Universitat Oberta de Catalunya

uoc.edu

CRG Centre for Genomic Regulation

Leveraging single-cell ATAC-seq data to gain insights into the cell-type selective component of the human pancreatic islet regulome

Autor: Andrés Castillo Bonilla

Máster universitario de Bioinformática y Bioestadística

Área 3

Director UOC: Dr. Diego Garrido Martín

Director CRG: Dr. Jorge Ferrer

Fecha Entrega: 8 de junio de 2021

FICHA DEL TRABAJO FINAL

Título del trabajo:	*Leveraging single-cell ATAC-seq data to gain insights into the cell-type selective component of the human pancreatic islet regulome*
Nombre del autor:	*Andrés Castillo Bonilla*
Nombre del consultor/a UOC:	*Diego Garrido Martín*
Nombre del consultor/a CRG:	*Jorge Ferrer*
Nombre del PRA:	*Laura Calvet Liñan*
Fecha de entrega (mm/aaaa):	06/2021
Titulación:	*Máster universitario en Bioinformática y Bioestadística UOC-UB*
Área del Trabajo Final:	*Bioinformática y Bioestadística Área 3*
Idioma del trabajo:	Ingés
Número de créditos:	15
Palabras clave	*single-cell ATAC-seq, human pancreatic islet, genome regulation*

Resumen del Trabajo (máximo 250 palabras): *Con la finalidad, contexto de aplicación, metodología, resultados i conclusiones del trabajo.*

ATAC-seq es esencial para perfilar la accesibilidad de la cromatina y caracterizar el panorama regulatorio transcripcional. Sin embargo, el reciente interés por el estudio de poblaciones celulares heterogéneas constituye un desafío para el ATAC-seq. Por lo tanto, el ATAC-seq unicelular surge como una respuesta a las limitaciones de ATAC-seq en masa cuando se estudia la heterogeneidad celular. Nuestro objetivo es caracterizar el componente de tipo celular de los

potenciadores utilizando datos scATAC-seq. Para lograr este propósito, a) anotamos potenciadores en regiones de cromatina abierta de tipo celular, b) estimamos el enriquecimiento de motivos entre potenciadores de tipo celular, c) detectamos potenciadores accesibles de tipo celular que muestran una unión robusta a TF y d) identificamos variantes asociadas a T2D que afectan la unión de TF a potenciadores de tipo celular. El análisis de enriquecimiento de motivos presentó grupos bien definidos de motivos enriquecidos en potenciadores de tipo celular. La recurrencia de motivos a través de potenciadores de tipo celular mostró que los potenciadores unidos a un TF dado eran consistentes con el agrupamiento de tipo celular observado en el análisis de enriquecimiento de motivos. Finalmente, la integración de los potenciadores de tipo celular que caracterizan la unión de TF con variantes genéticas de T2D nos permitió proponer el mecanismo molecular más probable subyacente a algunos loci de riesgo de T2D.

Abstract (in English, 250 words or less):

ATAC-seq is essential for profiling chromatin accessibility and characterizing the transcriptional regulatory landscape. However, the recent shift towards the study of heterogeneous cell populations poses a challenge for bulk ATAC-seq. Thus, single-cell ATAC-seq has emerged as a response to the limitations of bulk ATAC-seq when studying cellular heterogeneity. We aim to to characterize the cell-type-selective component of enhancers using scATAC-seq data. To achieve this purpose, we a) annotate regulome signatures across cell-type selective open chromatin regions, b) estimate TF motif enrichment among cell-type selective enhancers, c) detect accessible cell-type selective enhancers that show robust TF binding and d) identify T2D-associated SNPs affecting TF binding across cell-type selective enhancers. Motif enrichment analysis presented well-defined groups of TF motifs enriched across islet cell-type selective enhancers. TF motif occurrences across cell-type selective enhancers showed that enhancers bound by a given TF was consistent with the cell-type selective clustering observed in the TF motif enrichment analysis. Finally, the integration of TF-binding characterizing islet cell-type enhancers with fine-mapped T2D genetic variants allowed us to propose the most likely molecular mechanism underlying a few T2D risk loci.

Contents

1. Introduction

1.1. Context and justification

Bulk assay for transposase accessible chromatin sequencing (ATAC-seq) measurements offer comprehensive profiles of chromatin accessibility in a tissue-specific manner (Reddington et al., 2020). However, they are limited to disentangle tissue heterogeneity and the contribution of restricted cell-types into the regulatory landscape of human pancreatic islets. This is because bulk ATAC-seq produces aggregated profiles by averaging the signal over cell populations, masking cellular and regulatory heterogeneity (Rai et al., 2020; Shema, Bernstein, & Buenrostro, 2019). Consequently, the elevated cell-type heterogeneity of pancreatic islets could hinder the identification of accessible regulatory elements that can otherwise be identified with single cell ATAC-seq (scATAC-seq) data.

In this project we address islet cellular heterogeneity by characterizing the cell-type component of the human islet regulatory landscape using scATAC-seq data from beta, alpha, delta and acinar cells. When studying heterogeneous biological samples such as human pancreatic islets, single-cell analysis enables the identification of cell-type populations and regulatory elements (Buenrostro et al., 2018; Rai et al., 2020). This provides larger resolution to advance the molecular understanding of transcriptional regulation in tissues with large cellular complexity as human pancreatic islets. Thus, by adopting a cell-type-specific approach to examine distal regulatory elements in human islets, we could gain insights into the cellular diversity and gene regulatory mechanisms. Furthermore, characterizing the cell-type specific regulatory landscape of human islets not only allows elucidating the role that each cell type plays in the physiology of human islets, but it also offers a single-cell resolution view of metabolic disorders such as type 2 diabetes (T2D) (Chiou et al., 2019). Ultimately, we expect to obtain single-cell regulatory profiles to elucidate both the relationship between cell types and their contribution to pancreatic islet transcriptional regulation.

1.2. Objectives

General objective

Our general objective is to characterize the cell-type-selective component of distal human pancreatic islet regulatory elements, “enhancers”, using scATAC-seq data.

Specific objectives

1) To annotate regulome signatures identified in human islets among open chromatin regions in endocrine (beta, alpha and delta) and exocrine (acinar) cell types.

2) To estimate transcription factor (TF) motif enrichment in open chromatin regions for each islet cell type.

3) To leverage abovementioned motif enrichment analysis to detect cell-type selective distal regulatory regions, known as enhancer elements, that show robust TF binding.

4) To integrate genetic data from large-scale genetic association studies (GWAS) for T2D to identify variants associated with T2D risk that are likely to disrupt TF binding across cell-type specific enhancers identified in (3).

1.3. Approach and method to follow

In order to achieve the purpose of this project, we will leverage cell-type selective open chromatin regions identified in the host lab using unpublished scATAC-seq data from human pancreatic islets from one donor that were exposed at high glucose concentrations (11 mM). Sequencing was performed employing the 10x Genomics Chromium Single Cell ATAC platform. Data processing, identification and annotation of cell clusters, and peak differential analysis were performed by the host lab, using 10x Genomics Cell Ranger ATAC 1.2.0 (Satpathy et al., 2019; Zheng et al., 2017) for pre-processing and Signac 1.1.1 (Stuart, Srivastava, Lareau, & Satija, 2020), an extension of Seurat 4.0 (Hao et al., 2020), for downstream analyses. The resulting single-cell open-chromatin peaks are the starting point of our project.

To delineate single-cell regulatory profiles in the human pancreatic islet regulatory landscape we performed as follows by: a) annotating regulome signatures across cell-type selective open chromatin regions by overlapping scATAC-seq peaks with the Miguel-Escalada et al (2019) human islet regulome using BEDTools 2.30.0 (Quinlan & Hall, 2010). b) Estimating TF motif enrichment among cell-type selective enhancers using HOMER 4.11 (Heinz et al., 2010). c) Detection of accessible cell-type selective enhancers that show robust TF binding with FIMO 5.3.3 (Grant, Bailey,

& Noble, 2011). d) Identification of T2D-associated single-nucleotide variants (SNPs) affecting TF binding across cell-type selective enhancers using motifbreakR 2.4.0 (Coetzee, Coetzee, & Hazelett, 2015). To this end, we will also leverage fine-mapped variants from a large-scale T2D meta-analysis, generated by Mahajan et al (2018).

1.4. Work plan

Tasks and milestones

1. Work planning
    1.1. Contextualize and justify the project.
    1.2. Define the objectives.
    1.3. Outline the approach and the methodology to follow.
    1.4. Plan project milestones and timing.
    1.5. Write and submit CAT1.
2. Work development - characterization of single-cell regulatory profiles.
    2.1. Annotate regulome signatures across cell-type selective open chromatin regions.
    2.2. Perform motif enrichment analysis.
    2.3. Write and submit CAT2.
    2.4. Detect cell-type selective enhancers with robust TF binding.
    2.5. Identify variants affecting TF binding across cell-type selective enhancers.
    2.6. Write and submit CAT3.
3. Manuscript drafting and submission.
    3.1. Write the introduction and state of the art.
    3.2. Write the methodology and results.
    3.3. Write the discussion and conclusions.
    3.4. Last review and submission.
4. Project defense preparation.

Project schedule

	Tasks and milestones	Date	Weeks																
			1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17
CAT1	**1. Work planning**	01/03/21 - 16/03/21																	
	1.1 Contextualize and justify the project.	01/03/21 - 03/03/21																	
	1.2 Define the objectives.	04/03/21 - 06/03/21																	
	1.3 Outline the approach and the methodology to follow.	07/03/21 - 09/03/21																	
	1.4 Plan project milestones and timing.	10/03/21 - 12/03/21																	
	1.5 Write and submit CAT1.	13/03/21 - 16/03/21																	
CAT 2 & 3	**2. Work development - c**haracterization of single-cell regulatory profiles.	17/03/21 - 17/05/21																	
	2.1 Annotate regulome signatures across cell-type specific open chromatin regions.	17/03/21 - 04/04/21																	
	2.2 Perform motif enrichment analysis.	29/03/21 - 19/04/21																	
	2.3 Write and submit CAT2.	05/04/21 - 19/04/21																	
	2.4 Detect cell-type specific enhancers with robust transcription factor binding.	19/04/21 - 25/04/21																	
	2.5 Identify variants affecting transcription factor binding across cell-type specific enhancers.	26/04/21 - 02/05/21																	
	2.6 Write and submit CAT3.	03/05/21 - 17/05/21																	
CAT4	**3. Manuscript drafting and submission.**	10/05/21 - 08/06/21																	
	3.1 Write the introduction and state of the art.	10/05/21 - 16/05/21																	
	3.2 Write the methodology and results.	17/05/21 - 23/05/21																	
	3.3 Write the discussion and conclusions.	24/05/21 – 30/05/21																	
	3.4 Last review and submission.	31/05/21 - 08/06/21																	
CAT5	**4. Project defense preparation.**	31/05/21 - 23/06/21																	

1.5. Summary of products obtained

From our project we obtained and present in this manuscript the following results:

1) Sub-classification of the islet regulome according to cell-type selective open-chromatin regions.
2) Quantification of TF motif enrichments in cell-type selective clusters of enhancer elements.
3) Identification of cell-type selective accessible enhancer elements that show robust TF binding.
4) Identification of disease-associated genetic variants that are likely to disrupt TF binding across cell-type selective enhancer elements.

1.6. Brief description of the other chapters of the manuscript

Chapter 2: Introduces the state of the art or the level of development of the project topic.

Chapter 3: Describes the methodology followed throughout the project development.

Chapter 4: Presents the results obtained with this research proposal.

Chapter 5: Discusses the results within the context of the project and whether they meet the initial objectives or not, and future related research. In this section, limitations are also addressed.

Chapter 6: Enumerates the most relevant results and conclusions derived from this study.
Chapter 7: Glossary with the most relevant terms and acronyms used within the manuscript.
Chapter 8: List of references cited throughout the manuscript.
Chapter 9: Acknowledgments.
Chapter 10: Contains information that is not included in the manuscript's main body due to their extension and relevance.

2. State of the art

Nearly two decades after the human genome sequence was sequenced, many questions remain unanswered of how non-coding DNA directs spatial and temporal activation of gene expression. However, recent advances in regulatory genomics, also with the establishment of large consortia, has delivered comprehensive catalogues of non-coding gene regulatory elements along with the parallel development of novel genomic technologies (Andersson et al., 2014; ENCODE Project Consortium, 2012; FANTOM Consortium and the RIKEN PMI and CLST (DGT) et al., 2014). This progress in the understanding of the dynamic usage of non-coding transcriptional regulatory elements is essential to gain insights into development, cell and tissue function and identity (Shlyueva, Stampfel, & Stark, 2014), and disease pathophysiology (Chatterjee & Ahituv, 2017; Maurano et al., 2012; Miguel-Escalada, Pasquali, & Ferrer, 2015).

