovich TM, Mahajan A, Agarwala V, Gaulton KJ, Ma C,


Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MAR, Bender D, Maller J, Sklar P, de Bakker PIW, Daly MJ, et al. 2007. PLINK: a tool set for whole-genome association and population-based linkage analyses. Am J Hum Genet 81: 559–575.


Chapter 3

Functional characterization of thousands of type 2 diabetes-associated and chromatin-modulating variants under steady state and endoplasmic reticulum stress

Functional characterization of thousands of type 2 diabetes-associated and chromatin-modulating variants under steady state and endoplasmic reticulum stress

Shubham Khetan\textsuperscript{1,3}, Susan Kales\textsuperscript{2}, Romy Kursawe\textsuperscript{1}, Alexandria Jillette\textsuperscript{1}, Steven K. Reilly\textsuperscript{4}, Duygu Ucar\textsuperscript{1,3,5}, Ryan Tewhey\textsuperscript{2,6,7*}, and Michael L. Stitzel\textsuperscript{1,3,5*}

\textsuperscript{1}The Jackson Laboratory for Genomic Medicine, Farmington, CT, 06032, USA
\textsuperscript{2}The Jackson Laboratory for Mammalian Genetics, Bar Harbor, ME, 04609, USA
\textsuperscript{3}Department of Genetics and Genome Sciences, University of Connecticut, Farmington, CT, 06032, USA
\textsuperscript{4}Broad Institute of MIT and Harvard, Cambridge, MA, USA
\textsuperscript{5}Institute of Systems Genomics, University of Connecticut, Farmington, CT, 06032, USA
\textsuperscript{6}Graduate School of Biomedical Sciences and Engineering, University of Maine, Orono, ME, 04469, USA.
\textsuperscript{7}Tufts University School of Medicine, Boston, MA, 02111, USA

*Corresponding authors: ryan.tewhey@jax.org and michael.stitzel@jax.org

Keywords: Massively Parallel Reporter Assay (MPRA), human pancreatic islets, MIN6 beta cells, chromatin accessibility, ATAC-seq, chromatin accessibility quantitative trait locus (caQTLs), genome-wide association study (GWAS), single nucleotide polymorphism (SNP), Type 2 diabetes (T2D), endoplasmic reticulum (ER) stress, short interspersed nuclear elements (SINE), Alu elements.
Abstract: (199 words)

A major goal in functional genomics and complex disease genetics is to identify functional cis-regulatory elements (CREs) and single nucleotide polymorphisms (SNPs) altering CRE activity in disease-relevant cell types and environmental conditions. We tested >13,000 sequences containing each allele of 6,628 SNPs associated with altered in vivo chromatin accessibility in human islets and/or type 2 diabetes risk (T2D GWAS SNPs) for transcriptional activity in β cell under steady state and endoplasmic reticulum (ER) stress conditions using the massively parallel reporter assay (MPRA). Approximately 30% (n=1,983) of putative CREs were active in at least one condition. SNP allelic effects on in vitro MPRA activity strongly correlated with their effects on in vivo islet chromatin accessibility (Pearson r=0.52), i.e., alleles associated with increased chromatin accessibility exhibited higher MPRA activity. Importantly, MPRA identified 220/2500 T2D GWAS SNPs, representing 104 distinct association signals, that significantly altered transcriptional activity in β cells. This study has thus identified functional β cell transcription-activating sequences with in vivo relevance, uncovered regulatory features that modulate transcriptional activity in β cells under steady state and ER stress conditions, and substantially expanded the set of putative functional variants that modulate transcriptional activity in β cells from thousands of genetically-linked T2D GWAS SNPs.
**Introduction**

Studies over the past decade have identified millions of single nucleotide polymorphisms (SNPs) in the human population\(^1\). Genome-wide association studies (GWAS) have linked thousands of these SNPs to variability in physiological traits and disease susceptibility\(^2\), including type 2 diabetes (T2D)\(^3\). The overwhelming majority of GWAS SNPs reside in non-coding, regulatory regions of the genome\(^3,4\), implicating altered transcriptional regulation as a common molecular mechanism underlying disease risk. Compared to protein-coding regions, predicting regulatory functions of non-coding sequences, and the effect of SNPs within them, remains a significant challenge.

Approaches such as ATAC-seq identify regions of accessible chromatin as a marker of *in vivo* cis-regulatory elements (CREs)\(^5\). However, CREs identified by chromatin accessibility (ATAC-seq peaks) vary significantly in length. Moreover, ATAC-seq peaks are characterized by 1 or more summits, which have higher chromatin accessibility than flanking genomic regions also within ATAC-seq peaks\(^6\). Few studies have explored the relationship between transcriptional activity and open chromatin regions defined by ATAC-seq\(^7,8\). For instance, it is not known whether higher chromatin accessibility within ATAC-seq peaks also corresponds to enhanced transcriptional activity, or whether proximity to ATAC-seq peak summits influences the likelihood for a SNP to affect enhancer activity. Studies designed to identify chromatin accessibility quantitative trait loci (caQTL) are being increasingly applied to nominate SNPs altering CRE activity\(^9,10\). Previously, we identified 2,949 caQTLs in human islets\(^11\), which were significantly enriched for type 2 diabetes (T2D)-associated SNPs, highlighting the utility of this approach to nominate putative disease-associated SNPs affecting CRE activity in relevant cell types. However, caQTLs are correlative associations, and high throughput assays to directly test variant effects are needed to experimentally establish causality.
Massively parallel reporter assays (MPRAs) have been developed as a functional genomics platform to interrogate the enhancer potential of thousands of sequences simultaneously\textsuperscript{12} under a variety of (patho)physiologic conditions. By introducing nucleotide changes in a given sequence of interest, the effect of naturally occurring sequence variants in the human population on MPRA activity can also be elucidated\textsuperscript{13}. Several studies have employed MPRA to identify functional SNPs associated with red blood cell traits\textsuperscript{14}, adiposity\textsuperscript{15}, osteoarthritis\textsuperscript{16}, cancer and eQTLs\textsuperscript{13}.

Although studies over the past several years implicate altered islet CRE function in T2D genetic risk and progression, they have colocalized only \(\sim1/4\) of T2D-associated loci to altered chromatin accessibility and/or gene expression levels in islets\textsuperscript{11,17–20}. We hypothesize that this is partially because previous studies measured chromatin accessibility (caQTLs) and gene expression (eQTL) in islets cultured under steady state conditions\textsuperscript{11,21}, consequently missing important gene-environment interactions characteristic of a complex disorder like T2D. For example, endoplasmic reticulum (ER) stress, which is elicited by the high demands of insulin production and secretion on protein folding capacity in \(\beta\) cells, leads to (patho)physiologic adaptations. Low levels of ER stress are a signal for \(\beta\) cells to proliferate as a mechanism to meet higher insulin secretion demands\textsuperscript{22}. However, uncompensated ER stress induced by genetic and/or environmental perturbations leads to \(\beta\) cell failure and T2D\textsuperscript{23}.

Here, we applied MPRA to directly test >13,000 sequences, including those overlapping T2D GWAS and islet caQTL SNPs, for their ability to activate \(\beta\) cell transcription in steady state conditions and after exposure to the ER stress-inducing agent thapsigargin or DMSO solvent control. We identified motifs and features of sequences affecting MPRA activity in each condition and linked allelic effects on MPRA activity to T2D genetics and \textit{in vivo} islet chromatin accessibility. These results demonstrate the power of MPRA to elucidate \textit{in vivo} relevant \(\beta\) cell transcriptional
activating sequences and define the regulatory grammar of these sequences under steady state and stress-responsive conditions. We nominate 54 putative causal variants among multiple genetically-linked islet dysfunction and T2D GWAS SNPs for further study.

**Results**

**Selection and testing of sequences for MPRA activity in β cells**

To test putative islet β cell regulatory sequences for their ability to enhance transcription from a minimal promoter and to identify SNPs that alter regulatory activity, we generated an MPRA library containing: i) 1,910 SNPs significantly associated with changes in human islet chromatin accessibility (caQTLs)\(^1\); ii) 2,218 control SNPs overlapping human islet ATAC-seq peaks, which were reported to have no correlation with changes in human islet chromatin accessibility\(^1\); and iii) 2,500 index and genetically linked \((r^2>0.8)\) SNPs/indels corresponding to 259 T2D association signals in the NHGRI/EBI GWAS Catalogue (Figure 1a, Table S2, and Methods). In total, 13,628 two-hundred base pair (bp) sequences from the human genome were included in this MPRA library.

MIN6 mouse β cells have been extensively employed to study human islet regulatory sequences because human and mouse transcription factor (TF) expression and DNA binding motifs are extensively conserved. MIN6 cells express many β cell-specific TFs, such as Pdx1, Nkx6.1 and Foxa2, and in a comparison with nine human tissues/cell-types, the chromatin accessibility profile of MIN6 mouse β cells most resembled that of human islets (Supplementary Figure 1). To test human sequences for islet β cell transcriptional activity, we transfected the MPRA plasmid library into five independent batches of MIN6 β cells under standard culture conditions (25mM glucose, Methods). Thirty hours later, transfected cells were harvested for RNA isolation, GFP
mRNA capture, and Illumina sequencing of the regulatory sequence-associated barcodes (Figure 1b). As anticipated, principal component analysis indicated that RNA expression of the transfected MPRA libraries was highly correlated between the five biological replicates and clustered distinctly from the MPRA plasmid library input (Supplementary Figure 2, compare RNA vs. DNA). 2,224/13,628 sequences (16.3%), representing 1,372 distinct elements, exhibited significantly higher counts after transfection compared to the plasmid library input under standard culture conditions (FDR<1%; Figure 1c; Supplementary Table 2). We refer to these sequences as ‘MPRA active’ throughout the remainder of the manuscript. MPRA active sequences were significantly enriched for in vivo binding by human islet-specific TFs (PDX1, NKX6.1 and FOXA2; ChIP-seq; Figure 1d), suggesting that MPRA identifies human islet β cell regulatory sequences with in vivo relevance and validating MIN6 as a powerful cellular model to test human sequences for islet β cell transcriptional activity.

