Dissecting genetic regulatory mechanisms in human pancreatic islets to gain insights into type 2 diabetes pathophysiology

Goutham Atla

Faculty of Medicine, University of Barcelona

Dissecting genetic regulatory mechanisms in human pancreatic islets to gain insights into type 2 diabetes pathophysiology

Centre for Genomic Regulation (CRG)

Thesis submitted to University of Barcelona for the degree of Doctor of Philosophy

Doctorand
Goutham Atla

Directors
Jorge Ferrer Marrades i Sílvia Bonàs Guarch

Tutor
Josep Lluís Gelpí Buchaca

UNIVERSITAT DE BARCELONA

Abstract

Diabetes mellitus is a heterogeneous group of metabolic diseases characterized by impaired blood glucose homeostasis that affects more than 415 million people worldwide and is a leading cause of mortality. The most prevalent form of diabetes is Type 2 Diabetes (T2D) that accounts for 90% of diabetes cases. An interplay of environmental and genetic risk factors contributes to etiology of T2D via a progressive loss of pancreatic beta cell function coupled with insulin resistance. Genome Wide Association Studies (GWAS) identified more than 400 independent genetic loci associated with T2D risk, although the molecular mechanisms underlying these genetic signals remain poorly understood. A comprehensive understanding of gene regulation in human pancreatic islets and identifying the role of T2D risk variants on different components of gene regulation will enlighten our insights into T2D etiology.

In this work, we performed an in-depth characterization of human pancreatic islets transcriptional regulatory elements, attaining a greater granularity at transcriptional enhancers. We further identified glucose responsive enhancers which regulate glucose-dependent gene expression programs via three-dimensional chromatin interactions. This allowed us to gain insights into human islet transcriptional gene regulation and how glucose, a primary physiological stimulant of pancreatic islets, modulates human islet genome function.

We also generated comprehensive transcriptome annotations in human islets using short- and long-read sequencing data along with accurate maps of transcriptional start sites. This revealed islet-specific promoters, transcript isoforms and novel coding sequences. This underscored the importance of generating transcript models in disease relevant tissue to progress in the understanding of gene regulation.

Finally, these parallel efforts allowed us to create pioneer maps of genetic effects on human alternative splicing that revealed for the first time the noteworthy contribution of human islet mRNA splicing to T2D pathophysiology. These results have thus the potential to blossom in the discovery of novel T2D drug targets.

Acknowledgements

I will be grateful forever to Prof. Jorge Ferrer for giving me this opportunity and freedom to explore different research ideas.

I would like to thank all the current and past members of Ferrer lab. A special thanks to Irene Miguel Escalada for her persistent support throughout this journey. Thanks to Silvia Bonas Guarch for support, friendship and for teaching me human genetics. Thanks to all Ferrer lab members for amazing scientific discussions and constant feedback, without which this dissertation would have not been possible.

I would like to thank my parents, family members and my friends Kalyan, Naveen(s) and Victor for their moral support.

A big thanks to ZENCODE-ITN for funding a major part of my PhD work and for organising several scientific meetings which helped me to get a broad perspective of genomics research.

Table of Contents

List of figures

List of abbreviations and terms.

CAGE - Cap analysis of gene expression

ChIP – chromatin immunoprecipitation

ChIP-seq – chromatin immunoprecipitation sequencing

DNA – deoxyribonucleic acid

CDS - Coding sequence

CPP - Causal posterior probability

DIS - Disease impact score

ENCODE – encyclopedia of DNA elements

eQTL – expression quantitative trait locus

eRNA – enhancer RNA

GWAS – genome-wide association study

H3K27ac – histone 3 lysine 27 acetylation

H3K27me3 – histone 3 lysine 27 tri-methylation

H3K36me3 – histone 3 lysine 36 tri-methylation

H3K4me1 – histone 3 lysine 4 mono-methylation

H3K4me3 – histone 3 lysine 4 tri-methylation

IQR – Interquartile range

LD – linkage disequilibrium

NGS – next-generation sequencing

pcHi-C – promoter capture Hi-C

PIC – pre-initiation complex

PIR – promoter-interacting region

RBP - RNA-binding protein

RNA – ribonucleic acid

RNA Pol II – RNA polymerase II

RNA-seq – RNA sequencing

sQTL - Splicing quantitative trait loci

TAD – topological associating domain

TF – transcription factor

TFBS – transcription factor binding site

TC - Tag cluster

TSS – transcription start site

TWAS - Transcriptome-wide association study

1. Introduction

1.1. Pathophysiology of Type 2 Diabetes

1.1.1. Health burden of diabetes and sub-classification.

Diabetes mellitus (DM) is a heterogenous metabolic condition characterized by hyperglycemia that affects more than 415 million people around the world (InternationalDiabetesFederation, 2019) and is a leading cause of mortality. The alarming increase in the prevalence of diabetes speaks to limited preventive strategies and poor disease management, which results in long pre-diagnostic periods, eventually leading to serious life-threatening complications (cardiovascular disease, renal failure, blindness or lower limb amputation) (Feero et al., 2010). This major global health emergency requires significant progress in (i) the early identification of individuals at high risk and (ii) the improved individual response to available therapies. Thereafter, characterizing the array of molecular mechanisms with the most significant contributions to diabetes aetiology is pivotal to determine actionable components that will set the basis for refined preventive and therapeutic strategies.

The most prevalent form of diabetes is adult-onset type 2 diabetes (T2D) that accounts for nearly 90% of all diabetes worldwide (InternationalDiabetesFederation, 2019). The interplay between genetic and environmental factors has been reported to both influence T2D pathogenesis, and pancreatic islet dysfunction coupled with obesity-related insulin resistance(DeFronzo, 2004; Feero et al., 2010). Other common forms of diabetes are type 1 diabetes (T1D) that accounts for 5-10% of all diabetes (Daneman, 2006)and is largely driven by autoimmune destruction of pancreatic islets (Katsarou et al., 2017). Type 1 diabetes usually presents in the childhood along with circulating islet-cell antibodies. Finally, rarer monogenic forms, including maturity-onset diabetes of the young (MODY) or neonatal diabetes comprise 1-5% (Misra and Owen, 2018). Monogenic forms of diabetes share some hallmarks with T2D, such as mutations in genes that are essential for pancreatic islet function and identity. Taken together, although disparate pathological processes converge in T2D progression as aforementioned, pancreatic islets are central to the T2D pathogenesis and that of other diabetes forms.

1.1.2. Human pancreatic islets: organization, function and implications in diabetes.

The bulk of the pancreatic tissue is largely formed by the exocrine compartment that produces and delivers digestive enzymes to the gut. The endocrine compartment is enclosed in the islets of Langerhans and embodies a comparatively much smaller portion of the pancreas. However, endocrine islet cells are essential to maintain blood glucose homeostasis (Segerstolpe et al., 2016). Pancreatic islets are composed of five different cell-types: glucagon producing alpha cells, insulin producing beta cells, somatostatin producing delta cells, pancreatic polypeptide (PP) secreting or gamma cells, and ghrelin producing epsilon cells(Segerstolpe et al., 2016). Nevertheless, the core of pancreatic islets draws upon beta cells, which account for 50-70% of the islets (Dolenšek et al., 2015)Importantly, reduction of beta cell mass and dysfunction are key players in the development of diabetes(Kahn et al., 2006).

Pancreatic beta cells secrete insulin upon sensing increased glucose in the bloodstream (Figure 1.1). Glucose enters the beta cell through the GLUT2 transporters, is immediately phosphorylated by glucokinase encoded in the *GCK* gene, and subsequently metabolized. This result in an increase in the ratio of ATP to ADP. Elevated levels of cytosolic ATP close ATP-sensitive potassium (K_{ATP}) channels and leads to membrane depolarization, which stimulates calcium influx through voltage-dependent Ca^{2+} channels (Ashcroft and Rorsman, 2012). The accumulation of cytosolic Ca^{2+} triggers the insulin release that will promote the uptake of blood glucose by peripheral organs such as liver, skeletal muscle and adipose tissue.

Defects on insulin secretion are the main culprit of rare monogenic forms of diabetes, such as MODY or neonatal diabetes, but also of common forms like T2D, in this case in the context of obesity-associated insulin resistance. Neonatal diabetes is often caused by impaired beta cell depolarization due to mutations in the K_{ATP} channel that causes insulin secretion mis-regulation (Flanagan et al., 2009; Hattersley and Ashcroft, 2005) The majority of the cases of familial young-onset diabetes account for mutations in

Figure 1. 1 Glucose stimulated insulin secretion.

Adapted from *(León and Stanley, 2007)*

transcription factors (*HNF1A*, *HNF4A* and *HNF1B* genes), which determine beta cell fate and function, and the glucokinase *GCK* gene. Loss-of-function mutations in the *GCK* gene that results in decreased insulin secretion and gives rise to a mild hyperglycemia, which can be managed by diet alone. Transcription factor-associated mutations, mainly present in the *HNF1A* gene, result in a progressive decline of the beta cell function and hereby a deficit of insulin secretion(Shih et al., 2001; Yamagata et al., 2002; Yang et al., 2002) . Reduced beta cell function followed by the development of insulin resistance in muscle, liver and adipose tissues has also crucial roles in T2D pathogenesis (DeFronzo et al., 2015). Of note, T2D was considered for a long time as a disease mainly driven by insulin resistance associated with obesity, but the emergence of Genome Wide Association Studies (GWAS) has been instrumental for a paradigm shift. To date, more than 400 independent genetic signals have been identified to be associated with T2D risk (Mahajan et al., 2018a; Vujkovic et al., 2020). Although, most of these risk genetic variants are in non-coding regions, as we will discuss later on, genetic studies have revealed the role of beta cell function and identity in T2D pathogenesis (Gaulton 2010, Pasquali 2014 (Bonnefond et al., 2010; Feero et al., 2010; Thomsen et al., 2018). Thus, gaining insights into the mechanistic

underpinnings of gene regulatory networks that govern pancreatic islet beta cell development and function, and the interplay with T2D risk genetic signals will enlighten our understanding of disease pathogenesis and create new avenues for improved preventive and targeted therapeutic strategies.

1.2. The functional human genome

The human genome contains over 3.2 billion base pairs of nucleic acids as deoxyribonucleic acids (DNA) that encode the set of instructions for organism's development and function. Due to the smaller size of the nucleus, DNA is stored in a compact form, wrapped around nucleosomes, which are the basic units of chromatin (Figure 1.2). A nucleosome is composed of 2 copies of histone proteins H2A, H2B, H3 and H4, referred to as the histone octamer. Depending on the distance between nucleosomes, DNA can exist as (i) densely packed heterochromatin or (ii) loosely packed euchromatin. Heterochromatin was initially thought to be biochemically inactive while euchromatin is the active form, but it has been later suggested that the biochemical activity of DNA is not tightly subjected to this organization(Gilbert et al., 2004).

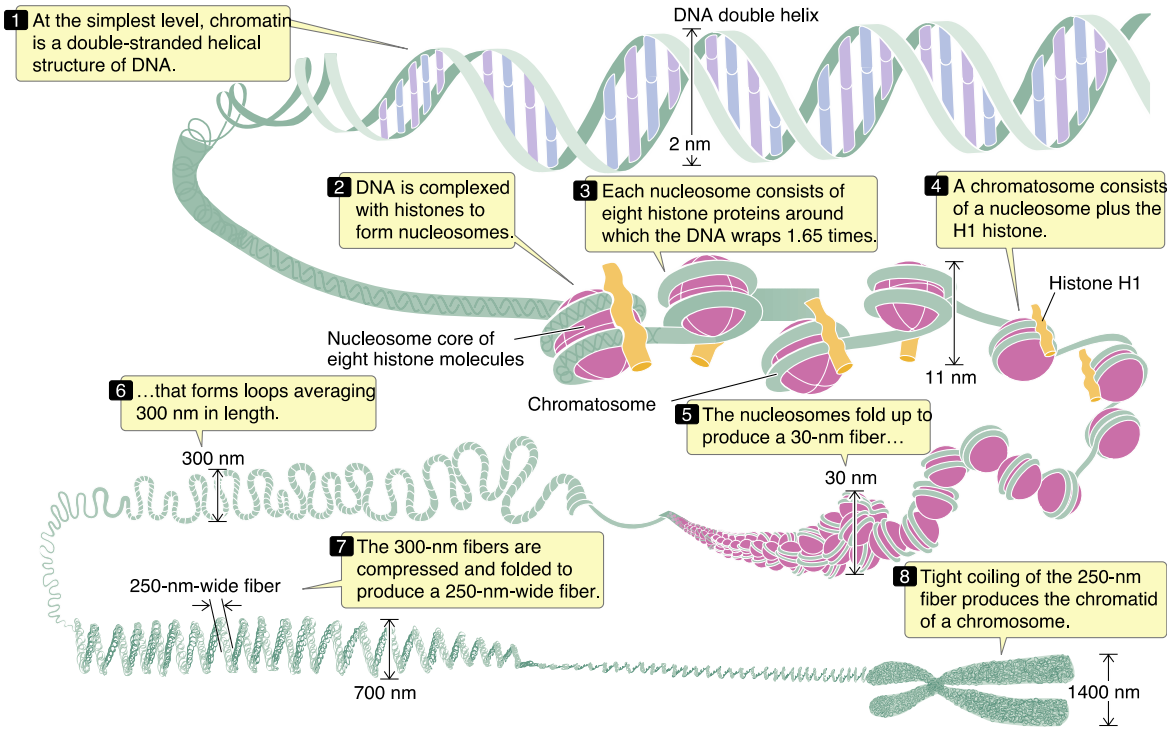

Figure 1. 2 Organization of DNA inside a nucleus.

Adapted from Pierce, Benjamin. Genetics: A Conceptual Approach, 4th edition.

The central dogma of molecular biology (CRICK, 1970) described a unidirectional flow of information where DNA is transcribed into ribonucleic acid (RNA), a process referred to as transcription, and then RNA is used as a template for protein synthesis, a process referred to as translation. One notable exception to this theory is reverse transcription of RNA by retroviruses, where RNA is used as a template for cDNA synthesis. Another exception has been the recognition that a large number of RNAs do not appear to encode for protein but may nevertheless exert regulatory functions (Carninci et al., 2005; Guttman et al., 2009; Mercer et al., 2012).

As DNA is tightly wrapped around nucleosomes, the genomic sequence is not usually accessible for transcription. Transcription of the human genome is a highly coordinated process that is preceded by the creation of an active chromatin environment. This active chromatin state is defined by the increasing accessibility of DNA sequences that are wrapped around histones by several proteins that initiate and stimulate transcription. Chromatin accessibility is primarily facilitated by post-translational modifications (PTMs) of histone proteins, in particular in the lysine residues close to amino acid termini of H3 and H4. The PTMs of histones is carried out by chromatin-modifying proteins such as histone acetyltransferases (HAT), histone deacetylases (HDAC), lysine methyltransferases (KMTs) and lysine deacetylases (KDMs). Depending on the type of PTMs, the underlying chromatin state may become more active or less active, and they are also associated with the disparate purposes of active chromatin, as we will discuss in detail in chapter 1.3.

1.2.1. Impact of alternative splicing on protein and transcriptome diversification.

The protein-coding parts of the human genome are transcribed to messenger RNA (mRNA), a process interchangeably termed gene transcription or gene expression. The resulting mRNA then can be translated to synthesize proteins that perform a vast number of biological functions. Of note, some of the RNAs may not encode for protein sequences. They are named non-coding RNAs (Carninci et al., 2005) and play key roles in genome regulation.

The human genome roughly contains 22,000 to 25,000 protein-coding DNA sequences in each tissue. However, the vast number of encoded proteins that the genome gives rise is far in excess of the original number of genes. A typical mRNA contains both exonic and intronic regions. Introns are not required for protein synthesis and are thereby spliced out from pre-mRNA molecules before being translated to proteins, by a mechanism known as alternative splicing(Sanford and Caceres, 2004) (Figure 1.3). Then, the exonic fraction of mRNA molecules are merged to form a template for protein synthesis. While many exons are part of pre-mRNA, some exons are alternatively spliced, which increases the breadth of mRNAs sequences and contribute to proteome diversity.

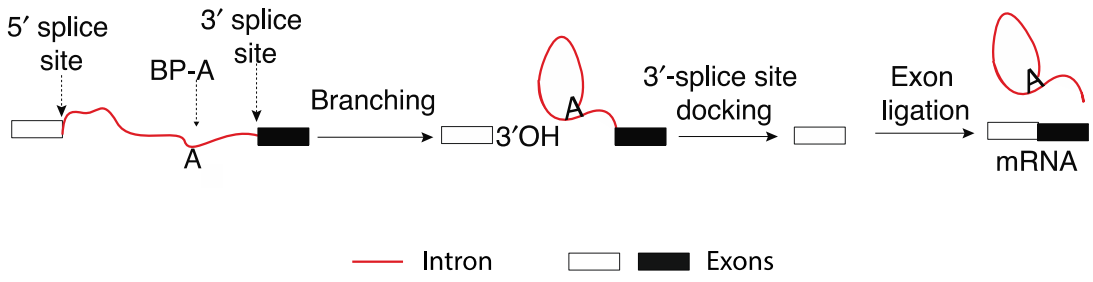

Figure 1. 3 A schematic of alternative splicing showing intron excision in a lariat fashion.

Alternative splicing not only involves differential usage of exons but also differential usage of splice sites and intron retention events (Figure 1.4).

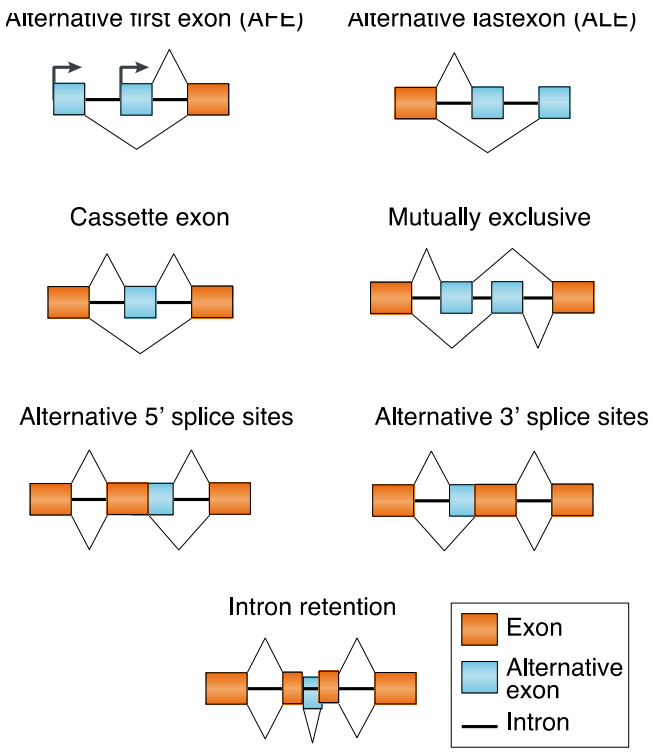

Figure 1. 4 Different types of alternative splicing events.
Adapted from Scotti and Swanson, 2016, nature reviews genetics.

Understanding splicing regulation is fundamental to gain insights into disease pathophysiology as many diseases are associated with mis-regulation of alternative splicing(Faustino and Cooper, 2003).

Alternative splicing machinery

Splicing is carried out by the spliceosome, a megadalton complex of U1, U2, U3, U4/U6 and U5 small ribonuclear proteins (snRNPs) along with multiple protein factors(Nagai and Fica, 2017; Wahl et al., 2009). Binding of snRNPs to pre-mRNA and stabilization of spliceosome is governed by intron and exon "definitions" of splicing. Intron definition of splicing predominates in organisms that have shorter introns, e.g., in yeast and many invertebrate species. In contrast, exon definition is preponderant in vertebrates, which usually have longer introns (median length of 1kb). This splicing complex is formed onto the pre-mRNA molecule in a stepwise manner (Figure 1.5). The introns contain splice site recognition sequences, the 5' and 3' splice sites (5'ss and 3'ss, respectively), that are GU and AG consensus sequences, respectively, along with a branch point sequence. The assembly of the splicing machinery begins with the recognition and binding of U1 snRNA through base-pairing

to the 5’ss of introns. This process concurs with the binding of SF1/BBP protein and U2 auxiliary factor (U2AF) to the branch point sequence and the polypyrimidine tract downstream of the branch point sequence. The U2AF has two subunits, a 65 kDa which interacts with SF1 and a 35 kDa subunit which recognizes and binds to the 3’ss. This yields the spliceosome complex ‘E’, which is a crucial step for the initial recognition of 5’ss and 3’ss. The U2 snRNP then displaces SF1 at the BPS in an ATP-dependent manner, resulting in the ATP-dependent complex (complex 'A'). This process is stabilized by SF3a and SF3b, which are heteromeric protein complexes of U2 snRNP. Then, U4, U5 and U6 are recruited as a tri-snRNP complex to from complex ‘B’, a catalytically inactive complex. The dissociation of U1 and U4 is next triggered by a series of conformational changes catalyzed by RNA helicases Brr2, Snu1 14, Prp3, among others, and this progresses to the activated complex B (‘B*’) spliceosome. Then, two subsequent catalytic steps occur. The first one releases U2-associated proteins SF3a and SF3b, exposing the BPS that attacks the 5’ss. This results in a free 5’ exon and an intron lariat intermediate ‘C1’ complex. The second catalytic step involves the 3’OH of the 5’ of the exon that attacks the 3’ss to generate the ‘C2’ complex. Splicing concludes by the dissociation of the remaining snRNPs, the ligation of the exons and the rapid degradation of the intron lariat (Figure 1.5).

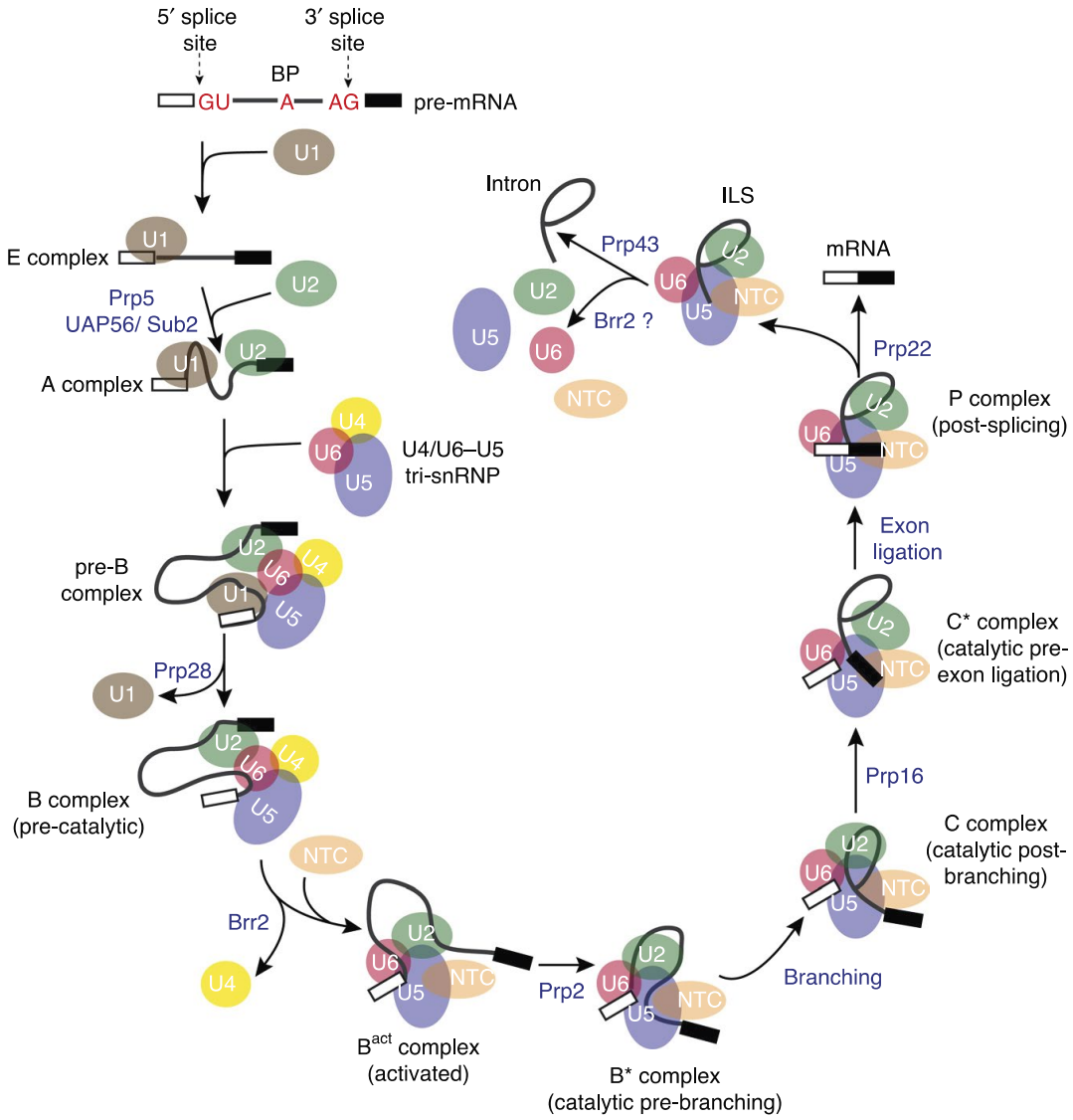

Figure 1. 5 A stepwise assembly of spliceosome.
Adapted from Kiyoshi Nagai et al. 2017

Regulation of alternative splicing

As splice site sequences are degenerate across the genome, their recognition and selection by the spliceosome is accompanied by flanking regulatory sequences known as 'splicing enhancers' and 'silencers' that could be located either in introns or in exons (Singh and Valcárcel, 2005) (Figure 1.6). Along with the snRNPs described above, the primary regulators of splicing are RNA recognition motif (RRM)-containing proteins (SR proteins) and SR-related proteins, which contain regions of alternating serine (S) and arginine (R) residues. SR proteins tend to enhance splicing by binding to purine-rich exonic regions of pre-mRNA that could act as exonic enhancers, thus

recruiting U1 splicing factor and U2 auxiliary factor to the 5' and 3' splice sites, respectively, and promoting the inclusion of the respective exon(Fu and Ares, 2014). In addition to SR proteins, other types of RNA binding proteins (RBPs) participate in splicing regulation, such as the heterogeneous ribonucleoprotein (hnRNP) family as well as RBPs containing RRN *k* homology domain (KH), zinc-fingers and many others (Lunde et al, 2007). The hnRNP binding on splicing silencer motifs usually have a repressive effect on splicing i.e. they act to antagonize the effects of SR proteins, and prevent exons to be included in the mRNA sequence. Polypyrimidine tract-binding proteins (PTB), which prefer to bind to polypyrimidine sequences, are the best-characterized of hnRNPs (Llorian et al., 2010; Xue et al., 2009)

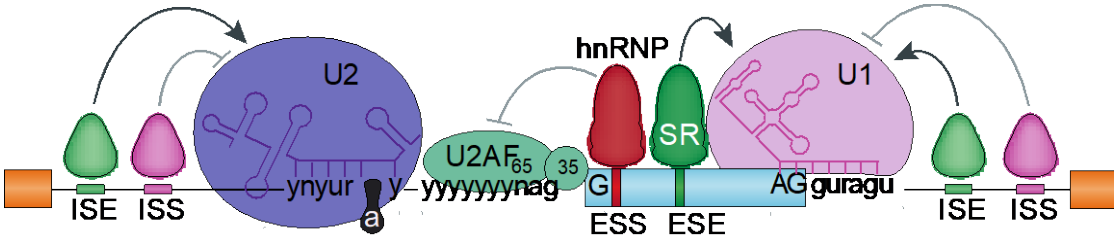

Figure 1. 6 Auxiliary splicing regulators.
A schematic representing splicing regulators. *cis*-acting Intronic and exonic splicing enhancers (ISE, ESE) and intronic and exonic splicing silences (ISS, ESS) are bound by *trans*-acting splicing regulators and govern the splicing reaction. Figure adapted from Scotti and Swanson, 2016, nature reviews genetics.

Even though alternative splicing is a ubiquitous mechanism of gene regulation, the regulation of alternative splicing is cell-type specific as it is a key player in development and tissue identity. The ubiquitous component of splicing is defined by very strong splice site sequences. In contrast, tissue-specific regulation of alternative splicing stems from weak splicing motifs. Additionally, cis-regulatory elements on the mRNA sequence enhance or suppress splicing depending on the binding of trans-acting RBPs, whose expression is regulated according to the cellular context(Baralle and Giudice, 2017; Pan et al., 2008; Wang et al., 2008).

Both gene transcription and alternative splicing are highly tissue-specific processes and previous studies indicate that they are interlinked at various levels (Drexler et al., 2019; Kornblihtt et al., 2013; Luco et al., 2011; Vargas et al., 2011).

Long noncoding RNAs

Along with protein-coding genes, the human genome also encodes for a plethora of non-coding RNAs as extensively reported in literature (Carninci et al., 2005; Guttman et al., 2009; Mercer et al., 2012). One class of non-coding RNAs are long non-coding RNAs (lncRNAs) that are arbitrarily defined as mRNA longer than 200bp that do not encode for proteins. Like protein-coding genes, lncRNAs are also poly-adenylated and undergo alternative splicing. The number of lncRNAs encoded by the human genome ranges from 15,000 to 25,000, depending on the source of the annotations. The majority of lncRNAs remain without a known clear role, but the regulatory potential of lncRNA has already been revealed but the characterization of some individual examples. One of the most notable examples is *xist,* a lncRNA that orchestrates mammalian X-chromosome inactivation (Penny et al., 1996). Since then, there have been many studies elucidating the role of lncRNA in diverse cellular function such as transcriptional regulation(Akerman et al., 2017), cell reprogramming(Sherstyuk et al., 2018), chromatin remodelling (Akhtar; Quinodoz and Guttman, 2014) and their mis-regulation has been reported to be implicated in several human diseases (Morán et al., 2012; Scheuermann and Boyer, 2013) including cancer (Lin and Yang, 2017).

Reference transcript annotations

To gain insights in the molecular underpinnings of the plethora of coding and non-coding transcripts in the human genome, it is essential to characterize them across tissues and cell-types. This fueled large initiatives to generate comprehensive catalogues of reference transcripts. Initially, human and vertebrate analysis and annotation (HAVANA) team at Sanger institute aimed at curating gene models from cDNAs, expressed sequence tags (ESTs) and protein sequences. The advent of the ENCODE project (Consortium, 2004), boosted the creation of a dedicated team, GENCODE (Harrow et al., 2006) strived for providing human gene annotations through the incorporation of manually curated HAVANA gene models along with experimental validations. In parallel, the Ensembl (Flicek et al., 2011) project took off with the purpose of annotating chordate species by the integration of gene annotations, multiple alignments, gene homology relationships and regulatory annotations. Currently, both GENCODE and Ensembl gene models are updated and cross-

referenced in parallel, and HAVANA team continues to manually curate their resulting gene annotations. FANTOM-CAT(Hon et al., 2017) is another resource that pursues creating a catalogue from of lncRNAs across tissues and cell types through the integration of several reference transcriptome annotations and by accurately determining the transcription start sites.

Although reference transcriptome annotations serve the purpose of providing an overall overview of abundant protein-coding and lncRNA genes, they fall behind in procuring novel insights into human disease pathophysiology that stem from the vast majority of tissue-specific and context-dependent gene expression programs. As an illustrative example, re-analysis of cancer genome atlas (TCGA) data in the MiTranscriptome project (Iyer et al., 2015) revealed that 79% of the 58,648 lncRNAs identified in this study were previously unannotated in reference transcriptome databases. Similarly, a major fraction of lncRNAs detected in human pancreatic islets were previously unannotated, most notably those that were tissue-specific (Akerman et al., 2017; Morán et al., 2012). Another remarkable example that re-analyzed 21,504 samples from the sequence read archive (SRA) database determined that ~65,000 junctions consistently identified in more than 1000 samples were not represented in reference annotations (Nellore et al., 2016).

There are several reasons for reference transcriptome annotations to fail to capture the transcriptome complexity observed in individual studies. First, transcriptome complexity could arise from genetic variation across individuals and also due to the inter-cellular somatic variation. Second, many of the transcriptional programs that regulate gene expression are cell-type specific and context-dependent, and therefore less studied tissues tend to have more poorly annotated transcripts. Third, post-transcriptional modifications such as mRNA processing and degradation are also cell-type and context-dependent. Thus, identifying gene and transcript models in disease-relevant tissues across multiple individuals will allow us to interrogate pertinent tissue-specific gene regulatory programs in order to expand our understanding of disease pathophysiology.

1.2.2. The non-coding genome

One of the main unanticipated results that the Human Genome Project revealed, and other large-scale initiatives substantiated (Dunham et al., 2012) was that more than 98% of the DNA sequence does not encode for protein-coding sequences. Until recently, the role of the vast non-genic fraction of the genome was almost entirely unknown. Non-coding genome was assumed to have no impact on the development and function of an organism, and therefore, to be under no selective pressure. Now, it is well established that the non-genic part of the genome contains regulatory elements that control gene transcription in a spatio- and temporal manner. This repertoire of non-coding elements that governs gene transcription are primarily composed of enhancers, promoters, silencers and insulators or boundary elements(Maston et al., 2006; Shlyueva et al., 2014). Promoter elements are short DNA sequences upstream from the gene transcription start site (TSS) and direct transcription initiation(Andersson and Sandelin, 2020; Haberle and Stark, 2018). In contrast, enhancer elements can be located several hundreds of kilobases away from their endogenous target genes, and thus, can activate or amplify gene transcription initiated by promoter sequences independently of their relative distance and orientation. Enhancers guide target gene expression by looping to the corresponding gene promoter with the assistance of an ensemble of transcriptional cofactors, such as Mediator, structural proteins like cohesin, CTCF and YY1 (Kagey et al., 2010; Weintraub et al., 2017). Enhancers and promoters are highly tissue-specific, harbour sequence determinants that recruit lineage-specific transcription factors (TFs), and hence coordinate genome activity in development, cell-type and tissue identity and disease.

Genome-wide identification of cis-regulatory elements

In contrast to heterochromatin, which is densely packed by nucleosomes, regions across the genome that usually embed regulatory elements and transcribed gene bodies are depleted of nucleosomes (Lee et al., 2004). These nucleosome depleted regions, known as ‘accessible chromatin’ or ‘open-chromatin regions’, are bound by TFs, RNA-polymerases and structural proteins that promote a higher-order genome organization that is central to gene transcriptional regulation (Figure 1.7). Several methods have been developed to identify accessible chromatin regions, but the rationale behind all of them is identifying DNA sequences that are susceptible to enzymatic methylation or

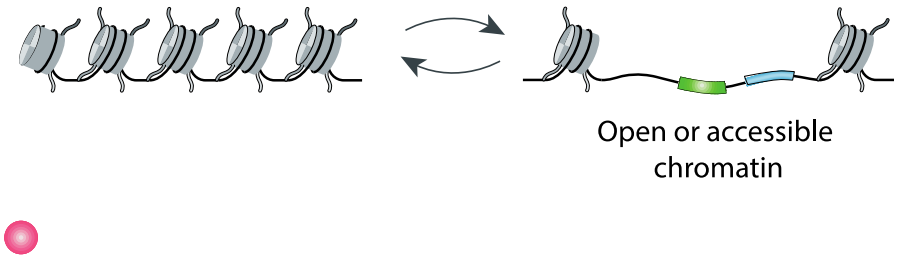

Figure 1. 7 Accessible chromatin.
DNA is usually wrapped around nucleosomes. The underlying DNA sequences is accessible by DNA binding proteins as and when necessary by opening of chromatin. Figure adapted from Shlyueva et al., 2014, Nature reviews genetics

cleavage. The first large-scale approach to identify open-chromatin regions was performed by hybridizing captured ends from DNA sequences sensitive to Deoxyribonuclease I (DNase I) endonuclease cleavage to tiling microarrays(Crawford et al., 2006). This method was initially implemented by the ENCODE (Encyclopedia Of DNA Elements) consortium in two cell-lines, CD4+ cells and a cyclic B-lymphoblastoid cell line, and was only capable to capture 1% of the selected genomic regions (Consortium, 2004). Open-chromatin regions identified by ENCODE are overwhelmingly located at gene transcription start sites, first exon, first intron and CpG islands, and average gene expression of nearby genes (<1kb) tend to be higher. The authors also observed that a fraction of the identified open-chromatin regions are specific to one or the other cell line, suggesting that accessible chromatin could be cell-type specific. Genome-wide maps of open-chromatin regions emerged with the availability of tiled genome-wide microarrays(Boyle et al., 2008). This first genome-wide method identified around 100,000 open-chromatin regions that represent ~2.1% of the human genome in CD4+ cells. While the observations remained consistent with the initial tiling array-based method, genome-wide maps by high-throughput sequencing also enabled examining nucleosome positioning. The major milestone in chromatin accessibility profiling was attained by the ENCODE project by high-throughput sequencing of open-chromatin regions in 125 different human cell and tissue types (Thurman et al., 2012). This landmark study identified around 2 million open-chromatin regions across cell-types. Interestingly, 3% of them are unique to a single cell-type, and the bulk of chromatin accessible regions (95%) are distal (>2.5kb) from the TSS and tend to be cell-type specific overall. This paper also revealed that,

while only 2-3% of the total genome is accessible, more than 90% of the TFs exclusively bind to open-chromatin regions. This indicates that the cis-regulatory elements are often found in an open-chromatin state.

Subsequently other methods were developed to map accessible chromatin, include FAIRE (formaldehyde assisted isolation or regulatory elements), which was first implemented at genome-wide scale to study human pancreatic islet regulatory elements and diabetes-relevant regulatory variants (Gaulton et al., 2010). More recently, Assay for Transposase Accessible chromatin using sequencing (ATAC-Seq) has become the most popular method to interrogate open chromatin regions(Buenrostro et al., 2013). It uses Tn5 transposase loaded with sequencing adapters, such that it simultaneously cleaves and ligates sequencing adapters to DNA fragments from open-chromatin regions. The fact that does not involve complex and time-consuming protocols (~2-3 hours), that can be applied to very few cells (~ 50,000 cells) and that has a very high signal to noise ratio explain the rise in popularity of this method. In addition, ATAC-Seq not only informs the open chromatin regions but also gives nucleosome patterns and the digital footprint of the binding of TFs to accessible chromatin regions.

While open-chromatin region maps allow identifying the landscape of accessible chromatin in a given cell-type/tissue, further annotation of the underlying chromatin states is crucial to learn the regulatory functions that these regions actually undertake.

Promoter elements

Gene transcription initiates at a precise location in the genome, the Transcription Start Site (TSS), which is the first base of the transcribed region. The RNA Polymerase along with General Transcription Factors (GTFs) forms a Pre-Initiation Complex (PIC) on the genomic region 50 bp upstream and 50 bp downstream from the TSS also known as "core promoter"(Smale and Kadonaga, 2003). This binding is facilitated by the underlying DNA sequence of the core-promoter. Traditionally, it was thought that all mammalian core-promoters have a TATA-box element and an initiator element, but genome-wide analysis of mouse and human promoter architecture revealed that only a fraction of them have a clear TATA-box (Carninci et al., 2006). While core-promoter elements help to assemble the transcription machinery and in the TSS recognition, the

rate at which genes are transcribed is determined by other proximal and distal regulatory elements. These regulatory elements are susceptible to TF binding and influence the dynamics of core-promoters i.e., transcription initiation or elongation rates. Proximal regulatory elements are usually found within 2.5kb upstream of the core-promoter and they are bound by one or more TFs. The 2.5kb genomic region around the gene TSS that spans the "core-promoter" and the "proximal-promoter", defines what we termed the "promoter" sequence.

Various experimental methods identify gene TSS based on sequencing the 5' end of nascent RNAs, which allows us to map promoter sequences in a given cell-type/tissue. One of the most widely adopted variants of this technique is Cap Analysis of Gene Expression (CAGE-Seq)(Takahashi et al., 2012), although many other complementary approaches also exist (Kruesi et al., 2013; Kwak et al., 2013; Lam et al., 2013; Mayer et al., 2015). CAGE data showed that gene sequences either embody closely spaced clusters of multiple TSS (typically <100bp away from each other) or a unique TSS at a single base-pair position. This observation enabled the subclassification of promoter sequences into 'broad' and 'sharp', that also involved different genomic contexts; e.g overrepresentation of TATA-boxes in sharp promoters, and CpG islands in broad promoters (Carninci et al., 2006). Developmentally active genes have been associated with broad promoters while tissue-specific genes are characterized by an enrichment in sharp promoters.

While 5'-end sequencing of nascent RNAs is the most popular method for TSS detection, the underlying chromatin state also has distinct signatures that identifies active transcription. Chromatin Immunoprecipitation Sequencing (ChIP-Seq) of histone modification marks showed that promoter sequences are highly enriched for tri-methylation of histone H3 lysine 27 residues (H3K27Ac) and tri-methylation of histone H3 lysine 4 residues (H3K4me3)(Heintzman et al., 2007; Roh et al., 2004) (Figure 1.8).

Transcriptional enhancers

Enhancers are distal regulatory elements that were first discovered in the early 80s as ~300-1000 bp regions that are upstream from the TATA-box, and amplify nearby gene transcription(Banerji et al., 1981; Benoist and Chambon, 1981; Grosschedl and Birnstiel, 1980). Enhancers act independently of their relative distance and orientation from their target gene. The tissue-specific nature of enhancer elements is identified by

the underlying DNA sequence context, which embeds specific sequence motifs that facilitate the association of lineage-specific TFs that are essential for enhancer activity (Banerji et al., 1983; Kundaje et al., 2015) . Cooperative occupancy of TFs is central to nucleosome repositioning and thus, regulation of enhancer activity (Spitz and Furlong, 2012; Tillo et al., 2010). Of note, chromatin accessibility at enhancers not always involve the cooperative binding of TFs; indeed, 'piooner' TFs also known as master regulators can directly bind nucleosomal DNA and ease enhancer activation by aiding in the subsequent association of other lineage TFs (Lambert et al., 2018; Magnani et al., 2011; Vaquerizas et al., 2009; Zaret and Carroll, 2011). Because of the complex interplay with TFs, transcriptional enhancers rely on a specific "grammar" that is essential for their activation and function. Two major models have been proposed for the enhancer lexicon. The first one is the "enhanceosome" model that proposes a very strict motif architecture in terms of motif organization and order in the enhancer DNA sequence. The other model is the "billboard" model where the underlying motif combination, order and spacing is flexible(Long et al., 2016; Thanos and Maniatis, 1995). Previous TFs co-operativity model was expected to be mirrored in the DNA sequence through a motif composite that would include the complete collection of motifs of each of the TFs in rigid order. However, large-scale studies of TF-TF interactions and their DNA binding preferences suggested that TF cooperative occupancy primarily occurs through a novel consensus motif that would be otherwise weakly bound by each individual TF(Jolma et al., 2015). Despite of this, genome-wide identification of transcriptional enhancers is primarily dependent on the underlying epigenome state. Active enhancers are distinctly marked by H3K27Ac and tri-methylation of histone H3 lysine 1 residues (H3K4Me1)(Heintzman et al., 2007, 2009) (Figure 1.8). Of note, subsets of enhancers that lack H3K27Ac enhancer hallmark and show a large presence of H3K27Me3, a repressive mark, have been identified as inactive or poised enhancers (Creyghton et al., 2010). Inactive enhancers can be fully activated by external stimuli and are thus, preferentially occupied by signal-dependent transcription factors(Heinz et al., 2015) (Figure 1.8).

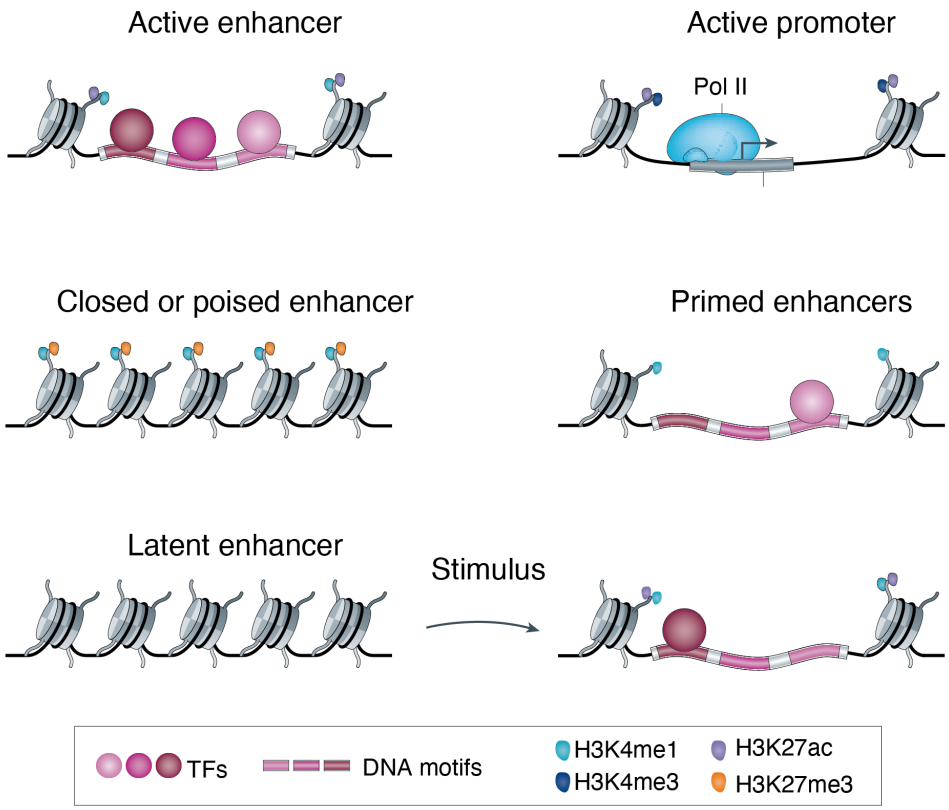

Figure 1. 8 Properties of enhancers and promoters.
Active enhancers are primarily marked by H3K27Ac and H3K4me1. Active promoters are marked by H3K27Ac and H3K4me3. Closed or poised enhancers are marked by H3K27me3 and H3K4Me1 while removing H3K27me3 makes poised enhancers into primed enhancers. Figure adapted from Shlyueva et al., 2014, Nature reviews genetics.

The number of active enhancers in a single cell type is far in excess of that of active genes, suggesting a synergistic or combinatorial activity of multiple different enhancers to control gene expression according to the cellular context. Another relevant feature of enhancer function relies on their distribution across the genome that is far from uniform. Cell-type specific genes have been reported to have a particular traction for enhancer-rich regions, variably known as clusters of open regulatory elements (COREs)(Gaulton et al., 2010), super-enhancers (Whyte et al., 2013), stretch enhancers(Parker et al., 2013), enhancer clusters(Pasquali et al., 2014; Whyte et al., 2013) or enhancer hubs(Miguel-Escalada et al., 2019). Mounting evidence showed that clusters of enhancers are highly occupied by lineage-specific TFs and transcriptional

co-activators such as Mediator complex, and thus are key players of cell-type identity and function.

Large efforts have been dedicated to the genome-wide identification of cis-regulatory elements across tissues. Early initiatives such as the NIH Roadmap Epigenomics Consortium(Kundaje et al., 2015) profiled DNA accessibility, histone modification marks and DNA methylation in 111 epigenome datasets from diverse cell-types and tissues and integrated these newly generated datasets with 16 epigenomes from the ENCODE project(Dunham et al., 2012). One of the major insights from this study was the tissue-specific nature of chromatin signatures associated with transcriptional enhancers, such as H3K4me1. In sharp contrast, chromatin fingerprints of active promoters and other transcribed regions are constitutively active across tissues. These cross-tissue genome-wide chromatin maps of enhancers and promoters allowed grouping them based on co-occurrence of chromatin accessibility. This unearthed modules of enhancers and promoters and revealed that nearby genes had relevant roles in the underlying tissue biology.

Cis-regulatory enhancer elements control the expression of genes that could be located far away, even up to 1 mega-base distance in the linear genome, and hence makes it challenging assigning enhancers to their target genes by *in-silico* approaches such as previous chromatin co-accessibility methods. Distal transcriptional regulation is mediated by chromatin loop formation that facilitates the physical contact of enhancer and target promoter elements. Thereafter, disentangling the tissue-specific three-dimensional organization of the chromatin within the nucleus was pivotal to expand our understanding of the mechanistic underpinnings underlying enhancer looping and to gain insights into how spatio- and temporal regulation is orchestrated.

1.2.3. Relationship between three-dimensional (3D) architecture of the chromatin and genome function.

Our genome is tightly packed into nuclei following a hierarchical organization. At the lowest scale, DNA sequence is wrapped around nucleosomes forming chromatin fibers, which will be further condensed as chromosomes. Chromosomes occupy distinct territories as revealed through advancements in microscopy and three-dimensional chromatin assays. These studies primarily identified two chromosomal compartments, an "A compartment" predominantly consisting of euchromatin, and a "B compartment" that mainly comprises heterochromatin (Cremer and Cremer, 2001; Geyer et al., 2011; Lieberman-Aiden et al., 2009; Stadhouders et al., 2019). Chromosomal compartments further segregates into Topologically Associated Domains (TADs), which encompasses chromatin regulatory loops such as enhancer and promoter contacts.

In order to study the three-dimensional architecture of the chromatin, there are two main experimental approaches:

(i) Imaging-based techniques, particularly Fluorescence in situ hybridization of DNA (DNA-FISH).
(ii) Chromosome Conformation Capture (3C) techniques, in particular different variants of High throughput chromosome conformation capture (Hi-C).

FISH techniques use fluorescently tagged DNA sequences as probes that hybridize to the target sequences of the genome. This involves enhancing cell membrane permeability and denaturing the DNA such that the fluorescently labelled probes enter the nucleus and binds to the regions of interest. Then, microscopy techniques are used to visualize the probes inside the nucleus, and the distance between two or more fluorescently labelled probes is measured. Due to the limited resolution and the number of probes that can be simultaneously visualized, the application of FISH techniques is circumscribed to examine long-range contacts between large nuclear domains such as TADs or chromosomes. There has been progress in the resolution and throughput of microscopy-based techniques, such as the development of cryo-FISH or

Oligopaints, but they have not been widely adapted for genome-wide studies of chromatin contacts due still abovementioned limitations.

Our knowledge of chromatin 3D organization has dramatically increased with the fruitful advances in 3C experimental techniques. 3C methods are based on cross-linking chromatin of spatially adjacent genomic loci followed by proximity ligation. Cross-linked chromatin is then digested using restriction enzymes and DNA ends are ligated in conditions that favor intramolecular proximity. These ligated fragments are known as a 3C library. Finally, interactions between two regions of interest are quantified from the 3C library using PCR and appropriate primer sequences. Thus, 3C is primarily focused on the interaction frequency of two genomic regions. Nevertheless, as a 3C library still contains all the genome-wide ligation events, it can be further employed to interrogate genome-wide chromatin loops (Kempfer and Pombo, 2020). Further advances such as circular chromosome conformation capture or chromosome conformation capture-on-chip (collectively called as 4C) techniques examined contact frequencies of one region of interest with the rest of the genome ("one versus all") (Kempfer and Pombo, 2020) Only with the emergence of Hi-C techniques, first genome-wide maps of long-range contacts ("all versus all") were generated. Hi-C first involves enzyme restriction of cross-linked genome with formaldehyde. Next, DNA ends are repaired and marked with biotin before ligation. Ligation ends are purified and non-specifically sheared using sonication. Finally, biotinylated junctions are isolated followed by paired-end sequencing (Berkum et al., 2010). Different variants of Hi-C such as "capture" based Hi-C methods were subsequently developed to primarily pull-down specific ligation products and therefore, enriching the library for targeted genomic regions.

As major feature of gene transcription is 3D organization of the chromatin, progress in 3D techniques has been essential to gain new insights into gene regulation. Combining Hi-C techniques with chromatin immunoprecipitation has shed light on chromatin modifiers, components of the transcription machinery or structural proteins that mediate gene regulatory loops. Several techniques such as ChIA-PET, Hi-ChIP, PLAC-seq etc allowed specific enrichment of Hi-C libraries by chromatin immunoprecipitation (ChIP) before ligation (Kempfer and Pombo, 2020).

Another method to genome-wide map chromatin contacts is genome architecture mapping (GAM) (Beagrie et al., 2017). Unlike Hi-C technologies, GAM is a ligation-free method. In GAM, nuclei are randomly sectioned into ultra-thin slices using cryosectioning. Total DNA from every slice of nuclei is amplified and barcoded independently, pooled and then, sequenced. Sequencing data is used to mathematically model co-segregation of two genomic regions i.e., if two genomic regions are in close proximity, they tend to be found more frequently than expected in the same nuclear slice. While GAM allows studying chromosome interactions without disrupting nuclear structures unlike that of Hi-C, which involves extraction and ligation steps, it also requires around 400 nuclear slices, each sequenced to 1 million reads to achieve Hi-C resolution. Depending on the resolution requirements, one could generate a few thousands of slices of nuclei.

Enhancer-promoter interactions

In this hierarchical chromatin folding, enhancer-promoter loops represent the smallest unit. Enhancer-promoter interactions predominately occur within TADs, which are sub-megabase domains of chromosome folding. TADs contain a high frequency of intradomain interactions that impose spatial insulation and prevents inter-domain interactions that could lead to aberrant gene regulation through unexpected contacts (Symmons et al., 2014, 2016). TAD formation is facilitated by cohesin and CTCF proteins, which are highly enriched and co-bound at TAD boundaries (Rao et al, Nora et al). This process has been convincingly described through the 'loop extrusion' model (Fudenberg 2016, sanborn 2015) (Figure 1.9). According to this model, cohesin extrudes DNA loops through its ring-like structure until it reaches two convergent CTCF bound regions. Although the loop-extrusion model has not been experimentally verified, it accounts for several lines of support, such as the observation that cohesin depletion and the subsequent loss of extrusion activity showed the disintegration of the majority of TAD boundaries (Rao et al 2017, Schwarzer 2017). Other authors showed that cells deficient in cohesin-unloading factor WAPL account for extended chromatin domains (Haarhuis Cell 2017, Wutz EMBO 2017). While the loop extrusion model can explain how TADs are formed, it fails to describe how intra-TAD enhancer-promoter regulation occurs. In contrast to TADs, enhancer-promoter loops are not particularly enriched for CTCF or cohesin. In line with this notion, depletion of CTCF and cohesin,

or inversion of CTCF sites, did not lead to wide-spread changes in gene expression profiles, suggesting that CTCF and cohesin maintain the TAD structure but does not control enhancer-promoter interactions. Recently, YY1 protein has been suggested to be involved in gene regulatory loops (Weintraub et al., 2017) by the specific binding to enhancers and promoters as shown across cell-types in both human and mouse (Weintraub et al., 2017). In concordance, deletion of the YY1-binding motif at Raf1 and Etv4 gene promoters lead to decreased interactions with their respective enhancers.

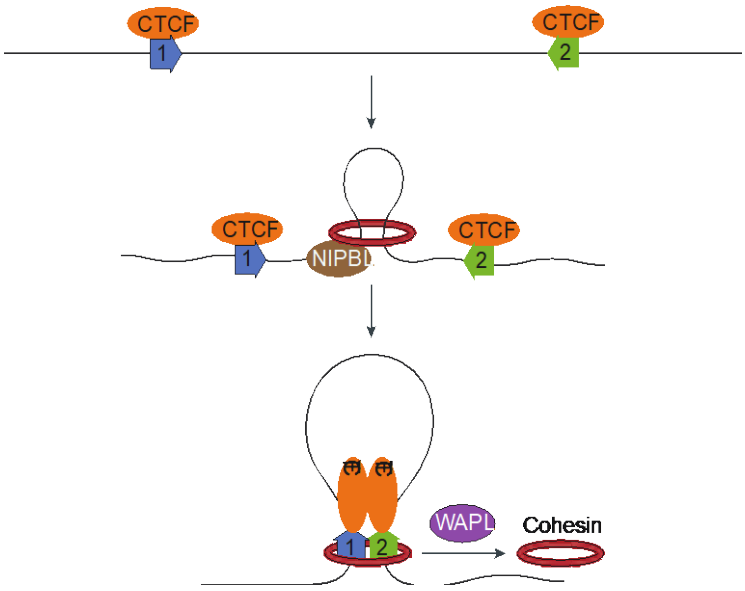

Figure 1. 9 Loop extrusion model.

According to the loop extrusion model, cohesin, after being loaded onto chromatin by NIPBL, progressively extrudes chromatin through its ring-shaped structure, resulting in a growing chromatin loop. Loop extrusion stops when cohesin encounters CCCTC binding factor (CTCF)-bound sites in a convergent orientation. Adapted from Schoenfelder, S., Fraser, P. (2019). Nature review genetics.

1.3. Non-genetic regulation of the genome function

Functional complexity of the human genome, reflected in cell-type diversification, stem from dynamic changes of gene expression profiles acquired during cellular development. Gene expression patterns are assumed to be determined by epigenetic regulation without involving changes in the DNA sequence. Two of the most important epigenetic mechanisms that modulate gene expression are DNA-methylation and histone modifications(Jaenisch and Bird, 2003), which shape the chromatin landscape during organism development. Epigenetic modifications can also occur during adult phases, as cells need to respond to external stimuli, environmental cues and to dynamically adapt to maintain their cellular identity and function. Upon stimulation or change in the cellular environment, alterations in the epigenetic landscape facilitate interactions between DNA and DNA-binding proteins, such as transcription factors, known to be key regulators of lineage-specific gene expression programs (Vaquerizas et al., 2009). DNA-TFs interactions with regulatory elements modulate cell-specific transcriptional programs and aid in the cellular response to external environmental perturbations. An illustrative example was the identification of latent enhancers that activate upon lipopolysaccharide (LPS) stimulation in terminally differentiated macrophages (Ostuni et al., 2013). Ostuni and colleagues observed that upon LPS stimulation, many of the annotated enhancers showed altered chromatin state as assayed by h3k4me1 and h3k27ac. Surprisingly, they found that some genomic regions, which were not marked as enhancers in the basal state, gained enhancer features and were also bound by lineage-determining transcription factors. Many of these latent enhancers did not go back to their native state but persisted and mediated a faster and stronger response upon re-stimulation.

Taken together, the identification of epigenetic alterations at enhancers and promoters, and by examining gene expression patterns after exposing cells to different environmental conditions, allow us to determine transcription factors whose expression and activity are key for cell homeostasis and function. This understanding has important implications to modulate cellular function for therapeutic purposes.

1.4. Genetic regulation of genome function

A major summit in modern biological research has been the completion of the first draft of the human genome sequence by the Human Genome Project (Lander et al., 2001; Venter et al., 2001). Since then, we witnessed a breathtaking process in genome science with the parallel technological development of genotyping arrays and Next Generation Sequencing (NGS), and the advent of large-scale data sharing initiatives that deepen democratization of scientific research. The emergence of these novel technologies fostered the rapid accumulation of sequence data from thousands of human subjects that has been used to catalogue genetic differences between individuals. A typical individual genome contains approximately 4-5 million single nucleotide variants (SNV) with respect to the reference genome(Auton et al., 2015). These SNVs can be categorized into common, rare and ultra-rare depending on the observed allele frequency of the minor allele (MAF) in the population. Variants with a MAF > 5% are considered common variants while variants with MAF < 1% are considered rare variants, although lower bound threshold are no longer fixed and are conditioned on the genetic resolution attained. Given that the human genome sequence encodes all the instructions for an organism development and function inevitably, genetic variation can cause phenotypic differences and underlie human diseases. To understand how genetic variation causes disease, first, we need to disentangle the relationship between individual genetic variation and disease conditions. With the establishment of genotyping arrays and NGS technologies, it is now possible to identify genetic variation across a large number of individuals, and that can be linked to phenotypic variation across individuals in a population.

1.4.1. Genome Wide Association Studies (GWAS) - Unravelling the genome-phenotype relationship underlying complex diseases

The etiology of complex traits and diseases, such as T2D, draws upon the aggregated effect of a large number of common genetic variants with small to modest effects, and a large involvement of environmental factors. Early efforts to study the genetic basis of complex traits and diseases recycled linkage analysis and candidate gene association studies, which were prolific in mapping the genetic causes of monogenic rare diseases

(Claussnitzer 2020). However, as highly penetrant rare genetic markers that co-segregate in large families do not dominate the genetic architecture of most common complex traits, and prior knowledge of biologically relevant genes was scarce, these approaches were largely unrewarding (Claussnitzer et al., 2020). Genome-wide association studies (GWAS) has emerged as the most successful approach to link genetic variation with the absence or presence of a complex disease or variance in a trait across individuals in a population. This statistical approach interrogates from one to several millions of genetic variants and looks for statistically significant differences in allele frequencies in large cohorts of patients and unaffected individuals. GWAS has been applied to various diseases and have been fruitful in giving clues about the genetic architecture of common diseases and traits(MacArthur et al., 2017; Visscher et al., 2017).

The 'missing heritability' conundrum

Despite the tremendous success of this approach, the overwhelmingly majority of GWAS risk variants only explain a modest proportion of the estimated heritability of a given disease or phenotype (Manolio, Nature 2009). This is known as the 'missing heritability' conundrum. One probable explanation of this gap of knowledge is that a myriad of common genetic variants that do not reach genome-wide significance due to their weak effects on disease risk are not readily detected in sample sizes used in current studies. Genotype imputation, a statistical approach that predicts the genotypes of variants that have not been directly typed in SNP arrays, boosted genetic resolution and increased statistical power by facilitating meta-analysis of GWAS summary statistics data from independent studies (Marchini and Howie, 2010). Nevertheless, the cumulative effects of all genome-wide common variants, even those that do not reach GWAS significance, was still lower than the heritability estimates from pedigree studies. Thus, several authors suggested that this missing fraction of complex trait heritability might rest upon low-frequency and rare variants that are not included in commercial arrays and thus, have not been extensively examined (Manolio et al., 2009; Visscher et al., 2012). Indeed, the recent availability of large whole-genome sequencing datasets allowed for the first time thoroughly assessing the role of rare and low-frequency variants. This has begun to fill the gap of the missing heritability for body mass index (BMI) and height (Wainschtein et al., 2019). Finally, it should also be noted

that genetic susceptibility of most common diseases is strongly influenced by environmental risk factors and studying the effect of genetic variants in the context of gene-environment interactions is likely to improve heritability estimates.

1.4.2. Translating GWAS discoveries into functional insights

The growing inventory of GWAS associations identified for hundreds of traits and diseases did not account so far for a comparable increase in novel insights into trait and disease biology. Two main reasons limited the biological interpretation of GWAS discoveries. First, genetic variants showing stronger associations are not usually the true causal variants. This limitation arises from the latent structure of genetic variation in the genome. Genetic variants can be segmented into haplotypes blocks of markers that frequently co-segregate together(Wall and Pritchard, 2003) . The degree of allelic co-dependency between two distinct loci is measured by Linkage Disequilibrium (LD). Within blocks of low frequency of recombination, a true disease causal variant will tend to be in high LD with adjacent markers. Thus, it will be inherited together, with genetic variants in relatively close proximity (Schaid et al., 2018). Following this rationale, any of the correlated markers within the same LD block could reach stronger associations with disease risk besides the true causal variant (Altshuler et al., 2008; Schaid et al., 2018). Of note, although this high correlation between genetic variants within a haplotype block eases GWAS discovery, it makes it thus challenging identifying the underlying causal variant and subsequently, gaining insights into disease pathophysiology. Genetic fine-mapping and functional genomics analysis in disease-relevant tissues emerged to discriminate likely causal variants from those that are merely correlated and do not account for a functional link with disease susceptibility.

Another major challenge in GWAS is that more than 80% of the disease risk variants fall in non-coding regions of the genome(Hindorff et al., 2009). Early studies showed that more than 60% of non-coding disease risk variants reside in cis-regulatory regions identified through chromatin accessibility analysis using DNase I hypersensitive sites (DHSs) (Maurano et al., 2012). Mounting evidence indicates that transcriptional enhancers, in particular clusters of enhancers, are enriched for disease risk variants(Kundaje et al., 2015; Pasquali et al., 2014) that might impact the binding

efficiency of TFs, and thus altering enhancer function. As we discussed in section 1.2, enhancer regulatory elements control the expression of genes that could be far away in the lineal genome. This hampers the assignment of non-coding GWAS variants to their respective downstream effector genes whose abnormal function impacts disease pathogenesis. Thus, progress in the identification of disease causal variants and effector transcripts could offer us novel insights into disease pathophysiology.

Identification of candidate causal variants

Genetic fine mapping nominates most-likely causal variant (or variants) of a disease association by estimating the posterior probability (PP) of a variant to be causal using Bayes factors. Variants are sorted in descending order of their causal posterior probabilities. Then, a 'credible set' of most likely causal variants is built by selecting those whose cumulative sum of posterior probabilities reach a certain threshold (cumulative PP ~ 95-99%) (Benner et al., 2016; Chen et al., 2015; Hormozdiari et al., 2014; Hutchinson et al., 2020; Lee et al., 2018; Wang et al., 2020) . Nevertheless, there are several factors that hinder the performance of fine-mapping approaches, such as the genetic resolution attained, the study size, the magnitude of the effect sizes or the number of causal variants (Schaid et al., 2018). As previously discussed, high local LD makes it challenging fine mapping causal variants. Trans-ethnic studies can exploit differences in LD structure between major ancestry groups to reach higher fine mapping resolution to uncover candidate causal variants. For instance, GWAS in individuals of African descent will benefit from the lower LD extension that will constrain the genomic space to search for the true causal variant (Cooper et al., 2008; Zaitlen et al., 2010)

In contrast with protein-coding variants, our ability to infer the magnitude of the deleterious consequences of regulatory candidate causal variants is poor due to our limited understanding of the enhancer grammar. *In-silico* pathogenic scores that assess the putative effects of a given genetic variant on cellular phenotypes can aid in refining genetically fine mapped datasets. The integration of these *in-silico* functional scores with fine mapped variants might allow selecting *bona fide* causal variants with clinically relevant implications for disease pathogenesis. (Figure 1.10). *In-silico* scores are primarily based on either evolutionary conservation, or functional and sequence attributes of the surrounding genomic context. Several approaches that use one or a

combination of these principles now allow researchers to estimate fitness consequences of fine mapped regulatory variants. Commonly adopted tools are: CADD (Kircher et al., 2014), GWAVA (Ritchie et al., 2014), DeepSea (Zhou and Troyanskaya, 2015), Eigen (Ionita-Laza et al., 2016) or LINSIGHT (Huang et al., 2017), among others.

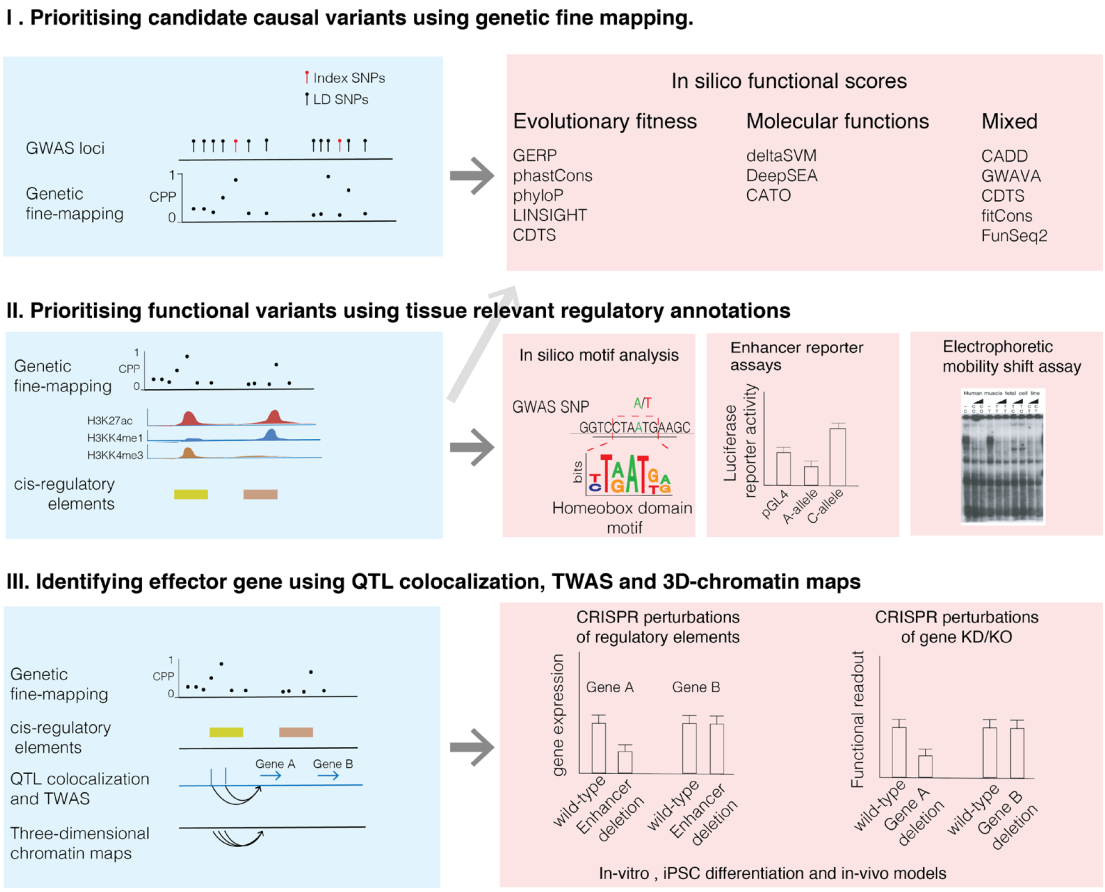

Figure 1. 10 A schematic of GWAS variant to gene workflow.
I) Genetic fine mapping approaches can be used to prioritize most likely candidate causal variants (credible set variants). The candidate causal variants can be further assessed for their likely function and pathogenicity based on a variety of in-silico functional scores. II) The credible set variants can further be narrowed down using epigenome data sets from disease relevant tissues. This can be validated by in-silico tools that can assess the disruption of sequence motifs by genetic variants and experimentally validated using enhancer reporter assays and electrophoretic mobility shift assays (EMSA). III) QTL colocalization and TWAS methods can be used to prioritize the candidate effector transcript in each GWAS loci along with the three-dimensional chromatin maps from disease relevant tissues. The effector candidate transcripts can be further validated by CRISPR perturbation of the variants and assess the impact on candidate gene expression. The candidate gene itself can be assessed for its functional role in the relevant tissues by knock-down and knock-out experiments in in-vitro and in-vivo.

Cis-regulatory elements like enhancers and promoters have distinct DNA sequences that promote lineage-specific transcription factor binding and regulate tissue-specific transcriptional programs(Long et al., 2016). The excess of genetic variants associated with complex traits and diseases in cis-regulatory elements(Gaulton et al., 2010; Kundaje et al., 2015; Thurman et al., 2012) indicates that their underlying molecular effects could involve altering transcription factor occupancy and thus, triggering defects in gene regulation. Hence, overlaying disease-risk variants with epigenome-based regulatory annotations from disease-relevant tissues eased causal variant prioritization (Figure 1.10). This approach has been successfully applied to understand the functional role of common regulatory genetic variants in several diseases (Cowper-Sal·lari et al., 2012; Farh et al., 2015; Gupta et al., 2017; Mumbach et al., 2017; Pasquali et al., 2014). Besides the direct overlay, regulome annotations and aforementioned *in silico* scores can also be used to directly improve fine mapping itself (Chen et al., 2016b; Kichaev et al., 2014; Pickrell, 2014) .

Identification of candidate effector genes

The identification of target effector genes of risk variants is an additional major handicap in the conversion of GWAS discoveries into mechanistic insights, as enhancers usually regulate distal genes. An illustrative example of this challenge is obesity-associated risk variants in the *FTO* locus. Data from 4C-Seq studies revealed that obesity-associated risk variants in the intron of the *FTO* gene are in spatially proximity of *IRX3* gene. Accordingly, *IRX3* gene expression in the brain(Smemo et al., 2014). Thereafter, along with the integration of regulome annotations, leveraging three-dimensional chromatin maps in disease relevant tissues might allow assigning candidate effector transcripts to GWAS risk variants (Figure 1.10). This approach has been widely used to predict target genes for several complex diseases (Arvanitis et al., 2020; Cowper-Sal·lari et al., 2012; Fachal et al., 2020; Fang et al., 2019; Greenwald et al., 2019; Javierre et al., 2016; Jung et al., 2019; Miguel-Escalada et al., 2019; Nott et al., 2019; Song et al., 2019). Of note, Javierre and colleagues introduced a Hi-C variant that increases the resolution to detect interactions between genes and their distal regulatory elements, known as promoter capture Hi-C data, in 17 primary blood cell-types. These novel chromatin interaction maps linked 2,604 genes to GWAS risk variants of 31 diseases and blood cell traits(Javierre et al., 2016). More recently, Jung

and colleagues generated a compendium of capture Hi-C data from 27 different human tissues and cell-types and assigned more than 50% of putative disease/trait-associated variants to at least one target gene (Jung et al., 2019). Although functional genomics based on 3C data has pushed forward our understanding of the molecular mechanisms underlying disease risk variants, it should be noted that, so far, only a handful of examples have been experimentally validated.

Mapping disease risk variants that affect gene expression using QTLs

Another strategy to gain insights into the role of non-coding genetic effects on genome function is assessing the relationship between genetic variants and molecular phenotypes from disease-relevant tissues. One of the most widely adopted approaches is expression quantitative trait loci (eQTL) (Figure 1.10) (Montgomery and Dermitzakis, 2011; Myers et al., 2007). eQTL studies test the association between allele frequencies of common genetic variants and gene expression measured in a given tissue across several individuals. Interestingly, these eQTL maps can be capitalized to prioritize non-coding variants likely affecting gene expression. Early studies indicated that eQTLs are tissue-specific by showing that 79.5% of the eQTLs mapped in three cell-types were unique, and only 8.5% are shared across cell-types (Dimas et al., 2009). These pioneer studies suggested that is essential generating eQTL maps in disease-relevant tissues due to their tissue-specific nature. Genotype-Tissue Expression (GTEx)(Consortium et al., 2015b; Lonsdale et al., 2013) aimed at identifying genetic variants associated with gene expression across several tissues including immune cell-types. So far, GTEx project identified thousands of genes under genetic control across 883 samples and 49 tissues in scaling phases, and consistently showed the eQTLs from disease-relevant tissues tend to be enriched among disease-associated risk variants. Taken together, eQTLs provides an opportunity to identify putative causal disease risk variants and their effector transcripts.

Other complementary approaches to eQTLs involve measuring molecular phenotypes that are informative of enhancer function, such as histone-QTLs (hQTLs), chromatin-accessibility QTLs (caQTLs) or methylation-QTLs (meQTLs) among others. The BLUEPRINT epigenome project (Chen et al., 2016a) characterized different molecular QTLs in matched samples, providing a unique opportunity to understand shared genetic effects across intermingled molecular traits. The authors assayed gene expression profiles, along with two histone marks, H3K27Ac and

H3K4me1 and genome-wide methylation status using Illumina 450k arrays. The authors observed that 43% of eQTLs were also hQTLs for h3k4me1 or h3k27ac. These shared genetic effects across molecular QTLs along allowed them to connect putative regulatory elements with their target genes.

Along with the eQTLs and complimentary hQTLs, genetic regulation of transcriptome variation also is driven from genetic effects on alternative splicing (sQTLs). Indeed, associated risk variants for several complex diseases have been shown to regulate alternative splicing in relevant tissues (Li et al., 2016, 2019; Raj et al., 2018a; Zhang et al., 2020) sQTLs provided additional candidate effector transcripts that otherwise could be missed in eQTL and hQTL analysis.

In order to take advantage of eQTL maps to gain insights into genes mediating disease predisposition, diverse approaches have been developed to consistently integrate GWAS and eQTL data (Figure 1.10).

Colocalization methods are one family of methods that link molecular QTLs to disease risk variants by seeking for a shared genetic marker that is simultaneously associated with the disease and the molecular phenotype (e.g., gene expression or alternative splicing). Colocalization approaches are thus able to nominate candidate genes and the tissue of action where the disease phenotype manifest. This approach has successfully provided novel insights into GWAS loci for celiac disease with the simple overlap of GWAS risk variants with whole blood eQTLs, which nominated candidate genes in 20 out of 38 loci (Dubois et al., 2010)One of the limitations of such approach is that they do not account for local LD structure, and thus, does not rule out spurious overlaps.

Nica et al. (Nica et al., 2010) proposed a method called Regulatory Trait Concordance (RTC) which assesses the residual effects of eQTLs when conditioned on GWAS risk variants. In contrast with early approaches, this method accounts for local LD structure but does not perform any formal test of the odds of colocalization against the null hypothesis.

Hypotheses tests were included in Sherlock(He et al., 2013), that matches genetic signatures from eQTLs with GWAS. The underlying patterns of genetic associations of a gene constitute these so-called “genetic signatures”. If a gene is mediating disease risk, genetic signatures of the gene should overlap, at least partially, with GWAS risk variants. Of note, this approach allows comparing genetic effects from two traits not only locally but genome-wide, facilitating the integration of *trans*-QTL

effects with GWAS risk variants. Sherlock is based on a Bayesian framework that does not depend on conservative p-value cutoffs and thus, can benefit from both strong and moderately associated variants while accounting for LD patterns (He et al., 2013). One of the limitations of Sherlock is that only uses SNPs that are associated with either gene expression or GWAS and does not account for variants that act against colocalization in a particular locus.

COLOC is the current state-of-the-art method that formally tests the null hypothesis against a shared genetic signal between two independent traits. It uses a Bayesian framework to calculate the posterior probability of a variant to be causal for both GWAS and eQTL phenotypes. COLOC tests five different tests hypothesis.

```
H0: No association with either trait N
H1: Association with trait 1, not with trait 2 N
H2: Association with trait 2, not with trait 1 N
H3: Association with trait 1 and trait 2, two independent SNPs N
H4: Association with trait 1 and trait 2, one shared SNP
```

For each of the above mutually exclusive configurations, a probability is calculated based on pre-selected informative probability priors, thus resulting in five posterior probabilities. A large posterior probability for H4 scenario indicates a shared causal genetic effect on both traits. It should be noted that all the above methods assume a single causal genetic variant in a locus. Alternative methods based on the original COLOC algorithm emerged, such as *gwas-pw* that employs empirical Bayes to estimate per-hypothesis priors.

Previous methods assume no more than a single causal variant at each locus, and this does not reconciliate with recent observations of widespread allelic heterogeneity (Hormozdiari et al., 2017; Jansen et al., 2017). To further address this inconsistency, methods like eCAVIAR that accommodate multiple causal variants were developed. For a given lead GWAS variant, a window around that SNP is selected to include M variants (e.g., 50). Then, for all the variants within the locus, eQTL marginal statistics are considered and eCAVIAR framework is applied. eCAVIAR also provides a posterior probability for the hypothesis of shared causal variants termed colocalization posterior probabilities (CLPP) i.e., a posterior probability that the variant is causal for both the traits. There are several advantages for using CLPP; CLPPs not only inform

us about the strength of GWAS and eQTL colocalizatio. If eQTLs from several tissues and GWAS from multiple traits are available, CLPPs also give us clues about the tissues of action for GWAS loci from several traits (Hormozdiari et al., 2016).

TWAS

Transcriptome-wide association studies (TWAS) are another distinct family of methods to prioritize candidate effector transcripts of GWAS risk variants. TWAS tests the association of gene expression levels with disease risk by estimating the heritable component of gene expression from large expression panels, that will be harnessed to 'impute' gene expression in individuals from GWAS datasets (Figure 1.11). Briefly, per-gene cis-heritable expression is estimated using predictive models that learn over panels where both gene expression and genotypes are available, such as GTEx data. Local SNPs at a certain distance from a given gene are considered (e.g., SNPs with in 1 megabase from gene TSS) model gene expression variation based on allelic counts. These predictive models can then be used to impute gene expression in GWAS individuals where such data has not been measured. Predicted gene expression is then tested for association with trait variation or disease susceptibility. Thereafter, this approach is able to prioritize candidate genes that might mediate disease risk (Wainberg et al., 2019)

Gamazon et al proposed PrediXcan method (Consortium et al., 2015a) that uses LASSO and elastic net models to predict the genetically regulated gene expression (GREx) component of whole-blood data(Battle et al., 2014) ('training set') to predict gene expression levels (treated as quantitative traits) in GEUVADIS LCLs(Consortium et al., 2013) and nine GTEx pilot tissues(Lonsdale et al., 2013) ('test sets'). These per-gene predictive models were then used to impute the gene expression in WTCCC study(Burton et al., 2007) and identified 41 associations for 5 diseases. One of the limitations of PrediXcan is that it requires individual-level genotype data from GWAS individuals, that are rarely available due to ethical policies. Thus, Gusev et al. adapted this TWAS framework to summary-statistics data in his method also known as Functional Summary-based Imputation (FUSION) (Gusev et al., 2016). FUSION uses five predictive models, Best Linear Unbiased Predictor (BLUP) and Bayesian Linear Mixed Model (BSLMM) on top of LASSO, elastic net and top eQTL SNPs. The inclusion of diverse predictive models allows dynamically selecting for each gene that

of that provides the highest accuracy, and thus, outperforms single eQTL based predictions.

Another set of statistical methods that have been developed to prioritize candidate effector genes mediating disease risk rests on Mendelian Randomization (MR) approaches. The rationale underlying MR is that if a genetic variant is associated with gene expression variation, and if that gene expression variation mediates disease risk, genotype differences for that genetic variant should result in cognate phenotypic variation across individuals of a population. As in classic MR, genetic variants are considered as an instrumental variable to test for the causative effect of an exposure (e.g., gene expression) on an outcome (disease phenotype). Zhu et. al implemented this rationale in their Summary data-based Mendelian Randomization (SMR) method, which requires summary statistics from GWAS and QTL studies(Zhu et al., 2016). One of the limitations of MR approaches is that linkage is misinterpreted as pleiotropy (Hemani et al., 2018). Thus, Zhu et al implemented the heterogeneity in dependent instruments (HEIDI) test that can distinguish pleiotropy from linkage using multiple SNP associations within a given locus. Recently, several methods emerged hinging on the same principle(Richardson et al., 2020; Schmidt et al., 2020), and also considering multiple-instrument and multiple-exposure MR models (Porcu et al., 2019). It must be noted that, even though these methods have been primarily developed with gene expression reference panels, they can be used with any molecular trait measured across large reference panels and with genotype data available.

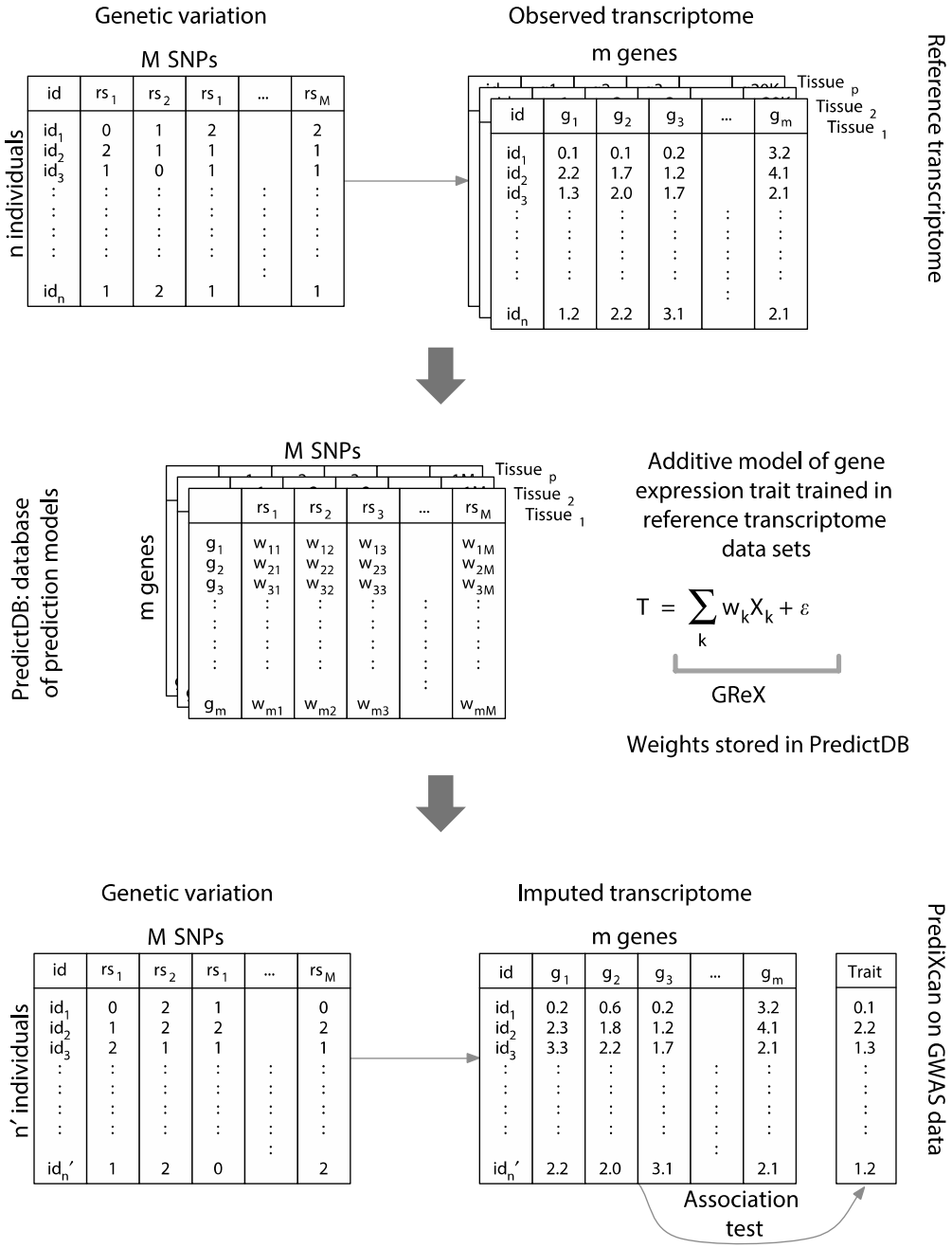

Figure 1. 11 A schematic representing TWAS approaches.
First a reference transcriptome data set with genotypes is used to compute the heritability of gene expression. The weights for each SNP toward heritability of gene expression are used to impute the gene expression in GWAS individuals. The imputed gene expression levels are then used to perform association test with the trait.

One of the key advantages of performing TWAS or MR based association tests is that multiple SNPs within the same *cis*-window are used to estimate the gene expression

heritable component by testing different models for prediction, which greatly increases the power over single-SNP eQTL associations. The second major advantage is that TWAS or MR can uncover novel disease risk loci that might have been missed by GWAS studies due to still unpowered study sizes. However, all novel biology proposed by TWAS and MR should be carefully considered as these methods cannot distinguish horizontal pleiotropy from vertical pleiotropy(Hemani et al., 2018; Wainberg et al., 2019). Thus, close evaluation of prioritized effector transcripts using several lines of evidence, such as drawing additional support from colocalization approaches, is recommended.

1.5. From T2D risk genetic variants to novel mechanistic insights.

T2D is a paradigm of complex multifactorial diseases and has proved to be a fertile ground to advance the field of complex disease genetics. Our catalogue of T2D risk associated variants has expanded to more than 400 independent genetic signals (Mahajan et al., 2018a; Newman et al., 1987). This astounding summit has been accomplished thanks to the establishment of pioneer trait-specific consortia, such as the DIAbetes Genetics Replication And Meta-analysis (DIAGRAM), that allowed extending samples sizes from few to hundreds of thousands of individuals. The success of this collaborative strategy has been reinforced by the emergence of national biobanks, such as the UK Biobank (Bycroft et al., 2018) or the Million Veteran Program (Gaziano et al., 2016) . The generalization of T2D GWAS across underrepresented ancestries has been of the outmost importance to foster genetic discovery but also to enhance the resolution of fine mapping approaches(Consortium et al., 2014a). This large inventory of T2D risk genetic signals has already provided fruitful insights into disease pathophysiology. The central role of human pancreatic islets to T2D pathogenesis (Ashcroft and Rorsman, 2012) has been revealed with the marked enrichment of T2D risk variants in the epigenome landscape, particularly in enhancer domains, of human pancreatic islets (Gaulton et al., 2010; Parker et al., 2013; Pasquali et al., 2014). The by far non-coding nature of T2D disease risk variants has challenged the identification of effector transcripts through which these risk variants impact on T2D pathophysiology. Advances in “Hi-C” technologies aided in the

identification of effector transcripts by detecting genes that are in close proximity to the T2D risk loci in the 3D space. The integration of promoter-capture Hi-C data with refined pancreatic islet enhancer annotations generated during this PhD thesis allowed assigning one or more candidate effector genes to 53 loci, and 75% of them accounted for distal and unexpected candidate gene(Miguel-Escalada et al., 2019). Other studies have successfully integrated both Hi-C and eQTL information in human pancreatic islets with GWAS data to identify effector transcripts(Greenwald et al., 2019).

Candidate gene assignments for T2D risk loci has also been successfully attained using Quantitative Expression Loci (eQTL) panels. Scaling efforts have been dedicated to expression and chromatin accessibility QTL studies in human pancreatic islets(Bunt et al., 2015; Fadista et al., 2014; Khetan et al., 2018; Varshney et al., 2017a), that have culminated with recent eQTL maps from the InsPIRE consortium in 420 human pancreatic islet samples from cadaveric organ donors(Viñuela et al., 2019). This large eQTL panel identified effector transcripts in 23 T2D GWAS loci, suggesting that a large fraction of the biology underlying T2D predisposition still needs to be discovered.

So far, three-dimensional chromatin studies and QTL approaches capitalized transcriptional regulation to provide effector transcripts associated with T2D risk. Importantly, it is now well established that genetic variants that alter alternative pre-mRNA splicing contribute significantly to disease risk in several human genetic diseases(Li et al., 2016). Thereafter, complementary studies that assess the effect of genetic variants on pre-mRNA splicing in human pancreatic islets might hasten this translation of GWAS discoveries into molecular insights that could be harnessed for therapeutic purposes.

2. Hypothesis and objectives

This thesis aims to provide a comprehensive understanding of both transcriptional and alternative splicing regulation in human pancreatic islets and to elucidate the relationship with T2D aetiology.

This main aim has been built upon the following hypotheses:

Hypothesis 1. Our insights into the human pancreatic islet transcriptional regulatory elements that control islet function and cell identity are still limited although their role is central to T2D pathogenesis.

Hypothesis 2. The tissue-specific component of the human islet transcriptome is underrepresented in reference annotations.

Hypothesis 3. The impact of common genetic variants on alternative pre-mRNA splicing in human islets has not been determined and could shed light on the molecular mechanisms concurring in T2D pathophysiology.

These hypotheses fuelled the following objectives, respectively:

Objective 1. To generate (1.a) a genome-wide atlas of islet regulatory elements by the integration of high-resolution maps of open-chromatin regions, histone modification marks and transcriptional regulators in human pancreatic islets, and to (1.b) unearth the components of the islet regulatory network that maintain beta cell function by using a glucose perturbation model in quasi-physiological conditions.

Objective 2. To refine the human islet transcriptome and to identify novel coding sequences encoding islet-specific peptides, which can be potentially targeted for therapeutic purposes.

Objective 3. To generate high-resolution maps of genetic effects on alternative pre-mRNA splicing and gene expression in human islets and to examine their distinct potential to ease on the identification of disease causal variants underlying T2D pathogenesis.

3. Methods

ATAC-Seq data analysis

13 human pancreatic islet donor samples were sequenced to a median depth of 30 million reads. Low quality bases and adaptor trimming was performed using TrimGalore v.0.4.1 (`--quality 15 --nextera`). Trimmed reads were aligned to hg19 using bowtie2 v.2.1.0 (`--no-unal`) allowing no mismatches. Uniquely mapped reads (Mapping quality, MAPQ ≥ 30) were retained using SAMtools v.1.2 and duplicate reads were removed (picard v.2.6.0). Reads falling in blacklisted regions and reads from mitochondrial genome were also discarded. Peaks were called using MACS2 (`--shift 100 --extsize=200 --keep-dup all –nomodel --p 0.01`) in 13 individual samples. We then pooled the bam files from these 13 samples and peaks were called on the pooled bam file using MACS2 (`--shift 100 --extsize=200 --keep-dup all –nomodel --q 0.05`). We then defined consistent peaks present in at least three samples as well as in the pooled set. Consistent ATAC peaks that showed multiple subpeaks in >3 islet samples were manually split, leading to $n = 241{,}481$ ATAC peaks. A final set of accessible chromatin regions ($n = 249{,}582$) was defined by adding regions lacking ATAC-seq peaks that showed either Mediator or CTCF binding ($n = 1{,}319$, $n = 9{,}596$ respectively) or were bound by at least two islet transcription factors ($n = 1{,}514$).

ChIP-Seq data alignment

The alignment step for histone modifications (H3K27ac, H3K4me1 and H3K4me3), Mediator, CTCF and SMCA1 (part of the cohesion complex) was similar as follows. Adaptor trimming was performed with cutadapt v.1.9.1 (`-m 20`). Trimmed reads were aligned to hg19 using bowtie2 v.2.1.0 (`--no-unal`) allowing no mismatches. Uniquely mapped reads (Mapping quality, MAPQ ≥ 30) were retained using SAMtools v.1.2 and duplicate reads were removed using picard v.2.6.0. Reads falling in blacklisted regions were also removed.

Consistent peaks – histone modifications

For H3K4me3 and H3K4me1, broad enriched regions were called with MACS2 (`--g hs --extsize=300 --keep-dup all --nomodel –broad`) and H3K27ac

narrow regions were called using MACS2 (`--g hs --extsize=300 --keep-dup all --nomodel`). To obtain a robust set of ChIP-seq peaks, we called peaks in individual human islet samples with relaxed stringency ($P < 0.01$), and in pooled samples using a stringent threshold (false discovery rate (FDR) $q < 0.01$). We then identified peaks present in at least three individual samples, or at least two samples if only three replicates were processed, as well as in the pooled set.

Consistent peaks – Mediator, SMCA1 and CTCF.

Narrow peaks were called using MACS2 (`--g hs --extsize=300 --keep-dup all`). To obtain a robust set of ChIP-seq peaks, we called peaks in individual human islet samples with relaxed stringency ($P < 0.01$), and in pooled samples using a stringent threshold (false discovery rate (FDR) $q < 0.05$). We then identified peaks present in at least three individual samples, or at least two samples if only three replicates were processed, as well as in the pooled set

Classification of human islet-accessible chromatin

We classified 249,582 consistent islet open chromatin regions using *k*-medians clustering of ChIP-seq signal distributions of H3K27ac, H3K4me1, H3K4me3, Mediator, cohesin and CTCF, using islet samples with the greatest signal to noise for these marks. Briefly, $-\log_{10}$ (*P* value) signal was calculated for each mark using 100 base pair bins across a 6-kb window centered on consistent open chromatin regions. MACS2 estimates a p-value at each base-pair by testing the ChIP signal against the corresponding local lambda derived from the control sample (input) with a Poisson model. Full details of this algorithm are available at https://github.com/macs3-project/MACS/wiki/Advanced%3A-Call-peaks-using-MACS2-subcommands#Step_6_Compare_ChIP_and_local_lambda_to_get_the_scores_in_pvalue_or_qvalue**.** *K*-median clustering (flexClust) was used to classify open chromatin regions into 14 clusters, which were manually merged into eight clusters based on the chromatin mark enrichment patterns. Each open chromatin class was ranked by CTCF binding to highlight a subset of CTCF-bound enhancers. Post-hoc analysis showed that human islet transcription start-sites defined by cap-analysis of gene expression (CAGE) were markedly enriched in regions classified as active promoters and, to a lesser extent, in class I enhancers.

Enhancer-promoter H3K27 acetylation correlations

We considered that H3K27 acetylation signal in enhancer-promoter target pairs should tend to show higher correlation values across tissues and human islet samples than unrelated pairs. We empirically combined data from multiple tissues and human islet samples to generate a single Spearman's Rho value for every possible enhancer-promoter pair in each islet TAD-like domain and found improved discrimination in functionally characterized enhancer-gene pairs. As control sets, for every enhancer with an assigned gene promoter, we randomly selected another gene promoter in the same TAD.

Correlations were calculated with H3K27ac ChIP-Seq from 14 human islet samples, including 7 samples exposed to 11mM glucose and 4mM glucose, and 51 tissues from the Epigenome Roadmap Consortium (aligned reads from ChIP-Seq samples and inputs downloaded from egg2.wustl.edu/roadmap/data/byFileType/alignments/consolidated/ and converted to BAM format using *bamToBed* from BEDTools). We selected datasets containing > 15 million usable reads, and uniformly subsampled to a maximum of 30 million usable reads. Active islet enhancers were uniformly extended to +/-750 bp from the centre of the peak. Active islet promoters were used without further modifications.

Reads mapping to islet active enhancers and promoters were quantified in all tissues and inputs using *featureCounts* from Rsubread R package and were sequence-depth normalized. Input signal was subtracted from the sample signal. Spearman's Rho values (scipy.stats.spearmanr) were calculated between all pairs of active enhancers and active promoters within TAD-like domains of human pancreatic islets function.

Glucose regulation of enhancers and their target genes.

We analyzed H3K27ac ChIP-seq and RNA-seq datasets from islets from 7 human donors cultured for three days in 11- or 4-mM glucose.

For RNA-seq of islets exposed to different glucose concentrations, 100 bp paired-end sequencing reads were aligned to masked hg19 genome using STAR aligner v2.3.0 (options: --*outFilterMultimapNmax* 1 --*outFilterMismatchNmax* 10). Gene level counts were obtained using FeatureCounts v1.5.0. (-s 2 -p). After removing genes that did not have at least 5 raw reads mapped in at least 3 replicates, a paired DESeq2 (v1.10.1) analysis was carried out to identify differentially regulated genes.

To assess glucose-regulation of enhancers, we defined H3K27ac-enriched regions, rather than using annotated enhancers that typically contain nucleosome-depleted sub-regions lacking H3K27ac enrichment. We thus defined consistent H3K27ac- enriched regions for each human donor and each glucose condition treatment using MACS2 (`--g hs --extsize=300 --keep-dup all –nomodel`). A total of 90,814 narrow H3K27ac-enriched regions were interrogated. The number of H3K27ac reads mapping to each peak was calculated using FeatureCounts (-- *ignoreDup* -O -- *minOverlap* 10). Then, paired DESeq2 (v1.10.1) analysis was used to assess differential signal strength. Peaks showing differential H3K27ac ChIP-seq signal at adjusted $P \leq 0.05$ were then mapped to annotated regulatory elements. Motif analysis was performed using homer(Heinz et al., 2010)

To calculate the enrichment of interactions between glucose-induced enhancers and glucose-induced genes, we considered all possible pairs of glucose-induced enhancers and genes within an islet TAD-like domain. For each enhancer-promoter pair we created a control pair with a distance-matched gene. We excluded experimental pairs when we could not find a distance-matched control. Then, we calculated a Fisher's exact test p-value to assess if glucose-induced enhancer and glucose-induced genes were enriched in high-confidence or imputed assignments. As an additional control, we assessed if glucose-induced enhancers preferentially contact glucose-repressed genes. We further examined whether gene promoters assigned to glucose-induced enhancers also show a significant glucose-dependent increase in H3K27ac levels. As a control, for every glucose-induced enhancer we chose a gene promoter that had the closest distance to the enhancer as the assigned gene promoter. The median distance for interacting gene promoters and control promoters was 200 kb (IQR 102-356 kb) and 167 kb (IQR 99-351 kb), respectively. The median distance for imputed gene promoters and control promoters was 114 kb (IQR 58-301 kb) and 134 kb (IQR 71-324 kb), respectively.

PacBio data analysis

Two human pancreatic islet donor samples were sequenced on PacBio IsoSeq platform. From each human islet samples, 4 libraries of different sizes (0.5-2kb, 1.5-3kb, 2.5-6kb, 4.5-10kb) were sequenced. This yielded around 400,000 full-length non-chimeric

reads in total. The full-length reads were aligned to hg19 reference genome using STARlong

```
STARlong --runThreadN 16 --outFilterMultimapNmax 1 --genomeDir Indexes --readFilesIn ${sample} --outSAMstrandField intronMotif --genomeLoad NoSharedMemory --runMode alignReads --outSAMattributes NH HI NM MD --outFilterMultimapScoreRange 1 --outFilterMismatchNmax 2000 --scoreGapNoncan -20 --scoreGapGCAG -4 --scoreGapATAC -8 --scoreDelOpen -1 --scoreDelBase -1 --scoreInsOpen -1 --scoreInsBase -1 --alignEndsType Local --seedSearchStartLmax 50 --seedPerReadNmax 100000 --seedPerWindowNmax 1000 --alignTranscriptsPerReadNmax 100000 --alignTranscriptsPerWindowNmax 10000 --outReadsUnmapped Fastx --readNameSeparator space
```

The resulting aligned sam file was converted to a gff file using `sam_to_gff3.py`. The 95% identical transcript models were then merged using `collapse_isoforms_by_sam.py`. Both scripts are run with default parameters as recommended and are available at https://github.com/Magdoll/cDNA_Cupcake

De-novo transcriptome assembly

PacBio, FANTOM cat and GENCODE GTF files were merged using gffcompare, and used as reference for StringTie de-novo assembly.

```
stringtie ${BAM} -o ${SAMPLE}_stringtie_output.gtf -p 16 -G ${reference_annotation} --fr -f 0.01
```

StringTie assemblies from 77 human islet samples culture in high glucose concentration and 53 culture in low glucose were merged using gffcompare.

```
gffcompare -r ${reference_annotation} -T -A -K -p STG -o ${OUTPUT_PREFIX} ${StringTie_assembly_GTFs}
```

Transcript expression was quantified using Salmon using the merge of StingTie assemblies as reference to build Salmon index.

```
salmon quant -p 12 --incompatPrior 0.0 --gcBias --posBias -i ${INDEX} -l A -1 ${FASTQ_R1} -2 ${FASTQ_R2} -o ${sample}_quantification
```

Single cell RNA-Seq data analysis

We compiled 4 publicly available scRNA-seq data sets of human pancreas and human pancreatic islets. We selected data sets that were sequenced using full-length transcript protocols which ensure maximum capture efficiency (Enge et al., 2017; Lawlor et al., 2017; Segerstolpe et al., 2016; Xin et al., 2016). We downloaded the individual fastq files and aligned the data to hg19 reference genome using STAR.

```
STAR –runThreadN 8 --outFilterMultimapNmax 1 --
outFilterMismatchNmax 10 --genomeDir Index --
readFilesCommand zcat --readFilesIn ${sample}_1.fastq.gz
${sample}_2.fastq.gz --outSAMstrandField intronMotif --
genomeLoad NoSharedMemory --outSAMattributes All --
quantMode TranscriptomeSAM --sjdbGTFfile
HI_transcriptome_v2.1.2.gtf
```

Gene level quantifications were obtained using salmon(Patro et al., 2017). Combat (Johnson et al., 2007) was used to correct the batch effects across 4 datasets and Seurat_v3.1.0 (Butler et al., 2018) was used to cluster the cells. A series of steps performed using Seurat are

`ScaleData` → `RunPCA` → `FindNeighbors` → `RunUMAP` → `findClusters`

We did not find any batch effects on clustering (Figure 3.17). Cell clusters were manually annotated based on the expression of marker genes (Figure 3.18).

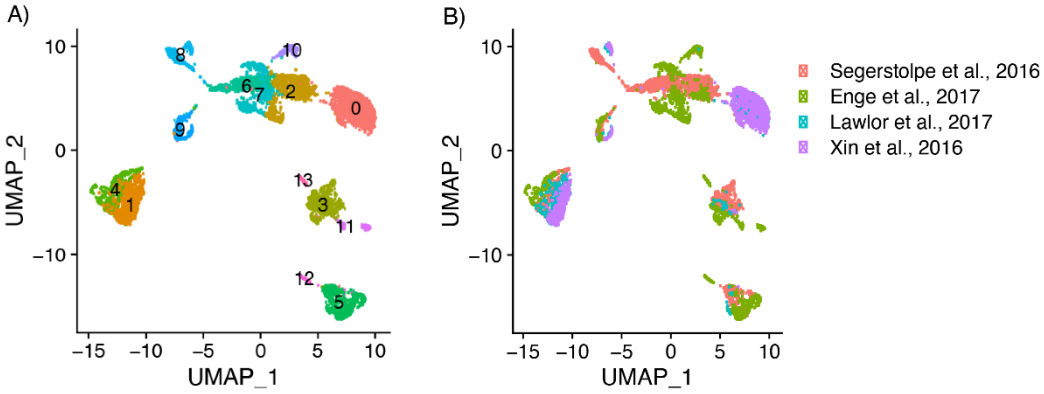

Figure 3. 2 Seurat analysis of scRNA data sets.
A) Clustering of all cells from different studies identified 12 clusters. B) Same clusters as shown in A) but the cells are colored according to the study, representing that the clustering is not driven by batch effects.

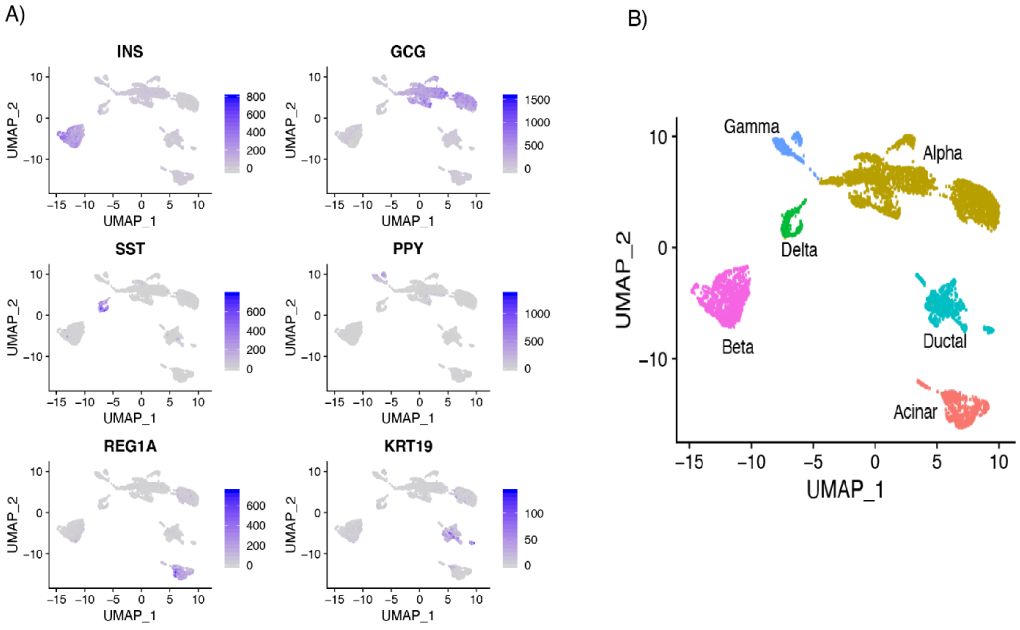

Figure 3. 1 Annotation of human islet scRNA-Seq clusters.
A) Expression of cell-type specific marker genes (*INS*: beta cells, *GCG*: Alpha cells, *SST*: Delta cells, *PPY*: Gamma cells, *REG1A*: Acinar cells, *KRT19*: Acinar cells) B) Based on the marker expression, cells were labeled with respective cell-type. This led us to identify 6 cell-types.

To obtain the cell-type specific expression of each gene, we took the mean expression of genes per cell-type and *Tau* score (Equation 3.1) (Kryuchkova-Mostacci and Robinson-Rechavi, 2016; Yanai et al., 2005) is calculated.

$$\tau = \frac{\sum_{i=1}^{n}\left(1-\widehat{x_i}\right)}{n-1};\ \widehat{x_i} = \frac{x_i}{\max\limits_{1\le i\le n}\left(x_i\right)}.$$

Equation 3.1 x_i is the expression of the gene in cell-type *i*. *n* is the number of cell-types

Tau score ranges from 0-1. A score of 0 represent ubiquitous genes and 1 represent cell-type specific genes. To identify cell-type specific genes, we selected genes with a *Tau* score ≥0.9 and subsequently performed a k-means clustering using cluster3.0 (Hoon et al., 2004) based on their mean expression levels. A heatmap representing the gene expression levels is plotted using Seaborn python package (using option z-score=True). For cell-type enriched genes, we chose all the genes whose *Tau* score is >0.5 and performed k-means clustering analysis and a heatmap is plotted as above.

Coding sequence analysis

We first identified human islet transcripts that contained an annotated CDS. We compared the annotated CDS sequence with human islet transcript sequence to identify various types of internal modifications. For the transcripts that do not contain an annotated CDS, we used TransDecoder https://github.com/TransDecoder/TransDecoder/wiki to identify a long (>= 100 amino acids) open reading frame (ORF). We then compared these ORFs against the human UniProtKB protein sequence database using BLAST-P which led to identification of 24,865 unannotated CDS. CDS with the end >50 bp upstream of the last exon-exon junction were considered to trigger NMD.

CAGE data analysis

Four human pancreatic islets were cultured in moderately high (11mMol) and low (4mMol) glucose conditions. no-amplification non-tagging (nAnTi)-CAGE protocol (Murata et al., 2014) was used to sequence 8 libraries using 100bp paired-end reads. Individual samples were aligned to hg19 reference genome using STAR aligner.

```
star --runThreadN 12 --outFilterMultimapNmax 1 --
outFilterMismatchNmax 10 --genomeDir $TMPDIR/Index --
readFilesCommand zcat --readFilesIn ${read1} ${read2} --
sjdbFileChrStartEnd $TMPDIR/Introns.bed --
outSAMstrandField intronMotif --genomeLoad NoSharedMemory
--outSAMattributes All
```

The resulting sam file was converted to bam file using samtools (Li et al., 2009). The genomic position of 5'-end of each read-1 (R1) of a paired-end read was extracted, which correspond to a TSS. The TSS from each sample were used to identify tag clusters (TC) using decomposition based peak identification (DPI) https://github.com/hkawaji/dpi1

Promoter width is defined as an interquantile range that captures 10-90% of the expression. A cumulative sum of TSS expression along the TC is calculated. The number of base-pairs between 10% and 90% of expression were defined as promoter width.

De-novo motif analysis was performed on sharp and broad promoters separately using homer (Heinz et al., 2010). A window of 500bp upstream and 250bp downstream from the promoter was chosen for motif analysis. Strand information was used depending on the directionality of promoter. All the promoters were used as background.

```
findMotifsGenome.pl input_bed hg19 ouput_folder -size
given -len 4,6,8,12 bits -p 8 -h -bg all_promoters.bed –
norevopp
```

For each gene with more than one promoter, relative promoter usage is calculated. Promoter with maximum expression is referred to as primary promoter. Promoter with next high expression is referred to as secondary promoter and expression from the remaining promoters is summed and referred to as 'Others'.

To identify human islet specific promoters, we analyzed 672 samples from FANTOM consortium. We downloaded bam files and quantified the expression (TPM) of human islet assigned promoters across all the samples. We used average TPM whenever the replicates are available from same tissue/cell-type. Then we calculated a z-score. A promoter with a z-score >3 is considered as islet specific promoter. For de-novo motif analysis, open-chromatin regions with active promoter signatures that

overlap with islet specific promoters were used (n=1851). As background, rest of the active promoters of human islets were used (n=11,827)

RNA-Seq data analysis for QTL study

We compiled publicly available genotype and RNA-seq datasets (GEO accession number GSE50244, EGA accession number EGAD00001001601), 112 unpublished samples obtained in the context of a collaboration with Piero Marchetti and sequenced through T2DSystems Horizon 2020 project, and 101 in-house samples from human islet donors without diagnosis of diabetes (after QC analysis, respectively, see below), yielding a total of 399 samples.

RNA-Seq alignment was performed using STAR aligner (Dobin et al., 2013). First, an index was generated for human reference genome version hg19. Then, Raw fastq files were aligned to using STAR (version) using the options `--outFilterMultimapNmax 1 --outSAMstrandField intronMotif --outSAMattributes All --twopassMode Basic`.

Samtools(Li et al., 2009) was used to convert the sam file format to bam format and then read-group information is added using Picard (http://broadinstitute.github.io/picard/). For EGAD00001001601, only bam files were available, hence the alignment step was not performed.

Genotype analysis

In-house samples were genotyped with distinct SNP arrays, Illumina Infinium OmniExpress 12 v1 and HumanOmni 2.5-8v1. Thus, we removed strand ambiguous variants and duplicate samples (prioritizing data from the SNP array with the largest genetic resolution) to harmonize all genotypes in a single dataset. Then, a three-step quality control of genotype data, involving two stages of SNP removal and one intermediate stage of sample exclusion, was conducted in each cohort. Genotyped SNPs were filtered if (i) minor allele frequency (MAF) < 0.01), (ii) missing genotype rate ≥ 5% and (iii) significantly deviated from Hardy-Weinberg equilibrium (HWE, p-value ≤ 1x10-6). Samples were excluded if (i) individual missing genotype rate ≥ 2%, (ii) cryptic relationships and sample duplicates (individuals with higher individual missingness genotype rate from pairs with pi ≥ 0.185), or (iii) showed >4 standard deviations from the mean according to the first four principal components in each given cohort.

For each cohort, we generated per-chromosome VCF files after checking for strand alignment against the Haplotype Reference Consortium (HRC) and 1000 Genomes (1000G) reference SNP list. HRC-1000G-check-bim.pl script with the -n option (to turn off the removal of variants showing MAF differences between the reference panel and the study genotypes) was used. We submitted resulting VCF files to the Michigan Imputation Server (https://imputationserver.sph.umich.edu/index.html): EAGLE2 was used for phasing, minimac3 for genotype imputation with the HRC r1.1 20 and the 1000G Phase 3 release reference panels, independently. For each dataset of imputed genotypes, we excluded variants with: (i) MAF < 1%, (ii) imputation-quality $R^2 < 0.7$, and/or (iii) HWE P ≤ $1x10^{-6}$. We extracted indels from the 1000G Phase3 imputed results, filtered them using the aforementioned criteria and merged with the filtered HRC imputed dataset.

Genotype principal component analysis

To identify individuals of divergent ancestry and to characterize population structure, we first selected a subset of genotyped SNPs that were common in all 4 data sets, that also passed all our QC filters and with MAF ≥ 1% and missingness < 5% across all the samples. We also excluded SNPs in high LD (pairwise r2 ≤ 0.1 within 1 Mb window), C/G and A/T SNPs to avoid strand mismatches, and those located in previously reported regions with long-range LD. We aggregated the 1000 Genomes Phase3 reference dataset using the set of overlapping variants. flashPCA tool was used to calculate genetic principal components.

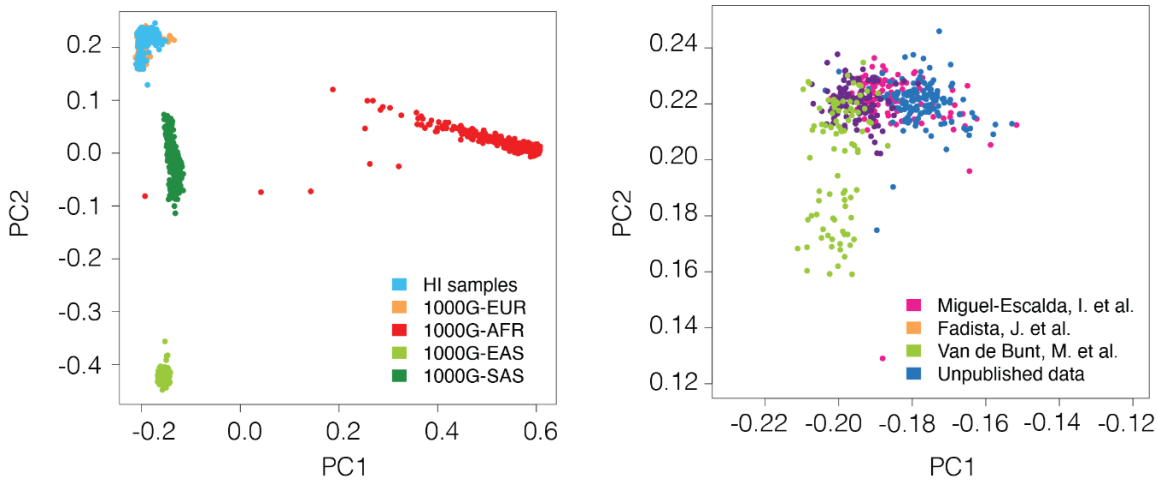

Figure. 3 Principal Component analysis (PCA) of population structure on islet samples.
We calculated PCs from genotypes with the aggregation of 1000 Genomes Phase 3 data. (A) The population structure of the islet samples included in this study was assessed by the comparison with major population groups from 1000 Genomes Phase 3 data using PC1 (x-axis) and PC2 (y-axis). (B) Differences in population structure between the four cohorts that are part of our complete panel of 399 islet transcriptomes through PC1 and PC2.

Correcting allelic bias mapping using WASP Pipeline

All reads from WASP pipeline (Geijn et al., 2015) was used to remove reads mapped with allelic bias. First the genotype vcf file was converted to a HDF5 file format using the following command.

```
snp2h5 --chrom chrom.hg19.txt --format vcf --haplotype haplotype.h5 --snp_index index.h5 --snp_tab tab.h5 --geno_prob geno_probs.h5 chr*.vcf.gz
```

Once the VCF file was converted to HDF5 format, the WASP alignment and correction steps were carried out. Briefly, first WASP identifies all the reads that may have mapping biases using `find_intersecting_snps.py`. For all the reads that overlap a heterozygous variant, the two allelic version of the reads are generated. The two allelic versions of the reads are mapped back to the reference genome using exact parameters as that of original alignment step. Then, `filter_remapped_reads.py` is used to filter out reads where one or more of allelic versions of the reads fail to map back to the same location as the original read.

VerifyBAMID(Jun et al., 2012) was used to assess the concordance between genotypes and RNA-Seq samples using options `--best --precise --`

`maxDepth 200`. Samples with more than 2% contamination (`CHIPMIX >> 0.02 and FREEMIX >> 0.02`) were removed.

Note: In-case of EGAD00001001601, only bam files were available, hence the initial alignment step was not performed.

Gene expression quantifications for QTL study

The in-house developed transcriptome annotations were used to quantify gene expression. featureCounts(Liao et al., 2014) was used to get the gene level qualifications using default parameters except using appropriate strandedness flag for each dataset and batches. Genes with less than 5 raw reads mapped in less than 5% of the samples were removed. Counts per million (CPM) normalization was performed using edgeR(Robinson et al., 2010) *cpm* function and then the normalized expression values were `log2` transformed. `Combat` (Johnson et al., 2007) was used to remove known batch effects (Figure 3.36) and 15 principal components (PCs) were calculated on Combat corrected gene expression using *prcomp* function in R-programming which were later used as covariates in eQTL analysis.

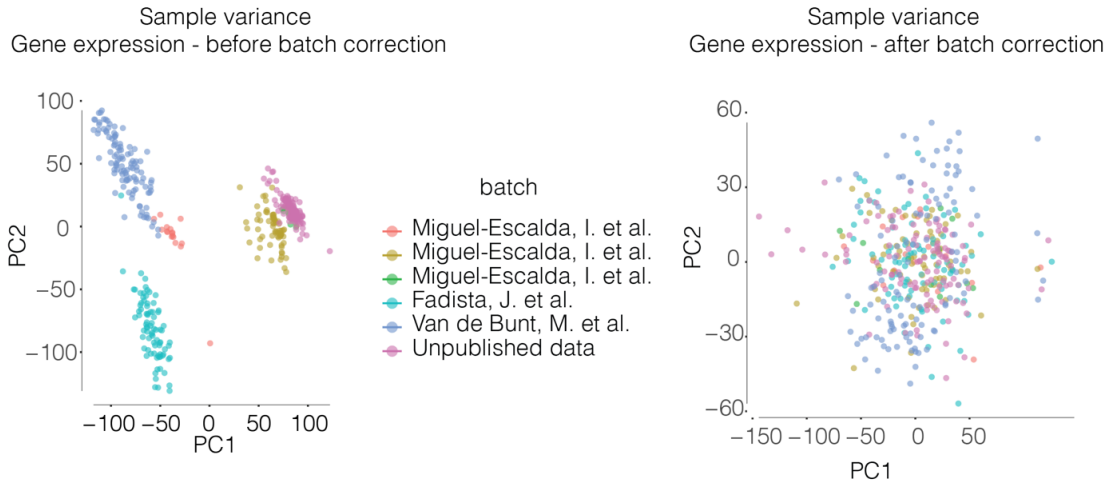

Figure 3. 4 Gene expression principal components before and after correcting batch effects.

Splicing activity quantification

To quantify the splicing activity, we used the annotation free method, *leafcutter* (Li et al., 2018). Briefly `bam2junc.sh` *from* `leafcutter` was used to quantify de-novo the number

of junctions spanning reads i.e., split mapped reads. We removed junctions that are not supported by at least 5 spliced reads in 10% of the samples and then clustered the junction spanning reads that are anchored on common junctions using `leafcutter_cluster.py` using the options -m 30 -l 500000 to get the read quantifications per junction and corresponding cluster information. `prepare_phenotype_table.py` was used to get relative junction usage (ratios) across samples. `Combat` was used to remove know batch effects (Figure 3.37) and 5 principal components were calculated on `Combat` corrected junction usage using `prcomp` function in R programming.

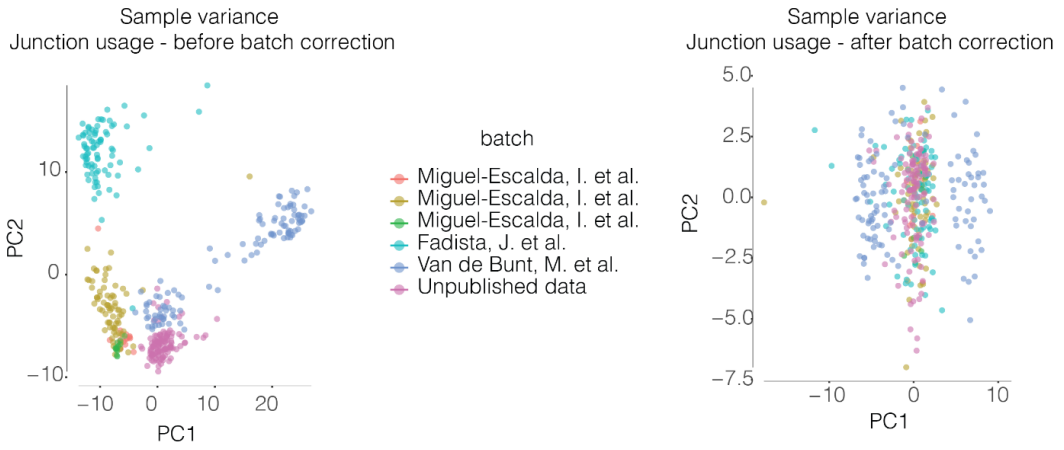

Figure 3. 5 Leafcutter junction usage principal components before and after correcting batch effects.

cis-e*QTL* mapping

cis-eQTL mapping was performed using QTLtools (Delaneau et al., 2017)for 399 samples with available genotype and RNA-seq data after quality control analysis using a cis-window of 500 kb up- and downstream of the transcription start site (TSS). 15 principal components derived from gene expression and 4 genetic principal components were used as covariates in the linear model. In order to identify best associated cis eQTL SNP-eGene pairs, QTLtools was run using the permutation pass mode (1000 permutations), and beta approximated permutation p-values were adjusted for multiple testing correction using Storey q-values implemented in the *qvalue* R package. We set the significance threshold at FDR < 0.01. We also calculated nominal p-values for all cis-SNPs within a 500kb window centered on the TSS of each gene (nominal pass mode from QTLtools). To identify all significant variant-gene pairs, we defined a genome-wide p-value threshold (*pt*), by considering the empirical p-value of the eGene closest to the 0.01 FDR threshold. A gene-based nominal p-value threshold was then calculated using *pt* and the beta distribution parameters from QTLtools. For each significant eGene, variants with a nominal p-value below the gene-level threshold were considered in subsequent analyses (significant nominal cis-eQTL variants).

cis-s*QTL* mapping

We performed cis-sQTL mapping as described in above using intron excision ratios *and a* cis-window of 50 kb up- and downstream of the junction. In case of cis-sQTLs,

only 5 PCs derived from splicing ratios and 4 genetic PCs were used in the linear model. In order to identify best associated cis sQTL SNP-junciton pairs, QTLtools was run using the permutation pass mode (1000 permutations), and beta approximated permutation p-values were adjusted for multiple test correction using Storey q-values implemented in the *qvalue* R package. We set the significance threshold at FDR q-value ≤ 0.01 resulting in 4,858 junctions with a significant sQTL. We also calculated nominal p-values for all cis-SNPs within a 50kb window around the junction (nominal pass mode from QTLtools). To identify all significant variant-junction pairs, we defined a genome-wide p-value threshold (*pt*), by considering the empirical p-value of the junctions closest to the 0.01 FDR threshold. A junction-based nominal p-value threshold was then calculated using *pt* and the beta distribution parameters from QTLtools. For each significant junction, variants with a nominal p-value below the gene-level threshold were considered in subsequent analyses (significant nominal cis-eQTL variants).

Annotation of sQTL junctions

As leafcutter is a de-novo-based method, we used the transcriptome annotation GTF file as a base to annotate the leafcutter derived junctions with respective genes. We used *gtf2leafcutter.pl* script from leafcutter's leafviz module to obtain the intron coordinates of all gene. The sQTL junctions were then mapped to the intron coordinates and annotated with respective gene names.

Magnitude of genetic effects on splicing

To quantify the magnitude of genetic effects on splicing, for each sQTL junction, we calculated the difference (delta-psi) in median junction usage of samples with homozygous reference and homozygous alternate alleles. This was plotted as a function of -log10(p-value).

Visualization of splicing events

The junctions identified from leafcutter were loaded into IGV (Robinson et al., 2011) for visualization.

Genomic regions enrichment analysis

We used GREGOR (Schmidt et al., 2015)to perform enrichment analysis of lead sQTLs and eQTLs in different genomic annotations using the following options R2THRESHOLD=0.7, LDWINDOWSIZE=50000 (for sQTLs), LDWINDOWSIZE=100000 (for eQTLs), MIN_NEIGHBOR_NUM=200 and POPULATION=EUR.

We used the islet-regulome annotations and genic annotations from GENCODE.

Comparison with GTEx

We obtained summary statistics data for eQTLs and sQTLs from 49 GTEx tissues (v8 release). We first listed eGenes and junctions with significant eQTLs and sQTLs, respectively, at FDR 0.01, consistent with our significance threshold. Significant variant-phenotype associations for each of the 49 tissues were filtered based on the previous feature sub-selection, and variant and junction coordinates (for sQTLs) were lifted down using liftOver from hg38 to hg19. For each GTEx tissue, we looked at the variant-phenotype (eGenes or Junctions) overlap with our islet eQTL and sQTLs, using nominal QTL variants. For example, for GTEx *x* eGene in *j* tissue, if any of the GTEx significant variants mapped any of our nominal eQTL variants for that *x* eGene, we considered that islet eQTL to be shared with that given *j* tissue. Same approach was implemented to sQTLs. We excluded from this analysis testis, given the pervasive number of eQTLs, and pancreas because it is a partial surrogate of pancreatic islets.

Credible set analysis

We used fine-mapping approaches to identify candidate causal variants that underlie cis-eQTL and sQTL *loci*. We identified 95% credible set variants using CAVIAR(Chen et al., 2015) software and allowing for one causal variant (*-c 1*). LD information between SNP pairs (i.e. the *r* matrix) was generated using PLINK (Purcell et al., 2007) v1.9 *–matrix –r* , and our effective 399 high-quality human islet samples used in the eQTL and sQTL identification (see Methods, *Genotype analysis* and *Correcting allelic bias mapping using WASP Pipeline* sections) as reference panel.

DeepSea annotations

The credible set variants of both eQTLs and sQTLs were assessed for their disease impact on the basis of their predicted transcriptional and post-transcriptional regulatory effects, using a deep-leaning model that is based on (a) (a) DeepSEA, trained on

transcriptional regulatory features (histone marks, DNAse I profiles and transcription factors, a total of 2,002 features), and (b) Seqweaver, trained on post-transcriptional regulatory features i.e RNA-binding proteins binding data based on CLIP experiments on 82 unique RBPs (ENCODE and other CLIP datasets) (Zhou et al., 2019). We performed in-silico mutagenesis on both eQTL and sQTL credible set variants using both DeepSEA and Seqweaver models and obtained the Disease Impact Scores (DIS) from respective models. Prior to in-silico mutagenesis, the strand information was added to sQTL credible sets based on the orientation of the gene. The credible sets were further stratified based on their location in the genome and the DIS from both models for each category of variants were shown as boxplots.

Quantile-quantile plots:

In order to estimate genomic inflation of T2D risk in transcriptomicquantitative trait loci (enrichment of small T2D GWAS pvalues among e and sQTLs), we generated quantile-quantile (Q-Q) plots using summary statistics from Mahajan, et. al 2018. We included variants with MAF $\geq$ 5% that were intersected with our nominal e and sQTLs (see Methods XX). To provide further support to the enrichment of sQTLs and eQTLs in T2D GWAS data, we generated 1000 permutations of subsets of control sQTL variants. Each control set of sQTL-like variants was generated by first identifying independent recombination regions, defined by Berisa and Pickrell, 2016, that comprised nominal eQTL or sQTL variants, respectively. Then, we shuffled non-overlapping genomic regions, that were created based on our nominal sQTL variants, across the genome, but excluding those recombination regions where either eQTL or sQTL variants were located, and blacklisted regions (wgEncodeDacMapabilityConsensusExcludable.bed.gz and wgEncodeDukeMapabilityRegionsExcludable.bed.gz). Among the set of shuffled recombination regions, we randomly sampled the same number of nominal sQTL variants. This was done 1000 times.

TWAS analysis:

FUSION software (Gusev et al., 2016) was used for the TWAS analysis. For gene expression, first the weights were computed using *FUSION.compute_weights.R* using options `--models`

top1,blup,bslmm,lasso,enet on the same data that is used for eQTL analysis. 15 PCs and 5 genetic PCS were used as covariates and variants with-in a *cis*-window of 500kb from TSS were used to compute the gene expression heritability. For splicing analysis, variants with-in 50kb from the junction boundaries were used and 5 PCs and 5 genetic PCs were used to compute the weights.

After computing weights, *FUSION.assoc_test.R* script was used to test for association of the computed weights and BMI-adjusted T2D GWAS summary statistics from one of the latest large-scale meta-analysis in 74,124 T2D and 824,006 controls (Mahajan et al., 2018b) . We only included GWAS data from variants that overlaid our ~6.5M high-quality imputed common genetic variants. The resulting p-values were corrected using Benjamini-Hochberg method.

Colocalization analysis across 403 independent T2D-GWASsignals.

We performed colocalization as implemented in gwas-pw at each of the selected independent T2D signals. To this end, we only considered the fraction of variants in $r^2 \geq 0.6$ with credible set variants with genetic posterior probability ≥ 0.01 for each independent T2D signal. LD calculations were performed using the genotypes of our ~399 high-quality islet samples, but if any of the selected credible set variants for a given independent signal were not included in our imputed genotypes, we then used 1000 Genomes Phase3 as the reference panel. Colocalization was performed across 1Mb genomic interval centered on the reported lead variant for a given GWAS locus. We nominated a region as a colocalized locus if the posterior probability for model 3 (presence of the same genetic variant associated with QTL and GWAS traits, "colocalization") was ≥ 0.9.

4. Results

4.1. Regulatory maps of human islets and effect of glycemic environment

The heritability of T2D has been shown to be enclosed in the islet regulatory landscape through large enrichments of T2D risk variants in human pancreatic islet enhancers (Pasquali et al., 2014). Previous definitions of human islet enhancers defined in Pasquali et al were based on open-chromatin regions defined using a combination of FAIRE-Seq and H2Z.A ChIP-seq data. However, FAIRE-Seq and H2Z.A ChIP-seq data tend to have higher signal to noise ratio that limits the resolution of the maps and thus, define broader genomic regions. Furthermore, only 2-3 samples were used in those maps. We reasoned that using high resolution Assay for Transposase-Accessible Chromatin using sequencing (ATAC-Seq) (Buenrostro et al., 2015) data from a larger number of individuals would allow us to better characterize the open-chromatin landscape in human pancreatic islets. We also sought to improve the enhancer definitions using additional data such as Mediator and Cohesin occupancy. Mediator has been shown to identify enhancers that are highly occupied by lineage specific transcription factors that regulate cell-specific gene expression programs(Whyte et al., 2013). Even though one primary role of Cohesin is to maintain the chromatin structure, it is also been observed that Cohesin regulates enhancer promoter interactions independent of CTCF (Schmidt et al., 2010).

To further investigate how the regulatory genome orchestrates the human islet function, we used a perturbation model where human pancreatic islets are exposed to varying glucose concentrations. Glucose is a primary physiological stimulus for pancreatic beta cell to secrete insulin. High glucose concentrations for prolonged periods of time have been shown to have both adverse and beneficial effects on beta cells. The adverse effects include increased oxidative stress leading to apoptosis (Poitout and Robertson, 2008; Poitout et al., 2010). On the other hand, the beneficial effects of glucose include its capacity to act as a mitogen. Studies in mouse and in human islets transplanted into mouse have reported that beta cells replicate upon glucose challenge (Alonso et al., 2007; Levitt et al., 2010), and that glucose stimulation is the underlying cause for beta cell replication (Porat et al., 2011). Furthermore, moderately high glucose concentrations can help beta cells functionally adapt to increased demands. Understanding the molecular mechanisms that control this adaptive response of beta cells to glucose could lead us to novel therapeutic targets. However,

the molecular basis of these this glucose adaptation remains poorly understood. Studies in mouse and rodent islets showed that carbohydrate response element binding protein (ChREBP, encoded by *MLXIPL* gene) has been shown to orchestrate some glucose-induced transcriptional changes (Metukuri et al., 2012; Schmidt et al., 2016). Although these studies laid the groundwork for a more proper understanding of glucose-induced transcriptional rewiring of beta cells, they have not been translated to human model systems. Furthermore, other glucose-dependent changes remain unexplored. Thus, we investigated the transcriptional and epigenetic changes to variation in glucose concentrations in human pancreatic islets.

4.1.1. High resolution human islet regulome annotations

We performed ATAC-Seq (Buenrostro et al., 2015) on 13 human pancreatic islet donor samples (median depth of 30 million reads) to define a set of consensus open-chromatin regions. We first identified stringent open-chromatin regions using MACS2 ($q < 0.05$) by pooling data from 13 samples. We also obtained less stringent ($P < 0.01$) open-chromatin regions in individual samples. We then defined consistent open chromatin regions if they were present in at least three samples as well as in the pooled set. This led us to identify 241,481 consistent human islet open-chromatin regions.

To annotate the underlying epigenome state, we generated ChIP-Seq data from histone modifications, structural proteins, and a transcriptional co-activator. We thus assayed H3k27Ac (a mark for active promoters and enhancers), H3K4me3 (a mark for active promoters), H3K4me1 (a mark for active enhancers), the Med1 subunit of Mediator (a transcriptional co-activator), CTCF, and SMC1A (subunit of cohesin complex, a chromatin structural protein that is also enriched in active regulatory elements) in human pancreatic islet donor samples. We followed a similar strategy as that of ATAC-seq analysis to define consistent peak sets for each ChIP-Seq data set using replicate samples. We observed that some regions that were not captured by ATAC-Seq were enriched with Mediator, CTCF or previously generated islet transcription factors' ChIP-Seq data (Pasquali et al., 2014); thereafter, we added those regions to the consistent accessible chromatin region set. The rationale was that it is unclear if all regulatory regions were necessarily call by ATAC-seq, and if some

genomic regions were bound by DNA-binding proteins technically this means that they are accessible. This led us to define a final set of 249,582 accessible chromatin regions. We then used an unsupervised k-medians clustering approach to group these open-chromatin regions based on the enrichment of combinations of different chromatin marks. Briefly, we defined a window of 6kb from the center of each accessible chromatin region. We then divided each 6kb region into 100bp bins and quantified the ChIP signal enrichments (as a -log10 p-value) of assayed chromatin marks. Finally, we performed a k-medians clustering of all accessible chromatin regions. This led us to group all accessible chromatin regions into active enhancers (n=45,683) that are enriched with H3K27ac and H3K4me1, active promoters (16,313) that are enriched with H3K27ac and HeK4me3, CTCF binding sites (n=29,915), inactive enhancers that show enrichment of H3K4me1 but lacks H3K27ac (n=66,029), and "inactive" regions which do not of any ChIP signal of assayed chromatin features (n=91,642) (Figure 3.6). H3K27ac and Mediator ChIP-Seq signal allowed us to further classify active enhancers into different subclasses, class I – III, which showed marked differences in the strength of these signals (Figure 3.6). We also observed, post-hoc, that active promoters are expectedly marked by strong human islet CAGE signal. These findings, therefore, defined a map of accessible chromatin regions in human islets with greater spatial and functional resolution than that provided in the team's previous epigenome maps (Pasquali et al., 2014)

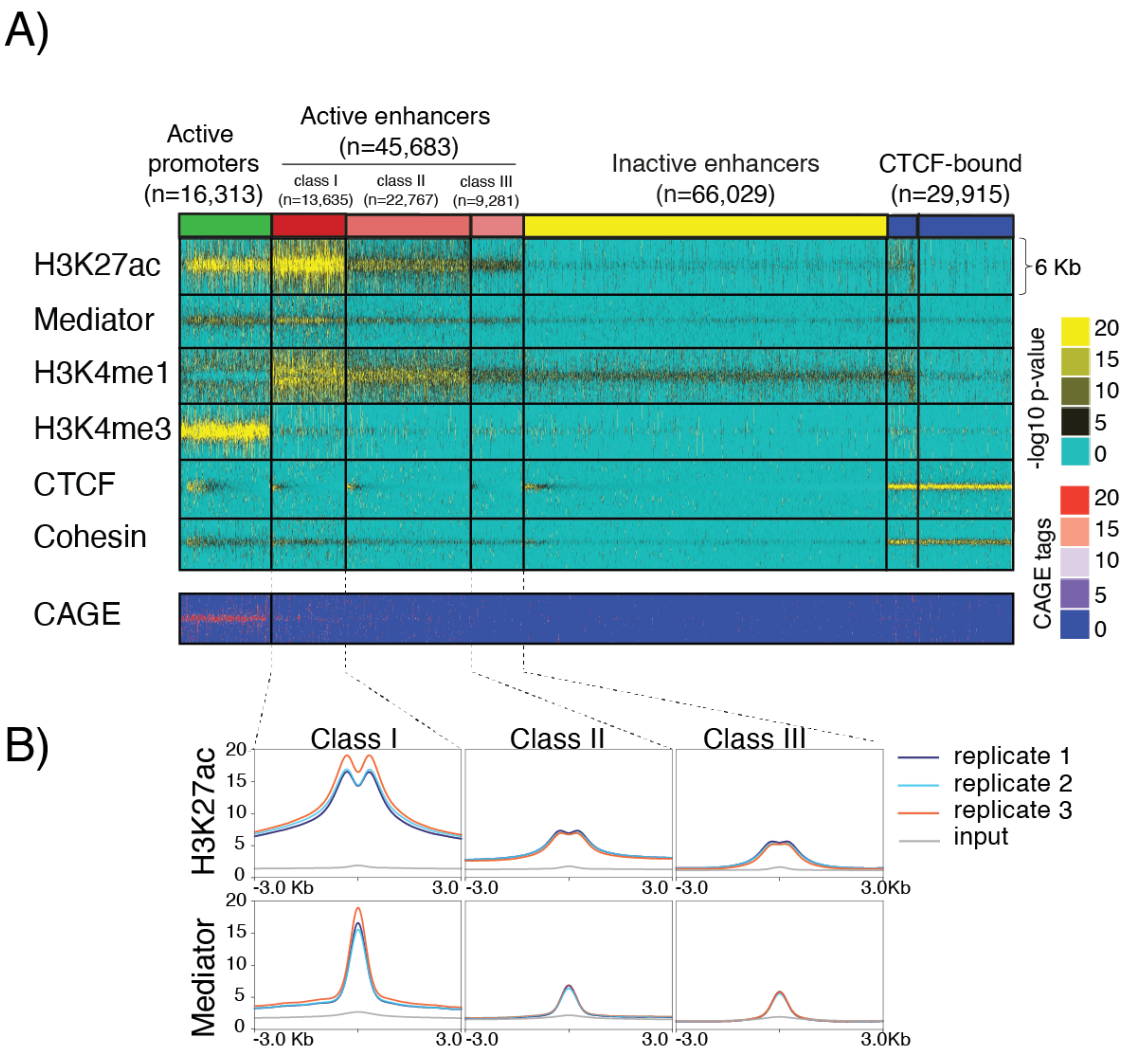

Figure 3. 6 High-resolution annotations of islet open chromatin.
A) ATAC-seq data from 13 islet samples were used to define consistent open chromatin regions, which were classified with k-medians clustering based on different combinations of epigenomic features. Mediator and H3K27ac binding patterns allowed subclassification of active enhancer classes I–III. Post-hoc analysis of islet CAGE tags confirmed that transcription start sites are highly enriched in promoters and weakly in class I enhancers. B) Average H3K27ac and Mediator signal centered on open chromatin regions for active enhancer subtypes in three human islet samples and input DNA.

4.1.2. Effect of glycemic environment on human islet regulome

To gain insights into the effect of variable glucose concentrations into the human islets regulatory landscape, we cultured human pancreatic islets from 7 organ donors in 11 mM (referred to as "high glucose") and 4 mM glucose ("low glucose") concentrations for 72h. Of note, 4- and 11-mM glucose represent glucose levels that can be observed under extreme physiological conditions. We then profiled gene expression (RNA-Seq) and H3K27Ac activity (ChIP-Seq)

High glucose induces gene expression programs beneficial for beta cells

We performed differential gene expression analysis between high and low glucose conditions that revealed 930 up-regulated genes, and 595 down-regulated genes in high glucose condition (adjusted p-value ≤ 0.05, absolute fold change ≥ 1.2, Figure 3.7). This suggested that human pancreatic islets undergo broad transcriptional changes upon glucose stimulation. Functional enrichment analysis showed that upregulated genes in high glucose condition are enriched for synaptic transmission, ion channels, and for beta cell differentiation genes such as *PDX1, NKX6-1* (Figure 3.3). We also observed that genes implicated in apoptotic pathways such as *DDIT3, JUN and TNF* are downregulated in high glucose conditions (Figure 3.8). This suggests that in our in-vitro human model system, glucose is showing a beneficial effect as it is increased expression of genes that are important for the correct function of pancreatic islet cells.

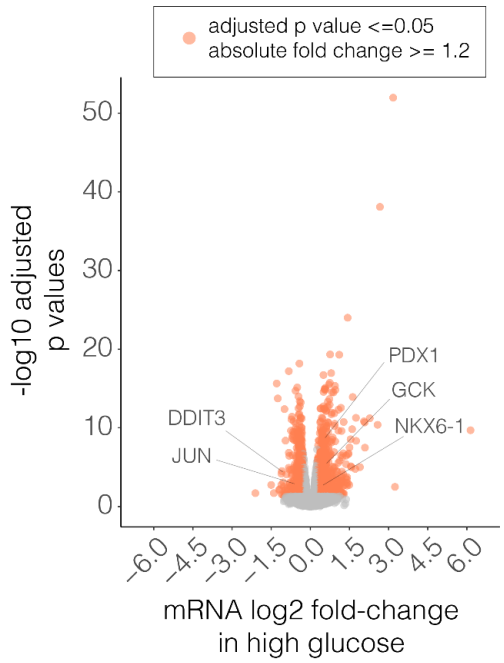

Figure 3. 7 Differential gene expression analysis between high and low glucose samples.

Volcano plot showing the gene expression log2 fold change against the -log10(adj p-value). A positive fold change indicates the a given gene is up regulated in high glucose condition.

Up regulated genes in high glucose

KEGG 2019 Human

Name	Adjusted p-value	Genes
Dopaminergic synapse	0.01124	*GNAS, SLC18A2, KCNJ6*
Nicotine addiction	0.03332	*SLC17A6, CACNA1A, GRIA4*
Type II diabetes mellitus	0.05050	*GCK, SLC2A2, PDX1*
Maturity onset diabetes of the young	0.05277	*PAX4, NKX6-1,IAPP*

GO Molecular Function 2018

Name	Adj. p-value	Genes
Voltage-gated cation channel activity (GO:0022843)	0.01657	*KCND1, KCNG3,CACNA1B*
Cation channel activity (GO:0005261)	0.04383	*SLC24A2, RYR2, TRPM5*
3',5'-cyclic-nucleotide phospho -diesterase activity (GO:0004114)	0.05060	*PDE3A, PDE3B, PDE8B*
Voltage-gated potassium channel activity (GO:0005249)	0.05542	*KCND1,KCNQ1, KCNJ16*

Down regulated genes in high glucose

KEGG 2019 Human

Name	Adjusted p-value	Genes
IL-17 signaling pathway	9e10-6	*CXCL6, HSP90AA1, CSF2*
Legionellosis	3e10-5	*HSPA1A, HSPA6, HSPA8*
p53 signaling pathway	5e10-5	*STEAP3, CDKN1A, SERPINB5*
Apoptosis	3e10-4	*JUN, TNF,ATF4,DDIT3*

GO Molecular Function 2018

Name	Adj. p-value	Genes
RNA binding (GO:0003723)	1.5e-11	*EIF4A2,CCT4,HSPA1B*
translation initiation factor activity (GO:0003743)	0.02	*EIF3E,EIFS2S,EIF3D*
vascular endothelial growth factor receptor 2 binding (GO:0043184)	0.09	*GREM1, VEGFA, VEGFB*
ubiquitin-protein transferase inhibitor activity (GO:0055105)	0.07	*RPL11, RPL5, RPL23*

Figure 3. 8 Enrichment of glucose regulated genes in functional annotations.
Genes induced by high glucose conditions are enriched for ion channels, synaptic transmission and beta cell differentiation genes. Genes repressed in high glucose conditions are genes primarily involved in apoptotic pathway.

High glucose predominately activates enhancers in human pancreatic islets

We performed differential analysis of H3k27ac activity to understand how glucose effects translate into the chromatin landscape. This revealed glucose-dependent changes in 2,847 H3k27ac regions (adjusted p-values ≤ 0.05), with 2,193 regions induced and 654 regions repressed in high glucose (Figure 3.9A). These H3k27ac-enriched regions were defined independently of enhancers or promoters, but the 2,193 H3k27ac-enriched regions induced at high glucose condition mapped to 3,065 active enhancers (of which 2,116 i.e., 69% were class-I enhancers) and 443 were active promoter elements.

We performed *de novo* motif analysis on 2,116 glucose induced class-I enhancers using HOMER (Heinz et al., 2010). As a background, we used the rest of non-glucose induced class-I enhancers (~10,000). This identified a homeobox domain motif (Figure 3.9B) to be enriched among glucose induced enhancers. This is consistent with the known up-regulation in high glucose condition of homeobox 1 genes, such as *PDX1* and *NKX6.1* that are crucial regulators of islet cell identity. Our results therefore showed that high glucose concentrations elicit quantitative chromatin changes in many human islet enhancers that are enriched with the homeobox domain motif and may thus be mediated by changes in transcription factors such as PDX1.

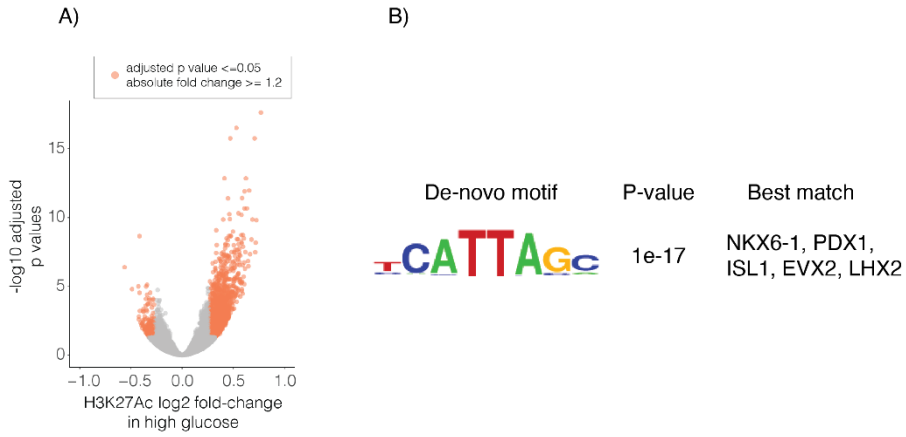

Figure 3. 9 Differential H3K27Ac activity between high and low glucose samples. A) Volcano plot showing the log2 fold-change against -log10 (p-values). A positive fold change indicates the region show higher H3K27Ac activity in high glucose. B) De-novo motif analysis in glucose induced class-I enhancers.

Glucose induced enhancers regulate glucose induced genes

To investigate if glucose-regulated enhancer activity is coupled with cognate glucose-regulated gene expression changes, we used promoter-capture Hi-C (pcHi-C) data generated after culturing human pancreatic islets in high glucose condition.

More than 80% of human islet enhancers have been linked to their target genes using targeted sequencing of promoter 3D chromatin interactions (pcHi-C) in human islets, either based on observed interactions or via imputations (Figure 3.10A), as described in (Miguel-Escalada et al., 2019).

To further evaluate if pcHi-C assignments of enhancers to their target genes truly informs us about functional interactions, we first calculated correlations of human islet enhancer and promoter activity based on H3k27ac data from a broad range of human tissues. We observed a high correlation in H3k27ac signal between human islet enhancers and their assigned gene promoters, and this correlation was greater than that of the same enhancers with gene promoters that were not connected by pcHi-C but resided in the same TAD (Figure 3.10B).

Next, we reasoned that if the epigenetic alteration at glucose-induced enhancers is coupled with glucose-regulated changes at gene expression level, we should observe an increased frequency of interactions between glucose-induced enhancer-gene pairs. Indeed, we observed that glucose-induced enhancers show enriched interactions with glucose-induced genes, compared with distance-matched non-glucose regulated genes (Odds ratio 2.7, $p= 4.9e^{-16}$ and OR 2.6, $p=6.4e^{-12}$) (Figure 3.10C). We also observed that glucose-induced enhancers do not show any increase in the frequency of interactions with glucose-repressed genes (OR 0.9, p-value 0.9 and OR 1.3, p=0.2) (Figure 3.10C). Likewise, gene promoters that are assigned to glucose-induced

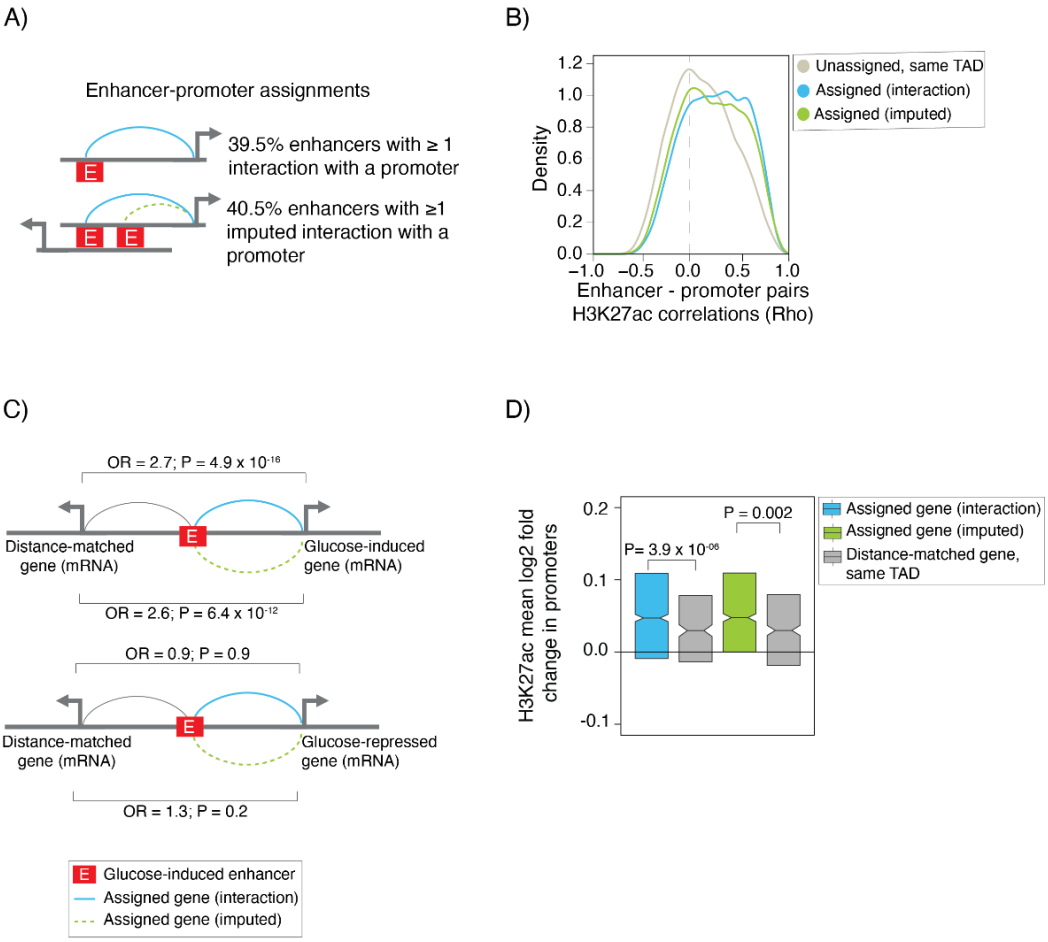

Figure 3. 10 Glucose-induced enhancer are linked to glucose-induced genes.

A) We assigned target genes to 39.5% of all 45,683 active enhancers through high-confidence interactions and imputed assignments to gene promoters for another 40% of all active enhancers. B) Functional correlations of enhancer–gene pairs assigned through high-confidence interactions (n = 18,637 pairs) or imputations (n = 28,695 pairs). Spearman's rho values for normalized H3K27ac signal in enhancer–promoter pairs across 14 human islet samples and 51 Roadmap Epigenomics tissues. Control enhancer–gene pairs were unassigned gene-enhancer pairs from the same TAD (n = 20,186 pairs). C) Genes assigned to glucose-induced enhancers showed concordant glucose-induced gene expression. Glucose-induced enhancers were enriched in high-confidence (n = 439) or imputed (n = 640) assignments to glucose-induced genes, compared with distance-matched genes from the same TAD (top). Glucose-induced enhancers showed no enrichment for assignments to genes that were inhibited by high glucose concentrations (n = 196 interacting and n = 218 imputed pairs) (bottom). OR odds ratio. P values were calculated with chi-square tests. D) Genes assigned to glucose-induced enhancers through high-confidence interactions (n = 275) or imputations (n = 321 pairs) were enriched for glucose-induced promoter H3K27ac, compared with control genes from the same TAD. Box plots represent interquartile ranges (IQRs), notches are 95% confidence intervals of median, P values are from Wilcoxon's two-sided signed ranked tests.

enhancers showed increased H3K27ac levels in high glucose condition than distance-matched genes in the same TAD ($p=10e^{-6}$, p=0.002) (Figure 3.10D). Altogether, this analysis suggested that glucose-induced changes of enhancer activity are accompanied with glucose-induced gene expression changes. It also provided a functional validation of enhancer-promoter assignments based on 3D chromatin interactions.

Glucose elicits domain-wide chromatin changes

Given the functional link between glucose-induced enhancers and their distal target genes, we wondered if glucose varying concentrations also elicit domain-wide chromatin changes.

To investigate this, we examined enhancer-hub domains. Enhancer hubs represent clusters of enhancers that shared showed 3D chromatin interactions with common genes. In a bit more detail, islet enhancer hubs were defined as three-dimensional regulatory units where multiple class-I enhancers (>3 class-I enhancers) are connected through pcHi-C interactions to one or more promoters in the same TAD (Miguel-Escalada et al., 2019) (Figure 3.11A).

We found that glucose-induced enhancers and genes were highly enriched in enhancer-hubs, compared with non-hub enhancers (Fisher's $P = 1.1 \times 10^{-7}$ and 2.2×10^{-16}, respectively). Of 297 glucose-induced H3K27ac regions that map to active promoters, 94 were contained in hubs, and 65% of these showed glucose-induced mRNA changes. We reasoned that if glucose draws out domain level changes, hub enhancers connected to glucose-induced genes should tend to show coordinated glucose-dependent changes. Indeed, we observed that hub enhancers assigned to glucose-induced promoters showed a widespread parallel increase in H3K27ac levels (Figure 3.11B). This was illustrated by the KIRREL3 hub (Figure 3.11C). This analysis revealed that varying glucose concentrations elicit chromatin changes in human islets at the level of broad regulatory domains. Together, this analysis suggests that varying glucose concentrations elicit a domain wide change in human pancreatic islets chromatin organization.

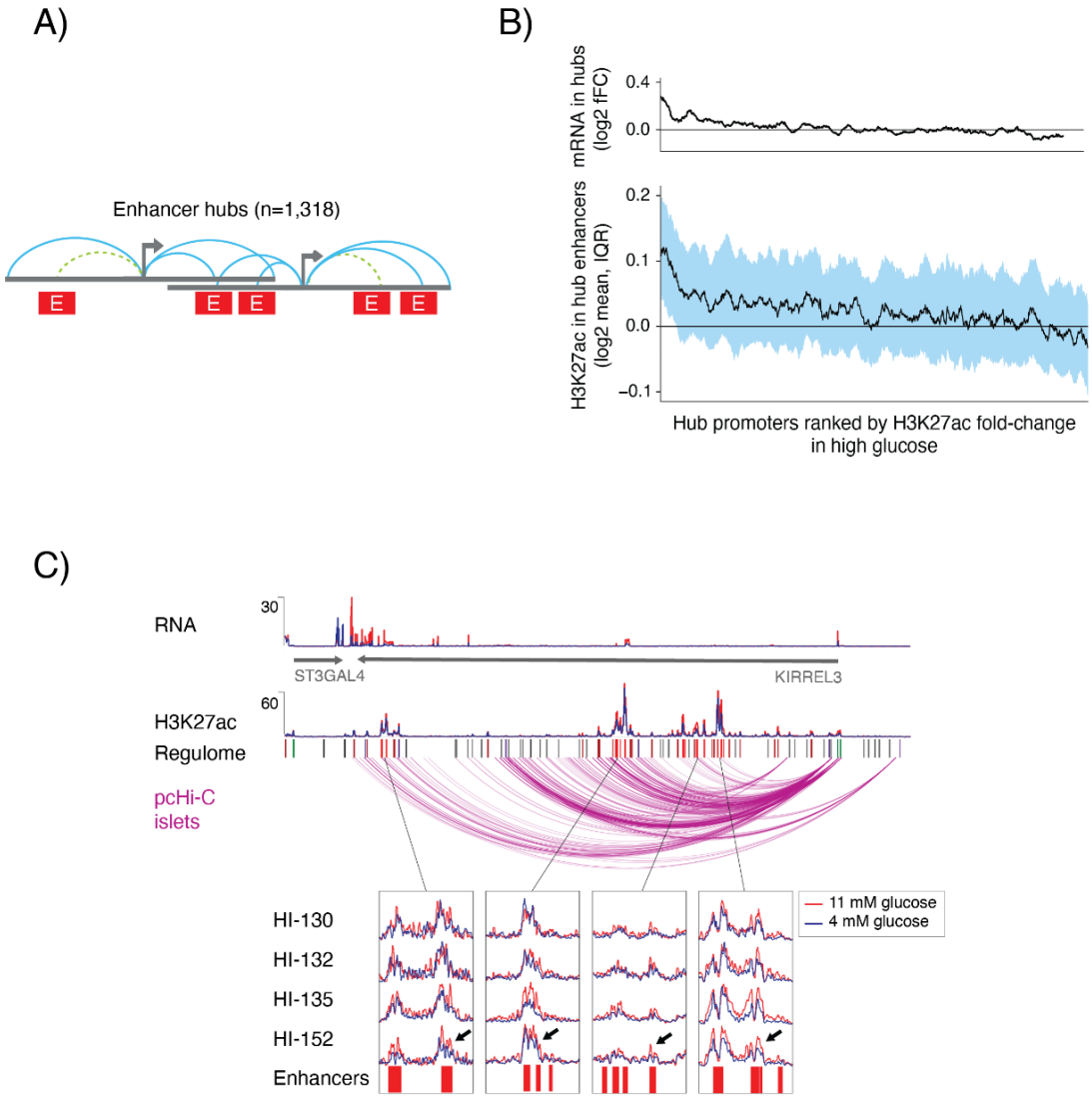

Figure 3. 11 Glucose elicits domain-wide chromatin changes
A) A schematic of enhancer-hubs. Enhancer-hubs are regulatory units where > 3 class-I enhancers are connected to one or more promoters. Red boxes represent enhancer elements, and turquoise and dashed green lines depict high-confidence and imputed enhancer assignments, respectively. B) Hub promoters were ranked by their median fold change (FC) in H3K27ac at high glucose, so that glucose-induced promoters are on the left side in the x-axis. Median mRNA levels for genes associated with each hub (top). Median glucose-dependent fold change of H3K27ac in enhancers from hubs connecting with each promoter, IQR values in blue shading (bottom). In both graphs, values are shown as running averages (window = 50). C) Coordinated glucose-induced H3K27ac in hub enhancers connected to KIRREL3. Top tracks show RNA and H2K27ac in one representative sample. Bottom insets highlight H2K27ac at 11 mM glucose (red) versus 4 mM (blue) in selected regions, showing coordinated glucose-induced changes in most hub enhancers, highlighted with black arrows (n = 4 human islet samples).

4.2. Comprehensive Transcriptome annotation of human pancreatic islets

4.2.1. Annotation of human pancreatic islet transcriptome

High quality annotation of the transcriptome is critical for the accuracy of genetic and genomic studies. Current annotations such as ENSEMBL or GENCODE have been built by compiling known gene transcript isoforms. As such, they are still biased towards transcript isoforms present in the most studied tissues and cell lines. Therefore, these annotations are still incomplete, in particular, cell-specific isoforms from less studied tissues are missing. To generate an accurate annotation of the human islet transcriptome, we decided to integrate RNA sequencing datasets from human islets.

Short read sequencing technologies produce millions of reads which can be used to reconstruct the transcript models from large panel of samples. However, such transcripts tend to be incorrect in their exonic composition and often produce inaccurate start and end positions(Steijger et al., 2013). Third generation sequencing platforms can sequence entire mRNA molecule and provide highly accurate transcript models(Sharon et al., 2013). But such technologies capture only highly abundant transcripts and are not easily scalable to large panel of human samples due to several issues. Thus, we designed a strategy to maximize the accuracy of our transcript models using long reads, short reads along with Cap Analysis of Gene Expression (CAGE) data from human pancreatic islets. An overview of the strategy is show in Figure 3.12.

Briefly, we first used long-read sequencing data to guide the de-novo transcript assembly from short reads. We generated approximately 500,000 full-length non-chimeric reads from two human pancreatic islet donor samples using Pacific Biosciences (PacBio) platform. This resulted in 79,020 non-redundant transcript models. PacBio transcript models along with the annotations from GENCODE and FANTOM-CAT(Harrow et al., 2012; Hon et al., 2017) were used as a template to perform de-novo transcript assembly from 8 billion paired-end short reads derived from 130 human pancreatic islet samples (average of 60 million paired-end reads per sample). Integrated analysis using Stringtie (Pertea et al., 2015) resulted in 10,487,818 transcripts (Figure 3.12).

Next, we used no-amplification non-tagging (nAnTi)-CAGE (Murata et al., 2014) from four human islet samples to annotate transcription start sites (TSS) with single base pair resolution. In contrast to previous studies where TSS have been linked

to transcripts based on proximity, we performed 100 bp paired-end sequencing, and used the CAGE 3' read information to accurately assign a CAGE TSS to a transcript where the 3'read was mapped. We retained transcripts that contained a CAGE TSS, as well as those with a minimum of expression of 0.1 TPM in at least 10 samples. This resulted in 202,593 transcript models that correspond to 19,812 genes (Figure 3.12).

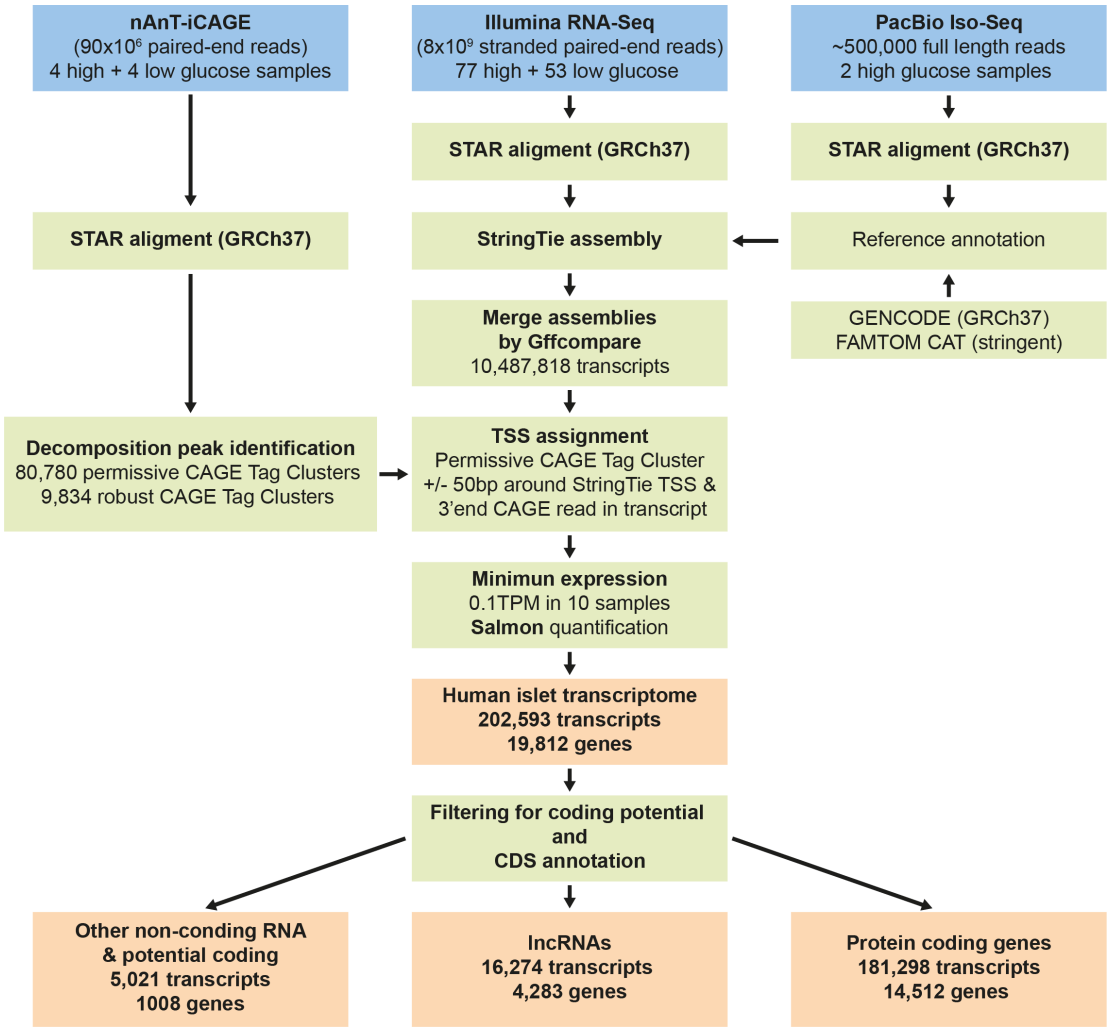

Figure 3. 12 An overview of transcriptome annotation workflow.

We further classified the transcripts into protein-coding genes, long non-coding RNAs, and others (e.g., miR, pseudogenes) based on their coding potential(Wang et al., 2013). We identified 14,512 protein coding genes, 4,283 lncRNA genes and around 1000 other type of genes.

To evaluate the novelty of our annotation, we compared it with one of the most recent GENCODE annotation (v34lift37), requiring a difference > 50 bp at the beginning and end of the transcript to call the 1st and last exon different from GENCODE. Out of all the transcripts, 4,046 were from completely unannotated genes, while 1,512 were from new spliced variant containing at least one exon not previously annotated. The majority of the newly annotated transcripts (79.4%) were novel splice variants of transcripts in which all exons had been previously annotated (Figure 3.13). This analysis, therefore, defined the known and novel human islet transcripts with accurate TSS.

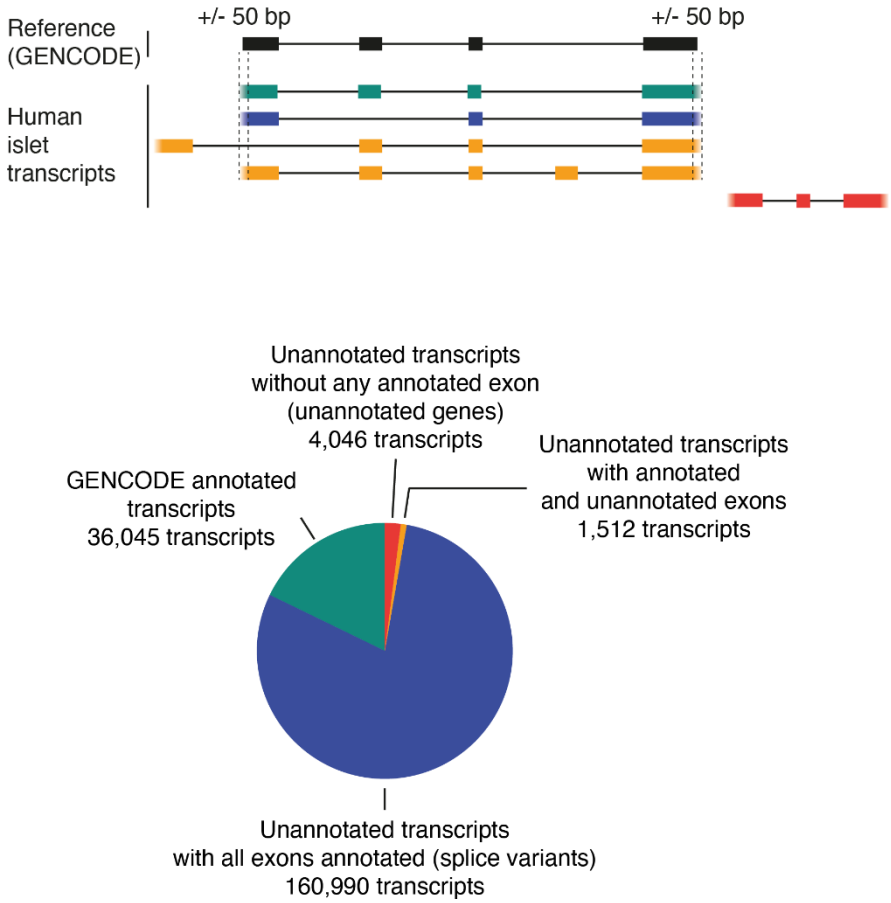

Figure 3. 13 Comparison of human islet isoforms with reference annotation maps. Top panel is a schematic showing the comparison of human islet annotations with reference annotations. Bottom pie chart indicates number of isoforms in each category.

4.2.2. Annotation of novel protein coding sequences.

One of the major challenges in transcriptome studies is to understand to what extent the vast number of alternatively spliced isoforms encode novel peptides. We therefore

decided to annotate human islet CDS based on our de-novo transcript models. To address this, we used a systematic approach. First, we identified transcripts that contained an annotated CDS with various internal modifications (Figure 3.14A). Then, for transcripts that do not contain an annotated CDS, we used TransDecoder to identify a long (>= 100 amino acids) open reading frame (ORF). We then compared these ORFs against the human UniProtKB protein sequence database which led to identification of 24,865 unannotated CDS. A substantial number of these seemingly novel CDS could be a byproduct of alternative splicing with low expression and/or be regulated by non-sense mediated decay (NMD). We therefore filtered out unannotated CDSs that were most likely to trigger NMD i.e., CDS with the end >50 bp upstream of the last exon-exon junction. We also removed CDS retaining introns as they are less likely to be translated. Finally, we required a minimum of expression of 1 TMP, and >20% of expression of all CDS transcripts of the gene (Figure 3.14B). This led us to identification of a total of 938 unannotated CDSs from annotated genes that have a significant expression and could represent new protein isoforms specifically expressed in human islets. The large majority of these new CDSs are generated by alternative splicing of already annotated exons. However, we identified 67 CDSs containing a coding exon that was not previously annotated by GENCODE. In ongoing experiments, in collaboration with Prof. Alan Attie and Prof. Lloyd Smith, we are comparing these predicted CDSs to proteoforms identified by mass spectrometry to validate the existence of these peptides.

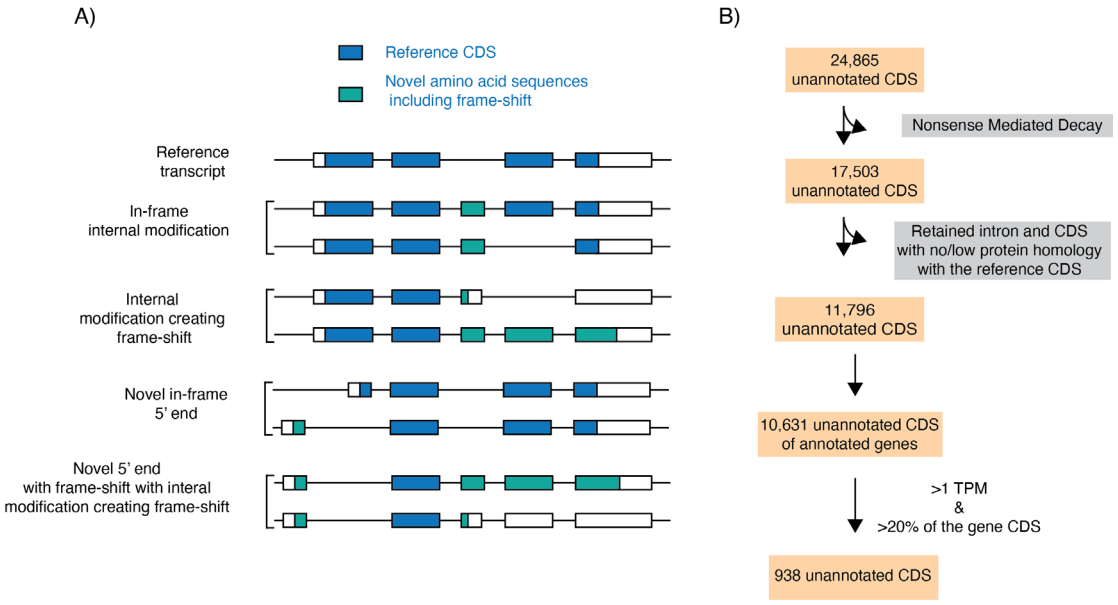

Figure 3. 14 Identification of unannotated coding sequences.
A) Schematic representing the identification of unannotated coding sequences from annotated genes. A reference CDS sequence that is contained within islet transcript was compared with respective human islet transcript sequence to identify various amino acid sequence modifications. B) Workflow to identify potentially novel protein coding CDS.

4.2.3. Annotation of islet transcriptome at single-cell resolution

With the advancement of single cell genomic technologies, it is now possible to measure gene expression levels in individual cells (Eberwine et al., 2014). Understanding cell-type specific gene expression patterns will provide insights into how the repertoire of genes contributes to cellular identity and function. We compiled 4 published single cell RNA-Seq (scRNA-Seq) data sets of human islets and human pancreas, based on Full-length SMART-Seq protocol (Enge et al., 2017; Lawlor et al., 2017; Segerstolpe et al., 2016; Xin et al., 2016). Briefly, we quantified the gene expression levels based on our transcriptome annotations. Dataset specific effects were removed using combat. The cells are clustered together using Seurat workflow. The clustering analysis defined transcriptomes that matched major islet cell types, namely Alpha, Beta, Delta and Gamma along with exocrine Acinar and Ductal cells. This revealed cell-type specific gene expression patterns of both protein coding and lncRNA genes.

We calculated a Tau-score of each gene (Yanai et al., 2005). Tau-score ranges from 0-1, 0 refers to ubiquitous expression and 1 refers to cell-type specific expression. We found 226 genes (of which 16 are lncRNA genes) that were expressed in only one pancreatic cell-type (Tau-score > 0.9). We also annotated 2986 genes (of which 502 are lncRNA genes) whose expression is enriched in one or more cell-types (Tau-score > 0.5) (Figure 3.15). Given the low capture efficiency and low depth of sequencing, it is challenging to annotate the cell-type specific isoform expression. There is an ongoing effort to characterize the cell-type specific isoform expression by pooling data from multiple cells using on k-nearest neighbors (KNN) approach. These results, therefore, revealed cell-type specific and cell-type enriched protein coding genes and lncRNA genes.

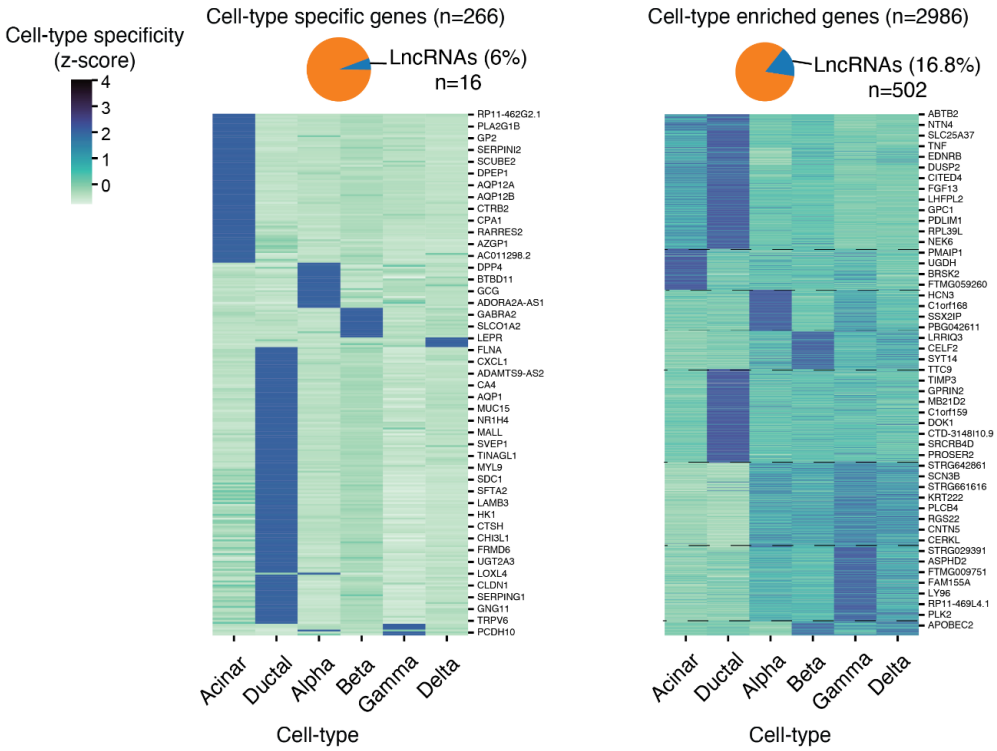

Figure 3. 15 Cell-type gene expression patterns.
Heatmaps representing cell-type specific and cell-type enriched gene expression patters. Pie-charts represents the number of protein-coding genes and lncRNA genes in each category.

4.2.4. Promoter landscape of human pancreatic islets

Gene transcription is initiated by RNA-Polymerase II at precise locations in the genome. The DNA regions surrounding transcriptional initiation sites are known as promoters. The promoter architecture is crucial for spatial, temporal and tissue specific gene expression patterns(Lenhard et al., 2012). The promoter landscape of human pancreatic islets is generally unknown.

The 5'-end of each CAGE paired-end read of annotated transcripts was defined as a TSS. TSS that were within 100 bp of each other were merged to form a tag cluster (TC) and the number of reads that formed a TC was used to determine the expression of the TC. We used decomposition based peak identification (DPI) method ((DGT) et al., 2014) to define a set of 61,337 permissive (>3 reads per TC) and 19,834 robust promoters (>11 reads per TC). As expected, we observed that the width of promoters shows a bi-modal distribution that can be separated into broad (>10.5 bp) and sharp promoters (<=10.5bp) (Carninci et al., 2006) (Figure 3.16A). The sharp promoters were enriched with the TATA-binding protein binding site motifs while the broad promoters were enriched with ETS domain motifs (Figure 3.16B). These results were expected for these promoter types(Carninci et al., 2006; (DGT) et al., 2014), and supported that our CAGE-based promoter definitions comply with known mammalian promoter architecture subtypes.

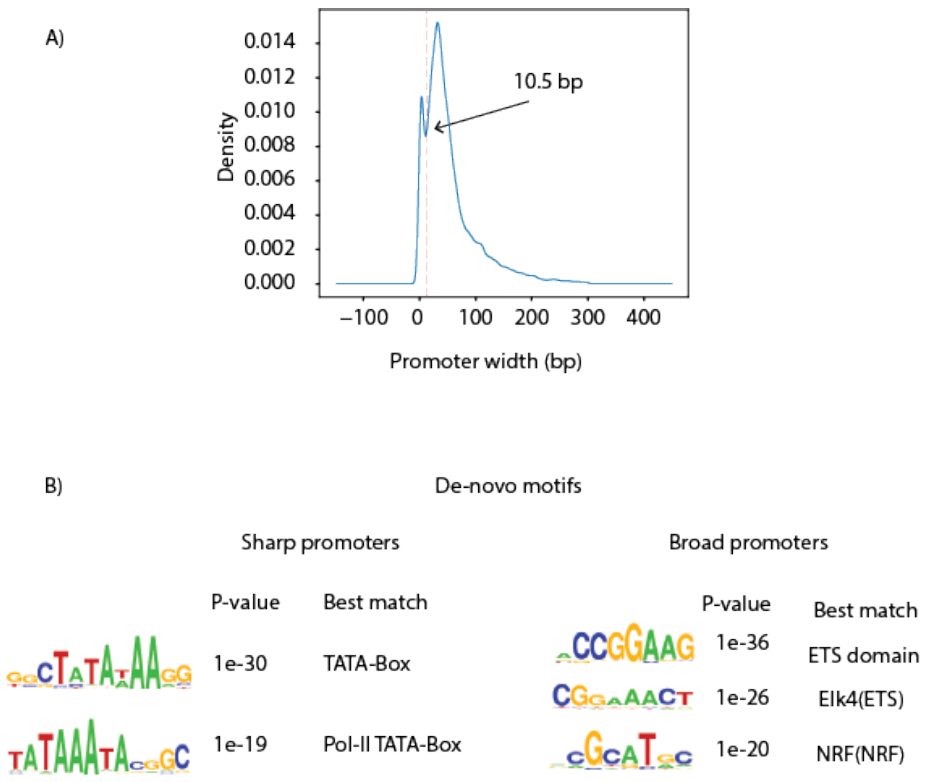

Figure 3. 16 Human islet promoter characteristics.
A) Distribution of width of all human islet CAGE-based promoters, an inter-quantile range between 10-90% of expression is plotted. The arrow points to bi-modal distribution that separates sharp (<10.5bp) and broad promoters (>10.5bp) B) De-novo motifs identified via homer on sharp and broad promoters.

The TC were assigned to a transcript if the transcript resided within 50 bp on the same strand, and the 3' read overlapped an exon of the transcript (Figure 3.17A). Using this approach, 83% of the robust TCs and 25% of the permissive TCs were assigned to at least one transcript (Figure 3.17B). A majority of the assigned TCs also contained an underlying epigenome-based active promoter signature, which further supported the accuracy of our TSS annotations (Figure 3.17C).

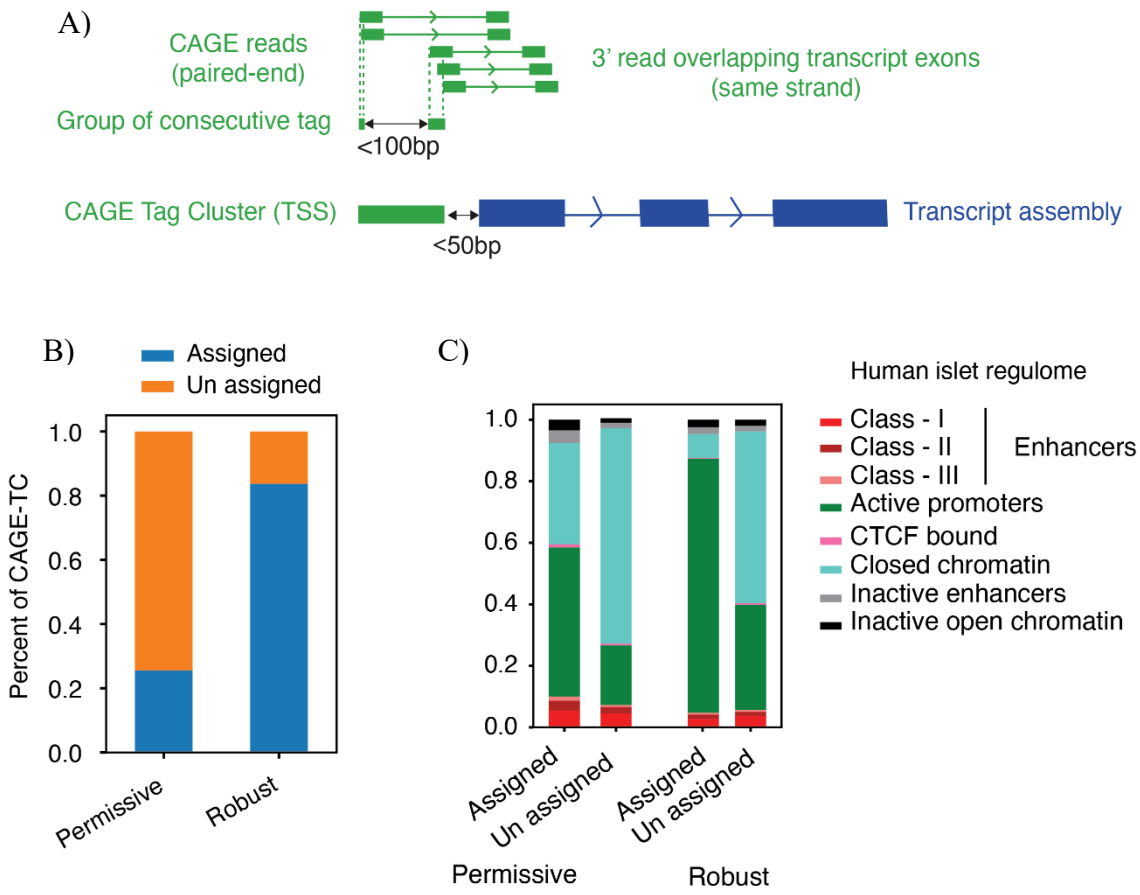

Figure 3. 17 Accurate annotation of transcription start sites (TSS).
A) A schematic representing TSS assignment. B) Proportion of robust and permissive TCs assigned to at least one transcript. C) Proportion of TSS overlapping with different epigenome-based regulatory annotations. Majority of the assigned TSS have an active promoter signature.

Thousands of novel TSS contribute to human islet gene expression

We further compared our TSS with reference annotations of transcriptional initiation sites. Among protein-coding genes, we observed ~1300 robust TSS and ~6000 permissive TCs that were >100 bp away from any predicted 5' region of GENCODE transcripts (Figure 3.18A). These represent previously unannotated promoters of protein coding genes.

We also compared our annotated lncRNA TSS to FANTOM-CAT(Hon et al., 2017), which is the most comprehensive catalogue of lncRNAs. We observed around 132 robust and 885 permissive TSS are at least 100bp away from annotated lncRNA TSS (Figure 3.18B).

Overall, we noticed that 2,566 unannotated TSS contributes to >20% of total expression of the gene (RNA-Seq) (Figure 3.18C). This suggests that a major portion of the human islets active TSS are not annotated in reference maps.

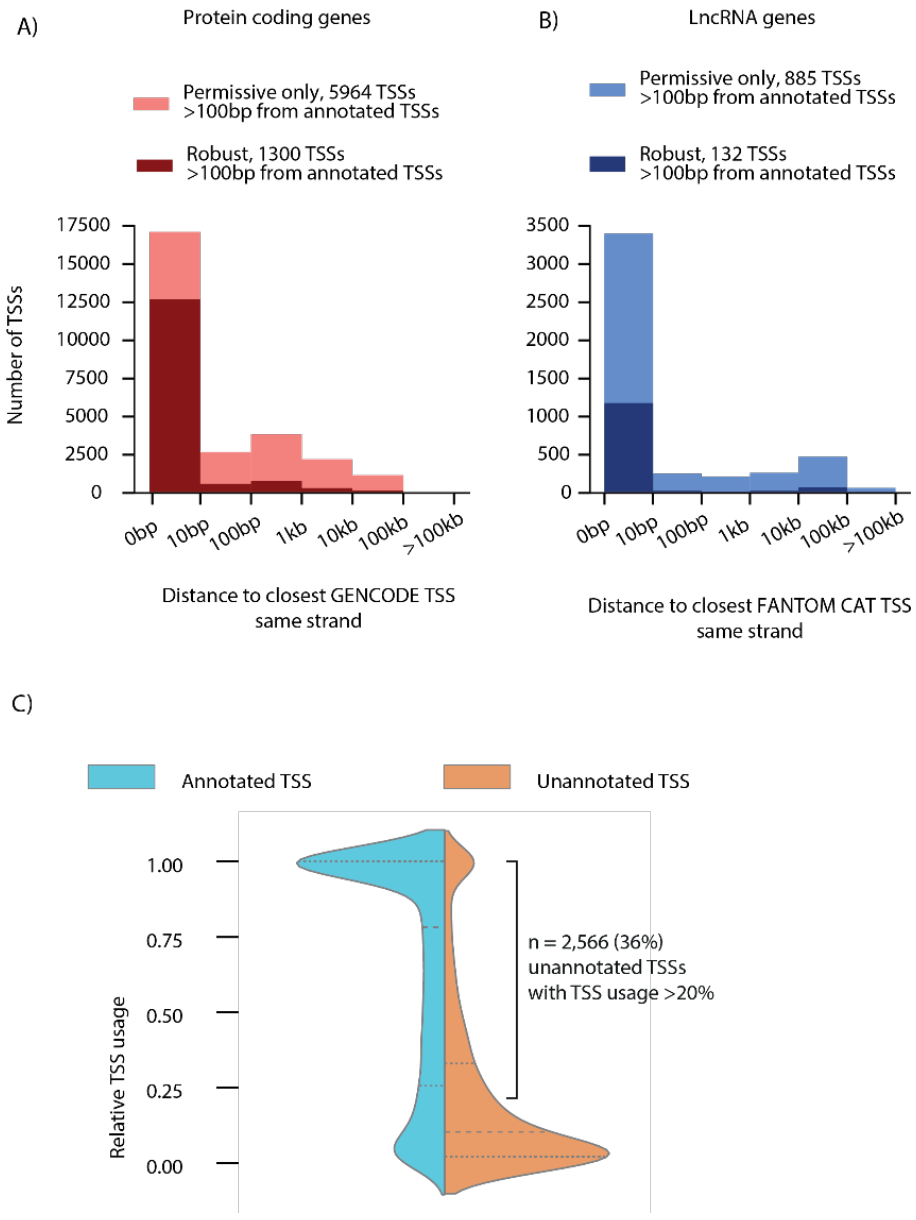

Figure 3. 18 Identification of unannotated TSS.
A) A bar plot representing number of human islets TSS of protein coding genes and their distance to GENCODE TSS. B) A bar plot representing number of human islets TSS of lncRNA genes and their distance to FANTOM-TSS. C) Relative usage of annotated and unannotated TSS of each gene. X-axis represents the density of TSS and y-axis represents the relative TSS usage.

Functional validation of a novel pancreatic islet promoter

To evaluate if our annotations are truly capturing functional TSS, we performed a clustered regularly interspaced short palindromic repeats (CRISPR) activation experiment targeting the newly identified dominant promoter of NKX6-1 gene, an important regulator of pancreatic islet differentiation (Aigha and Abdelalim, 2020) (Figure 3.19A). We designed guide RNAs (gRNAs) targeting two alternative NKX6-1 promoters. HEK293 cells were transfected in biological triplicates using FugeneHD (Promega) with a plasmid expressing dCas9-VPR activator and a plasmid expressing five guide RNAs targeting the GENCODE (v34lift37) annotated TSS region, five guide RNAs targeting the dominant human islet promoter and five control non targeting guide RNAs. RNA was isolated 72 h after transfection, retrotranscribed and NKX6-1 gene expression was measured using RT-qPCR(Balboa et al., 2015). This CRISPRa experiments showed that the newly annotated dominant promoter shows a marked increase in NKX6-1 expression (p=0.003) while the GENCODE annotated TSS show only a marginal increase in NKX6-1 expression (Figure 3.19B). This illustrates the functional value of our TSS and highlights the importance of annotating accurate TSS to aid in epigenomic modification studies.

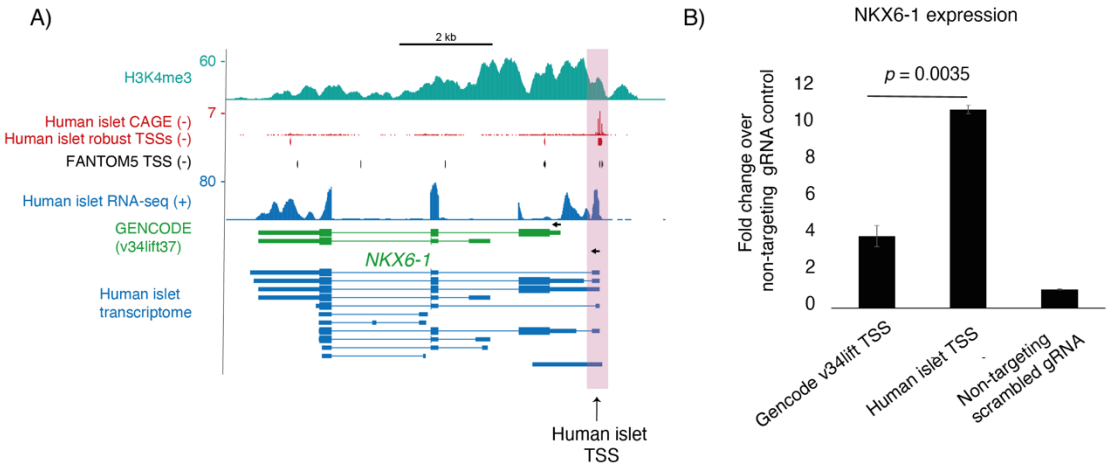

Figure 3. 19 Transcriptional activation of alternative NKX6-1 promoters
A) Genome browser screenshot illustrating the dominant TSS of NKX6-1 identified through our CAGE data. B) CRISPRa experiments in HEK293 cells showing the NKX6-1 expression after targeting two alternative TSS.

Widespread alternative promoters contribute to human islet gene expression

We found 3495 genes (23% of all active genes in pancreatic islets) that have 2 or more independent active promoters (Figure 3.20C). 3296 of these genes correspond to protein-coding genes.

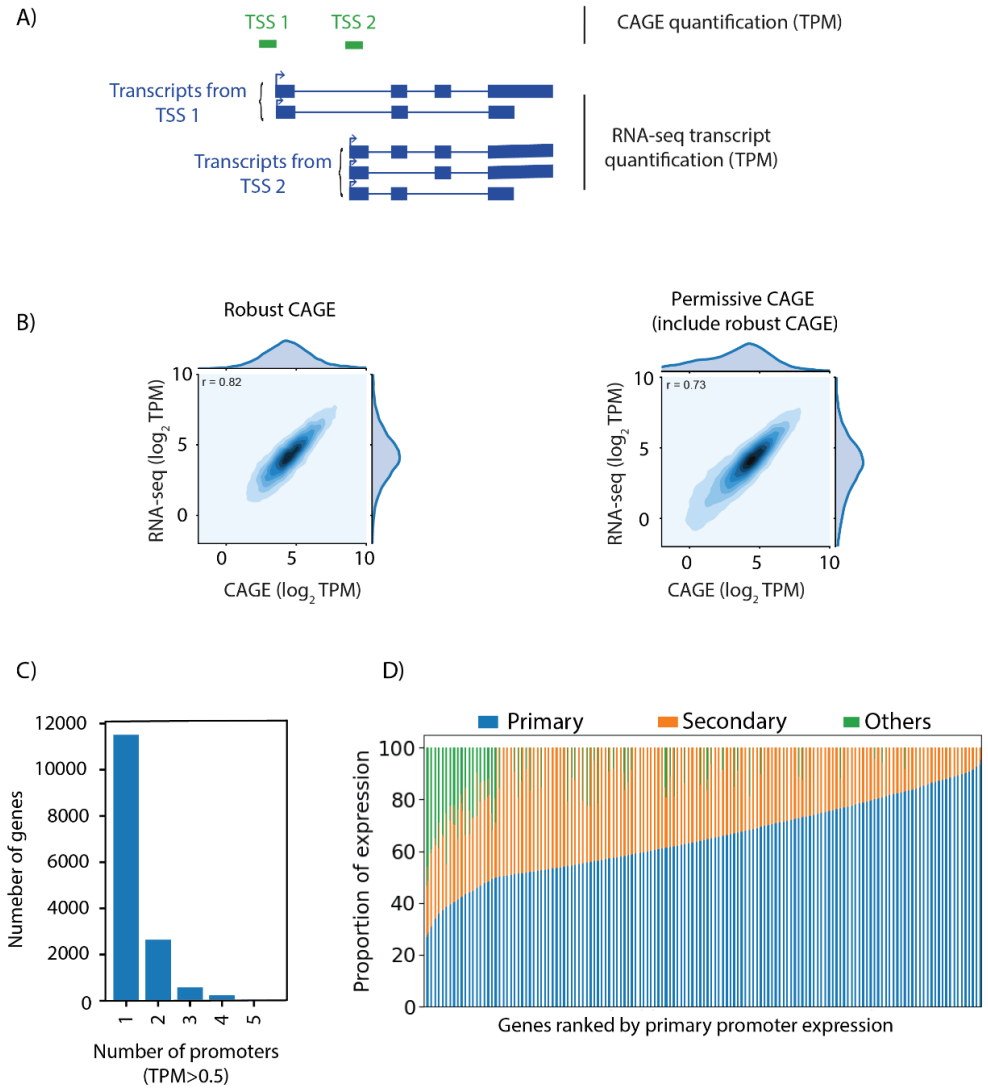

Figure 3. 20 Alternative promoter usage in human pancreatic islets.
A) A schematic representing linking CAGE-based promoter expression to RNA-Seq expression. Average expression of all transcripts that are assigned to a TSS is quantified. B) The CAGE-based promoter expression is plotted against the average transcript expression from RNA-Seq. A spearnman's correlation (r) is shown inside the plot. C) Number of independent promoters identified per gene. D) For each gene with more than 1 promoter, relative promoter expression is calculated. For each gene, promoters are ranked based on their contribution to total gene expression. On x-axis, genes are ranked in ascending order according to primary expression and proportion of expression is plotted on y-axis.

To assess if these were likely to be biologically relevant alternative promoters, we evaluated the relative contribution of different promoters to each gene, which we refer to as *relative promoter usage*. To this end, we first assessed if CAGE-based promoter activity provided reliable estimates of gene expression levels. For each CAGE TC, we calculated the average mRNA expression levels from all the transcripts that originated from that TC (Figure 3.20A) and performed a correlation analysis between CAGE-based promoter expression vs mRNA. We found a strong correlation between the promoter expression and mRNA expression (R2 = 0.82 and R2 0.73 for robust and permissive TC, respectively) (Figure 3.20B) Next, we assessed the relative contribution of promoters for each gene (Figure 3.20C). We found that more than 80% of genes that have a secondary promoter, show >20% expression from a secondary promoter (Figure 3.20D).

Analysis of islet-specific promoters

Identifying tissue specific promoters can give us novel biological insights into cell-specific transcriptional regulation. To this end, we quantified the expression of all 31,967 transcript-assigned islet promoters in 672 samples from FANTOM human tissues and primary cells. We calculated a z-score that related the expression value of each promoter in islets to that of other samples. This revealed that 5427 promoters (16% of all islet promoters) were islet-specific (z-score>3), of which 47% are robust promoters (Figure 3.21A,B). Out of 5427 islet-specific promoters, 1052 do not have evidence of transcriptional initiation sites in any of the 672 samples (<0.1 TPM across all FANTOM samples) thus represents islet selective promoters.

To gain insights into the sequence composition of islet-specific promoters, we used homer (Heinz et al., 2010) to perform *de novo* motif analysis on the upstream open-chromatin regions containing a characteristic active promoter chromatin signature, using rest of active promoters as a background sequence. Comparison of *de novo*-enriched motifs with known motifs revealed that homeobox domain, CTCF, ETS domains and zinc-finger domain containing TFs were enriched among islet specific promoters (Figure 3.21C). We would like to make a note that a more thorough analysis is being carried out to further analyze the sequence composition of cell-specific promoters, and to link islet-specific transcription factor expression with the enriched motifs in islet specific promoters. These findings, therefore, have so far shown that i) Human islet TSS are incorrectly annotated ii) Human islets use more than one promoter

to drive gene expression iii) Thousands of islet-selective and islet specific promoters harbor specific sequence determinants.

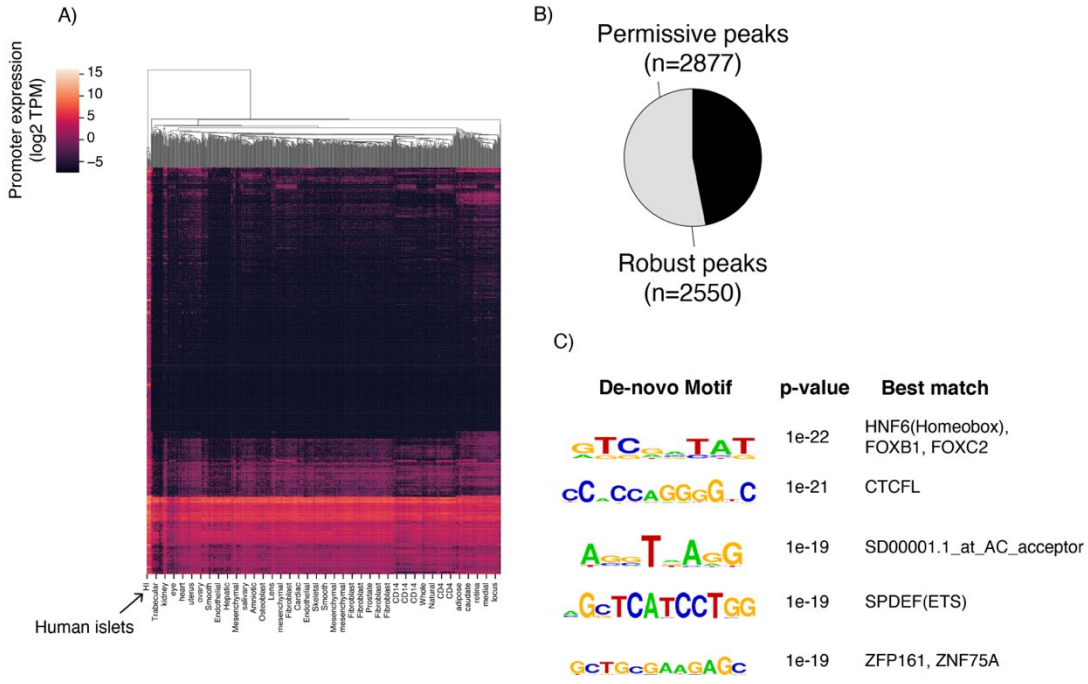

Figure 3. 21 Islet specific promoters.
A) A heatmap of promoter expression (log2 TPM) of islet-specific promoters (z-score >3) across 680 CAGE samples (includes 8 human islet samples). x-axis represents the samples and y-axis represents the promoters. Human islet samples are highlighted with an arrow at the bottom. B) A pie chart representing number of robust and permissive promoters that are islet specific. C) De-novo motifs detected on islet-specific robust promoters.

4.3. Genetic regulation of alternative splicing and gene expression.

4.3.1. Widespread effects of genetic variants on human islet splicing

The growing inventory of T2D risk associated genetic variants that GWAS have identified over 15 years did not result in connate transformative biological insights into T2D pathophysiology. As we noted in the Introduction, section 1.4, several limitations hinder the biological interpretation of GWAS results, such as high local LD or the fact that the majority of risk variants fall in the non-coding genome and lack a direct address to disease-causal target genes. In parallel, most of the efforts dedicated to characterize the molecular mechanisms underlying T2D non-coding variants only assessed the impact on islet transcriptional regulation, neglecting the effects of genetic variation on alternative splicing. Nevertheless, genetic effects on alternative splicing have been reported to underlie the etiology of several diseases(Li et al., 2016; Raj et al., 2018a). To address these limitations, we generated a catalogue of genetic effects on alternative splicing (sQTLs) and gene expression variation (eQTLs) in human pancreatic islets. This data allowed us to unveil the distinct contribution of genetic effects into alternative

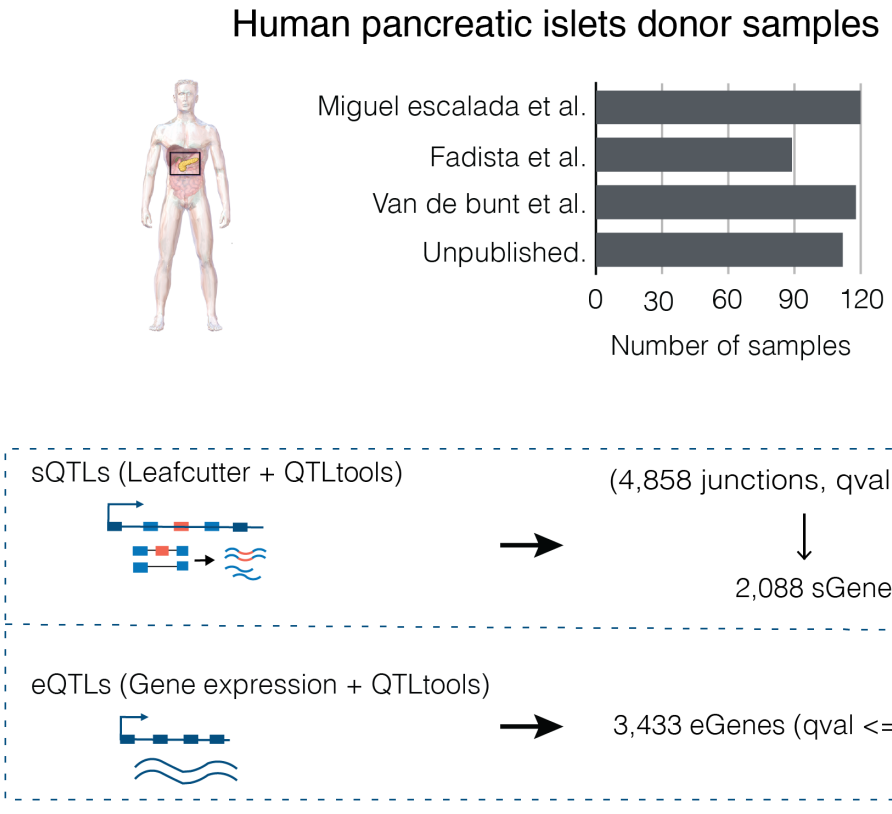

Figure 3. 22 QTL discovery in human pancreatic islets.

splicing regulation in contrast with transcriptional regulation in human islets, as well as assessing their independent contribution to T2D pathophysiology.

W aggregated unpublished and publicly available RNA-Seq and genotype data totaling 399 high-quality human pancreatic islet RNA-Seq samples (Bunt et al., 2015; Fadista et al., 2014; Miguel-Escalada et al., 2019) (Figure 3.22).

Samples were sequenced at an average depth of 50 million paired-end 100bp reads. We quantified relative junction usage using leafcutter, an annotation free method (Li et al., 2018)(Liao et al., 2014). Briefly, leafcutter uses split mapped reads to infer junction positions. A graph is then constructed based on the overlapping junctions that have a common donor or acceptor site to form a cluster of junctions. Relative junction usage in each cluster is calculated in an analogous manner to the percent spliced in (PSI) metric. Junctions that were not supported by at least 5 reads in 5% of the samples were removed before applying the leafcutter algorithm.

In parallel, we quantified mRNA expression using featureCounts (Liao et al., 2014)

After correcting for known and unknown covariates, we performed QTL analysis using ~6.5 million high-quality imputed common variants. This led to the identification of 4,858 cis-sQTLs ($q\text{-}val < 0.01$) (Figure 3.22). We mapped the 4,858 sQTL junctions to 2,088 distinct genes (sGenes, ~6% of which were lncRNA genes) in which common variants cause splicing variation in human pancreatic islets. In parallel, we identified significant cis-eQTLs ($q\text{-}val<0.01$) in 3,433 out of 16,070 tested eGenes.

We found that sQTL junctions encompassed major types of splice variants (Figure 3.23A). For example, we identified instances in which common genetic variation had major effects on the usage of the first exon (*RNF6*), caused alternative splice sites in an exon or complex splicing variation in genes that map to T2D loci (*THADA*, *KCNK16A*, respectively), or led to mutual exclusion of exons in *SLC7A2*, thus leading to a splicing variation that is known to affect the function of this amino acid transporter gene (Smith et al., 1997) (Figure 3.23B)

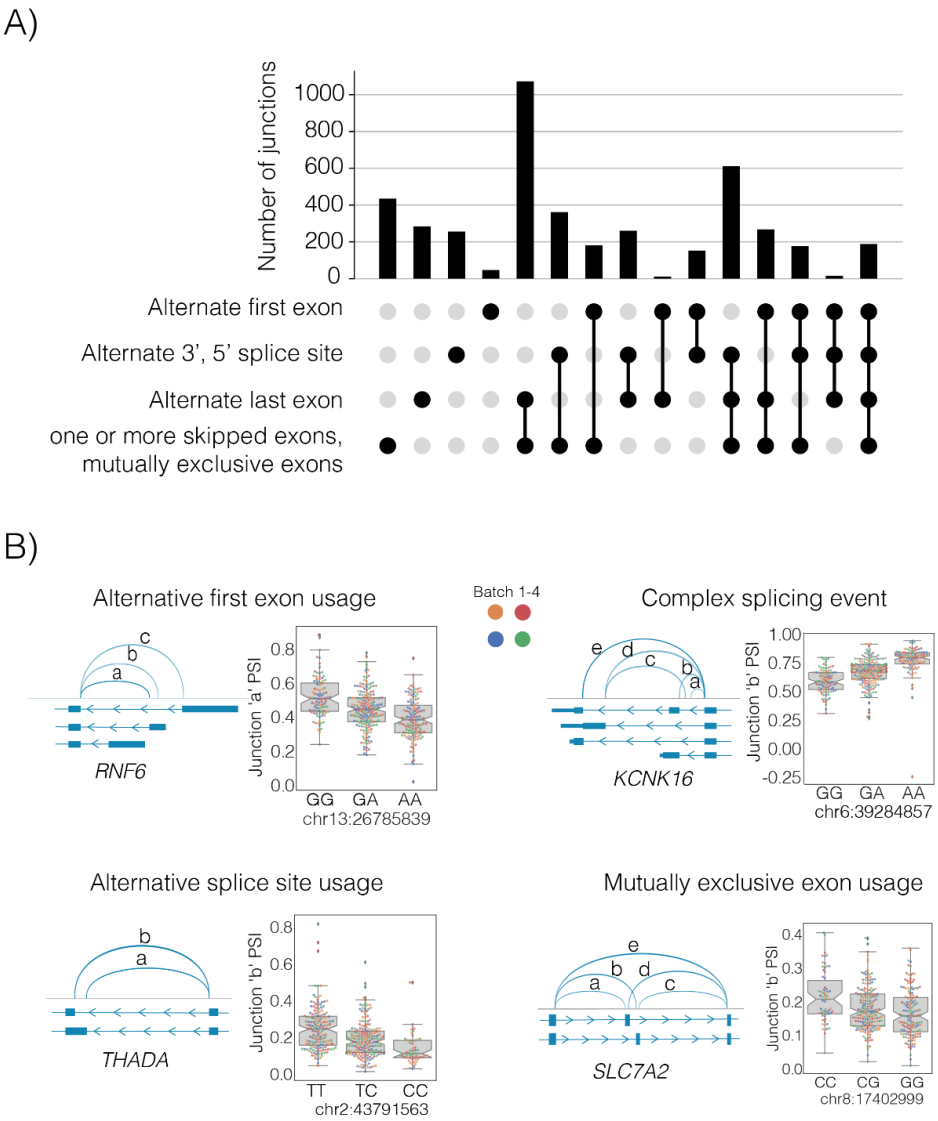

Figure 3. 23 Types of alternate splicing event under genetic effects.
A) All significant sQTL junctions were categorized based on their participation in different types of splicing events. B) Examples showing various types of splicing events under genetic control. Box plot represents the junction PSI stratified by lead sQTL genotype.

We benchmarked islet splice variants against GENCODE annotations and found that 23% of the sQTL junctions were unannotated (Figure 3.24). This overlap, however, was increased to 90% in comparison with our human islet transcript maps, further supporting that human islet transcripts are incompletely annotated. Our sQTLs, therefore, detect splicing variation in annotated and unannotated islet transcripts.

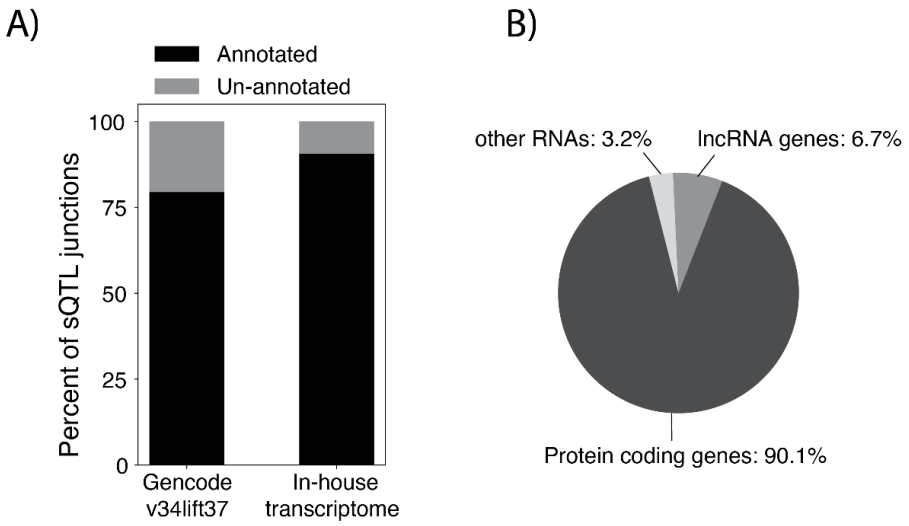

Figure 3. 24 Annotation of sQTL junctions.
A) Percent of sQTL junctions that could be annotated using GENCODE and in-house transcriptome annotations. B) Percent of protein coding, lncRNA genes and other RNAs that sQTL junction belongs to.

We further compared islet sQTLs to previously reported exon-QTLs from the largest human islet eQTL study to date (Viñuela et al., 2020), and found that only 18% of sQTL junctions were flanked by exons from exon-QTLs. Furthermore, when sQTL and exon-QTLs affected the same gene, there was limited LD between the lead sQTL and exon-QTL variant: for 45.2% of overlapping genes the LD between all lead sQTLs and exonQTLs identified for that given gene showed $r^2 < 0.6$ (Figure 3.25A, B). In contrast, 36.9% of genes that harbor eQTLs and exon-QTLs showed low LD correlation ($r^2 < 0.6$), suggesting a larger degree of genetic sharing between eQTLs and exon-QTLs (Figure 3.25C, D). This finding then suggests that sQTLs, which directly measure splice junction variation, and exon-QTLs, which measure exon levels and can thus be influenced by variables unrelated to RNA splicing, capture fundamentally different events.

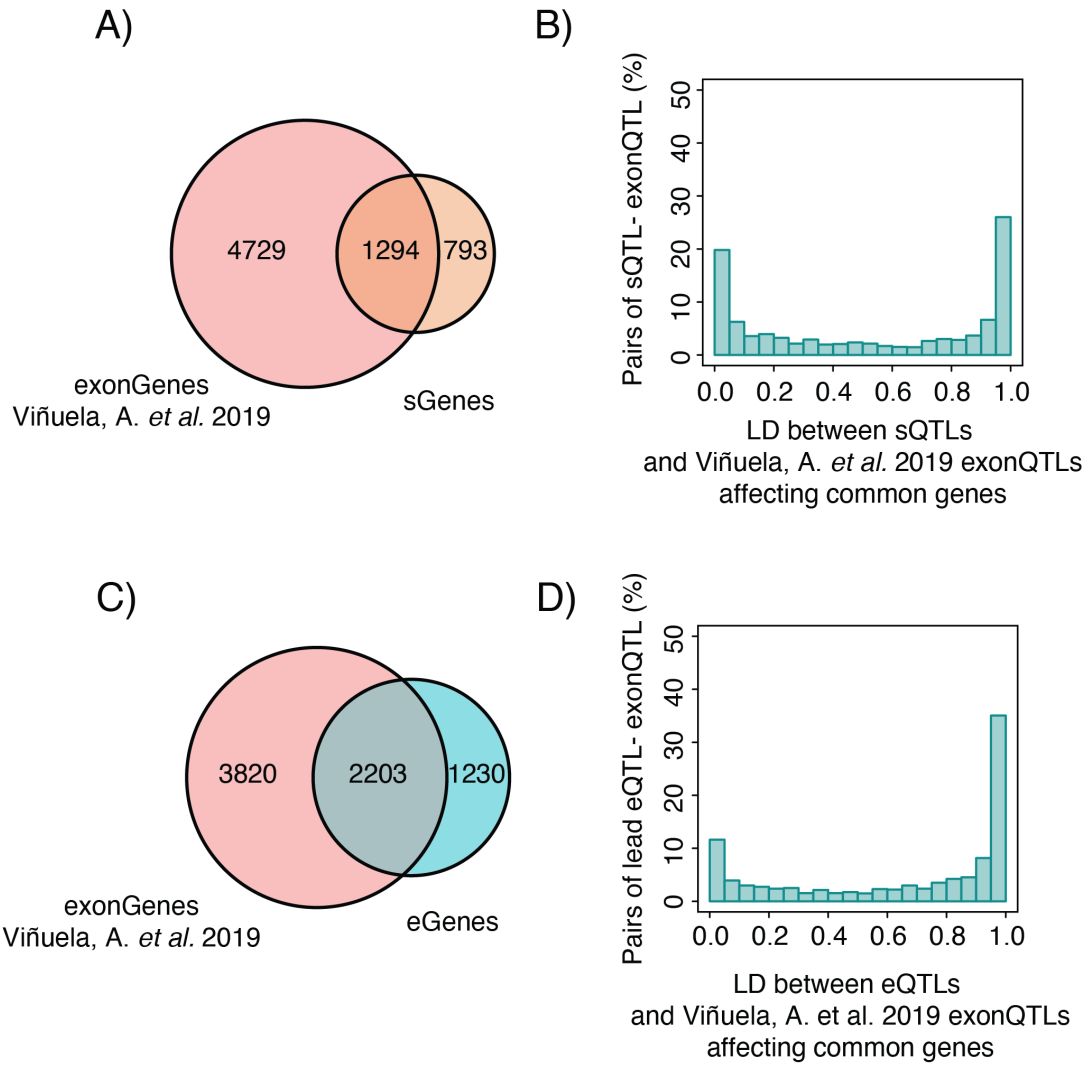

Figure 3. 25 Comparision of sQTLs and eQTLs with exon-QTLs.
A) Overlap between number of exonGenes and sGenes. B) Linkage disequilibrium (LD r2) between the lead exonQTL and lead sQTL for the overlapping genes. C) Overlap between number of exonGenes and eGenes. D) Linkage disequilibrium (LD r2) between the lead exonQTL and lead eQTL for the overlapping genes.

To assess the magnitude of genetic effects on isoform usage, we quantified the absolute difference between median junction usage of individuals with reference and alternate allele of lead sQTL. This showed that 25% of sGenes had >10% shift in transcript usage depending on the genotype (Figure 3.26A).

We further quantified the percent of transcripts containing an sQTL junction that resulted in either truncating variants or omitted the stop codon, which we refer to as nonsense mediated decay/non-stop decay (NMD/NSD) (Figure 3.26B). This showed

that ~23% of transcripts that contain an sQTL junctions led to NMD/NSD, while 22% preserved the ORF.

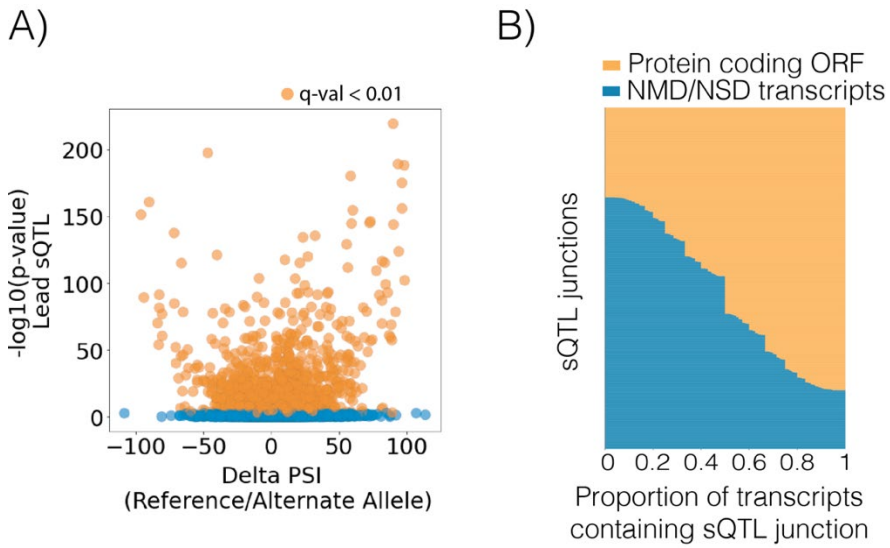

Figure 3. 26 Magnitude of genetic effects on splicing.
A) Delta-PSI of sQTL junction w.r.t reference and alternate allele is shown x-axis with -log10(p-vaues) on y-axis. B) Proportion of transcripts containing sQTL junctions that undergo either non-sense mediated decay/non-stop decay (NMD/NSD) or leads to a protein coding ORFs.

We also assessed if sQTLs preferentially affect distinct gene programs. We found that sGenes are enriched in annotations that belong to co-expression networks that control islet-cell function (Figure 3.27),

ARCHS4 Tissues	Adjusted P-value
PANCREATIC ISLET	6.38E-08

ARCHS TF co-expression modules	Adjusted P-value
GLIS3	2.70E-08
ZNF254	7.50E-06
LCORL	1.10E-05
NKX2-2	7.50E-05
PDX1	1.30E-04

Figure 3. 27 Functional enrichment of sGenes.

Enrichr was used to identify enrichment terms of various biological functions based on All RNA-seq and ChIP-seq sample and signature search (ARCHS4) web resource *(Lachmann et al., 2018)*. Only significant enrichments (p-adjusted <0.01) are shown.

In sum, our results disclosed widespread effects of common genetic variants on alternative splicing of human pancreatic islet transcripts, which cannot be directly measured by genetic effects on exon level variation. Our sQTL data also provided further evidence about the limitations of reference transcriptome annotations in identifying the tissue-specific component of the transcriptome. This hereby underscores the importance of generating transcriptome annotations in disease-relevant tissues to gain insights into disease biology. Finally, we also observed that genes whose splicing is genetically controlled tend to be key players of islet-cell identity and function.

4.3.2. sQTLs and eQTLs reveal distinct forms of transcriptome variation

Although sQTLs and eQTLs are intended to measure different types of events, it is unclear to what extent splicing variation affects steady state mRNA levels, or transcriptional mechanisms affect splicing. We thus examined the degree of genetic sharing between gene expression and splicing regulation. We found that only 34% of sGenes (715 genes) also harbor a significant eQTL. (Figure 3.28A). We further observed that for those 715 common genes, the lead eQTL and sQTL frequently showed low linkage disequilibrium (r^2 <0.6 for 56.5% of genes, < 0.1 for 23.6% of genes (Figure 3.28B). This suggests that, in most genes that harbored both eQTL and sQTLs, these were driven by different variants that drive distinct processes. This is illustrated by the *RGS1* gene, which has an intronic sQTL variant that impacts exon inclusion but has no effect on mRNA levels, and a distal eQTL which impacts total gene expression but not exon inclusion (Figure 3.28C).

In keeping with these findings, eQTLs and sQTLs were enriched in different functional genomic annotations. sQTLs were predominantly enriched in 5', 3' splice sites and exons, whereas eQTLs showed higher enrichment in active promoters and enhancers (Figure 3.29A). This is consistent with the notion that sQTLs impact *cis*-regulatory elements that govern splicing, while eQTLs impact transcriptional cis-regulatory elements, thereby showing that both represent fundamentally independent mechanisms of genetic variation.

We further examined the extent to which genetic effects on splicing or expression differed across tissues. We observed that ~60% of lead sQTLs were found in less than 5 GTEx (Consortium, 2020) tissues, compared to ~30% of lead eQTLs, suggesting greater islet-specificity of sQTLs (Figure 3.29B).

Taken together, our results reveal two separable layers of genetic influences on the human islet transcriptome that is reflected in terms of the underlying consequences in the functional non-coding genome and their distinct tissue-specific nature.

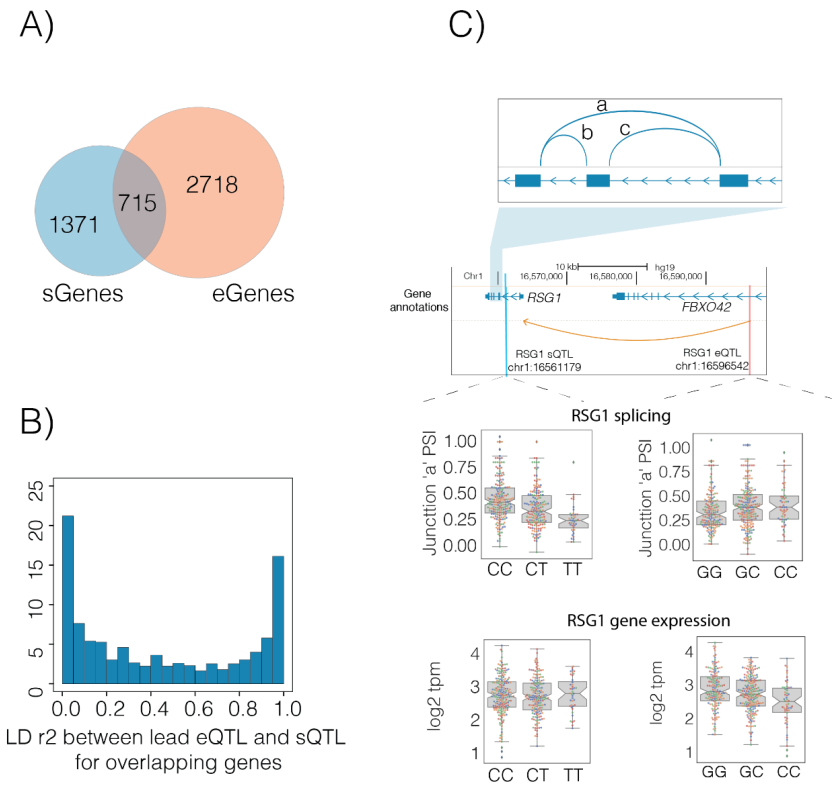

Figure 3. 29 Distinct genetic effects on gene expression and alternative splicing.

A) Overlap of sGenes and eGenes. B) Distribution of LD (measured as r2) between lead sQTL and lead eQTL for the 715 common genes from panel-A. C) An example illustrating distinct genetic effects on gene expression and alternative splicing. RSG1 lead eQTL is distal and located in intron of FBXO42 gene which does not have any association with the splicing of RSG1 gene. The intronic lead sQTL of RSG1 gene does not have any association with RSG1 gene expression.

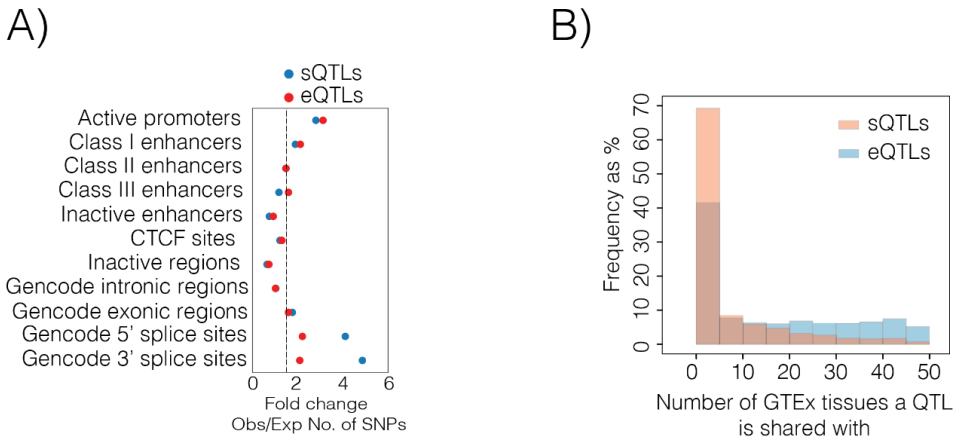

Figure 3. 28 Degree of genetic sharing of eQTLs and sQTLs across tissues.

A) Enrichment of sQTLs and eQTLs in different functional genomic regions. eQTLs are more enriched in active promoters and class-I enhancers where as sQTLs are more enriched in 5' and 3' splice sites and exonic regions. B) Number of GTEx tissues that an eQTL or sQTL is sharee with.

4.3.3. Identification of candidate causal variants

One of the main limitations that frustrates the conversion of GWAS results into molecular insights is that extensive local linkage disequilibrium hampers the identification of true causal variants (Altshuler et al., 2008; Schaid et al., 2018). Genetic fine mapping approaches aid in narrowing down the most likely candidate causal variants in each locus. We thus derived 95% credible sets of putative causal *cis* eQTL and sQTLs using CAVIAR (Hormozdiari et al., 2016) and reasoned that causal posterior probabilities (CPP) estimated for each marker should inform about their likelihood for being causal. To this end, we compared the CPP distribution of e- and sQTL credible set variants across different functional annotations. This showed that amongst 95% credible set sQTL variants, those located in 5'- and 3' splice sites showed higher CPP compared to intergenic variants (Mann-Whitney p-values = $1.36x10^{-19}$, $1.13x10^{-21}$, respectively) (Figure 3.30A). We also observed that exonic and intronic credible set variants showed higher CPP compared to intergenic variants (Mann-Whitney p-value = 4.55^{-80} and 0.001, respectively). This genetic analysis fulfilled functional expectations, because variants that influence splicing are known to alter sequence motifs of splice acceptor/donor sites, as well as exonic/intronic splicing enhancers and silencers. It therefore validated the ability of credible sets to prioritize causal sQTL variants.

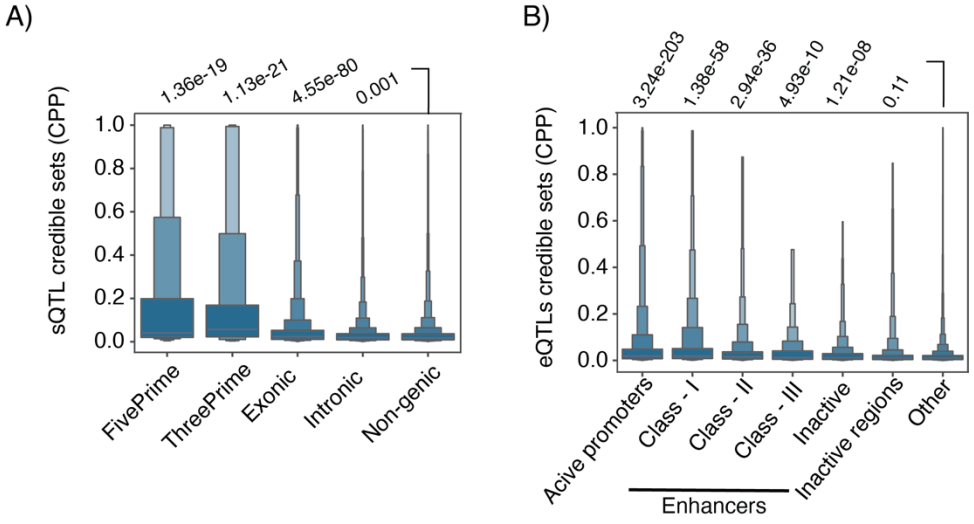

Figure 3. 30 Distribution of DIS of sQTLs and eQTLs.
A) Distribution of CPP of sQTLs in genic annotations. sQTL credible sets show higher CPP in 5' and 3' splice sites and exonic regions. B) eQTLs show higher CPP in active promoters and active enhancers (primarily class-I enhancers) compared to eQTLs in non-open-chromatin regions.

In parallel, we assessed the distribution of CPP probabilities for eQTL variants across islet regulatory annotations. We observed higher CPPs for fine-mapped eQTLs that overlap human islet active promoters and enhancers (mainly in class I enhancers) (Mann-Whitney p-values $3.24x10^{-203}$, $38x10^{-58}$, respectively) (Figure 3.30B). This is consistent with the notion that promoter and distal regulatory elements play a key role in gene transcriptional regulation.

To further characterize *bona fide* causal regulatory variants, we calculated the disease impact scores (DIS) for all credible set variants using DeepSEA models (Zhou et al., 2019). DeepSEA trains deep learning models on chromatin data (a model known as DeepSea) and RNA-binding proteins data (model known as Seqweaver), to predict underlying regulatory sequence preferences, and to subsequently perform in-silico mutagenesis. Then, DeepSEA combines this information with human gene mutation database to derive disease impact scores (DIS) (Zhou et al., 2019). We reasoned that even though such models used datasets from non-islet tissues, they could provide additional insights into causal functional variants.

Consistent with functional expectations, sQTL credible set markers in 5'- and 3' splice sites had highest disease impact scores, followed by exonic variants (Figure 3.A).

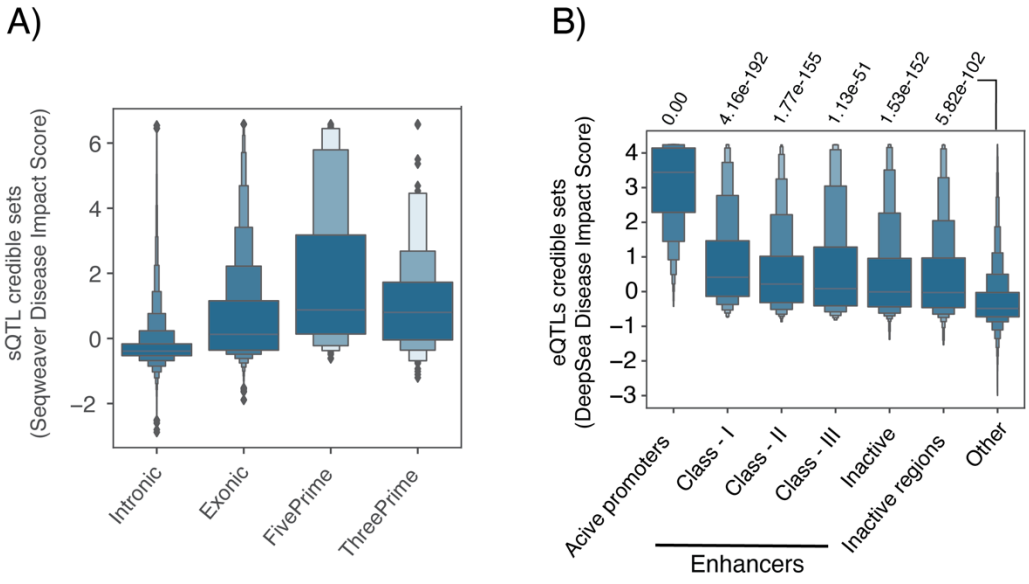

Figure 3. 31 Distribution of DIS of sQTLs and eQTLs.
A) Distribution of DIS of sQTLs in genic annotations. sQTL credible sets show higher DIS in 5' and 3' splice sites and exonic regions. B) eQTLs show higher DIS in active promoters and active enhancers (primarily class-I enhancers) compared to eQTLs in non-open-chromatin regions.

Furthermore, we looked at the impact of credible set variants on individual TFs and RBPs. For each sQTL credible set variant (CPP > 0.01), we identified the RBP whose sequence has a maximum impact based on in-silico mutagenesis. Then, top 10 most frequently disrupted RBPs in each category were chosen. The same analysis was done for eQTL credible sets (CPP >0.01) with individual TFs. This analysis showed that 3'-splice site credible set sQTLs showed recurrent disruptions of core-splicing components such as *U2AF1, U2AF2* and branch point motifs (Wahl et al., 2009) , whereas credible set sQTLs in introns and exons disrupted motifs of auxiliary regulators of splicing such as SR and SR-related proteins (*SRSF3, SRSF9*) (Fu and Ares, 2014), heterogenous ribonucleoprotein proteins (*HNRNPA1,HNRNPK,HNRNPC),* and polypyrimidine tract binding proteins (*PTBP2)* (Llorian et al., 2010; Xue et al., 2009) (Figure 3.32A).

On the other hand, eQTL credible set variants in active promoters and class I enhancers showed higher disease impact scores compared to variants in non-open chromatin regions of human pancreatic islets (Figure 3.31B). Expectedly, promoter eQTLs were recurrently disrupting motifs of promoter associated chromatin regulators such as *CHD1* and *RBBP5* (Murawska and Brehm, 2014; Narlikar et al., 2013), while other credible set eQTLs disrupted sequence motifs of CTCF and cohesion complex components (Bailey et al., 2015; Ong and Corces, 2014) , as well as FOXA and GATA TF motifs that are known to be involved in pancreatic development and pathophysiology of diabetes(Greenwald et al., 2018; Pasquali et al., 2014; Shaw-Smith et al., 2014; Viger et al., 2008) (Figure 3.32B).

Taken together, our credible set analysis, coupled with deep learning analysis of disease impact scores and functional annotations, provides a collection of *bona fide* candidate causal variants that are likely to drive splicing and expression variation in human pancreatic islets.

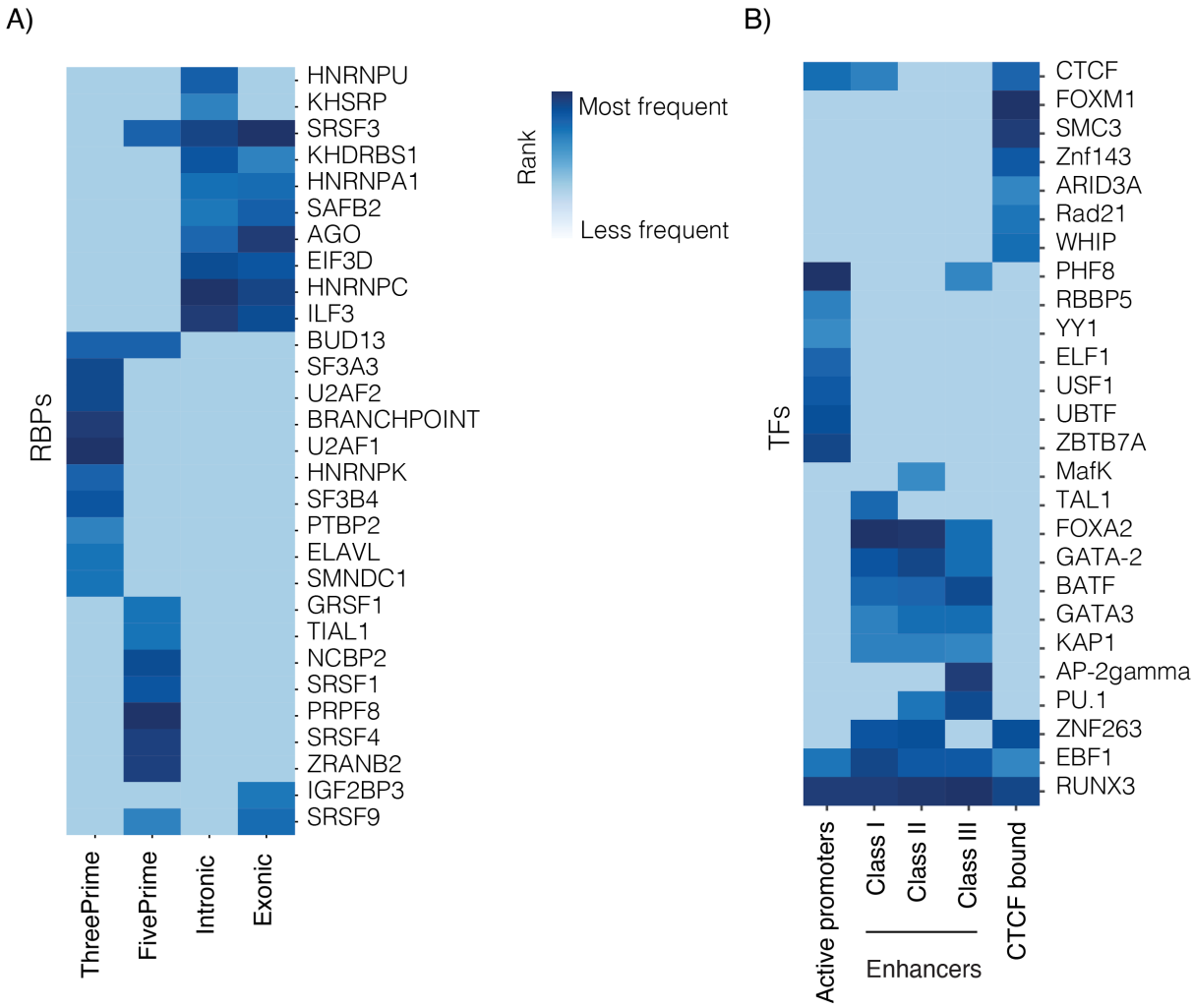

Figure 3. 32 Frequently disrupted RNA-binding proteins and transcription factor sequences by sQTLs and eQTLs.

This analysis is performed after dividing each credible set variant into different categories based on their genomic location. A) Frequently disrupted RNA-binding protein sequences by sQTLs credible set variants as predicted by Seqweaver B) Frequently disrupted transcription factor sequences by eQTL credible set variants as predicted by DeepSea model.

4.4. Interpretation of T2D GWAS signals through TWAS and Colocalization

sQTLs are enriched among T2D risk variants

Genetic susceptibility for T2D has been consistently linked to variants that influence transcription in human islets, based on enrichments of T2D risk variants in islet-specific regulatory annotations(Miguel-Escalada et al., 2019; Parker et al., 2013; Pasquali et al., 2014; Thurner et al., 2018; Varshney et al., 2017b) and human islet eQTL studies(Bunt et al., 2015; Fadista et al., 2014; Viñuela et al., 2020). However, the relationship between T2D susceptibility and islet splicing is poorly understood.

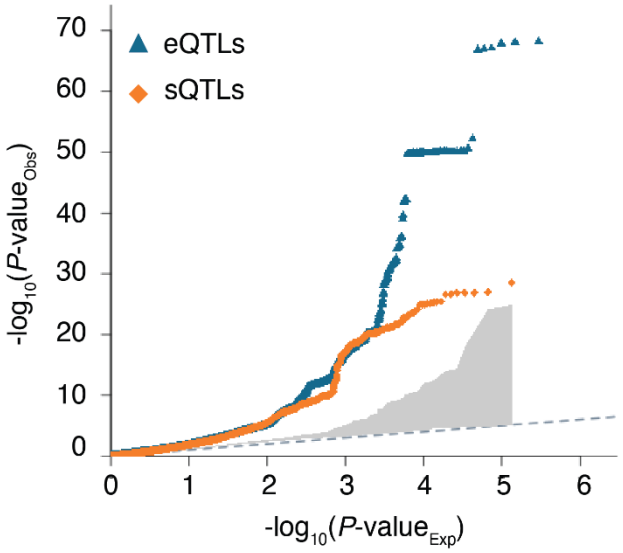

Figure 3. 33 Quantile-Quantile plot (QQ plot) for T2D risk across eQTLs and sQTLs. Expected $-\log_{10}$ p-values under the null hypothesis are represented in the x axis, while observed $-\log_{10}$ p-values are represented in the y axis.

We first examined the enrichment of T2D susceptibility variants in islet eQTLs and sQTLs. To this end, we used quantile-quantile plots that compare the distribution of T2D association p-values from one of the largest BMI-adjusted meta-analysis for T2D(Mahajan et al., 2018b) against the expected null distribution (Figure 3.30). We identified a strong inflation of lower T2D association p-values for eQTLs, consistent with the expected enrichment of islet transcriptional regulatory variants. Remarkably, our results also showed T2D risk inflation for sQTLs (Figure 3.33). This provides for

the first time an indication that a subset of noncoding variants could contribute to T2D genetic susceptibility through their effects on RNA splicing in human islets.

Transcriptome-Wide Association Study reveals novel T2D risk loci

To gain further insights into the genetic mechanisms underlying T2D genetic signals, we integrated genetic effects on gene expression and splicing in human islets with T2D GWAS data using (1) Transcriptome-Wide Association Studies (TWAS) (Gusev et al., 2016), which as discussed in the introduction, imputes gene expression in GWAS data to identify genes whose splicing or expression in human islets is associated with T2D risk, and (2) traditional colocalization approaches that match QTL and GWAS genetic signals by estimating the probability that both association signals are due to the same causal variant (Giambartolomei et al., 2014; Hormozdiari et al., 2016; Pickrell et al., 2016).

We first sought to nominate candidate effector transcripts for T2D susceptibility variants through TWAS as implemented in FUSION(Gusev et al., 2016). We thus leveraged our expression and splicing datasets to impute gene expression (eTWAS) and splicing ratios (sTWAS) into T2D summary statistics. This identified 44 genes showing eTWAS associations with T2D, and 37 annotated genes (65 splicing events) with sTWAS associations with T2D at after multiple test correction (p-value significance at 1.75×10^{-5} and 8.61×10^{-6}, after correcting for 2,851 genes and 5,804 splicing junctions, respectively) (Figure 3.34). We observed that 40/44 with eTWAS and 32/37 genes sTWAS associations were in known T2D GWAS loci. Besides outlining genes of interest for known T2D risk variants, the second main potential of TWAS is discovering novel risk loci that do not reach stringent thresholds for statistical significance in GWAS yet show genetic transcriptome effects that enable the detection of significant TWAS associations. We found 4 genes that identified novel T2D risk regions via eTWAS; *CWF19L1, PCBD1, PXK* and *CTC-228N24.2* (Figure 3.34B). *PCBD1* is of obvious interest because recessives mutations have been reported to cause early-onset diabetes and it is the co-factor for *HNF1A*, a gene that is mutated in monogenic diabetes and carries variants associated with polygenic T2D (Bonnefond and Froguel, 2015; Simaite et al., 2014). An additional five novel T2D risk loci were found through sTWAS in *ERO1LB, SCAMP3, NHSL1, FAM57A* and *ZNF277* genes (Figure 3.34A). As expected, all of them accounted for suggestive p-values in GWAS (best GWAS lead variant p-value < 5×10^{-5}).

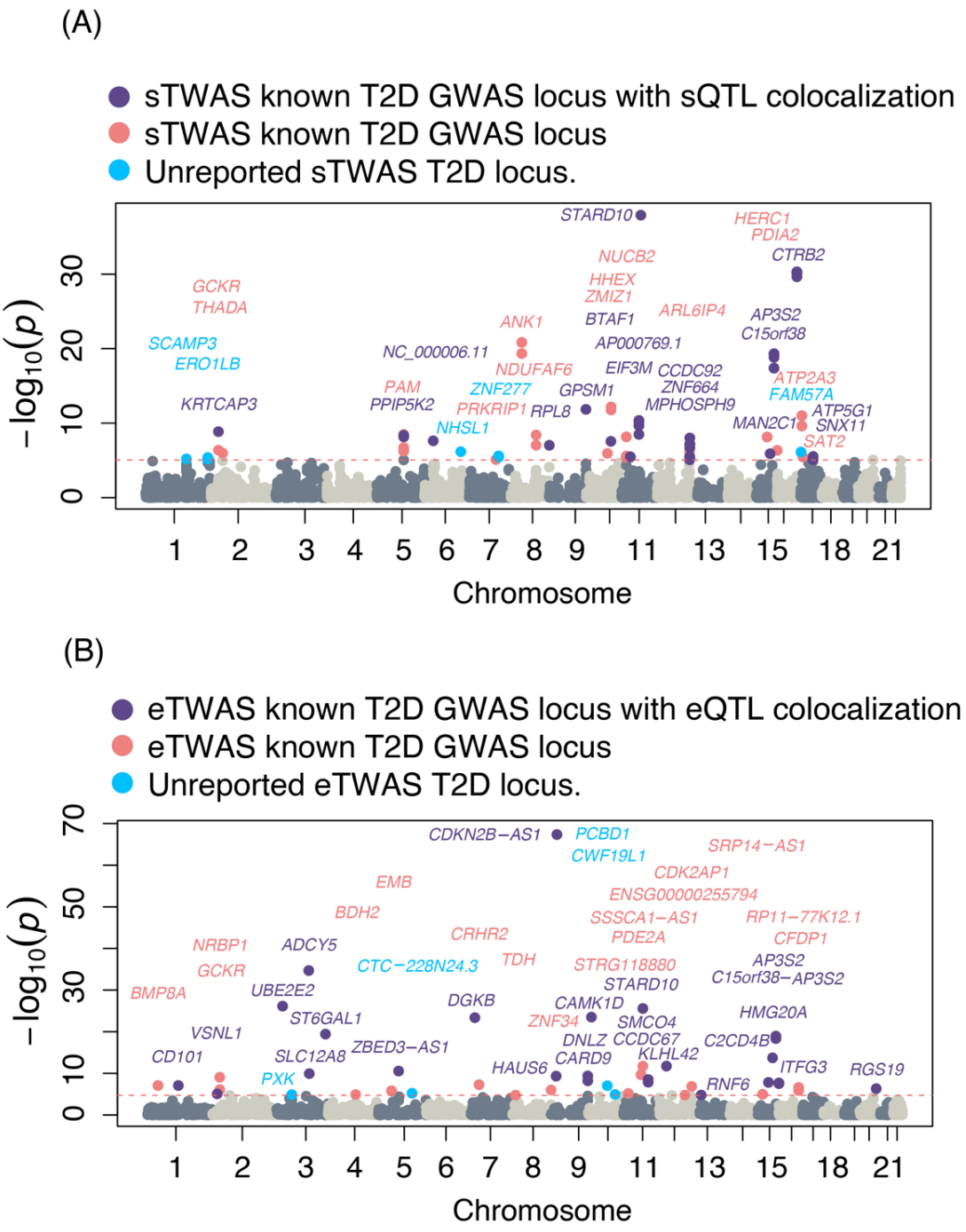

Figure 3. 34 Manhattan plots of islet gene expression (eTWAS) and splicing (sTWAS) associations for T2D risk.

Y-axis represents the −log10 p-values that were colored in dark and light grey for non-significant associations in alternate chromosomes, respectively. Significant eTWAS or sTWAS associations in known T2D GWAS loci are depicted in red. Those that also attained strong support from colocalization (any of the approaches implemented, see Methods). Were colored in purple. Those associations in loci that have not been reported by Mahajan, A. et al (2019), were depicted as blue dots. (A) shows sTWAS results, and (B) presents eTWAS results.

As TWAS suffers from limitations to distinguish genetic sharing from both traits from linkage (Hemani et al., 2018; Ndungu et al., 2020; Wainberg et al., 2019), we carried out colocalization analysis of TWAS variants with T2D risk variants that show substantial statistical association evidence (GWAS p-value < $1x10^{-5}$). Colocalization probabilities (PP4 ≥ 0.8) that indicate a strong support for shared signal between T2D risk and gene expression variation were found at 24 out of 44 genes, and between T2D susceptibility and variation in splicing for 33 out of 65 splicing events in 37 genes. In addition, as shown in Figure 3.35, we observed an overall trend of high PP4 values supporting sTWAS and eTWAS associations. This is in sharp contrast with the depletion of high PP3 values that suggest linkage as the underlying cause of the association signal between T2D susceptibility and our islet expression and splicing datasets. In line with this, among the fraction of novel T2D risk loci identified either by eTWAS or sTWAS, only three genes did not show strong colocalization (PP4 ≥ 0.8). Coupled with the fact that the majority of our novel associations are within robust T2D GWAS significant loci, our results suggests that the novel T2D risk associations identified here have a potential role in the pathophysiology of T2D.

Independent colocalization analysis

We performed additional colocalization analysis between each of the 403 independent GWAS signals (Mahajan et al., 2018b) and our significant islet eQTL and sQTL maps. We applied colocalization as implemented in *gwas-pw* (Pickrell et al., 2016), that draws upon the original coloc algorithm but does not rely on user-defined priors. We identified candidate effector transcripts with robust colocalization evidence (posterior probability of shared association between both phenotypes ≥ 0.9) for 25 and 9 independent T2D GWAS signals using eQTLs and sQTLs, respectively. We further compared TWAS and *gwas-pw* results using eQTL and sQTL data. This comparison showed that both approaches converged in at least one common candidate effector transcript (Figure 3.31) except for two T2D independent GWAS signals. We then examined additional candidate effector transcript genes identified by *gwas-pw* alone. *gwas-pw* identified eight candidate effector transcripts that were not detected in eTWAS: *PTGFRN*, *B3GALNT2*, *CEP68*, *IGF2BP2*, *HI-LNC77, RPL8, PLEKHA1* and RP11-282018.3. Colocalization between T2D risk variants and sQTLs identified two additional candidate genes that were not captured by sTWAS: *KIF9* and *CTBP1*. Taken together,

the implementation of both TWAS and colocalization approaches offered a comprehensive catalog of target effector genes whose expression or splicing in human pancreatic islets is linked to genetic effects influencing T2D predisposition.

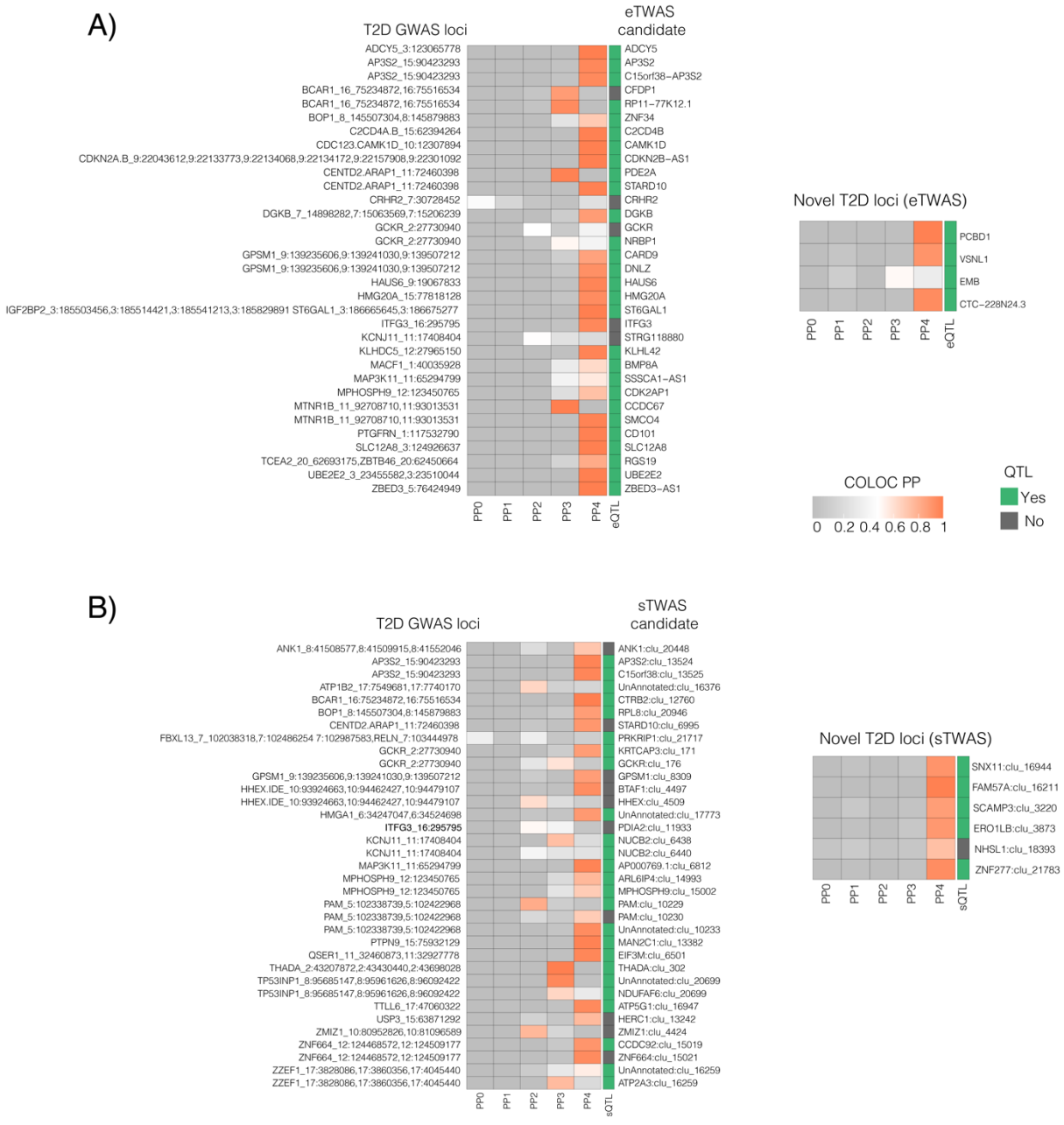

Figure 3. 35 A heatmap representing the colocalization posterior probabilities for TWAS associations.
A) Heatmap showing the colocalization posterior probabilities for eTWAS analysis. B) Heatmap showing the colocalization posterior probabilities for sTWAS analysis. Separate heatmaps are shown for known T2D loci and novel T2D loci identified by TWAS. Heatmap color bar indicates colocalization posterior probability. Green vertical bar indicates if the TWAS prioritized candidate gene is also a e/sQTL in human islets.

Integrative analysis of islet QTLs to highlight effector transcripts of T2D risk

We next quantified the actual gain in novel molecular insights into T2D susceptibility that our expression and splicing datasets provided. To this end, we compared our complete collection of candidate genes identified by TWAS and colocalization approaches, with the set of candidate effector transcripts nominated in the largest human islet eQTL study (Viñuela et al., 2020). The integration of islet QTLs from both studies pointed to candidate effector transcripts for 100 T2D independent GWAS signals. For 27 of these independent T2D GWAS signals, candidate effector genes were nominated by both our transcriptome studies, as well previous studies. In total, 42 candidate effectors were identified for these 27 signals, 22 of which were genes that showed eQTLs/eTWAS in both studies, namely *CEP68*, *UBE2E2*, *ADCY5*, *SLC12A8*, *IGF2BP2*, *DGKB*, *HAUS6*, *GPSM1*, *CARD9*, *DNLZ*, *CAMK1D, PLEKHA1*, *STARD10*, *PDE2A*, *CCDC67*, *KLHL42*, *CCDC92*, *RNF6*, *HMG20A*, *C15orf38-AP3S2*, *AP3S2* and *ITFG3*. Importantly, the current study pointed to splicing or expression variation in candidate effector genes for additional 45 of the 100 T2D independent signals (Figure 3.36).

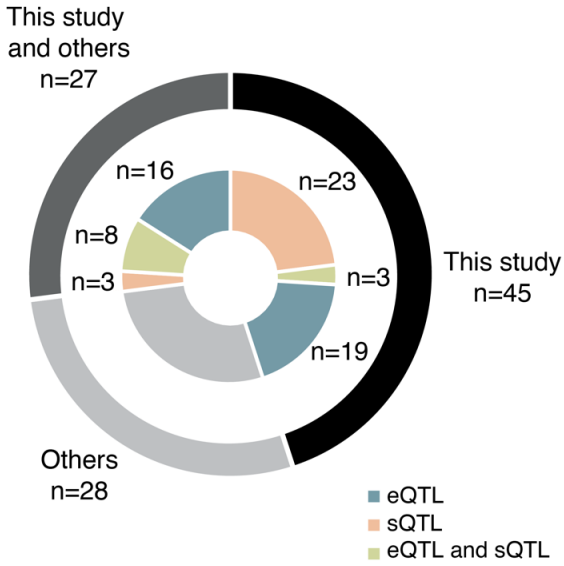

Figure 3. 36 Candidate effector transcript genes for 100 T2D risk loci assigned by the current study and/or previous studies.

The outer circle depicts T2D independent signals with a candidate effector transcript nominated solely by the current study (black), this study and eQTL maps from Viñuela, A. et al 2020 (dark grey), or only by the latter study (light grey). The inner circle breaks down each of the previous fraction that includes target effector transcripts nominated in this study according to the underlying molecular mechanism: islet expression (blue), islet alternative splicing (orange) and both (green).

Overall, we identified candidate effector transcripts for 72 independent T2D signals, increasing the number of T2D loci with candidate effector transcripts by 1.8-fold (Figure 3.33). Importantly, this increase in T2D known loci with assigned target genes was also notably driven by the integration of allelic variation with effects on islet splicing

Credible set analysis helps prioritize causal T2D risk variants

Given that the human pancreatic islet QTL credible set analysis (Section 3.3.3) allowed prioritization of candidate causal variants, we rationalized that QTL credible sets should also help us prioritize T2D risk causal variants. Therefore, first we sought for convergence of GWAS credible set and sQTL/eQTL credible set posterior probabilities (CPP). To investigate this in an unbiased manner, for each of the 403 GWAS signals we selected 99% credible set variants along with variants that are in moderate LD (r2 > 0.1) with lead GWAS variant (Mahajan et al., 2018a). We then annotated each variant into three mutually exclusive categories: (i) GWAS credible set only; (ii) GWAS credible set in LD (r2>0.1) with a lead QTL, but not in a QTL credible set; (iii) GWAS credible set that is also in a QTL credible set. We then plotted the distribution of GWAS CPP for each of the three categories of variants. This analysis showed that the GWAS credible set variants that are also in sQTL credible sets show a higher GWAS CPP compared to GWAS credible set variants that showed no overlap at all, or were only in LD with a lead sQTL (Mann-Whitney p = 3.30×10^{-64} and 3.15×10^{-11}, respectively) (Figure 3.37A). Likewise, we observed that GWAS credible set variants that are also eQTL credible sets show higher GWAS CPP compared to GWAS credible set variants that showed no overlap at all, or were only in LD with a lead eQTL (Mann-Whitney p = 1.19×10^{-143} and 4.3×10^{-5}, respectively) (Figure 3.37B). We also assessed the distribution of deep learning-based Disease Impact Scores in the same categories and found marginally increased, but non-significant, Seqweaver and DeepSea scores in GWAS credible set variants that overlapped with sQTL or eQTL credible sets (p = 0.1, and p=0.06, respectively) (Figure 3.37C, D).

Overall, this integrative analysis provided further independent support for a role of splicing and expression QTLs in T2D susceptibility variants and suggests that candidate causal variants that impact gene regulation in human pancreatic islets can guide us to prioritize T2D risk causal variants.

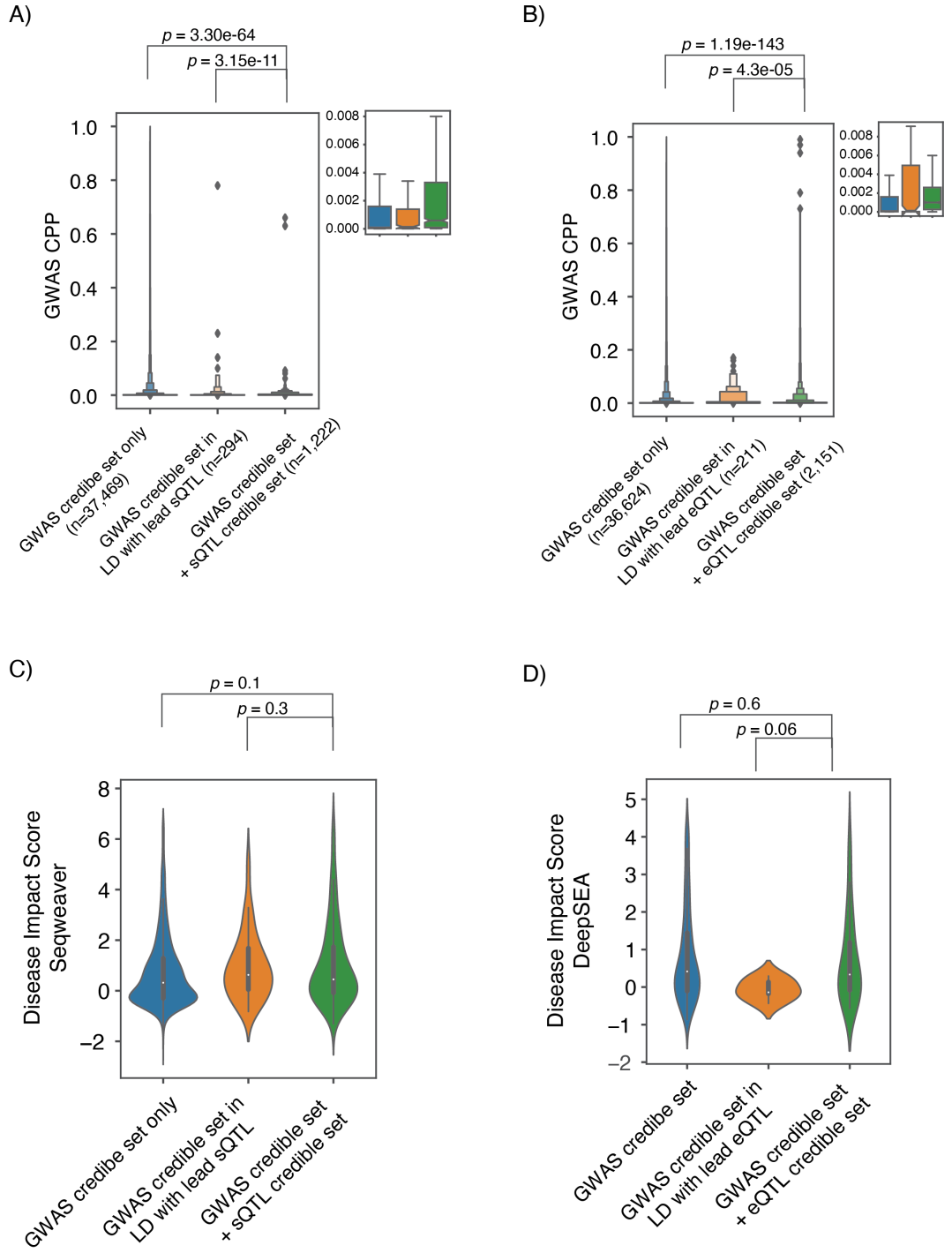

Figure 3. 37 QTL credible sets prioritize T2D risk candidate causal variants.

A) Distribution of GWAS causal posterior probabilities of the variants that are either in LD (r2>0.1) with a GWAS index variant or in credible set of GWAS. The GWAS CPP is plotted based on the relation with sQTL credible set variants. B) Similar plot but in relation with the eQTL credible set. Interquartile range is shown in the small box plots C) Seaweaver disease impact scores for the exonic variants according to the same category as in A. D) DeepSEA disease impact score for variants in class-I enhancers according the same categories as in B. Mann-Whitney p-values are shown.

Discussion

Refining the regulatory landscape of human pancreatic islets.

Genome-wide maps of open-chromatin regions in the human genome and their in-depth characterization of their distinct regulatory potential into regulome annotations is fundamental to advance our understanding of tissue-specific gene regulation. Human pancreatic islet open chromatin regions were previously defined based on a combination of FAIRE-Seq and H2ZA data that resulted in low-resolution maps of chromatin accessibility(Gaulton et al., 2010). Moreover, previous annotations of regulatory elements (Pasquali et al., 2014) did also not profile key transcriptional regulators such as Mediator and Cohesin.

Considering these major limitations, we profiled open-chromatin regions in human pancreatic islets using high resolution ATAC-Seq data from several individuals. We then annotated open-chromatin regions by epigenome datasets of histone modifications, Mediator, CTCF and Cohesin. This not only allowed us to refine human islet regulatory annotations but also to sub-classify active enhancers (Class I-III) based on high Mediator presence, revealing a finer granularity in enhancer function.

Our novel high-resolution regulome annotations are central to unearth the molecular mechanisms underlying T2D pathophysiology with a polygenic risk model based on genetic variants overlaying islet enhancer annotations that identifies individuals at high genetic risk due to beta-cell dysfunction.(Miguel-Escalada et al., 2019) Remarkably, these novel regulome annotations also have relevant implications to expand our understanding of the genetic causes of monogenic diabetes forms. The majority of individuals clinically diagnosed with rare monogenic DM forms do not present causal mutations in known protein-coding genes (Hattersley and Patel, 2017; Shields et al., 2010). Recent work showed that recessive islet enhancer mutations are the most common cause of isolated pancreatic agenesis (Consortium et al., 2014b) suggesting that the role of islet enhancers spans beyond T2D-heritability to rare and severed diabetes forms. Thereafter, we expect that our islet regulome annotations could be leveraged to gain insights into the genetic basis of monogenic diabetes through the non-coding genome. From a gene regulatory point of view, this study also deepens into enhancer hierarchy by the identification of class-I enhancer, which are highly enriched with Mediator signal and are key players in our novel definitions of islet specific three-dimensional regulatory units that control islet-specific gene expression programs

Nevertheless, we should note that temporal regulation elicited by transcriptional enhancers during development has not been accurately captured in this study. Our open-chromatin regions and regulatory maps are primarily defined based on data from adult donor samples. Thereafter, regulatory elements that are distinctly active at particular developmental stage could be missed. Although the dearth in available islet human tissues at early developmental stages, our on-going efforts to improve our regulatory annotations using juvenile samples will enlighten the role of islet regulation during pancreatic islet differentiation.

Human pancreatic islets are composed of various cell-types, predominately beta and alpha cells. Our regulatory maps primarily reflect the bulk of the human islet tissue, and thus, we did not reach enough resolution to characterize the cell-type specific nature of regulatory elements. Preliminary analysis based on our scATAC data from human islets has shown that class-I enhancers are more specific to beta cells while inactive enhancer regions are more specific to rare islet-cell populations, such as delta cells. Thus, further analysis is required to dissect the cell-type specific component of regulatory annotations to have a more precise understanding of gene regulation in human islets in order to gain insights into diabetes pathophysiology.

Glucose dependent genome regulation in human pancreatic islets.

Blood glucose is a primary stimulant for beta cells to secret insulin. In mouse and rodent islet studies (Alonso et al., 2007; Levitt et al., 2010; Porat et al., 2011), prolonged exposure to high glucose has shown to induce beta cell proliferation to meet the increasing demands of insulin secretion. Such studies are lacking in human model systems. To gain molecular mechanisms into adaptive response of human pancreatic islets to prolonged glucose concentrations, we cultured human pancreatic islets in 11mmMol (high) and 4mMol (low) glucose concentrations for 72 hours, where 11mMol glucose mimics prolonged glucose levels that occur in physiological conditions. We then assayed gene expression (RNA-Seq) and chromatin activity (H3K27ac).

By analyzing gene expression data, we observed that human pancreatic islets undergo major transcriptional changes. High glucose concentrations induce beta cell differentiation genes while repress genes involved in apoptotic pathways. This is contrary to glucotoxic models, that do not echo physiological glucose variation, and thus lead to apoptosis (Poitout and Robertson, 2008; Poitout et al., 2010).

By analyzing chromatin activity, we observed that human islet enhancers, in particular class-I are predominately induced by a high glucose condition. Remarkably, glucose-induced enhancers account for cognate changes in gene expression for their target genes. We further show that, glucose effects on the islet regulatory landscape are not confined to individual enhancer-gene pairs but elicit a domain wide change as observed in our enhancer hub definitions. This suggests that glucose may rewire enhancer-promoter interactions at a broad domain level. However, we have not profiled chromatin interactions in low glucose samples, which limits our insights into the effect of glucose on re-arranging regulatory domain interactions.

This analysis paved a way to identify transcriptional programs that are involved in the adaptive response of human pancreatic islets to the glucose stimuli, and a detailed study on higher sample sizes is required extend the primary insights achieved in this PhD thesis.

Accurate transcriptome annotations of human pancreatic islets

Reference transcriptome annotations such as GENCODE are mainly driven to annotate abundant protein-coding and lncRNA genes. Thus, they are not particularly powered to elucidate the mechanistic underpinnings of human disease pathophysiology that stem from tissue-specific and context-dependent gene expression programs(Akerman et al., 2017; Iyer et al., 2015; Morán et al., 2012; Nellore et al., 2016)

In this work, we leveraged billions of short reads and thousands of transcript models from long-read sequencing from human pancreatic islets to assemble transcript models of protein-coding and lncRNA genes. We further used CAGE data to accurately annotate TSS that also revealed dominant and alternative promoters of human islet transcripts. This analysis uncovered novel isoforms for known protein-coding genes and new transcripts without an assigned known gene. We re-analyzed several scRNA-Seq data sets to characterize the cell-type specific component of human islet genes. One of the major concerns (Tress et al., 2017) that arises from the vast number of isoforms identified is whether they encode for novel peptides or not. We undertook a systematic approach that allow us detecting novel coding sequences that may encode for tissue-specific peptides. We leveraged large catalogues of promoter expression data from FANTOM to identify islet selective and islet specific promoters.

Accurate transcript models are fundamental to understand tissue biology, as they aid in characterizing transcription regulatory programs and provide a biological

interpretation to guide genetic studies. For example, we demonstrated that the induction of NKX6-1 gene expression via CRISPRa on annotated TSS did not elicit gene expression levels. In sharp contrast, our dominant promoter definitions based on CAGE data showed a significant induction in gene expression upon CRIPSR activation. This is a noteworthy example of the importance of annotating tissue-specific transcript models that facilitate genomic and genetic perturbation studies. A recent study(Cummings et al., 2020) used transcript expression levels to infer the pathogenicity of coding mutations. We could envision that our human islet transcript models could be capitalized to provide a more comprehensive interpretation of coding mutations in rare monogenic forms of diabetes.

Even though we created transcriptome annotations based on long-read sequencing technologies, the majority of our transcript models may have been built from short reads. One of the limitations of long-read sequencing technologies is that they mainly capture abundant transcripts. Moreover, more than 60% of the human islet transcriptome is composed of mRNAs that are transcribed from *INS* and *GCG* genes. This makes it challenging for long-read sequencing technologies to detect even moderately abundant transcripts. Several strategies are now available to deplete selected mRNA molecules before sequencing (Gu et al., 2016), but the accuracy of such techniques needs to be evaluated. Thus, very deep sequencing of human islet transcriptomes in comparatively larger study sizes could further improve the accuracy of our transcript annotations.

The little progress in examining the potential of the vast number of transcript models arises from technological limitations in measuring protein abundance. Current methods are limited to the detection of abundant short peptides but not intact proteins. This hinders our ability to assess whether novel isoforms encode novel protein sequences or not. We are currently collaborating with Prof. Alan Attie and Prof. Lloyd Smith, who are experts in human islet proteomics and proteoform detection(Schaffer et al., 2019) to further investigate to what extent intact proteins from our novel coding sequences are functionally relevant.

Gene transcription is initiated at precise locations in the genome, the TSS, by the involvement of a sequence of 40bp upstream and downstream of TSS known as as the core promoter. Core promoters contains sequence determinants that facilitate transcription initiation by RNA-PolII machinery(Smale and Kadonaga, 2003). There is

growing evidence that suggests that promoter elements not only direct the RNA-Pol II machinery to the TSS, but they also receive *cis*-regulatory inputs and can determine the responsiveness of gene transcription (Engström et al., 2007; Zabidi et al., 2015). Thus, analyzing the sequence composition of tissue specific promoters gives clues about the determinants of transcriptional regulation. Our current analysis of sequence determinants of islet-specific promoters is restricted to only a subset of robust promoters that have an epigenome-based active promoter signature. An unbiased thorough analysis is still needed to uncover the sequence determinants of human islet specific promoters.

Genetic effects on human pancreatic islets transcriptome provides insights into Type 2 Diabetes genetic signals.

As we discussed in the Introduction, to unlock the real potential of GWAS discoveries to pave a way to personalized medicine, we need to push forward our understanding of the consequences of common genetic variation on the different components of human islet gene regulation and its implications in human disease. Multiple dedicated efforts have only focused on the impact of common genetic variation on transcriptional regulation in human pancreatic islets and the role in T2D predisposition. Although, it is now well established that genetic variation that alter alternative pre-mRNA splicing have significant contributions to disease risk in several human genetic diseases(Li et al., 2016; Raj et al., 2018b; Walker et al., 2019), allelic effects on human islet alternative splicing have not been profiled and its implications in T2D are not largely understood.

These limitations drove us to create maps of genetic effects on mRNA expression and alternative splicing in a panel of ~400 human pancreatic islet samples. This led us to identify wide-spread genetic effects on both islet mRNA expression (eQTLs) and alternative splicing (sQTLs). Our in-depth characterization of eQTL and sQTL variants revealed that they represent two separable layers of genetic control on human islet gene regulation, accounting for a distinct tissue-specific nature and contributing to gene regulation through recognizable different molecular mechanisms. We then assessed the contribution of islet alternative splicing to T2D susceptibility by the integration of sQTL and T2D GWAS data. This revealed that genetic effects on islet alternative splicing contributes to T2D heritability. Following these evidences, we leveraged TWAS and colocalization approaches to uncover molecular target genes

underlying T2D risk loci. Of note, we showed that the increase in the number of independent T2D signals with candidate effector transcripts that this study provided was notably driven by the inclusion of sQTL data. This highlights the importance of underappreciated non-coding genetics effects on alternative splicing in human pancreatic islets to gain insights into T2D pathophysiology. Of note, the implementation of TWAS approaches allowed us to expand our understanding of the genetic basis of T2D by identifying novel T2D genetic loci that yet have not been identified by GWAS approaches. Although these novel T2D loci should be carefully examined due to the inherent TWAS limitations to distinguish pleiotropy from linkage, we envisage that their implementation on other molecular traits will harvest additional new knowledge of themolecular mechanisms underlying T2D risk. Remarkably, we also observed that the fraction of T2D credible set variants that are also in QTL credible sets show a marked increase in posterior probabilities for T2D risk. Following this notion, we rationalized that defects on human islet gene regulation might have higher impact on T2D risk than genetic effects on the regulatory landscape of other disease-relevant tissues or environmental effects. This could be further assessed based on the credible set size; a marked decrease in the credible set size of GWAS loci that colocalize with islet QTLs vs the rest could suggest that genetic effects on islet function and identity are not only the major contributors to T2D heritability overall, but also have higher functional deleterious consequences. This has relevant implication on fine-mapping approaches and perturbation genetic screens.

We want now to discuss several limitations of the present study and how we foresee that could be addressed in on-going efforts arising from this project or in other future studies. First, our sQTL maps are based on measuring splicing activity using short-read sequencing, although attaining accurate measures of isoform variant using this of technology is challenging. Second, islet splicing activity was measured using LeafCutter that measures local splicing events based on the relative junction usage. We want to underscore that although this might inform us about splicing activity, it does not quantify the direct impact on isoform expression, as each junction could belong to multiple isoforms. This is one of the largest bottlenecks of our data, as it limits our understanding of the impact of alternative junction usage on a particular isoform, and the implications in T2D pathophysiology. A potential way to overcome this short-coming is by the quantification of the expression of each isoform using long read-sequencing technologies across hundreds of individuals, which is far from reality to

date. Third, the majority of our human islet samples are from adult donors, and thus do not allow us to uncover genetic effects that impact islet-cell identity and function during development. A recent study revealed that developmental specific genetic effects on splicing and gene expression are implicated in neurological and psychiatric diseases (Walker et al., 2019) . Collecting samples from juvenile individuals at a sufficient study size to robustly carry out QTL analysis is arduous, Thereby, interrogating the effect of genetic variants on gene expression and splicing using human induced pluripotent stem cell (iPSC) models from individuals with different genetic backgrounds remains an attractive alternative strategy (Zhang et al., 2020). Fourth, cell-type specific genetic effects have been shown to provide insights into complex trait disease genetics (Kim-Hellmuth et al., 2020). Several methods (Jew et al., 2020; Newman et al., 2015; Wang et al., 2019) are now available to deconvolute the cell-type composition of bulk heterogenous tissues that can be used as an interaction term to identify cell-type specific genetic effects. Our initial analysis following this approach did not reveal any novel insights, but we hypothesize that this could be due to the fact that our human islet data is composed of a more uniform distribution of predominant cell-types. Fifth, we did not interrogate the contribution of low-frequency variants into the regulation of islet expression and splicing. Thereafter, we could not comprehensively link human islet splicing to all the spectrum of risk alleles that contribute to T2D pathophysiology. Extreme changes in gene expression and alternative splicing triggered by rare genetic variants have already been observed across tissues(Ferraro et al., 2020; Li et al., 2017). We predict that the aggregation of human islet samples in comparatively larger datasets, as well as the replacement of SNP arrays by WGS, will allow identifying rare variants underlying extreme changes but also mild effects in islet expression and splicing, and assessing the relationship with T2D pathophysiology. Finally, so far, this study did not provide any experimental evidence to directly implicate splicing mis-regulation into T2D pathophysiology. We are currently carrying out two experiments. First to validate the impact of candidate cause sQTL variants on alternative splicing and to also measure the impact of splicing on beta cell function using glucose-responsive EndoC βH3 cell line (Benazra et al., 2015)

Conclusions

This thesis advanced our molecular understanding of the regulatory molecular mechanisms underlying cell-identity and function in in human pancreatic islets and its implications into T2D pathophysiology.

- High-resolution genome-wide maps of open-chromatin sites that integrate multidimensional epigenomic datasets provide a larger granularity of the functional regulome in human pancreatic islets, particularly of transcriptional enhancers.
- The integration of regulome annotations, capture Hi-C data (pcHi-C in this study) and T2D GWAS data from large-scale meta-analysis led to the identification of novel targets for T2D loci.
- Capitalizing a glucose perturbation model in human isles led to the detection of functional regulatory domains that underlie adaptive response of human pancreatic islets and thus, to maintain islet-cell homeostasis and function.
- Transcriptome annotation in human pancreatic islets revealed vast number of previously unannotated transcripts, promoters and coding sequences.
- Maps of genetic effects on islet mRNA splicing and gene expression (sQTLs and eQTLs, respectively) showed distinct layers of genetic control from non-coding common genetic variants.
- eQTLs and sQTLs identify distinct molecular mechanisms with recognizable different degree of genetic sharing across tissues, and impact independent machinery of the functional non-coding genome involving independent regulatory networks.
- The integration of fine mapped QTL data, *in-silico* functional scores and genome annotations provide a compendium of *bona fide* candidate causal variants that are likely to impact splicing and expression variation in islets.
- Leveraging large panel human islets transcriptome data with the joint effort of TWAS and colocalization approaches provided an exhaustive catalogue of candidate target genes that might enlighten our understanding of T2D pathophysiology.

- Fine-mapped eQTL and sQTL data integrated with T2D credible set variants prioritize *bona fide* candidate causal variants that are more likely to show functional consequences in genetic perturbation screens.

Bibliography

Aigha, I.I., and Abdelalim, E.M. (2020). NKX6.1 transcription factor: a crucial regulator of pancreatic β cell development, identity, and proliferation. Stem Cell Res Ther *11*, 459.

Akerman, I., Tu, Z., Beucher, A., Rolando, D.M.Y., Sauty-Colace, C., Benazra, M., Nakic, N., Yang, J., Wang, H., Pasquali, L., et al. (2017). Human Pancreatic β Cell lncRNAs Control Cell-Specific Regulatory Networks. Cell Metab *25*, 400–411.

Akhtar, M.S. and A. Dosage Compensation of the X Chromosome: A Complex Epigenetic Assignment Involving Chromatin Regulators and Long Noncoding RNAs.

Alonso, L.C., Yokoe, T., Zhang, P., Scott, D.K., Kim, S.K., O'Donnell, C.P., and Garcia-Ocana, A. (2007). Glucose Infusion in Mice: A New Model to Induce -Cell Replication. Diabetes *56*, 1792–1801.

Altshuler, D., Daly, M.J., and Lander, E.S. (2008). Genetic Mapping in Human Disease. Science *322*, 881–888.

Andersson, R., and Sandelin, A. (2020). Determinants of enhancer and promoter activities of regulatory elements. Nat Rev Genet *21*, 1–17.

Arvanitis, M., Tampakakis, E., Zhang, Y., Wang, W., Auton, A., Agee, M., Aslibekyan, S., Bell, R.K., Bryc, K., Clark, S.K., et al. (2020). Genome-wide association and multi-omic analyses reveal ACTN2 as a gene linked to heart failure. Nat Commun *11*, 1122.

Ashcroft, F.M., and Rorsman, P. (2012). Diabetes Mellitus and the β Cell: The Last Ten Years. Cell *148*, 1160–1171.

Auton, A., Abecasis, G.R., Altshuler, D.M., Durbin, R.M., Abecasis, G.R., Bentley, D.R., Chakravarti, A., Clark, A.G., Donnelly, P., Eichler, E.E., et al. (2015). A global reference for human genetic variation. Nature *526*, 68–74.

Bailey, S.D., Zhang, X., Desai, K., Aid, M., Corradin, O., Cowper-Sal·lari, R., Akhtar-Zaidi, B., Scacheri, P.C., Haibe-Kains, B., and Lupien, M. (2015). ZNF143 provides sequence specificity to secure chromatin interactions at gene promoters. Nat Commun *6*, 6186.

Balboa, D., Weltner, J., Eurola, S., Trokovic, R., Wartiovaara, K., and Otonkoski, T. (2015). Conditionally Stabilized dCas9 Activator for Controlling Gene Expression in Human Cell Reprogramming and Differentiation. Stem Cell Rep *5*, 448–459.

Banerji, J., Rusconi, S., and Schaffner, W. (1981). Expression of a β-globin gene is enhanced by remote SV40 DNA sequences. Cell *27*, 299–308.

Banerji, J., Olson, L., and Schaffner, W. (1983). A lymphocyte-specific cellular enhancer is located downstream of the joining region in immunoglobulin heavy chain genes. Cell *33*, 729–740.

Baralle, F.E., and Giudice, J. (2017). Alternative splicing as a regulator of development and tissue identity. Nat Rev Mol Cell Bio *18*, nrm.2017.27.

Battle, A., Mostafavi, S., Zhu, X., Potash, J.B., Weissman, M.M., McCormick, C., Haudenschild, C.D., Beckman, K.B., Shi, J., Mei, R., et al. (2014). Characterizing the genetic basis of transcriptome diversity through RNA-sequencing of 922 individuals. Genome Res *24*, 14–24.

Beagrie, R.A., Scialdone, A., Schueler, M., Kraemer, D.C.A., Chotalia, M., Xie, S.Q., Barbieri, M., Santiago, I. de, Lavitas, L.-M., Branco, M.R., et al. (2017). Complex multi-enhancer contacts captured by genome architecture mapping. Nature *543*, 519–524.

Benazra, M., Lecomte, M.-J., Colace, C., Müller, A., Machado, C., Pechberty, S., Bricout-Neveu, E., Grenier-Godard, M., Solimena, M., Scharfmann, R., et al. (2015). A human beta cell line with drug inducible excision of immortalizing transgenes. Mol Metab *4*, 916–925.

Benner, C., Spencer, C.C.A., Havulinna, A.S., Salomaa, V., Ripatti, S., and Pirinen, M. (2016). FINEMAP: efficient variable selection using summary data from genome-wide association studies. Bioinformatics *32*, 1493–1501.

Benoist, C., and Chambon, P. (1981). In vivo sequence requirements of the SV40 early promoter region. Nature *290*, 304–310.

Berkum, N.L. van, Lieberman-Aiden, E., Williams, L., Imakaev, M., Gnirke, A., Mirny, L.A., Dekker, J., and Lander, E.S. (2010). Hi-C: A Method to Study the Three-dimensional Architecture of Genomes. J Vis Exp Jove 1869.

Bonnefond, A., and Froguel, P. (2015). Rare and Common Genetic Events in Type 2 Diabetes: What Should Biologists Know? Cell Metab *21*, 357–368.

Bonnefond, A., Froguel, P., and Vaxillaire, M. (2010). The emerging genetics of type 2 diabetes. Trends Mol Med *16*, 407–416.

Boyle, A.P., Davis, S., Shulha, H.P., Meltzer, P., Margulies, E.H., Weng, Z., Furey, T.S., and Crawford, G.E. (2008). High-Resolution Mapping and Characterization of Open Chromatin across the Genome. Cell *132*, 311–322.

Buenrostro, J.D., Giresi, P.G., Zaba, L.C., Chang, H.Y., and Greenleaf, W.J. (2013). Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position. Nat Methods *10*, 1213 1218.

Buenrostro, J.D., Wu, B., Chang, H.Y., and Greenleaf, W.J. (2015). ATAC-seq: A Method for Assaying Chromatin Accessibility Genome-Wide. Curr Protoc Mol Biology *109*, 21.29.1-21.29.9.

Bunt, M. van de, Fox, J.E.M., Dai, X., Barrett, A., Grey, C., Li, L., Bennett, A.J., Johnson, P.R., Rajotte, R.V., Gaulton, K.J., et al. (2015). Transcript Expression Data from Human Islets Links Regulatory Signals from Genome-Wide Association Studies for Type 2 Diabetes and Glycemic Traits to Their Downstream Effectors. Plos Genet *11*, e1005694.

Burton, P.R., Clayton, D.G., Cardon, L.R., Craddock, N., Deloukas, P., Duncanson, A., Kwiatkowski, D.P., McCarthy, M.I., Ouwehand, W.H., Samani, N.J., et al. (2007). Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. Nature *447*, 661–678.

Butler, A., Hoffman, P., Smibert, P., Papalexi, E., and Satija, R. (2018). Integrating single-cell transcriptomic data across different conditions, technologies, and species. Nat Biotechnol *36*, 411–420.

Bycroft, C., Freeman, C., Petkova, D., Band, G., Elliott, L.T., Sharp, K., Motyer, A., Vukcevic, D., Delaneau, O., O'Connell, J., et al. (2018). The UK Biobank resource with deep phenotyping and genomic data. Nature *562*, 203–209.

Carninci, P., Kasukawa, T., Katayama, S., Gough, J., Frith, M.C., Maeda, N., Oyama, R., Ravasi, T., Lenhard, B., Wells, C., et al. (2005). The Transcriptional Landscape of the Mammalian Genome. Science *309*, 1559–1563.

Carninci, P., Sandelin, A., Lenhard, B., Katayama, S., Shimokawa, K., Ponjavic, J., Semple, C.A.M., Taylor, M.S., Engström, P.G., Frith, M.C., et al. (2006). Genome-wide analysis of mammalian promoter architecture and evolution. Nat Genet *38*, 626–635.

Chen, L., Ge, B., Casale, F.P., Vasquez, L., Kwan, T., Garrido-Martín, D., Watt, S., Yan, Y., Kundu, K., Ecker, S., et al. (2016a). Genetic Drivers of Epigenetic and Transcriptional Variation in Human Immune Cells. Cell *167*, 1398-1414.e24.

Chen, W., Larrabee, B.R., Ovsyannikova, I.G., Kennedy, R.B., Haralambieva, I.H., Poland, G.A., and Schaid, D.J. (2015). Fine Mapping Causal Variants with an Approximate Bayesian Method Using Marginal Test Statistics. Genetics *200*, 719–736.

Chen, W., McDonnell, S.K., Thibodeau, S.N., Tillmans, L.S., and Schaid, D.J. (2016b). Incorporating Functional Annotations for Fine-Mapping Causal Variants in a Bayesian Framework Using Summary Statistics. Genetics *204*, 933–958.

Claussnitzer, M., Cho, J.H., Collins, R., Cox, N.J., Dermitzakis, E.T., Hurles, M.E., Kathiresan, S., Kenny, E.E., Lindgren, C.M., MacArthur, D.G., et al. (2020). A brief history of human disease genetics. Nature *577*, 179–189.

Consortium, T.E.P. (2004). The ENCODE (ENCyclopedia Of DNA Elements) Project. Science *306*, 636–640.

Consortium, T.Gte. (2020). The GTEx Consortium atlas of genetic regulatory effects across human tissues. Science *369*, 1318–1330.

Consortium, Dia.G.R.A.M. (DIAGRAM), Consortium, A.G.E.N.T. 2 D. (AGEN-T., Consortium, S.A.T. 2 D. (SAT2D), Consortium, M.A.T. 2 D. (MAT2D), Consortium, T. 2 D.G.E. by N. sequencing in multi-E.S. (T2D-G., Mahajan, A., Go, M.J., Zhang, W., Below, J.E., Gaulton, K.J., et al. (2014a). Genome-wide trans-ancestry meta-analysis provides insight into the genetic architecture of type 2 diabetes susceptibility. Nat Genet *46*, 234–244.

Consortium, Gte., Gamazon, E.R., Wheeler, H.E., Shah, K.P., Mozaffari, S.V., Aquino-Michaels, K., Carroll, R.J., Eyler, A.E., Denny, J.C., Nicolae, D.L., et al. (2015a). A gene-based association method for mapping traits using reference transcriptome data. Nat Genet *47*, 1091–1098.

Consortium, I.P.A., Weedon, M.N., Cebola, I., Patch, A.-M., Flanagan, S.E., Franco, E.D., Caswell, R., Rodríguez-Seguí, S.A., Shaw-Smith, C., Cho, C.H.-H., et al. (2014b). Recessive mutations in a distal PTF1A enhancer cause isolated pancreatic agenesis. Nat Genet *46*, 61–64.

Consortium, T.G., Lappalainen, T., Sammeth, M., Friedländer, M.R., Hoen, P.A.C. 't, Monlong, J., Rivas, M.A., Gonzàlez-Porta, M., Kurbatova, N., Griebel, T., et al. (2013). Transcriptome and genome sequencing uncovers functional variation in humans. Nature *501*, 506–511.

Consortium, T.Gte., Ardlie, K.G., Deluca, D.S., Segre, A.V., Sullivan, T.J., Young, T.R., Gelfand, E.T., Trowbridge, C.A., Maller, J.B., Tukiainen, T., et al. (2015b). The Genotype-Tissue Expression (GTEx) pilot analysis: Multitissue gene regulation in humans. Science *348*, 648–660.

Cooper, R.S., Tayo, B., and Zhu, X. (2008). Genome-wide association studies: implications for multiethnic samples. Hum Mol Genet *17*, R151–R155.

Cowper-Sal·lari, R., Zhang, X., Wright, J.B., Bailey, S.D., Cole, M.D., Eeckhoute, J., Moore, J.H., and Lupien, M. (2012). Breast cancer risk–associated SNPs modulate the affinity of chromatin for FOXA1 and alter gene expression. Nat Genet *44*, 1191–1198.

Crawford, G.E., Davis, S., Scacheri, P.C., Renaud, G., Halawi, M.J., Erdos, M.R., Green, R., Meltzer, P.S., Wolfsberg, T.G., and Collins, F.S. (2006). DNase-chip: a high-resolution method to identify DNase I hypersensitive sites using tiled microarrays. Nat Methods *3*, 503–509.

Cremer, T., and Cremer, C. (2001). Chromosome territories, nuclear architecture and gene regulation in mammalian cells. Nat Rev Genet *2*, 292–301.

Creyghton, M.P., Cheng, A.W., Welstead, G.G., Kooistra, T., Carey, B.W., Steine, E.J., Hanna, J., Lodato, M.A., Frampton, G.M., Sharp, P.A., et al. (2010). Histone H3K27ac separates active from poised enhancers and predicts developmental state. Proc National Acad Sci *107*, 21931–21936.

CRICK, F. (1970). Central Dogma of Molecular Biology. Nature *227*, 561–563.

Cummings, B.B., Karczewski, K.J., Kosmicki, J.A., Seaby, E.G., Watts, N.A., Singer-Berk, M., Mudge, J.M., Karjalainen, J., Satterstrom, F.K., O'Donnell-Luria, A.H., et al. (2020). Transcript expression-aware annotation improves rare variant interpretation. Nature *581*, 452–458.

Daneman, D. (2006). Type 1 diabetes. Lancet *367*, 847–858.

DeFronzo, R.A. (2004). Pathogenesis of type 2 diabetes mellitus. Med Clin N Am *88*, 787–835.

DeFronzo, R.A., Ferrannini, E., Groop, L., Henry, R.R., Herman, W.H., Holst, J.J., Hu, F.B., Kahn, C.R., Raz, I., Shulman, G.I., et al. (2015). Type 2 diabetes mellitus. Nat Rev Dis Primers *1*, 15019.

Delaneau, O., Ongen, H., Brown, A.A., Fort, A., Panousis, N.I., and Dermitzakis, E.T. (2017). A complete tool set for molecular QTL discovery and analysis. Nat Commun *8*, 15452.

(DGT), F.C. and the R.P. and C., Forrest, A.R.R., Kawaji, H., Rehli, M., Baillie, J.K., Hoon, M.J.L. de, Haberle, V., Lassmann, T., Kulakovskiy, I.V., Lizio, M., et al. (2014). A promoter-level mammalian expression atlas. Nature *507*, 462–470.

Dimas, A.S., Deutsch, S., Stranger, B.E., Montgomery, S.B., Borel, C., Attar-Cohen, H., Ingle, C., Beazley, C., Arcelus, M.G., Sekowska, M., et al. (2009). Common Regulatory Variation Impacts Gene Expression in a Cell Type–Dependent Manner. Science *325*, 1246–1250.

Dobin, A., Davis, C.A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., Batut, P., Chaisson, M., and Gingeras, T.R. (2013). STAR: ultrafast universal RNA-seq aligner. Bioinformatics *29*, 15–21.

Dolenšek, J., Rupnik, M.S., and Stožer, A. (2015). Structural similarities and differences between the human and the mouse pancreas. Islets *7*, e1024405.

Drexler, H.L., Choquet, K., and Churchman, L.S. (2019). Splicing Kinetics and Coordination Revealed by Direct Nascent RNA Sequencing through Nanopores. Mol Cell *77*, 985-998.e8.

Dubois, P.C.A., Trynka, G., Franke, L., Hunt, K.A., Romanos, J., Curtotti, A., Zhernakova, A., Heap, G.A.R., Ádány, R., Aromaa, A., et al. (2010). Multiple common variants for celiac disease influencing immune gene expression. Nat Genet *42*, 295–302.

Dunham, I., Kundaje, A., Aldred, S.F., Collins, P.J., Davis, C.A., Doyle, F., Epstein, C.B., Frietze, S., Harrow, J., Kaul, R., et al. (2012). An integrated encyclopedia of DNA elements in the human genome. Nature *489*, 57–74.

Eberwine, J., Sul, J.-Y., Bartfai, T., and Kim, J. (2014). The promise of single-cell sequencing. Nat Methods *11*, 25–27.

Enge, M., Arda, H.E., Mignardi, M., Beausang, J., Bottino, R., Kim, S.K., and Quake, S.R. (2017). Single-Cell Analysis of Human Pancreas Reveals Transcriptional Signatures of Aging and Somatic Mutation Patterns. Cell *171*, 321 330.e14.

Engström, P.G., Sui, S.J.H., Drivenes, Ø., Becker, T.S., and Lenhard, B. (2007). Genomic regulatory blocks underlie extensive microsynteny conservation in insects. Genome Res *17*, 1898–1908.

Fachal, L., Aschard, H., Beesley, J., Barnes, D.R., Allen, J., Kar, S., Pooley, K.A., Dennis, J., Michailidou, K., Turman, C., et al. (2020). Fine-mapping of 150 breast cancer risk regions identifies 191 likely target genes. Nat Genet *52*, 56–73.

Fadista, J., Vikman, P., Laakso, E.O., Mollet, I.G., Esguerra, J.L., Taneera, J., Storm, P., Osmark, P., Ladenvall, C., Prasad, R.B., et al. (2014). Global genomic and transcriptomic analysis of human pancreatic islets reveals novel genes influencing glucose metabolism. Proc National Acad Sci *111*, 13924–13929.

Fang, H., Beckmann, G., Bountra, C., Bowness, P., Burgess-Brown, N., Carpenter, L., Chen, L., Damerell, D., Egner, U., Fang, H., et al. (2019). A genetics-led approach defines the drug target landscape of 30 immune-related traits. Nat Genet *51*, 1082–1091.

Farh, K.K.-H., Marson, A., Zhu, J., Kleinewietfeld, M., Housley, W.J., Beik, S., Shoresh, N., Whitton, H., Ryan, R.J.H., Shishkin, A.A., et al. (2015). Genetic and epigenetic fine mapping of causal autoimmune disease variants. Nature *518*, 337–343.

Faustino, N.A., and Cooper, T.A. (2003). Pre-mRNA splicing and human disease. Gene Dev *17*, 419–437.

Feero, W.G., Guttmacher, A.E., and McCarthy, M.I. (2010). Genomics, Type 2 Diabetes, and Obesity. New Engl J Medicine *363*, 2339–2350.

Ferraro, N.M., Strober, B.J., Einson, J., Abell, N.S., Aguet, F., Barbeira, A.N., Brandt, M., Bucan, M., Castel, S.E., Davis, J.R., et al. (2020). Transcriptomic signatures across human tissues identify functional rare genetic variation. Science *369*, eaaz5900.

Flanagan, S.E., Clauin, S., Bellanné-Chantelot, C., Lonlay, P. de, Harries, L.W., Gloyn, A.L., and Ellard, S. (2009). Update of mutations in the genes encoding the pancreatic beta-cell KATP channel subunits Kir6.2 (KCNJ11) and sulfonylurea receptor 1 (ABCC8) in diabetes mellitus and hyperinsulinism. Hum Mutat *30*, 170–180.

Flicek, P., Amode, M.R., Barrell, D., Beal, K., Brent, S., Chen, Y., Clapham, P., Coates, G., Fairley, S., Fitzgerald, S., et al. (2011). Ensembl 2011. Nucleic Acids Res *39*, D800–D806.

Fu, X.-D., and Ares, M. (2014). Context-dependent control of alternative splicing by RNA-binding proteins. Nature Reviews Genetics *15*.

Gaulton, K.J., Nammo, T., Pasquali, L., Simon, J.M., Giresi, P.G., Fogarty, M.P., Panhuis, T.M., Mieczkowski, P., Secchi, A., Bosco, D., et al. (2010). A map of open chromatin in human pancreatic islets. Nat Genet *42*, 255–259.

Gaziano, J.M., Concato, J., Brophy, M., Fiore, L., Pyarajan, S., Breeling, J., Whitbourne, S., Deen, J., Shannon, C., Humphries, D., et al. (2016). Million Veteran Program: A mega-biobank to study genetic influences on health and disease. J Clin Epidemiol *70*, 214–223.

Geijn, B. van de, McVicker, G., Gilad, Y., and Pritchard, J.K. (2015). WASP: allele-specific software for robust molecular quantitative trait locus discovery. Nat Methods *12*, 1061–1063.

Geyer, P.K., Vitalini, M.W., and Wallrath, L.L. (2011). Nuclear organization: taking a position on gene expression. Curr Opin Cell Biol *23*, 354–359.

Giambartolomei, C., Vukcevic, D., Schadt, E.E., Franke, L., Hingorani, A.D., Wallace, C., and Plagnol, V. (2014). Bayesian Test for Colocalisation between Pairs of Genetic Association Studies Using Summary Statistics. PLoS Genetics *10*.

Gilbert, N., Boyle, S., Fiegler, H., Woodfine, K., Carter, N.P., and Bickmore, W.A. (2004). Chromatin Architecture of the Human Genome Gene-Rich Domains Are Enriched in Open Chromatin Fibers. Cell *118*, 555–566.

Greenwald, W.W., Chiou, J., Yan, J., Qiu, Y., Dai, N., Wang, A., Nariai, N., Aylward, A., Han, J.Y., Kadakia, N., et al. (2018). Pancreatic islet chromatin accessibility and conformation defines distal enhancer networks of type 2 diabetes risk. Biorxiv 1 38.

Greenwald, W.W., Chiou, J., Yan, J., Qiu, Y., Dai, N., Wang, A., Nariai, N., Aylward, A., Han, J.Y., Kadakia, N., et al. (2019). Pancreatic islet chromatin accessibility and conformation reveals distal enhancer networks of type 2 diabetes risk. Nat Commun *10*, 2078.

Grosschedl, R., and Birnstiel, M.L. (1980). Spacer DNA sequences upstream of the T-A-T-A-A-A-T-A sequence are essential for promotion of H2A histone gene transcription in vivo. Proc National Acad Sci *77*, 7102–7106.

Gu, W., Crawford, E.D., O'Donovan, B.D., Wilson, M.R., Chow, E.D., Retallack, H., and DeRisi, J.L. (2016). Depletion of Abundant Sequences by Hybridization (DASH): using Cas9 to remove unwanted high-abundance species in sequencing libraries and molecular counting applications. Genome Biol *17*, 41.

Gupta, R.M., Hadaya, J., Trehan, A., Zekavat, S.M., Roselli, C., Klarin, D., Emdin, C.A., Hilvering, C.R.E., Bianchi, V., Mueller, C., et al. (2017). A Genetic Variant Associated with Five Vascular Diseases Is a Distal Regulator of Endothelin-1 Gene Expression. Cell *170*, 522-533.e15.

Gusev, A., Ko, A., Shi, H., Bhatia, G., Chung, W., Penninx, B.W.J.H., Jansen, R., Geus, E.J.C. de, Boomsma, D.I., Wright, F.A., et al. (2016). Integrative approaches for large-scale transcriptome-wide association studies. Nat Genet *48*, 245–252.

Guttman, M., Amit, I., Garber, M., French, C., Lin, M.F., Feldser, D., Huarte, M., Zuk, O., Carey, B.W., Cassady, J.P., et al. (2009). Chromatin signature reveals over a thousand highly conserved large non-coding RNAs in mammals. Nature *458*, 223–227.

Haberle, V., and Stark, A. (2018). Eukaryotic core promoters and the functional basis of transcription initiation. Nat Rev Mol Cell Bio *19*, 621–637.

Harrow, J., Denoeud, F., Frankish, A., Reymond, A., Chen, C.-K., Chrast, J., Lagarde, J., Gilbert, J.G., Storey, R., Swarbreck, D., et al. (2006). GENCODE: producing a reference annotation for ENCODE. Genome Biol *7*, S4.

Harrow, J., Frankish, A., Gonzalez, J.M., Tapanari, E., Diekhans, M., Kokocinski, F., Aken, B.L., Barrell, D., Zadissa, A., Searle, S., et al. (2012). GENCODE: The reference human genome annotation for The ENCODE Project. Genome Res *22*, 1760–1774.

Hattersley, A.T., and Ashcroft, F.M. (2005). Activating Mutations in Kir6.2 and Neonatal Diabetes: New Clinical Syndromes, New Scientific Insights, and New Therapy. Diabetes *54*, 2503–2513.

Hattersley, A.T., and Patel, K.A. (2017). Precision diabetes: learning from monogenic diabetes. Diabetologia *60*, 769–777.

He, X., Fuller, C.K., Song, Y., Meng, Q., Zhang, B., Yang, X., and Li, H. (2013). Sherlock: Detecting Gene-Disease Associations by Matching Patterns of Expression QTL and GWAS. Am J Hum Genetics *92*, 667–680.

Heintzman, N.D., Stuart, R.K., Hon, G., Fu, Y., Ching, C.W., Hawkins, R.D., Barrera, L.O., Calcar, S.V., Qu, C., Ching, K.A., et al. (2007). Distinct and predictive chromatin signatures of transcriptional promoters and enhancers in the human genome. Nat Genet *39*, 311–318.

Heintzman, N.D., Hon, G.C., Hawkins, R.D., Kheradpour, P., Stark, A., Harp, L.F., Ye, Z., Lee, L.K., Stuart, R.K., Ching, C.W., et al. (2009). Histone modifications at human enhancers reflect global cell-type-specific gene expression. Nature *459*, 108–112.

Heinz, S., Benner, C., Spann, N., Bertolino, E., Lin, Y.C., Laslo, P., Cheng, J.X., Murre, C., Singh, H., and Glass, C.K. (2010). Simple Combinations of Lineage-

Determining Transcription Factors Prime cis-Regulatory Elements Required for Macrophage and B Cell Identities. Mol Cell *38*, 576–589.

Heinz, S., Romanoski, C.E., Benner, C., and Glass, C.K. (2015). The selection and function of cell type-specific enhancers. Nat Rev Mol Cell Bio *16*, 144–154.

Hemani, G., Bowden, J., and Smith, G.D. (2018). Evaluating the potential role of pleiotropy in Mendelian randomization studies. Human Molecular Genetics *27*, R195–R208.

Hindorff, L.A., Sethupathy, P., Junkins, H.A., Ramos, E.M., Mehta, J.P., Collins, F.S., and Manolio, T.A. (2009). Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. P Natl Acad Sci Usa *106*, 9362–9367.

Hon, C.-C., Ramilowski, J.A., Harshbarger, J., Bertin, N., Rackham, O.J.L., Gough, J., Denisenko, E., Schmeier, S., Poulsen, T.M., Severin, J., et al. (2017). An atlas of human long non-coding RNAs with accurate 5′ ends. Nature *543*, 199–204.

Hoon, M.J.L. de, Imoto, S., Nolan, J., and Miyano, S. (2004). Open source clustering software. Bioinformatics *20*, 1453–1454.

Hormozdiari, F., Kostem, E., Kang, E.Y., Pasaniuc, B., and Eskin, E. (2014). Identifying Causal Variants at Loci with Multiple Signals of Association. Genetics *198*, genetics.114.167908.

Hormozdiari, F., van de Bunt, M., Segrè, A.V., Li, X., Joo, J.W.J., Bilow, M., Sul, J.H., Sankararaman, S., Pasaniuc, B., and Eskin, E. (2016). Colocalization of GWAS and eQTL Signals Detects Target Genes. The American Journal of Human Genetics *99*.

Hormozdiari, F., Zhu, A., Kichaev, G., Ju, C.J.-T., Segrè, A.V., Joo, J.W.J., Won, H., Sankararaman, S., Pasaniuc, B., Shifman, S., et al. (2017). Widespread Allelic Heterogeneity in Complex Traits. Am J Hum Genetics *100*, 789–802.

Huang, Y.-F., Gulko, B., and Siepel, A. (2017). Fast, scalable prediction of deleterious noncoding variants from functional and population genomic data. Nat Genet *49*, 618–624.

Hutchinson, A., Asimit, J., and Wallace, C. (2020). Fine-mapping genetic associations. Hum Mol Genet ddaa148-.

InternationalDiabetesFederation (2019). International Diabetes Federation. International Diabetes Federation 9th edn.

Ionita-Laza, I., McCallum, K., Xu, B., and Buxbaum, J.D. (2016). A spectral approach integrating functional genomic annotations for coding and noncoding variants. Nat Genet *48*, 214–220.

Iyer, M.K., Niknafs, Y.S., Malik, R., Singhal, U., Sahu, A., Hosono, Y., Barrette, T.R., Prensner, J.R., Evans, J.R., Zhao, S., et al. (2015). The landscape of long noncoding RNAs in the human transcriptome. Nat Genet *47*, 199–208.

Jaenisch, R., and Bird, A. (2003). Epigenetic regulation of gene expression: how the genome integrates intrinsic and environmental signals. Nat Genet *33*, 245–254.

Jansen, R., Hottenga, J.-J., Nivard, M.G., Abdellaoui, A., Laport, B., Geus, E.J. de, Wright, F.A., Penninx, B.W.J.H., and Boomsma, D.I. (2017). Conditional eQTL analysis reveals allelic heterogeneity of gene expression. Hum Mol Genet *26*, 1444–1451.

Javierre, B.M., Burren, O.S., Wilder, S.P., Kreuzhuber, R., Hill, S.M., Sewitz, S., Cairns, J., Wingett, S.W., Várnai, C., Thiecke, M.J., et al. (2016). Lineage-Specific Genome Architecture Links Enhancers and Non-coding Disease Variants to Target Gene Promoters. Cell *167*, 1369-1384.e19.

Jew, B., Alvarez, M., Rahmani, E., Miao, Z., Ko, A., Garske, K.M., Sul, J.H., Pietiläinen, K.H., Pajukanta, P., and Halperin, E. (2020). Accurate estimation of cell composition in bulk expression through robust integration of single-cell information. Nat Commun *11*, 1971.

Johnson, W.E., Li, C., and Rabinovic, A. (2007). Adjusting batch effects in microarray expression data using empirical Bayes methods. Biostatistics *8*, 118–127.

Jolma, A., Yin, Y., Nitta, K.R., Dave, K., Popov, A., Taipale, M., Enge, M., Kivioja, T., Morgunova, E., and Taipale, J. (2015). DNA-dependent formation of transcription factor pairs alters their binding specificity. Nature *527*, 384.

Jun, G., Flickinger, M., Hetrick, K.N., Romm, J.M., Doheny, K.F., Abecasis, G.R., Boehnke, M., and Kang, H.M. (2012). Detecting and Estimating Contamination of Human DNA Samples in Sequencing and Array-Based Genotype Data. Am J Hum Genetics *91*, 839–848.

Jung, I., Schmitt, A., Diao, Y., Lee, A.J., Liu, T., Yang, D., Tan, C., Eom, J., Chan, M., Chee, S., et al. (2019). A compendium of promoter-centered long-range chromatin interactions in the human genome. Nat Genet *51*, 1442–1449.

Kagey, M.H., Newman, J.J., Bilodeau, S., Zhan, Y., Orlando, D.A., Berkum, N.L. van, Ebmeier, C.C., Goossens, J., Rahl, P.B., Levine, S.S., et al. (2010). Mediator and cohesin connect gene expression and chromatin architecture. Nature *467*, 430–435.

Kahn, S.E., Hull, R.L., and Utzschneider, K.M. (2006). Mechanisms linking obesity to insulin resistance and type 2 diabetes. Nature *444*, 840–846.

Katsarou, A., Gudbjörnsdottir, S., Rawshani, A., Dabelea, D., Bonifacio, E., Anderson, B.J., Jacobsen, L.M., Schatz, D.A., and Lernmark, Å. (2017). Type 1 diabetes mellitus. Nat Rev Dis Primers *3*, 17016.

Kempfer, R., and Pombo, A. (2020). Methods for mapping 3D chromosome architecture. Nat Rev Genet *21*, 207–226.

Khetan, S., Kursawe, R., Youn, A., Lawlor, N., Jillette, A., Marquez, E.J., Ucar, D., and Stitzel, M.L. (2018). Type 2 Diabetes Associated Genetic Variants Regulate Chromatin Accessibility in Human Islets. Diabetes *67*, db180393.

Kichaev, G., Yang, W.-Y., Lindstrom, S., Hormozdiari, F., Eskin, E., Price, A.L., Kraft, P., and Pasaniuc, B. (2014). Integrating Functional Data to Prioritize Causal Variants in Statistical Fine-Mapping Studies. Plos Genet *10*, e1004722.

Kim-Hellmuth, S., Aguet, F., Oliva, M., Muñoz-Aguirre, M., Kasela, S., Wucher, V., Castel, S.E., Hamel, A.R., Viñuela, A., Roberts, A.L., et al. (2020). Cell type–specific genetic regulation of gene expression across human tissues. Science *369*, eaaz8528.

Kircher, M., Witten, D.M., Jain, P., O'Roak, B.J., Cooper, G.M., and Shendure, J. (2014). A general framework for estimating the relative pathogenicity of human genetic variants. Nat Genet *46*, 310–315.

Kornblihtt, A.R., Schor, I.E., Alló, M., Dujardin, G., Petrillo, E., and Muñoz, M.J. (2013). Alternative splicing: a pivotal step between eukaryotic transcription and translation. Nat Rev Mol Cell Bio *14*, 153–165.

Kruesi, W.S., Core, L.J., Waters, C.T., Lis, J.T., and Meyer, B.J. (2013). Condensin controls recruitment of RNA polymerase II to achieve nematode X-chromosome dosage compensation. Elife *2*, e00808.

Kryuchkova-Mostacci, N., and Robinson-Rechavi, M. (2016). A benchmark of gene expression tissue-specificity metrics. Brief Bioinform bbw008.

Kundaje, A., Meuleman, W., Ernst, J., Bilenky, M., Yen, A., Heravi-Moussavi, A., Kheradpour, P., Zhang, Z., Wang, J., Ziller, M.J., et al. (2015). Integrative analysis of 111 reference human epigenomes. Nature *518*, 317–330.

Kwak, H., Fuda, N.J., Core, L.J., and Lis, J.T. (2013). Precise Maps of RNA Polymerase Reveal How Promoters Direct Initiation and Pausing. Science *339*, 950–953.

Lachmann, A., Torre, D., Keenan, A.B., Jagodnik, K.M., Lee, H.J., Wang, L., Silverstein, M.C., and Ma'ayan, A. (2018). Massive mining of publicly available RNA-seq data from human and mouse. Nat Commun *9*, 1366.

Lam, M.T.Y., Cho, H., Lesch, H.P., Gosselin, D., Heinz, S., Tanaka-Oishi, Y., Benner, C., Kaikkonen, M.U., Kim, A.S., Kosaka, M., et al. (2013). Rev-Erbs repress macrophage gene expression by inhibiting enhancer-directed transcription. Nature *498*, 511–515.

Lambert, S.A., Jolma, A., Campitelli, L.F., Das, P.K., Yin, Y., Albu, M., Chen, X., Taipale, J., Hughes, T.R., and Weirauch, M.T. (2018). The Human Transcription Factors. Cell *172*, 650–665.

Lander, E.S., Linton, L.M., Birren, B., Nusbaum, C., Zody, M.C., Baldwin, J., Devon, K., Dewar, K., Doyle, M., FitzHugh, W., et al. (2001). Initial sequencing and analysis of the human genome. Nature *409*, 860–921.

Lawlor, N., George, J., Bolisetty, M., Kursawe, R., Sun, L., Sivakamasundari, V., Kycia, I., Robson, P., and Stitzel, M.L. (2017). Single-cell transcriptomes identify human islet cell signatures and reveal cell-type–specific expression changes in type 2 diabetes. Genome Res *27*, 208 222.

Lee, C.-K., Shibata, Y., Rao, B., Strahl, B.D., and Lieb, J.D. (2004). Evidence for nucleosome depletion at active regulatory regions genome-wide. Nat Genet *36*, 900–905.

Lee, Y., Luca, F., Pique-Regi, R., and Wen, X. (2018). Bayesian Multi-SNP Genetic Association Analysis: Control of FDR and Use of Summary Statistics. Biorxiv 316471.

Lenhard, B., Sandelin, A., and Carninci, P. (2012). Metazoan promoters: emerging characteristics and insights into transcriptional regulation. Nature Reviews Genetics *13*.

León, D.D.D., and Stanley, C.A. (2007). Mechanisms of Disease: advances in diagnosis and treatment of hyperinsulinism in neonates. Nat Clin Pract Endoc *3*, 57–68.

Levitt, H.E., Cyphert, T.J., Pascoe, J.L., Hollern, D.A., Abraham, N., Lundell, R.J., Rosa, T., Romano, L.C., Zou, B., O'Donnell, C.P., et al. (2010). Glucose stimulates human beta cell replication in vivo in islets transplanted into NOD–severe combined immunodeficiency (SCID) mice. Diabetologia *54*, 572 582.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., Durbin, R., and Subgroup, 1000 Genome Project Data Processing (2009). The Sequence Alignment/Map format and SAMtools. Bioinformatics *25*, 2078–2079.

Li, X., Kim, Y., Tsang, E.K., Davis, J.R., Damani, F.N., Chiang, C., Hess, G.T., Zappala, Z., Strober, B.J., Scott, A.J., et al. (2017). The impact of rare variation on gene expression across tissues. Nature *550*, 239–243.

Li, Y.I., Geijn, B. van de, Raj, A., Knowles, D.A., Petti, A.A., Golan, D., Gilad, Y., and Pritchard, J.K. (2016). RNA splicing is a primary link between genetic variation and disease. Science *352*, 600–604.

Li, Y.I., Knowles, D.A., Humphrey, J., Barbeira, A.N., Dickinson, S.P., Im, H.K., and Pritchard, J.K. (2018). Annotation-free quantification of RNA splicing using LeafCutter. Nat Genet *50*, 151–158.

Li, Y.I., Wong, G., Humphrey, J., and Raj, T. (2019). Prioritizing Parkinson's disease genes using population-scale transcriptomic data. Nat Commun *10*, 994.

Liao, Y., Smyth, G.K., and Shi, W. (2014). featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. Bioinformatics *30*, 923–930.

Lieberman-Aiden, E., Berkum, N.L. van, Williams, L., Imakaev, M., Ragoczy, T., Telling, A., Amit, I., Lajoie, B.R., Sabo, P.J., Dorschner, M.O., et al. (2009). Comprehensive mapping of long-range interactions reveals folding principles of the human genome. Sci New York N Y *326*, 289–293.

Lin, C., and Yang, L. (2017). Long Noncoding RNA in Cancer: Wiring Signaling Circuitry. Trends Cell Biol *28*, 287–301.

Llorian, M., Schwartz, S., Clark, T.A., Hollander, D., Tan, L.-Y., Spellman, R., Gordon, A., Schweitzer, A.C., Grange, P. de la, Ast, G., et al. (2010). Position-dependent alternative splicing activity revealed by global profiling of alternative splicing events regulated by PTB. Nat Struct Mol Biol *17*, 1114–1123.

Long, H.K., Prescott, S.L., and Wysocka, J. (2016). Ever-Changing Landscapes: Transcriptional Enhancers in Development and Evolution. Cell *167*, 1170–1187.

Lonsdale, J., Thomas, J., Salvatore, M., Phillips, R., Lo, E., Shad, S., Hasz, R., Walters, G., Garcia, F., Young, N., et al. (2013). The Genotype-Tissue Expression (GTEx) project. Nat Genet *45*, 580–585.

Luco, R.F., Allo, M., Schor, I.E., Kornblihtt, A.R., and Misteli, T. (2011). Epigenetics in Alternative Pre-mRNA Splicing. Cell *144*, 16–26.

MacArthur, J., Bowler, E., Cerezo, M., Gil, L., Hall, P., Hastings, E., Junkins, H., McMahon, A., Milano, A., Morales, J., et al. (2017). The new NHGRI-EBI Catalog of published genome-wide association studies (GWAS Catalog). Nucleic Acids Res *45*, D896–D901.

Magnani, L., Eeckhoute, J., and Lupien, M. (2011). Pioneer factors: directing transcriptional regulators within the chromatin environment. Trends Genet *27*, 465–474.

Mahajan, A., Taliun, D., Thurner, M., Robertson, N., Torres, J., Rayner, N., Payne, A., Steinthorsdottir, V., Scott, R., Grarup, N., et al. (2018a). Fine-mapping type 2 diabetes loci to single-variant resolution using high-density imputation and islet-specific epigenome maps. Nat Genet *50*, 1505–1513.

Mahajan, A., Taliun, D., Thurner, M., Robertson, N.R., Torres, J.M., Rayner, N.W., Payne, A.J., Steinthorsdottir, V., Scott, R.A., Grarup, N., et al. (2018b). Fine-mapping type 2 diabetes loci to single-variant resolution using high-density imputation and islet-specific epigenome maps. Nat Genet *50*, 1505–1513.

Manolio, T.A., Collins, F.S., Cox, N.J., Goldstein, D.B., Hindorff, L.A., Hunter, D.J., McCarthy, M.I., Ramos, E.M., Cardon, L.R., Chakravarti, A., et al. (2009). Finding the missing heritability of complex diseases. Nature *461*, 747–753.

Marchini, J., and Howie, B. (2010). Genotype imputation for genome-wide association studies. Nat Rev Genet *11*, 499–511.

Maston, G.A., Evans, S.K., and Green, M.R. (2006). Transcriptional Regulatory Elements in the Human Genome. Genom Hum Genetics *7*, 29–59.

Maurano, M.T., Humbert, R., Rynes, E., Thurman, R.E., Haugen, E., Wang, H., Reynolds, A.P., Sandstrom, R., Qu, H., Brody, J., et al. (2012). Systematic localization of common disease-associated variation in regulatory DNA. Sci New York N Y *337*, 1190–1195.

Mayer, A., di Iulio, J., Maleri, S., Eser, U., Vierstra, J., Reynolds, A., Sandstrom, R., Stamatoyannopoulos, J.A., and Churchman, L.S. (2015). Native Elongating Transcript Sequencing Reveals Human Transcriptional Activity at Nucleotide Resolution. Cell *161*, 541–554.

Mercer, T.R., Gerhardt, D.J., Dinger, M.E., Crawford, J., Trapnell, C., Jeddeloh, J.A., Mattick, J.S., and Rinn, J.L. (2012). Targeted RNA sequencing reveals the deep complexity of the human transcriptome. Nat Biotechnol *30*, 99–104.

Metukuri, M.R., Zhang, P., Basantani, M.K., Chin, C., Stamateris, R.E., Alonso, L.C., Takane, K.K., Gramignoli, R., Strom, S.C., O'Doherty, R.M., et al. (2012). ChREBP Mediates Glucose-Stimulated Pancreatic -Cell Proliferation. Diabetes *61*, 2004–2015.

Miguel-Escalada, I., Bonàs-Guarch, S., Cebola, I., Ponsa-Cobas, J., Mendieta-Esteban, J., Atla, G., Javierre, B.M., Rolando, D.M.Y., Farabella, I., Morgan, C.C., et al. (2019). Human pancreatic islet three-dimensional chromatin architecture provides insights into the genetics of type 2 diabetes. Nat Genet *51*, 1137–1148.

Misra, S., and Owen, K.R. (2018). Genetics of Monogenic Diabetes: Present Clinical Challenges. Curr Diabetes Rep *18*, 141.

Montgomery, S.B., and Dermitzakis, E.T. (2011). From expression QTLs to personalized transcriptomics. Nat Rev Genet *12*, 277–282.

Morán, I., Akerman, I., Bunt, M. van de, Xie, R., Benazra, M., Nammo, T., Arnes, L., Nakić, N., García-Hurtado, J., Rodríguez-Seguí, S., et al. (2012). Human β cell transcriptome analysis uncovers lncRNAs that are tissue-specific, dynamically regulated, and abnormally expressed in type 2 diabetes. Cell Metab *16*, 435 448.

Mumbach, M.R., Satpathy, A.T., Boyle, E.A., Dai, C., Gowen, B.G., Cho, S.W., Nguyen, M.L., Rubin, A.J., Granja, J.M., Kazane, K.R., et al. (2017). Enhancer connectome in primary human cells identifies target genes of disease-associated DNA elements. Nat Genet *49*, 1602–1612.

Murata, M., Nishiyori-Sueki, H., Kojima-Ishiyama, M., Carninci, P., Hayashizaki, Y., and Itoh, M. (2014). Transcription Factor Regulatory Networks, Methods and Protocols. Methods Mol Biology *1164*, 67–85.

Murawska, M., and Brehm, A. (2014). CHD chromatin remodelers and the transcription cycle. Biochem Soc Symp *2*, 244–253.

Myers, A.J., Gibbs, J.R., Webster, J.A., Rohrer, K., Zhao, A., Marlowe, L., Kaleem, M., Leung, D., Bryden, L., Nath, P., et al. (2007). A survey of genetic human cortical gene expression. Nat Genet *39*, 1494–1499.

Nagai, K., and Fica, S.M. (2017). Cryo-electron microscopy snapshots of the spliceosome: structural insights into a dynamic ribonucleoprotein machine. Nat Struct Mol Biology *24*, 791.

Narlikar, G.J., Sundaramoorthy, R., and Owen-Hughes, T. (2013). Mechanisms and Functions of ATP-Dependent Chromatin-Remodeling Enzymes. Cell *154*, 490–503.

Ndungu, A., Payne, A., Torres, J.M., Bunt, M. van de, and McCarthy, M.I. (2020). A Multi-tissue Transcriptome Analysis of Human Metabolites Guides Interpretability of Associations Based on Multi-SNP Models for Gene Expression. Am J Hum Genet.

Nellore, A., Jaffe, A.E., Fortin, J.-P., Alquicira-Hernández, J., Collado-Torres, L., Wang, S., III, R.A.P., Karbhari, N., Hansen, K.D., Langmead, B., et al. (2016). Human splicing diversity and the extent of unannotated splice junctions across human RNA-seq samples on the Sequence Read Archive. Genome Biol *17*, 266.

Newman, A.M., Liu, C.L., Green, M.R., Gentles, A.J., Feng, W., Xu, Y., Hoang, C.D., Diehn, M., and Alizadeh, A.A. (2015). Robust enumeration of cell subsets from tissue expression profiles. Nat Methods *12*, 453–457.

Newman, B., Selby, J.V., King, M.-C., Slemenda, C., Fabsitz, R., and Friedman, G.D. (1987). Concordance for Type 2 (non-insulin-dependent) diabetes mellitus in male twins. Diabetologia *30*, 763–768.

Nica, A.C., Montgomery, S.B., Dimas, A.S., Stranger, B.E., Beazley, C., Barroso, I., and Dermitzakis, E.T. (2010). Candidate Causal Regulatory Effects by Integration of Expression QTLs with Complex Trait Genetic Associations. Plos Genet *6*, e1000895.

Nott, A., Holtman, I.R., Coufal, N.G., Schlachetzki, J.C.M., Yu, M., Hu, R., Han, C.Z., Pena, M., Xiao, J., Wu, Y., et al. (2019). Brain cell type–specific enhancer–promoter interactome maps and disease-risk association. Science *366*, 1134–1139.

Ong, C.-T., and Corces, V.G. (2014). CTCF: an architectural protein bridging genome topology and function. Nat Rev Genet *15*, 234–246.

Ostuni, R., Piccolo, V., Barozzi, I., Polletti, S., Termanini, A., Bonifacio, S., Curina, A., Prosperini, E., Ghisletti, S., and Natoli, G. (2013). Latent Enhancers Activated by Stimulation in Differentiated Cells. Cell *152*, 157–171.

Pan, Q., Shai, O., Lee, L.J., Frey, B.J., and Blencowe, B.J. (2008). Deep surveying of alternative splicing complexity in the human transcriptome by high-throughput sequencing. Nat Genet *40*, 1413–1415.

Parker, S.C.J., Stitzel, M.L., Taylor, D.L., Orozco, J.M., Erdos, M.R., Akiyama, J.A., Bueren, K.L. van, Chines, P.S., Narisu, N., Program, N.C.S., et al. (2013). Chromatin stretch enhancer states drive cell-specific gene regulation and harbor human disease risk variants. Proc National Acad Sci *110*, 17921–17926.

Pasquali, L., Gaulton, K.J., Rodríguez-Seguí, S.A., Mularoni, L., Miguel-Escalada, I., Akerman, İ., Tena, J.J., Morán, I., Gómez-Marín, C., Bunt, M. van de, et al. (2014). Pancreatic islet enhancer clusters enriched in type 2 diabetes risk-associated variants. Nat Genet *46*, 136–143.

Patro, R., Duggal, G., Love, M.I., Irizarry, R.A., and Kingsford, C. (2017). Salmon provides fast and bias-aware quantification of transcript expression. Nat Methods *31*, 3881 10.

Penny, G.D., Kay, G.F., Sheardown, S.A., Rastan, S., and Brockdorff, N. (1996). Requirement for Xist in X chromosome inactivation. Nature *379*, 131–137.

Pertea, M., Pertea, G.M., Antonescu, C.M., Chang, T.-C., Mendell, J.T., and Salzberg, S.L. (2015). StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. Nat Biotechnol *33*, 290–295.

Pickrell, J.K. (2014). Joint Analysis of Functional Genomic Data and Genome-wide Association Studies of 18 Human Traits. Am J Hum Genetics *94*, 559–573.

Pickrell, J.K., Berisa, T., Liu, J.Z., Ségurel, L., Tung, J.Y., and Hinds, D.A. (2016). Detection and interpretation of shared genetic influences on 42 human traits. Nat Genet *48*, 709–717.

Poitout, V., and Robertson, R.P. (2008). Glucolipotoxicity: Fuel Excess and β-Cell Dysfunction. Endocr Rev *29*, 351–366.

Poitout, V., Amyot, J., Semache, M., Zarrouki, B., Hagman, D., and Fontés, G. (2010). Glucolipotoxicity of the pancreatic beta cell. Biochimica Et Biophysica Acta Bba - Mol Cell Biology Lipids *1801*, 289–298.

Porat, S., Weinberg-Corem, N., Tornovsky-Babaey, S., Schyr-Ben-Haroush, R., Hija, A., Stolovich-Rain, M., Dadon, D., Granot, Z., Ben-Hur, V., White, P., et al. (2011). Control of Pancreatic β Cell Regeneration by Glucose Metabolism. Cell Metab *13*, 440–449.

Porcu, E., Rüeger, S., Lepik, K., Agbessi, M., Ahsan, H., Alves, I., Andiappan, A., Arindrarto, W., Awadalla, P., Battle, A., et al. (2019). Mendelian randomization integrating GWAS and eQTL data reveals genetic determinants of complex and clinical traits. Nat Commun *10*, 3300.

Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M.A.R., Bender, D., Maller, J., Sklar, P., Bakker, P.I.W. de, Daly, M.J., et al. (2007). PLINK: A Tool Set for Whole-Genome Association and Population-Based Linkage Analyses. Am J Hum Genetics *81*, 559–575.

Quinodoz, S., and Guttman, M. (2014). Long noncoding RNAs: an emerging link between gene regulation and nuclear organization. Trends Cell Biol *24*, 651–663.

Raj, T., Li, Y.I., Wong, G., Humphrey, J., Wang, M., Ramdhani, S., Wang, Y.-C., Ng, B., Gupta, I., Haroutunian, V., et al. (2018a). Integrative transcriptome analyses of the aging brain implicate altered splicing in Alzheimer's disease susceptibility. Nat Genet *50*, 1584–1592.

Raj, T., Li, Y.I., Wong, G., Humphrey, J., Wang, M., Ramdhani, S., Wang, Y.-C., Ng, B., Gupta, I., Haroutunian, V., et al. (2018b). Integrative transcriptome analyses of the aging brain implicate altered splicing in Alzheimer's disease susceptibility. Nat Genet *50*, 1584–1592.

Richardson, T.G., Hemani, G., Gaunt, T.R., Relton, C.L., and Smith, G.D. (2020). A transcriptome-wide Mendelian randomization study to uncover tissue-dependent regulatory mechanisms across the human phenome. Nat Commun *11*, 185.

Ritchie, G.R.S., Dunham, I., Zeggini, E., and Flicek, P. (2014). Functional annotation of noncoding sequence variants. Nat Methods *11*, 294–296.

Robinson, J.T., Thorvaldsdóttir, H., Winckler, W., Guttman, M., Lander, E.S., Getz, G., and Mesirov, J.P. (2011). Integrative genomics viewer. Nat Biotechnol *29*, 24–26.

Robinson, M.D., McCarthy, D.J., and Smyth, G.K. (2010). edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics *26*, 139–140.

Roh, T., Ngau, W.C., Cui, K., Landsman, D., and Zhao, K. (2004). High-resolution genome-wide mapping of histone modifications. Nat Biotechnol *22*, 1013–1016.

Sanford, J.R., and Caceres, J.F. (2004). Pre-mRNA splicing: life at the centre of the central dogma. J Cell Sci *117*, 6261–6263.

Schaffer, L.V., Millikin, R.J., Miller, R.M., Anderson, L.C., Fellers, R.T., Ge, Y., Kelleher, N.L., LeDuc, R.D., Liu, X., Payne, S.H., et al. (2019). Identification and Quantification of Proteoforms by Mass Spectrometry. Proteomics *19*, 1800361.

Schaid, D.J., Chen, W., and Larson, N.B. (2018). From genome-wide associations to candidate causal variants by statistical fine-mapping. Nature Reviews Genetics *19*.

Scheuermann, J.C., and Boyer, L.A. (2013). Getting to the heart of the matter: long non-coding RNAs in cardiac development and disease. Embo J *32*, 1805–1816.

Schmidt, A.F., Finan, C., Gordillo-Marañón, M., Asselbergs, F.W., Freitag, D.F., Patel, R.S., Tyl, B., Chopade, S., Faraway, R., Zwierzyna, M., et al. (2020). Genetic drug target validation using Mendelian randomisation. Nat Commun *11*, 3255.

Schmidt, D., Schwalie, P.C., Ross-Innes, C.S., Hurtado, A., Brown, G.D., Carroll, J.S., Flicek, P., and Odom, D.T. (2010). A CTCF-independent role for cohesin in tissue-specific transcription. Genome Res *20*, 578–588.

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

Schmidt, S.F., Madsen, J.G.S., Frafjord, K.Ø., Poulsen, L. la C., Salö, S., Boergesen, M., Loft, A., Larsen, B.D., Madsen, M.S., Holst, J.J., et al. (2016). Integrative Genomics Outlines a Biphasic Glucose Response and a ChREBP-RORγ Axis Regulating Proliferation in β Cells. Cell Reports *16*, 2359 2372.

Segerstolpe, Å., Palasantza, A., Eliasson, P., Andersson, E.-M., Andréasson, A.-C., Sun, X., Picelli, S., Sabirsh, A., Clausen, M., Bjursell, M.K., et al. (2016). Single-Cell Transcriptome Profiling of Human Pancreatic Islets in Health and Type 2 Diabetes. Cell Metab *24*, 593 607.

Sharon, D., Tilgner, H., Grubert, F., and Snyder, M. (2013). A single-molecule long-read survey of the human transcriptome. Nat Biotechnol *31*, 1009–1014.

Shaw-Smith, C., Franco, E.D., Allen, H.L., Batlle, M., Flanagan, S.E., Borowiec, M., Taplin, C.E., Velden, J. van A. der, Cruz-Rojo, J., Nanclares, G.P. de, et al. (2014). GATA4 Mutations Are a Cause of Neonatal and Childhood-Onset Diabetes. Diabetes *63*, 2888–2894.

Sherstyuk, V.V., Medvedev, S.P., and Zakian, S.M. (2018). Noncoding RNAs in the Regulation of Pluripotency and Reprogramming. Stem Cell Rev Rep *14*, 58–70.

Shields, B.M., Hicks, S., Shepherd, M.H., Colclough, K., Hattersley, A.T., and Ellard, S. (2010). Maturity-onset diabetes of the young (MODY): how many cases are we missing? Diabetologia *53*, 2504–2508.

Shih, D.Q., Screenan, S., Munoz, K.N., Philipson, L., Pontoglio, M., Yaniv, M., Polonsky, K.S., and Stoffel, M. (2001). Loss of HNF-1 Function in Mice Leads to Abnormal Expression of Genes Involved in Pancreatic Islet Development and Metabolism. Diabetes *50*, 2472–2480.

Shlyueva, D., Stampfel, G., and Stark, A. (2014). Transcriptional enhancers: from properties to genome-wide predictions. Nat Rev Genet *15*, 272–286.

Simaite, D., Kofent, J., Gong, M., Ruschendorf, F., Jia, S., Arn, P., Bentler, K., Ellaway, C., Kuhnen, P., Hoffmann, G.F., et al. (2014). Recessive Mutations in PCBD1 Cause a New Type of Early-Onset Diabetes. Diabetes *63*, 3557–3564.

Singh, R., and Valcárcel, J. (2005). Building specificity with nonspecific RNA-binding proteins. Nat Struct Mol Biol *12*, 645–653.

Smale, S.T., and Kadonaga, J.T. (2003). THE RNA POLYMERASE II CORE PROMOTER. Annu Rev Biochem *72*, 449–479.

Smemo, S., Tena, J.J., Kim, K.-H., Gamazon, E.R., Sakabe, N.J., Gómez-Marín, C., Aneas, I., Credidio, F.L., Sobreira, D.R., Wasserman, N.F., et al. (2014). Obesity-

associated variants within FTO form long-range functional connections with IRX3. Nature *507*, 371–375.

Smith, P.A., Sakura, H., Coles, B., Gummerson, N., Proks, P., and Ashcroft, F.M. (1997). Electrogenic arginine transport mediates stimulus-secretion coupling in mouse pancreatic beta-cells. J Physiology *499*, 625–635.

Song, M., Yang, X., Ren, X., Maliskova, L., Li, B., Jones, I.R., Wang, C., Jacob, F., Wu, K., Traglia, M., et al. (2019). Mapping cis-regulatory chromatin contacts in neural cells links neuropsychiatric disorder risk variants to target genes. Nat Genet *51*, 1252–1262.

Spitz, F., and Furlong, E.E.M. (2012). Transcription factors: from enhancer binding to developmental control. Nat Rev Genet *13*, 613–626.

Stadhouders, R., Filion, G.J., and Graf, T. (2019). Transcription factors and 3D genome conformation in cell-fate decisions. Nature *569*, 345–354.

Steijger, T., Abril, J.F., Engström, P.G., Kokocinski, F., Abril, J.F., Akerman, M., Alioto, T., Ambrosini, G., Antonarakis, S.E., Behr, J., et al. (2013). Assessment of transcript reconstruction methods for RNA-seq. Nat Methods *10*, 1177–1184.

Symmons, O., Uslu, V.V., Tsujimura, T., Ruf, S., Nassari, S., Schwarzer, W., Ettwiller, L., and Spitz, F. (2014). Functional and topological characteristics of mammalian regulatory domains. Genome Res *24*, 390–400.

Symmons, O., Pan, L., Remeseiro, S., Aktas, T., Klein, F., Huber, W., and Spitz, F. (2016). The Shh Topological Domain Facilitates the Action of Remote Enhancers by Reducing the Effects of Genomic Distances. Dev Cell *39*, 529–543.

Takahashi, H., Lassmann, T., Murata, M., and Carninci, P. (2012). 5′ end–centered expression profiling using cap-analysis gene expression and next-generation sequencing. Nat Protoc *7*, 542–561.

Thanos, D., and Maniatis, T. (1995). Virus induction of human IFNβ gene expression requires the assembly of an enhanceosome. Cell *83*, 1091–1100.

Thomsen, S.K., Raimondo, A., Hastoy, B., Sengupta, S., Dai, X.-Q., Bautista, A., Censin, J., Payne, A.J., Umapathysivam, M.M., Spigelman, A.F., et al. (2018). Type 2 diabetes risk alleles in PAM impact insulin release from human pancreatic β-cells. Nat Genet *50*, 1122 1131.

Thurman, R.E., Rynes, E., Humbert, R., Vierstra, J., Maurano, M.T., Haugen, E., Sheffield, N.C., Stergachis, A.B., Wang, H., Vernot, B., et al. (2012). The accessible chromatin landscape of the human genome. Nature *489*, 75–82.

Thurner, M., Bunt, M. van de, Torres, J.M., Mahajan, A., Nylander, V., Bennett, A.J., Gaulton, K.J., Barrett, A., Burrows, C., Bell, C.G., et al. (2018). Integration of human pancreatic islet genomic data refines regulatory mechanisms at Type 2 Diabetes susceptibility loci. Elife *7*, e31977.

Tillo, D., Kaplan, N., Moore, I.K., Fondufe-Mittendorf, Y., Gossett, A.J., Field, Y., Lieb, J.D., Widom, J., Segal, E., and Hughes, T.R. (2010). High Nucleosome Occupancy Is Encoded at Human Regulatory Sequences. Plos One *5*, e9129.

Tress, M.L., Abascal, F., and Valencia, A. (2017). Alternative Splicing May Not Be the Key to Proteome Complexity. Trends Biochem Sci *42*, 98 110.

Vaquerizas, J.M., Kummerfeld, S.K., Teichmann, S.A., and Luscombe, N.M. (2009). A census of human transcription factors: function, expression and evolution. Nat Rev Genet *10*, 252–263.

Vargas, D.Y., Shah, K., Batish, M., Levandoski, M., Sinha, S., Marras, S.A.E., Schedl, P., and Tyagi, S. (2011). Single-Molecule Imaging of Transcriptionally Coupled and Uncoupled Splicing. Cell *147*, 1054–1065.

Varshney, A., Scott, L.J., Welch, R.P., Erdos, M.R., Chines, P.S., Narisu, N., Albanus, R.D., Orchard, P., Wolford, B.N., Kursawe, R., et al. (2017a). Genetic regulatory signatures underlying islet gene expression and type 2 diabetes. Proc National Acad Sci *114*, 201621192 21.

Varshney, A., Scott, L.J., Welch, R.P., Erdos, M.R., Chines, P.S., Narisu, N., Albanus, R.D., Orchard, P., Wolford, B.N., Kursawe, R., et al. (2017b). Genetic regulatory signatures underlying islet gene expression and type 2 diabetes. Proc National Acad Sci *114*, 2301–2306.

Venter, J.C., Adams, M.D., Myers, E.W., Li, P.W., Mural, R.J., Sutton, G.G., Smith, H.O., Yandell, M., Evans, C.A., Holt, R.A., et al. (2001). The Sequence of the Human Genome. Science *291*, 1304–1351.

Viger, R.S., Guittot, S.M., Anttonen, M., Wilson, D.B., and Heikinheimo, M. (2008). Role of the GATA family of transcription factors in endocrine development, function, and disease. Mol Endocrinol Baltim Md *22*, 781–798.

Viñuela, A., Varshney, A., Bunt, M. van de, Prasad, R.B., Asplund, O., Bennett, A., Boehnke, M., Brown, A., Erdos, M.R., Fadista, J., et al. (2019). Influence of genetic variants on gene expression in human pancreatic islets – implications for type 2 diabetes. Biorxiv 655670.

Viñuela, A., Varshney, A., Bunt, M. van de, Prasad, R.B., Asplund, O., Bennett, A., Boehnke, M., Brown, A.A., Erdos, M.R., Fadista, J., et al. (2020). Genetic variant effects on gene expression in human pancreatic islets and their implications for T2D. Nat Commun *11*, 4912.

Visscher, P.M., Brown, M.A., McCarthy, M.I., and Yang, J. (2012). Five Years of GWAS Discovery. Am J Hum Genetics *90*, 7–24.

Visscher, P.M., Wray, N.R., Zhang, Q., Sklar, P., McCarthy, M.I., Brown, M.A., and Yang, J. (2017). 10 Years of GWAS Discovery: Biology, Function, and Translation. Am J Hum Genetics *101*, 5–22.

Vujkovic, M., Keaton, J.M., Lynch, J.A., Miller, D.R., Zhou, J., Tcheandjieu, C., Huffman, J.E., Assimes, T.L., Lorenz, K., Zhu, X., et al. (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. Nat Genet *52*, 680–691.

Wahl, M.C., Will, C.L., and Lührmann, R. (2009). The Spliceosome: Design Principles of a Dynamic RNP Machine. Cell *136*, 701–718.

Wainberg, M., Sinnott-Armstrong, N., Mancuso, N., Barbeira, A.N., Knowles, D.A., Golan, D., Ermel, R., Ruusalepp, A., Quertermous, T., Hao, K., et al. (2019). Opportunities and challenges for transcriptome-wide association studies. Nat Genet *51*, 592–599.

Wainschtein, P., Jain, D.P., Yengo, L., Zheng, Z., Consortium, Topm.A.W.G., Trans-Omics for Precision Medicine, Cupples, L.A., Shadyab, A.H., McKnight, B., Shoemaker, B.M., Mitchell, B.D., et al. (2019). Recovery of trait heritability from whole genome sequence data. Biorxiv 588020.

Walker, R.L., Ramaswami, G., Hartl, C., Mancuso, N., Gandal, M.J., Torre-Ubieta, L. de la, Pasaniuc, B., Stein, J.L., and Geschwind, D.H. (2019). Genetic Control of Expression and Splicing in Developing Human Brain Informs Disease Mechanisms. Cell *179*, 750-771.e22.

Wall, J.D., and Pritchard, J.K. (2003). Haplotype blocks and linkage disequilibrium in the human genome. Nat Rev Genet *4*, 587–597.

Wang, E.T., Sandberg, R., Luo, S., Khrebtukova, I., Zhang, L., Mayr, C., Kingsmore, S.F., Schroth, G.P., and Burge, C.B. (2008). Alternative isoform regulation in human tissue transcriptomes. Nature *456*, 470–476.

Wang, G., Sarkar, A., Carbonetto, P., and Stephens, M. (2020). A simple new approach to variable selection in regression, with application to genetic fine mapping. J Royal Statistical Soc Ser B Statistical Methodol.

Wang, L., Park, H.J., Dasari, S., Wang, S., Kocher, J.-P., and Li, W. (2013). CPAT: Coding-Potential Assessment Tool using an alignment-free logistic regression model. Nucleic Acids Res *41*, e74–e74.

Wang, X., Park, J., Susztak, K., Zhang, N.R., and Li, M. (2019). Bulk tissue cell type deconvolution with multi-subject single-cell expression reference. Nat Commun *10*, 453 9.

Weintraub, A.S., Li, C.H., Zamudio, A.V., Sigova, A.A., Hannett, N.M., Day, D.S., Abraham, B.J., Cohen, M.A., Nabet, B., Buckley, D.L., et al. (2017). YY1 Is a Structural Regulator of Enhancer-Promoter Loops. Cell *171*, 1573 1588.e28.

Whyte, W.A., Orlando, D.A., Hnisz, D., Abraham, B.J., Lin, C.Y., Kagey, M.H., Rahl, P.B., Lee, T.I., and Young, R.A. (2013). Master Transcription Factors and Mediator Establish Super-Enhancers at Key Cell Identity Genes. Cell *153*.

Xin, Y., Kim, J., Okamoto, H., Ni, M., Wei, Y., Adler, C., Murphy, A.J., Yancopoulos, G.D., Lin, C., and Gromada, J. (2016). RNA Sequencing of Single Human Islet Cells Reveals Type 2 Diabetes Genes. Cell Metab *24*, 608 615.

Xue, Y., Zhou, Y., Wu, T., Zhu, T., Ji, X., Kwon, Y.-S., Zhang, C., Yeo, G., Black, D.L., Sun, H., et al. (2009). Genome-wide Analysis of PTB-RNA Interactions Reveals a Strategy Used by the General Splicing Repressor to Modulate Exon Inclusion or Skipping. Mol Cell *36*, 996–1006.

Yamagata, K., Nammo, T., Moriwaki, M., Ihara, A., Iizuka, K., Yang, Q., Satoh, T., Li, M., Uenaka, R., Okita, K., et al. (2002). Overexpression of Dominant-Negative Mutant Hepatocyte Nuclear Factor-1 in Pancreatic -Cells Causes Abnormal Islet Architecture With Decreased Expression of E-Cadherin, Reduced -cell Proliferation, and Diabetes. Diabetes *51*, 114–123.

Yanai, I., Benjamin, H., Shmoish, M., Chalifa-Caspi, V., Shklar, M., Ophir, R., Bar-Even, A., Horn-Saban, S., Safran, M., Domany, E., et al. (2005). Genome-wide midrange transcription profiles reveal expression level relationships in human tissue specification. Bioinformatics *21*, 650–659.

Yang, Q., Yamagata, K., Fukui, K., Cao, Y., Nammo, T., Iwahashi, H., Wang, H., Matsumura, I., Hanafusa, T., Bucala, R., et al. (2002). Hepatocyte Nuclear Factor-1 Modulates Pancreatic -Cell Growth by Regulating the Expression of Insulin-Like Growth Factor-1 in INS-1 Cells. Diabetes *51*, 1785–1792.

Zabidi, M.A., Arnold, C.D., Schernhuber, K., Pagani, M., Rath, M., Frank, O., and Stark, A. (2015). Enhancer–core-promoter specificity separates developmental and housekeeping gene regulation. Nature *518*, 556–559.

Zaitlen, N., Paşaniuc, B., Gur, T., Ziv, E., and Halperin, E. (2010). Leveraging Genetic Variability across Populations for the Identification of Causal Variants. Am J Hum Genetics *86*, 23–33.

Zaret, K.S., and Carroll, J.S. (2011). Pioneer transcription factors: establishing competence for gene expression. Gene Dev *25*, 2227–2241.

Zhang, Y., Yang, H.T., Kadash-Edmondson, K., Pan, Y., Pan, Z., Davidson, B.L., and Xing, Y. (2020). Regional Variation of Splicing QTLs in Human Brain. Am J Hum Genetics.

Zhou, J., and Troyanskaya, O.G. (2015). Predicting effects of noncoding variants with deep learning–based sequence model. Nat Methods *12*, 931–934.

Zhou, J., Park, C.Y., Theesfeld, C.L., Wong, A.K., Yuan, Y., Scheckel, C., Fak, J.J., Funk, J., Yao, K., Tajima, Y., et al. (2019). Whole-genome deep-learning analysis identifies contribution of noncoding mutations to autism risk. Nat Genet *51*, 973–980.

Zhu, Z., Zhang, F., Hu, H., Bakshi, A., Robinson, M.R., Powell, J.E., Montgomery, G.W., Goddard, M.E., Wray, N.R., Visscher, P.M., et al. (2016). Integration of

summary data from GWAS and eQTL studies predicts complex trait gene targets. Nature Genetics *48*, 481–487.