Eukaryotic DNA is organized in the cell nucleus into chromatin, which preserves and compacts the genetic information but also controls gene expression (Klemm, Shipony, & Greenleaf, 2019; Wolffe, 2000). Chromatin is highly compacted into structural units named "nucleosomes", formed by DNA wrapped around a histone octamer core, enabling the genome to be assembled into the cell's small nucleus. Nucleosome occupancy across the genome defines chromatin accessibility, which precedes transcription of the human genome, and varies between cell types and tissues (Kaplan et al., 2009); e.g. a low nucleosome occupancy (nucleosome-depleted regions) translates into a high chromatin accessibility. Chromatin accessible regions, also known as "open chromatin" regions, are hereby targeted by TFs, RNA polymerases and other structural proteins and co-factors that result in a higher-order

genome organization essential for gene transcriptional regulation. The accessibility of the chromatin is largely facilitated by diverse post-translational modifications of histone proteins that will identify the distinct purposes of active chromatin. Two of the main players that coordinate gene transcription are enhancers and promoters (Andersson & Sandelin, 2020). Promoter elements identify short genomic sequences at the vicinity of the transcription start sites (TSS), which initiate gene transcription (Shlyueva et al., 2014). Genomic regions that embody promoter regulatory elements are characterized by the enrichment of acetylation of histone H3 lysine 27 residues (H3K27Ac) and tri-methylation of histone H3 lysine 4 residues (H3K4me3) epigenomic signatures (Andersson & Sandelin, 2020; Shlyueva et al., 2014). In sharp contrast with promoter elements, transcriptional enhancers are ~300-1000 bp DNA fragments that are often located hundreds of kilobases away from their endogenous gene targets. The three-dimensional re-organization of the chromatin allows enhancers to loop to the promoter sequence of their target gene and thus, to guide gene expression activation (Kagey et al., 2010). Active enhancers are characterized by H3K27ac and H3K4me1 modifications in flanking nucleosome histones, among other hallmarks (Andersson & Sandelin, 2020; Kagey et al., 2010). To identify this repertoire of transcriptional regulatory elements, several experimental methods have been developed to map chromatin accessibility based on the susceptibility of these DNA fragments to enzymatic cleavage or methylation.

Assay for Transposase-Accessible Chromatin using sequencing (ATAC-seq), developed by Buenrostro et al. (2013) as an alternative to Micrococcal Nuclease sequencing (MNase-seq) (Schones et al., 2008), Formaldehyde-Assisted Isolation of Regulatory Elements sequencing (FAIRE-seq) (Giresi, Kim, McDaniell, Iyer, & Lieb, 2007) and DNase I hypersensitive sites sequencing (DNase-seq) (Boyle et al., 2008), is currently one of the most powerful and widely used chromatin accessibility profiling methods (Yan, Powell, Curtis, & Wong, 2020). ATAC-seq evaluates genome-wide DNA accessibility using a genetically engineered hyperactive enzyme known as Tn5 transposase (Reznikoff, 2008; Shashikant & Ettensohn, 2019). This enzyme simultaneously cuts DNA and inserts high-throughput sequencing adaptors, with preference for nucleosome-depleted chromatin regions. DNA fragments are then purified and amplified via PCR, creating DNA sequencing libraries that are enriched for open chromatin regions. These libraries are then sequenced by next generation sequencing (NGS). ATAC-seq data analysis then follows four major steps (Yan et al.,

2020); (1) The pre-analysis step, where reads are evaluated for quality assessment and aligned to the reference genome assembly. (2) The core analysis or peak calling, where regions with a high density of aligned reads are identified, indicating accessible regions which are also referred to as peaks. (3) Advanced downstream analysis, with the focus on peaks, motifs, nucleosomes, and TF footprints. Finally, (4) integrative multiomics approaches allows the characterization of the underlying regulatory networks.

The success of ATAC-seq is driven by the low-input requirements, the simplicity and sensitivity of ATAC-seq (Buenrostro et al., 2013; Buenrostro, Wu, Chang, & Greenleaf, 2015). Simultaneous chromatin fragmentation and insertion of sequencing adaptors executed by Tn5 transposase simplifies the experimental protocol, which can be completed in a single day. The method's high sensitivity enables it to accurately perform even on small cellular samples ranging from 500 to 50,000 cells.

ATAC-seq has proven to be an essential player for profiling chromatin accessibility and characterizing the genomic landscape of transcriptional regulatory elements such as enhancers and promoters (Buenrostro et al., 2013; Buenrostro, Wu, Chang, et al., 2015; Yan et al., 2020). However, our recent shift towards the study of heterogeneous cell populations posed a challenge for bulk chromatin accessibility assays. Bulk ATAC-seq can generate comprehensive chromatin accessibility profiles in a tissue-specific manner (Reddington et al., 2020), but the aggregated profiles that delivers by averaging the signal over cell populations masks cellular and regulatory heterogeneity (Rai et al., 2020; Shema et al., 2019). This hampers the understanding of how diverse individual cell types contribute to the regulatory networks maintaining cell and tissue homeostasis. Consequently, single-cell ATAC-seq (scATAC-seq) has emerged as a response to the limitations of bulk ATAC-seq when studying cellular heterogeneity (Buenrostro, Wu, Litzenburger, et al., 2015; Cusanovich et al., 2015).

Single-cell ATAC-seq allows the identification of chromatin accessibility and regulatory elements for thousands of single cells within and across cell-type populations (Baek & Lee, 2020; Buenrostro et al., 2018; Rai et al., 2020), with a wide range of available sequencing technologies (Baek & Lee, 2020; Buenrostro, Wu, Litzenburger, et al., 2015; Xi Chen, Miragaia, Natarajan, & Teichmann, 2018; Xingqi Chen et al., 2018; Cusanovich et al., 2015; Lareau et al., 2019; Mezger et al., 2018; Mulqueen et al., 2019; Rubin et al., 2019; Satpathy et al., 2018). Nevertheless, three main protocols are used to generate single-cell ATAC libraries after exposing

individual cells to Tn5 transposase (Baek & Lee, 2020; H. Chen et al., 2019). These include (1) barcoding individual single cells by the split-and-pool method allowing the identification of reads from each cell, (2) extraction and labeling of single cell DNA using microfluidic droplet-based technologies or (3) depositing single cells into a multi-well plate or array. Post-sequencing analyses (e.g., quality control, alignment and peak calling) are similar to those of bulk ATAC-seq, but they differ in other downstream analyses (Baek & Lee, 2020; Yan et al., 2020). Unlike bulk ATAC-seq, after the preprocessing of sequencing reads and the quality control, cells with good quality are selected to create a cell-by-feature matrix that is used for downstream analysis such as clustering, cell identity annotation, determination of differential accessibility, and estimation of regulatory networks (Baek & Lee, 2020; H. Chen et al., 2019).

Diversity is one of the most characterizing aspects of life, and as any other organ or tissue, the human pancreas is made up of diverse and highly specialized cell types. The largest fraction of the pancreatic tissue is embodied in the exocrine specialized tissue (exocrine acini ducts), formed by acinar cells. In contrast, the comparatively smaller endocrine compartment is confined in the islets of Langerhans and is essential to maintain blood glucose homeostasis (Segerstolpe et al., 2016). Hormone-secreting cells in the endocrine compartment are formed by glucagon-producing alpha cells, insulin-producing beta cells, somatostatin-producing delta cells, pancreatic polypeptide (PP) producing gamma cells, and ghrelin-producing epsilon cells (Chiou et al., 2019; Rai et al., 2020). ScATAC-seq can ease deciphering the cellular heterogeneity of the pancreatic tissue by characterizing their distinct regulatory profiles, delivering new clues about their contribution to the pancreatic function and identity (Baek & Lee, 2020; Buenrostro, Wu, Litzenburger, et al., 2015). Importantly, a single-cell resolution of gene regulatory mechanisms in human pancreatic islets has already prove to be a fertile ground to gain novel insights into the pathophysiology of metabolic disorders such as diabetes mellitus (Chiou et al., 2021).

Type 2 diabetes (T2D) is the most prevalent form of diabetes mellitus, a group of chronic metabolic disorders characterized by elevated blood glucose levels (International Diabetes Federation, 2019). Pancreatic islet dysfunction and insulin resistance are the two central pathological processes of the multifactorial nature of T2D (American Diabetes Association, 2020; McCarthy, 2010). Despite the highly polygenic inheritance of T2D (Mahajan et al., 2018a; Vujkovic et al., 2020), the large enrichment of T2D-predisposing genetic variants in islet regulatory annotations

highlights the central role of pancreatic islets in diabetes pathophysiology (Miguel-Escalada, Bonàs-Guarch, Cebola, Ponsa-Cobas, Mendieta-Esteban, Atla, Javierre, Rolando, Farabella, Morgan, García-Hurtado, et al., 2019; Pasquali et al., 2014; Thurner et al., 2018). However, the distinct role of endocrine cell types into T2D pathophysiology has not been extensively explored. Although more than 400 T2D-risk genetic variants have been identified in large-scale genetic association studies (GWAS) (Mahajan et al., 2018a; Vujkovic et al., 2020), the conversion to novel molecular insights has been limited. One of the main bottlenecks that frustrates the translation of GWAS genetic discoveries into molecular insights are the high amounts of (i) local linkage disequilibrium (LD) (that is, high correlation between neighbouring genetic markers). The identification of the true causal variant underlying a GWAS association is hereby hindered by high local LD between adjacent genetic markers (Schaid, Chen, & Larson, 2018). Statistical approaches, known as "fine-mapping", have been developed to overcome this limitation by identifying the minimum set of SNPs ("credible sets") with a 95-99% cumulative posterior probability of including the true causal variant (Wellcome Trust Case Control Consortium et al., 2012). However, the overwhelmingly majority of GWAS risk variants fall in non-coding regions and far away from coding sequences, which impairs the identification of an obvious target gene (Maurano et al., 2012). The integration of genome-wide maps of regulatory elements and chromatin interactions has been resourceful in aiding fine-mapping approaches to identify most likely causal regulatory variants and target genes (Miguel-Escalada, Bonàs-Guarch, Cebola, Ponsa-Cobas, Mendieta-Esteban, Atla, Javierre, Rolando, Farabella, Morgan, García-Hurtado, et al., 2019). Providing a single-cell perspective of gene regulation is now essential to refine the molecular interpretation of non-coding T2D risk GWAS associations.

This project focuses on leveraging scATAC-seq data in human pancreatic islets to characterize the cell-type-specific component of human pancreatic islet gene regulation. Our main goal is to obtain islet cell-type selective regulatory profiles by the integration of single-cell chromatin accessibility maps with islet regulome annotations to elucidate both the relationship between cell types and their contribution to the pancreatic islet-cell identity and function. Finally, by connecting single-cell epigenomic annotations with T2D GWAS results, we aim to reveal novel molecular insights into into T2D pathophysiology

3. Methodology

3.1. Data sources

In this project we leveraged five different datasets: (i) unpublished cell-type accessible chromatin peaks in human pancreatic islets, (ii) islet regulome annotations and (iii) enhancer-to-gene assignments identified in Miguel-Escalada et al (2019), (iv) unpublished cis-eQTLs mapped in 399 human pancreatic islet samples, and finally (v) fine-mapped variants from one of the largest meta-analysis for type 2 diabetes (Mahajan et al 2018).

Cell-type enriched and specific accessible chromatin peaks for beta, alpha, delta and acinar cells (see Figure 1) were identified by the host lab using unpublished human pancreatic islets scATAC-seq data from one donor sample. Cell-type enriched and specific peaks were identified after peak differential analysis; enriched peaks are more often open on a given cell-type but may also be open on other cell-types, and specific peaks are specifically open in a given cell-type and not open in the rest of the cell-types. In this project we primarily focus on enriched peaks since the low number of open chromatin regions specific for a given cell-type (see Figure 1) can limit the statistical power of our study.

Single cell sequencing libraries were generated from human pancreatic islets from a single donor that were exposed at high glucose concentrations (11 mM). Libraries were sequenced using the 10x Genomics Chromium Single Cell ATAC platform. Then, cell-type peaks of open chromatin regions were identified by the host lab after quality control, identification and annotation of cell-type clusters, and peak differential analysis. The host lab used 10x Genomics Cell Ranger ATAC 1.2.0 (Satpathy et al., 2019; Zheng et al., 2017) to demultiplex Illumina BCL files into FASTQ files, and Signac 1.1.1 (Stuart et al., 2020), an extension of Seurat 4.0 (Hao et al., 2020), for the rest of downstream analyses.

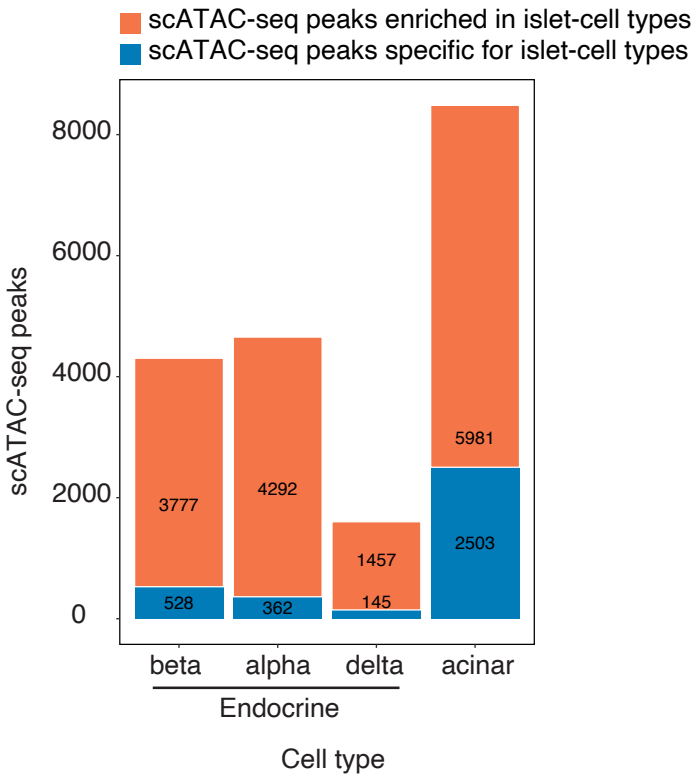


Figure 1 Barplot showing open chromatin peaks identified by the host lab using scATAC-seq, specific (in blue) or enriched (in orange) in a given cell-type in human pancreatic islets.