Identification of β cell regulatory elements responding to ER stress

The high secretory burden of insulin production, processing, and secretion makes β cells particularly susceptible to ER stress. In fact, activation of the unfolded protein response (UPR) and ER stress has been implicated in the genetic etiology and pathophysiology of both monogenic\textsuperscript{24,25} and type 2 diabetes\textsuperscript{23}. Thapsigargin (TG) blocks calcium transport into the ER lumen and is widely used to induce ER stress in cells\textsuperscript{26–33}. Therefore, to identify sequences whose β cell transcriptional activity is modulated by ER stress, MIN6 cells transfected with the MPRA library were grown in standard culture media (25 mM glucose) supplemented with 250 nM TG or DMSO solvent control for 24 hours (Figure 2a). As expected, exposing MIN6 cells to 250 nM TG induced expression of ER stress response genes such as Ddit3 (Chop), Hspa5, and Edem1, and reduced Ins2 expression (Figure 2b). Compared to the DMSO solvent control, ER stress
decreased MPRA activity of 656 sequences (mapping to 449 elements) and increased MPRA activity of 328 sequences (mapping to 275 elements), respectively, at FDR <1% (Figure 2c).

We compared these differentially active sequences to identify TF motifs that may be mediating their transcriptional activity differences. Elements enriched for sequence binding motifs of β cell-specific TFs that mediate insulin gene transcription and beta cell function\textsuperscript{34,35}, such as MAFA, FOXA2 and PAX6, had lower MPRA activity under ER stress (Figure 2d). Concordantly, we observed a significant decrease in insulin gene expression (Figure 2b), suggesting ER stress leads to inactivation of these β cell-specific TFs. It is known that uncompensated ER stress leads to preferential ATF4 translation and induction of \textit{DDIT3/CHOP}\textsuperscript{36,37}. Consistently, we found that motifs for these TFs were enriched in elements with higher MPRA activity after ER stress induction (Figure 2d). Activity measurements with MPRA, therefore, recapitulate TF dynamics during ER stress and identify β cell regulatory elements that respond to ER stress.

\textbf{ATAC-seq peak summits are enriched for TF binding motifs and MPRA activity}

In total, 30% of sequences containing one or both SNP alleles were MPRA active in at least one treatment condition (Figure 3a, n=1983/6628, FDR ≤1%). MPRA active sequences were significantly enriched for 114 TF binding motifs (Supplementary Table 3), including those of islet TFs such as HNF1, MAFA, RFX5, and FOXA2. Human-mouse sequence similarity did not influence the probability that a tested human sequence was MPRA active in MIN6 (black dashed line in Figure 3b; Supplementary Figure 3). As anticipated, sequences overlapping regions with \textit{in vivo} evidence of transcriptional regulatory potential in human islets (ATAC-seq peaks) were significantly more likely to be identified as MPRA active regardless of the sequence similarity between human and mouse (Figure 3b).
Notably, the majority (60%) of elements accessible in human islets were not MPRA active (Figure 3b). Islet ATAC-seq peaks vary in length (100-3500 bps). Moreover, within an ATAC-seq peak, there are differences in chromatin accessibility levels at nucleotide resolution, where peak summits are locations with higher read counts than the flanking regions (schematic in Figure 3c; example in Figure 5a). Peak summits are significantly enriched in TF binding motifs compared to flanking regions within the same ATAC-seq peak (Figure 3c and Supplementary Figure 4). Therefore, for MPRA library sequences overlapping ATAC-seq peaks, we examined the distance of SNPs, which we designed to be in the center of all MPRA sequences tested, to the summit of ATAC-seq peaks in which they reside. Sequences closer to ATAC-seq peak summits were significantly more likely to be MPRA active (Figure 3d). However, this effect was observed only when SNPs were less than 100 bps from the ATAC-seq peak summit (dashed green line in Figure 3d). Since the MPRA library tests the 100 base pairs flanking both sides of a given SNP of interest, ATAC-seq peak summits were included in the MPRA sequences only if they were ≤100 bp from the SNP. Therefore, MPRA sequences overlapping ATAC-seq peaks were categorized based on whether the summit was within 100 bps of the SNP or not, i.e., whether the ATAC-seq peak summit was included in the sequence tested with MPRA or not. Sequences that included the ATAC-seq peak summit were indeed more likely to be active with MPRA, irrespective of the number of times a genomic region was accessible in a cohort of 19 islet donors (Figure 3e). Interestingly, inclusion of the ATAC-seq peak summit was predictive of MPRA activity even for genomic regions with accessibility in only one of 19 islet donors.

Together, these analyses indicate that sequences overlapping human islet ATAC-seq peaks were more likely to be active with MPRA in MIN6 mouse β cells than sequences outside peaks, regardless of sequence similarity between human and mouse. ATAC-seq peak summits are enriched for TF binding motifs, and exclusion of
ATAC-seq peak summits among the elements overlapping ATAC-seq peaks, significantly decreased the probability of observing MPRA activity.

**SNPs proximal to ATAC-seq summits have a higher probability of altering MPRA activity and chromatin accessibility**

To identify SNPs that modulate regulatory element activity, we assessed allelic differences (skew) in MPRA activity (schematic in Figure 1b) under each MIN6 beta cell culture condition (standard, 250nM TG, and DMSO solvent control). In total, 879 SNPs exhibited allelic skew at FDR <10% (Table S2; Supplementary Figure 5). Importantly, when allelic skew was detected in more than one experimental condition, the direction-of-effect was concordant for >98.5% of the SNPs (Supplementary Figure 5). Assessing allelic skew for SNPs across all conditions increased the power to detect allelic skew and identified both highly reproducible and condition-specific SNP allelic effects on MPRA activity.

SNPs associated with chromatin accessibility changes in islets (caQTLs) were significantly closer to the ATAC-seq peak summits than control SNPs, which are SNPs that reside in ATAC-seq peaks\textsuperscript{11} but were not associated with chromatin accessibility changes in islets (Figure 4a). Consequently, caQTL SNPs were more likely to be MPRA active than were control SNPs (Figure 4b). Among elements with MRPA activity, caQTL SNPs were also more likely to exhibit allelic skew than control SNPs (Figure 4c) and showed a significantly greater magnitude of skew between alleles (Figure 4d). Together, these results suggest that caQTL SNPs are closer to ATAC-seq peak summits and more likely to modulate the magnitude of both in vitro MPRA activity and in vivo chromatin accessibility, likely by disrupting TF binding motifs that are enriched in these regions.
Allelic effects on MIN6 beta cell MPRA activity and *in vivo* islet chromatin accessibility are significantly correlated

After identifying SNPs with allelic skew in MPRA activity, we investigated the relationship between SNP effects on *in vivo* islet chromatin accessibility (caQTLs) and *in vitro* MPRA activity. Figure 5a shows an example caQTL (rs17396537) for which islet donors with homozygous GG genotypes (n=6) exhibit higher chromatin accessibility than islet donors with GC (n=11) or CC (n=2) genotypes at this variant. Concordantly, the sequence containing the ‘G’ allele for rs17396537 exhibited higher MPRA activity than the one containing the ‘C’ allele (Figure 5a, b). 297/1910 lead caQTL SNPs demonstrated significant allelic skew in MPRA activity. Overall, caQTL and MPRA directions-of-effect for these 297 SNPs were highly correlated (Pearson R = 0.526, Figure 5b), and a significant concordance of direction was observed (i.e., alleles associated with higher chromatin accessibility also had higher MPRA activity) (n = 246/297; 82.8%).

A subset of lead caQTL SNPs (17.2%) showed discordant allelic effects on chromatin accessibility levels and MPRA activity (blue; Figure 5B). We hypothesized that discordance may be driven by a nearby SNP that has MPRA activity antagonistic to the lead caQTL SNP. For example, the reference allele of the lead caQTL SNP, rs1515555, is associated with *in vivo* higher chromatin accessibility but lower MPRA activity (Figure 5b). This region contains a second SNP, rs1515556, 24 bps away from rs1515555 and in tight LD ($r^2 = 0.99$). We therefore assessed sequences containing each of the four possible allelic combinations of these SNPs in the MPRA library for additive, antagonistic, or synergistic transcriptional effects. In comparison to rs1515555:rs1515556 REF:REF sequence (G:T in Figure 5c), the ALT:REF (A:T) sequence had higher MPRA activity, while the REF:ALT (G:C) sequence had lower MPRA activity (G:C). However, due to high LD between the 2 SNPs, both of these
combinations (‘A:T’ and ‘G:C’; ALT:REF and REF:ALT; Figure 5c) are rarely observed in the human population. Sequences with either REF:REF (G:T) or ALT:ALT (A:C) at rs1515555:rs1515556 are more likely to be observed in individuals in the population. Comparison of these haplotypes revealed concordant effects on both chromatin accessibility and MPRA activity and underscores the importance of accounting for haplotypes when characterizing allelic effects.

For 94 lead caQTL SNPs, we tested for potential interactions with neighboring SNPs by including all 4 allelic combinations in the MPRA library. However, these were too few to clearly implicate antagonism between neighboring SNPs for the observed discordance between chromatin accessibility and MPRA activity. Therefore, we cannot rule out discordant SNPs as merely being false positives or mechanistic difference due to technical constraints of the assay. However, MPRA provides a unique opportunity to test and dissect the contributions of neighboring SNP to in vivo effects.

Allelic effects on MPRA activity and chromatin accessibility are therefore significantly correlated. Importantly, these results demonstrate that that MPRA detects in vivo relevant allelic effects on β cell transcriptional activity.