Annotations of the human pancreatic islet regulome were obtained from Miguel-Escalada et al (2019). This recent work from members of the host lab harnessed bulk ATAC-seq data to generate genome-wide maps of open chromatin regions in human pancreatic islets. Open chromatin regions were classified into distinct epigenome annotations, such as active promoters and active enhancers, by implementing k-medians clustering into chromatin immunoprecipitation (ChIP)-seq datasets including H3K27ac, H3K4me1, H3K4me3, Mediator, cohesin and CTCF. Active enhancers were subclassified into three categories I, II and III, based on Mediator, cohesin and H3K27ac occupancy patterns (from higher to lower activity, respectively). Of note, in this project we aggregated the three categories (active enhancers I, II and III) since bulk ATAC-seq can be hampered in capturing regulatory elements specific from minor cell populations that can otherwise be detected in scATAC-seq data. Active promoters were defined by H3K27ac and H3K4me3 marks but they have not been considered in our analysis.

Enhancer-gene assignments identified in Miguel-Escalada et al (2019) and unpublished islet cis-eQTLs were leveraged to assign target genes to cell-type selective enhancers. In Miguel-Escalada et al (2019) enhancer elements were first connected to their gene targets by leveraging a (i) high-resolution genome-wide map of chromatin interactions between islet gene promoters and their regulatory elements using promoter capture Hi-C (Javierre et al., 2016). Due to conservative detection thresholds or the limitation of Hi-C methods for short-range interactions, the authors (ii) imputed additional enhancer-gene assignments that were missed. Islet cis-eQTL mapping was performed by the host lab using QTLtools (Delaneau et al., 2017) in 399 human pancreatic islet samples using a cis-window of 500 kb up- and downstream of the TSS. Further details about RNA-seq processing, gene expression quantification, genotype QC and imputation will be provided in the manuscript in preparation by the host lab. In the linear model, 15 PCs derived from gene expression and 4 genetic PCs were used as covariates. Best associated cis eQTL SNP-eGene pairs, were identified using the permutation pass mode (--permute 1000 --window 500000). Beta approximated permutation p-values were adjusted for multiple testing correction using Storey q-values implemented in the qvalue R package (Storey, Bass, Dabney, Robinson, & Warnes, 2021) and significance threshold was set at FDR q-value ≤ 0.01 (3,433 eGenes FDR ≤ 1%). Nominal p-values for all cis-SNPs were calculated within a 500kb window centered on the TSS of each gene (--nominal 1 --window 500000). Significant variant-gene pairs were identified based on a genome-wide p-value threshold (pt) by considering the empirical p-value of the eGene closest to the 0.05 FDR threshold. A gene-based nominal p-value threshold was then calculated using pt and the beta distribution parameters from QTLtools. For 3,433 significant eGenes, variants with a nominal p-value below the gene-level threshold were considered in subsequent analyses (named from now on as nominally significant cis-eQTL variants). Nominally significant cis-eQTLs were intersected with islet enhancer elements (see Command 1). For overlapping islet eQTL-enhancer pairs, the eGene was assigned as the target gene that the enhancer is likely to regulate.

```
bedtools intersect –a <bulk islet regulome regions> –b <islet cis–eQTLs> –wa –wb
```

Command 1

Finally, fine-mapped genetic variants across 381 independent T2D signals identified in a large-scale meta-analysis for T2D (Mahajan et al 2018) in 898,130 individuals of European ancestry (74,124 T2D cases and 824,006 controls) were integrated with our cell-type selective epigenomic annotations.

3.2. Cell-type characterization of the islet regulome

To provide single-cell resolution to islet regulome annotations we followed four general steps: (1) we annotated regulome signatures by the overlap with cell-type selective open chromatin regions, (2) we estimated the enrichment of known TF motifs among islet cell-type selective enhancers, (3) we identified islet cell-type selective enhancers that show robust TF binding, and finally (4) we integrated genetic data to identify variants that are likely to disrupt TF binding across cell-type selective enhancers and TF binding regions identified in (3).

Regulome annotations were integrated with open chromatin regions that are selective for endocrine (beta, alpha and delta) and exocrine (acinar) cell types. This was accomplished by overlapping cell-type selective scATAC-seq peaks with the Miguel-Escalada et al (2019) human islet regulome using the intersect command (see Command 2) from the BEDTools 2.30.0 software (Quinlan & Hall, 2010). We grouped human islet regulome active enhancers I, II and III and used them jointly. Even though the distinct analyses performed in this study were based on islet cell-type enriched scATAC-seq peaks, we also annotated active enhancers across islet cell-type specific scATAC-seq peaks.

```
bedtools intersect –a <bulk regulome regions> –b <cell–type open chromatin peaks>
```

Command 2

Additionally, we integrated enhancer-to-gene assignments from Miguel-Escalada et al (2019) and significant nominal cis-eQTL variants from the host lab by overlapping them with cell-type selective active enhancer elements in human pancreatic islets using the intersect command from bedtools (see Command 3). The **-wa** and the **-wb** options were set to write the original entries of both intersecting files. This allowed us to assign a target gene based on the eQTL mapping and promoter capture Hi-C assignments to those overlapping cell-type selective enhancer elements.

Then we used enrichr (Xie et al., 2021) to search for ontologies, pathways and cell-types associated with genes regulated by islet cell-type selective enhancers.

```
bedtools intersect –wa –wb –a <enhancer–to–gene assignments> –b <cell–type annotated enhancers>
```

Command 3

TF motif enrichment analysis was performed among enhancers overlapping islet-selective scATAC-seq peaks using HOMER 4.11 (Heinz et al., 2010), which is based on a differential motif discovery algorithm. Motif enrichment was also estimated across the smaller fraction of enhancers overlapping islet cell-type specific open chromatin regions. This provided further support of the islet regulatory cell-type component revealed through open chromatin regions enriched in a given islet cell-type. To assess the robustness of the results, we used two different backgrounds for enrichment analysis, (i) the HOMER software default background and (ii) a custom background containing open chromatin regions that do not show enhancer epigenomic signatures. The HOMER default background is generated by selecting random regions from the genome until the total number of regions is 50000 or 2x the total number of peaks that are being analysed for each test. To execute the motif enrichment analysis, we used the HOMER findMotifsGenome.pl function (see Command 4). The fragment size used for motif finding was set to the exact size of the input regions with the option **-size given**, the UCSC human genome assembly hg19 was assigned as reference and the **-mask** option was used to mask out the repeat sequences in the genome. When the **-bg** option is not defined, HOMER selects the default background.

```
findMotifsGenome.pl <cell–type enhancer peaks> <genome> <output directory> –size given –mask –bg <background regions>
```

Command 4

For the identification of TF motif occurrences in enhancer elements, enhancers within cell-type selective scATAC-seq peaks were transformed into DNA sequences using the getfasta command (see Command 5) from the BEDTools 2.30.0 software (Quinlan & Hall, 2010). The hg19 reference genome was used to extract the sequences.

```
bedtools getfasta –fi <genome> –bed <cell–type enhancer peaks>
```

Command 5

The FIMO 5.3.3 software (Grant et al., 2011) was employed to identify cell-type selective enhancers that show robust transcription factor binding. To that end, individual known TF motif occurrences are scanned across islet cell-type selective enhancer DNA sequences obtained as aforementioned (see Command 6). We included transcription binding motifs, represented as position probability matrices, based on our previous TF motif enrichment results. The –parse-genomic-coord FIMO option was set to check for UCSC style genomic coordinates.

```
fimo –oc <output directory> ––parse–genomic–coord <motif file> <cell–type enhancer sequences>
```

Command 6

Islet cell-type selective enhancers with robust transcription factor binding were finally identified by intersecting (see Command 7) TF motif occurrences detected by FIMO with the annotated active enhancers across islet cell-type enriched peaks, resulting in TF motif-enhancer assignments.

```
bedtools intersect –wa –wb –a <FIMO TF motif coordinates> –b <cell–type annotated enhancer>
```

Command 7

Finally, we identified known TF-binding that is potentially disrupted by common single-nucleotide variants (SNVs or SNPs, from now on). MotifbreakR 2.4.0 (Coetzee et al., 2015) was implemented on fine-mapped T2D-associated variants identified in Mahajan et al (2018). We overlapped T2D candidate causal variants with TF binding regions in islet-cell selective enhancers, and we estimated allele-specific effects of these candidate T2D causal variants on individual TF-binding (see Command 8 for chosen parameters; see complete code in Annex Command 1). MotifbreakR assesses if the sequence that surrounds a variant matches a known TF binding site and evaluates the amount of information that is gained or lost by one allele vs. another. The background frequencies (A=0.270182, C=0.2290216, G=0.2297711, T=0.2710253) were calculated from pancreatic islet enhancers (see Command 8); this

affects the calculation of motif disruptions. The **threshold** option was set to establish 5e-5 as the maximum p-value for a match to be called. The resulting variants were filtered by pct (pct > 0.8) ensuring that 80% of the motif matches the DNA sequence for the reference or alternate allele, and by the strength of the effect (effect = strong).

Thus, the integration of TF-binding that characterizes islet cell-type selective enhancers with fine-mapped genetic variants allowed us to identify the most likely molecular mechanism underlying a particular T2D association.

```
motifbreakR(snpList = list of snps,
            filterp = TRUE, #to filter by p-value
            pwmList = list of motifs to be interrogated,
            threshold = 5e-5, #maximum p-value for a match to be called
            method = "ic",
            bkg = c(A=0.270182, C=0.2290216, G=0.2297711, T=0.2710253),
            BPPARAM = BiocParallel::SerialParam())
```

Command 8

4. Results

Annotation of regulome signatures across cell-type selective open chromatin regions. After intersecting human pancreatic islet regulome annotations (Miguel-Escalada, Bonàs-Guarch, Cebola, Ponsa-Cobas, Mendieta-Esteban, Atla, Javierre, Rolando, Farabella, Morgan, García-Hurtado, et al., 2019) with open chromatin regions (scATAC-seq peaks) enriched or specific for islet cell-types, we obtained an islet cell-type classification of active enhancers (see Figure 2). Within the endocrine proportion (beta, alpha and delta cell types) of the annotated enhancers, we observe that open chromatin regions enriched for beta and alpha cells account for the largest fraction of overlapping enhancer elements in comparison to delta cells. This is concordant for both cell-type enriched or specific scATAC-seq peaks. However, the low count of cell-type specific peaks directed our analysis towards islet enhancers that overlap islet cell-type enriched peaks ("islet cell-type selective enhancers"). scATAC-seq peaks enriched in acinar cells accounted for <30% of all enhancers overlapping any cell-type enriched scATAC peak, with more than 70% of them falling in open chromatin regions selectively active in endocrine cell-types, as expected. Nevertheless, we leveraged active enhancers selectively active in acinar cells to provide further evidence to the cell-type regulatory component connected to human islet endocrine cell populations.

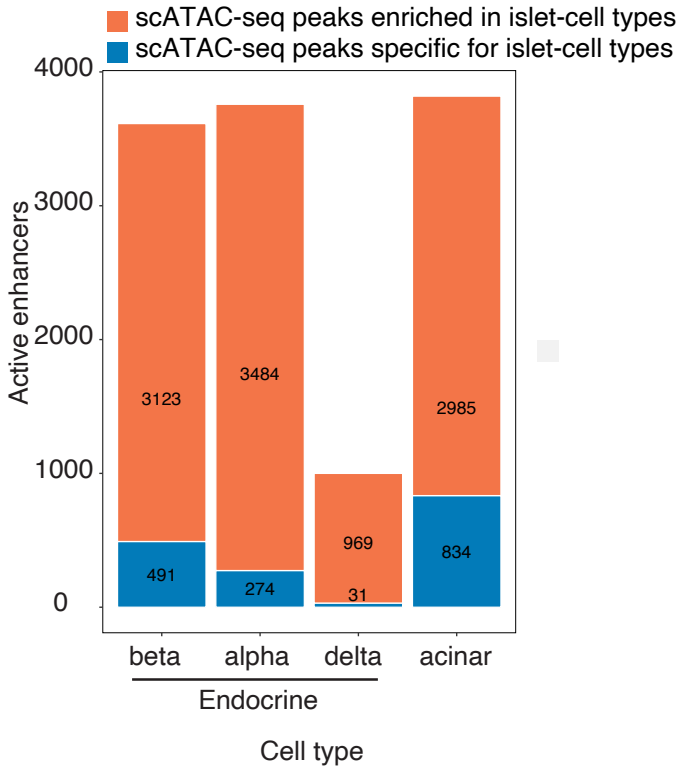


Figure 2 Barplot showing absolute number of active enhancers overlapping islet cell-type enriched (in orange) and specific (in blue) scATAC-seq peaks (islet cell-type enriched/specific open chromatin regions).