**Functional identification of T2D SNPs altering β cell regulatory element activity**

A key challenge to translate GWAS associations into a mechanistic understanding of the genes and pathways altered by T2D risk variants is identifying the SNP(s) that affect functional or active CREs among tens to hundreds of genetically linked, associated SNPs in each locus. The ability to functionally annotate groups of SNPs with empirical measurements of CRE modulation would help prioritize variants and expedite our ability to identify T2D causal alleles. To uncover transcriptional effects of T2D-associated SNP alleles, we tested sequences containing each allele of 2,500 index and genetically linked (r²>0.8) SNPs/indels for 259 association signals for T2D and
related quantitative traits reported in the NHGRI/EBI GWAS Catalog. One or both alleles of 492 SNPs/indels were active by MPRA. Active sequences were enriched for the binding motifs of TFs that play important roles in β cell maturation and insulin secretion (Figure 6a), such as liver x receptor (LXR)\textsuperscript{38,39}, thyroid hormone receptor (THR\textsubscript{a} and THR\textsubscript{b})\textsuperscript{40,41}, retinoic acid receptor (RAR\textsubscript{a})\textsuperscript{42}, and BCL11A\textsuperscript{43}. Allelic effects on MPRA activity were observed for approximately half of these elements (n=220/492), corresponding to 104 distinct T2D-associated loci (Supplementary Table 4). Importantly, this list included T2D-associated SNPs that were previously identified for their \textit{in vivo} effects on islet chromatin accessibility and/or \textit{in vitro} effects on luciferase reporter assays, such as rs7903146 (TCF7L2)\textsuperscript{44,45}, rs1635852 (JAZF1)\textsuperscript{46}, rs10428126 (IGF2BP2)\textsuperscript{11,47}, and rs12189774 (VEGFA)\textsuperscript{48} (Figure 6b, Table 1). Importantly, MPRA has substantially expanded the set of T2D-associated GWAS SNPs with empiric effects on transcriptional activation in β cells, and therefore nominated them for the first time as the putative causal-functional variants in these T2D loci. These include rs2881632 (SDHAF4), rs13096599 (KBTBD8), rs7783500 (GCC1), rs72697237 (NOTCH2), rs13405776 (THADA), rs17748864 (PEX5L), rs7957197 (OASL), and rs13026123 (SRBD1), which are highlighted in Table 1.

For 54/104 T2D-associated signals, one SNP among those tested exhibited allelic effects on MPRA activity (black points in Figure 6b). For example, rs987964 was the only SNP of 49 tested in the ZBTB20 locus (Figure 6c; Supplementary Table 5) that exhibited allelic effects on MPRA activity. For this and 53 additional T2D-associated signals, MPRA thus provides empiric evidence that the variant directly modulates CRE activity and assists in prioritizing variants for additional investigation as putative causal alleles (Supplementary Table 5). The number of SNPs with allelic skew was found to generally correlate with the number of SNPs tested per locus (Figure 6b). For example, 10/45 tested SNPs in LD with the index SNP rs12304921 in the HIGD1C locus had an
allelic skew in MPRA activity (Figure 6d), not all of which were in the same direction with respect to the T2D risk alleles. Investigating the relative importance and potential interactions between these ten putative causal variants will be important to establish causality for this locus. These results underscore the value of assessing MPRA activity in disease-relevant cell types as a tool to nominate putative causal T2D SNPs based on systematic testing of their transcription-modulating effects in beta cells.

**ER stress modulates MPRA activity of T2D SNP-containing sequences**

MPRA activity of some putative CREs were exclusive to one out of the three conditions tested. For example, ~5% of those tested were identified as MPRA active exclusively in ER stressed beta cells (Figure 3a, n=106). However, we found that ER stress elicited substantial quantitative changes in MPRA activity of 724 CREs (Figure 2c). To investigate both the impact of ER stress on MPRA activity and how SNPs modify this response, we examined elements meeting two criteria: 1) allelic effects on MPRA activity was observed in at least one condition; and 2) a minimum of one allele showed differences in MPRA activity under ER stress (Figure 7a).

We previously characterized islet caQTLs from individuals with intact insulin expression and secretion, and observed an enrichment of islet-specific TF binding motifs at caQTL regulatory elements. ER stress, however, leads to the inactivation of β cell specific TFs (Figure 2d), with a concomitant decrease in Ins2 gene expression (Figure 2b). As might be expected based on these observations, steady state islet caQTL-containing sequences predominantly exhibited lower MPRA activity in TG-treated cells than the DMSO solvent controls. (Figure 7a, red dashes and 7b, red). In contrast, T2D SNP-containing elements were significantly more likely to have higher MPRA activity under ER stress (Figure 7a, green and 7b). Given the majority of SNPs tested in this set
are non-causal/functional ‘passenger’ alleles, this effect may reflect general CRE architecture across T2D-associated loci.

Curiously, only 13% (30/220) of T2D-associated SNPs with allelic skew in MPRA activity overlapped islet ATAC-seq peaks. We hypothesized that this may partially be due to mappability issues since 63.1% (139 / 220) of T2D-associated SNPs with allelic skew in MPRA activity overlapped repetitive sequences. Since there are no mappability issues with MPRA, we first asked if repetitive sequences have MPRA activity. Among the various classes of repetitive sequences, only elements overlapping short interspersed nuclear elements (SINEs) were significantly more likely to be active (Supplementary Figure 6a). Since >50% of elements overlapping SINEs in the MPRA library harbored T2D-associated SNPs (Supplementary Figure 6b and S6c), we next asked whether T2D-associated elements overlapping SINEs showed higher activity under ER stress. Indeed, >75% of T2D-associated elements with higher MPRA activity under ER stress overlapped SINEs, a significantly higher proportion compared to T2D-associated elements with lower or no change in MPRA activity under ER stress (Figure 7c). Alu elements, the most common SINEs in the human genome, are primate-specific. When we investigated conservation in 20 mammalian genomes, T2D-associated elements with higher MPRA activity under ER stress were indeed less likely to be conserved in all non-primate mammalian species for which comparisons were made (Figure 7d).

In conclusion, we have identified 220 elements whose MPRA activity is disrupted by T2D-associated SNPs. 40% (n=86) of these T2D-associated SNPs had significantly higher MPRA activity under ER stress, >75% of which overlapped SINEs.

Discussion
In this study, we used MPRA to test >13,000 sequences containing each allele of 6628 SNPs for MPRA activity in MIN6 β cells under standard culture conditions and after ER stress or paired solvent control exposures. In total, 30% (n=1983/6628) of putative CREs tested increased transcriptional activity of a minimal promoter. SNP alleles, including those associated with altered in vivo chromatin accessibility in human islets and with T2D genetic risk by GWAS, altered MPRA activity for approximately half of these elements (n=879/1983).

MPRA activity exhibited a striking, positive correlation with in vivo islet chromatin accessibility. As anticipated, elements accessible in vivo were far more likely to have MPRA activity in vitro. However, MPRA refined our understanding of the specific content and location of DNA sequences within ATAC-seq peaks that drive β cell transcriptional activation. Likelihood of an element showing MPRA activity increased as a function of its proximity to ATAC-seq peak summits, and SNPs closer to ATAC-seq peak summits were more likely to alter MPRA activity and chromatin accessibility. These results help to prioritize sequences within open chromatin regions for their importance in regulating enhancer activity by disrupting TF binding. Interactions between neighboring SNPs sometimes led to discordant effects between chromatin accessibility and activity measured with MPRA. In these instances, MPRA helped to determine the relative contributions of neighboring SNPs (e.g., additive or antagonistic effects) on transcriptional activation.

T2D-associated SNPs are overwhelmingly non-coding, and primarily alter regulatory activity rather than protein structure and function. A key challenge in T2D genetics is to identify the putative causal SNP(s) from among tens to hundreds of genetically linked, noncoding variants per association signal. This is the first study to systematically test thousands of T2D-associated SNP alleles for their effects on transcriptional activity in β cells. MPRA identified 492 elements in T2D-associated loci as
active. 220/492 SNP alleles, representing 104 distinct T2D association signals, significantly altered β cell MPRA activity. MPRA recapitulated, and thereby confirmed, allelic effects of several T2D SNPs that have been characterized previously using targeted, low throughput luciferase assays, such as rs10428126 and rs2943656 at the IGF2BP2 and IRS1 locus, respectively. For 54/104 T2D association signals, only one SNP showed an allelic skew among all the SNPs tested, implicating them as the putative causal SNP within their respective locus. Although not an exhaustive test of all credible set SNPs identified to date, this approach has substantially expanded the list of T2D-associated SNP alleles that empirically alter β cell regulatory sequence activity. These results should facilitate targeted follow-up experiments and integrated genomic analyses for a better understanding of the functional genetics of islet (dys)function and T2D risk and strongly motivate future studies of more comprehensive panels of T2D-associated SNPs for their transcription-modulating effects in beta cells and other diabetes-relevant cell types under steady state, stimulatory, and/or stress conditions.

Islet dysfunction and beta cell failure in T2D results from both genetic and environmental risk factors. MPRA of sequences containing T2D-associated SNPs, when tested under control and ER stress conditions, revealed the effects of these factors on β cell transcriptional activation. A significant proportion (40%; 86/220) of sequences containing T2D-associated index or linked SNPs exhibiting allelic skew under control conditions had significantly higher MPRA activity under ER stress, suggesting – i) MIN6 cells cultured under standard, high glucose (25 mM) conditions are already under a basal level of ER stress, and ii) many T2D-associated SNPs overlap CREs relevant for β cells to respond to ER stress. Uncompensated ER stress was found to lead to the inactivation of β cell-specific TFs causing the downregulation of insulin transcription and secretion\textsuperscript{50–54}. ER folding and trafficking capacity may therefore be a major factor determining how much insulin can be released by β cells under elevated blood glucose.
levels, before stress ensues. SNPs at regulatory elements modulating the expression of genes relevant to meet higher demands for insulin synthesis may impair or improve ER capacity, ultimately determining the threshold at which ER stress ensues, leading to β cell failure.

Finally, we uncovered evidence that repetitive element sequences, most notably SINEs, elicited robust transcriptional activation in β cells and that multiple T2D-associated SNPs residing in these sequences modulated their transcriptional activity. The MPRA results obtained in this study thus suggest that repetitive element-containing sequences may play important roles in modulating stimulus and/or stress-responsive β cell transcriptional programs and remind us to consider the potentially important roles that repetitive elements, and SNPs within them, may play in the genetics of islet (dys)function, diabetes risk and progression. Studies over the past few years demonstrating repetitive element-mediated oncogene activation and modulation of chromatin structure have contributed to an emerging appreciation of the importance of these sequences in epigenetic and transcriptional regulation\textsuperscript{65–63}. Recently, Hernandez et al found that Alu elements are transcriptionally induced by cellular stress, including thermal and ER stress, and that the corresponding SINE RNAs function as critical transcriptional switches during stress\textsuperscript{64}. Future studies to elucidate the target genes of these and other MPRA active, SINE-containing regulatory sequences, will be necessary to fully understand the functional consequences of sequence variation in these transcriptionally active sequences and the potential role(s) of exaptation\textsuperscript{57,65,66} in the genetics of islet (dys)function and T2D.

\textbf{Figure Legends}

\textbf{Figure 1.} Massively Parallel Reporter Assay (MPRA) identifies beta cell transcription activating sequences. (a) Features of the 6,628 two hundred base pair
(bp) elements selected for MPRA library construction and testing. 4,329 elements overlap islet ATAC-seq peaks, of which 1,910 contain SNPs associated with altered in vivo chromatin accessibility (caQTLs). 2,500 elements contain T2D-associated SNPs (index or linked ($r^2>0.8$)) reported in the NHGRI/EBI GWAS catalog, 201 of which overlap islet ATAC-seq peaks. **(b)** Schematic of MPRA approach. A mock SNP, rs123456789, is used as an example to show how MPRA activity and allelic skew is measured for each 200 bp regulatory element. mP = minimal promoter. **(c)** Log$_2$ fold-change in barcode/sequence counts in plasmid input (DNA) and gfp mRNA transcripts (RNA) 30 hr after transfection of MPRA library into MIN6 cells grown in standard culture conditions (DMEM, 25 mM glucose). Red points indicate sequences with MPRA activity (FDR<1%; n = 2224/13,628). **(d)** Barplot showing the odds of MPRA active sequences overlapping in vivo human islet TF ChIP-seq peaks compared to MPRA-inactive sequences. ***p<0.001; **p<0.01, Fisher’s Exact Test.

**Figure 2.** MPRA identifies endoplasmic reticulum (ER) stress-responsive transcriptional activating sequences. **(a)** Experimental design to identify ER stress-responsive transcription activating sequences. **(b)** RT-qPCR of ER stress response genes (Ddit3, Hspa5, Edem1, and Hsp90b1), Ins2, and the control gene Actb in MIN6 cells exposed to the ER stress inducer thapsigargin (TG, 250 nM) or dimethylsulfoxide (DMSO) solvent control for 24 hours. Bar plots show mean +/- SEM from three biological replicates. *p<0.05; ** p<0.01; *** p<0.001, paired t-test. **(c)** Heatmap of sequences with higher (far right column, red bar; n=328, mapping to 275 elements) or lower MPRA activity (blue bar; n=656, mapping to 449 elements) under ER stress (FDR<1%). Red annotation bars to the left of the heatmap indicate sequences identified as MPRA active in DMSO and/or TG based on their sequence counts in RNA vs. plasmid DNA input. **(d)** Comparison of transcription factor (TF) binding motifs enriched in elements with higher
(n=275) or lower (n=449) MPRA activity under ER stress. Red and blue dots denote TF motifs significantly enriched in elements with higher or lower MPRA activity in ER stressed MIN6 beta cells, respectively (FDR<1%). Yellow dots denote TF motifs with no significant enrichment in either comparison.

Figure 3. ATAC-seq peak summits are enriched for MPRA activity. (a) Venn diagram showing the number of elements with at least one allele identified as MPRA active in standard culture, TG, and DMSO conditions (FDR<1%). Note that elements identified as MPRA active in ≥1 experimental condition may still exhibit quantitative differences in MPRA activity under ER stress. (b) The probability that an element has significant MPRA activity (y-axis; at least 1 allele; any experimental condition) is shown as a function of sequence similarity (x-axis) between human and mouse genomes (expected probability is the black dotted line). Elements overlapping islet ATAC-seq peaks (brown line) had a significantly higher probability of MPRA activity than elements that did not overlap ATAC-seq peaks (green line), regardless of human-mouse sequence similarity. ***, **, and * indicate Bonferroni-corrected Fisher's Exact Test p<0.001, 0.01, 0.05, respectively. (c) 200 bp genomic regions around human islet ATAC-seq peak summits were compared to 200 bp regions also within ATAC-seq peaks, but flanking the peak summit. ATAC-seq peak summits were significantly enriched for many TF motifs compared to genomic regions that overlap ATAC-seq peaks but are further away from the summit. (d) Elements where the SNP (center of all tested loci) was ± 100 bp (green dashed lines) from ATAC-seq peak summit were more likely to be identified as MPRA active. This effect is restricted to SNPs within 100 bps of the ATAC-seq peak summit because our MPRA library was designed to test only the 200 bp sequence on either side of a given SNP of interest. ATAC-seq peak summits were therefore included in the sequences tested only if the distance of the SNP was within 100 bps of the ATAC-
(e) (Top) The probability of an element (at least one allele) being identified as MPRA active (y-axis) binned by the number of islet donors (out of 19 total) in whom the ATAC-seq peak was present (x-axis). Elements where the ATAC-seq peak summit was included in the sequence tested, i.e., distance of SNP to ATAC-seq peak summit was less than 100 bps (red), had a significantly higher probability of being identified as MPRA active. * indicates FDR < 10%, Fisher’s Exact Test. (Bottom) Stacked barplot of the number of times an ATAC-seq peak overlapping the MPRA sequence tested is detected in the cohort (n = 19 islet donors).

**Figure 4. SNPs closer to ATAC-seq peak summits are more likely to affect MPRA activity and chromatin accessibility.** (a) caQTL SNPs (blue bars) are closer to ATAC-seq peak summits than control SNPs (red bars), irrespective of MPRA activity (pairwise Wilcoxon tests with Bonferroni correction for multiple testing). (b) Elements containing caQTL SNPs are more likely to be identified as MPRA active (any experimental condition; Fisher’s Exact Test). (c) Among SNPs with at least 1 allele identified as MPRA active, caQTLs are significantly more likely to show allelic skew (any experimental condition), suggesting polymorphisms proximal to ATAC-seq peak summits are more likely to affect MPRA activity (Fisher’s Exact Test). (d) Among SNPs with allelic skew in MPRA activity, caQTLs have a significantly larger difference in MPRA activity between alleles than control SNPs (Wilcoxon Test). For all panels, *, **, and *** indicate adjusted p < 0.05, p < 0.01, and p < 0.001, respectively.

**Figure 5. Alleles associated with higher chromatin accessibility also have higher MPRA activity.** (a) (Left) caQTL example: Islet samples GG homozygous for rs17396537 (green; average ATAC-seq read counts) were associated with significantly higher chromatin accessibility than samples with GC (pink) or CC (black) genotypes.
Sequence with the reference ‘G’ allele at rs17396537 displayed significantly higher MPRA activity than the alternate ‘C’ allele. **(b)** Correlations between allelic effects on *in vivo* islet chromatin accessibility (x-axis) and MPRA activity (y-axis) for sequences containing islet caQTL SNPs. Quadrants 1 and 3 (red) correspond to SNPs where the allele associated with higher chromatin accessibility also had higher MPRA activity (concordant). The number of SNPs in each quadrant is indicated. Pearson R = 0.526, p value < 2.2e-16) among the caQTL SNPs with significant allelic skew in MPRA activity (FDR<10%). Brackets indicate points beyond the axis limits. For example, there are a total of 135 SNPs in the bottom left quadrant, 2 of which are beyond the axis limits. REF=hg19 reference allele; ALT=alternate allele. **(c)** As shown in panel b, the lead caQTL SNP rs1515555 demonstrated apparently discordant allelic effects between chromatin accessibility and MPRA activity. Since there is another SNP, rs1515556, 24 bps from the lead caQTL SNP ($r^2=0.99$), sequences containing all four allelic combinations were tested in the MPRA library. The alternate allele of rs1515555 increased MPRA activity, whereas that of rs1515556 decreased MPRA activity. Ref/Ref for rs1515555 and rs1515556 exhibited higher MPRA activity than the Alt/Alt combination, which is concordant with its higher chromatin accessibility associated with Ref/Ref for rs1515555 and rs1515556.

**Figure 6. Functional identification of T2D SNPs altering β cell regulatory element activity.** **(a)** TF binding motifs enriched at 492 elements containing T2D-associated SNPs with significant MPRA activity (at least one allele). Red dots denote significantly enriched TF binding motifs (FDR<1%). **(b)** MPRA identifies a subset of T2D-associated index and genetically linked ($r^2\geq0.8$) SNPs representing 259 signals from the NHGRI/EBI GWAS catalog (x-axis; log scale) that demonstrate significant allelic effects on MPRA activity (y-axis) in one or more condition tested. A total of 220 SNPs in linkage
disequilibrium with 104 T2D-associated index SNPs show significant allelic effects on MPRA activity (FDR<10%). Jitter in the plot is used to visually separate individual points. **(c)** (Top). In total, 49 SNPs in LD ($r^2 > 0.8$) with the T2D-associated index SNP rs73230612 were tested with MPRA (ZBTB20 locus). Allelic skew in MPRA activity was detected in the same direction across all 3 experimental conditions (standard culture, DMSO and TG) at 1 SNP only, rs987964. (Bottom; zoom-in) The 200 bp genomic region spanning rs987964 is specific to humans. **(d)** In total, 45 SNPs in LD ($r^2 > 0.8$) with the T2D-associated index SNP rs12304921 were tested with MPRA (HIGD1C locus). Allelic skew in MPRA activity was detected at 10 SNPs.