After classifying enhancers according to islet cell-type selective chromatin accessibility, we assigned target genes by leveraging pcHi-C enhancer-gene assignments (Miguel-Escalada, Bonàs-Guarch, Cebola, Ponsa-Cobas, Mendieta-Esteban, Atla, Javierre, Rolando, Farabella, Morgan, García-Hurtado, et al., 2019) and islet cis-eQTL variants (manuscript in preparation by the host lab). We performed enrichment analysis for functional annotations and ontologies with enrichR (Kuleshov et al., 2016) in target genes assigned to islet cell-type selective enhancers. We show in Table 1 that genes connected to endocrine cell-type selective enhancers were enriched (although not significantly after multiple test correction, in most cases) for functional annotations that are essential for pancreatic function, endocrine cell differentiation and development, and diabetes. Note that, while highly ranked annotations across genes linked to each group of endocrine cell-type selective enhancers revealed a broad endocrine functional profile, we did not achieve

appropriate resolution to unearth a cell-type selective profile. However, target genes assigned to alpha cell-selective enhancers were enriched in pathways related to glucagon secretion. Target genes connected to acinar cell-selective enhancers showed enrichment for annotations associated with cellular stress and apoptosis.

Table 1 Top functional biological terms enriched in target genes assigned to islet cell-type selective enhancers (see complete list in Annex Table 1).

	Gene-set library	Top functional biological terms for gene sets	P value	Adjusted P values	Combined score	Gene examples
Beta cells	**KEGG 2021 Human**	Maturity onset diabetes of the young	0.00944	1.000	13.72	NEUROD1 ,PDX1 ,SLC2A2 ,HES1
	KEGG 2021 Human	Type II diabetes mellitus	0.00969	1.000	10.57	KCNJ11 ,ABCC8 ,PRKCE ,PDX1
	Descartes Cell Types and Tissue 2021	Islet endocrine cells in Pancreas	0.02314	1.000	5.62	NECAB2 ,CERKL ,NKX2-2-AS1 ,DDC
Alpha cells	**Elsevier Pathway Collection**	alpha-Cell to beta-Cell Interconversion (Hypothesis)	0.00022	0.347	46.88	NEUROD1 ,CXCL12 ,MAF ,MAFB
	Elsevier Pathway Collection	L-cell: GCG, PYY and 5-HT Release	0.00418	1.000	19.17	CASR ,FFAR4 ,GNAS ,FFAR2
	GO Biological Process 2018	type B pancreatic cell differentiation (GO:0003309)	0.00016	0.703	139.27	PDX1 ,RFX3 ,INSM1 ,DLL1
	ARCHS4 Tissues	PANCREATIC ISLET	0.00005	0.006	12.19	USP6NL ,SCOC ,EHF ,FAM159B
Delta cells	**KEGG 2021 Human**	Insulin secretion	0.00377	1.000	14.29	CAMK2B ,CHRM3 ,RYR2 ,CAMK2D
	GO Biological Process 2018	regulation of type B pancreatic cell development (GO:2000074)	0.00006	0.175	233.92	GSK3B ,RHEB ,RFX3 ,NKX6-1
	ARCHS4 Tissues	BETA CELL	0.01388	1.000	5.18	EHF ,TRIO ,TMEM200A ,TESK1
	ARCHS4 Tissues	PANCREATIC ISLET	0.02602	1.000	4.33	USP6NL ,SCOC ,EHF ,FAM159B
Acinar cells	**WikiPathway 2021 Human**	Apoptosis-related network due to altered Notch3 in ovarian cancer WP2864	0.00116	0.661	17.73	VAV3 ,APP ,SOCS3 ,CDKN1A
	CCLE Proteomics 2020	ASPC1 PANCREAS TenPx29	0.00402	1.000	7.25	SH2D4A ,ACY1 ,CD82 ,PWWP2B

TF motif enrichment analysis. Motif enrichment analysis revealed well-defined groups of known TF motifs distinctly enriched across islet cell-type selective enhancers. (see Figure 3). Clustering of beta and alpha cell-selective enhancers was largely driven by a recurrent enrichment for TF motifs from members of the Forkhead box (FOX) family, among others. In particular, FOXA1 and FOXA2 were consistently enriched in alpha cell-selective enhancers, and in a lower degree in beta cell-selective enhancers. These FOXA family members are essential for alpha cell function and differentiation, glucagon biosynthesis and for beta cell secretory and metabolic activity (N. Gao et al., 2008; Nan Gao et al., 2010; Heddad Masson et al., 2014; Lee, Sund, Behr, Herrera, & Kaestner, 2005). We also identified a large enrichment of motifs from members of the RFX TF family in endocrine active enhancers, such as RFX6 in alpha cell-selective enhancers, which is involved in the determination of the endocrine cell lineage (Bramswig & Kaestner, 2011; Chandra et al., 2014; Smith et al., 2010). Endocrine cell-selective enhancers showed TF enrichment for members from the NKX family such as NKX6.1, which have also been reported to participate in alpha-cell formation and glucagon biosynthesis (Henseleit et al., 2005). TF motifs from PDX1 and PBX2 transcription factors were found to be enriched across endocrine cell-selective enhancers but mostly across enhancers selectively active in beta and delta cells. These results are in line with previously observed activity of PDX1 and PBX2 in the stimulation of somatostatin expression (Ampofo, Nalbach, Menger, & Laschke, 2020).

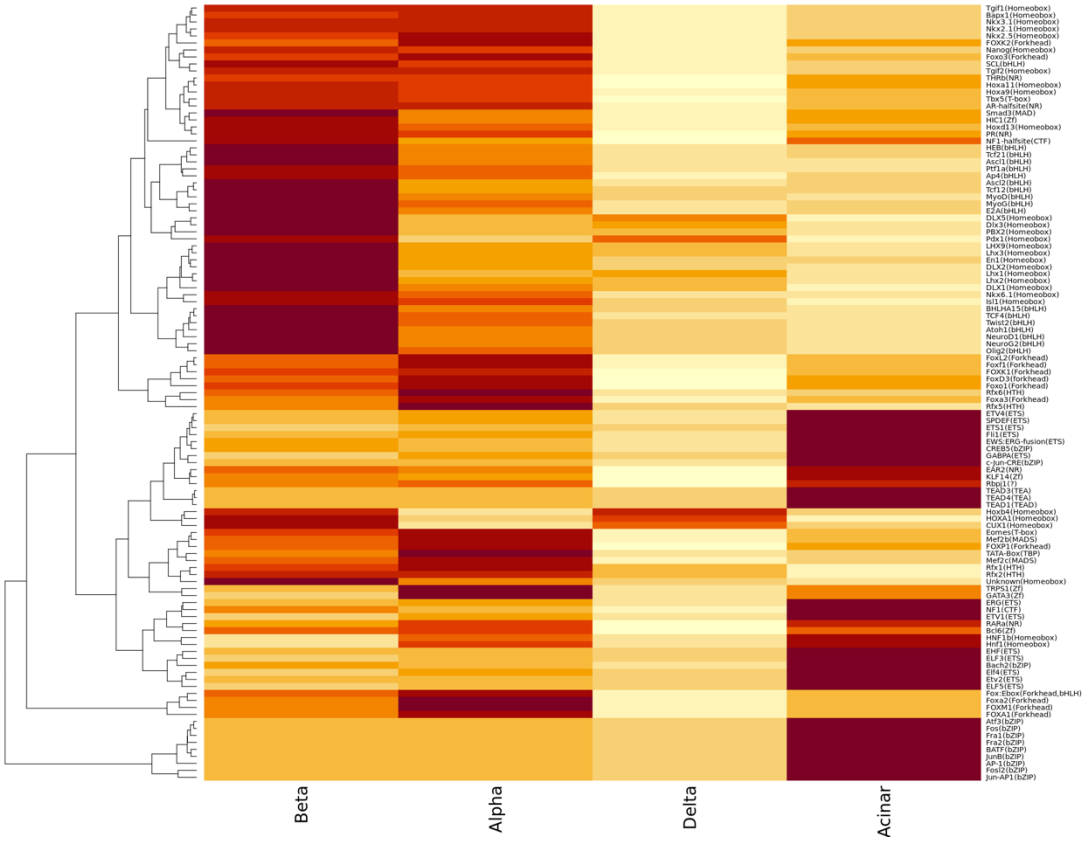


Figure 3 Heatmap representing enrichment of known TF motifs in cell-type selective active enhancers vs. the custom background (open chromatin regions that do not show enhancer epigenomic signatures). Top 50 enriched TF motifs for each cell type were selected for plotting. High enrichment (based on the enrichment p-value) is in red, and low enrichment in yellow.

HOXB4 and HOXA1 TF motifs are highly enriched in beta and delta cell-selective enhancers. Previous work reported that HOX TFs may be involved in pancreatic development (Gray, Pandha, Michael, Middleton, & Morgan, 2011). Other interesting TF motifs are LHX1, PTF1A and NEUROD1, which are significantly enriched in beta-selective enhancers, and ISL1 TF motifs, enriched in both beta and alpha cell-selective enhancers; these TFs are involved in pancreatic development and glucose homeostasis (Bethea et al., 2019; Dong, Provost, Leach, & Stainier, 2008; Gray et al., 2011; Mastracci, Anderson, Papizan, & Sussel, 2013). Within the minor fraction of delta cell-selective enhancers (9% of all cell-type selective enhancers, see Figure 2) we observed very low TF enrichment except for some TF motifs abovementioned. We rationalized that the low number of scATAC-seq peaks identified

in delta cells (see Figure 1) could hinder our statistical power. Of note, we also noticed that acinar cell-selective enhancers were largely enriched for FOS, FRA and JUN TF motifs, which are involved in the response against stress-induced cell death (Vaz et al., 2012; Zhou et al., 2007). The limited fraction of acinar cells captured in this analysis are most likely to be the result from exocrine contamination, and hereby, they might have suffered from cellular stress during human pancreatic isolation.

Analysis of individual motif occurrence. We selected a subset of TF motifs that were representative across islet cell-type selective enhancers based on previous TF motif enrichments. Results show that the cell-type selective component of the enhancers bound by a given TF (see Figure 4) is consistent with the cell-type selective clustering revealed in the TF motif enrichment analysis (see Figure 3).

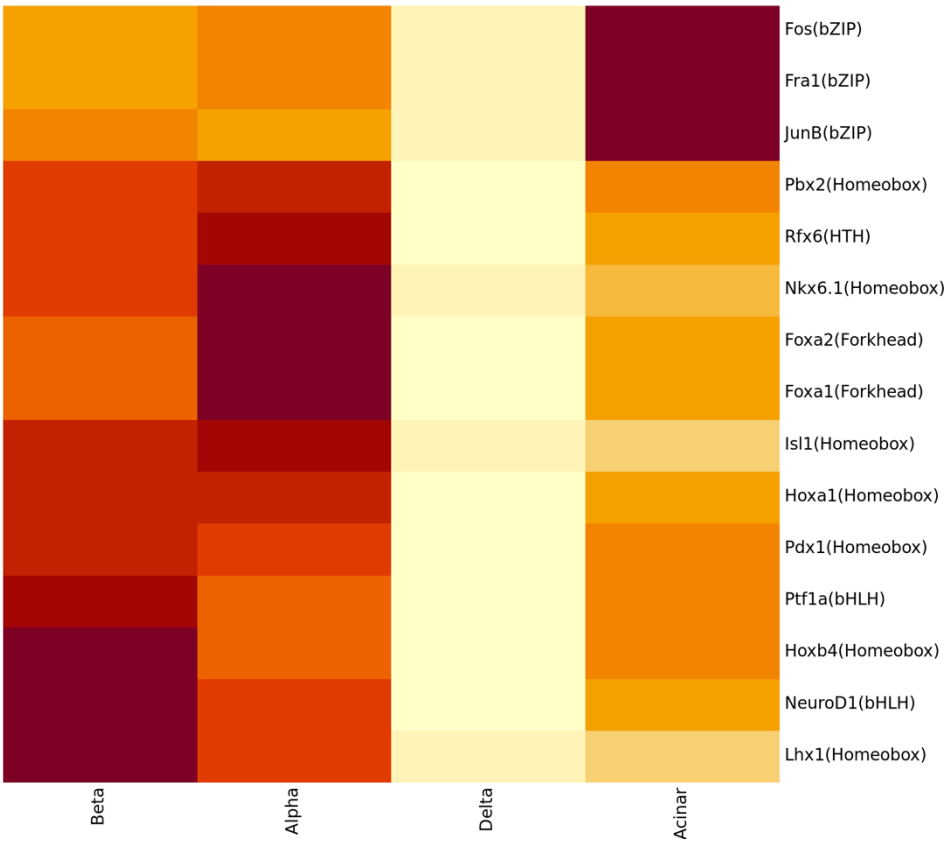


Figure 4 Heatmap showing known TF motif occurrences across islet cell-type selective enhancers. Lower to higher occurrences are represented from light yellow to dark red, respectively. TF motifs were selected based on previous TF motif enrichment results (see Figure 3).

LHX1, NEUROD1 and HOXb4 TF motifs show a high occurrence in beta cell-selective enhancers. Across alpha cell-selective enhancers, we observe an elevated occurrence of NKX6.1, FOXA1 and FOXA2 TF motifs. High TF motif occurrences for FOS, FRA1 and JUNB were identified in acinar cell enhancers. Finally, very low levels of TF motif occurrences were detected across delta cell-selective enhancers, consistent with the limited power already observed in TF motif enrichment analysis for this endocrine cell population.

Identification of T2D risk genetic variants affecting TF binding across islet cell-type selective enhancers. To predict the effect of T2D risk variants on islet cell-type gene regulation, we first overlapped fine-mapped T2D variants from one of the largest T2D GWAS meta-analysis (Mahajan et al., 2018a) with TF binding regions previously identified in islet-cell selective enhancers. This detected 22 candidate T2D causal variants within islet cell-type associated TF binding sites. After estimating the allele-specific effects of these 22 variants on the corresponding TF-binding sites using motifbreakR (Coetzee et al., 2015), 6 candidate T2D causal variants (rs180980072, rs115077735, rs703977, rs386111, rs34584161 and rs190513637) were predicted to disrupt TF binding for NEUROD1, HOXA1, LHX1 and FOS across beta, alpha and acinar selective enhancers (see Annex Table 2).

For example, at the *RNF6/CDK8* locus rs34584161 impacts on FOS TF binding in beta-cell selective active enhancers (see Annex Table 2). We observe that the T2D [A] risk increasing allele (effect size = 0.05) favors FOS TF binding (pct = 0,99). FOS TF motifs have recently been reported to be enriched in chromatin accessible regions associated with hormone-low endocrine cell-type states, including insulin low-secretory beta cells (Chiou et al., 2021). Furthermore, according to in-house and a recently published eQTL dataset in ~400 human pancreatic islet samples (Viñuela et al., 2020) (see figure 5) the rs34584161 [A] T2D risk allele is also associated with *RNF6* and *CDK8* increased gene expression levels. *CDK8* has been proposed as a negative regulator of insulin secretion and as a repressor proapoptotic neuropeptides expression during metabolic stress (Xue, Scotti, & Stoffel, 2019). Taken together, this suggests that the rs34584161 variant within a beta-cell selective enhancer contributes to T2D pathophysiology via the FOS-related regulatory network and by impacting on *RNF6* and *CDK8* gene expression. Further experiments to disentangle the effect on

insulin secretion and response to cell stress are necessary to elucidate the underlying molecular mechanism.

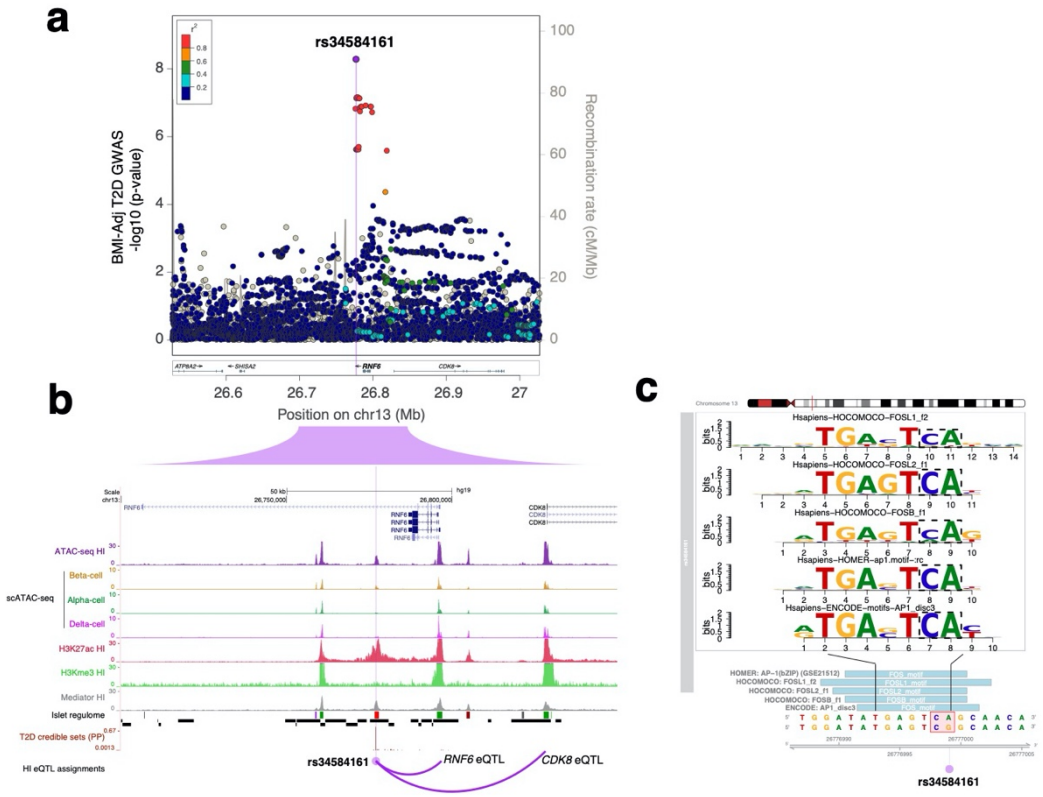


Figure 5. (a) Regional association signal locuszoom plot for the *RNF6/CDK8* locus centered on the rs34584161 T2D risk variant. Each dot represents a variant, with its p-value from a BMI-adjusted T2D meta-analysis on a -$\log_{10}$ scale in the y-axis. The x-axis represents the genomic position (hg19). Each variant is coloured by the LD (r^2) with rs34584161. (b) Human islet ATAC-seq, scATAC-seq across endocrine cell-types and CHIP-seq datasets for H3K27ac, H3K4me3 and Mediator are represented across islet regulome annotations. Gene assignments for rs34584161 are shown as purple archs based on eQTL maps in human islets. (c) TF binding disruption of FOS motifs by rs34584161. The position of the SNP within the motif are indicated by a purple dot and the red box. Motif logos for the robustly disrupted TF motifs are shown.

Another example is the rs180980072 variant, which impacts on NEUROD1 TF binding in an alpha cell-selective active enhancer (see Annex Table 2). It should be noted that although this enhancer shows an increased chromatin accessibility in alpha cells, we observe scATAC-seq signal in other endocrine cell types. Additional

comprehensive analyses will provide further resolution to the true islet cell-type component of this and other regulatory elements. Contrary to the last example, here the rs180980072 [A] effect allele protects for T2D (effect size = -0.19) and favors NEUROD1 TF binding (pct = 0.97). Chromatin interaction maps from pcHi-C in human islets (see Annex Figure 1) links the enhancer where rs180980072 falls to *ITGA1* and *PELO* genes. This suggests a candidate regulatory influence from the alpha selective enhancer containing the rs180980072 variant over the expression of *ITGA1* and *PELO*. Of note, *PELO* has been identified in a recent genome-wide CRISPR screen to positively regulate insulin secretion (Grotz et al., 2021) and *ITGA1* has been reported as a beta-cell surface marker that successfully performs to enrich for functional stem-cell derived beta cells (Veres et al., 2019).

5. Discussion

Untwining the single-cell regulatory profile of gene regulation is one of the most challenging goals ahead. New technological developments such as scATAC-seq have expended our capacity to bring gene regulatory frames at a cellular resolution. This creates new opportunities to unravel the cell-type selective contributions to cell identity and function, or disease pathophysiology. Our project aims to characterize single-cell regulatory profiles in human pancreatic islets and elucidate the distinct roles of islet cell-types to human pancreatic islet transcriptional regulation. We sub-classified human islet active enhancers according to their cell-type selective chromatin accessibility and quantified TF motif enrichments in groups of islet cell-type selective enhancers. Finally, we identified cell-type selective active enhancers bound by TF that were a proxy of islet cell-type selective regulatory programmes, which also were impacted by T2D risk alleles.

Our classification of enhancer elements according to islet-cell selective chromatin accessibility attained the appropriate resolution to discriminate between islet gene regulation in the endocrine vs. the exocrine component. However, as observed by the enrichment of functional annotations and ontologies in genes linked to islet cell-selective enhancers, we were not able to unearth a cell-type selective profile. Several limitations could explain this lack of endocrine cell-type resolution. First, we relied on cell-type enriched chromatin accessibility peaks to sub-classify enhancer elements. Although a given enhancer might show increased chromatin

accessibility in a given cell type, with our current dataset, we cannot discard enhancer activity in other endocrine cell-types. Second, we reasoned that the overwhelmingly lack of significant enrichments (based on adjusted p-values) for the different ontologies and annotations can be partially explained by the narrowed gene count due to the limited number of available cell-type selective enhancers. Of note, we did not leverage cell-type selective promoters, which might have provided a more comprehensive single-cell perspective of human islet regulation. Furthermore, expanding this pilot study, by harnessing human islet scATAC-seq data in additional donors, will provide further resolution. Finally, we also relied on islet eQTL data to connect enhancer elements to target genes, which could be confounded by LD (Schaid et al., 2018). Thus, we might have connected enhancer elements to non-relevant target genes that may dilute our understanding of the cell-type selective component of human pancreatic gene regulation.

We should note that linking single-cell chromatin accessibility on enhancers and promoters, and by also integrating scRNA-seq data of candidate target genes, could provide a larger granularity in the cell-type characterization of gene regulatory networks in human pancreatic islets. This can be accomplished by employing CICERO (Pliner et al., 2018). This tools leverages single-cell chromatin accessibility data to predict cis-regulatory interactions (such as those between enhancers and promoters). This will allow us to assign enhancers to their endogenous target genes by following single-cell profiles.

Motif enrichment analysis and the identification of motif occurrence across cell-type selective enhancers revealed several TFs that could delineate islet cell-type regulatory networks. However, TF motif redundancy was observed between cell-types hindering the identification of consistent islet cell-type TF networks. One approach to overcome this limitation is harnessing manually-curated PWM (position weighted matrices) databases for the TF motif enrichment analysis. Furthermore, alternative motif enrichment software more appropriate for scATAC-seq data, such as chromVAR (Schep, Wu, Buenrostro, & Greenleaf, 2017), could have refined our results.

Our integrative approach combining fine-mapped T2D variants, eQTL data and cell-type selective enhancers allowed us to propose the molecular mechanism underlying T2D genetic susceptibility. However, due to the multifactorial nature of T2D, other tissues and biological pathways could be contributing to the disease pathophysiology. Therefore, not all T2D signals could be explained by perturbations

in the islet regulatory landscape. Thus, our data might have been more relevant to dissect genetic associations related to other islet-related traits, such as fasting glycemia or measures of beta-cell function based on oral glucose tolerance tests, among others. However, high-resolution fine-mapped data was not available.

Despite all these limitations, we were able to characterize cell-type selective regulatory profiles by identifying T2D-associated variants that modulate TF binding sites, which are distinctly enriched in sets of cell-type selective enhancers. This and other on-going efforts are essential to bridge the genotype to phenotype gap and invigorate drug discovery.

6. Conclusions

After the completion of this master project the following conclusions were drawn. First, despite TF motif redundancy, motif enrichment analysis presented well-defined groups of known TF motifs distinctly enriched across islet cell-type selective enhancers. Second, the landscape of TF motif occurrences across cell-type selective enhancers was consistent with the cell-type selective clustering observed in the TF motif enrichment analysis. This provides further support to our definitions of islet-cell selective TF binding. Third, the integration of TF-binding that characterizes islet cell-type selective enhancers with fine-mapped T2D genetic variants allowed us to propose the most likely molecular mechanism underlying some T2D risk loci. Fourth, the small count of enhancers leveraged in this study could have limited the performance of our approach, masking cell-type selective regulatory networks operating across the islet regulome. And last, further analyses should focus on extending the sample size, as well as on joint analysis of cell-type selective enhancers along with their target promoters, and on the integration of scRNA-seq data. These and other efforts will aid in elucidating the cell-type selective regulatory component of gene regulation in human pancreatic islets.