**Figure 7.** ER stress increases MPRA activity of 40% of T2D-associated SNP containing sequences with allelic skew. **(a)** Among the SNPs with allelic skew (any condition), the heatmap includes those with at least 1 allele showing significantly higher or lower MPRA activity under ER stress vs. DMSO (row annotation (right); red and blue bars, respectively). For each SNP (rows), z-scores are obtained by centering and scaling the normalized RNA/DNA ratios for the reference and alternate alleles (5 replicates each: DMSO and TG). Annotation horizontal bars on the left indicate whether an element (row) harbors a caQTL or T2D-associated SNP. Note that the majority of elements with higher MPRA activity under ER stress contain T2D-associated SNPs. **(b)** SNPs with allelic skew (any condition) are categorized as having at least 1 allele with – i) lower, ii) no difference in, or iii) higher MPRA activity under ER stress vs. DMSO (x-axis). SNPs in each of the above categories are further fractioned as being – i) caQTLs (red), ii) control SNPs (yellow), or iii) T2D-associated (green). A significant proportion of caQTL SNPs with allelic skew were significantly more likely to show reduced MPRA activity, and significantly less likely to show higher MPRA activity, under ER stress. A significant proportion of T2D-associated SNPs with allelic skew displayed the opposite trend: they
were significantly less likely to show reduced MPRA activity, and significantly more likely to show higher MPRA activity (86/220), under ER stress. (c) Fraction of T2D-associated elements or 10,000 random genomic regions overlapping SINE repeats in the human genome. The T2D-associated elements are categorized based on whether the elements show – i) MPRA activity, ii) Allelic skew in MPRA activity, or iii) higher MPRA activity under ER stress. A significant fraction of T2D-associated elements with higher activity under ER stress overlap SINEs. (d) The probability that elements containing SNPs with allelic skew are conserved with sequence similarity greater than 20% (y-axis) is plotted for 20 mammals whose genomes are available (x-axis). Red and yellow dots show comparisons for caQTL and control SNPs, respectively. T2D-associated SNPs with allelic skew were split based on whether they had significantly higher MPRA activity under ER stress (n=86; purple), or no change (n=110; green). Compared to 10,000 random 200 bps regions from the human genome, elements overlapping caQTL and control SNPs were significantly more likely, and T2D-associated elements with higher MPRA activity under ER stress significantly less likely, to be conserved in non-primate mammalian species (Fisher’s exact test; FDR<5%). For Panels B and C, Fisher’s Exact Test p-value (corrected for multiple testing) < 0.001 is indicated as ***, ** indicates p-value < 0.01, and * indicates p-value < 0.05.

**Supplementary Figure Legends**

**Supplementary Figure 1**: The chromatin accessibility profile of MIN6 mouse β cells most resembles that of human islets. (a) ATAC-seq peaks from 9 human tissues were mapped to the mouse genome (mm9) using the UCSC Genome Browser liftover tool and compared to MIN6 ATAC-seq peak locations. (b) Heatmap showing the enrichment of islet-specific TF motifs at human ATAC-seq peaks uniquely overlapping MIN6 ATAC-seq peaks.
Supplementary Figure 2: QC Metrics of MPRA libraries. (a) Scatter plot of the first 2 principal components, which together explain >99% of the variation in 5 plasmid and 5 RNA MPRA replicates. RNA replicates were obtained after transfection of the MPRA library into MIN6 cells under standard culture conditions. (b) Pairwise Pearson correlation coefficients are shown as a heatmap with unsupervised row and column clustering.

Supplementary Figure 3: Human-Mouse sequence similarity of elements with MPRA activity. Human-mouse sequence similarity for elements with (red) or without (black) MPRA activity is plotted. Human elements with 0% sequence similarity did not liftover to the mouse genome (mm9) with even 1% sequence similarity. A bootstrap hypothesis test of equality did not find a statistically significant difference between elements with (red) or without (black) MPRA activity for human-mouse sequence similarity. The upper and lower end-points for equality is indicated as a blue reference band.

Supplementary Figure 4: Enrichment of islet-specific TF binding motifs at islet ATAC-seq peak summits. Islet ATAC-seq peak summits are significantly enriched for many islet-specific TF motifs compared to genomic regions that overlap ATAC-seq peaks but are further away from the summit.

Supplementary Figure 5: Allelic skew in MPRA activity for 879 SNPs across the 3 experimental conditions. (Top) Venn diagrams showing the number of SNPs with allelic skew in MPRA activity (FDR<10%) in each condition. (Bottom) Scatter plot of log fold change in MPRA activity for SNPs with allelic skew in 2 experimental conditions.
SNPs for which comparisons are made are indicated as pink in the corresponding Venn diagram above each scatter plot.

**Supplementary Figure 6: The majority of elements overlapping SINEs harbor T2D-associated SNPs.** (a) Barplot showing the odds of MPRA active sequences (at least one allele; any experimental condition) overlapping long interspersed nuclear element (LINE), long terminal repeat (LTR), or short interspersed nuclear element (SINE) repetitive element classes. ***p< 0.001, Fisher’s Exact test. (b) Fraction of elements containing i) caQTL, ii) control, or iii) T2D-associated SNPs that overlap SINEs. 10,000 random genomic regions from hg19 are included for context. (c) Stacked barplot showing the fraction of elements in the MPRA library overlapping SINEs in the human genome.

**Methods**

**MPRA library design:** 200 base pair sequences, with 100 bps flanking each side of 6628 SNPs were included in our MPRA library. The SNPs belong to 3 categories:

1. Islet chromatin accessibility quantitative trait loci (caQTLs): These are SNPs we previously identified as having a significant association with altered *in vivo* chromatin accessibility in islet samples\(^\text{11}\). For 1806 caQTLs, only the lead SNP was included in the MPRA library. For 94 caQTLs, the 200-bp sequences contained two SNPs in LD less than 25 bp apart. For these sequences, all four allelic combinations were synthesized and tested.

2. Control SNPs: SNPs that overlapped islet ATAC-seq peaks but did not significantly alter accessibility of those peaks were also synthesized and tested\(^\text{11}\).
Since islet caQTLs were identified in a relatively small cohort of individuals (n=19), the following criteria were used to include SNPs for which the caQTL study was likely more powered to detect associations with chromatin accessibility:

- unadjusted p value > 0.2,
- minor allele frequency > 0.125

SNPs overlapping individual-specific peaks or sharing peaks with other SNPs were further filtered out. Of the remaining 15,178 SNPs, 2218 were randomly selected to be included in the MPRA library.

3. T2D-associated SNPs/indels: 200 bp sequences overlapping SNPs (n=2299), small insertions (n=72) and small deletions (n=129) in linkage disequilibrium ($r^2>0.8$) with T2D-associated index SNPs (n=259) from the NHGRI/EBI GWAS Catalog (accessed January 19th, 2017) were synthesized and tested, as previously described. Briefly, T2D-associated GWAS SNPs were pruned using PLINK version 1.9 to identify SNPs in high linkage disequilibrium ($r^2>0.8$).

The vast majority of SNPs and sequences tested belonged to only 1 of the 3 categories. However, T2D-associated SNPs overlapping 13 ATAC-seq peaks were significantly associated with chromatin accessibility in islets (caQTLs). Therefore, for analysis purposes, whenever SNPs were required to belong to only 1 of the 3 categories above (such as Figures 1A, 7A and 7B), they were not categorized as caQTLs, but as being T2D-associated only.

**MPRA library construction:** The MPRA library was constructed as previously described. Briefly, oligos were synthesized (Agilent Technologies) as 230 bp sequences containing 200 bp of genomic sequences and 15 bp of adaptor sequence on
either end. Unique 20 bp barcodes were added by PCR along with additional constant sequence for subsequent incorporation into a backbone vector by Gibson assembly. The oligo library was expanded by electroporation into *E. coli*, and the resulting plasmid library was sequenced by Illumina 2 X 150 bp chemistry to acquire oligo-barcode pairings. The library underwent restriction digestion, and GFP with a minimal TATA promoter was inserted by Gibson assembly resulting in the 200 bp oligo sequence positioned directly upstream of the promoter and the 20 bp barcode falling in the 3’ UTR of GFP. After expansion within *E. coli* the final MPRA plasmid library was sequenced by Illumina 1 X 31 bp chemistry to acquire a baseline representation of each oligo-barcode pair within the library. Barcodes mapping to more than 1 sequence were discarded from all downstream analyses. **Note:** 2 separate batches of the MPRA library were prepared. The first batch was used to perform MPRA under standard culture conditions. This MPRA library was then electroporated into *E. coli* to obtain a second batch of the MPRA library, which was used for the DMSO-TG experiments.

**MPRA library transfection into MIN6 cells:** 10 million MIN6 cells were seeded in each of seven 15 cm² dishes. The cells were 60-70% confluent the next day. Each 15 cm² dish was replaced with 20 ml of fresh media, and transfected with 7ug of the MPRA plasmid library using 55ul Lipofectamine 2000 (38% transfection efficiency). Six hours after transfection, media was either i) not changed (MPRA under standard culture conditions), ii) replaced with media containing 250 nM Thapsigargin dissolved in 0.025% DMSO, or iii) replaced with media containing 0.025% DMSO. Thirty hours after transfection, cells were trypsinized and collected by centrifugation. Cell pellets were frozen at -80°C. For each condition (standard culture / DMSO / 250 nM TG), MIN6 cells were transfected on five separate days to generate biological replicates.
RNA isolation and MPRA RNA-seq library generation: RNA was extracted from frozen cell pellets using the Qiagen RNeasy Midi kit. Following DNase treatment, a mixture of 3 GFP-specific biotinylated primers (Supplementary Table 1; #120, #123 and #126) were used to immunoprecipitated GFP transcripts using Streptavidin C1 Dynabeads (Life Technologies). Following another round of DNase treatment, cDNA was synthesized from GFP mRNA using SuperScript IV and purified with AMPure XP beads. Quantitative PCR using primers specific for GFP (Supplementary Table 1; #34 and #52) was used to determine the cycle at which linear amplification begins for each replicate. Replicates were diluted to approximately the same concentration based on the qPCR results, and PCR with primers #34 and #52 was used to amplify barcodes associated with the ~13.5k sequences included in the MRPA library for each replicate (9 cycles for standard culture, and 13 cycles for DMSO / 250 nM TG). A second round of PCR (6 cycles) was used to add Illumina sequencing adaptors to the DNA/RNA replicates. The resulting MPRA barcode libraries were spiked with 5% PhiX and sequenced using Illumina single-end 31 bp chemistry (with 8 bp index read), clustered at 80-90% maximum density.