7. Glossary

ATAC-seq – assay for transposase-accessible chromatin using sequencing

ChIP – chromatin immunoprecipitation

ChIP-seq – chromatin immunoprecipitation sequencing

DNA – deoxyribonucleic acid

eQTL – expression quantitative trait locus

GWAS – genome-wide association study

H3K27ac – histone 3 lysine 27 acetylation

H3K4me1 – histone 3 lysine 4 mono-methylation

H3K4me3 – histone 3 lysine 4 tri-methylation

LD – linkage disequilibrium

pcHi-C – promoter capture Hi-C

scATAC-seq – single-cell ATAC sequencing

scRNA-seq – single-cell RNA sequencing

TF – transcription factor

TFBS – transcription factor binding site

TSS – transcription start site

8. References

American Diabetes Association. (2020). 2. Classification and Diagnosis of Diabetes: Standards of Medical Care in Diabetes-2020. *Diabetes Care*, *43*(Suppl 1), S14–S31. https://doi.org/10.2337/dc20-S002

Ampofo, E., Nalbach, L., Menger, M. D., & Laschke, M. W. (2020). Regulatory Mechanisms of Somatostatin Expression. *International Journal of Molecular Sciences*, *21*(11), 4170. https://doi.org/10.3390/ijms21114170

Andersson, R., Gebhard, C., Miguel-Escalada, I., Hoof, I., Bornholdt, J., Boyd, M., … Sandelin, A. (2014). An atlas of active enhancers across human cell types and tissues. *Nature*, *507*(7493), 455–461. https://doi.org/10.1038/nature12787

Andersson, R., & Sandelin, A. (2020). Determinants of enhancer and promoter activities of regulatory elements. *Nature Reviews. Genetics*, *21*(2), 71–87. https://doi.org/10.1038/s41576-019-0173-8

Baek, S., & Lee, I. (2020). Single-cell ATAC sequencing analysis: From data preprocessing to hypothesis generation. *Computational and Structural Biotechnology Journal*, *18*, 1429–1439. https://doi.org/10.1016/j.csbj.2020.06.012

Bethea, M., Liu, Y., Wade, A. K., Mullen, R., Gupta, R., Gelfanov, V., … Hunter, C. S. (2019). The islet-expressed Lhx1 transcription factor interacts with Islet-1 and

contributes to glucose homeostasis. *American Journal of Physiology-Endocrinology and Metabolism*, *316*(3), E397–E409. https://doi.org/10.1152/ajpendo.00235.2018

Boyle, A. P., Davis, S., Shulha, H. P., Meltzer, P., Margulies, E. H., Weng, Z., … Crawford, G. E. (2008). High-Resolution Mapping and Characterization of Open Chromatin across the Genome. *Cell*, *132*(2), 311–322. https://doi.org/10.1016/j.cell.2007.12.014

Bramswig, N. C., & Kaestner, K. H. (2011). Transcriptional regulation of α-cell differentiation. *Diabetes, Obesity and Metabolism*, *13*, 13–20. https://doi.org/10.1111/j.1463-1326.2011.01440.x

Buenrostro, J. D., Corces, M. R., Lareau, C. A., Wu, B., Schep, A. N., Aryee, M. J., … Greenleaf, W. J. (2018). Integrated Single-Cell Analysis Maps the Continuous Regulatory Landscape of Human Hematopoietic Differentiation. *Cell*, *173*(6), 1535-1548.e16. https://doi.org/10.1016/j.cell.2018.03.074

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

Buenrostro, J. D., Wu, B., Chang, H. Y., & Greenleaf, W. J. (2015). ATAC-seq: A Method for Assaying Chromatin Accessibility Genome-Wide. *Current Protocols in Molecular Biology*, *109*(1). https://doi.org/10.1002/0471142727.mb2129s109

Buenrostro, J. D., Wu, B., Litzenburger, U. M., Ruff, D., Gonzales, M. L., Snyder, M. P., … Greenleaf, W. J. (2015). Single-cell chromatin accessibility reveals principles of regulatory variation. *Nature*, *523*(7561), 486–490. https://doi.org/10.1038/nature14590

Chandra, V., Albagli-Curiel, O., Hastoy, B., Piccand, J., Randriamampita, C., Vaillant, E., … Scharfmann, R. (2014). RFX6 Regulates Insulin Secretion by Modulating Ca2+ Homeostasis in Human β Cells. *Cell Reports*, *9*(6), 2206–2218. https://doi.org/10.1016/j.celrep.2014.11.010

Chatterjee, S., & Ahituv, N. (2017). Gene Regulatory Elements, Major Drivers of Human Disease. *Annual Review of Genomics and Human Genetics*, *18*, 45–63. https://doi.org/10.1146/annurev-genom-091416-035537

Chen, H., Lareau, C., Andreani, T., Vinyard, M. E., Garcia, S. P., Clement, K., … Pinello, L. (2019). Assessment of computational methods for the analysis of

single-cell ATAC-seq data. *Genome Biology*, *20*(1), 241. https://doi.org/10.1186/s13059-019-1854-5

Chen, Xi, Miragaia, R. J., Natarajan, K. N., & Teichmann, S. A. (2018). A rapid and robust method for single cell chromatin accessibility profiling. *Nature Communications*, *9*(1), 5345. https://doi.org/10.1038/s41467-018-07771-0

Chen, Xingqi, Litzenburger, U. M., Wei, Y., Schep, A. N., LaGory, E. L., Choudhry, H., … Chang, H. Y. (2018). Joint single-cell DNA accessibility and protein epitope profiling reveals environmental regulation of epigenomic heterogeneity. *Nature Communications*, *9*(1), 4590. https://doi.org/10.1038/s41467-018-07115-y

Chiou, J., Zeng, C., Cheng, Z., Han, J. Y., Schlichting, M., Huang, S., … Gaulton, K. J. (2019). Single cell chromatin accessibility reveals pancreatic islet cell type- and state-specific regulatory programs of diabetes risk. *BioRxiv*. https://doi.org/10.1101/693671

Chiou, J., Zeng, C., Cheng, Z., Han, J. Y., Schlichting, M., Miller, M., … Gaulton, K. J. (2021). Single-cell chromatin accessibility identifies pancreatic islet cell type– and state-specific regulatory programs of diabetes risk. *Nature Genetics*, *53*(4), 455–466. https://doi.org/10.1038/s41588-021-00823-0

Coetzee, S. G., Coetzee, G. A., & Hazelett, D. J. (2015). MotifbreakR: An R/Bioconductor package for predicting variant effects at transcription factor binding sites. *Bioinformatics*, *31*(23), 3847–3849. https://doi.org/10.1093/bioinformatics/btv470

Cusanovich, D. A., Daza, R., Adey, A., Pliner, H. A., Christiansen, L., Gunderson, K. L., … Shendure, J. (2015). Multiplex single-cell profiling of chromatin accessibility by combinatorial cellular indexing. *Science*, *348*(6237), 910–914. https://doi.org/10.1126/science.aab1601

Delaneau, O., Ongen, H., Brown, A. A., Fort, A., Panousis, N. I., & Dermitzakis, E. T. (2017). A complete tool set for molecular QTL discovery and analysis. *Nature Communications*, *8*(1), 15452. https://doi.org/10.1038/ncomms15452

Dong, P. D. S., Provost, E., Leach, S. D., & Stainier, D. Y. R. (2008). Graded levels of Ptf1a differentially regulate endocrine and exocrine fates in the developing pancreas. *Genes & Development*, *22*(11), 1445–1450. https://doi.org/10.1101/gad.1663208

ENCODE Project Consortium. (2012). An integrated encyclopedia of DNA elements

in the human genome. *Nature*, *489*(7414), 57–74. https://doi.org/10.1038/nature11247

FANTOM Consortium and the RIKEN PMI and CLST (DGT), Forrest, A. R. R., Kawaji, H., Rehli, M., Baillie, J. K., de Hoon, M. J. L., … Hayashizaki, Y. (2014). A promoter-level mammalian expression atlas. *Nature*, *507*(7493), 462–470. https://doi.org/10.1038/nature13182

Gao, N., LeLay, J., Vatamaniuk, M. Z., Rieck, S., Friedman, J. R., & Kaestner, K. H. (2008). Dynamic regulation of Pdx1 enhancers by Foxa1 and Foxa2 is essential for pancreas development. *Genes & Development*, *22*(24), 3435–3448. https://doi.org/10.1101/gad.1752608

Gao, Nan, Le Lay, J., Qin, W., Doliba, N., Schug, J., Fox, A. J., … Kaestner, K. H. (2010). Foxa1 and Foxa2 Maintain the Metabolic and Secretory Features of the Mature β-Cell. *Molecular Endocrinology*, *24*(8), 1594–1604. https://doi.org/10.1210/me.2009-0513

Giresi, P. G., Kim, J., McDaniell, R. M., Iyer, V. R., & Lieb, J. D. (2007). FAIRE (Formaldehyde-Assisted Isolation of Regulatory Elements) isolates active regulatory elements from human chromatin. *Genome Research*, *17*(6), 877–885. https://doi.org/10.1101/gr.5533506

Grant, C. E., Bailey, T. L., & Noble, W. S. (2011). FIMO: scanning for occurrences of a given motif. *Bioinformatics*, *27*(7), 1017–1018. https://doi.org/10.1093/bioinformatics/btr064

Gray, S., Pandha, H. S., Michael, A., Middleton, G., & Morgan, R. (2011). HOX genes in pancreatic development and cancer. *Journal of the Pancreas*, *12*(3), 216–219. https://doi.org/https://doi.org/10.6092/1590-8577/3284

Grotz, A. K., Navarro-Guerrero, E., Bevacqua, R. J., Baronio, R., Thomsen, S. K., Nawaz, S., … Gloyn, A. L. (2021). A genome-wide CRISPR screen identifies regulators of beta cell function involved in type 2 diabetes risk. *BioRxiv*, 2021.05.28.445984. https://doi.org/10.1101/2021.05.28.445984

Hao, Y., Hao, S., Andersen-Nissen, E., Mauck, W. M., Zheng, S., Butler, A., … Satija, R. (2020). Integrated analysis of multimodal single-cell data. *BioRxiv*. https://doi.org/10.1101/2020.10.12.335331

Heddad Masson, M., Poisson, C., Guérardel, A., Mamin, A., Philippe, J., & Gosmain, Y. (2014). Foxa1 and Foxa2 Regulate α-Cell Differentiation, Glucagon Biosynthesis, and Secretion. *Endocrinology*, *155*(10), 3781–3792.

https://doi.org/10.1210/en.2013-1843

Heinz, S., Benner, C., Spann, N., Bertolino, E., Lin, Y. C., Laslo, P., … Glass, C. K. (2010). Simple Combinations of Lineage-Determining Transcription Factors Prime cis-Regulatory Elements Required for Macrophage and B Cell Identities. *Molecular Cell*, *38*(4), 576–589. https://doi.org/10.1016/j.molcel.2010.05.004

Henseleit, K. D., Nelson, S. B., Kuhlbrodt, K., Hennings, J. C., Ericson, J., & Sander, M. (2005). NKX6 transcription factor activity is required for α- andβ-cell development in the pancreas. *Development*, *132*(13), 3139–3149. https://doi.org/10.1242/dev.01875

International Diabetes Federation. (2019). *IDF Diabetes Atlas, 9th edn*. Brussels, Belgium: International Diabetes Federation, 2019.

Javierre, B. M., Burren, O. S., Wilder, S. P., Kreuzhuber, R., Hill, S. M., Sewitz, S., … Fraser, P. (2016). Lineage-Specific Genome Architecture Links Enhancers and Non-coding Disease Variants to Target Gene Promoters. *Cell*, *167*(5), 1369-1384.e19. https://doi.org/10.1016/j.cell.2016.09.037

Kagey, M. H., Newman, J. J., Bilodeau, S., Zhan, Y., Orlando, D. A., van Berkum, N. L., … Young, R. A. (2010). Mediator and cohesin connect gene expression and chromatin architecture. *Nature*, *467*(7314), 430–435. https://doi.org/10.1038/nature09380

Kaplan, N., Moore, I. K., Fondufe-Mittendorf, Y., Gossett, A. J., Tillo, D., Field, Y., … Segal, E. (2009). The DNA-encoded nucleosome organization of a eukaryotic genome. *Nature*, *458*(7236), 362–366. https://doi.org/10.1038/nature07667

Klemm, S. L., Shipony, Z., & Greenleaf, W. J. (2019). Chromatin accessibility and the regulatory epigenome. *Nature Reviews Genetics*, *20*(4), 207–220. https://doi.org/10.1038/s41576-018-0089-8

Kuleshov, M. V, Jones, M. R., Rouillard, A. D., Fernandez, N. F., Duan, Q., Wang, Z., … Ma'ayan, A. (2016). Enrichr: a comprehensive gene set enrichment analysis web server 2016 update. *Nucleic Acids Research*, *44*(W1), W90-7. https://doi.org/10.1093/nar/gkw377

Lareau, C. A., Duarte, F. M., Chew, J. G., Kartha, V. K., Burkett, Z. D., Kohlway, A. S., … Buenrostro, J. D. (2019). Droplet-based combinatorial indexing for massive-scale single-cell chromatin accessibility. *Nature Biotechnology*, *37*(8), 916–924. https://doi.org/10.1038/s41587-019-0147-6

Lee, C. S., Sund, N. J., Behr, R., Herrera, P. L., & Kaestner, K. H. (2005). Foxa2 is

required for the differentiation of pancreatic α-cells. *Developmental Biology*, *278*(2), 484–495. https://doi.org/10.1016/j.ydbio.2004.10.012

Mahajan, A., Taliun, D., Thurner, M., Robertson, N. R., Torres, J. M., Rayner, N. W., … McCarthy, M. I. (2018a). Fine-mapping type 2 diabetes loci to single-variant resolution using high-density imputation and islet-specific epigenome maps. *Nature Genetics*, *50*(11), 1505–1513. https://doi.org/10.1038/s41588-018-0241-6

Mahajan, A., Taliun, D., Thurner, M., Robertson, N. R., Torres, J. M., Rayner, N. W., … McCarthy, M. I. (2018b). Fine-mapping type 2 diabetes loci to single-variant resolution using high-density imputation and islet-specific epigenome maps. *Nature Genetics*, *50*(11), 1505–1513. https://doi.org/10.1038/s41588-018-0241-6

Mastracci, T. L., Anderson, K. R., Papizan, J. B., & Sussel, L. (2013). Regulation of Neurod1 Contributes to the Lineage Potential of Neurogenin3+ Endocrine Precursor Cells in the Pancreas. *PLoS Genetics*, *9*(2), e1003278. https://doi.org/10.1371/journal.pgen.1003278

Maurano, M. T., Humbert, R., Rynes, E., Thurman, R. E., Haugen, E., Wang, H., … Stamatoyannopoulos, J. A. (2012). Systematic localization of common disease-associated variation in regulatory DNA. *Science (New York, N.Y.)*, *337*(6099), 1190–1195. https://doi.org/10.1126/science.1222794

McCarthy, M. I. (2010). Genomics, type 2 diabetes, and obesity. *The New England Journal of Medicine*, *363*(24), 2339–2350. https://doi.org/10.1056/NEJMra0906948

Mezger, A., Klemm, S., Mann, I., Brower, K., Mir, A., Bostick, M., … Greenleaf, W. (2018). High-throughput chromatin accessibility profiling at single-cell resolution. *Nature Communications*, *9*(1), 3647. https://doi.org/10.1038/s41467-018-05887-x

Miguel-Escalada, I., Bonàs-Guarch, S., Cebola, I., Ponsa-Cobas, J., Mendieta-Esteban, J., Atla, G., … Ferrer, J. (2019). Human pancreatic islet three-dimensional chromatin architecture provides insights into the genetics of type 2 diabetes. *Nature Genetics*, *51*(7), 1137–1148. https://doi.org/10.1038/s41588-019-0457-0

Miguel-Escalada, I., Bonàs-Guarch, S., Cebola, I., Ponsa-Cobas, J., Mendieta-Esteban, J., Atla, G., … Ferrer, J. (2019). Human pancreatic islet three-

dimensional chromatin architecture provides insights into the genetics of type 2 diabetes. *Nature Genetics*, *51*(7), 1137–1148. https://doi.org/10.1038/s41588-019-0457-0

Miguel-Escalada, I., Pasquali, L., & Ferrer, J. (2015). Transcriptional enhancers: functional insights and role in human disease. *Current Opinion in Genetics & Development*, *33*, 71–76. https://doi.org/10.1016/j.gde.2015.08.009

Mulqueen, R., DeRosa, B., Thornton, C., Sayar, Z., Torkenczy, K., Fields, A., … Adey, A. (2019). Improved single-cell ATAC-seq reveals chromatin dynamics of in vitro corticogenesis. *BioRxiv*, 637256. https://doi.org/10.1101/637256

Pasquali, L., Gaulton, K. J., Rodríguez-Seguí, S. A., Mularoni, L., Miguel-Escalada, I., Akerman, İ., … Ferrer, J. (2014). Pancreatic islet enhancer clusters enriched in type 2 diabetes risk-associated variants. *Nature Genetics*, *46*(2), 136–143. https://doi.org/10.1038/ng.2870

Pliner, H. A., Packer, J. S., McFaline-Figueroa, J. L., Cusanovich, D. A., Daza, R. M., Aghamirzaie, D., … Trapnell, C. (2018). Cicero Predicts cis-Regulatory DNA Interactions from Single-Cell Chromatin Accessibility Data. *Molecular Cell*, *71*(5), 858-871.e8. https://doi.org/10.1016/j.molcel.2018.06.044

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

Rai, V., Quang, D. X., Erdos, M. R., Cusanovich, D. A., Daza, R. M., Narisu, N., … Collins, F. S. (2020). Single-cell ATAC-Seq in human pancreatic islets and deep learning upscaling of rare cells reveals cell-specific type 2 diabetes regulatory signatures. *Molecular Metabolism*, *32*, 109–121. https://doi.org/10.1016/j.molmet.2019.12.006

Reddington, J. P., Garfield, D. A., Sigalova, O. M., Karabacak Calviello, A., Marco-Ferreres, R., Girardot, C., … Furlong, E. E. M. (2020). Lineage-Resolved Enhancer and Promoter Usage during a Time Course of Embryogenesis. *Developmental Cell*, *55*(5), 648-664.e9. https://doi.org/10.1016/j.devcel.2020.10.009

Reznikoff, W. S. (2008). Transposon Tn 5. *Annual Review of Genetics*, *42*(1), 269–286. https://doi.org/10.1146/annurev.genet.42.110807.091656

Rubin, A. J., Parker, K. R., Satpathy, A. T., Qi, Y., Wu, B., Ong, A. J., … Khavari, P. A. (2019). Coupled Single-Cell CRISPR Screening and Epigenomic Profiling

Reveals Causal Gene Regulatory Networks. *Cell*, *176*(1–2), 361-376.e17. https://doi.org/10.1016/j.cell.2018.11.022

Satpathy, A. T., Granja, J. M., Yost, K. E., Qi, Y., Meschi, F., McDermott, G. P., … Chang, H. Y. (2019). Massively parallel single-cell chromatin landscapes of human immune cell development and intratumoral T cell exhaustion. *Nature Biotechnology*, *37*(8), 925–936. https://doi.org/10.1038/s41587-019-0206-z

Satpathy, A. T., Saligrama, N., Buenrostro, J. D., Wei, Y., Wu, B., Rubin, A. J., … Chang, H. Y. (2018). Transcript-indexed ATAC-seq for precision immune profiling. *Nature Medicine*, *24*(5), 580–590. https://doi.org/10.1038/s41591-018-0008-8

Schaid, D. J., Chen, W., & Larson, N. B. (2018). From genome-wide associations to candidate causal variants by statistical fine-mapping. *Nature Reviews. Genetics*, *19*(8), 491–504. https://doi.org/10.1038/s41576-018-0016-z

Schep, A. N., Wu, B., Buenrostro, J. D., & Greenleaf, W. J. (2017). chromVAR: inferring transcription-factor-associated accessibility from single-cell epigenomic data. *Nature Methods*, *14*(10), 975–978. https://doi.org/10.1038/nmeth.4401

Schones, D. E., Cui, K., Cuddapah, S., Roh, T.-Y., Barski, A., Wang, Z., … Zhao, K. (2008). Dynamic Regulation of Nucleosome Positioning in the Human Genome. *Cell*, *132*(5), 887–898. https://doi.org/10.1016/j.cell.2008.02.022

Segerstolpe, Å., Palasantza, A., Eliasson, P., Andersson, E.-M., Andréasson, A.-C., Sun, X., … Sandberg, R. (2016). Single-Cell Transcriptome Profiling of Human Pancreatic Islets in Health and Type 2 Diabetes. *Cell Metabolism*, *24*(4), 593–607. https://doi.org/10.1016/j.cmet.2016.08.020

Shashikant, T., & Ettensohn, C. A. (2019). Genome-wide analysis of chromatin accessibility using ATAC-seq (pp. 219–235). https://doi.org/10.1016/bs.mcb.2018.11.002

Shema, E., Bernstein, B. E., & Buenrostro, J. D. (2019). Single-cell and single-molecule epigenomics to uncover genome regulation at unprecedented resolution. *Nature Genetics*, *51*(1), 19–25. https://doi.org/10.1038/s41588-018-0290-x

Shlyueva, D., Stampfel, G., & Stark, A. (2014). Transcriptional enhancers: from properties to genome-wide predictions. *Nature Reviews. Genetics*, *15*(4), 272–286. https://doi.org/10.1038/nrg3682

Smith, S. B., Qu, H.-Q., Taleb, N., Kishimoto, N. Y., Scheel, D. W., Lu, Y., …

German, M. S. (2010). Rfx6 directs islet formation and insulin production in mice and humans. *Nature*, *463*(7282), 775–780. https://doi.org/10.1038/nature08748

Storey, J. D., Bass, A. J., Dabney, A., Robinson, D., & Warnes, G. (2021). qvalue: Q-value estimation for false discovery rate control. https://doi.org/10.18129/B9.bioc.qvalue

Stuart, T., Srivastava, A., Lareau, C., & Satija, R. (2020). Multimodal single-cell chromatin analysis with Signac. *BioRxiv*. https://doi.org/10.1101/2020.11.09.373613

Thurner, M., van de Bunt, M., Torres, J. M., Mahajan, A., Nylander, V., Bennett, A. J., … McCarthy, M. I. (2018). Integration of human pancreatic islet genomic data refines regulatory mechanisms at Type 2 Diabetes susceptibility loci. *ELife*, *7*. https://doi.org/10.7554/eLife.31977

Vaz, M., Machireddy, N., Irving, A., Potteti, H. R., Chevalier, K., Kalvakolanu, D., & Reddy, S. P. (2012). Oxidant-Induced Cell Death and Nrf2-Dependent Antioxidative Response Are Controlled by Fra-1/AP-1. *Molecular and Cellular Biology*, *32*(9), 1694–1709. https://doi.org/10.1128/MCB.06390-11

Veres, A., Faust, A. L., Bushnell, H. L., Engquist, E. N., Kenty, J. H.-R., Harb, G., … Melton, D. A. (2019). Charting cellular identity during human in vitro β-cell differentiation. *Nature*, *569*(7756), 368–373. https://doi.org/10.1038/s41586-019-1168-5

Viñuela, A., Varshney, A., van de Bunt, M., Prasad, R. B., Asplund, O., Bennett, A., … McCarthy, M. I. (2020). Genetic variant effects on gene expression in human pancreatic islets and their implications for T2D. *Nature Communications*, *11*(1), 4912. https://doi.org/10.1038/s41467-020-18581-8

Vujkovic, M., Keaton, J. M., Lynch, J. A., Miller, D. R., Zhou, J., Tcheandjieu, C., … Saleheen, D. (2020). Discovery of 318 new risk loci for type 2 diabetes and related vascular outcomes among 1.4 million participants in a multi-ancestry meta-analysis. *Nature Genetics*, *52*(7), 680–691. https://doi.org/10.1038/s41588-020-0637-y

Wellcome Trust Case Control Consortium, Maller, J. B., McVean, G., Byrnes, J., Vukcevic, D., Palin, K., … Donnelly, P. (2012). Bayesian refinement of association signals for 14 loci in 3 common diseases. *Nature Genetics*, *44*(12), 1294–1301. https://doi.org/10.1038/ng.2435

Wolffe, A. (2000). Chromatin Structure. In *Chromatin* (Third Edit, pp. 7–172).

Academic Press. https://doi.org/10.1016/B978-012761914-9/50004-0

Xie, Z., Bailey, A., Kuleshov, M. V., Clarke, D. J. B., Evangelista, J. E., Jenkins, S. L., … Ma'ayan, A. (2021). Gene Set Knowledge Discovery with Enrichr. *Current Protocols*, *1*(3). https://doi.org/10.1002/cpz1.90

Xue, J., Scotti, E., & Stoffel, M. (2019). CDK8 Regulates Insulin Secretion and Mediates Postnatal and Stress-Induced Expression of Neuropeptides in Pancreatic β Cells. *Cell Reports*, *28*(11), 2892-2904.e7. https://doi.org/10.1016/j.celrep.2019.08.025

Yan, F., Powell, D. R., Curtis, D. J., & Wong, N. C. (2020). From reads to insight: a hitchhiker's guide to ATAC-seq data analysis. *Genome Biology*, *21*(1), 22. https://doi.org/10.1186/s13059-020-1929-3

Zheng, G. X. Y., Terry, J. M., Belgrader, P., Ryvkin, P., Bent, Z. W., Wilson, R., … Bielas, J. H. (2017). Massively parallel digital transcriptional profiling of single cells. *Nature Communications*, *8*(1), 14049. https://doi.org/10.1038/ncomms14049

Zhou, H., Gao, J., Lu, Z. Y., Lu, L., Dai, W., & Xu, M. (2007). Role of c-Fos/JunD in protecting stress-induced cell death. *Cell Proliferation*, *40*(3), 431–444. https://doi.org/10.1111/j.1365-2184.2007.00444.x

Acknowledgments

We thank Irene Miguel-Escalada (Regulatory Genomics and Diabetes, Centre for Genomic Regulation, The Barcelona Institute of Science and Technology, Barcelona. CIBERDEM.) and Goutham Atla (Regulatory Genomics and Diabetes, Centre for Genomic Regulation, The Barcelona Institute of Science and Technology, Barcelona. CIBERDEM.) for producing and providing the data; Silvia Bonàs-Guarch (Regulatory Genomics and Diabetes, Centre for Genomic Regulation, The Barcelona Institute of Science and Technology, Barcelona. CIBERDEM.) for guiding and supervising this project, and the Centre for Genomic Regulation for supporting this project through their International Master Internships Programme.