MPRA data analysis: Data from the MPRA was analyzed as previously described\textsuperscript{13}. Briefly, the sum of the barcode counts for each oligo within replicates was median normalized, and oligos showing differential expression relative to the plasmid input were identified by modeling a negative binomial distribution with DESeq2 and applying a false discovery rate (FDR) threshold of 1%. For sequences that displayed significant MPRA activity, a paired t-test was applied on the log-transformed RNA/plasmid ratios for each experimental replicate to test whether the reference and alternate allele had similar activity. An FDR threshold of 10% was used to identify SNPs with a significant skew in MPRA activity between alleles (allelic skew). Because the MPRA testing standard
culture conditions was performed with a separate MRPA library preparation. Therefore, the DMSO-TG MPRA results were not directly compared to MPRA performed under standard culture conditions.

**Annotating repetitive elements tested with MPRA:** The ‘RepeatMasker’ track for hg19 was downloaded from the UCSC genome browser. Among the ten different classes of repeats, only three classes (long interspersed nuclear element (LINE), long terminal repeat (LTR), and short interspersed nuclear element (SINE)) overlapped more than 100 elements tested with MPRA. Therefore, only these three classes of repeats were assessed for associations with MRPA activity.

**Mapping human regulatory sequences tested with MPRA to mammalian genomes:** Liftover tool in the UCSC genome browser was used to map human sequences (hg19) tested with MPRA to 20 mammalian genomes (with a minimum ratio of 0.20 bases that must remap; allowing for multiple output regions). The 20 mammalian genomes are: papAnu2 (Baboon), felCat5 (Cat), PanTro6 (Chimpanzee), BosTau7 (Cow), canFAM3 (Dog), loxAfr3 (Elephant), nomLeu3 (Gibbon), gorGor3 (Gorilla), equCab2 (Horse), mm9 (mouse), ponAbe2 (Orangutan), aiMel1 (Panda), susScr11 (Pig), ochPri3 (Pika), oryCun2 (Rabbit), m5 (Rat), rheMac8 (Rhesus), oviAri3 (Sheep), sorAra2 (Shrew), speTri2 (Squirrel). Human sequences that did not liftover to the genome assembly of a given species were subsequently classified as not conserved (with a minimum ratio of 0.20 bases that must remap; allowing for multiple output regions).

To obtain human-mouse sequence similarity measures, liftover was performed 99 times with the minimum ratio of bases that must remap ranging from 0.01 to 1.00 in increments of 0.01 (allowing for multiple output regions). The R package ‘sm’ was used to plot density of human-mouse sequence similarity and perform non-parametric bootstrap.
hypothesis tests of equality. Human sequences that did not liftover to the mm9 mouse genome with even 1% sequence similarity were classified as having 0% sequence similarity.

Cross-species mapping of human ATAC-seq peaks to MIN6 ATAC-seq peaks: Human and mouse ATAC-seq data were processed as previously described\textsuperscript{11}. Briefly, low quality portions of reads were trimmed using Trimmomatic\textsuperscript{70} and aligned to the hg19 or mm9 genome assembly using Burrows Wheeler Aligner-MEM. For each replicate, duplicate reads were removed after shifting. Technical replicates were merged using SAMtools and peaks were called using MACS2\textsuperscript{6} (with parameters -callpeak --nomodel -f BAMPE) at FDR<1%. ATAC-seq peak summit positions were obtained from MACS2 output files. The liftover tool in the UCSC genome browser was used to map human ATAC-seq peaks to the mouse (mm9) genome using a minimum ratio of 0.10 bases that must remap (not allowing for multiple output regions). Using bedtools, human ATAC-seq peaks mapping to the mouse genome were then overlapped with MIN6 ATAC-seq peaks.

Transcription factor (TF) motif enrichment: Homer findMotifsGenome.pl script was used to investigate TF motifs enriched in a given set of sequences. Elements with lower MPRA activity under ER stress were used as background to identify TF motifs enriched in elements with higher MPRA activity under ER stress, and vice-versa (parameters: hg19, -size given). 2,008 T2D-associated elements with no MPRA activity were used as background to identify TF motifs enriched in the 492 T2D SNP-containing sequences with significant MPRA activity (parameters: hg19, -size given). For the cross-species analysis of ATAC-seq peaks, ATAC-seq peaks shared with other human cell types were
used as background to identify TF motifs enriched in unique ATAC-seq peaks (parameters: mm9, -size given).

**Analysis of islet ChIP-seq data:** Chromatin immunoprecipitation sequencing (ChIP-seq) data from Pasquali et al.\textsuperscript{17} were aligned to the hg19 reference human genome as previously described\textsuperscript{44}. Elements tested with MPRA were then overlapped with ChIP-seq peaks to conduct Fisher’s exact tests using R.

**CentriMo analysis of ATAC-seq peak summits:** CentriMo software in the MEME Suite was used to identify TF motifs enriched in the 100 bp genomic regions flanking human islet ATAC-seq peak summits\textsuperscript{71}. Negative controls were 200 bp genomic regions overlapping ATAC-seq peaks but flanking the human islet ATAC-seq peak summits (default parameters).

**Acknowledgements.** We thank members of the Stitzel and Ucar labs for helpful discussion and critiques during study design and execution and Francis S. Collins, D. Leland Taylor, Christine Beck and Taneli Helenius for helpful comments on the manuscript. This work was supported by W81XWH-18-0401 (to MLS and DU) and R00HG008179 (to RT).

**References**


Figure 1

(a) Distribution of mean DNA count for different categories of genetic variants:
- caQTLs (n=1910)
- control SNPs (n=2218)
- T2D-Associated SNPs/Indels (n=2500, 259 Signals)

(b) Schematic representation of RNA-DNA ratio and its relationship with allele-specific expression:
- RNA = DNA (not Enhancer)
- RNA > DNA (is Enhancer)

(c) Scatter plot showing the relationship between log2(RNA/DNA) and mean DNA count.

(d) Bar chart illustrating the odds of MRPA active sequences overlapping human islet ChIP-seq peaks for various transcription factors:
- CTCF
- H2A.Z
- H3K27ac
- PDX1
- MAFB
- NKX6.1
- FOXA2

Note: The diagram includes a representation of transfection of MIN6 beta cells with plasmid pools containing different barcodes and gene expression markers.
Figure 2

Plasmid Pool

- rs123456789-allele-A mP GFP Barcode-A
- rs123456789-allele-B mP GFP Barcode-B

- 6,628 SNPs
- 13,628 sequences

Transfect MIN6 mouse beta cells

Standard Culture

- 6 hours

- 24 hrs

0.025% DMSO

- solvent control

250 nM Thapsigargin (TG)

Harvest cells for RNA

- 30 hrs after transfection

DNA RNA

DMSO

- RNA

TG

- RNA

Replicates 1-5

Replicates 1-5

Replicates 1-5

MPRA activity

- Lower in TG (n=656)
- Higher in TG (n=328)

b

log(2)(TG / DMSO)

Normalized to Gapdh

Actb

Lns2

Ddit3/Chop

Hspa5

Edem1

Hsp90b1

- ns

* ****

- ns

- ns

- ns

- ns

- ns

- ns

- ns

- ns

- ns

d

log(2)(Enrichment) DMSO-TG

CEBP

Atf4

Chop

LXRE

THRα

Mef2a

n-Myc

MafA

CRE

CLOCK

RXR

Foxa2

Rfx1

USF1

X-box

TFE3

PAX6

Hnf1

86
**Figure 3**

**a** Probability of an element showing MPRA activity

**b** Overlap Islet ATAC-seq peak (n=4329)
- - - - All elements tested
- - - - Doesn't Overlap ATAC-seq peak (n=2299)

**c** Model ATAC-seq peak

**d** Elements overlapping ATAC-seq peaks are:

**e** Elements tested:

- Included ATAC-seq peak summit (<100 bp)
- Excluded ATAC-seq peak summit (>100 bp)

**Legend**

- n.s.
- * p < 0.05
- **p < 0.01
- *** p < 0.001

**Graphs**

- Probability of an element showing MPRA activity
- Density
- Median distance from islet ATAC-seq peak summit to sequence center
- # Elements
Figure 4

(a) Absolute (median) distance of SNP to ATAC-seq peak summit

(b) Probability of an element showing MPRA activity

(c) Probability an MPRA active element has allelic skew in activity

(d) MPRA Activity log₂ (Higher Allele / Lower Allele)

Control SNPs +
caQTL SNPs +
MPRA Active +

Probability of an element showing MPRA activity
# MPRA Active
# Elements =

707 2218 778 1910

Probability an MPRA active element has allelic skew in activity
# AllelicSkew
# MPRA Active =

283 707 376 778

MPRA Activity log₂ (Higher Allele / Lower Allele)

Control SNPs (n=283) +
caQTL SNPs (n=376) +

(n=283)

(n=376)
**Figure 5**

*a* Pearson R = 0.526  ***

**b**

N=24[+2]  
N=111

rs17396537  
GG n=6  
GC n=11  
CC n=2

rs1515555  
rs10428126  
IGF2BP2

rs17396537  
ATAC-seq Peak  
Summit  
200 bp element

rs17396537  
0.0 0.2 0.4
G
REF
C
ALT
log2(RNA/DNA)
*

rs17396537  
GG  
GC  
CC
n=2  
n=11  
n=6

(G/C)

Summit
89

rs17396537

rs1515555

rs10428126

rs2943656

rs1515556

G
REF
C
ALT

log2(RNA/DNA)
*

rs1515556

GG  
GC  
CC

n=2  
n=11  
n=6

(G/C)

Summit
89

Ref > Alt  
Ref < Alt

caQTL effect size

MPRA log2 (Alt / Ref)
Change in MPRA activity under ER stress of SNPs with allelic Skew

- **(n=110)
- *(n=31)
- +(n=86)
- *(n=67)
- n.s.

Enhancer Activity under ER stress of T2D-associated SNPs with Allelic skew

- Higher Activity under ER stress
- Higher Activity

 SNP with Allelic skew are conserved (Sequences are conservation if sequence similarity > 20%)

- caQTLs (n=376)
- Controls (n=283)
<table>
<thead>
<tr>
<th>IndexSNP</th>
<th>Locus</th>
<th>SNP</th>
<th>REF_ALLELE</th>
<th>ALT_ALLELE</th>
<th>MAF_AFR</th>
<th>MAF_AMR</th>
</tr>
</thead>
<tbody>
<tr>
<td>rs7903146</td>
<td>TCF7L2</td>
<td>rs7903146</td>
<td>C</td>
<td>T</td>
<td>0.28</td>
<td>0.25</td>
</tr>
<tr>
<td>rs9472138</td>
<td>VEGFA</td>
<td>rs12189774</td>
<td>G</td>
<td>A</td>
<td>0.19</td>
<td>0.21</td>
</tr>
<tr>
<td>rs3842770</td>
<td>INS-IGF2</td>
<td>rs115420895</td>
<td>G</td>
<td>A</td>
<td>0.06</td>
<td>0.01</td>
</tr>
<tr>
<td>rs7578326</td>
<td>IRS1</td>
<td>rs2943656</td>
<td>A</td>
<td>G</td>
<td>0.6</td>
<td>0.73</td>
</tr>
<tr>
<td>rs10507349</td>
<td>RNF6</td>
<td>rs4630391</td>
<td>C</td>
<td>T</td>
<td>0.08</td>
<td>0.36</td>
</tr>
<tr>
<td>rs1535500</td>
<td>KCNK16</td>
<td>rs12663159</td>
<td>C</td>
<td>A</td>
<td>0.86</td>
<td>0.49</td>
</tr>
<tr>
<td>rs864745</td>
<td>JAZF1</td>
<td>rs1635852</td>
<td>T</td>
<td>C</td>
<td>0.22</td>
<td>0.4</td>
</tr>
<tr>
<td>rs864745</td>
<td>JAZF1</td>
<td>rs849133</td>
<td>C</td>
<td>T</td>
<td>0.21</td>
<td>0.4</td>
</tr>
<tr>
<td></td>
<td></td>
<td>rs10811660</td>
<td>G</td>
<td>A</td>
<td>0.05</td>
<td>0.14</td>
</tr>
<tr>
<td>rs10811661</td>
<td>CDKN2B-AS1</td>
<td>rs10811661</td>
<td>T</td>
<td>C</td>
<td>0.05</td>
<td>0.14</td>
</tr>
<tr>
<td></td>
<td></td>
<td>rs10965246</td>
<td>T</td>
<td>C</td>
<td>0.07</td>
<td>0.14</td>
</tr>
<tr>
<td></td>
<td></td>
<td>rs10428126</td>
<td>T</td>
<td>C</td>
<td>0.6</td>
<td>0.29</td>
</tr>
<tr>
<td></td>
<td></td>
<td>rs9808924</td>
<td>G</td>
<td>A</td>
<td>0.8</td>
<td>0.32</td>
</tr>
<tr>
<td>rs1470579</td>
<td>IGF2BP2</td>
<td>rs9854769</td>
<td>A</td>
<td>G</td>
<td>0.8</td>
<td>0.31</td>
</tr>
<tr>
<td></td>
<td></td>
<td>rs7630554</td>
<td>A</td>
<td>G</td>
<td>0.79</td>
<td>0.31</td>
</tr>
<tr>
<td></td>
<td></td>
<td>rs35188816</td>
<td>CA</td>
<td>C</td>
<td>0.78</td>
<td>0.31</td>
</tr>
<tr>
<td>rs1048886</td>
<td>SDHAF4</td>
<td>rs2881632</td>
<td>C</td>
<td>A</td>
<td>0.4</td>
<td>0.22</td>
</tr>
<tr>
<td>rs12496230</td>
<td>KBTBD8</td>
<td>rs13096599</td>
<td>T</td>
<td>C</td>
<td>0.02</td>
<td>0.2</td>
</tr>
<tr>
<td>rs6467136</td>
<td>GCC1</td>
<td>rs7783500</td>
<td>G</td>
<td>A</td>
<td>0.54</td>
<td>0.4</td>
</tr>
<tr>
<td>rs10923931</td>
<td>NOTCH2</td>
<td>rs72697237</td>
<td>A</td>
<td>T</td>
<td>0.36</td>
<td>0.12</td>
</tr>
<tr>
<td>rs7578597</td>
<td>THADA</td>
<td>rs13405776</td>
<td>C</td>
<td>T</td>
<td>0.34</td>
<td>0.07</td>
</tr>
<tr>
<td>rs7630877</td>
<td>PEX5L</td>
<td>rs17748864</td>
<td>C</td>
<td>T</td>
<td>0.24</td>
<td>0.27</td>
</tr>
<tr>
<td>rs7957197</td>
<td>OASL</td>
<td>rs7957197</td>
<td>T</td>
<td>A</td>
<td>0.17</td>
<td>0.15</td>
</tr>
<tr>
<td>rs34669198</td>
<td>SRBD1</td>
<td>rs13026123</td>
<td>T</td>
<td>A</td>
<td>0.09</td>
<td>0.09</td>
</tr>
</tbody>
</table>
### Allele with higher MPRA activity under

<table>
<thead>
<tr>
<th>MAF_ASN</th>
<th>MAF_EUR</th>
<th>Standard Culture</th>
<th>DMSO</th>
<th>250 nM TG</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.03</td>
<td>0.31</td>
<td>-</td>
<td>T</td>
<td>-</td>
</tr>
<tr>
<td>0.11</td>
<td>0.29</td>
<td>-</td>
<td>G</td>
<td>G</td>
</tr>
<tr>
<td>0</td>
<td>0.0013</td>
<td>A</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>0.94</td>
<td>0.62</td>
<td>-</td>
<td>G</td>
<td>G</td>
</tr>
<tr>
<td>0.44</td>
<td>0.24</td>
<td>C</td>
<td>C</td>
<td>C</td>
</tr>
<tr>
<td>0.43</td>
<td>0.49</td>
<td>A</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>0.21</td>
<td>0.54</td>
<td>-</td>
<td>C</td>
<td>-</td>
</tr>
<tr>
<td>0.21</td>
<td>0.53</td>
<td>C</td>
<td>C</td>
<td>-</td>
</tr>
<tr>
<td>0.44</td>
<td>0.16</td>
<td>-</td>
<td>G</td>
<td>-</td>
</tr>
<tr>
<td>0.44</td>
<td>0.16</td>
<td>-</td>
<td>C</td>
<td>-</td>
</tr>
<tr>
<td>0.44</td>
<td>0.17</td>
<td>C</td>
<td>C</td>
<td>-</td>
</tr>
<tr>
<td>0.28</td>
<td>0.29</td>
<td>T</td>
<td>T</td>
<td>-</td>
</tr>
<tr>
<td>0.29</td>
<td>0.29</td>
<td>A</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>0.29</td>
<td>0.29</td>
<td>G</td>
<td>G</td>
<td>G</td>
</tr>
<tr>
<td>0.29</td>
<td>0.28</td>
<td>G</td>
<td>G</td>
<td>G</td>
</tr>
<tr>
<td>0.28</td>
<td>0.27</td>
<td>-</td>
<td>CA</td>
<td>-</td>
</tr>
<tr>
<td>0.12</td>
<td>0.19</td>
<td>A</td>
<td>A</td>
<td>A</td>
</tr>
<tr>
<td>0.08</td>
<td>0.18</td>
<td>C</td>
<td>C</td>
<td>C</td>
</tr>
<tr>
<td>0.67</td>
<td>0.42</td>
<td>G</td>
<td>G</td>
<td>G</td>
</tr>
<tr>
<td>0.03</td>
<td>0.11</td>
<td>T</td>
<td>T</td>
<td>T</td>
</tr>
<tr>
<td>0.01</td>
<td>0.09</td>
<td>T</td>
<td>T</td>
<td>T</td>
</tr>
<tr>
<td>0.15</td>
<td>0.34</td>
<td>T</td>
<td>T</td>
<td>T</td>
</tr>
<tr>
<td>0</td>
<td>0.2</td>
<td>T</td>
<td>T</td>
<td>T</td>
</tr>
<tr>
<td>0.1</td>
<td>0.13</td>
<td>T</td>
<td>T</td>
<td>T</td>
</tr>
<tr>
<td>References</td>
<td>Comments</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>-----------------------------</td>
<td>-------------------------------</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Stitzel 2010, Gaulton 2010</td>
<td>Validated with MPRA</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Miguel-Escalada 2019</td>
<td>Validated with MPRA</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Khetan 2018</td>
<td>Validated with MPRA</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Khetan 2018</td>
<td>Validated with MPRA</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Khetan 2018</td>
<td>MPRA has further clarified haplotype</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Varshney 2017</td>
<td>MPRA has further clarified haplotype</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Fogarty 2013</td>
<td>Validated with MPRA</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Wesolowska-Andersen 2019, bioRxiv</td>
<td>Validated with MPRA</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Khetan 2018</td>
<td>Validated with MPRA</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>WW Greenwald 2019</td>
<td>Validated with MPRA</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Novel
Chapter 4

Discussion & Future Directions
T2D-Associated Genetic Variants Regulate Chromatin Accessibility in Human Islets.