Annex

```
library(motifbreakR)
library(MotifDb)
library(BSgenome)
library(BSgenome.Hsapiens.UCSC.hg19)
library(SNPlocs.Hsapiens.dbSNP142.GRCh37)

data(motifbreakR_motif)

# read in Single Nucleotide Variants
pca.snps <-c("rs115077735","rs180980072","rs190513637",
             "rs34584161","rs386111","rs703977")

# import rsid snips "pca.snps"
snps.mb <- snps.from.rsid(rsid = pca.snps,
                          dbSNP = SNPlocs.Hsapiens.dbSNP142.GRCh37,
                          search.genome = BSgenome.Hsapiens.UCSC.hg19)

# execute motifbreakr
results <- motifbreakR(snpList = snps.mb, filterp = TRUE,
                       pwmList = motifbreakR_motif,
                       threshold = 5e-5,
                       method = "ic",
                       bkg = c(A=0.270182, C=0.2290216,
                             G=0.2297711, T=0.2710253),
                       BPPARAM = BiocParallel::SerialParam())

# calculate p-values
results <- calculatePvalue(results)

# filter by effect and pct
results <- results[results$effect == "strong" & (results$pctRef > 0.80 |
results$pctAlt > 0.80),]
```

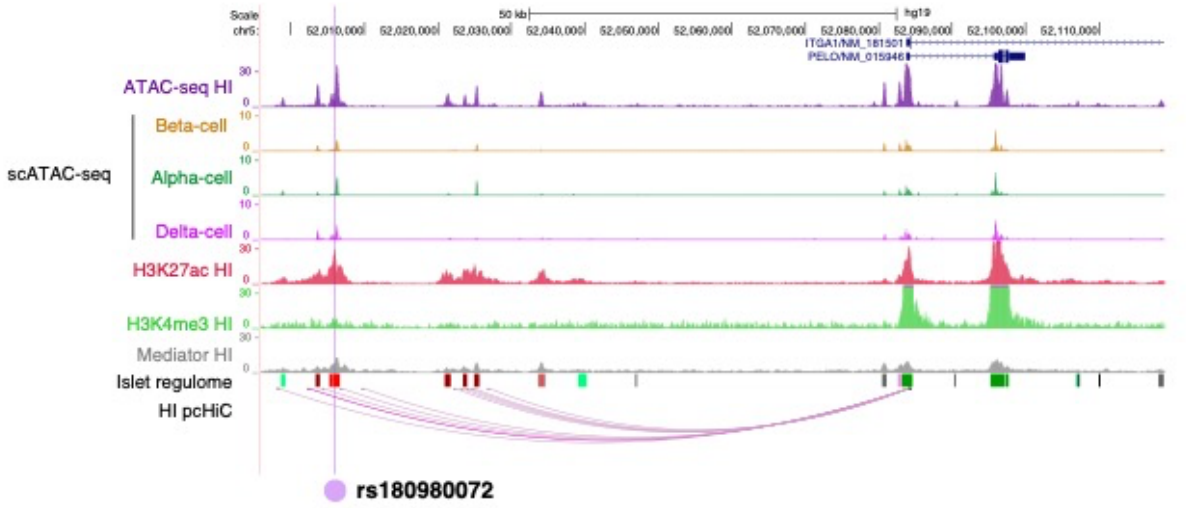
Annex Command 1

Annex Table 1. Complete list of top functional biological terms enriched in target genes assigned to islet cell-type selective enhancers.

Cell type	Gene-set library	Top functional biological terms for gene sets	P value	Adjusted P values	Combined score	Genes
		Various types of N-glycan biosynthesis	0.0016	0.498	18.95	CHST9,ALG9,ALG2,ALG3
		Maturity onset diabetes of the young	0.0094	1.000	13.72	NEUROD1,PDX1,SLC2A2,HES1
	KEGG 2021 Human	Type II diabetes mellitus	0.0097	1.000	10.57	KCNJ11,ABCC8,PRKCE,PDX1
		N-Glycan biosynthesis	0.0214	1.000	7.76	DPAGT1,ALG9,ALG2,ALG3
		Insulin secretion	0.0663	1.000	4.13	CAMK2B,CHRM3,RYR2,CAMK2D
		insulin-like growth factor I binding (GO:0031994)	0.0005	0.458	57.35	IGFBP1,IGFBP5,IGFBP4,ITGB4
		insulin-like growth factor binding (GO:0005520)	0.0009	0.458	43.52	IGFBP1,IGFBP5,IGFBP4,ITGB4
Beta cells	GO Molecular Function 2018	insulin-like growth factor II binding (GO:0031995)	0.0025	0.794	70.29	IGFBP1,IGFBP5,IGFBP4,IGFBP3,
		RNA polymerase II transcription corepressor activity (GO:0001106)	0.0094	1.000	13.72	TLE1,CITED2,CTBP1,ZMYND8
		glucocorticoid receptor binding (GO:0035259)	0.0112	1.000	26.43	NR4A2,NR4A1,NRIP1,YWHAH
		Islet endocrine cells in Pancreas	0.0231	1.000	5.62	NECAB2,CERKL,NKX2-2AS1,DDC
		Neuroendocrine cells in Lung	0.0257	1.000	6.09	KCNH2,OPRD1,TMEM132D,GPX2
	Descartes Cell Types and Tissue 2021	Chromaffin cells in Intestine	0.0894	1.000	3.30	RAB3B,CERKL,LRRC10B,LSAMP
		Stromal cells in Lung	0.2525	1.000	1.83	PDGFRA,MIR1245A,OCA2,MYOCD
		Stromal cells in Intestine	0.3749	1.000	1.19	EDAR,EPN3,PDGFRA,NPY
		alpha-Cell to beta-Cell Interconversion (Hypothesis)	0.0002	0.347	46.88	NEUROD1,CXCL12,MAF,MAFB
		L-cell: GCG, PYY and 5-HT Release	0.0042	1.000	19.17	CASR,FFAR4,GNAS,FFAR2
	Elsevier Pathway Collection	Transcription Factors in beta-Cell Neogenesis (Rodent Model)	0.0042	1.000	19.17	NEUROD1,MAFB,PDX1,PAX6
		Prostate Cancer	0.0101	1.000	7.68	GSK3B,CDKN1A,TGFB1I1,PTEN
		Endocannabinoids Role in Sleep Regulation	0.0118	1.000	20.20	DAGLA
		type B pancreatic cell differentiation (GO:0003309)	0.0002	0.703	139.27	PDX1,RFX3,INSM1,DLL1
		neurotrophin signaling pathway (GO:0038179)	0.0012	1.000	34.40	NTRK1,SORT1,MAGI2,DDIT4,
Alpha cells	GO Biological Process 2018	sympathetic ganglion development (GO:0061549)	0.0018	1.000	57.71	NELL2,FZD3,SEMA3A,INSM1
		positive regulation of endothelial cell migration (GO:0010595)	0.0021	1.000	13.79	NELL2,ADAM17,NOS3,ATOH8
		ganglion development (GO:0061548)	0.0037	1.000	38.16	NELL2,FZD3,SEMA3A,INSM1
		PANCREATIC ISLET	0.0001	0.006	12.19	USP6NL,SCOC,EHF,FAM159B
		BETA CELL	0.1092	1.000	2.38	EHF,TRIO,TMEM200A,TESK1
	ARCHS4 Tissues	ALPHA CELL	0.5788	1.000	0.54	ERRFI1,EHF,TMEM200A,RGSL1
		PREFRONTAL CORTEX	0.9998	1.000	0.00	PPP5D1,TRIO,ZMYND8,RGSL1
		CEREBRAL CORTEX	1.000	1.000	0.00	HPSE2,HDAC11,AQP4,LDLRAD4
		Insulin secretion	0.0037	1.000	14.29	CAMK2B,CHRM3,RYR2,CAMK2D
		Non-homologous end-joining	0.0081	1.000	30.68	FEN1,RAD50,XRCC4
	KEGG 2021 Human	GABAergic synapse	0.0303	1.000	7.08	KCNJ6,SLC12A5,GAD1,GNAI3
		Thyroid cancer	0.0315	1.000	9.60	NTRK1,TCF7L2,CDKN1A,RXRA
		Other types of O-glycan biosynthesis	0.0319	1.000	8.64	GALNT7,COLGALT2,GALNT3,GALNT2
		regulation of type B pancreatic cell development (GO:2000074)	0.0001	0.175	233.92	GSK3B,RHEB,RFX3,NKX6-1
		adenylate cyclase-activating adrenergic receptor signaling pathway (GO:0071880)	0.0013	0.762	40.62	ADRA1B,ADRA2A
Delta cells	GO Biological Process 2018	type B pancreatic cell differentiation (GO:0003309)	0.0018	0.762	72.73	PDX1,RFX3,INSM1,DLL1
		negative regulation of MAP kinase activity (GO:0043407)	0.0018	0.762	19.95	DUSP4,PTPN1,DUSP2,RGS14
		cellular response to corticosteroid stimulus (GO:0071384)	0.0021	0.762	44.38	BCL2L11,AKR1C3,SSTR2,NR3C1
		BETA CELL	0.0138	1.000	5.18	EHF,TRIO,TMEM200A,TESK1
		PANCREATIC ISLET	0.0260	1.000	4.33	USP6NL,SCOC,EHF,FAM159B
	ARCHS4 Tissues	PREFRONTAL CORTEX	0.2087	1.000	1.69	PPP5D1,TRIO,ZMYND8,RGSL1
		CINGULATE GYRUS	0.4593	1.000	0.79	APP,SCOC,HPSE2,HDAC11
		SPINAL CORD (BULK)	0.4948	1.000	0.71	HPSE2,AQP4,HS6ST1,ANTXR1
		Apoptosis-related network due to altered Notch3 in ovarian cancer WP2864	0.0012	0.661	17.73	VAV3,APP,SOCS3,CDKN1A
		Integrated breast cancer pathway WP1984	0.0046	0.806	9.03	ATF1,ZMYND8,ODC1,PTEN
	WikiPathway 2021 Human	Pathogenic Escherichia coli infection WP2272	0.0048	0.806	12.17	TUBAL3,TUBB,PRKCA,ACTB
		Bladder cancer WP2828	0.0064	0.806	12.77	CDKN1A,RASSF1,RPS6KA5,MYC
		Gastrin signaling pathway WP4659	0.0071	0.806	8.68	GSK3B,JUN,CDKN1A,MEF2B
		insulin-like growth factor I binding (GO:0031994)	0.0005	0.484	56.89	IGFBP1,IGFBP5,IGFBP4,IGFBP3
		insulin-like growth factor binding (GO:0005520)	0.0010	0.484	43.16	IGFBP1,IGFBP5,IGFBP4,IGFBP3
Acinar cells	GO Molecular Function 2018	insulin-like growth factor II binding (GO:0031995)	0.0026	0.832	69.77	IGFBP1,IGFBP5,IGFBP4,IGFBP3
		cadherin binding involved in cell-cell adhesion (GO:0098641)	0.0110	1.000	15.37	ANXA2,PAK4
		retinoic acid receptor binding (GO:0042974)	0.0136	1.000	15.67	NR4A2,MBD4,NCOA6
		ASPC1 PANCREAS TenPx29	0.0040	1.000	7.25	SH2D4A,ACY1,CD82,PWWP2B
		SNU719 STOMACH TenPx29	0.0142	1.000	5.38	ATF1,HERPUD2,SCOC,DHRS11
	CCLE Proteomics 2020	JHH4 LIVER TenPx40	0.0211	1.000	5.02	FKBP10,TGFB1I1,BHLHE41,JADE1
		LS180 LARGE INTESTINE TenPx27	0.0405	1.000	3.81	USP6NL,SCOC,PWWP2B,RTCB
		HEYA8 OVARY TenPx27	0.0466	1.000	3.75	RAB3B,BCAR3,SCOC,NRP2

Annex Table 2. List of candidate T2D causal variants estimated with motifbreakR and filtered by pct (pct > 0.8) and strength of the effect (effect = strong). Ref = reference allele for the variant; Alt = alternate allele for the variant.

chr	Locus	rsid	position	Effect Allele	Effect size	Ref Allele	Alt Allele	PctRef	PctAlt	Ref P-value	Alt P-value	AlleleDiff	Enhancer location	Enhancer type	TF motif	Cell type	Motif database
chr5	ITGA1-rs62357230	rs180980072	52005870	A	-0,19	A	T	0,97	0,82	1,72E-05	1,08E-03	-1,46	chr5:52005770:52006703	Active enhancers I	NeuroD1	alpha	HOMER
chr9	GLIS3-rs510807	rs115077735	4137685	A	0,16	G	A	0,85	0,99	2,64E-04	9,54E-07	1,20	chr9:4137032:4138334	Active enhancers II	HOXA1	acinar	HOCOMOCO
chr10	ZMIZ1-rs703972	rs703977	80944230	T	0,08	T	G	0,92	0,81	2,37E-05	5,09E-04	-1,07	chr10:80943759:80944788	Active enhancers I	Lhx1	beta	ENCODE-motif
chr11	KCNQ1-rs445084	rs386111	2933605	A	0,00	A	G	0,97	0,82	8,58E-06	1,84E-03	-1,33	chr11:2933530:2934500	Active enhancers III	NeuroD1	acinar	HOCOMOCO
chr13	RNF6-rs34584161	rs34584161	26776999	A	0,05	A	G	0,99	0,83	1,62E-05	1,12E-03	-1,51	chr13:26776649:26777631	Active enhancers I	Fos	beta	HOMER
chr13	RNF6-rs34584161	rs34584161	26776999	A	0,05	A	G	0,96	0,82	4,01E-05	9,37E-04	-1,54	chr13:26776649:26777631	Active enhancers I	Fos	beta	ENCODE-motif
chr17	ZZEF1-rs1043246	rs190513637	3977886	A	0,07	A	G	0,98	0,84	6,68E-06	6,57E-04	-1,37	chr17:3977635:3978684	Active enhancers I	NeuroD1	beta	HOMER
chr17	ZZEF1-rs1043246	rs190513637	3977886	A	0,07	A	G	0,97	0,81	1,62E-05	2,07E-03	-1,33	chr17:3977635:3978684	Active enhancers I	NeuroD1	beta	HOCOMOCO



Annex Figure 1. Human islet ATAC-seq, scATAC-seq across endocrine cell-types, and CHIP-seq datasets for H3K27ac, H3K4me3 and Mediator are shown across islet regulome annotations. Gene assignments based on pcHi-C data connecting the rs180980072-containing enhancer to *PELO* and *ITGA1* genes are shown as purple as purple arches.