T2D results when pancreatic beta cells are unable to secrete adequate amounts of insulin in response to elevated blood glucose levels. Two prominent changes in islets have been found to correlate with the T2D disease-state, a decline in beta cell numbers\(^1\)\(^-\)\(^4\), and a loss of beta cell identity\(^5\)\(^-\)\(^7\). We hypothesized that either of these changes in islets from T2D donors will be reflected as changes in chromatin accessibility. However, when we compared islet samples from diabetic and non-diabetic donors, few differences in chromatin accessibility were found. This may be due to:

1. The small size of our cohort. Larger and more homogenous cohorts are needed to comprehensively document the changes in islet chromatin accessibility associated with the T2D disease-state.

2. The heterogeneous composition of islets. Islets are composed of at least 5 different types of cells, alpha, beta, delta, gamma and epsilon. Bulk islet ATAC-seq measurements could be masking subtle changes in the relative proportion of these cell-types, or their identity\(^8\). Single cell ATAC-seq measurements will be important to document both changes in relative cell proportions and in cell-type specific chromatin accessibility patterns due to the T2D disease-state.

While cross-sectional studies documenting changes in T2D islets have provided useful insights into the cellular mechanisms possibly causing islet dysfunction, they are insufficient to understand why beta cell numbers decline in the first place, or why some individuals are more at risk for islet dysfunction and failure than others. To answer these questions, it is important to understand how genetic differences between individuals in the population affect islet function.
Therefore, in a cohort of 19 cadaveric donors, we employed a quantitative trait loci (QTL) approach\textsuperscript{9} to identify 2,949 SNPs, genotypes at which were significantly correlated with changes in islet chromatin accessibility. These islet caQTLs included SNPs at 13 T2D-associated GWAS signals, suggesting changes in regulatory element activity by T2D-associated SNPs are reflected as changes in islet chromatin accessibility. It will be important to perform islet caQTL studies with saturating sample sizes to obtain an exhaustive list of T2D-associated GWAS SNPs affecting islet chromatin accessibility. Since a significant fraction of ATAC-seq reads failed to map to the human genome, largely owing to repetitive elements, it will also be important that Illumina sequencing of ATAC-seq fragments be performed using longer read lengths (‘2 X 150 bp’ instead of ‘2 X 75 bp’) in the future to refine our understanding of how repetitive elements affect islet chromatin accessibility and regulatory element activity. However, several questions would still remain unanswered. We discuss them below, along with strategies for addressing some of them:

1. Do islet caQTL T2D-associated SNPs affect regulatory element activity in beta cells exclusively, or are they active in other cell types comprising islets as well? Cell-type specific ATAC-seq profiles are again required to answer this question since islets are composed of at least 5 different types of cells, and it is unclear which cell-types are affected by SNPs identified as islet caQTLs.

2. What are the TF binding motifs disrupted by islet caQTL T2D-associated SNPs? While high-throughput approaches are not yet feasible, low throughput electrophoretic mobility shift assays (EMSA) to identify TF binding motifs disrupted by SNPs will be important to characterize the upstream signaling pathways.

3. What are the genes whose expression is affected by islet caQTL T2D-associated SNPs? To answer this question, islet caQTL approaches will need to be complemented by 3D chromatin interaction maps and eQTL approaches in islets.
One limitation of islet caQTL approaches is that they identify correlative associations. Another limitation is that a large fraction of islet caQTLs (>30%) are characterized by multiple linked SNPs in the same ATAC-seq peak, and it is unclear whether 1 or more SNPs among them are functional. We used a high throughput MPRA\textsuperscript{10} to overcome these limitations by directly testing islet caQTL SNPs for allelic differences in beta cell regulatory element activity. We found that SNPs significantly correlated with islet chromatin accessibility changes, i.e., caQTLs, were highly predictive of changes in regulatory element activity. Moreover, MPRA was effective in identifying functional SNPs among neighbors in LD at a given ATAC-seq peak. Therefore, MPRA and islet caQTL are complementary approaches to identify functional T2D-associated SNPs, and characterize them for their in vivo effects on the islet epigenome.

**T2D-associated SNPs reside in regulatory elements responsive to changes in the ER load of beta cells.**

In Chapter 2, SNPs at 13 T2D-associated GWAS signals were correlated to changes in islet chromatin accessibility. While islet caQTL studies with saturating sample sizes will be important to know how many T2D-associated SNPs in total affect islet chromatin accessibility under normoglycemic conditions, there is data to already suggest that the answer is not going to be many more than 30 because:

1. Even though a comprehensive set of 154,437 open chromatin regions were identified using ATAC-seq on islet samples from 19 cadaveric donors, only 30 out of 220 T2D-associated SNPs identified as functional in beta cells overlapped islet ATAC-seq peaks, and
2. Functional T2D-associated SNPs that did not overlap islet ATAC-seq peaks were significantly more likely to have higher MPRA activity under ER stress.

Therefore, our data suggests that the majority of T2D-associated SNPs reside in regulatory elements responsive to changes in the cellular environment of beta cells, such as ER stress, which we and others identified as an important environmental risk factor to unravel the genetic risk for T2D\(^1,11^-13\). Uncompensated ER stress was found to lead to the inactivation of beta cell-specific TFs causing the downregulation of insulin transcription and secretion\(^14^-16\). ER folding and trafficking capacity, therefore, may be a major factor determining how much insulin can be released by beta cells before stress ensues. SNPs at regulatory elements modulating the expression of genes relevant to meet higher demands for insulin synthesis may impair or improve ER capacity, ultimately determining the threshold at which ER stress ensues, leading to beta cell failure and T2D.

However, our MPRA study, in turn, suffers from several limitations as well. We discuss them below, along with strategies for overcoming them:

1. The current limits on oligo synthesis technologies allow MPRA to only test 200 bp genomic regions. As shown in Chapter 3, when the 200 bp genomic region tested with MPRA excluded ATAC-seq peak summits, the probability of observing MPRA was significantly decreased. Therefore, it will be important for future studies to test longer genomic loci, either by tiling the genomic sequence flanking a SNP, or by using longer oligos as and when such technologies become available.

2. MPRA is an episomal assay, in which the test sequence is fully accessible to TFs. This fails to capture the nucleosomal dynamics that play important roles in transcriptional regulation. Future MPRA studies should be complemented by lenti-MPRA...
investigations\textsuperscript{17,18} to understand whether pioneer factors and chromatin remodelers are affected by T2D-associated SNPs.

3. MPRA in Chapter 3 studied human sequences in a mouse beta cell-line. This was because the growth and transfection conditions in human beta cell-lines, EndoC-ßH1 and -ßH3, were optimized long after the MPRA studies in Chapter 3 were initiated, optimized, and completed. While these human beta cell-lines are still slow to grow (doubling time: \( \sim 7 \) days) and do not transflect robustly (\( \sim 10\text{-}15\% \) in our hands), making it difficult to perform MPRA, it will be important for future studies to study human sequences in human cell-lines.

4. Since islets are composed of at least 5 different types of cells, it will be important to assess MPRA activity of sequences in the other cell-types that comprise islets, such as alpha and delta cells. Additional testing of sequences in non-islet tissues/cell-types, such as adipose, skeletal muscle and liver, will be important to assess the involvement of these diverse cell-types in the genetic risk of T2D.

5. Finally, since caQTL approaches require islets from \( \geq 100 \) individuals, wherein power is still highly dependent on minor allele frequency, we also used MPRA to show that ER stress is an important cellular context under which regulatory elements harboring T2D-associated SNPs become active. However, it will be important to characterize chromatin accessibility in islets exposed to ER stress, and correlate resulting changes to SNPs in the human population. Characterizing such ER stress response islet caQTL will be critical to link T2D-associated SNPs to \textit{in vivo} changes in the islet epigenome.
Future directions

We have made significant contributions in understanding how non-coding variants modulate risk for T2D. Using caQTL approaches, SNPs at 13 T2D-associated GWAS signals were correlated to changes in islet chromatin accessibility. Using MPRA, 220 functional SNPs at 104 distinct T2D-associated GWAS signals were identified. While a single SNP was nominated as the putative causal variant for 54 GWAS signals, >1 functional SNPs were detected for 50 T2D association signals, suggesting T2D risk from individual GWAS signals may result from a combined effect of multiple functional SNPs. Given that T2D is a complex disorder with both genetic and environmental risk factors, we also successfully identified ER stress in beta cells as an important environmental context for the genetic risk for T2D. Finally, we have implicated a role of SNPs at SINEs in mediating both ER stress in beta cells and the genetic risk of T2D. However, there are several challenges that remain in understanding how non-coding variants modulate risk for a complex disorder, such as T2D:

1. When T2D risk from individual GWAS signals result from a combined effect of multiple functional SNPs, do the multiple functional SNPs all affect the same target gene? Do the multiple functional SNPs at a given GWAS signal reside in regulatory elements responsive to different cellular environments? A combination of response islet caQTL and eQTL approaches, along with MPRA is required to answer these questions.

2. Why do small changes in the expression of target genes lead to a higher risk for T2D? Elucidating functions of these target genes in islets, under appropriate cellular environments, such as ER stress, will be critical to answer this question.
3. RNA from SINE elements are transcriptionally induced by cellular stress, including thermal and ER stress, and the corresponding SINE RNAs have been shown to function as critical transcriptional switches during stress\textsuperscript{19}. Using MPRA, we implicated SNPs in SINE elements for their roles in mediating changes to both the ER stress response in beta cells, and the genetic risk of T2D. However, since MPRA cannot distinguish enhancers from promoters, further studies are needed to understand how exactly do SNPs at SINEs affect transcriptional regulation in islets.

4. In addition to ER stress, various other forms of cellular stress in beta cells, such as inflammatory\textsuperscript{20,21}, oxidative\textsuperscript{22,23} and xenobiotic\textsuperscript{24,25} stress, have been linked to a higher risk for T2D. Understanding how islets respond to these cellular stresses will be critical to have a complete understanding of the influence of SNPs in increasing or decreasing tolerance to them to modulate risk for T2D.

This thesis has significantly advanced our understanding of T2D genetic risk by identifying functional SNPs in beta cells, and linking them to activity under ER stress. We hope that future studies will use these insights to better understand the genetic risk of T2D.

References:


