Understanding the role of islet development in type 2 diabetes susceptibility



Marta Pérez Alcántara

Merton College

University of Oxford

A thesis submitted for the degree of

Doctor of Philosophy

Trinity Term 2019

To Pedro

I believe that when I die I shall rot, and nothing of my ego will survive. I am not young and I love life. But I should scorn to shiver with terror at the thought of annihilation. Happiness is nonetheless true happiness because it must come to an end, nor do thought and love lose their value because they are not everlasting. Many a man has borne himself proudly on the scaffold; surely the same pride should teach us to think truly about man's place in the world. Even if the open windows of science at first make us shiver after the cosy indoor warmth of traditional humanizing myths, in the end the fresh air brings vigour, and the great spaces have a splendour of their own.

Bertrand Russell's *What I believe*

Acknowledgements

I want to thank several people that have helped me to complete this thesis.

My supervisors, Prof. Mark McCarthy and Prof. Ben Davies for allowing me the privilege to work with you and for your support, guidance, encouragement and inspiring criticism.

"My" postdoc Agata Wesolowska-Andersen for helping me and sharing fun times in the office every day through these last two years. Previous postdocs Martijn van de Bunt and Nicola Beer for pushing me through the anxiety and feelings of inadequacy of the first months of my DPhil.

Juan Fernandez-Tajes, Anubha Mahajan, Jason Torres and Matthias Thurner for helping me with various analyses and giving me data and advice whenever I needed it.

Fernando Abaitua, Phalguni Rath, Katia Mattis, Nicole Krentz, Claire Duff, Antje Grotz, Grace Yu, and so many other members of the Davies and Gloyn groups for continuous support in the lab madness.

Lauren Wedekind for arranging out of the office times so I wouldn't go mad writing. Zammy Fairhurst-Hunter for conference escapades, nights out, writing help and yoga to keep me sane. The whole McCarthy group for so much fun and conference dinners to remember.

The anonymous people on ResearchGate and StackOverflow that answered hundreds of questions on lab work, genomics and programming throughout these years.

My parents Antonio and Yayo for your love and for giving me the opportunity to be where I am today. My brother Curro, Pilar and the rest of my family.

And Pedro Victori for giving me strength and unwaveringly believing that I am cut out for being a scientist. I wouldn't have finished these four years' worth of work without you by my side.

Chapter 2 of this thesis reproduces part of the data published in the article "Patterns of differential gene expression in a cellular model of human islet development, and relationship to type 2 diabetes predisposition" (Perez-Alcantara et al., 2018). Any work that is not wholly my own has been indicated in the relevant sections.

Abstract

Undestanding the role of islet development in type 2 diabetes susceptibility
Marta Perez Alcantara

Merton College
For the degree of Doctor of Philosophy
Trinity Term 2019

Most variants associated with type 2 diabetes (T2D) predisposition in genome-wide association studies (GWAS) act through defects in insulin secretion. In principle, these could result from deficiencies in islet development and/or mature islet function: most functional studies have focused on the latter. To explore the contribution of disturbed islet development to T2D pathogenesis, I characterised the transcriptome (using RNA-seq), accessible chromatin (ATAC-seq), H3K27ac histone marks (ChIP-seq) and DNA methylation patterns (using the 850k Illumina array) of three human iPSC lines (from different donors) differentiated in vitro across 7 stages (n=24) of pancreatic beta cell development.

Principal component analysis, enriched pathways and transcription factors regulating the transcriptome and DNA regulatory marks analysed indicate that the hiPSC-derived model from which the data originated recapitulates key aspects of beta cell development, and achieves more adult beta cells (though still immature) than previously-published differentiation models. Analysis of differentially expressed genes and regulatory marks within each stage indicated that both the most adult stage (beta-like cells) and less differentiated stages were enriched in T2D-associated signals, pointing to the contribution of both developing and adult islets to T2D risk. Novel potential regulators of islet development were identified (e.g. *REST* or *CUX2*), including genes within loci associated with T2D by GWAS (e.g. *PROX1*, *ONECUT1/HNF6*) and TFs whose binding was disrupted by T2D-associated variants (e.g. *CTCF, ZBTB33*). The DNA regulatory information was used to lever the posterior probabilities of association of variants within each T2D-associated locus, refining fine-mapping at these developmentally active regions. Integration of DNA regulatory marks with transcriptomic information facilitated the identification of active enhancers in islet development and their possible target genes, including candidate gene effectors of T2D-associated variants.

One of these candidate genes, *PROX1*, was functionally tested in islet development in the laboratory, aiming to modulate its expression using CRISPR activation, interference, and knock-out. *PROX1* was successfully upregulated in the undifferentiated hiPSC system using CRISPRa. Finally, it was possible to replicate the differentiation model in different cell lines (KOLFC2) than those that originated the *omics* data (SB). These new cell lines had the CRISPRa/i machinery integrated in the genome, providing more tools to study the effect of candidate genes in islet development and T2D.

List of abbreviations

- **ABC** Activity By Contact
- **ATAC-seq** Assay For Transposase-Accessible Chromatin sequencing
- **ATP** Adenosine Triphosphate.
- **BLC** Beta-like Cells
- **CDCV** Common Disease Common Variant
- **ChIP** Chromatin Immunoprecipitation
- **chromHMM** Chromatin Hidden Markov Model
- **CI** Confidence Interval
- **CRISPR** Clustered Regularly Interspaced Short Palindromic Repeats
- **CT** Cycle Threshold
- **dCas9** Deactivated Cas9
- **DE** Definitive Endoderm
- **DIAGRAM** Diabetes Genetics Replication And Meta-Analysis
- **DMR** Differentially Methylated Region
- **EN** Endocrine-like cells
- **ENCODE** The Encyclopedia of DNA Elements
- **ENGAGE** European Network For Genetic And Genomic Epidemiology
- **EP** Endocrine Progenitors
- **eQTL** Expression Quantitative Trait Loci
- **ESC** Embryonic Stem Cell
- **FACS** Fluorescence-Activated Cell Sorting
- **FAIRE-seq** Formaldehyde-Assisted Isolation of Regulatory Elements Sequencing
- **FC** Fold Change
- **FDR** False Discovery Rate
- **FE** Fold Enrichment
- **FG** Fasting Glucose
- **GCK** Glucokinase
- **GT** (Primitive) Gut Tube

- **GWAS** Genome-Wide Association Study
- **HDAC** Histone Deacetylase
- **HMR** High Methylated Region
- **IDR** Irreproducible Discovery Rate
- **indels** Insertions and Deletions
- **iPSC** Induced Pluripotent Stem Cell
- **LD** Linkage Disequilibrium
- **lincRNA** Long Intergenic Noncoding RNA
- **LMR** Low Methylated Region
- **MAF** Minor Allele Frequency
- **MODY** Maturity Onset Diabetes of the Young
- **mQTL** Methylation Quantitative Trait Loci
- **PBS** Phosphate Buffer Saline
- **PCA** Principal Component Analysis
- **PCR** Polymerase Chain Reaction
- **PE** Pancreatic Endoderm
- **PF** Posterior Foregut
- **PPA** Posterior Probability of Association
- **QC** Quality Control
- **qPCR** quantitative Real-Time PCR
- **SNP** Single Nucleotide Polymorphism
- **T1D** Type 1 Diabetes
- **T2D** Type 2 Diabetes
- **TAD** Topologically Associated Domain
- **TF** Transcription Factor
- **TFBS** Transcription Factor Binding Site
- **TSS** Transcription Start Site
- **WGBS** Whole-Genome Bisulfite Sequencing

List of contents

1 Introduction ……..……..…….. 1

1.1 Pathophysiology and epidemiology of type 2 diabetes ……. 1

1.2 Diabetes aetiology …….. 3

1.3 Genetics of T2D …….. 4

1.4 DNA regulatory marks in T2D ……..... 5

1.5 Pancreatic islets are central to T2D risk …… 6

1.6 Multi-*omics* integration ……...…….. 12

1.7 hiPSC-derived models of islet development to study diabetes pathogenesis…….....……...……... 15

1.8 CRISPR in functional genomics …….. 20

1.9 Approaches to characterizing *omics* data ……...…….. 21

1.10. Aims of this thesis ……...…….. 27

2 Generation and transcriptomic characterization of an hiPSC model of beta cell differentiation ……..……...…… 29

2.1 Introduction …… 29

2.2 Methods ……...…….. 32

2.3 Results ……... 38

2.4 Discussion …… 56

3 Chromatin modifications in an hiPSC model of beta cell differentiation …….. 58

3.1 Introduction ……... 58

3.2 Methods …… 61

3.3 Results ……..... 73

3.4 Discussion …… 118

4 Integration of transcriptomic and regulatory information in an islet development model ……..... 121

4.1 Introduction ……... 121

4.2 Methods …… 124

4.3 Results ……... 126

4.4 Discussion …… 148

5 Studying the role of a T2D-associated candidate gene in a beta cell differentiation model using CRISPR 152
5.1 Introduction 152
5.2 Materials and methods 155
5.3 Results 168
5.4 Discussion 186
6 General discussion 190
Appendices 198
Bibliography 230

1. Introduction

1.1. Pathophysiology and epidemiology of type 2 diabetes

Diabetes is a group of metabolic diseases characterised by high blood glucose. The most common form of the disease is type 2 diabetes (T2D), with over 90% of diabetes cases being classified as such (IDF, 2017). Hyperglycaemia in T2D is caused by defects in insulin secretion of the beta cells in the pancreatic islets of Langerhans and by insulin resistance of the peripheral tissues (Taylor, 2013). Insulin resistance is overcompensated by insulin production to maintain normoglycemia, but a steady increment in blood glucose levels in prediabetic patients over time reflects the inability of the beta cells to cope. Other diabetes forms include type 1 diabetes, characterized by a total or partial insulin secretion deficiency provoked by the auto-immune response against pancreatic beta cells. There are also monogenic, rare forms of diabetes, subclassified as neonatal diabetes, maturity-onset diabetes of the young [MODY] or syndromic diabetes. Unlike T2D patients, which are diagnosed at an average age of 57.8 ± 8.7 years in developed countries (Zoungas *et al.*, 2014), type 1 diabetes and monogenic diabetes patients tend to receive a diagnosis when they are young.

A person is considered to have diabetes when the fasting plasma glucose increases from the healthy range of 4.0-7.0 mM over 7.0 mM. Hyperglycaemia in diabetes patients initially manifests in polydipsia (excessive thirst), polyuria (abnormally high production of urine), and weight loss. In the most severe cases, diabetes causes retinopahty (leading to blindness), nephropathy (reduced kidney function), peripheral neuropathy (nerve damage in feet and hands) and cardiovascular problems including necrosis of the limbs (ADA, 2018). These conditions can be life-threatening if left untreated.

Over 450 million adults worldwide suffer from some form of diabetes, with an age-adjusted prevalence of 8.6%, almost half of which are undiagnosed. This translated in 5

million deaths of individuals aged 20-99 attributed to the disease, and estimated healthcare costs of over 850 billion dollars in 2017 (Cho *et al.*, 2018). The number of diabetes patients around the world is predicted to reach 642 million in 2040, with an increased age-adjusted prevalence in almost every country (**Figure 1.1**). Prevalence has steadily increased in the last 50 years, promoted by lifestyle changes such as the consumption of hypercaloric and hyperglycaemic meals and lower activity. This obesogenic environment drives up the frequency of overweight individuals in the population, who in turn are at a larger risk of developing T2D.

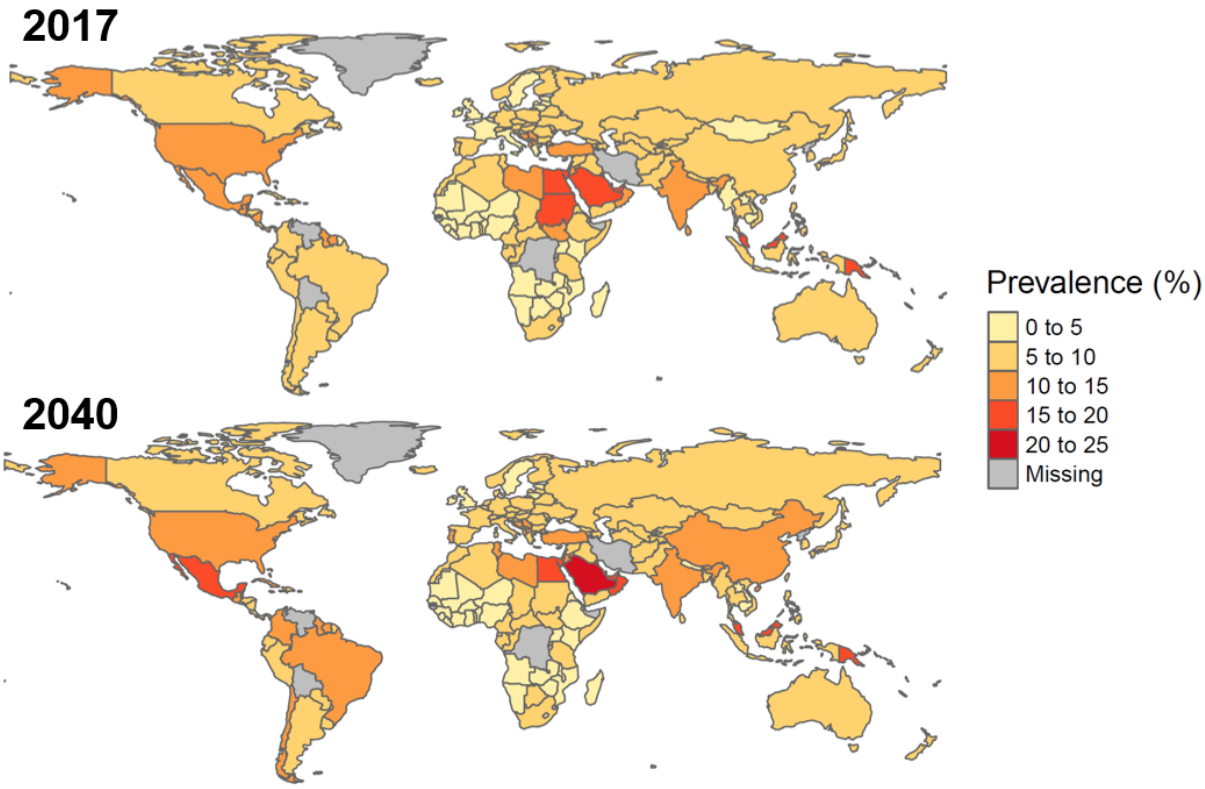


Figure 1.1. Diabetes is a global epidemic. Age-adjusted prevalence per country in 2017 and estimates for 2040. Global age-adjusted prevalence in adults (ages 20-99): 8.6% in 2017 and 9.1% in 2040 (Cho *et al.*, 2018). Data provided by and reproduced with the permission of the International Diabetes Federation.

1.2. Diabetes aetiology

The insulin hormone reduces the concentration of glucose in the blood by mediating its uptake by the peripheral tissues, either for immediate use as an energy source (e.g. in skeletal muscle) or for storage as glycogen in the liver (Aronoff *et al.*, 2004). The insulin sensitive tissues then provide feedback to the beta cells, which in turn regulate their production of the hormone. As I mentioned above, T2D is caused by insulin resistance of the peripheral tissues and decreased insulin secretion by the beta cells (Taylor, 2013), although the contribution of both processes to the syndrome varies from patient to patient (Zheng *et al.*, 2018). Insulin resistant tissues have reduced glucose uptake from the blood, which in turn increments insulin production by the beta cells to try to maintain glucose homeostasis. As insulin insensitivity increases, so does insulin production until it reaches a physiological limit. Blood glucose increases as a consequence, in severe cases developing into T2D when the function of beta cell starts to deteriorate (Skyler *et al.*, 2017).

Beta cells are able to sense blood glucose and respond producing and releasing the needed insulin in a process called Glucose-Stimulated Insulin Secretion (GSIS). Glucose enters the cell via GLUT2 transporters and is converted to Glucose-6-phosphate by the glucokinase, which in turn is transformed to ATP in the mitochondria as part of the glycolysis. The increase in ATP closes ATP-sensitive potassium channels, which leads to depolarization of the membrane and entrance of calcium into the beta cells via the calcium channels. This influx of calcium finally triggers insulin exocytosis (Meda and Schuit, 2013).

T2D is a complex disease, influenced by both environmental and genetic factors. Important risk factors towards acquiring insulin sensitivity are excessive caloric intake and inactivity, which in turn cause obesity and inflammation. These can be treated with lifestyle changes like diet and exercise, or in more severe cases with medication like metformin, which reduces gluconeogenesis in the liver and increases insulin sensitivity

by yet unclear mechanisms (Rena *et al.*, 2017). These environmental factors also impair beta cell function by unclear causes that might involve beta cell death or de-differentiation, leading to reduced GSIS (Kolb and Martin, 2017). Once beta cell damage has occurred it is difficult to regenerate their function, although new strategies for endogenous beta cell proliferation and differentiation from iPSCs *in vitro* have emerged in recent years (Zhou and Melton, 2018).

1.3. Genetics of T2D

Type 2 diabetes is a polygenic disease, with 243 loci (403 SNPs) associated with it so far resulting from multiple genome-wide association studies (GWAS) (Mahajan *et al.*, 2018). GWAS efforts have been driven by the common disease common variant (CDCV) hypothesis: that the genetic susceptibility to develop diseases that are prevalent among humans is caused, for each phenotype, by multiple genetic variants with low penetrance present at high frequency in the population (Reich and Lander, 2001). GWAS can establish causal inference because the analysed exposure variables (the common variants in the genome) are not confounded or influenced by external factors, so the mechanistic causality flows only in the genotype to phenotype direction, even though a variant could affect multiple phenotypes. They also have been more productive than candidate gene studies in detecting high-frequency genetic variants of low and medium effect in disease risk. Still, T2D-associated GWAS variants explain 18% of the chip heritability (Mahajan *et al.* 2018a), about half of the heritability estimates from twin studies, as is common in complex phenotypes. This “missing heritability” (Manolio *et al.*, 2009), does not seem to be hiding in rare variants in the case of T2D: whole genome sequencing of 2,657 Europeans diabetes patients and controls and exome sequencing in 12,940 individuals with varied ancestry did not find that rare variants contributed a large fraction of the heritability, even after increasing the sample size by imputation of 11,645 T2D cases and 32,769 controls (Fuchsberger *et al.*, 2016). Exome sequencing of 20,791 cases and 24,440 controls of T2D found a similar result (Flannick *et al.*, 2019).

Teasing out the biological mechanisms underlying the association signals has proven difficult, since most GWAS hits fall outside coding regions and often do not signal specific variants but a broader region due to linkage disequilibrium. Some functional T2D variants had been identified so far, including several coding variants at the *GCRK* and *SLC30A8* loci (Beer *et al.*, 2009, Flannick *et al.*, 2014). In other variants, their high posterior probabilities in the 99% credible sets (that is, the set of variants that account for 99% of the posterior probability of driving the association at each locus) make these good candidates for follow-up analyses (Mahajan *et al.* 2018a, Mahajan *et al.* 2018b).

1.4. DNA regulatory marks in T2D

To understand how the non-coding variants might be contributing to disease it is essential to identify and understand regulatory regions. Since most cells in the body from the point of conception, save somatic mutations, share identical DNA sequence, developmental differences between tissues and cell types are due to other processes. These are DNA regulatory marks, which include DNA methylation, non-coding RNA, and histone modifications. Often their remodelling result in changes in nucleosome occupancy and 3-D contacts which in turn could modify the availability of transcription factor binding sites at enhancers or promoters, with concomitant alterations in gene expression networks, although the directionality between these processes is not established (Kilpinen and Dermitzakis, 2012). Recent endeavours have started annotating these sequences, which have been linked to disease (The ENCODE Project Consortium, 2007 & 2012; Roadmap Epigenomics Consortium, 2015). Their influence in T2D has been demonstrated in several studies that showed altered regulatory marks at the promoters and enhancers of genes regulating insulin secretion in diabetes patients compared to healthy individuals (Yang *et al.*, 2011; Dayeh *et al.*, 2014). Information on these tissue-specific regulatory regions can also be incorporated, recalculating the posterior probabilities of T2D-associated variants within the 99% credible sets and improving fine-mapping efforts (Mahajan *et al.*, 2018a).

1.5. Pancreatic islets are central to T2D risk

While insulin secretion and insulin resistance are both part of the aetiology of T2D, a large fraction of the genetic variation affects the former, highlighting the importance of the pancreatic islet (Scott *et al.*, 2007; Zeggini *et al.*, 2007; Voight *et al.*, 2010; Pasquali *et al.*, 2014; Dimas *et al.*, 2014; Wood *et al.*, 2017). Most of the analysis of the contribution of islets to T2D risk have been performed in adult islets, but there is evidence that some of the signals might influence islet (and particularly beta cell) development (Travers *et al.*, 2013). The clearest evidence comes from monogenic forms of diabetes, caused by penetrant mutations in genes which mostly impact islet development (Schwitzgebel *et al.*, 2014). An increasingly larger number of these genes also reside in T2D-associated loci (**Figure 1.2**) (Fuchsberger *et al.*, 2016). Furthermore, several studies have observed that islet regulatory elements are enriched in T2D associated variants (Thurner *et al.*, 2018; Parker *et al.*, 2013; Pasquali *et al.*, 2014), highlighting the importance of these regulatory processes and of islet biology in T2D risk. In the following pages I will detail how previous studies, using transcriptomics and DNA regulatory information, have come to this conclusion.

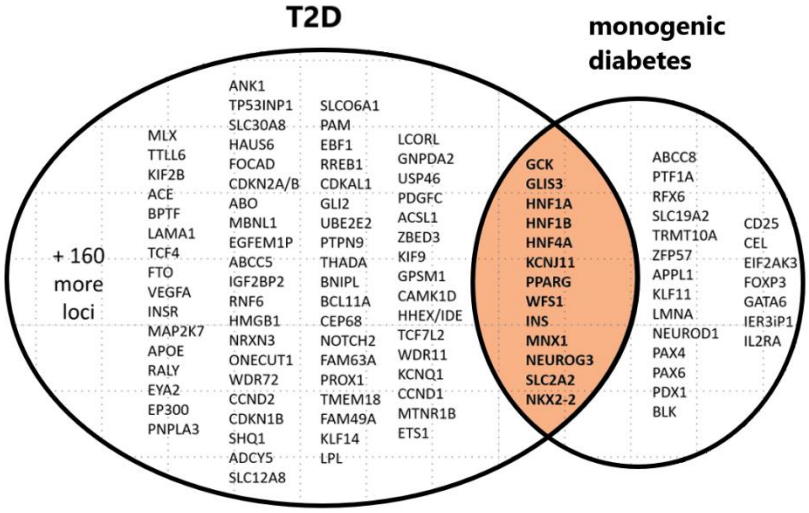


Figure 1.2. Overlap between T2D-associated loci and causal genes for monogenic diabetes.T2D loci names are those genes nearest to the GWAS index SNP, which may not be the causal genes. Adapted from Florez: *The genetics of T2D and related traits* (2016).

1.5.1. Transcriptomics analysis in islet biology and T2D

Sequence variants detected by GWAS can be tested for significant associations to gene expression changes, discovering expression quantitative trait loci (eQTL) (Morloy *et al.*, 2004, GTEx Consortium, 2015 & 2017). Fadista and colleagues (2014) showed how RNA-seq from 89 human islets could be used to discover 640 cis-eQTLs (eQTLs influencing genes within 250kb), five of which were GWAS SNPs associated with T2D or glycaemic traits. Van de Bunt and colleagues (2015) mapped eQTLs within 14 T2D-associated loci in islets, providing evidence for effector genes at these regions. Experimental follow-up work on one of these candidates (*ZMIZ1*), modifying its expression in dispersed human islets and beta cells, caused altered insulin secretion.

Differential expression analyses in relation to T2D have been performed interrogating particularly islet biology. Nica and colleagues (2013) compared the transcriptomes of 7 whole islets, 11 beta cell preparations and 5 beta cell-depleted samples. They detected over 500 genes uniquely expressed in beta cells compared to other islet subtypes, which were enriched in neuronal pathways, possibly reflecting the common evolutionary origin of neurons and beta cells (Atouf *et al.*, 1997; Schwartz *et al.*, 2010; Ohta *et al.*, 2011).

Co-expression networks have been used in T2D in the study by Keller and colleagues (2008) in which they compared gene expression patterns of mice strains susceptible to diabetes vs controls before and after they induced diabetes by making them obese. The authors identified co-expressed genes enriched in cell cycle pathways that were correlated with islet proliferation and were predictors of diabetes.

Micro RNAs (miRNAs) and long noncoding RNAs (lncRNAs), whose discovery and study has been facilitated by moving from microarrays to RNA-seq, have both been highlighted as important regulatory elements and have been linked to a variety of diseases (Krol *et al.*, 2010). miRNAs are known to repress gene expression, and van de Bunt and colleagues (2013) analysed the influence of miRNAs acting on T2D-associated genes in human islets and beta cells, showing how predicted targets of miRNAs in islets were moderately enriched in T2D-associated signals ($p=0.01$, $q=0.1$), including six loci

(*AP3S2, KCNK16, NOTCH2, SCL30A8, VPS26A*, and *WFS1*) where they overlapped potential causal variants.

Concerning lncRNAs, Morán and colleagues (2012) investigated their role in in T2D and beta cell development sequencing purified human islets. They defined T2D genes using the tool MAGENTA (allocating association signals to likely effector genes considering their size and distance to the variants) and found that lncRNAs were enriched for association with these candidates ($p = 0.02$), including in 6 loci that have been linked to beta cell dysfunction (*PROX1, WFS1, HNF1A, KCNJ11, ARAP1* and *HNF4A*). They observed that islet-specific lncRNAs were found preferably nearby protein-coding genes important for islet function, such as *MAFB, FOXA2, RFX6* or *ISL1*. They also showed dynamic variation of lncRNA levels during differentiation of a hiPSC-derived beta cell differentiation model, highlighting them as possible candidate regulators of beta cell development.

Despite efforts to isolate specific tissues or cell types (e.g. by cell surface markers using FACS), there is inevitably a great heterogeneity captured by bulk RNA-seq. The trend in transcriptomics is moving towards single cell RNA-seq to uncover cell subtypes and biology that might have been muddled in previous analyses. Xin and colleagues (2016) performed single-cell RNA-seq on 1,492 islet cell subtypes: alpha, beta, delta and pancreatic progenitors from 12 non-diabetic and 6 T2D donors. They identified 245 differentially expressed genes between cases and controls, the majority of which (92%) have not been previously implicated in cell function or growth. Only 18 of them were in loci previously associated with T2D by GWAS. After comparing with single-cell RNA-seq data from mouse islets, they concluded that mice and human differ in a few key genes including MAFB and SOX4, which were completely absent in the murine model. This highlights the need of human endocrine pancreas models for study of islet differentiation. In the same line, Lawlor and colleagues (2017) performed RNA-seq on a smaller set of single cells (638) from the same islet subtypes, comparing the transcriptomes of 5 non-diabetic and 3 T2D donors. They identified a total of 410 differentially expressed genes between T2D and non-diabetic patients, which in delta cells made for 90% of previously

undetected differentially expressed transcripts. In contrast, a larger study of bulk RNA-seq in 84 non-diabetic and 19 T2D islets gave only 19 differentially expressed genes (Solimena *et al.*, 2017). This represents an example of the potential of single-cell analysis, and how failure to correctly stratify cell types hides potentially important biological differences between individuals affected by T2D and controls.

1.5.2. DNA methylation analysis in islet biology and disease

Since most variants associated to complex diseases are located in intergenic regions, it has been proposed that they are acting through regulatory mechanisms which include variation in methylation levels. In fact, regions associated with several diseases by GWAS are enriched in differentially methylated regions between healthy and affected individuals (Ziller *et al.*, 2013; Bell *et al.*, 2016) or in methylation quantitative trait loci (mQTL) (Gamazon *et al.*, 2013), often in disease-relevant tissues and cell types (Ziller *et al.*, 2013).

Changes in T2D risk might be partially mediated by the effect of disease-associated variants in methylation levels, since around half of T2D-associated SNPs cause alterations in CpG sites susceptible to methylation (Dayeh *et al.*, 2013; Elliott *et al.*, 2017). In an analogous way to eQTLs, several studies have identified methylation quantitative trait loci, which in turn could be affecting gene expression (Wagner *et al.*, 2014) and other epigenetic marks such as histone modifications (Banovich *et al.*, 2014). Olsson and colleagues (2014) performed the first mQTL analysis in human pancreatic islets, linking 1,465 SNPs to 383 CpG sites in 89 human islets. 12 mQTLs acting in cis mapped to T2D-associated SNPs, in loci such as *PDX1, ADCY5, KCNJ11* and *INS*, which harbour important genes in islet development and function.

The importance of the developing pancreatic islet in T2D is highlighted by the study of Travers and colleagues (2013): genes in the T2D-associated *KCNQ1* locus show imprinting in the developing (but not adult) islets, likely mediating changes in T2D risk

through effects of the maternally inherited allele pre-birth (Kong *et al.*, 2009). A comprehensive analysis (480K CpGs) of the methylation differences between islets of diabetic (n=15) and non-diabetic patients (n=34) showed decreased methylation in 97% of differentially methylated CpGs after FDR filtering. Another islet study found that most of the differentially methylated regions were found in “open sea”, intergenic regions away from CpG islands, overlapping 17 T2D-associated loci including *FTO, KCNQ1, IRS1* and *TCF7L2* (Dayeh *et al.*, 2014).

1.5.3. Accessible regions in islet biology and disease

Early evidence of the importance of chromatin accessibility in islet biology comes from the era before genomic studies were available. Chakrabarti and colleagues (2002) found limited binding, measured by chromatin immunoprecipitation (ChIP), of the essential pancreatic progenitor TF Pdx1 to promoters of selected genes *in vivo* (transfecting Pdx1 into the cells) compared to *in vitro* (measuring Pdx1 binding to promoter fragments), indicating that chromatin inaccessibility could impede binding at unavailable promoters.

Soon genome-wide approaches started to map the accessible landscape of pancreatic islets. Gaulton and colleagues (2010) used FAIRE-seq to identify regions of open chromatin in 3 samples of human islets and found approx. 80,000 peaks in 3,000 islet-specific clusters of accessible chromatin when compared to 5 other cell types. These clusters harboured multiple binding sites of key islet TFs: *MAFB, FOXA2, PDX1* and *NKX6-1.* When they co-localised T2D and fasting glucose (FG)-associated variants in these clusters they found 38 SNPs at 8 loci. One of these was rs7903146 from *TCF7L2*, the locus with the strongest association with T2D (Grant *et al.*, 2007; Saxena *et al.*, 2006; Cauchi *et al.*, 2007, Mahajan *et al.*, 2018a). Analysis of 31 islet samples found strong allelic imbalance of chromatin accessibility at this variant, and using a luciferase assay the authors discovered that it affects enhancer activity. A similar study by Stitzel and colleagues (2010) profiled accessible regions in four islets with DNase-seq and also combined H3K4me1, H3K4me3 and H3K79me2 histone marks and CTCF binding. They

identified over 100,000 DNase I peaks, and comparison with other four human cell lines revealed that approx. 35% were islet-specific. rs7903146 from *TCF7L2* showed again an allelic effect on enhancer activity in islets.

Recent studies have found more T2D-associated variants and at the same time have improved their resolution using fine-mapping. Thurner and colleagues (2018) used this genetic data and epigenetic annotations to refine causal variants. They found that the number of SNPs that had been fine-mapped in each 99% credible set could be reduced incorporating just ATAC-seq peak annotations. Then they tested for allelic imbalance of these SNPs in open chromatin and found three variants showing significant effects at *CDC123* / *CAMK1D*, *KLHDC5* and *ADCY5* loci, all located within putative enhancers.

1.5.4. Histone marks in islet biology and T2D

To test the role of histone acetylation in pancreas development, Haumaitre and colleagues (2008) inhibited histone deacetylases (HDACs) in the differentiating pancreas of rats, culturing explants of pancreatic buds in vitro. They compared the global acetylation levels of histones H3 and H4 to controls, finding that treatment with inhibitors of class I and II histone deacetylases upregulated beta cell development. Meanwhile, treatment with inhibitors of only class I HDACs reduced beta cell mass and promoted differentiation of alpha cells and pancreatic progenitors. Class II HDAC expression is restricted to certain cellular subtypes and interacts with specific transcription factors, which partially explains the difference in treatment outcomes, despite general increment of endocrine progenitors (Ngn3+ cells) after treatment with both kinds of HDACs inhibitors. This work shows the importance of global patterns of histone acetylation in islet development.

Histone methyltransferases also play an important role in islet development. Ezh2 is part of the Polycomb Repressive Complex 2 (PRC2), which has histone methyltransferase activity and catalyzes the transference of methyl groups to H3K27 (Aloia *et al.*, 2013).

This H3K27me3 repressive mark is found often near developmental regulators (Lee *et al.*, 2006) and consequently van Arensbergen and colleagues (2010) profiled this mark during mouse pancreas development. They found that H3K27me3 repressive mark is lost from promoters of important genes in pancreatic progenitors and subsequent cells of the endocrine lineage including beta cells (eg. *Pdx1, Foxa2, Neurod1*). Later studies by Xu and colleagues (2011 & 2014) in which they deleted *Ezh2* in mouse posterior foregut showed an expansion of cells from the pancreatic versus the hepatic lineage, while deletion in pancreatic progenitors increased the number of beta cells *in vivo* and *in vitro*. Conversely, adult mice with a beta cell-specific deletion of *Ezh2* develop mild diabetes and have reduced insulin production due to a decrease in beta cell mass (Chen *et al.*, 2009). All these studies show how global patterns of histone modifications control the balance between the pancreatic and hepatic cell fate and beta cell proliferation.

1.6. Multi-*omics* integration

1.6.1. Multi-*omics* interactions in systems biology

Biological systems are subject to complex relationships between their various components. DNA sequence, epigenetic modifications of genome and histones, gene expression, protein levels and other layers of molecules contribute to the emergent properties of cells and organisms. Technological advances have made the simultaneous study of the various -*omics* more available, allowing for the analysis of their interactions – the so called “systems biology”. The relationships between the different datasets are so complex that often inference across them is not accurate. For example, comparison of transcriptomics and proteomics data has discovered that the linear correlation between protein abundance and gene expression is not high (Pearson’s r = 0.4), due to post-transcriptional and translational regulation (Vogel *et al.*, 2012). At the epigenetic level, a study of 125 cell types discovered that only 40% of sites with variable methylation also presented chromatin accessibility variability, >97% of which presented lower accessibility with increased DNA methylation. This means that decreased methylation

does not always imply increased accessibility, which can be negatively regulated by other means such as histone modifications (Thurman *et al.*, 2012).

The effect of a genetic variant varies depending on the genomic and DNA regulatory environment, so the different layers of data can be integrated to find the link between the variants associated with disease and the gene regulatory networks that underlie the changes in risk (Ritchie *et al.*, 2015). These kind of multi-level analyses also make up for potential loss of information from one of the *-omics* datasets, providing several layers of evidence that point to the same molecular pathways. On the other hand, they require complex statistical analyses as there are usually a large number of variables to model and little replicates, and are difficult to visualize.

1.6.2. Multi-*omics* analysis in islet biology and T2D

Integrative *-omics* approaches have been applied to the task of linking disease-associated variants to their biological effectors, including in T2D. QTL studies highlighted previously would be a simple integration model of two datasets, genomic and transcriptomic or DNA regulatory data, but as the technologies become more affordable more layers of information are being analysed at the same time. Combination of RNA-seq, ChIP-seq and histone marks can reveal the role of enhancers and transcription factors in cell identity. Pasquali and colleagues (2014) immunoprecipitated five TFs important for beta cells and other islet cell types (FOXA2, MAFB, NKX2-2, NKX6-1 and PDX1) and found that they co-localized with islet-specific enhancer clusters, while their binding to promoters is not associated to cell-type specific transcription events. Integration with RNA-seq data indicated that the islet enhancers were directing the expression of islet-specific genes. Interestingly, a median of 3 TFs bound at these islet-specific active enhancer clusters, showing how the interplay of several TFs is important for enhancer function.

An integrated way of leveraging multiple pieces of epigenomic information is to build chromatin state maps using a chromatin hidden Markov model (ChromHMM) (Ernst and Kellis, 2012). This uses machine learning to capture the overlaps of diverse DNA regulatory marks (histone modifications, chromatin accessibility and methylation) at each genomic position to create meaningful biological annotations, for example to classify regions as active or poised Transcription Start Sites (TSS), and can serve to clarify the functional effect of variants associated with disease (Wang et al., 2016). Thurner and colleagues (2018) used this approach to build high-resolution chromatin maps of human islets, combining whole-genome bisulfite sequencing (WGBS), ATAC-seq and histone annotations. T2D-associated GWAS signals were enriched in enhancer signals, but only those that showed a combination of H3K4me1 peaks with H3K27ac or ATAC-seq peaks, marks of strong and open enhancers. Using these marks to refine fine-mapping of T2D signals helped reduce the number of potential variants per locus, including at *ADCY5* and *CDC123*, which are implicated by this study in islet enhancer activity. These results are in line with the study of Parker and colleagues (2013), that also used this approach but combining exclusively histone marks from 10 cell types, showing an enrichment of T2D signals in larger-than-average, islet specific “stretch enhancers”. RNA-seq data linked the activity of said enhancers to nearby genes in a tissue-specific fashion. The ChromHMM approach has been used in the same way in a recent study of 112 islets by Varshney and colleagues (2017), showing that islet-specific cis-eQTLs are also enriched in islet stretch enhancers and that footprints from transcription factors generated from ATAC-seq data point to strong activity of RFX TFs at T2D GWAS loci.

1.6.3. Integrating *-omics* data to link enhancer activity to gene expression

If most variants associated by GWAS to disease are in intergenic regions, to have an effect in risk they must affect regulatory processes, and it is likely that they often fall in enhancers. Often mapping the action of these regions to target genes is not straightforward as they can influence multiple genes (and, at the same time, one gene

can be affected by multiple enhancers), often not the ones located closest to them, and their activity changes under different stimuli and across different cell types. Therefore, the classic mapping of enhancer activity onto strictly the closest genes is inaccurate.

Chromatin conformation information can be incorporated to multi-*omics* analyses to try to map these interactions (Krivega *et al.*, 2012; Sexton *et al.*, 2015). For example, Greenwald and colleagues (2019) generated Hi-C chromatin maps in three islets and used publicly available ATAC-seq and histone ChIP-seq data to generate chromatin maps with ChromHMM. Active promoters and enhancers were most enriched for interactions within chromatin loops. Using them to map proximal and distal islet enhancers to their target promoters, they observed that genes linked to multiple enhancers had higher expression (measured by RNA-seq) on average, and that subset included genes important for islet function (e.g. *MAFB*, *FOXA2, NKX6-1*). Islet enhancers encompassed variants at 30 T2D-associated loci which could be linked to a reduced set of gene candidates, including *CAMK1D, IGF2BP2,* and *ABCC8*. Laboratory validation showed that inactivating the homolog of *IGF2BP2* in mouse beta cells caused impaired insulin secretion with high fat diet.

1.7. hiPSC-derived models of islet development to study diabetes pathogenesis

1.7.1. Human endocrine pancreas development

Current understanding of islet development has primarily been derived from studying murine models. However, there are several core differences in the processes involved in development between human and mice (Jennings *et al.*, 2015) and in adult islets both at the level of protein-coding gene expression and regulatory lncRNAs (Benner *et al.*, 2014) and in the macroscopic organization of islet subtypes, the exocrine tissue and the circulatory system that irrigates the organ (Dolensek *et al.*, 2015). These findings motivate the endeavours to differentiate human embryonic-derived and, more recently,

induced pluripotent stem cells (iPSC) into adult pancreatic islets (D'Amour *et al.*, 2005; D'Amour *et al.*, 2006; Bruin *et al.*, 2014; Pagliuca *et al.*, 2014; Rezania *et al.*, 2014; Xie *et al.*, 2013; Russ *et al.*, 2015; Nostro *et al.*, 2015; van de Bunt *et al.*, 2016; Perez-Alcantara *et al.*, 2018; Nair *et al.*, 2019).

The islets contain diverse hormone-producing cells including beta, alpha, delta, gamma and epsilon cells; which respectively secrete insulin, glucagon, somatostatin, pancreatic polypeptide and ghrelin hormones into the bloodstream. The islets are scattered among the exocrine pancreas that produces digestive enzymes. The pancreas starts to form from the endodermal layer, folding into the primitive gut tube within the first month of embryonic development (25-27 days post conception [dpc]). This tube divides into foregut, midgut and hindgut; it will be the foregut that will differentiate into the pancreatic fate, starting to form ventral and dorsal buds of pancreatic endoderm at 29-31 dpc, and into the liver. From the endoderm, pancreatic progenitors grow forming protuberances of tip and trunk domains shortly after entering the first month post conception. The pancreatic progenitors in the trunk will give rise to endocrine cell subtypes in islets thanks to the transient expression of *NEUROG3*. Conversely, those in the tip will differentiate into acinar cells (Pan and Wright 2011; Jennings *et al.*, 2015). Markers of each stage and key genes involved in tissue specification are shown in **Figure 1.3.**

1.7.2. iPSC-derived models of pancreas development

Perfecting in vitro models of iPSC-derived endocrine pancreas development would enable investigations of loci influencing islet differentiation without using embryonic tissue from either humans or animals. iPSCs can be reprogrammed from somatic cells by inducing the expression of the so-called "Yamanaka factors": *OCT4/POU5F1*, *SOX2*, *KLF4*, and *MYC* (Takahashi *et al.*, 2006). These cells then acquire self-renewal capacity and may be differentiated into any type of tissue, including pancreatic beta cells.

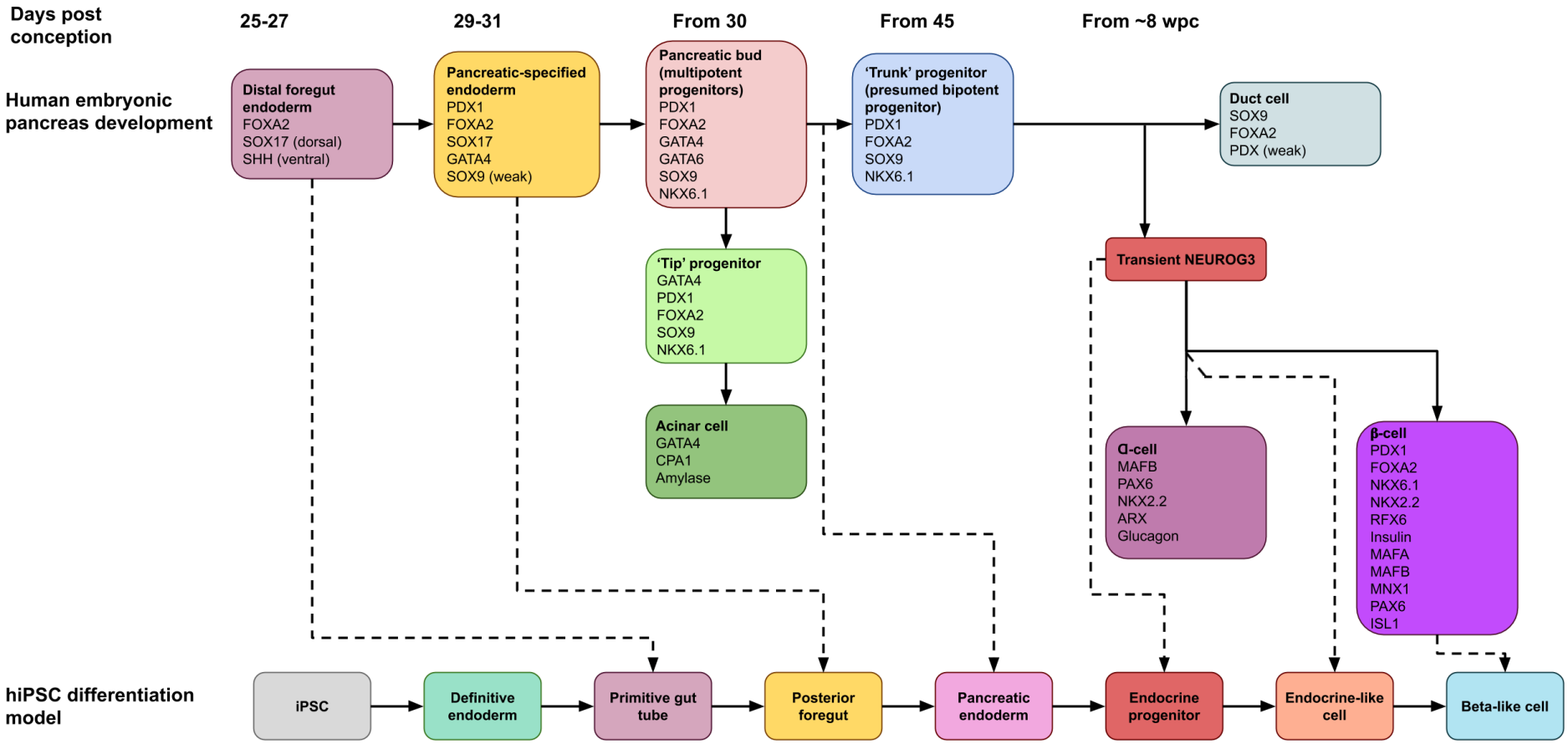


Figure 1.3. Human pancreas development. Schematic with key steps and genes that take part in human pancreas differentiation (top), with marker genes expressed, and comparison with the beta cell differentiation model used in this thesis (bottom). Dotted lines link stages of development *in vitro* and *in vivo* that are approximately analogous. Adapted from Jennings and colleagues (2015). wpc: weeks post conception.

Beta cell differentiation protocols are constantly improving, achieving endocrine cells with increasing insulin secretory capacity. **Figure 1.3** compares the differentiation model used in this thesis with current knowledge of the steps of pancreas development (although the stages of development *in vitro*, defined by the expression of marker genes, do not exactly correspond to those *in vivo*). These models are developed by screening for molecules that modulate the genes and pathways observed in *in vivo* human and mouse pancreas development. The study of D'Amour and colleagues (2005) was the first to differentiate in vitro human embryonic stem cells to the definitive endoderm thanks to the action of activin. Further research achieved efficient differentiation towards PDX1-positive pancreatic progenitors by addition into the media of retinoic acid, SHH inhibitors and Fgf10 (D'Amour *et al.*, 2006; Kroon *et al.*, 2008; Rezania *et al.*, 2012).

From this point, research has focused on trying to achieve cells that exclusively secrete insulin and that can modulate secretion in response to varying concentrations of glucose (Bruin *et al.*, 2014; Pagliuca *et al.*, 2014; Rezania *et al.*, 2014; Xie *et al.*, 2013; Russ *et al.*, 2015; Nostro *et al.*, 2015; van de Bunt *et al.*, 2016; Perez-Alcantara *et al.*, 2018; Nair *et al.*, 2019). For that, it is essential to start from an abundant population of PDX1 and NKX6-1 positive cells. Pancreatic progenitors transplanted into rodents also improved beta cell function and identity, showing that there are environmental cues in the pancreas that are not yet understood (Xie *et al.*, 2013). A 3-D environment replicating *in vivo* disposition of cells in the pancreatic trunk also seems to improve beta cell differentiation (Pagliuca *et al.*, 2014; Nair *et al.*, 2019). Variability of differentiation capacity between cell lines (Kajiwara *et al.*, 2012; Rouhani *et al.*, 2014) and epigenetic memory (Kim *et al.*, 2011; Bar-Nur *et al.*, 2011; Thurner *et al.*, 2017) further complicates this process.

Despite the limitations, iPSC-derived beta cells can already be used as disease modelling platforms by studying the associated variants in disease, particularly with the help of recent advances in genome editing (Beer and Gloyn, 2016). Cell culture from iPSC-derived beta cells may also serve as a cost-effective platform for drug screening in diabetes treatment research, and could provide material for transplants for diabetic

patients, something that has been tried with promising results in mice (Pagliuca *et al.*, 2014; Rezania *et al.*, 2014; Millman *et al.*, 2016).

1.7.3. *Omics* analysis of iPSC-derived islet differentiation models

Several studies presenting updates in beta cell differentiation protocol made use of *omics* data, particularly transcriptomics, to characterize differentiating cells and specifically the expression of markers of endocrine pancreas at each stage. Comparison between the differentiation models and the developing pancreas of human embryos is essential to discover new factors directing the process and applying this knowledge to improve the current models. Cebola and colleagues (2015) did this comparing RNA-seq, H3K27ac and H3K4me1 marks of multipotent pancreatic progenitors to tissue samples from human pancreatic buds. The combination of histone marks allowed them to map active enhancers present in the *in vivo* differentiation, which were enriched for TEAD binding sites, a transcription factor involved in cell proliferation (Lin *et al.*, 2017). ChIP-seq of this TF and disruption of TEAD function in pancreatic progenitors grown *ex vivo* showed altered expression of genes regulating progenitor cell growth and significantly limited the size of the pancreatic buds compared to untreated controls (Cebola *et al.*, 2015).

Xie and colleagues (2013) performed this kind of analysis on hESC differentiated towards beta cells in vitro until the pancreatic endoderm stage, and compared the profiles of cells further differentiated *in vitro* or *in vivo* (in mice grafts), using RNA-seq, H3K27me3 and H3K4me3 ChIP-seq. These histone marks signals repressed and active TSS respectively and appear together in poised promoters. In this study 78% of bivalent promoters resolved their state at the final stage differentiated *in vivo*, but a high proportion of those promoters remained poised in the cells differentiated *in vitro*, showing that the latter retain an immature profile that is likely responsible for the lower expression of key beta cell genes.

1.8. CRISPR in functional genomics

1.8.1. The CRISPR-Cas9 system

Genome editing has been widely used to investigate gene function, genomic regulation and the effect of specific variants in disease. The CRISPR-Cas9 system has been a revolution in genome engineering, with applications that include DNA deletion, insertion, labelling and modulation of gene expression (Doudna and Charpentier, 2014). The Cas9 enzyme uses its catalytic centre to cut double-stranded DNA at specific target sites located next to a protospacer adjacent motif (PAM) sequence, using the guide of a 20-nucleotide RNA (gRNA). For modulating the expression of selected genes, a deactivated Cas9 protein (dCas9), that has no catalytic activity, is used either fused to activation or to repression domains (Qi *et al.*, 2013; Gilbert *et al.*, 2014).

1.8.2. CRISPR as a tool for interrogating disease pathology

CRISPR can be used to edit variants and candidate genes in iPSC-derived pancreas differentiation models and provide insight into T2D development. For example, this technique was used to generate a double knockout of *RFX6*, which altered the formation of pancreatic progenitors in vitro via reduced *PDX1* transcription (Zhu *et al.*, 2016). *RFX6* is implicated as well in the maintenance adult beta cell function by regulating insulin secretion (Chandra *et al.*, 2014, Piccand *et al.*, 2014). T2D GWAS loci are also enriched in RFX footprints (Varshney *et al.*, 2017), providing several levels of evidence that this factor is mediating T2D risk.

Recent advances in genome-wide CRISPR screening have made it easier to interrogate the function of many genes and potential regulatory regions on a selected phenotype. For example, Rubin and colleagues (2019) developed Perturb-ATAC, a method in which they transduced gRNAs targeting the promoters of TFs and chromatin modifiers in 2,627 B lymphoblasts that had dCas9-KRAB repression machinery integrated in their genome.

Then they performed ATAC-seq on single cells and sequenced the gRNA barcodes to know which perturbation had occurred. They found that repressing the expression of several factors had a similar effect in changing accessibility at many TF binding motifs (e.g. EZH2 and SPI), possibly reflecting their common regulation or interactions. Comparison of the correlations in accessibility at pairs of TF motifs in control versus altered samples can therefore uncover gene networks in which the altered TF plays a central regulatory role.

This method can also uncover TFs, regulatory sites and gene networks important in differentiation. Rubin and colleagues (2019) also profiled accessible sites during keratinocyte differentiation, finding 3 modules of accessibility at TF binding sites, coinciding with the timeframe of differentiation. From those modules they selected 5 TFs presenting highly dynamic changes in accessibility across differentiation to perform Perturb-ATAC on them, this time with a KO screen in Cas9+ cells. They also did KO combinations of two of those TFs at the same time, finding synergistic and additive effects of several factors on accessibility at many motifs. E.g. *KLF4* and *ZNF750* presented high target correlation and ChIP-seq on both shown that they in fact bound the same genomic positions. In addition, several KO TFs impeded correct differentiation. This makes Perturb-ATAC or similar methods of CRISPR screening useful at uncovering TF regulatory mechanisms and candidate regulators of diverse cellular processes, and can be readily applied for endocrine pancreas development.

1.9. Approaches to characterizing *omics* data

Recent years have witnessed the birth and growth of high-throughput *omics* techniques. They all interrogate different aspects of cellular complexity and can be integrated to elucidate the relationships and feedbacks that occur between them. I'm going to focus on specific *omics* technologies that will be used throughout this thesis.

1.9.1. RNA-sequencing

Genome-wide measurement of gene expression (transcriptomics) has improved the understanding of cell physiology in health and disease (The ENCODE Project Consortium, 2007 & 2012; Roadmap Epigenomics Consortium, 2015). The scientific community has moved from microarrays of cDNA (Schena *et al.*, 1995) to the direct sequencing of RNA (RNA-seq), which does not require any prior knowledge of expressed sequences or open reading frames. This has facilitated the discovery of coding, and particularly noncoding, RNAs (Weber *et al.*, 2015). The latter have been shown to dominate the transcriptomic landscape, since most of the genomic sequence appears to be transcribed while only around 3% of it encodes proteins (The ENCODE Project Consortium, 2012). Additional advantages of RNA-seq are the ability to map transcription boundaries and exon-exon junctions with single-base precision, identification of variants in the genome, greater range of detection of transcripts than microarrays, and high reproducibility between technical and biological replicates (Wang *et al.*, 2009).

RNA-sequencing has the following steps. First, the isolation of a population of RNA, which can be total or a fraction (e.g. 3' polyadenylated if one wants to select only messenger RNAs, or ribosomal RNA depleted to include also noncoding RNAs). Second, the RNA is reverse transcribed to complementary DNA (cDNA), creating a library with adaptors attached to one or both ends. Finally, the library can be sequenced over one or several lanes.

1.9.2. DNA methylation

DNA methylation is a stable epigenetic mark involved in repression of gene expression, X-chromosome inactivation, imprinting, splicing and silencing of repetitive DNA sequences (Li *et al.*, 1993; Jaenisch and Bird, 2003; Jones, 2012; Maunakea *et al.*, 2010). DNA methylation is absent or presents huge variation in genomic location and abundance throughout eukaryotic species (Suzuki and Bird, 2008). In vertebrates, this

process happens mostly in cytosine residues directly upstream a guanine (CpG). Although ~70% of all CpGs are methylated in human somatic cells, large stretches of CpGs known as CpG islands are mostly unmethylated (Bird, 1986). Humans have ~30,000 CpG islands (Lander *et al.*, 2001), more likely found upstream of gene promoters: ~60% of genes have them (Bird, 1986; Antequera and Bird, 1993). In humans, mice and other mammals general demethylation followed by *de novo* methylation occurs during embryonic development (Jaenisch *et al.*, 1982), which casts doubt into the role of trans-generational inheritance in mammals.

It is currently unclear to what extent regulatory activity can be inferred from methylation at isolated CpGs. Methylation marks are highly correlated across large stretches of DNA, and their functional analysis usually occurs at the level of the previously mentioned CpG islands, shores (Irizarry *et al.*, 2009), shelves (Sandoval *et al.*, 2011) or other genomic blocks (Jaffe *et al.*, 2012). In fact, the majority of tissue-specific methylation occurs not in CpG islands but at the shores, regions up to 2kb distance from the former (Irizarry *et al.*, 2009). The importance of these regions in tissue-specific regulation is highlighted by the fact that 70% of them are differentially methylated in induced (reprogrammed) pluripotent stem cells (iPSCs) compared to embryonic stem cells (Ji *et al.*, 2010). This in turn could affect the efficiency of cell reprogramming into iPSCs and potentially posterior differentiation down alternative developmental pathways (Ma *et al.*, 2014)

Several methods of DNA methylation profiling have been developed over the years. While whole genome bisulfite sequencing or other genome-wide methods provide higher coverage, microarray solutions are very popular. These include the Infinium arrays from Illumina (Weisenberger *et al.*, 2008), which also facilitate analysis since there are several widely used pipelines developed for them (Leti *et al.*, 2018).

1.9.3. Assay for Transposase-Accessible Chromatin and sequencing

DNA is densely packed inside the nucleus of the cell and organised around histones to form nucleosomes. These are condensed into 30nm fibres of chromatin (Robinson *et al.*, 2006) arranged in a non-random manner with the nuclear lamina or with other regions of the genome (forming topologically associated domains, TADs). These structures ensure that the DNA to be transcribed into RNA and proteins or at regulatory regions is accessible, while keeping inactive regions packed. Regulation of DNA accessibility is dynamic and varies between tissues and cell types, as well as in response to external stimuli or during embryonic development (Bell *et al.*, 2011).

Nucleosome-depleted regions ("accessible" or "open" chromatin) can be detected by several means, including DNase-seq (Song *et al.*, 2011), FAIRE-seq (Giresi *et al.*, 2007) and ATAC-seq (Buenrostro *et al.*, 2013). ATAC-seq employs the prokaryotic Tn5 transposase to link adaptors preferably into regions of accessible chromatin (**Figure 1.4**) simplifying the pre-sequencing protocol compared to the aforementioned methods while requiring 100 to 1000 times less starting material.

It has been estimated that 5% of all accessible regions are located at promoters, and the rest are split evenly between introns and intergenic regions. TSS at promoters are usually highly accessible regardless of organism or cell type (Thurman *et al.*, 2012; Tsompana *et al.*, 2014), which means that there are other elements controlling gene expression, the most obvious being transcription factor activity.

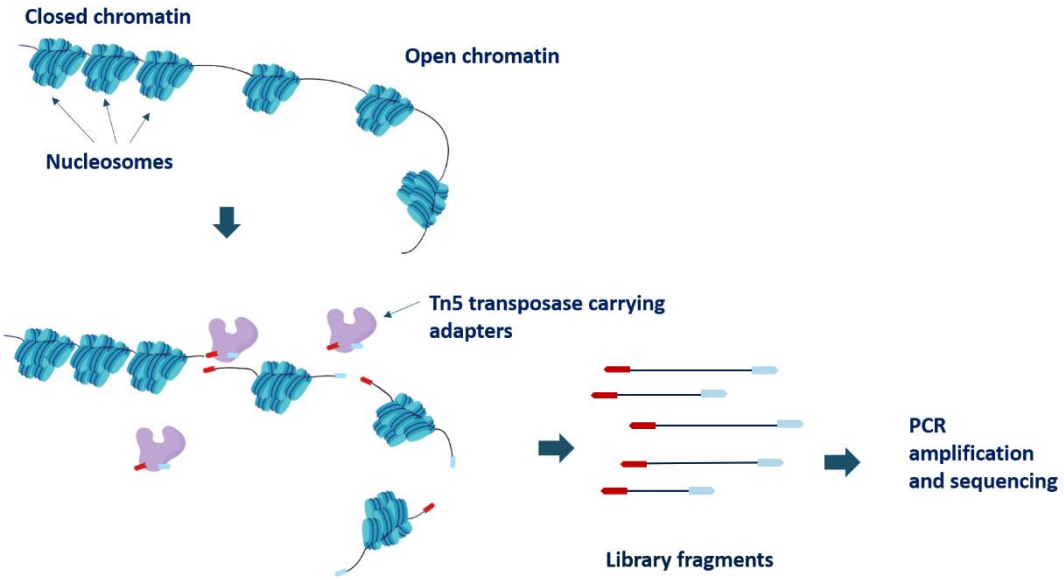


Figure 1.4. ATAC-seq outline. This method probes regions of open chromatin due to the insertion of sequencing adapters (red and light blue) by a transposase (purple), generating the sequencing library that will be amplified before sequencing.

1.9.4. Chromatin immunoprecipitation and sequencing

The nucleosome particle mentioned in the previous section is composed of histones, onto which 147 bp of DNA are wrapped. The histones are organised in an octamer, with a canonical tetramer of histones H3 and H4 surrounded by two dimers of histones H2A and H2B (Kornberg *et al.*, 1999). Each histone has a "tail" of 20-35 amino acids that can be modified in diverse ways. For example, lysines and arginines can be methylated, and can accept from one to three methyl groups. Tens of modifications have been discovered for each core histone (Zhao *et al.*, 2015), and a summary of some of the most studied can be found in **Figure 1.5**. There are also different histone variants, with altered amino acids or whole domains that perform specific functions e.g. during DNA replication (Venkatesh *et al.*, 2015).

These histone modifications alter chromatin structure and accessibility, establishing a pattern that signals which regions are silenced and inaccessible, or accessible and available e.g. for transcription or DNA repair. They are also dynamic and subject to the

action of enzymes that might apply different groups in the same amino acid, or several instances of the same modification (Strahl and Allis, 2000).

Some marks have been specially studied in relation to activation or repression of transcription and modulation of regulatory processes. Active chromatin is marked by H3K27ac and H3K4me3. Conversely, repressed chromatin is characterized by the presence of H3K27me3 and H3K9me3, among other modifications (Venkatesh and Workman, 2015). The transference of acetyl and methyl groups to these aminoacids is performed by histone acetyltransferases and histone methyltransferases, while the opposite process is mediated by histone deacetylases and histone demethylases (Bannister and Kouzarides, 2011). H3K27ac is the most profiled and has been associated with activation of enhancer (Creyghton *et al.*, 2010) and promoter activity (Wang *et al.*, 2008). H3K27ac is particularly interesting because it signals enhancer activity when it appears together with H3K4me1, while H3K4me1 on its own marks “poised” or “bivalent” enhancers that are primed for quick activation during development and that commit to active or repressed state by the end of this process (Mikkelsen *et al.*, 2007; Creyghton *et al.*, 2010). Other marks can differentiate poised promoters, such as H3K27me3 and H3K4me3, which on their own can be found in repressed and active promoters respectively (Xie *et al.*, 2013). The combination of several histone marks with measures of chromatin accessibility can be used to implement complex models of chromatin states which are bound to give more information about the functional characteristics of regulatory regions of the genome and their role in disease.

ChIP can be used to target histones that present a particular amino acid modification (Johnson *et al.*, 2007). Proteins are first crosslinked to DNA, which in turn is sheared by sonication or by enzymatic methods. The antibodies, hybridized to magnetic beads, target the histones and precipitate them with the help of a magnet. Finally, the DNA is de-crosslinked from the proteins. Whole DNA can be sequenced after ligation of adapters.

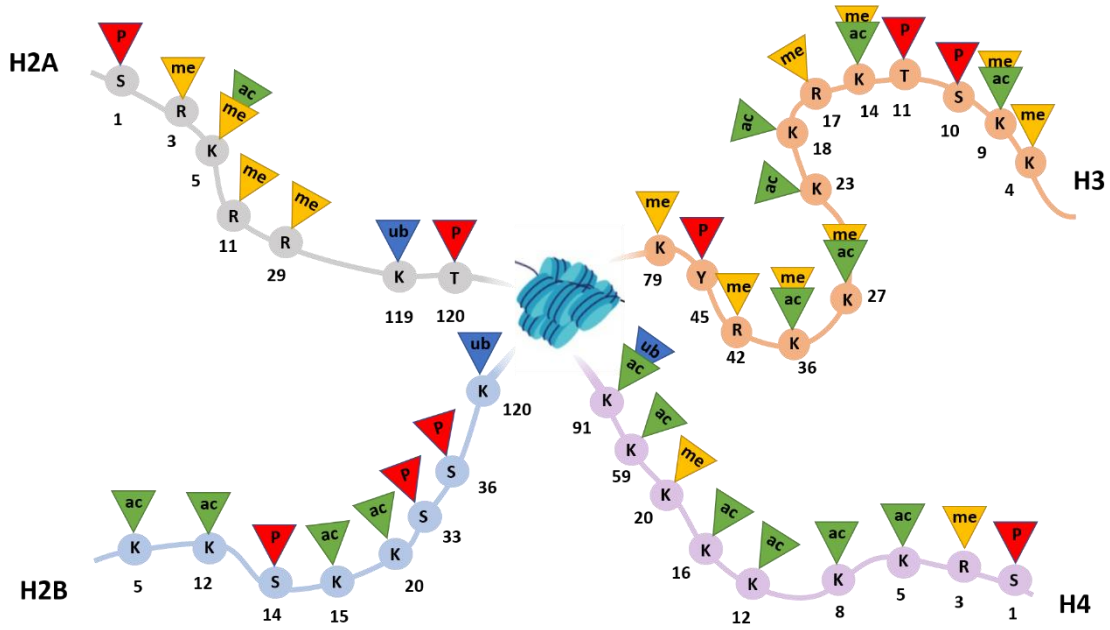


Figure 1.5. Common core histone modifications. ac: acetylation; me: methylation, P: phosphorylation; ub: ubiquitination.

1.10. Aims of this thesis

The main aim of this thesis is to investigate the contribution of islet development to T2D risk, using different sets of *omics* data at every stage of differentiation from a hiPSC model of beta cell development.

To do that, I will aim to:

- Characterize the transcriptome and part of the DNA regulatory landscape (methylation, H3K27ac histone modifications and open chromatin regions) across the stages of differentiation of an hiPSC-derived beta cell development model.
- Study how T2D-associated variants might be influencing changes in disease risk through transcriptomic and DNA regulatory mechanisms.

- Integrate gene expression and DNA regulation to find important genes involved in islet development and T2D risk, including candidate gene effectors within T2D-associated loci.
- Validate *in vitro* the role of novel candidate genes that are highlighted by the *in silico* analyses.

2. Generation and transcriptomic characterization of an hiPSC model of beta cell differentiation

2.1. Introduction

2.1.1. Role of islet development in type 2 diabetes

As described in the introductory chapter, genome-wide association studies (GWAS) have greatly improved our understanding of the genetic contribution to type 2 diabetes (T2D). As most GWAS signals fall outside coding regions, the biological mechanisms underlying these disease association signals remain largely unknown. Despite this, various analyses across associated loci have demonstrated the importance of pancreatic islets in T2D risk (Kahn *et al.,* 2001; Lyssenko *et al.*, 2008; Ashcroft and Rorsman, 2012; Pasquali *et al.*, 2014; van de Bunt *et al.*, 2015).

Most of the follow-up work from GWAS, particularly functional studies, have been performed in adult islets (whether whole islets or their various subpopulations, particularly beta cells) and the immortalized human beta cell line EndoC-βH1 (Parker *et al.*, 2013; Fadista *et al.*, 2014; Pasquali *et al.*, 2014; van de Bunt *et al.*, 2015; Wang *et al.* 2016; Ackermann *et al.*, 2016; Varshney *et al.*, 2017; Ravassard *et al.*, 2011; Lawlor *et al.*, 2017; Lawlor *et al.*, 2019). However, there is evidence that islet development might be influenced by genetic variants associated with T2D:

- With some exceptions (Schwitzgebel *et al.*, 2014), monogenic diabetes genes impact islet and pancreas development, and several of them also reside in or near T2D-associated loci (see **Figure 1.2**) (Fuchsberger *et al.*, 2016).

- T2D-associated variants at the *KCNQ1* locus present allelic imbalance in early islet development, causing monoallelic expression of three imprinted genes (*CDKN1C*, *KCNQ1* and *KCNQ1OT1*) (Travers *et al.,* 2013). The effect on T2D risk was observed to come only from the maternally inherited allele at this locus (Kong *et al.*, 2009), hinting at a developmental role.

- Other genes within T2D-associated loci are clearly expressed during development (e.g. *PROX1* in pancreatic progenitors [Wang *et al.*, 2005]) or regulate differentiation of earlier precursor cells that give rise to the liver and pancreas (e.g. *HHEX* [Bort *et al.*, 2004]).

This supports the role for islet development in T2D pathophysiology, as changes in the composition or the number of islets through developmental defects could lead to altered functional islet mass and increased T2D risk later in life.

2.1.2. *In vitro* differentiation models of islet development

Current understanding of islet development has primarily been derived from studying murine models, since access to human fetal material is restricted. However, there are several core differences between human and mice islet development, both at the gene expression and physiological levels (Kim *et al.*, 2009; Steiner *et al.*, 2010; Benner *et al.*, 2014). For example, glucagon-producing cells are detected early in the mouse pancreatic bud (in a period of rapid cell proliferation called the “primary transition”), while in humans these cells develop much later (Jennings, 2013; Pan and Brissova, 2014). These differences can result in murine models that do not recapitulate human diabetic phenotypes. One well-known example is *GATA6*: *Gata6* $^{+/-}$ mice have no apparent phenotypic alterations (Morrisey *et al.*, 1998), while GATA6 haploinsufficiency in humans often (but not always) causes pancreas agenesis and diabetes (Lango Allen *et al.*, 2011; Bonnefond *et al.*, 2012; Tiyaboonchai *et al.*, 2017).

These findings motivate the recent endeavours to differentiate induced pluripotent stem cells (iPSC) into adult pancreatic islets that, while increasingly recapitulating adult β-cell behaviour, lack complete functionality (Bruin *et al.*, 2014; Pagliuca *et al.*, 2014; Rezania *et al.*, 2014; Russ *et al.*, 2015; van de Bunt *et al.*, 2016). Perfecting *in vitro* models of endocrine pancreas development will allow more detailed interrogation of the genes and pathways influencing islet development and function in humans. Mechanistic analyses of the contribution of candidate regulators of islet development to long-term islet function

is enhanced by recent advances in CRISPR-based approaches that allow their experimental manipulation in *in vitro* developmental models (Beer and Gloyn, 2016).

iPSC-derived beta cells could also prove a cost-effective platform in the screening of drugs for diabetes treatment, and could provide functional tissue for transplants for diabetic patients, something that is being studied in mice (Pagliuca *et al.*, 2014; Rezania *et al.*, 2014; Millman *et al.*, 2016) and in humans (Viacyte Study).

In this chapter I aim to:

- Describe how the iPSC-derived model of beta cell differentiation (from which I derive all *omics* analysis) was generated, and show how it recapitulates the stages of islet development.
- Investigate its transcriptomic landscape to gain clues about the genetics of human endocrine pancreas development.
- Use the transcriptomic information at every stage to try to find deficiencies in the differentiation model, to help improve it.
- Use patterns of gene expression in T2D-associated loci to examine the contribution of islet development to T2D risk.

2.2. Methods

Methods 2.2.1 and 2.2.2 were carried out by other members of the laboratory (with attributions in each section) while the rest were entirely my own work.

2.2.1. *In vitro* iPSC differentiation towards beta-like cells

Human iPSCs were cultured and differentiated as described in Perez-Alcantara *et al.* (2018) by Dr Christian Honoré at Novo Nordisk (Maaloev, Denmark). *In vitro* differentiation involved the addition of recombinant growth factors and small molecules to iPSCs, to sequentially generate cells representing developmental stages of the endocrine pancreas. In order: definitive endoderm (DE), primitive gut tube (GT), posterior foregut (PF), pancreatic endoderm (PE), endocrine progenitors (EP), endocrine-like cells (EN), and beta-like cells (BLC). Three iPSC lines reprogrammed from fibroblasts from different healthy donors of European ancestry (SBAd2.1, SBAd3.1 and SBNeo1.1) were differentiated once, in parallel, using the same culture and differentiation media.

2.2.2. Generation of RNA-seq data

After RNA extraction from cells from each donor and differentiation stage, library preparation and sequencing was performed by the Oxford Genomics Centre (Wellcome Centre Human Genetics, University of Oxford) as described previously (van de Bunt *et al.*, 2016). Briefly, polyadenylated transcripts were isolated with the NEBNext PolyA mRNA Magnetic Isolation Module (E7490 L; New England Biolabs), followed by library preparation using the NEBNext Ultra Directional RNA Library Prep Kit for Illumina (New England Biolabs E7420 L) with 12 cycles of PCR and custom 8 bp indexes. All libraries were multiplexed and sequenced with the TruSeq PE Cluster Generation Kit v3 and TruSeq SBS Kit v3 (PE-401-3001 and FC-401-3001 respectively; both Illumina, San Diego, CA, USA) over 7 lanes of Illumina HiSeq2000 as 100-nucleotide paired-end reads.

RNA-seq libraries were sequenced to a mean read depth of 148 (+/- 12) million reads per sample. Raw sequencing reads were mapped by Dr Martijn van de Bunt from the

McCarthy group to human genome build hg19, with GENCODE v19 (https://www.gencodegenes.org/releases/19.html) as transcriptome reference, using STAR v.2.5 (Dobin *et al.*, 2015), followed by gene-level quantification with featureCounts from the Subread package v.1.5 (http://subread.sourceforge.net/) (Liao *et al.*, 2014).

2.2.3. Differential expression and principal component analysis of islet development stages

All steps from this point were performed by me. Resulting gene counts from the previous section were filtered to include only genes that reached one count per million (CPM) in all donors of at least one differentiation stage. Only autosomal protein-coding and long intergenic non-coding RNA (lincRNA) genes annotated in Ensembl Genes release 88 (http://mar2017.archive.ensembl.org/index.html) (Aken *et al.*, 2016) were retained (15,221 in total) for normalization and differential expression analysis. Gene counts were normalised and transformed to log-CPM using the *voom* function within the *limma* R package (version 3.32.5) (Law *et al.*, 2014), and contrasts were fitted with iPSC as the baseline, adjusting for donor effect, as input for the eBayes function in *limma* (Ritchie *et al.*, 2015) for the differential expression analysis. All the reported p-values of differential expression were adjusted for false discovery rate (*q* values) using the Benjamini and Hochberg method. Differentially-expressed genes (q<0.01) with absolute $\log_2$ fold change ($\log_2$FC)>1 were assigned to the stage in which they were most upregulated compared to the baseline iPSC profile. When the $\log_2$FC was negative for all contrasted stages, the gene was assigned to iPSC.

For the differential analysis between the previous (van de Bunt *et al.*, 2016) and current protocols, only genes detected in both experiments and expressed at >=1 CPM in all donors of at least one stage in at least one of the experiments were retained, resulting in 15,464 protein coding genes and lincRNAs. The differential expression analysis was performed exclusively for the stages shared between the protocols (iPSC, DE, GT, PF, PE and EN). Of note, sample numbers differed between the protocols with three samples (SB Ad2.1, SB Ad3.1, SB Neo1.1) differentiated with the current protocol, and two samples (SB Ad2.1, SB Ad3.4) with the previously published protocol.

For the principal component analysis, samples differentiated with the current protocol, the previous protocol (van de Bunt *et al.*, 2016), human embryonic stem cells (hESC) differentiated by Xie and colleagues (2013) and adult islets from Nica and colleagues (2013) were included. In Xie and colleagues' data, human embryonic stem cells had been differentiated *in vitro* along the DE, GT, PF, PE, late PE and polyhormonal stages, and matured *in vivo* (in mice) from PE to functional endocrine cells. In Nica and colleagues' study, they sequenced the transcriptome of adult islets, adult beta cells and adult non-beta cell subtypes. Genes detected in all four experiments and expressed at >1 CPM in all donors of at least one stage in at least one of the experiments were retained, resulting in 16,013 protein coding genes and lincRNAs. Counts were normalised and transformed to log-CPM using the *voom* function within the *limma* package in R, and corrected for batch effects of the studies before principal component analysis using *removeBatchEffect* in *limma*.

2.2.4. Pathway analysis

Differentially expressed genes in each stage were tested for enrichment in Reactome pathway terms using *a hypergeometric* test with the *enrichPathway* function of the ReactomePA package (v. 1.28.0; Yu *et al.*, 2016) in R. All the genes tested for differential expression were used as background. Terms that were significant (5%) after Benjamini and Hochberg p-value adjustment were retained.

2.2.5. ChIP-seq and transcription factor motif enrichment

Regulating transcription factors (TFs) were extracted from the lists of differentially expressed genes assigned to each stage using the iRegulon (v1.3) Cytoscape plugin (Verfaillie *et al.*, 2014; Shannon *et al.*, 2003). Significantly enriched TFs were detected based on 1,120 ChIP-seq tracks from ENCODE and 9713 non-redundant TF binding motifs from 7 species. Regions 10 kb around the transcription start site of each gene were considered, and enrichment was run with default parameters. Motifs and ChIP-seq tracks were ranked based on Normalized Enrichment Score (NES), with only those with

NES>3 (corresponding to an FDR = 3% - 9%) being considered. Enriched motifs were then matched to transcription factors known to bind them.

2.2.6. Clustering of gene expression profiles

Correlation of gene expression patterns across all stages was calculated using the weighted gene co-expression network analysis (WGCNA) package (v.1.51; Langfelder and Horvath, 2008) in R. Counts were normalized using variance stabilizing normalization and the network was constructed by hierarchical clustering of adjacency-based dissimilarity. This divided the genes into 29 co-expressed modules.

2.2.7. T2D and fasting glucose gene enrichment

Enrichment analysis was implemented in two ways: as a hypergeometric test in R (using all genes tested for differential expression as background) or using the gene scoring function in MAGENTA (Segrè *et al.*, 2010) followed by a gene set enrichment analysis (GSEA) (Mootha *et al.*, 2003; Subramanian *et al.*, 2005).

For the hypergeometric test, the differentially expressed genes from each differentiation stage were tested for enrichment in genes mapping to T2D or fasting glucose (FG) GWAS signals, which were defined as protein coding genes and lincRNAs located within 0, 50, 100, 200 or 500 kb surrounding the credible intervals for T2D/FG-associated loci. Credible intervals were defined by the boundaries of the 99% credible sets of variants (that collectively encompass the 99% posterior probability of association with the trait) (Maller *et al.*, 2012) from a meta-analysis of 32 GWAS (898,130 European subjects imputed to HRC [Mahajan et al., 2018a], 243 loci) and from ENGAGE (46,694 European subjects imputed to 1000 Genomes [Horikoshi *et al.*, 2015], 16 loci), respectively. This resulted in lists of 963, 1222, 1417, 1811 and 2892 genes and lincRNAs present in the background for T2D; and 14, 27, 39, 57 and 101 for FG.

Fifteen loci influencing hyperglycaemia, beta-cell function (defined as reduced insulin secretion and fasting hyperglycaemia), and insulin processing (reduction in fasting proinsulin levels); or that were significantly associated in a GWAS for intravenous

glucose test for insulin secretion (Dimas *et al.*, 2014; Wood *et al.*, 2017) (**Table 2.1**), were pooled as “beta-cell function”.

The enrichment analysis by hypergeometric test in R was performed using all genes tested for differential expression as background. A distribution of *p* values was calculated for each stage from 10,000 random samplings (without replacement) of groups of genes, of the same size as the gene sets tested, from the background, which after applying the hypergeometric test gave a distribution of random *p* values. Then a permuted *p* value was calculated for each stage, defined as the fraction of *p* values from the random distribution that were more extreme (smaller) than the *p* value from the differentially expressed gene set for each stage.

For the analysis with MAGENTA and GSEA, SNPs from the T2D GWAS meta-analysis (243 loci) (Mahajan et al., 2018a) were mapped to genes and a ranked list of *p* values was generated for each gene using 10,000 permutations and gene boundaries of 110 kb upstream and 40 kb downstream of each transcript. This gene score list was tested in GSEA for enrichment using 1,000 permutations, in two ways: for enrichment of the complete gene score list in sets of differentially expressed genes for each stage, or conversely for enrichment of these stages in significant gene scores ($p<0.05$ by MAGENTA). All the reported *p* values for GSEA are adjusted for false discovery rate (*q* values).

2.2.8. Assessment of deleterious mutations and p53 mutations

Combined Annotation-Dependent Depletion (CADD) was used to measure variant deleteriousness across the whole genome for all three donors (Kircher *et al.*, 2014; Rentzsch *et al.*, 2018). This method infers deleteriousness by generating annotations from dozens of genomic features and comparing those from fixed alleles in humans to simulated *de novo* mutations. It assumes that deleterious variants are depleted by natural selection in real human variation, but not in simulated variation.

To detect deleterious p53 mutations, single nucleotide variants were identified from RNA-seq data using *rnaseqmut* (v. 0.7, https://github.com/davidliwei/rnaseqmut) and then

compared with a library of p53 mutations at the DNA binding domain generated in cell culture and grown in mice (Kotler *et al.*, 2018). Deleteriousness was measured as an increase in cell proliferation compared to wild-type p53.

Locus	**Dimas *et al.*, 2014**	**Wood *et al.*, 2017 (Genome wide)**
ADCY5	BC	Yes
ARAP1	PI	Yes
CDKAL1	BC	Yes
CDKN2A_B	BC	Yes
DGKB	BC	-
GCK	HG	-
HHEX_IDE	BC	-
HNF1A	UC	Yes
IGF2BP2	UC	Yes
*KCNQ1*_rs163184	UC	Yes
MTNR1B	HG	Yes
PROX1	BC	-
SLC30A8	BC	Yes
TCF7L2	BC	Yes
THADA	BC	-

Table 2.1. Beta cell function loci, as defined by: Dimas and colleagues (2014) as involved in hyperglycaemia (HG), beta cell function (BC) or insulin processing (PI); or Wood and colleagues' (2017) following GWAS of insulin secretion after intravenous glucose tolerance test. UC: uncharacterized.

2.3. Results

2.3.1. Quality control of SB cell lines and of differentiation

Dr Christian Honoré differentiated in parallel iPSCs from three independent donors (SBAd2.1, SBAd3.1 and SBNeo1) through seven stages of islet development (see **Section 2.2.1**) and confirmed by FACS the presence of protein markers of each developmental stage. He also showed that cell lines SB Ad3.1 and SB Neo1.1 achieved beta cell function measured by a static insulin secretion assay in response to extracellular glucose (Perez-Alcantara *et al.*, 2018).

After sequencing, to ensure that the cell lines were not harbouring deleterious mutations that could compromise proliferation and differentiation, I searched for mutations in the *TP53* gene (a known regulator of apoptosis and proliferation that is often mutated in cancer [Wawryk-Gawda *et al.*, 2014]), comparing those with a library generated *in vitro* (Kotler *et al.*, 2018). None of the cell lines presented deleterious mutations in over 5% of RNA-seq reads at any stage. Inference of deleteriousness at all other genomic locations highlighted by CADD (see **Section 2.2.8**) indicated the presence of a mutation (PHRED score = 15.43, within predicted top 10% deleterious mutations genome-wide) in the *WRAP53* gene in SBNeo1.1 donor replicates. 40% of reads included the mutation at the iPSC stage and 45% at the BLC stage. This gene regulates p53 expression and therefore it is potentially controlling cell proliferation. Although a direct link between this mutation and the differentiation of this cell line could not be stablished, SBNeo1.1 donor replicates presented a lower differentiation efficiency by FACS at all stages (Perez-Alcantara *et al.*, 2018).

I then confirmed whether the differentiated cells followed normal islet development, profiling gene expression patterns of key marker genes at each stage. The pluripotency of the cells was confirmed by the high levels of expression of *POU5F1* / *OCT4* in iPSCs (**Figure 2.1**) (Pan *et al.*, 2002). Definitive endoderm (DE) marker *SOX17* appeared in the next stage, followed by the determinant of the gut tube (GT) *FOXA2* (Pagliuca *et al.*, 2013) and markers of pancreatic endoderm and progenitors (PF to EP in this model)

PDX1 and *NKX6-1* (Jennings *et al.*, 2015). The insulin gene (*INS*) along with genes encoding other members of the insulin processing and secretion pathway (e.g. *ABCC8, GCK*, etc.) indicate the presence of polyhormonal endocrine (EN) and beta-like cell (BLC) stages.

Principal component analysis of the transcriptome showed that the beta-like cells generated in the current study clustered more closely with *in vivo*-matured islet-like cells (Xie *et al.*, 2013) than cells from differentiations derived with a previously-published protocol by my laboratory (van de Bunt *et al.*, 2016) (**Figure 2.2 a**). Xie and colleagues matured endocrine cells in mice grafts *in vivo* after differentiating them through the preceding stages *in vitro*, constituting the current benchmark for most functionally mature endocrine pancreas-like cells. Of note, the previous differentiations only reached to what now would be akin to the EN stage (indicated in **Figure 2.2 a** with triangles). **Figure 2.2 b** indicates how recent advances in differentiation have improved mostly the later stages: the current and previous protocols from the McCarthy group, and cells differentiated by Xie and colleagues (2013) cluster well until the GT stage, while differences in the transcriptome start to emerge from the PF stage. Meanwhile, the EP stage from this protocol is more similar to polyhormonal cells from Xie and colleagues than the EN stage from the previous protocol. This increased maturity in the EP and EN stages from the current protocol could be due to the higher expression of key markers of endocrine progenitor cells including *NEUROG3* (**Figure 2.3**), essential to generate adult beta cells (Jennings *et al.*, 2015). Nevertheless, comparison with the transcriptome of adult islets and their subtypes (Nica *et al.,* 2013) indicates that even the most developed *in* vitro-derived beta cells are still immature (**Figure 2.3 c**).

Differential expression analysis comparing transcriptomic profiles obtained from differentiations under current and previous protocols showed increasing divergence with differentiation stage (from 17 genes showing differential expression in iPSCs to 2095 at the endocrine-like cell stage). Reactome pathway analysis indicated that genes displaying increased expression at the endocrine-like cell stage (in comparisons of the current vs previous protocols) were enriched for terms including ‘regulation of insulin

secretion' (q=1×10^{-5}), driven by the increased expression of genes like *INS*, *KCNJ11* and *ABCC8* (**Appendix Table 2.1**). Another significantly enriched term in the current differentiation was 'regulation of beta-cell development' (q=2×10^{-3}), driven by the higher expression of *NEUROD1*, *HNF1A* and *PAX4*, among others. This reveals how advances in differentiation protocols are reflected in the transcriptome, particularly in the later stages of differentiation where there is a clear increase in the expression of genes essential for beta cell function and identity. This is the case for beta cell transcription factor *MAFA* (Zhang *et al.*, 2005; Wang *et al.*, 2007; Hang *et al.*, 2014), which was completely absent in the previous differentiation protocol, and the pancreatic hormone *INS*, whose high expression (~10^3 fold higher than that obtained with the previous *in vitro* differentiation protocol) indicates a better differentiation towards the last stage of beta cell development (**Figure 2.3**).

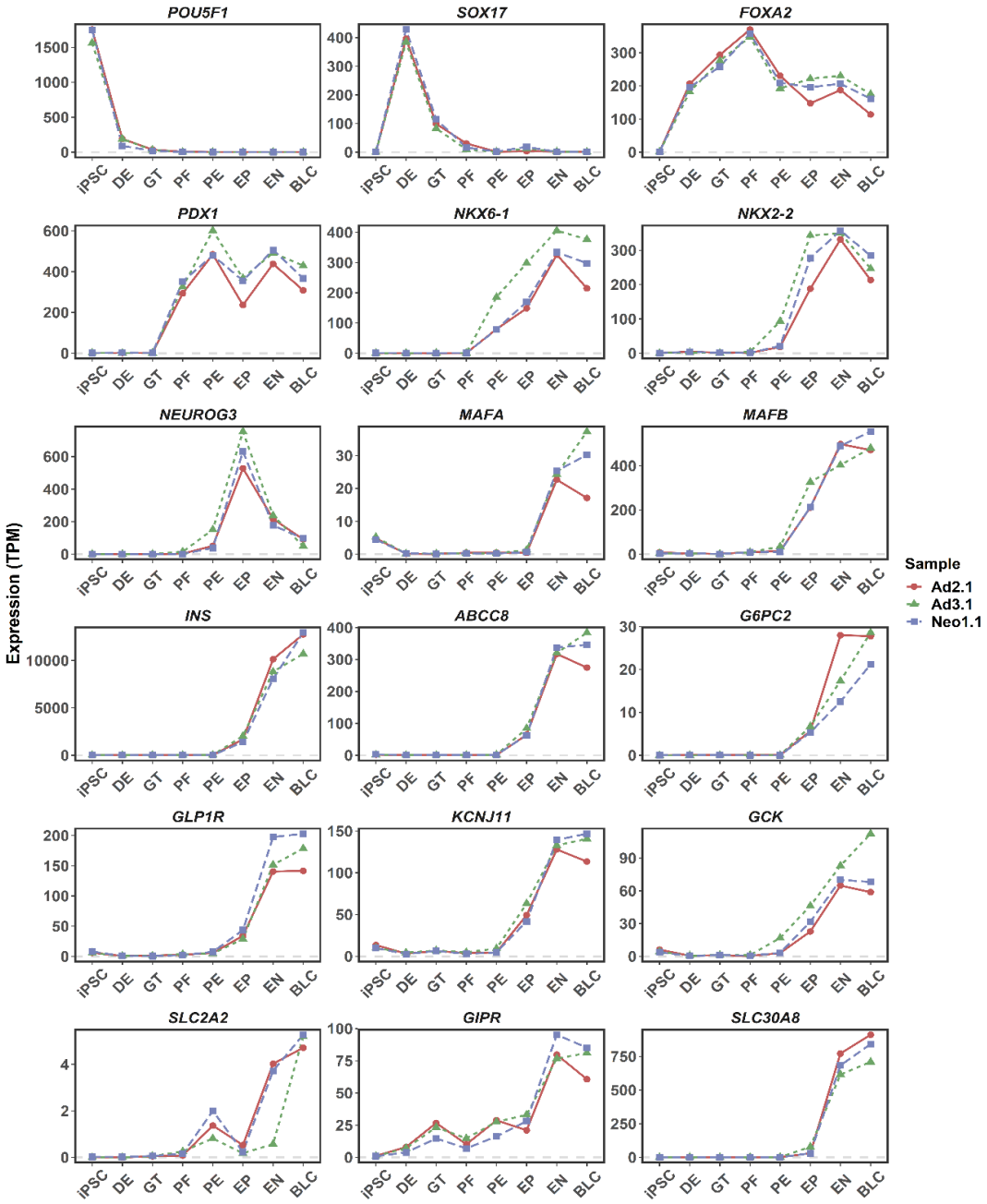


Figure 2.1. Expression patterns (in transcripts per million) of islet development genes and cell marker genes of endocrine cells: *POU5F1* / *OCT4* (iPSC), *SOX17* (DE), *FOXA2* (DE onwards), *PDX1* (PF onwards), *NKX6-1* (PE onwards), *NEUROG3* (EP), *MAFA* and *INS* (EN onwards), and additional beta cell markers, follow the expected patterns of expression. iPSC, induced pluripotent stem cell; DE, definitive endoderm; GT, primitive gut tube; PF, posterior foregut; PE, pancreatic endoderm; EP, endocrine progenitor; EN, endocrine-like cell; BLC, beta-like cell.

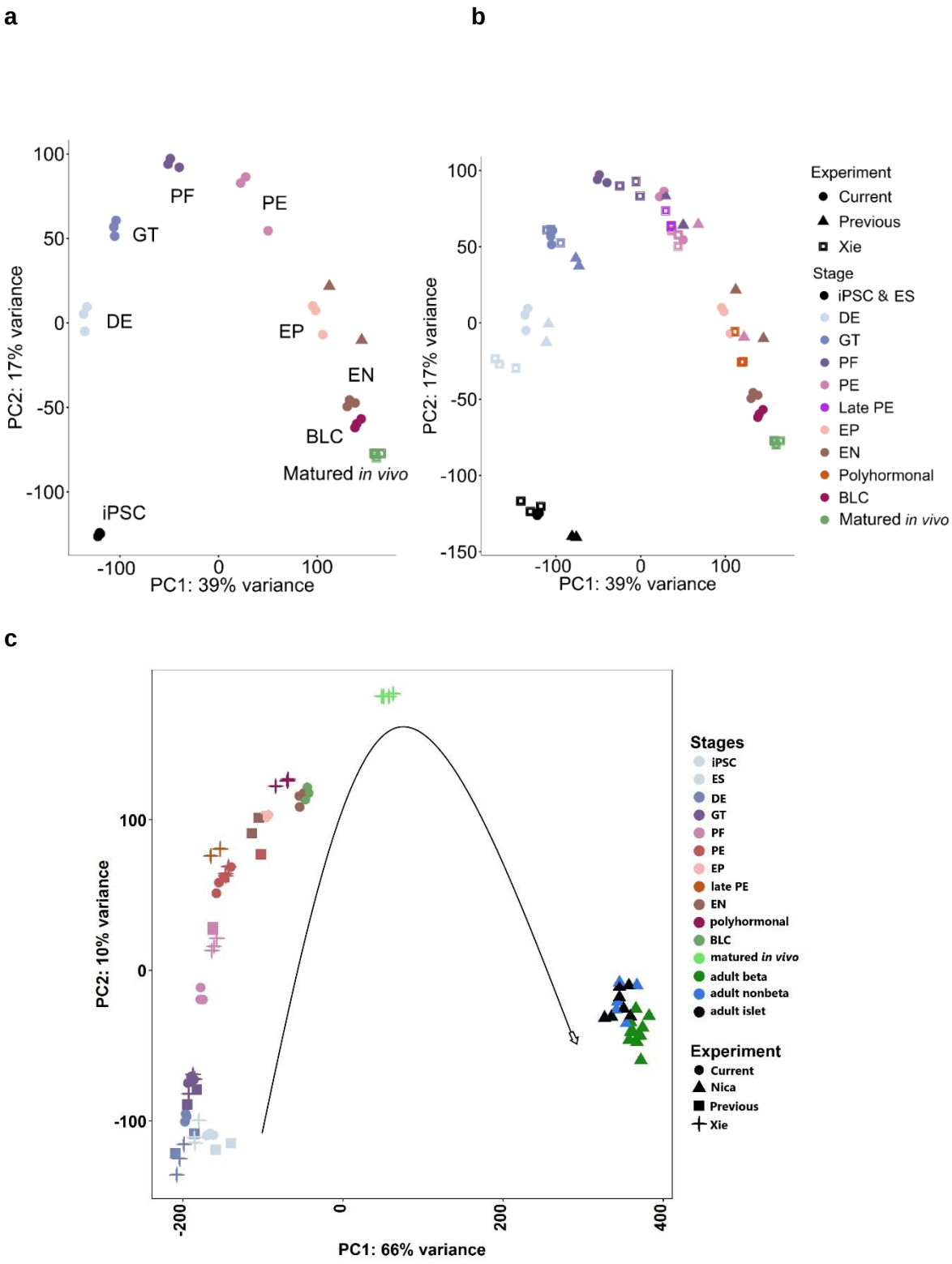


Figure 2.2. Principal component analysis of whole-transcriptome data derived from multiple differentiated human islet-like cell models. [*Legend continues on next page*]

Figure 2.2. [*Continues from previous page*]

(a) Data include all stages from the current differentiation protocol [Current], the most mature stage (EN) of a previously published differentiation protocol [Previous] (van de Bunt *et al.*, 2016), and cells derived via *in vivo* maturation by Xie and colleagues (2013) [Xie].
(b) Data includes all stages from the current differentiation protocol, all stages from the previously published protocol, and all cells differentiated by Xie and colleagues.

The first two principal components (PC1, PC2) have been calculated using normalised gene counts for all stages of all three studies and corrected for batch effects. iPSC, induced pluripotent stem cell; DE, definitive endoderm; GT, primitive gut tube; PF, posterior foregut; PE, pancreatic endoderm; EP, endocrine progenitor; EN, endocrine-like cell; BLC, beta-like cell. Stages from the current study: iPSC, DE, GT, PF, PE, EP, EN, BLC. Stages from the previously-reported study: iPSC, DE, GT, PF, PE, EN. Stages from Xie and colleagues' *in vivo* maturation study: ES, DE, GT, PF, PE, late PE, polyhormonal, matured *in vivo*.

(c) Data includes all stages from the current differentiation protocol, all stages from the previously published protocol, all cells differentiated by Xie and colleagues, and adult islets and islet subpopulations from Nica and colleagues (2013) [Nica]. The arrow indicates the direction of development, from iPSCs to islets.

The first two principal components (PC1, PC2) have been calculated using normalised gene counts for all stages of all four studies and corrected for batch effects. iPSC, induced pluripotent stem cell; DE, definitive endoderm; GT, primitive gut tube; PF, posterior foregut; PE, pancreatic endoderm; EP, endocrine progenitor; EN, endocrine-like cell; BLC, beta-like cell. Stages from the current study: iPSC, DE, GT, PF, PE, EP, EN, BLC. Stages from the previously-reported study: iPSC, DE, GT, PF, PE, EN. Stages from Xie and colleagues' *in vivo* maturation study: ES, DE, GT, PF, PE, late PE, polyhormonal, matured *in vivo*. Samples from Nica and colleagues' study: adult islet, adult nonbeta, and adult beta.

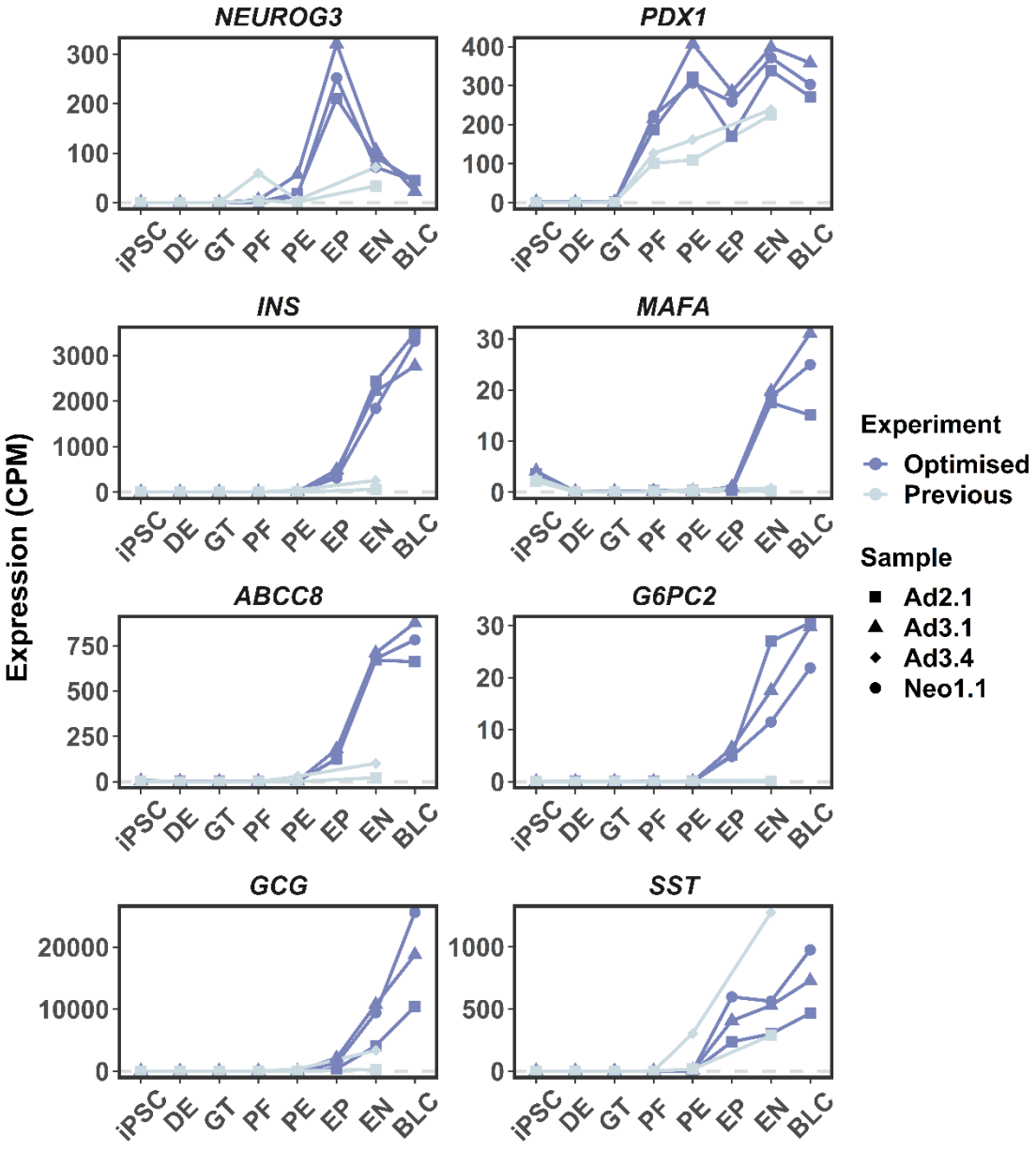


Figure 2.3. Comparison of expression patterns of EP, EN and BLC markers in current and previous protocols. Comparison of expression patterns (in counts per million) throughout all islet developmental stages from samples generated in this study, and reported previously (van de Bunt *et al.*, 2016). Lines SB Ad3.1 and Ad3.4 are different clones reprogrammed from the same individual. iPSC, induced pluripotent stem cells; DE, definitive endoderm; GT, primitive gut tube; PF, posterior foregut; PE, pancreatic endoderm; EP, endocrine progenitor; EN, endocrine-like cell; BLC, beta-like cell.

2.3.2. Characterization of the global transcriptomic landscape

To characterise the transcriptomic landscape of each developmental stage in the *in vitro*-differentiated cells produced in this study, I assigned significantly differentially expressed genes to the stage at which they were most upregulated: if the expression peaked in iPSCs, the gene was assigned to that stage (see **Section 2.2.3**). I detected 9,409 significantly differentially expressed genes (q<0.01, absolute $\log_2$FC>1) across all stages, ranging in number from 623 in the primitive gut tube stage to 2,773 in iPSCs. Known developmental marker genes, such as *NEUROG3* in endocrine progenitors (Rukstalis and Habener, 2009) and *INS* in beta-like cells, were correctly assigned to their canonical stages (**Table 2.2**).

Other important genes include *PDX1*, marker of the dorsal pancreas (Burlison *et al.*, 2008; Jennings *et al.*, 2013; Shih *et al.*, 2015), and *NKX6*-1, a key pancreatic endoderm transcription factor (Sander *et al.*, 2000; Schaffer *et al.*, 2010; Jennings *et al.*, 2013). It is important to keep in mind that these genes might be expressed in other stages (e.g. *PDX1* is detected as early as the PF stage, see **Figure 2.1**), which highlights the limitations of this differential expression analysis strategy. Nevertheless, this process assigns globally the expected set of genes to each stage as highlighted by gene pathway analysis (**Figure 2.4**), which showed enrichment in terms such as 'regulation of beta-cell development' in endocrine-like cells (q=0.02), driven by transcription factors such as *NEUROD1* and *NXK2-2*, and 'regulation of insulin secretion' in beta-like cells (q=2.0×10^{-4}), represented among others members of this pathway by *ABCC8*, which codes for a subunit of the potassium channel that controls insulin release.

Taken together, these results indicate that this differentiation model reflects known pancreas development biology, making it a good approximation for studying the role of genes in T2D-associated loci and monogenic diabetes genes in this process.

2.3.3. Monogenic diabetes genes in beta cell development

The expression patterns of monogenic diabetes genes can point towards stages at which disruption of islet development has long-term consequences for glucose homeostasis.

Of 32 genes that have been implicated in monogenic or syndromic diabetes, 24 were differentially expressed in at least one stage of the in vitro-differentiated model. Ten mapped to the latest beta-like cell stage, but the other 15 showed significant upregulation earlier in differentiation (**Table 2.3**). *GATA6*, for example, was differentially expressed at the definitive endoderm stage ($log_2FC=9.5$, $q=7.6\times10^{-11}$), whereas *GATA4* was differentially expressed in posterior foregut cells ($log_2FC=8.2$, $q=1.9\times10^{-11}$). Both have well-known functions in mice pancreas development (Carrasco *et al.*, 2012), but with limited studies in human organogenesis. Some differences with murine models have been observed: heterozygous *Gata6* knockout mice present no apparent changes in phenotype, while humans heterozygous mutants display a varied symptomatology of diabetes (Jennings *et al.*, 2015; Beer and Gloyn, 2016), including pancreas agenesis, an extreme phenotype has also been observed in a *GATA4* mutant (De Franco *et al.*, 2013; Shaw-Smith *et al.*, 2014; D'Amato *et al.*, 2010). Recent studies (Shi *et al.*, 2017; Tiyaboonchai *et al.*, 2017) reveal a role of GATA6 in definitive endoderm induction, consistent with the pattern of expression observed in the data from this differentiation model. The later expression of *GATA4* could contribute to the less severe phenotype of individuals carrying *GATA4* vs *GATA6* mutations (**Figure 2.5**) (Lango Allen *et al.*, 2011; Shaw-Smith *et al.*, 2014).

The differentiation model used in this study also sheds light on the developmental role of monogenic diabetes genes with lesser described roles. *LMNA*, for example, encodes a nuclear membrane protein involved in chromatin structure and nuclear stability: it has been implicated in the function and development of many tissues (Zuela *et al.*, 2012). The diabetes in carriers of the *LMNA* mutation is mostly driven by altered adipose tissue deposition and insulin resistance (Boschmann *et al.*, 2010). However, the profile of *LMNA* expression during in vitro islet differentiation (peaking in pancreatic endoderm; $log_2FC=1.1$, $q=3.1\times10^{-3}$) may indicate an additional impact on islet development (Kowluru, 2000).

2.3.4. Transcription factors controlling differentiation gene networks

The developmental competence of differentiating cells is in part driven by a subset of transcription factors that initiate and regulate changes in response to external stimuli, as highlighted by the many monogenic diabetes genes that are also transcription factors. To identify potential upstream transcriptional regulators active at each stage of islet development, I clustered genes by their expression patterns with WGCNA and determined the enrichment of transcription factor-binding motifs and ChIP sequencing signals near differentially expressed genes using iRegulon (see **Appendix Table 2.2**).

Assigned stage	Marker gene
iPSC	*POU5F1, NANOG*
DE	*SOX17, CXCR4*
GT	*HNF1B*
PF	*FOXA2*
PE	*SOX9*
EP	*NEUROG3*
EN	*NEUROD1, PDX1, NKX6-1, NKX2-2*
BLC	*INS, MAFA*

Table 2.2. Marker genes and stage to which they have been assigned by differential expression. Each marker was assigned to the stage in which they present the maximum log2FC compared to iPSC (see **Section 2.2.3**).

Assigned stage	Monogenic diabetes gene
iPSC	*ALMS1*
DE	*GATA6*
GT	*PPARG, HNF1B*
PF	*GATA4*
PE	*HNF4A, PTF1A, INSR, LMNA*
EP	RFX6, MNX1, UCP2, NEUROG3
EN	NEUROD1, PDX1
BLC	INS, ABCC8, KCNJ11, WFS1, HNF1A, GCK, GLIS3, INS, HADH, SLC2A2

Table 2.3. Differentially expressed genes identified as causal for monogenic diabetes by Fuchsberger *et al.* (2016).

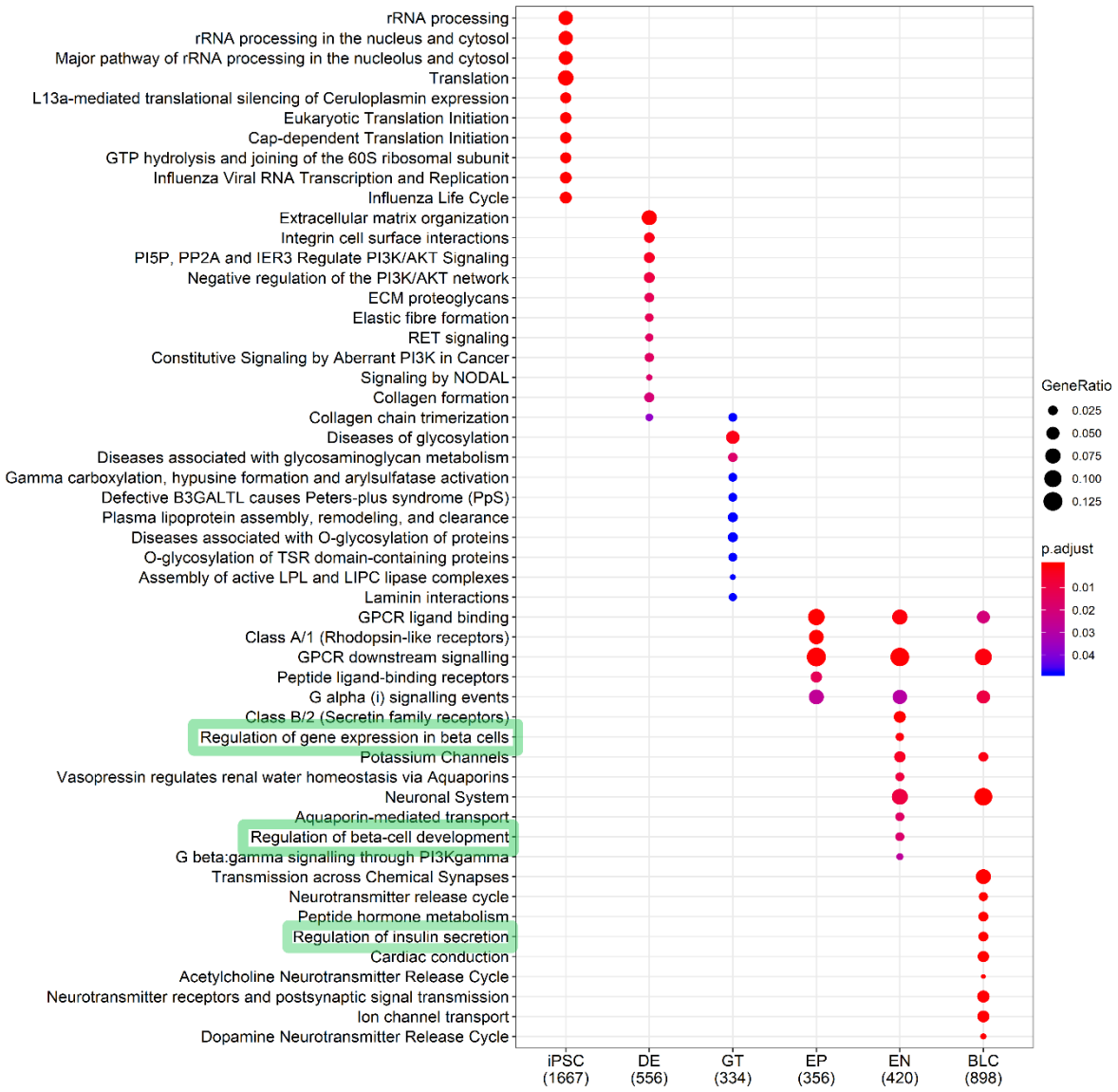


Figure 2.4. Pathway enrichment analysis of differentially expressed genes assigned to each stage. Enrichment of terms from the Reactome pathway (y-axis) was calculated per stage (x-axis, total genes in parentheses). Only top 10 significant (Benjamini-Hochberg adjusted p-value < 0.05) terms per stage are shown. In green: pathways related to beta cell development and function. “GeneRatio”: genes found per term over total genes in all terms per stage. “p.adjust”: Benjamini-Hochberg adjusted p-value.

This analysis confirmed the impact of well-established developmental transcriptional regulators such as the monogenic diabetes gene *HNF1B*, which showed iRegulon enrichment of its targets at the primitive gut tube stage (NES = 5.7). Some of these HNF1B targets also have known effects on pancreas development (*SMAD7* [Smart *et al.*, 2006], *ID2* [Hua *et al.*, 2006]), on mature islet function and on the development of other tissues that also arise from the gut tube (*GGCX*) (Ferron *et al.*, 2015). *HNF1A*, with possible overlapping functions, presents the same transcription factor binding motifs (see **Appendix Table 2.2**).

Analysis of the sets of stage-specific differentially expressed genes also highlighted the targets of transcription factors with less-well studied roles in human islet development. For example, expression of the transcriptional repressor *REST* peaks in the intermediate steps of *in vitro* differentiation and declines at the endocrine-like cell and beta-like cell stages (**Figure 2.6**), with reciprocal expression patterns seen among its predicted targets (enrichment analysis in iRegulon: NES in BLC stage = 4.9-7.9; NES in EN stage = 4.7-5.3; see **Appendix Table 2.2**). These targets include genes encoding neurexins (*NRXN1*, *NRXN2*) and subunits of the glutamate receptor channels *(GRIA1*, *GRIA2*, *GRID1*, *GRIK2*) implicated in insulin exocytosis (Mosedale *et al.*, 2012; Suckow *et al.*, 2008; Wu *et al.*, 2012). Correlation of gene expression with WGCNA assigns *REST* to the same cluster as *TCF7L2* and other genes from the Wnt signalling pathway, such as *TCF7*, *TCF3* and *TCF12* (Cadigan and Waterman, 2012). This pathway is important for islet development and is targeted in many *in vitro* differentiation protocols (Pagliuca *et al.*, 2014; Rezania *et al.*, 2014). These data therefore indicate that REST is likely to be an important transcriptional regulator of human islet development, both in intermediate (pancreatic endoderm, endocrine progenitor) and later (endocrine-like cell, beta-like cell) (Atouf *et al.*, 1997; Jensen, 2004) stages of differentiation, as has also been recently suggested by studies in mice and humans (Martin *et al.*, 2015; Solimena *et al.*, 2018).

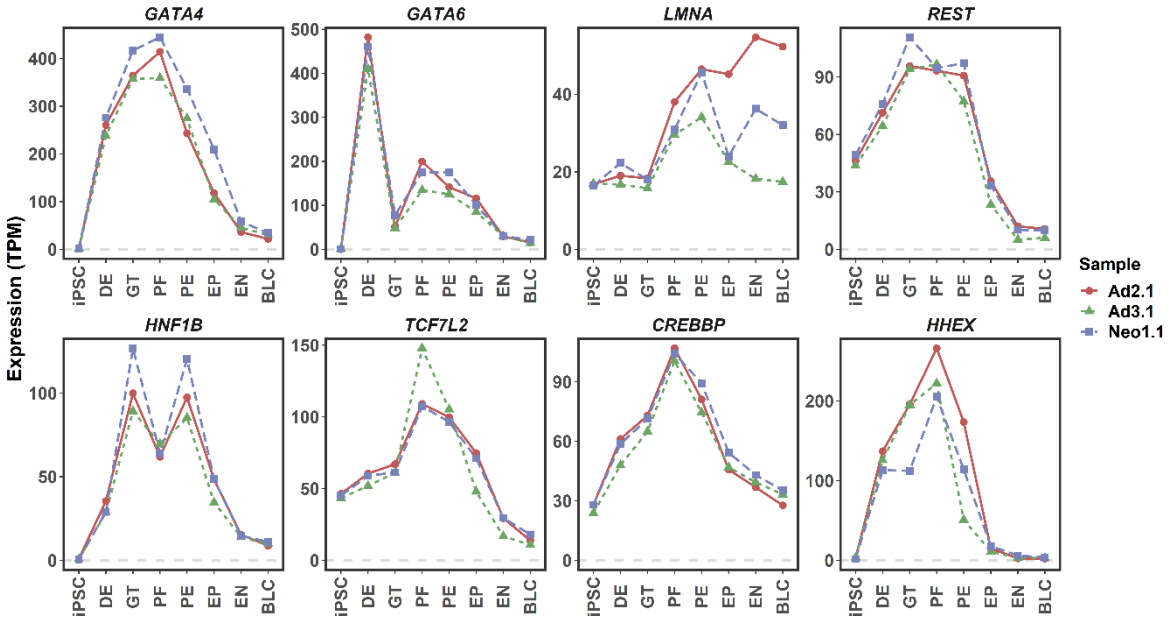


Figure 2.5. Expression patterns (in transcripts per million) of various genes highlighted in the main text. iPSC, induced pluripotent stem cell; DE, definitive endoderm; GT, primitive gut tube; PF, posterior foregut; PE, pancreatic endoderm; EP, endocrine progenitor; EN, endocrine-like cell; BLC, beta-like cell.

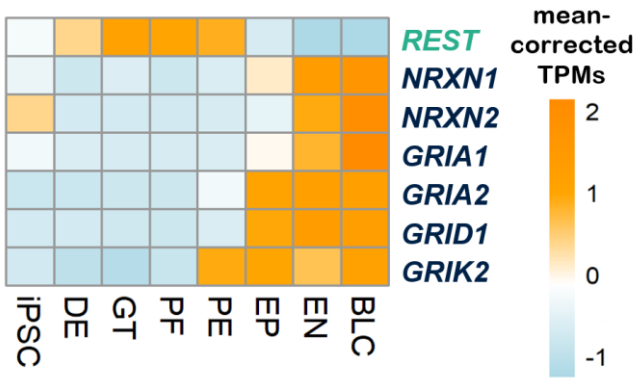


Figure 2.6. Patterns of expression of REST and its predicted targets during beta cell development. Differences of gene expression in mean-corrected transcripts per million, for the means of the three replicates per stage. iPSC, induced pluripotent stem cell; DE, definitive endoderm; GT, primitive gut tube; PF, posterior foregut; PE, pancreatic endoderm; EP, endocrine progenitor; EN, endocrine-like cell; BLC, beta-like cell.

TCF7L2 maps to the T2D-associated locus with the largest common effect on disease risk (Mahajan et al., 2018a). Analysis of TCF7L2 targets (as assessed by ChIP sequencing with iRegulon) shows marked enrichment at the posterior foregut stage (NES=3.4) that mirrors that of *TCF7L2* expression ($\log_2$FC=1.2; q=8.5×10^{-10}). *TCF7L2* shows enrichment in ChIP-seq signals in the same stage where it is most differentially expressed (PF) and the stage immediately before (GT). Some of its targets are genes involved in metabolic syndrome (*NR2F2*) (Boutant *et al.*, 2012), acinar cell identity (*MEIS2*) (Swift *et al.*, 1998), and postnatal alpha and beta cell proliferation (*SERPINB1* and *FOXP1*, respectively) (Ouaamari *et al.*, 2016; Spaeth *et al.*, 2015). The peak of expression of all these genes in this intermediate stage suggests a possible role for them under control of *TCF7L2* in early pancreas development (see **Appendix Table 2.2**).

The expression of several other Wnt family members also peaks at the posterior foregut stage; these include the coactivator *CREBBP*, the binding sites of which are significantly enriched in T2D-associated loci (Morris *et al.*, 2012), and *HHEX*, which maps to a prominent T2D-risk locus and is implicated in foregut development (Ferron *et al.*, 2015; Arterbery and Bogue, 2016). In the developing embryo, cells of the posterior foregut can differentiate into liver as well as endocrine pancreas (Gittes, 2009). Alleles associated with risk of T2D within the *TCF7L2* and *HHEX* loci may influence early expression of these genes, which could affect development in multiple metabolic tissues. This view is supported by cellular and murine studies indicating that *TCF7L2* regulates beta cell development and function (Le Bacquer *et al.*, 2011), including via indirect effects in supporting tissues (Sakhneny *et al.*, 2018), as well as affecting hepatic function (Boj *et al.*, 2012). Similarly, *Hhex* is essential for the differentiation of the posterior foregut into the liver in mice (Arterbery and Bogue, 2016), yet is also thought to regulate delta cell identity and function in islets (Zhang *et al.*, 2014).

In summary, analysis of transcription factor target genes at each stage of beta cell differentiation detects known and novel regulators of beta cell development, including possible T2D candidate gene effectors.

2.3.5. Per-stage enrichment in genes within T2D-associated loci

Although >200 genetic loci showing association with T2D susceptibility have been identified, most map to non-coding regions of the genome and are likely to exert their effects through altered regulation of nearby genes. I examined the transcriptomic data for evidence of developmental stage-specific enrichment of genes near these loci.

I first concentrated on genes whose coding sequence was at least partly contained within 99% credible intervals from T2D GWAS fine-mapping efforts on the basis that these represented a set of genes likely to be substantially enriched for T2D effector transcripts. Of the 963 genes mapping within credible sets, 568 were differentially expressed at any stage, and from those most (448; 79%) showed differential expression that peaked before the final beta-like cell stage (max. = 155 at iPSC and min. = 28 at PE). GSEA, which considers the strength of association at T2D GWAS signals (see **Section 2.2.7**), did not demonstrate enrichment at any stage (**Figure 2.7 a**), neither when the GWAS genes also implicated in monogenic diabetes (**Table 2.5**) were excluded. Using a complementary GSEA approach that ranked the strength of differential expression of each gene (in *q* value) per stage, I tested all differentially expressed genes assigned to each stage for enrichment among T2D GWAS loci; this analysis highlighted the beta-like cell stage (q=0.016, **Figure 2.7 a**). This enrichment remained significant (q=0.026) after monogenic diabetes genes had been excluded.

As an additional analytical approach, I performed a hypergeometric test for enrichment in the same set of 963 T2D credible interval genes. As opposed to the GSEA method above, this analysis does not consider the strength of differential expression (or of association with T2D) above the significance threshold. This test demonstrated that genes showing differential expression at the endocrine-like cell (EN) stage were enriched (compared with background) for location within T2D credible intervals ("T2D (all)" permuted *p* value=0.006; **Figure 2.7 b**). Excluding the monogenic diabetes genes, and those that fell in the same credible interval, from the differentially expressed genes at each stage maintained the significance (permuted *p* value=0.006).

I repeated the enrichment test using a subset of 15 T2D GWAS loci for which the evidence from physiological studies points most emphatically to risk of T2D mediated via reduced insulin secretion (**Table 2.4**) (Dimas *et al.*, 2014; Wood *et al.*, 2017). In this analysis, differentially expressed genes at the beta-like cell stage became significantly enriched ("T2D (beta-cell)" permuted *p* value=0.02; **Figure 2.7 b**). This enrichment was eliminated after excluding the monogenic diabetes genes and those within the same credible interval. Using the same approach of sampling from the hypergeometric distribution, I also detected enrichment for this stage in genes mapping to credible intervals for 16 loci significantly associated with fasting glucose (FG) (permuted *p* value=0.0002; **Figure 2.7 b**).

T2D-associated signals falling in non-coding regions have a presumed regulatory function: some may map to tissue-specific enhancers acting some distance away from their effector transcripts (Gaulton *et al.*, 2015). However, consistent with observations that many regulatory GWAS effects operate at relatively short distances (van de Bunt *et al.*, 2015; Rusu *et al.*, 2017; Carrat *et al.*, 2017; Roman *et al.*, 2017; Kycia *et al.*, 2018), I found attenuation of these enrichment signals as I extended the analyses to include genes mapping at increasing distance from the credible intervals (see **Section 2.2.7**), both for genes in all T2D credible intervals and for the subset implicated in beta cell function (**Figure 2.8**).

In conclusion, two different enrichment approaches have highlighted the BLC stage and the more immature EN stage as significantly enriched in T2D-associated signals, while only the BLC stage was enriched in genes within loci associated with fasting glucose.

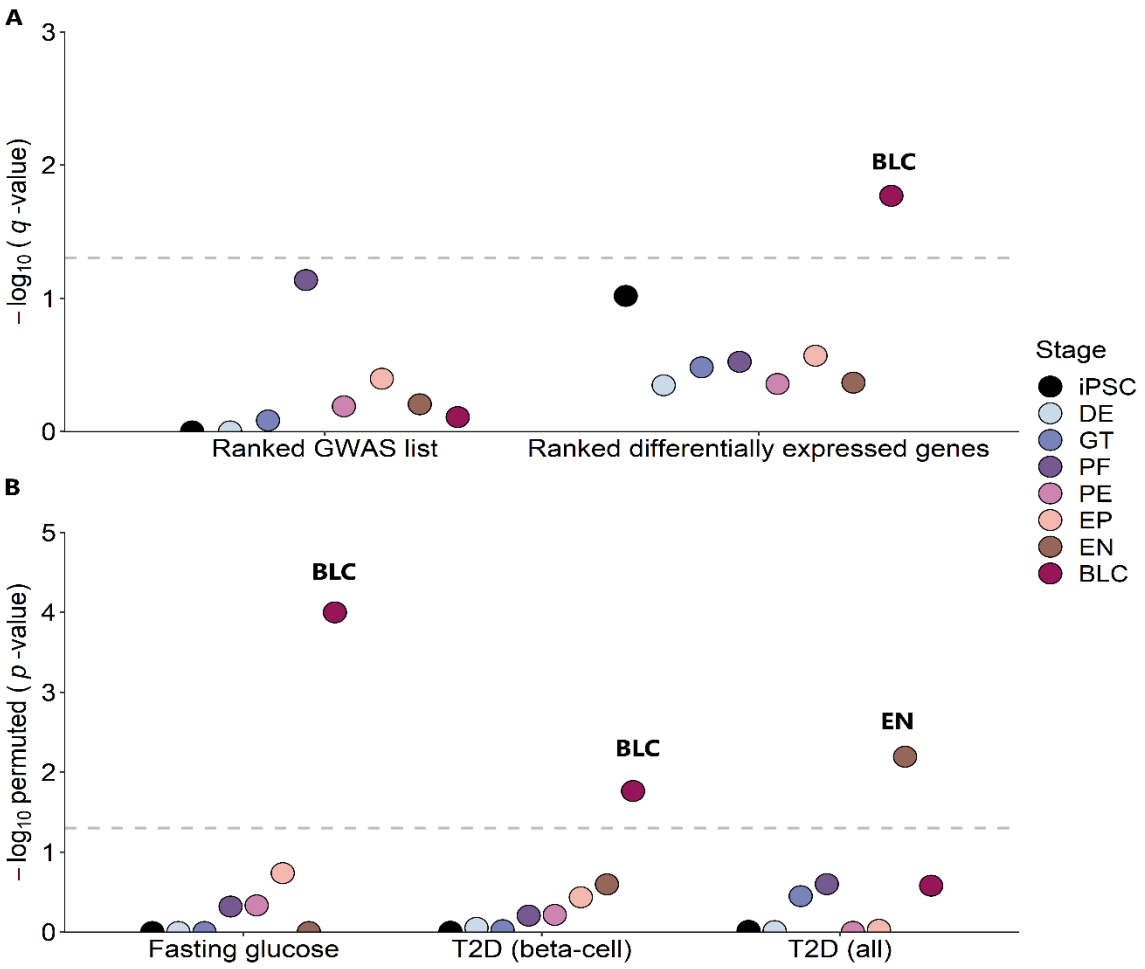
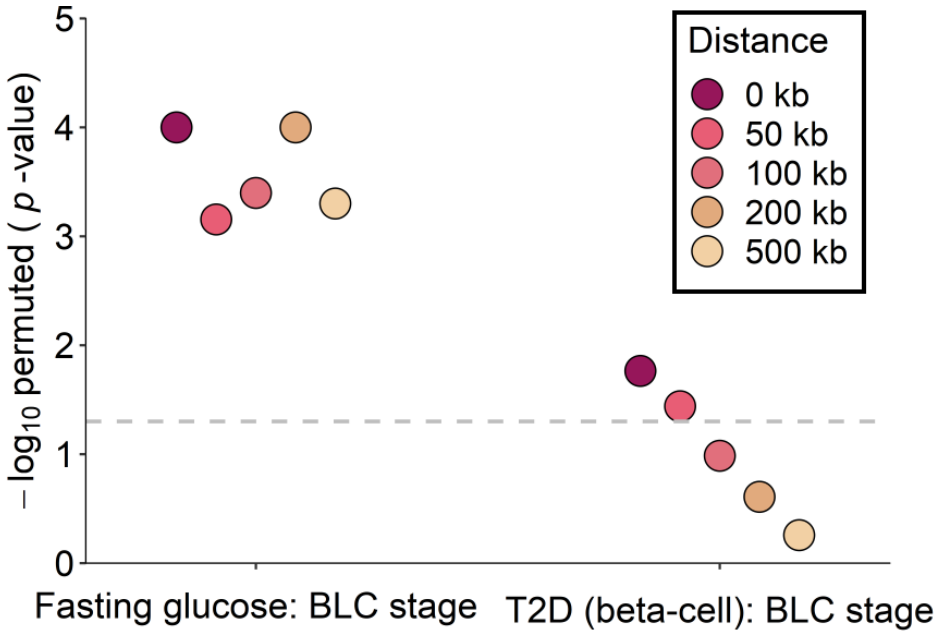


Figure 2.7. Both developing and mature islet-like cells are enriched for genes within T2D-associated loci. (**a**) Results from the GSEA. SNPs from the T2D GWAS meta-analysis of 32 GWAS (Mahajan *et al*., 2018a) (243 loci) were mapped to genes, and T2D association scores were calculated for each gene using MAGENTA. Two complementary analyses were performed: enrichment of all genes ordered by their MAGENTA scores in sets of differentially expressed genes for each stage (Ranked T2D GWAS list), and enrichment of differentially expressed genes per stage (ordered by *q* value) in significant ($p<0.05$ by MAGENTA) gene scores (Ranked differentially expressed genes). The *y*-axis represents the results of the GSEA in FDR-adjusted *p* values (*q* values, $-\log_{10}$). The horizontal grey dashed line marks the 5% significance threshold. (**b**) Results for the hypergeometric enrichment analysis. Enrichment was tested for all differentially expressed genes per stage in the 243 T2D credible intervals [T2D (all)] from a meta-analysis of 32 GWAS (Mahajan *et al.,* 2018a) and the 16 fasting glucose credible intervals (Fasting glucose) from ENGAGE (Horikoshi *et al.*, 2015), and for all differentially expressed genes in only physiological T2D loci [T2D (beta cell); Table 2.4). I consider beta cell function loci as 15 loci influencing hyperglycaemia, beta cell function and insulin processing (Dimas *et al.*, 2014; Wood *et al.*, 2017). The *y*-axis represents the results of the hypergeometric test in permuted *p* values ($-\log_{10}$). The horizontal grey dashed line marks the 5% significance threshold. T2D, T2D; iPSC, induced pluripotent stem cell, DE, definitive endoderm; GT, primitive gut tube; PF, posterior foregut; PE, pancreatic endoderm; EP, endocrine progenitor; EN, endocrine-like cell; BLC, beta-like cell.

a



b

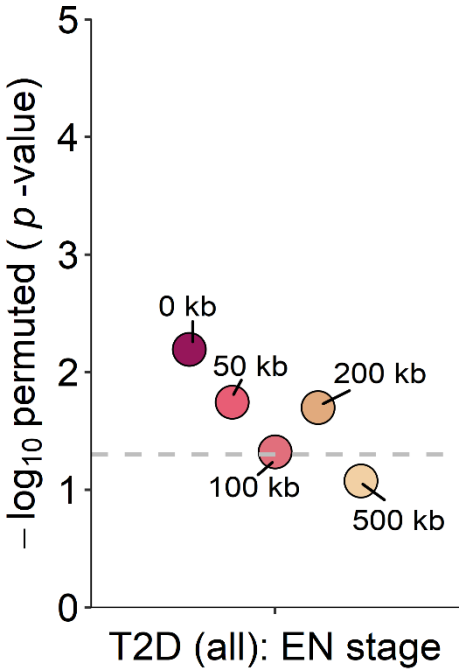


Figure 2.8. Enrichment results in genes inside variable distances around the T2D and fasting glucose credible regions, for differentially expressed genes assigned to the BLC stage (a) or EN stage (b). Enrichment was tested for all differentially expressed genes in the 243 T2D credible intervals ("T2D [all]") from a meta-analysis of 32 GWAS (Mahajan et al., 2018a) and the 16 fasting glucose credible intervals ("Fasting glucose") from ENGAGE (Horikoshi *et al.*, 2015), and for all differentially expressed genes in only physiological T2D loci ("T2D [beta-cell]"), Table 2.4). I consider beta cell function loci 15 loci influencing hyperglycaemia, beta-cell function, and insulin processing (Dimas *et al.*, 2014; Wood *et al.*, 2017). The y-axis represents the results of the hypergeometric test in permuted *p* values (-log10). The horizontal grey dashed line marks the 5% significance threshold.

2.4. Discussion

The main results from this chapter are:

- The hiPSC-derived differentiation model used in this thesis recapitulates islet development at the transcriptomic level.
- *REST* and *CREBBP* are novel potential regulators of islet development.
- The EN and BLC stage are enriched in genes mapping to T2D-associated loci.

In this study, I conducted a transcriptomic analysis of an *in vitro* differentiated cellular model that I have shown is able to recapitulate stages of islet development. The comparative analysis of cells generated using improved versions of differentiation protocols, alongside endocrine cells matured *in vivo* (Xie *et al.,* 2013) (the current benchmark for most functionally-mature endocrine pancreas-like cells) revealed how these advances in differentiation are reflected in the transcriptome. This was particularly true for the later stages of differentiation where there was a clear increase in the expression of genes essential for beta cell function and identity, such as *INS* and *MAFA*.

I have shown how these models of islet development can highlight genes and pathways involved in pancreas development and diabetes. By integrating the differential expression data with genomic annotations on transcription factor binding and clustering of longitudinal expression, I identified novel potential regulators (particularly members of the Wnt signalling pathway like *REST* and *CREBBP*) orchestrating gene expression patterns within the different developmental stages. Such transcriptomic analysis can also illuminate the mechanisms of action for monogenic diabetes genes and inform the search for novel MODY genes that influence the same pathways.

The notable overlap between monogenic diabetes genes and those mapping within T2D-associated loci supports the hypothesis that some component of T2D susceptibility arises through impairment of islet development (Fuchsberger *et al.*, 2016). The final stage in the islet development model (featuring cells expressing genes encoding the machinery to support glucose-stimulated insulin secretion) was enriched for genes

mapping to GWAS signals for both T2D and FG, while the hypergeometric approach for all genes within T2D loci indicated that the more immature EN stage was also enriched. These data are consistent with the concept that T2D-associated loci act both on the adult islet and during earlier developmental stages.

Two caveats of the enrichment analysis become obvious when the data is compared to the previously published analysis of this model (Perez-Alcantara *et al.*, 2018). There, the posterior foregut stage was found to be enriched using the GSEA approach that ranks the T2D-associated genes by MAGENTA, while in the re-analysis using the most recent GWAS results the enrichment is lost. This has occurred because GSEA is sensitive to the position of this genes in the ranking, and with more loci in the analysis there were more "T2D-associated genes" included in this ranked list. Due to this change, the genes that were assigned to the PF stage by MAGENTA (e.g. *HHEX* or *TCF7L2*) were not located at the very top of the ranked list anymore, thus reducing the enrichment score of this stage. The second problem was that the assignment of differentially expressed genes to a specific stage led to a wide distribution of the "T2D-associated genes" among the immature stages (DE to EN), which diluted the power to detect significant enrichment at stages before the beta-like cell stage. A method that included in each stage every gene that presented a $\log_2$FC over a certain threshold compared to the iPSC baseline, without assigning it exclusively to a stage, would have resulted in more "T2D-associated genes" in each stage to be detected by the enrichment analysis.

Amongst the over 200 T2D risk loci there were several examples where data point to candidate effector transcripts that play a role earlier in development, either pushing the cells towards liver or endocrine pancreas commitment (Gittes, 2009). *TCF7L2* and *HHEX* were examples of genes most highly expressed during this early developmental stage, and T2D-risk alleles at these loci influencing early expression of these genes could influence development in multiple metabolic tissues, as supported by studies in mice and cellular models. Investigating these and other diabetes-relevant genes in stem cell-derived models can help decipher the role of multi-organ developmental effects on diabetes pathogenesis.

3. Chromatin modifications in an hiPSC model of beta cell differentiation

3.1. Introduction

3.1.1. Regulatory function of chromatin modifications

As described in the first chapter, chromatin modifications regulate gene expression through histone repositioning, DNA methylation and histone modifications, among other processes (Campbell and Hoffmann, 2016). These marks contribute to the activation or inhibition of transcription, either at the promoters or at locations further away, often in a cell type-specific manner and creating complex synergies among themselves and with the products of the genes they regulate. In this chapter I study a subset of modifications including chromatin accessibility (measured by ATAC-seq), DNA methylation (using the Illumina 850k EPIC array) and H3K27ac histone marks (by ChIP-seq). ATAC-seq and H3K27ac ChIP-seq give as an output peaks with variable counts that provide a continuous measure of accessibility (e.g. capacity to be bound by transcription factors) and regulatory activity respectively. Both are enriched at promoter and enhancers (Thurman *et al.*, 2012; Tsompana and Buck, 2014). DNA methylation measurement by the EPIC array on the other hand gives a continuous distribution of intensity between zero (fully unmethylated) and one (fully methylated) at each fixed probe position. Although it is generally believed that methylated promoters have reduced gene expression (Suzuki and Bird, 2008; Jones, 2012; Schübeler, 2015; Korthauer and Irizarry, 2018) the inverse is less clear: often genes with unmethylated promoters are not expressed (Neiman *et al.*, 2017). The causal directionality of these regulatory processes and gene expression is not established (Kilpinen and Dermitzakis, 2012).

3.1.2. The influence of chromatin modifications in T2D susceptibility and islet development

Several studies have used chromatin modifications to try to understand the contribution of the regulome to T2D susceptibility. Chromatin modification patterns in islets can facilitate disease study, because functional analyses of T2D-associated SNPs have found that many are linked to variation in insulin secretion, beta cell function and islet-specific gene

expression (Scott *et al.*, 2007; Zeggini *et al.*, 2007; Voight *et al.*, 2010; Pasquali *et al.*, 2014; Dimas *et al.*, 2014; Wood *et al.*, 2017; Varshney *et al.*, 2017) and enriched in islet-specific regulatory regions (Roadmap Epigenomics Consortium, 2015; Trynka *et al.*, 2013; Stitzel *et al.*, 2015). Accordingly, islet studies using open chromatin data consistently detected an enrichment in enhancers and promoters, particularly in clusters or "stretch" enhancers (Parker *et al.*, 2013; Pasquali *et al.*, 2014; Gaulton *et al.*, 2015). This includes variants that act as quantitative trait loci for chromatin accessibility (Khetan *et al.*, 2018).

Thurner and colleagues (2018) recently evaluated whole genome bisulfite DNA methylation data and ATAC-seq from islets of 10 and 17 human donors respectively and found a similar enrichment in islet enhancers (characterized by ATAC-seq peaks and low methylated regions) that was driven mostly by accessible chromatin. They also used this information to lever the prior probabilities in the T2D credible sets from GWAS, which resulted in a reduction of the number of likely causal variants per locus. Testing them for allelic imbalance resulted in significant effects in chromatin accessibility at enhancer variants in *CDC123* / *CAMK1D*, *KLHDC5* and *ADCY5* loci. A recent multi-omic analysis of the immortalized beta cell line EndoCβ1 (Lawlor *et al.*, 2019) performed allelic imbalance for both H3K27ac and ATAC-seq, identifying 50 and 185 significant SNPs respectively, and then used ChIA-PET interactions to link the enhancer regions to genes. None of those overlapped the three variants from Thurner and colleagues, which may reflect the modest correlation of chromatin accessibility and H3K27ac profiles between this cell type and adult islets (Pearson's $r<0.70$). Based on the enrichment in binding of important transcription factors at regulatory regions in beta cell differentiation (e.g. PDX1) in EndoCβ1 compared to adult islets, this suggests a more immature state of these cells.

From the perspective of DNA methylation, it was found that 97% of differentially methylated CpG sites between T2D patients and healthy controls at T2D loci were hypomethylated, supporting the role of these modifications in the mediation of T2D risk (Dayeh *et al.*, 2014; Volkmar *et al.*2012). These differences in methylation affect specific genes and influence their function, including at important islet function and development genes such as *INS* (Kuroda *et al.*, 2009; Yang *et al.*, 2011), *PDX1* (Yang *et al.*, 2012), *GLP1R* (Hall *et al.*, 2013)

and *KCNQ1* (Travers *et al.*, 2013), all in T2D or monogenic diabetes associated variants. Nevertheless, due to the redundant role of regulatory marks in the genome (Ahsendorf *et al.*, 2017), other processes such as chromatin accessibility could be more influential towards these changes in gene expression and disease risk (Thurner *et al.*, 2018).

HiPSC models of islet development can also be used to interrogate the influence of chromatin modifications in this process. Xie and colleagues (2013) examined RNA-seq, H3K27me3 and H3K4me3 chromatin marks at several stages of an islet differentiation model. They observed a more islet-like transcriptome and better insulin secretion in response to glucose in cells differentiated *in vivo* compared to *in vitro*. Furthermore, when they looked closely at the histone marks they observed that a significant proportion of the poised or bivalent promoters that are usually resolved towards activation or repression in differentiated islets remained poised *in vitro* compared *to in vivo*, showing how these chromatin marks can influence gene expression and alter differentiation.

3.1.3. Aims of this chapter

The main aim of this chapter was to study the possible influence of regulatory marks in islet differentiation and T2D risk. To do this, I have:

- Assessed the quality of the ATAC-seq, H3K27ac ChIP-seq and 850k array methylation data in the hiPSC-derived model of beta cell differentiation.
- Characterized the regulatory landscape at every step of iPS cell differentiation towards beta cells at the level of each dataset.
- Evaluated the overlap and enrichment of the regulatory regions in T2D-associated signals.
- For ATAC-seq data, used the regulatory information to lever the probabilities of the T2D-associated variants to improve fine-mapping.

3.2. Methods

All sample preparation steps for ATAC-seq and DNA methylation before sequencing were performed by Dr Nicola Beer or Dr Matthias Thurner of the McCarthy group, while I performed most of the steps of H3K27ac ChIP. Library preparation and sequencing for all samples was performed by High-Throughput Genomics group at the Wellcome Centre for Human Genetics (Oxford, UK).

The three datasets originate from the same differentiations that originated the RNA-seq data (see **Chapter 2**), from iPSC through the following stages of endocrine pancreas development, in order: definitive endoderm (DE), primitive gut tube (GT), posterior foregut (PF), pancreatic endoderm (PE), endocrine progenitors (EP), endocrine-like cells (EN), and beta-like cells (BLC). Each stage had replicates originating from 3 donors (SBAd3.1, SBAd2.1 and SBNeo1.1) for a total of 24 samples harvested.

Plots were created with ggplot2 (Wickham *et al.*, 2016) in R (v. 3.6.1) unless otherwise stated.

3.2.1. ATAC-seq

3.2.1.1. Sample preparation

Samples were prepared for ATAC-seq by Dr Nicola Beer as previously described (Buenrostro *et al.*, 2013).

3.2.1.2. Sequencing

Libraries for the 24 samples (3 donors per each of 8 stages of differentiation) were sequenced as multiplexed pools on Illumina HiSeq2500 using 50-bp paired-end sequencing to a mean read depth of 111 (sd. = 40) million reads pairs per sample.

3.2.1.3. Peak calling

I performed all analysis methods for ATAC-seq from this point. The raw sequencing reads were processed with the Kundaje lab ATAC-seq pipeline (https://github.com/kundajelab/atac-seq-pipeline), following the ENCODE Consortium recommendations. The pipeline trims the sequencing adaptors, maps the reads to the

human reference genome (build hg19) using Bowtie2 (v. 2.2.6), marks duplicated alignments with Picard tools (v. 1.126), converts the BAM files to tagAlign format using BEDTools (v. 2.26) (Quinlan and Hall, 2010) and calls regions enriched for reads (“peaks”) using MACS2 (v. 2.1.0).

Signal-to-noise was measured by transcription start site (TSS) enrichment score, which is the fold change of average read depth centred at every TSS over the average read depth in 100 bp blocks 1 kb up and downstream.

3.2.1.4. Creating a consensus set of ATAC-seq peaks

Conservative peaks that passed the Irreproducible Discovery Rate (IDR) at the 10% significance threshold for every stage of differentiation (see the start of **Methods**) were overlapped and merged at each genomic location using BEDTools (v. 2.26), creating a consensus peak set of 133,116 peaks. Counts were extracted at each position with featureCounts (v. v1.5.0-p2) from the Subread package (http://subread.sourceforge.net/) (Liao *et al.* 2014). Peaks in autosomal chromosomes with more than 1 count per million (CPM) in all donors of at least one differentiation stage were retained (128,427 peaks).

Counts were normalised using *voom* within the limma package (v 3.38.3; Ritchie *et al.*, 2015) in R (v. 3.6.0) (R Core Team, 2017) before linear correlation and principal component analysis. *voom* normalizes data by modelling the mean-variance relationship per peak and fitting it to the log-CPMs, to account for the unequal variances of the peaks.

3.2.1.5. Clustering of ATAC-seq profiles

Raw counts were normalised using the *varianceStabilizingTransformation* function within the DESeq2 package (v. 1.18.1; Love *et al.*, 2014). Correlation of peak count profiles was calculated using the weighted gene co-expression network analysis (WGCNA) package (v. 1.66; Langfelder and Horvath, 2008) in R, after removing the SBAd2.1 replicates for EP and BLC, as these were driving some of the clustering modules as outliers. Non-default parameters for WGCNA’s *blockwiseModules* function were: power=12, networkType = ”signed hybrid”, minModuleSize=120, mergeCutHeight=0.30.

3.2.2. H3K27ac ChIP-seq

3.2.2.1. Sample preparation

For all three donor replicates in 8 stages of differentiation, cells were harvested and centrifuged, and 1 ml room-temperature PBS with 28 ul 36% formaldehyde was added for a final concentration of 1%. Cells were incubated whilst mixing for 10 min at room temperature. To stop the reaction, 114 ul 1.25 M glycine were added to each sample and the samples were incubated on a rotating wheel for 5 minutes. Tubes were spun 200xg for 10 min 4ºC and kept on ice from this point on. Supernatant was discarded and cells washed twice in 1ml cold PBS, pelleting cells at 200 x g for 10 min 4ºC. The pellet was then snap-frozen and kept at -80 ºC for two years. Dr Nicola Beer from the McCarthy group prepared the samples up to this point.

I performed all subsequent steps. Samples were thawed on ice and chromatin was immunoprecipitated following the protocol of the MAGnify Chromatin Immunoprecipitation System kit (#49-2024, Invitrogen). Briefly, 2 ug of H3K27ac antibody (#ab4729, Abcam) was coupled to 10 ul magnetic beads in 100ul dilution buffer per sample, mixing on rotating wheel at 4 ºC for 7 h. At the same time, samples were lysed following the kit's instructions, transferred to AFA fiber microtubes (#520045, Covaris) and then sonicated in a Covaris M220 sonicator with the following program, pre-tested to shear DNA to <500 bp fragments:

a. Duty cycle: 5%
b. Peak power: 75
c. Cycles per bust: 200
d. Cycle time: 60 seconds
e. Cycles: 8
f. Temperature: 4 ºC

After sonication, samples were transferred to DNA LoBind tubes and spun 20,000 x g at 4ºC for 5 min. Supernatant was removed to fresh DNA LoBind tubes and dilution buffer with protease inhibitors (from the kit) was added to each tube. 80% volume (90 ul) of diluted chromatin was added to microtubes with the antibody coupled to magnetic beads and incubated overnight rotating at 4ºC. The remaining diluted chromatin for each sample was

kept in fridge as input control. The crosslinking of the samples was reversed using the kit's buffers, and purified DNA using QIAquick PCR purification kit (# 28104, Qiagen).

For reverse crosslinking of the input controls, the following elution buffer stock was prepared:

Reagent	**Supplier**	**Final concentration**
Tris-HCl (pH 7.5) (1 M)	Life Tech	20 mM
EDTA (pH 8.0) (0.5 M)	Ambion	5 mM
NaCl	Ambion	50 mM

180 µl of elution buffer stock and 20 ul 10% SDS were added to each tube. Input controls were incubated at 37°C for 30 min before adding 2 µl of Proteinase K (20 mg/µl). After mixing, controls were incubated at 68°C overnight, followed by the addition of 180 ul elution buffer and 20 ul autoclaved milliQ water.

To purify the DNA of input controls, 500 ul of cold 25:24:1 phenol:cholorform:isoamyl alcohol (#P3803, Sigma-Aldrich) were added per control, mixed well and centrifuged at 4°C 20,000 x g 1 min. The upper layer was transferred to a new tube and the previous step was repeated. 20 mg glycogen, 50µl NaAc pH 5.2 and two volumes of 100 % EtOH were added to each tube, which were incubated at -80°C for 30 mins and then centrifuged for 4°C 20,000 x g 30 min. After discarding the supernatant, the pellet was washed with 1 ml 70% ice cold EtOH and centrifuged for 4°C 20,000 x g 10 mins. Finally, the supernatant was discarded and the pellet dried and resuspended, along with the DNA of the matching ChIP samples, in 15 ul autoclaved milliQ water.

3.2.2.2. Sequencing

DNA was prepared as described in **Section 3.2.1.2**. Samples PF and BLC SBNeo1.1 failed the library preparation step, while all input controls were prepared successfully. The remaining samples (n = 22) and input controls (n = 24) per donor and differentiation stage were sequenced as multiplexed pools on four lanes of Illumina HiSeq4000 using 75-bp paired-end sequencing.

3.2.2.3. Peak calling

The raw sequencing reads were processed with the ENCODE histone ChIP-seq pipeline v3 (https://github.com/ENCODE-DCC/chip-seq-pipeline2) after trimming adaptors with Trim Galore (v.0.4.4, https://github.com/FelixKrueger/TrimGalore) using Cutadapt (v. 1.16) (Martin, 2011). The subsequent steps for ChIP samples are analogous to those of the ATAC-seq Kundaje pipeline (**Section 3.2.1.3**), using paired input controls as background to calculate each peak's significance (p-value threshold = 0.01).

3.2.2.4. Creating a consensus set of H3K27ac peaks

"Optimal" naïve overlap peaks (creating from calling peaks on pooled pseudoreplicates) for every stage were overlapped and merged at each genomic location using BEDTools (v. 2.26), creating a consensus set of 280,717 peaks. Counts were extracted at each position with featureCounts (v. v1.5.0-p2) from the Subread package (Liao *et al.* 2014). Peaks in autosomal chromosomes with more than 1 count per million (CPM) in all donors of at least one differentiation stage were retained.

3.2.2.5. Peak count normalization

For Pearson correlation and principal component analysis, counts were normalised using *voom* within the limma package (v 3.38.3; Ritchie *et al.*, 2015) in R.

3.2.3. DNA methylation

3.2.3.1. Islet isolation

In addition to the samples from the iPSC-differentiated model (see the start of **Methods**), 11 adult human islets were retrieved from deceased donors. Six samples ("ISL") were collected by members of the Oxford DRWF Human Islet Isolation Facility, while five ("R") were collected by members of Dr. Patrick McDonald's group at the Alberta Diabetes Institute in Edmonton, Canada. Islet purity was determined by dithizone labelling at the islet isolation facility. High purity samples (>70%) were selected for DNA methylation analysis.

3.2.3.2. Sample preparation and 850k (EPIC) array data processing

For every adult islet and iPSC-differentiated sample (11 of the former and 24 of the latter: 32 in total), DNA was extracted and bisulfite conversion was performed as in Thurner *et al.* (2018) by Dr Matthias Thurner of the McCarthy group. Differentiation samples BLC-SBAd2.1, EP-SBAd2.1 and EP-SBNeo1.1 did not have enough DNA yield to proceed to the next step. All remaining adult islets and differentiation samples were then hybridized to Infinium Methylation EPIC arrays following the Illumina Infinium HD Methylation Assay.

3.2.3.3. Quality control and normalization

I performed all analysis methods from this point. Results from the EPIC array runs were loaded from .idat files using the minfi package (v. 1.30.0; Aryee *et al.*, 2014) in R. Probes were mapped to the human reference genome (build hg19) using the package IlluminaHumanMethylationEPICanno.ilm10b2.hg19 (v. 0.6.0; Hansen, 2016) in R. Low quality probes, probes mapping to the sex chromosomes and cross-reactive probes (Pidsley *et al.*, 2016) were filtered out. Signal intensities of Infinium Type I and II probes were normalised separately for every colour channel using quantile normalization (**Figure 3.1 a**). Beta values were calculated for each probe following the formula:

$$\beta = \frac{Meth}{Meth + Unmeth} \times 100$$

This value was used for plotting and interpretation of methylation intensity at each position. M values were used instead for differential methylation analysis, correlation and principal component analysis (**Figure 3.1 b**):

$$M = log2\frac{\beta}{1-\beta}$$

This ensures the homoscedasticity of the overall methylation intensity distribution and makes it more appropriate for linear modelling (Du *et al.*, 2010).

3.2.3.4. Principal component analysis

Principal component analysis (PCA) of differentiating beta cells was performed removing batch effects present in SBad3.1 replicates using the *removeBatchEffect* function from the limma package (v. 3.40.6; Ritchie *et al.*, 2015) in R. Batch effects were not removed for the

joint analysis of differentiating and adult islets, as most of the variance was driven by biological differences between the former and the latter.

3.2.3.5. Selection of low and high methylated regions

For every donor per stage, mean beta values per probe were calculated and used as input for the *methSeg* function from the methylKit package (v. 1.10.0; Akalin *et al.*, 2012) in R. This function found consecutive, similarly methylated probes using a segmentation algorithm and returned an average value for the whole segment. Minimum segment size was 2 probes. Segments were classified as low methylated regions (LMR; mean beta < 0.5) or high methylated regions (HMR; mean beta > 0.5).

3.2.3.6. Differential methylation analysis of probes and regions

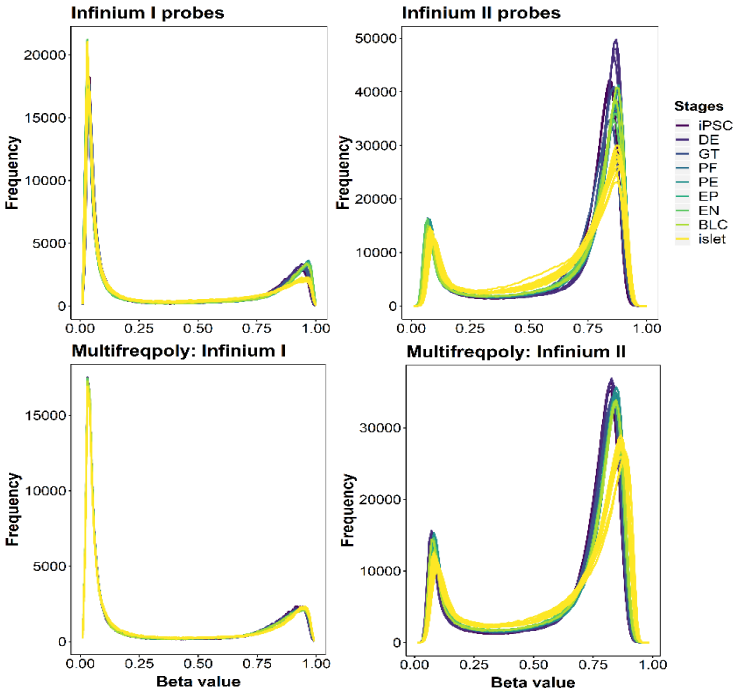
Differentially methylated probes (DMPs) were calculated using M values and fitting contrasts in a linear model with iPSC as the baseline (for differentiation stages) or BLC stage (for adult islets), adjusting for donor effect. These contrasts were used as input for the *eBayes* function in limma for the differential expression analysis**.**

To calculate differentially methylated regions (DMRs), individual CpGs were annotated with the function *cpg.annotate* and used as input for the *dmrcate* function from the DMRcate package (v. 1.20.0; Peters *et al.*, 2015) in R. Non-default parameter for dmrcate was C=2. A minimum of 2 consecutive DMPs less than 1 Kb apart were considered for calling a region. False Discovery Rate (FDR) cut-off, using Benjamini-Hochberg adjustment, for individually significant probes was set to 0.01.

3.2.3.7. Annotation of methylated regions and DMRs

HMRs, LMRs and DMRs were annotated by proximity to CpG island and promoter features and used as input for fGWAS. Annotations and associated plots were performed with the package annotatr (v.1.10.0; Cavalcante and Sartor, 2017) in R using reference genome hg19.

a



b

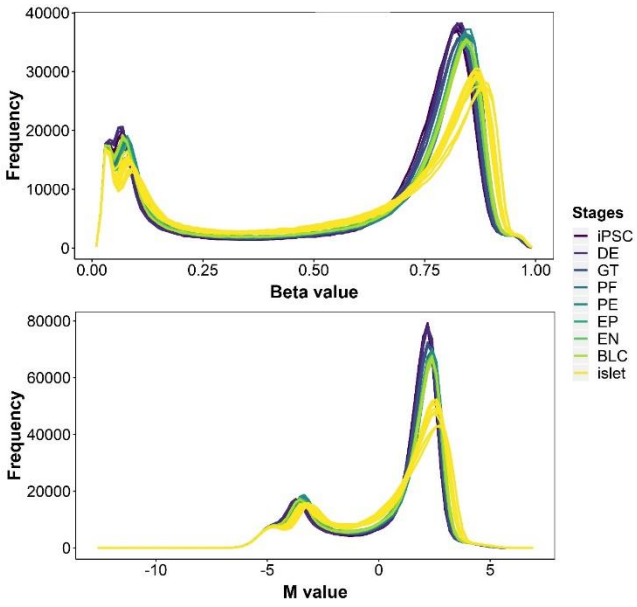


Figure 3.1. Methylation intensity distribution plots. Beta value distributions reflect the frequency (y-axis) of methylated CpGs (high beta values, x-axis) and unmethylated CpGs (low beta values) before (**a**, top) and after (**a**, bottom) normalisation of signal intensities and filtering of probes. The differences in distributions could not be appreciated when plotting the two sets of Illumina probes (Infinium I and II) together (**b**, top), but Infinium I probes had on average lower methylation values (**a**, left) than Infinium II probes (**a**, right). Beta (**b**, top) and M (**b**, bottom) value distributions were calculated after filtering and quantile normalization of the probes.

3.2.4. Significance of overlaps between DNA regulatory marks and genomic features

Significance of overlaps among DNA regulatory marks or with genomic features was calculated using permutation tests using the function *overlapPermTest* from the package regioneR (v. 1.16.5; Gel *et al.*, 2016) in R. ATAC-seq peaks, H3K27ac peaks or low methylated regions were randomly moved 100 times to different genomic coordinates, maintaining their size, and the significance of the true overlap was evaluated against the distribution created this way.

3.2.5. Publicly available data

Publicly available data for ATAC-seq and H3K27ac ChIP-seq peaks (both from the same samples) were taken from ENCODE (www.encodeproject.org, ENCODE Project Consortium, 2012). Their characteristics are listed in the following table:

Type	**Name**	**Biological Replicates**	**ATAC-seq accession**	**H3K27ac ChIP-seq accession**
Cell line	A549	3	ENCSR220ASC	ENCSR783SNV
Tissue	Adult pancreas	3	ENCSR515CDW, ENCSR765MXG, ENCSR152PSA	ENCSR748TFF, ENCSR596PFU, ENCSR520BIM
Tissue	Adult stomach	1	ENCSR337UIU	ENCSR133NBJ
Tissue	Adult thyroid	1	ENCSR914DTI	ENCSR716FUH
Tissue	Adult breast epithelium	1	ENCSR654UYP	ENCSR034ZKE
Tissue	Adult adrenal gland	1	ENCSR548KIL	ENCSR094VJC

Adult islets were not studied as H3K27ac and ATAC-seq peaks from the same samples were not available.

3.2.6. Genomic annotations and pathway enrichment

ATAC-seq and H3K27ac peaks were annotated to genomic regions and plots generated using the *annotatePeak* function of the ChIPseeker package (v. 1.20.0; Yu *et al.*, 2015). in R, defining the promoter region as 2000 bp upstream and 500 bp downstream the TSS using reference genome hg19. Reactome pathway enrichment was calculated using a hypergeometric test within the *enrichPathway* function of the ReactomePA package (v. 1.28.0; Yu *et al.*, 2016) in R, annotating each peak to the closest TSS and using all genes as background. Only terms with Benjamini-Hochberg adjusted p-value < 0.05 were considered significant.

Hyper- and hypomethylated DMRs were tested separately for pathway enrichment per contrast using the *annotatePeak* function followed by *enrichPathway* as described above, using all genes mapping closest to probes from the EPIC array as background.

3.2.7. T2D enrichment in ATAC-seq peak profiles

T2D-associated SNPs were tested for enrichment in ATAC-seq peaks, H3K27ac peaks, ATAC-seq WGCNA modules and methylated regions using fGWAS (v. 0.3.6) (Pickrell, 2014), running it within a Python (v. 3.7.3) handler (https://github.com/kjgaulton/pipelines/tree/master/fgwas_workflow). T2D-associated SNPs were taken from the European ancestry DIAMANTE analysis (~ 900,000 participants; Mahajan *et al.*, 2018a).

3.2.8. Enrichment in transcription factor binding motifs

Each set of DNA regulatory regions was tested for enrichment in transcription factor binding motifs using *findMotifsGenome.pl* from HOMER (v. 4.10) (Heinz *et al.* 2010). I used a list of 1,614 unique position weight matrices (PWMs) created by Grud *et al.* (2018). Briefly, they retained every PWMs from HOMER, Cis-BP (Weirauch *et al.* 2014) and HOCOMOCO (Kulakovskiy *et al.* 2016) that could be assigned to a human transcription factor, merging high correlated PWMs using hierarchical clustering. The PWMs for most of the remaining known transcription factors in the genome were inferred using ZifRC (Najafabadi *et al.* 2015) and by DNA-binding domain similarity. Only those PWMs corresponding to transcription

factors with expression higher than 1 CPM were retained in our analysis (1,330 PWMs). Motifs were scanned by *findMotifsGenome.pl* within the given (variable) size of the peaks. *De novo* motifs were then matched to the filtered list of PWMs, considering significant those motifs with enrichment p-values < 10^{-10} and matching scores equal or over 0.70. Motifs belonging to the same family (with similar PWMs) were grouped together.

3.2.9. Redefining 99% credible sets

The parameter estimates from the best joint model of fGWAS were used to calculate functional credible sets based on the enriched ATAC-seq peak modules. These estimates were used as priors to recalculate the genetic posterior probabilities in the 99% credible sets (that is, variants in each locus that collectively account for 99% posterior probability of association with T2D) as in Mahajan *et al.* (2018). Regions with one distinct signal were expanded 500kb up and downstream from the index variant position, creating non-overlapping intervals of ~1 Mb size.

fGWAS assumes that there is no more than one causal variant per block of 5,000 SNPs, but there are some where more than one distinct signal has been fine-mapped. Because of this, the association p-values of the secondary signals were recalculated by Dr. Anubha Mahajan at the McCarthy group by conditioning out those with higher significance. In this way, functional posterior probabilities were calculated for each likely causal variant in the credible set. The fGWAS call had the following parameters:

-i [specific annotated GWAS] -cc -print -onlyp -w [ATAC-seq module 1]+ [ATAC-seq module 2]+ … + [ATAC-seq module n] -bed [specific 1Mb regions]

3.2.10. Transcription factor binding at SNPs

T2D-associated variants within each regulatory region (ATAC-seq peaks, H3K27ac peaks or hypomethylated DMRs) were scanned for binding of transcription factors (same PWMs as in **Section 3.2.8**) using *findMotifsGenome.pl* from HOMER, centering the search on each variant and extending DNA sequences up and downstream by 15 bp. Only TFs with probability of match > 95% and that overlapped the SNP were retained.

T2D-associated variants within each regulatory region (ATAC-seq peaks, H3K27ac peaks or hypomethylated DMRs) were tested for altered binding of transcription factors (TFs) depending on their allele with SNPSelect from SNP2TFBS (https://ccg.epfl.ch/snp2tfbs/snpselect.php; Kumar *et al.*, 2017). Motifs used were from JASPAR Core Vertebrate 2014 PWM (Mathelier *et al.*, 2014) from MEME motif database version 12.1 (Bailey and Elkan, 1994).

3.2.11. Allelic imbalance

Allelic imbalance was calculated for ATAC-seq data at each T2D 99% credible set variant per stage (see **Section 3.2.9**) using ASEQ (v. 1.1.11). Only variants for which 2 donor replicates were heterozygous were considered, and a minimum of 5 sequencing reads overlapping each allele were needed. The significance of allelic imbalance was calculated using a negative binomial distribution test, keeping only variants with FDR-adjusted p-value of 0.05. This resulted in 143 variants presenting significant allelic imbalance for chromatin accessibility.

3.3. Results

3.3.1. Quality control

Throughout this chapter stages of beta cell differentiation per donor replicate (see beginning of **Methods**) will be referred to in the format *stage-donor*.

3.3.1.1. Quality control of ATAC-seq data

A total of 24 samples were processed with the ATAC-seq method, and reads were mapped and filtered before peak calling (see **Section 3.2.1.2** and **3.2.1.3**). 99.7% of the reads on average could be mapped to the hg19 genome and all samples met ENCODE standards (**Figure 3.2 a**). Of those, an average of 45% could not be paired. Of the remaining filtered reads, a mean of 32% (sd. = 8 %) were duplicates, and after removing those, 22% of the initial reads (sd. = 4%) on average remained. While ENCODE currently has no minimum percentage requirements for ATAC-seq non-duplicated and mapped reads, these translated to a mean of 24 (sd. = 8.9) million usable fragments per sample and thus, 16 samples out of 24 had fewer than the ideal number (**Figure 3.2 b**). Nevertheless, this is a broad recommendation dependant on other quality control parameters, and therefore no samples were discarded at this point. Raw sequencing reads correlated (Pearson's r = 0.88) with the number of reads passing all quality control steps.

Fragment size distribution had a large proportion of reads pairs corresponding to nucleosome-free fragments (<100 bp) and allowed the visualization of nucleosome occupancy periodicity for most samples (**Figure 3.3**), indicating that their read pairs had good length quality. Three samples presented lower quality: EP-SBNeo1.1 and BLC-SBAd2.1 only showed periodicity up to the single nucleosome level, and for EP-SBAd2.1 there was no apparent periodicity.

Transcription start site (TSS) enrichment scores (see **Section 3.2.1.3**) for all samples were > 10, the ideal ENCODE threshold indicating low signal-to-noise, except for EP-SBad2.1 and EN-SBad2.1 which had "concerning" but acceptable levels (TSS scores = 4.2 and 6.4 respectively).

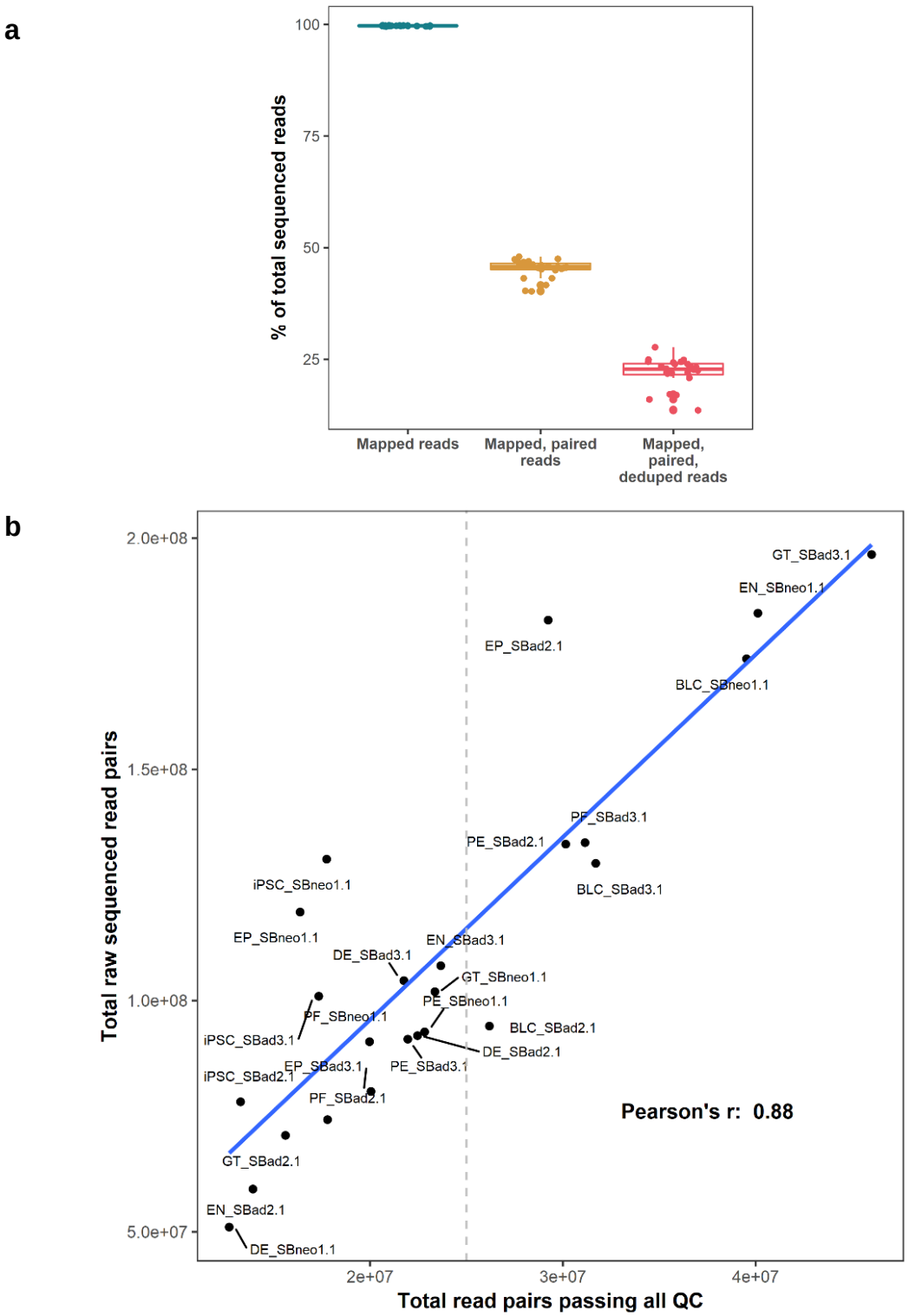


Figure 3.2. ATAC-seq quality control of reads. a) Boxplots of the percentage of reads remaining at every QC filtering step: after mapping (left), pairing (centre) and removing duplicates (right). Dots represent individual samples. **b**) Scatterplot of raw read pairs sequenced (y axis) vs total read pairs passing all QC (x axis) shows high Pearson´s correlation. The dotted line denotes the minimum number of read pairs recommended by ENCODE (25 million).

Peaks within each stage and donor contained a mean of 25% of reads (sd. = 6%) that passed all quality control measures (**Figure 3.4**), with a mean number of 239,000 (sd. = 17,000) peaks. ENCODE has no recommendations regarding these two parameters. Samples with more peaks and less reads, and conversely those with more reads and less peaks called, had a smaller fraction of reads in peaks, reflecting how a large amount of background reads affected the peak calling process.

Peaks that were consistent among biological replicates for every stage of differentiation (IDR > 10%, see **Section 3.2.1.4**) were retained, and overlapping peaks from this set across stages were merged to get a "consensus set" of peaks. This set had 133,116 peaks across stages, of which ~16,500 (12%) were shared among all (**Figure 3.5**). There was a mean of ~50,000 peaks per stage, with 12% on average that were unique to each stage ("stage-specific"; min = 3% in EP, max = 26 % in iPSC). As expected, peaks that were not stage-specific were more frequently shared among neighbouring developmental stages.

At every peak from the consensus set, reads that passed all QC were counted at every stage and replicate to get a measure of raw reads per peak, analogous to the gene counts from the previous chapter. Four per cent of all peaks were filtered out and counts normalized (see **Section 3.2.1.4)**, leaving 128,427 peaks for downstream analysis. Linear correlation (**Figure 3.6**) showed that samples within the same stage had Pearson's r above 0.8, except for EP replicate SBAd2.1 (which had 0.63 and 0.68 with SBAd3.1 and SBneo1.1 respectively). This was caused in part by the smaller fraction of reads in peaks, produced by high background reads that in turn reduced the number of peaks passing IDR, subsequently producing the smaller consensus peak set. When overlapped with the count files it caused higher disparity in the read counts in the consensus set of peaks among replicates with noisy backgrounds.

Principal component analysis of normalized counts (**Figure 3.7**) show how the stages are aligned from iPSC until the more adult stages EN and BLC, the last two clustering closer to each other probably due to their similar profiles. Samples cluster by stage and not by replicate, highlighting the distinct nature of the chromatin landscape at every step of beta

cell development. Nevertheless, samples EP and BLC SBAd2.1 appear further away from the replicates of their respective stages, consistent with the observations in the Pearson correlation plot and the quality of fragment length (**Figures 3.6 and 3.3**).

Taken together, quality control measures indicate good overall quality of the ATAC-seq samples. No sample presented concerning quality scores throughout all measures, and therefore all of them were analysed at a qualitative (presence or absence of peaks at a genomic position) and quantitative level, as will be shown in the next pages.

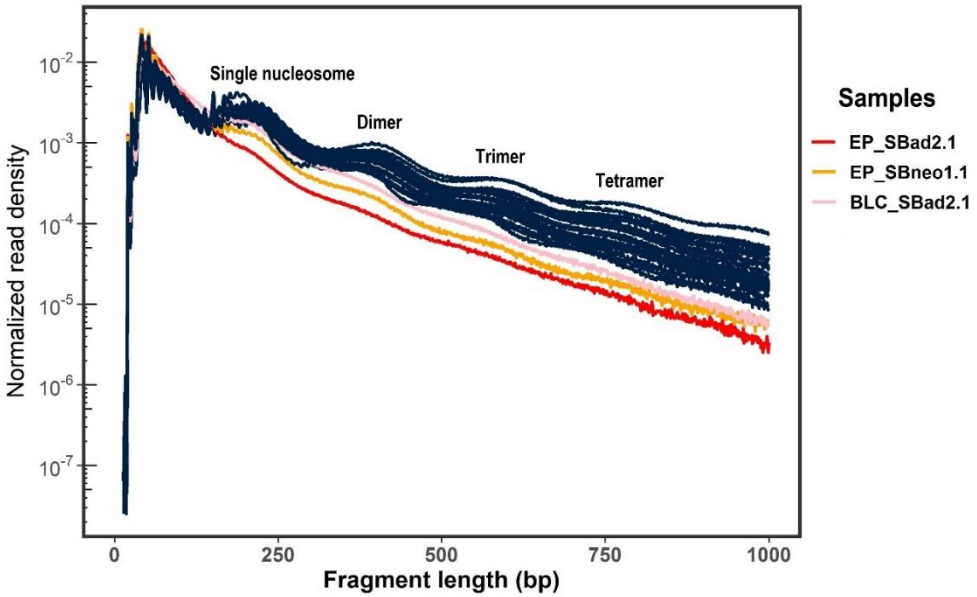


Figure 3.3. ATAC-seq fragment size distribution. Read density (normalized to all reads per sample, y-axis) distribution per fragment length (x-axis). Periodicity of fragment length indicates nucleosome-free reads (<100 bp) and those corresponding to stretches of DNA occupied by one ("single nucleosome"), two ("dimer") or more nucleosomes. Samples for which the fragment length distribution corresponding to nucleosome occupancy is less apparent are highlighted, and samples with good fragment quality are coloured black. Only fragments up to 1000 bp shown.

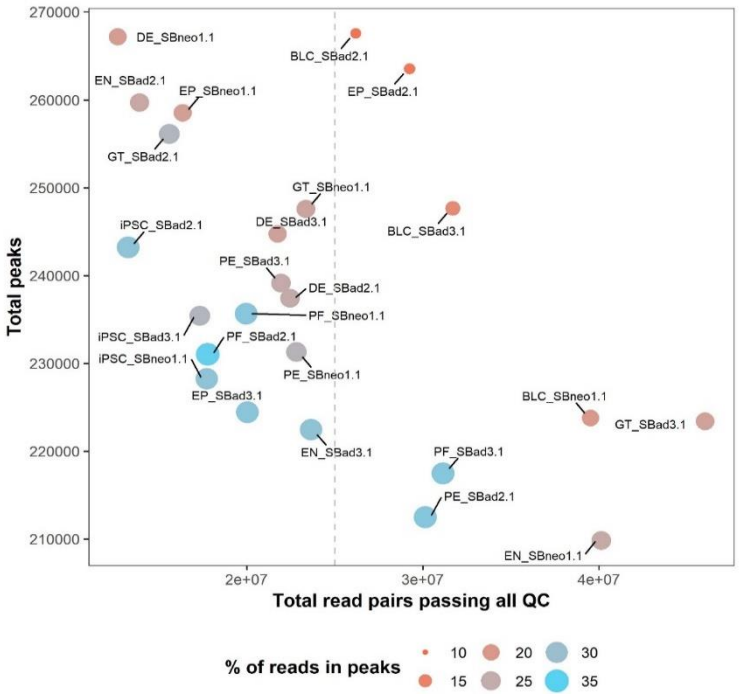


Figure 3.4. Quality control of peak calling. Relationship between read (x-axis) and peak number (y-axis). Area: percentage of reads in peaks (the recommended value by ENCODE being 30%, while >20% is acceptable). The dotted line denotes the minimum number of read pairs recommended by ENCODE (25 million).

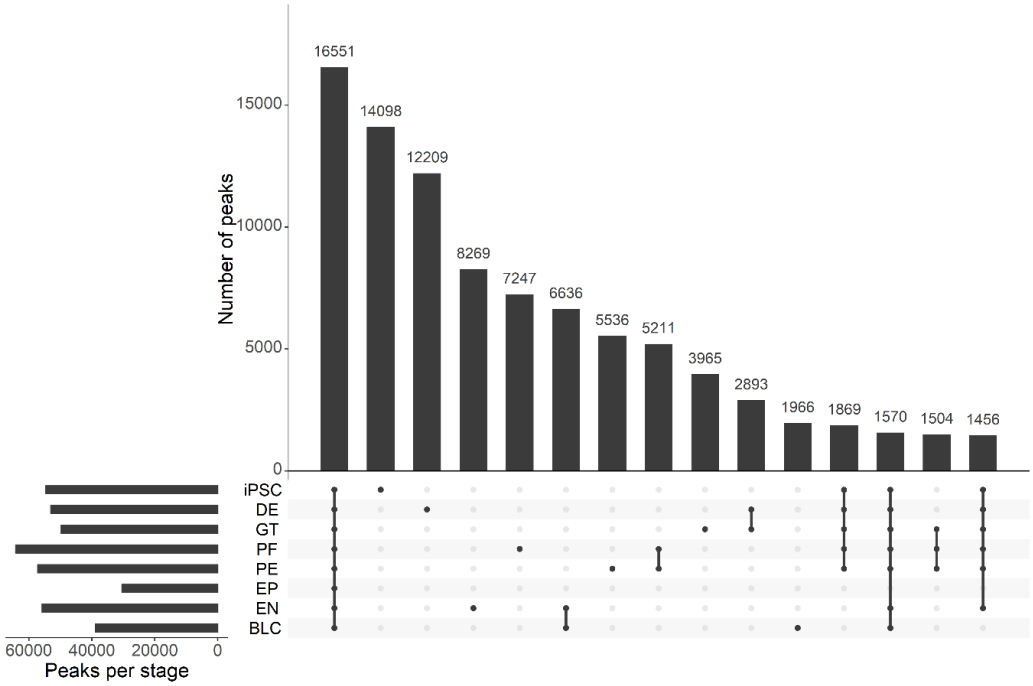


Figure 3.5. Consensus set of ATAC-seq peaks. Dotted lines indicate the number of peaks belonging to each stage combination. The most abundant category was that of the peaks that were present in all stages. The number of peaks per stage is shown in the left barplot. Only the 15 most abundant peak sets are shown. EP stage had only 885 stage-specific peaks (not shown).

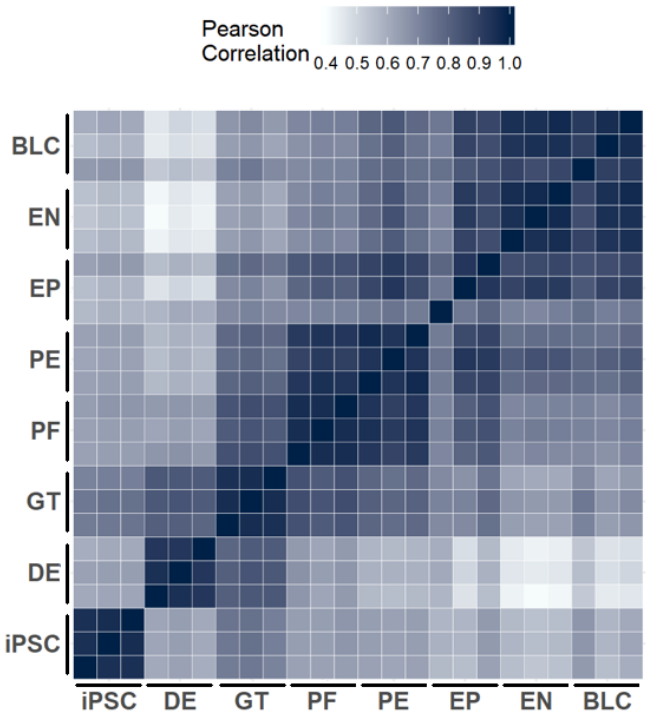


Figure 3.6. Pairwise ATAC-seq peak count correlation. Replicates show high (>0.8) linear correlation (measured by Pearson's r) of normalised read counts within stages, except for the EP stage. Replicates are ordered within each stage (down to up and left to right) as SBad2.1, SBad3.1 and SBneo1.1.

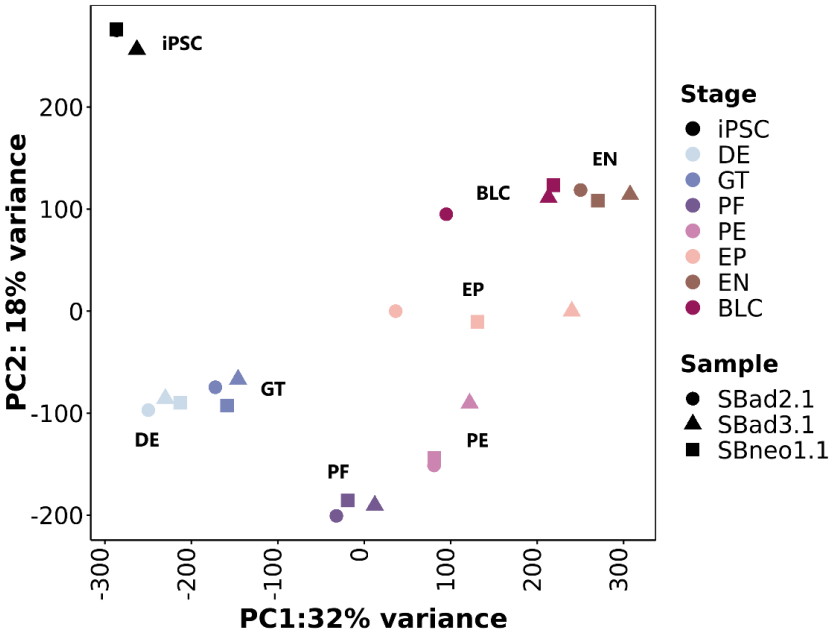


Figure 3.7. Principal component analysis of ATAC-seq data. Shown are the first two principal components of the normalized counts per peak from the 1 CPM-filtered consensus set.

3.3.1.2. Quality control of H3K27ac ChIP-seq data

Three donor replicates per stage from iPSC to BLC (see start of **Methods**) were immunoprecipitated with antibodies targeting the H3K27ac histone mark (see **Section 3.2.2.1)**. Of those, 22 ChIP samples (PF and BLC SBNeo1.1 failed the library preparation step, see **Section 3.2.2.2**) were sequenced, along with 24 paired input controls that were not immunoprecipitated, to a mean depth of 28 (sd. = 2.3) million reads pairs. Reads were mapped and filtered before peak calling (see **Section 3.2.2.3**). 97% of the reads on average could be mapped, and of those a mean of 5.5% could not be paired. Of the remaining filtered reads, an average of 63% (sd. = 14 %) reads in ChIP samples were duplicates, while input controls had a mean of 15% (sd. = 3 %) duplicated reads. On average, 31% of all reads (sd. = 12%) passed all quality control parameters in ChIP samples (**Figure 3.8**), while in input controls (which were not immunoprecipitated and thus had more DNA for sequencing) the percentage was 69% (sd. = 7.7%). As with ATAC-seq, ENCODE currently has no minimum percentage requirements for H3K27ac ChIP-seq non-duplicated and mapped reads, but recommend a minimum of 20 million usable read pairs.

After removing duplicates, a mean of 9 (sd. = 3.5) million read pairs per ChIP sample remained, 18 (sd. = 2.0) million for input controls. This meant that no sample or control passed this threshold but, as it occurred with the ATAC-seq guidelines, failing to meet these recommendations are not enough reason to discard any sample at this point. A relationship between raw sequenced reads and those that pass all QC is shown in **Figure 3.9.** Despite ChIP samples having a higher number of raw reads, they had less usable ones than input controls after passing all QC due to the latter having higher percentage of duplicates.

Despite these concerns, the use of paired input controls as background in the peak calling process ensured only high confidence peaks (and not those which are product of artefacts during cross-linking, sonication and sequencing) were called for each individual donor and stage (see **Section 3.2.2.3**). A mean of 17% of reads (sd. = 4.4%) that passed all quality control measures fell in peaks (**Figure 3.10**), and samples had a mean of 183,404 (sd. = 53,283) peaks. Unlike for ATAC-seq, the official ENCODE histone pipeline disables the IDR step of histone ChIP-seq peaks due to artefacts in the calculations of the peak ranks.

Because of this, an overlap of biological and pooled replicates was calculated instead for each stage. From these, overlapping peaks across stages were merged (see **Section 3.2.2.4**). This set had 280,717 peaks across stages, with a mean of ~50,560 peaks per stage, 51% being specific to each one (min = 42% in EN, max = 63 % in DE). In contrast to ATAC-seq consensus peaks, only 981 peaks were shared among all stages and stage-specific peaks were most abundant (**Figure 3.11**).

Peaks with counts lower than 1 count per million (CPM) and those in sex chromosomes were filtered out (see **Section 3.2.2.4**), removing 2% of all peaks and leaving 274,687 for downstream analysis. Counts were normalized and transformed to log-CPM for linear correlation of the consensus peak set (**Figure 3.12**). Samples within the same stage show low Pearson's r (below 0.71), possibly due to the few shared peaks between donors and among stages, and high background noise of unspecific binding.

Principal component analysis of normalized counts (**Figure 3.13 a**) showed how, although most replicates per stage grouped together and follow a temporal progression from iPSC to the later stages of differentiation, the variance in the first principal component is driven by an outlier (EP-SBAd2.1). Removing this outlier (**Figure 3.13 b**) positioned the DE donor replicates further away in the variance space instead. BLC replicates did not cluster together nor were placed in the principal component space as they should.

Due to the behaviour of the normalized counts highlighted by these two figures, I decided against performing any quantitative analysis of H3K27ac peaks and investigated them only based on their presence or absence per stage.

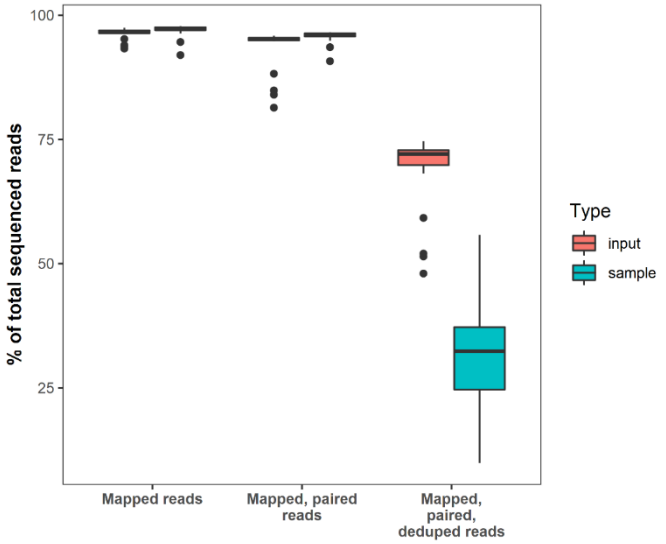


Figure 3.8. H3K27ac ChIP-seq quality control of reads. Boxplots of the percentage of reads remaining at every QC filtering step: after mapping (left), pairing (centre) and removing duplicates (right).

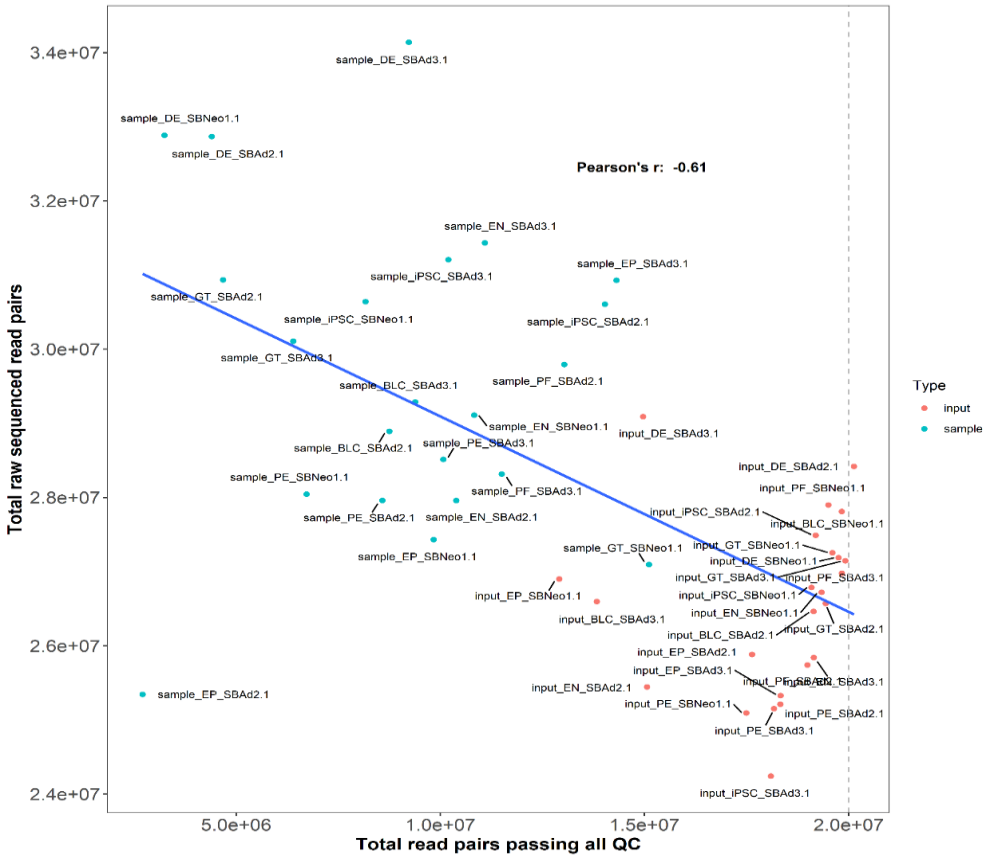


Figure 3.9. H3K27ac ChIP-seq quality control of reads. Scatterplot of raw read pairs sequenced (y axis) vs total read pairs passing all QC (x axis) shows high inverse Pearson´s correlation. The dotted vertical line denotes the minimum number of read pairs recommended by ENCODE (20 million). Blue dots: samples. Red dots: input controls.

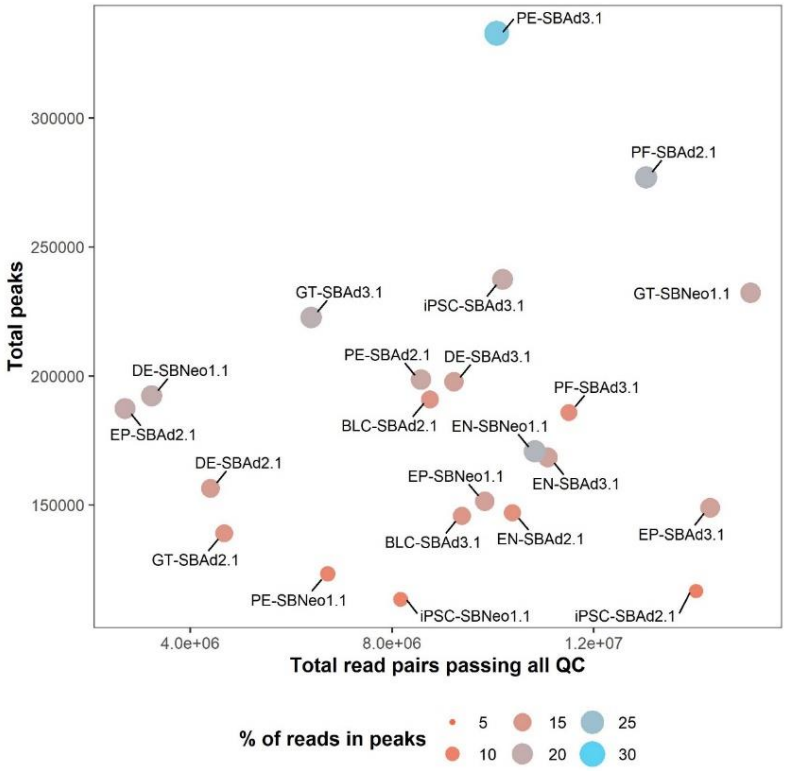


Figure 3.10. Quality control of peak calling. Relationship between read (x-axis) and peak number (y-axis). Area: percentage of reads in peaks, with lower values in red.

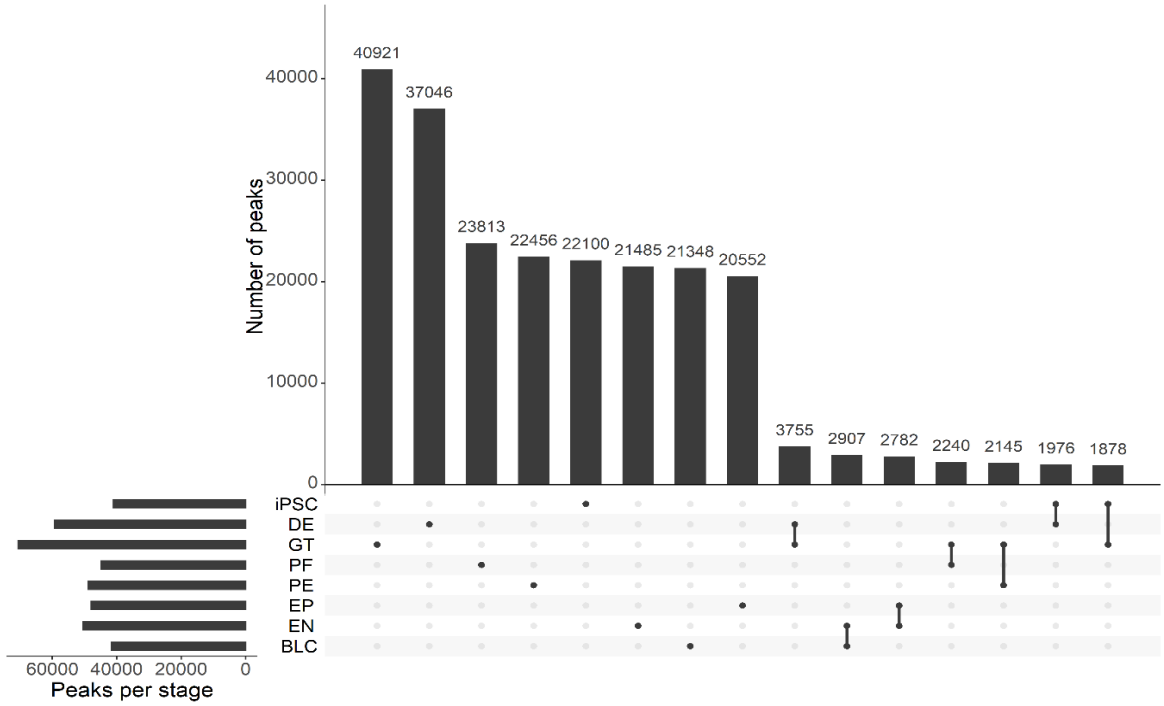


Figure 3.11. Consensus set of H3K27ac ChIP-seq peaks. Dotted lines indicate the number of peaks belonging to each stage combination, with only the 15 most abundant shown. The number of peaks per stage is shown in the left barplot.

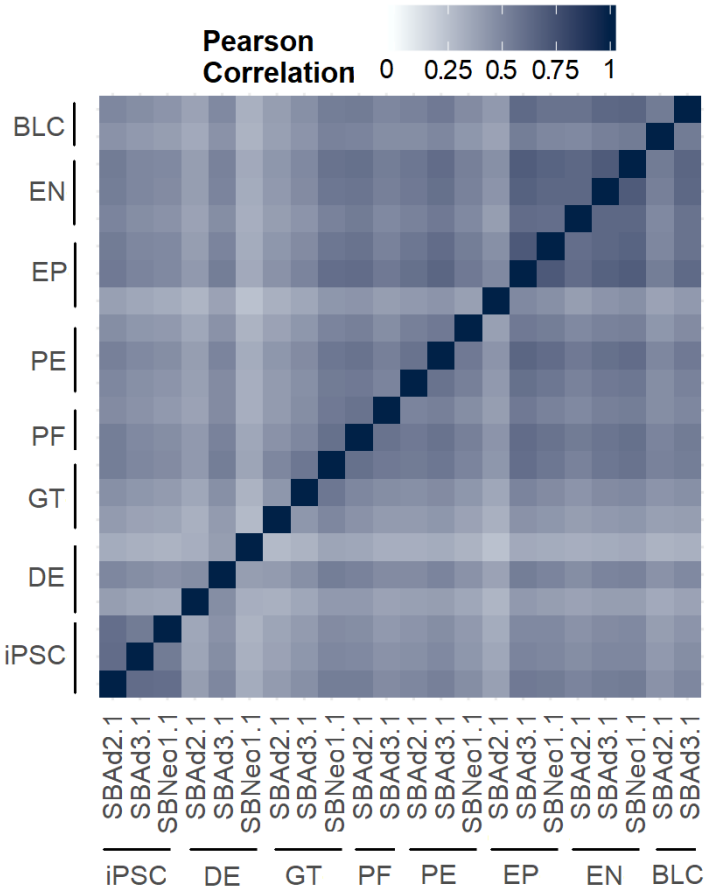
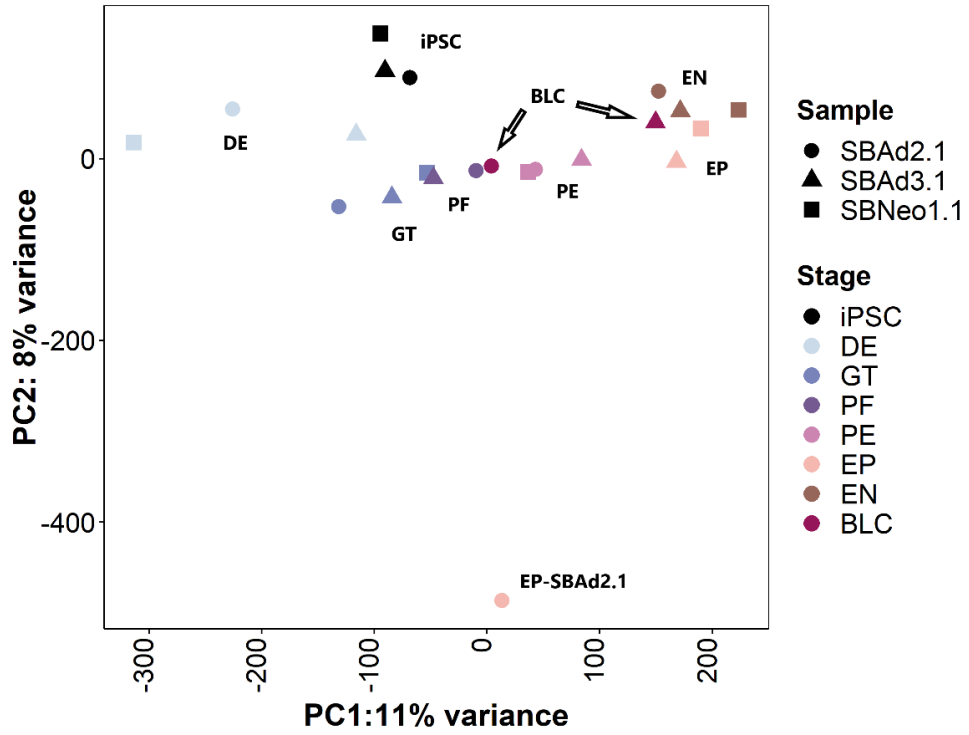


Figure 3.12. Pairwise H3K27ac ChIP-seq peak count correlation. Donor replicates show low (< 0.71) linear correlation (measured by Pearson's r) within stages. Replicates are ordered within each stage (down to up and left to right) as SBad2.1, SBad3.1 and SBneo1.1. Peak read counts have been normalized and transformed to log-CPMs.

a



b

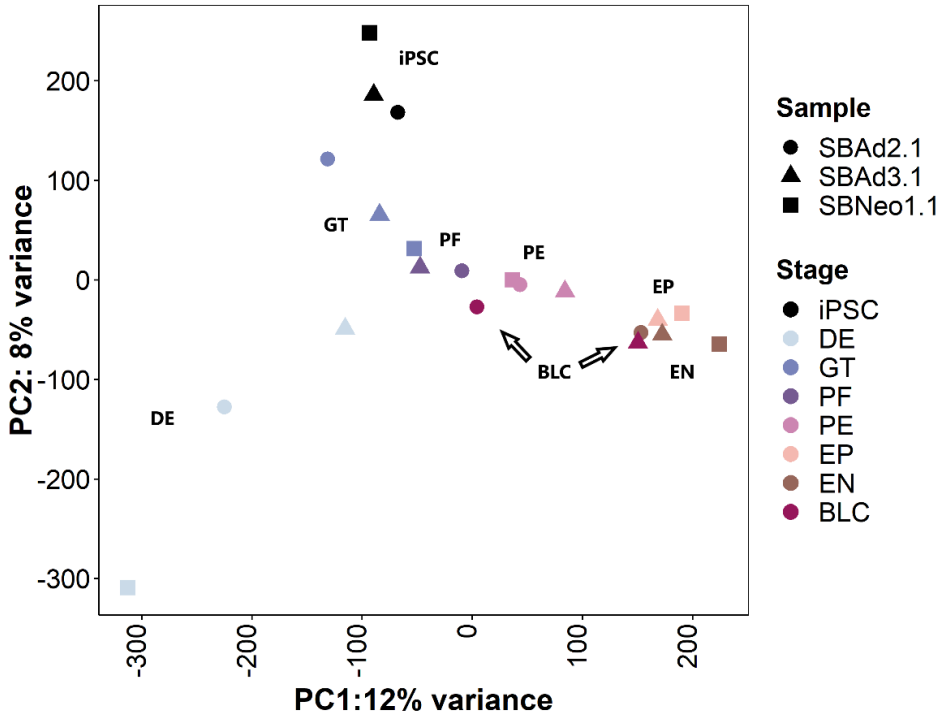


Figure 3.13. Principal component analysis of H3K27ac ChIP-seq data. Shown are the first two principal components of the normalized counts per peak, with all sequenced samples (**a**), and after removing the EP-SBAd2.1 outlier (**b**).

3.3.1.3. Quality control of the EPIC methylation array

11 adult islets from two different sources and 21 iPSC-derived samples differentiated towards beta-like cells (see the start of **Methods** and **Section 3.2.3.1** and **3.2.3.2**) were processed with the Illumina 850K (EPIC) methylation array, which covers 865,859 cytosines genome-wide. Differentiation samples had three replicates per stage sequenced except for EP (EP from donor SBAd2.1 present only) and BLC (SBAd3.1 and SBNeo1.1 donor replicates present) due to limitations in DNA availability.

To evaluate the quality of the methylation array I examined the intensities of the control probes in the green and red channels. All types of probes presented the right intensity in the appropriate channel, except for hybridization controls, in which probe labels for high and low intensity were exchanged (this was a known issue for this batch of Illumina EPIC arrays). However, these did not compromise the performance of the methylation array.

All samples passed the quality control threshold of <1% of failed probes, which were identified as those with higher detection p-value than 0.01. The filtering process followed by quantile normalization (see **Section 3.2.3.3.**) removed 65,243 probes from the analysis (approx. 7% of the total), and reduced methylation intensity differences between replicates (see **Figure 3.1**).

Samples presented similar methylation levels overall (**Figure 3.14**), with slightly less high methylated CpGs (beta >0.75) in adult islets (46% vs 50-51% in iPSC and differentiating stages) and an increased number of medium-low methylated CpGs (0.25 > beta > 0.50; 9% vs 5-7% in iPSC and differentiating stages). These similarities are also reflected in the high pairwise correlation of methylation intensities at CpGs among all stages and samples (Pearson's r > 0.87, **Figure 3.15;** see **Section 3.2.3.3**). As expected, differentiating beta cells and adult islets (ISL and R) are clearly distinct, which can also be observed in the PCA plot (**Figure 3.16 a**). Nevertheless, the alignment over the first principal component indicated that the methylome of the most differentiated stage (BLC) was more akin to that of an adult islet than the other stages. Adult islets isolated in two different locations (Oxford and Alberta, see **Section 3.2.3.1**) were not separated in the first two principal components.

Differentiating samples cluster by stage (**Figure 3.16 b**), and most of those were distinct, except for the two most differentiated (EN, BLC) which (as was shown by the ATAC-seq PCA) had a similar methylation profile.

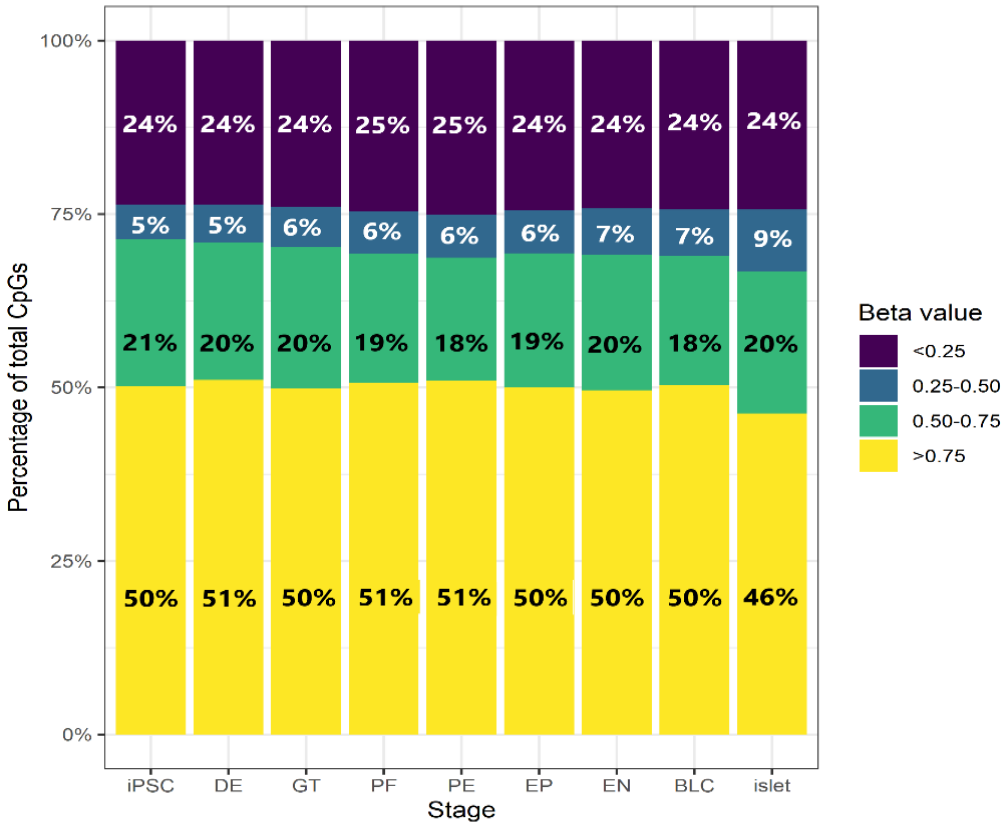


Figure 3.14. Distribution of methylation levels per stage. Methylation levels are represented by beta values, which have been binned in four different categories from low methylation (beta < 0.25) to high methylation (beta > 0.75) per stage (x-axis). y-axis represents the percentage of each category over the total of CpG sites.

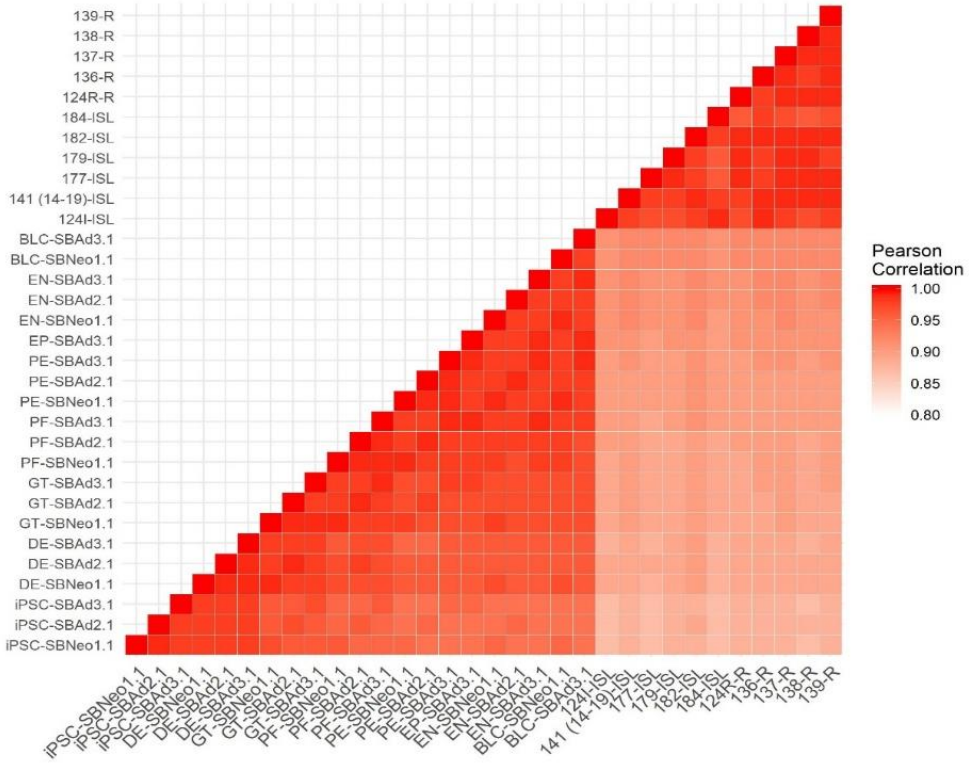


Figure 3.15. Correlation of methylation intensities across CpGs. All samples showed a high pairwise correlation of methylation intensities (measured by Pearson's r of M values) at each probe genome-wide (r > 0.87). There was a clear distinction between the differentiating beta cells (iPSC to BLC, down to up and left to right) and the adult islets (ISL and R).

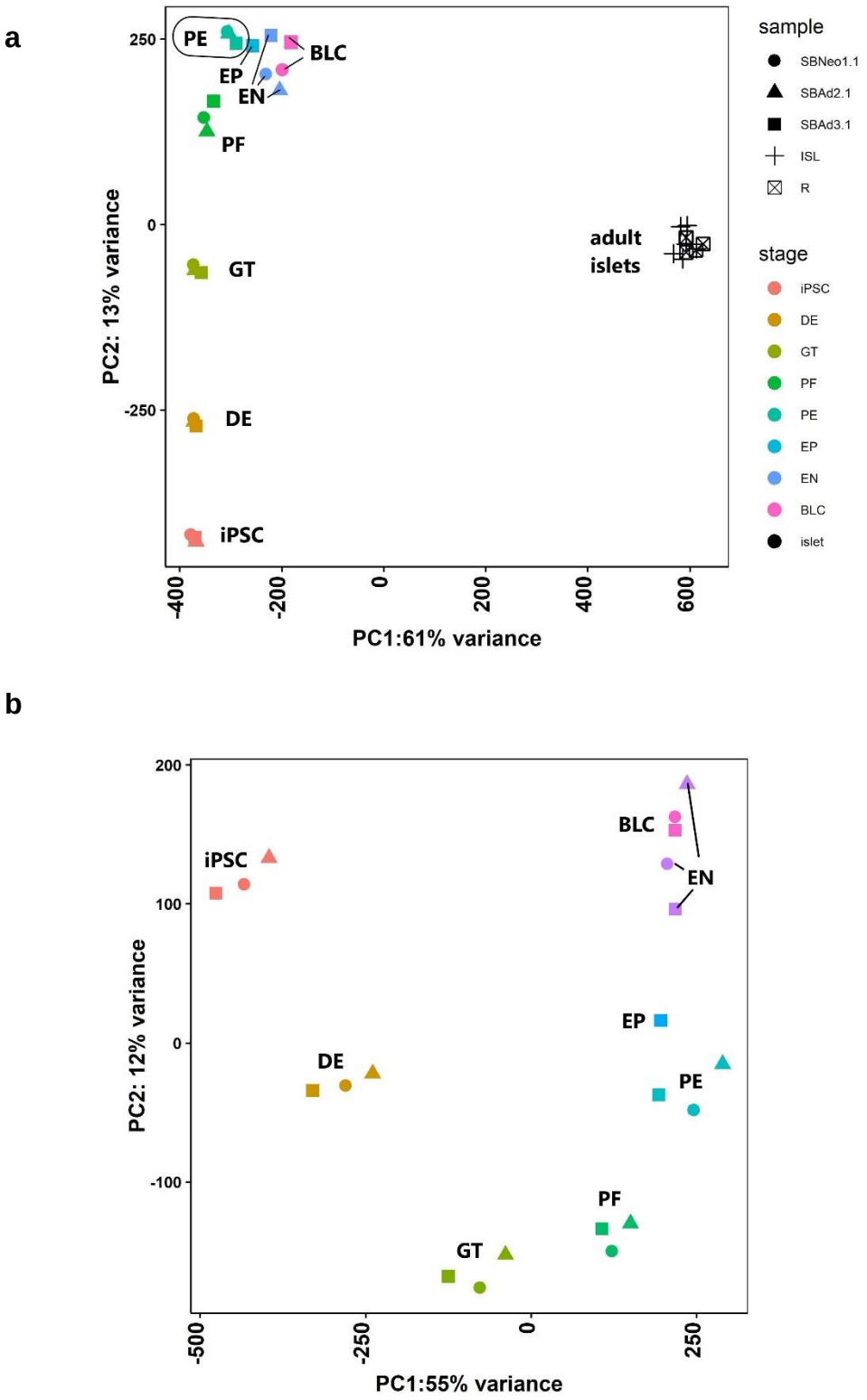


Figure 3.16. Principal component analysis of the methylome. The two first principal components are shown for all samples (**a**) and differentiating beta cells only (**b**).

3.3.2. Distribution of chromatin regulatory marks across the genome

I investigated the genomic location of the ATAC-seq and H3K27ac peaks, and the methylation levels at each probe from the array, and how those varied at every stage. Accessible regions were located mostly in distal intergenic regions (mean across stages = 34%, **Figure 3.17 left**), followed by promoters (31%) and introns (27%). Most of the peaks (20%) were located between 10 and 100 kb away from the TSS, while 16% were within 1kb and 11% were further away from 100kb. 54% of the tallest peaks (>10 CPM, 37,190 peaks) overlapped the promoters. Therefore, the most accessible chromatin overlaps the promoters, particularly within 1kb from the TSS, while presumed regulatory regions often present smaller peaks. As expected, the significance of the overlap of accessible chromatin regions (peaks over 1 CPM) with promoters is very significant and higher than expected by chance (p-value < 0.01, see **Section 3.2.4**).

Conversely, H3K27ac peaks mapped mostly to introns (46%) and distal intergenic regions (42%), while just 7% of peaks on average were at promoters (**Figure 3.17 right**), although the overlap with these regions was still significant (p-value < 0.01). This meant that most of the peaks (53%) were 10 – 100 kb from the TSS, while 24% were further away than 100kb and only 6% were within 1 kb. As was shown in **Figure 3.11**, most of these peaks are found exclusively at one stage (they are “stage-specific”), suggesting that these regulatory marks act preferentially in this dataset as cell type-specific enhancers at intergenic and intronic regions. However, this could be an artefact of the lower quality of my H3K27ac data.

ATAC-seq and H3K27ac ChIP-seq peaks from publicly available data (**Figure 3.18, Section 3.2.5**) indicated that they were more frequently found in intergenic regions and only a low percentage (5% for ATAC and 7% for ChIP-seq) in promoters, which was more similar to the H3K27ac results from the differentiating islets. However, H3K27ac and ATAC-seq peaks described in this chapter were a ten-fold smaller set of high quality regions, retained after IDR comparison of replicates (a process that the public data shown here did not go through). Therefore, the overlap with genomic features of the public data and the data from differentiating beta cells were not directly comparable.

Looking at methylation data, most CpGs in the EPIC methylation array were located at introns (35%), followed by promoters (31%) and intergenic regions (24%). In relation to CpG-rich regions known as CpG islands (Jones, 2012), these were the 2nd most abundant location (19%), while most CpGs fell in "open sea" regions (56%), > 4 kb away from islands. These features presented differences in the distribution of methylation levels: promoters were largely hypomethylated while other genomic features were mostly hypermethylated (**Figure 3.19 a**). CpG island annotations showed these to be hypomethylated, while regions > 2kb away from islands were hypermethylated and shores (< 2kb from islands; Irizarry *et al.*, 2009) presented a wider distribution of methylation values (**Figure 3.19 b**).

It has been observed that, while methylation remains largely unchanged through the adult life, in early development genome-wide demethylation and reprogramming occurs (Jaenisch *et al.* 1982). To assess the changes in the methylome in the differentiating cells, I called differentially methylated probes (DMPs) using the M values calculated previously, comparing each stage of differentiation to the iPSC baseline (see **Section 3.2.3.6)**. I also compared the methylation status of adult islets to the more immature BLCs, using the BLC as baseline for the methylation changes. Contiguous DMPs with similar and significant fold-changes in methylation values (in the same direction of gain or reduction of methylation intensity) were joined together in differentially methylated regions (DMRs, see **Section 3.2.3.6**).

89,790 hypomethylated and 51,122 hypermethylated DMRs were called, overlapping mostly and significantly (p-value < 0.01) with promoters (38% for all DMRs, ranging from 35-42% for hyper-and hypomethylated DMRs respectively; **Figure 3.20 a and b**). These DMRs had a median size of 613 bp (min = 3 bp, max = 18 kb), and encompassed 355,833 hypomethylated and 182,788 hypermethylated DMPs across all stages. For the differentiating stages, the minimum amount of DMRs was detected for the DE stage, with numbers increasing up to the EN stage (**Figure 3.20 a**). Adult islets were more dissimilar to the most differentiated BLC stage than the latter were compared to iPSCs (38,059 vs 19,279 total DMRs), likely due to the very different origins of the differentiating and adult islet samples.

Despite the significance of methylation changes at probes and regions, it was difficult to rationalize what these would mean in terms of changes in DNA regulatory processes, as often the changes in methylation were very small (median absolute fold change in average beta value of the region= 0.06) and they remained within their initial high (beta value > 0.5) or low (beta value < 0.5) average methylation values. I therefore clustered contiguous CpGs by their mean methylation profiles per stage and in adult islets (see **Section 3.2.3.5**), dividing the resulting segments into high methylated regions (HMR, average beta > 0.5) or low methylated regions (LMR, average beta < 0.5). There were 27,903 LMR and 22,883 HMR in total and, as expected, these regions were often larger than DRMs, but they presented a very different size distribution depending on the methylation status: LMRs had a median size of 3 kb (min = 2 bp, max = 7 Mb), while HMRs had a median size of 500 kb (min = 2 bp, max = 130 Mb). This indicated that LMRs were more focused at specific genomic locations, usually overlapping promoters (**Figure 3.20 b**) while larger stretches of the genome remained overmethylated and inactive.

In conclusion, most of the H3K27ac peaks of differentiating islets can be found at distal regulatory regions, while DMRs and LMRs were more often found and promoters, and accessible chromatin split between putative enhancers and promoters. Finally, comparing all datasets together, the ATAC-seq peaks, H3K27ac peaks and low methylated regions overlap significantly more often than expected by chance (p-value < 0.01 at every stage of differentiation, see Section 3.2.4), indicating the interplay between these signals.

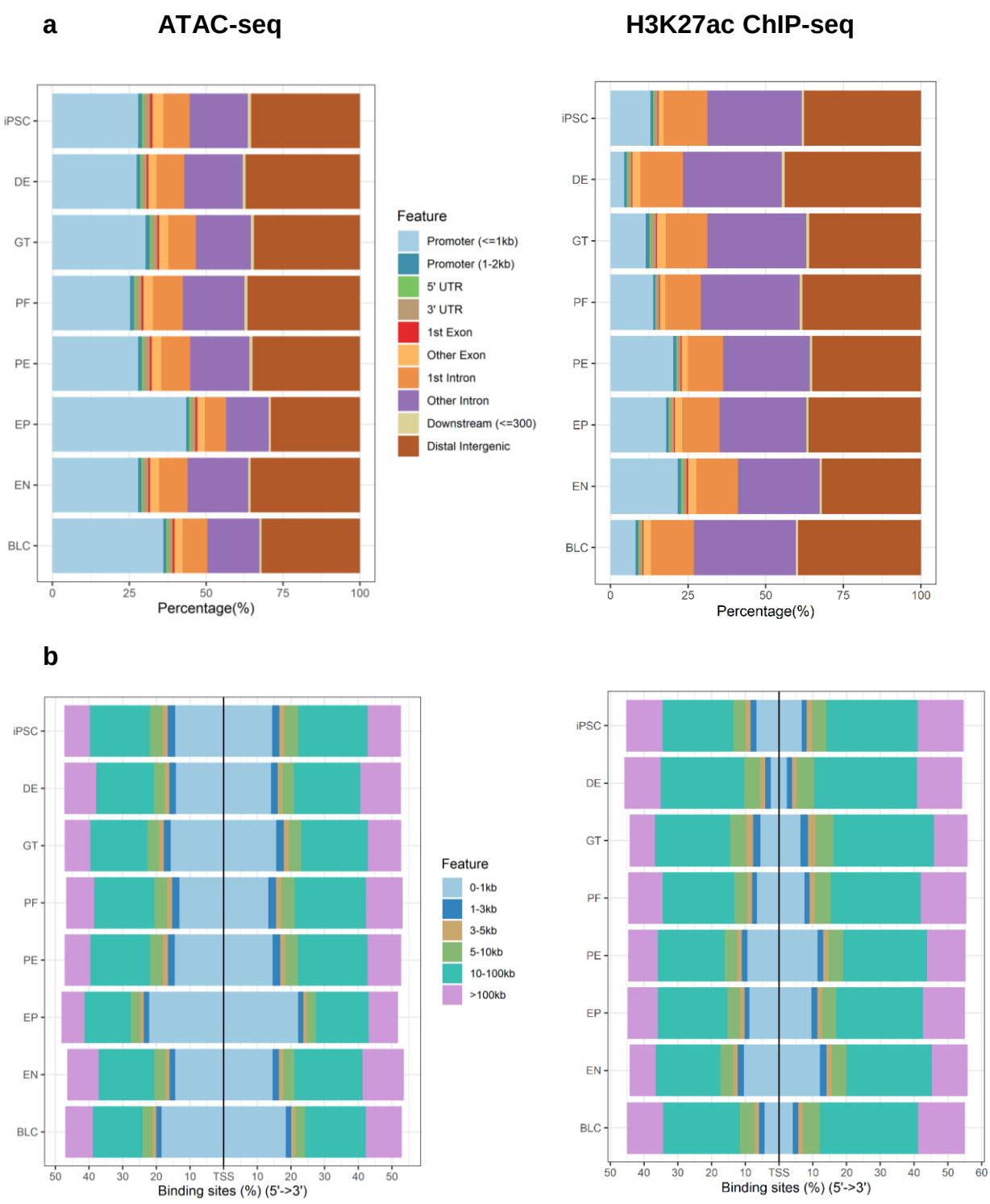


Figure 3.17. Genomic annotation of ATAC-seq and H3K27ac peaks from the beta cell differentiation model. Distribution of genomic features (a) or distance to TSS (b) per stage at open chromatin peaks (left) and H3K27ac peaks (right).

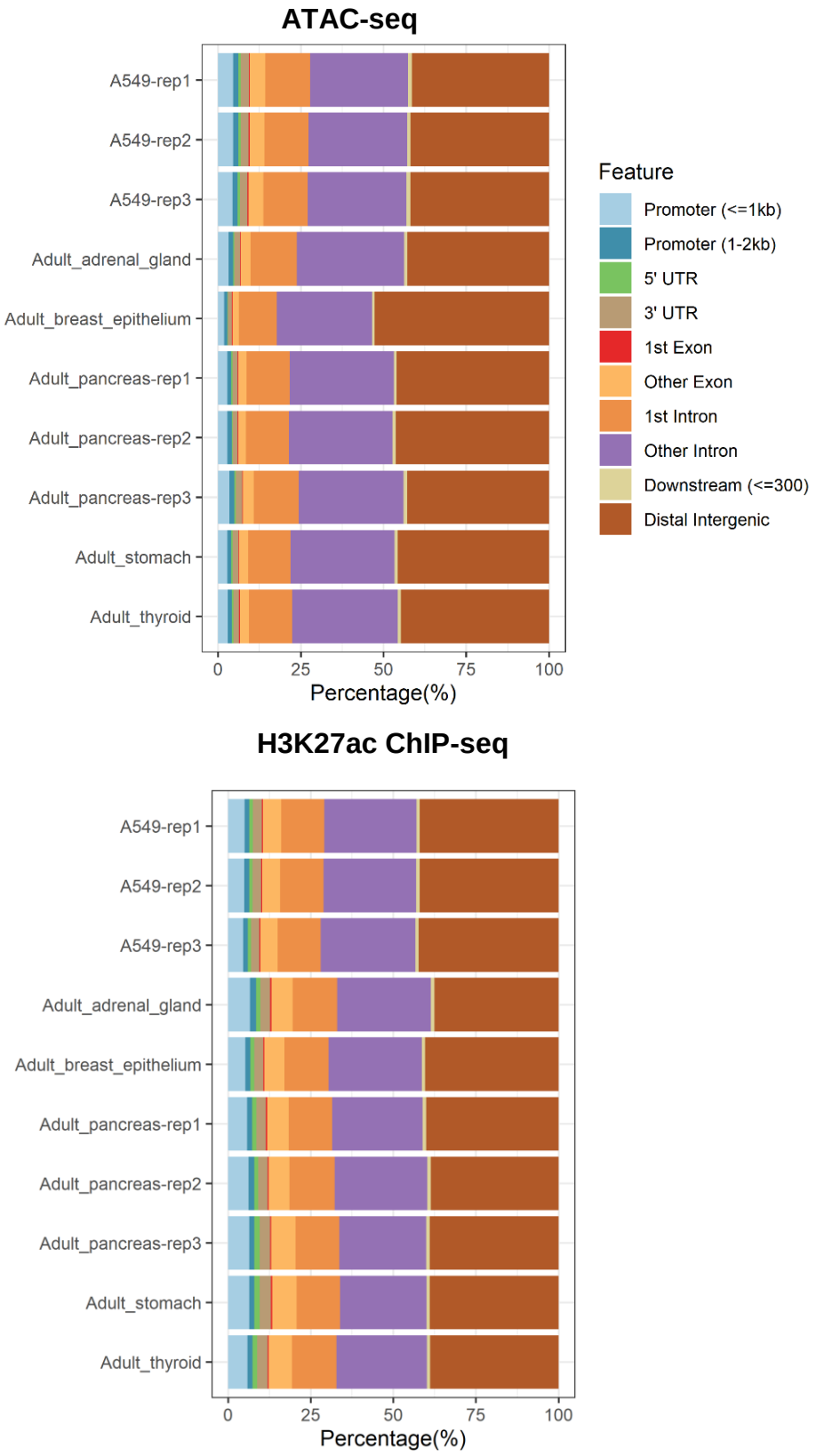
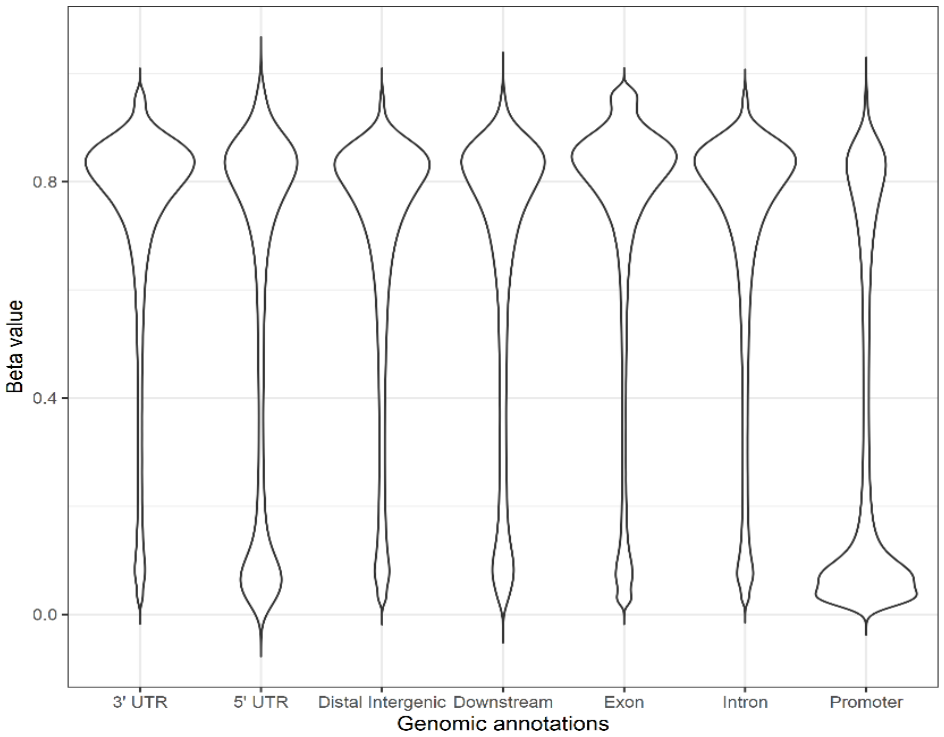


Figure 3.18. Genomic annotation of ATAC-seq and H3K27ac peaks from publicly available data. Distribution of genomic features per tissue or cell type at open chromatin peaks (above) and H3K27ac peaks (below).

a



b

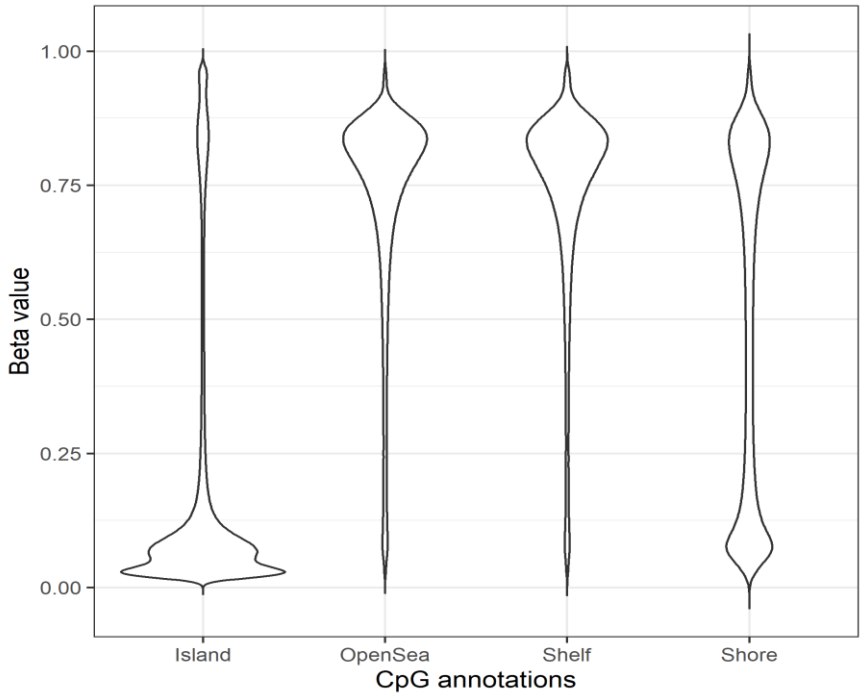
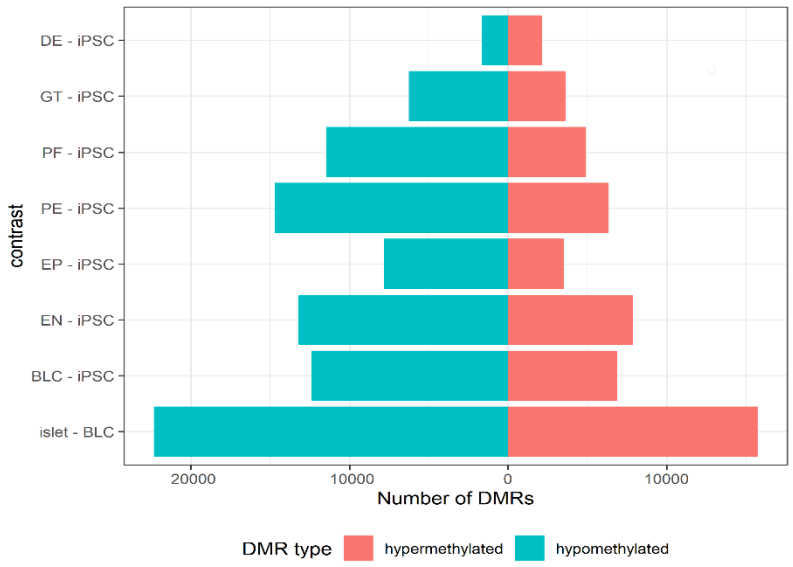
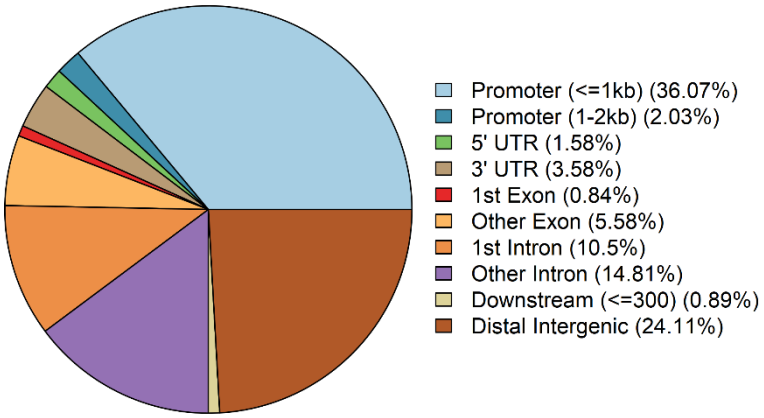


Figure 3.19. Methylation intensities per annotation for all samples. Methylation intensity (represented by beta values, y-axis) distributions of genomic features (x-axis) in relation to TSS (**a**) or to CpG islands (**b**). Promoters were more frequently hypomethylated (low beta), while the other genomic features were often hypermethylated (high beta). CpG islands were mostly hypomethylated, while features further away had higher methylation intensities. Shore: < 2kb from islands. Shelf: 2 – 4kb from islands. Open sea: > 4 kb away from islands.

a



b



c

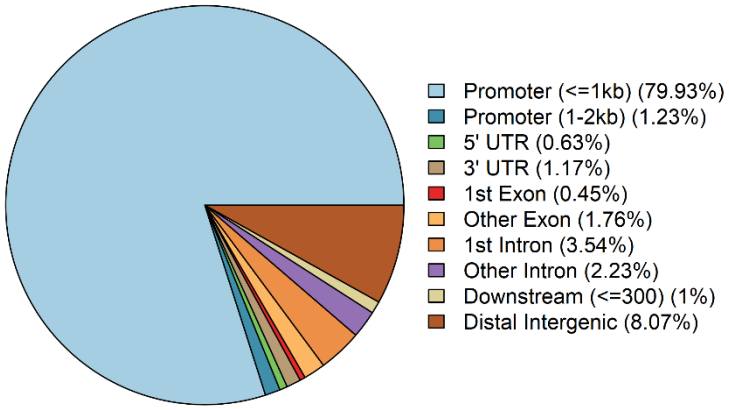


Figure 3.20. Overview of differentially methylated and low methylated regions (DMRs and LMRs). a) Number of hypermethylated and hypomethylated DMRs (x-axis) per contrast (y-axis): each differentiation stage vs iPSC and adult islets vs BLC. b) Genomic annotations for both hypermethylated and hypomethylated DMRs for all contrasts. c) Genomic annotations for LMRs in all differentiation stages and adult islets.

3.3.3. Regulatory marks are enriched in pathways involved in islet development and function

The differences and similarities of the biological pathways and transcription factor networks recapitulated by each DNA regulatory mark per stage were assessed next. To investigate the pathways, I assumed that each mark acted on the closest gene (see **Section 3.2.6)**, due to the absence of chromatin interaction data or other information with which to link the marks to their effectors.

In the case of ATAC-seq peaks, because a sizeable proportion of them are shared among all differentiation stages (**Figure 3.5**), I removed those to capture more specific signatures. The enrichment of their closest genes in Reactome pathways was calculated, compared to the background of all genes, with a hypergeometric test (see **Section 3.2.6)**. This revealed an enrichment in many terms (**Figure 3.21**), including “Regulation of beta cell development” from the GT stage to the beta-like cell stage (BH-adjusted p-value < 0.02), driven by peaks near 13 known monogenic diabetes genes and other important genes in islet development (e.g. *NEUROD1*, *PTF1A*, *NEUROG3, NKX6-1*, *HNF1A*, *PDX1*, etc.). Nevertheless, most of the enriched terms were shared across all stages, even those related to the biology of more adult differentiation stages (e.g. “Regulation of insulin secretion”).

Therefore, to avoid the loss of information produced by selecting only a certain set of ATAC-seq peaks while preserving the differences in chromatin accessibility between the stages, I clustered the peaks based on their accessibility profiles using WGCNA (see **Section 3.2.1.5**). This resulted in 15 modules of variable size (**Figure 3.22 a** and **b**), including the “grey” module that is where the peaks are assigned when they do not cluster with the average profile of any other module. The eigenvector (or “eigengene” in WGCNA terms) from the first principal component of all peaks in each module constitutes the representative profile of the peaks in that stage. Pathway enrichment analysis of genes mapping closest to the peaks within these modules (**Figure 3.23**) revealed better enrichment for insulin pathway-related terms, including “Potassium channels”, in the blue and black module which peak their accessibility in the latest stages of development. Enrichment of the term

“Regulation of beta cell development” was also maintained in the yellow module, with peak accessibility in the middle and late differentiation stages.

Genes mapping close to stage-specific H3K27ac peaks (that constitute the majority of the ChIP-seq peaks, see **Figure 3.11**) in the same way also revealed an enrichment (**Figure 3.24**), in “Regulation of insulin secretion”, in this case only in the EN stage (BH-adjusted p-value = 0.007). Only the iPSC stage was enriched for the term “Transcriptional regulation of pluripotent stem cells” (BH-adjusted p-value = 2×10^{-4}), indicating that the differentiated nature of the subsequent stages is revealed by their histone acetylation profile. At the same time, all stages shared enrichment for the “Signalling by receptor tyrosine kinases” pathway (BH-adjusted p-value < 2×10^{-4}), growth factor and cytokine receptors essential in embryonic development that would be activated during the differentiation process (Hubbard and Miller, 2007), which was also observed in all stages and modules of ATAC-seq pathway enrichment (**Figures 3.21 and 3.23**).

Concerning the pathway enrichment of methylation signals, annotation of DMRs or DMPs to nearby genes did not reveal significant (adjusted p-value < 0.05) enrichment of terms at any stage compared to the background of the genes in the neighbourhood of all CpG positions in the array. Examination of low methylated regions (LMR) did reveal an enrichment (**Figure 3.25**) in the development pathways mentioned above (“Regulation of beta cell development” from iPSC to PE and “Signalling by receptor tyrosine kinases” in the same stages: BH-adjusted p-value for both pathways < 0.04), despite there being less LMRs than DMRs (see previous section). This could indicate that these regions capture regulatory processes of differentiating islets globally in a more comprehensive and interpretable way than changes in methylation. High methylated regions were not tested due to their large size, meaning their annotation to the closest gene would likely be ambiguous.

In summary, the genes more likely to be regulated by these DNA marks were consistently involved in beta-cell related biological pathways, which are captured better by the chromatin accessibility information at each stage of pancreas development.

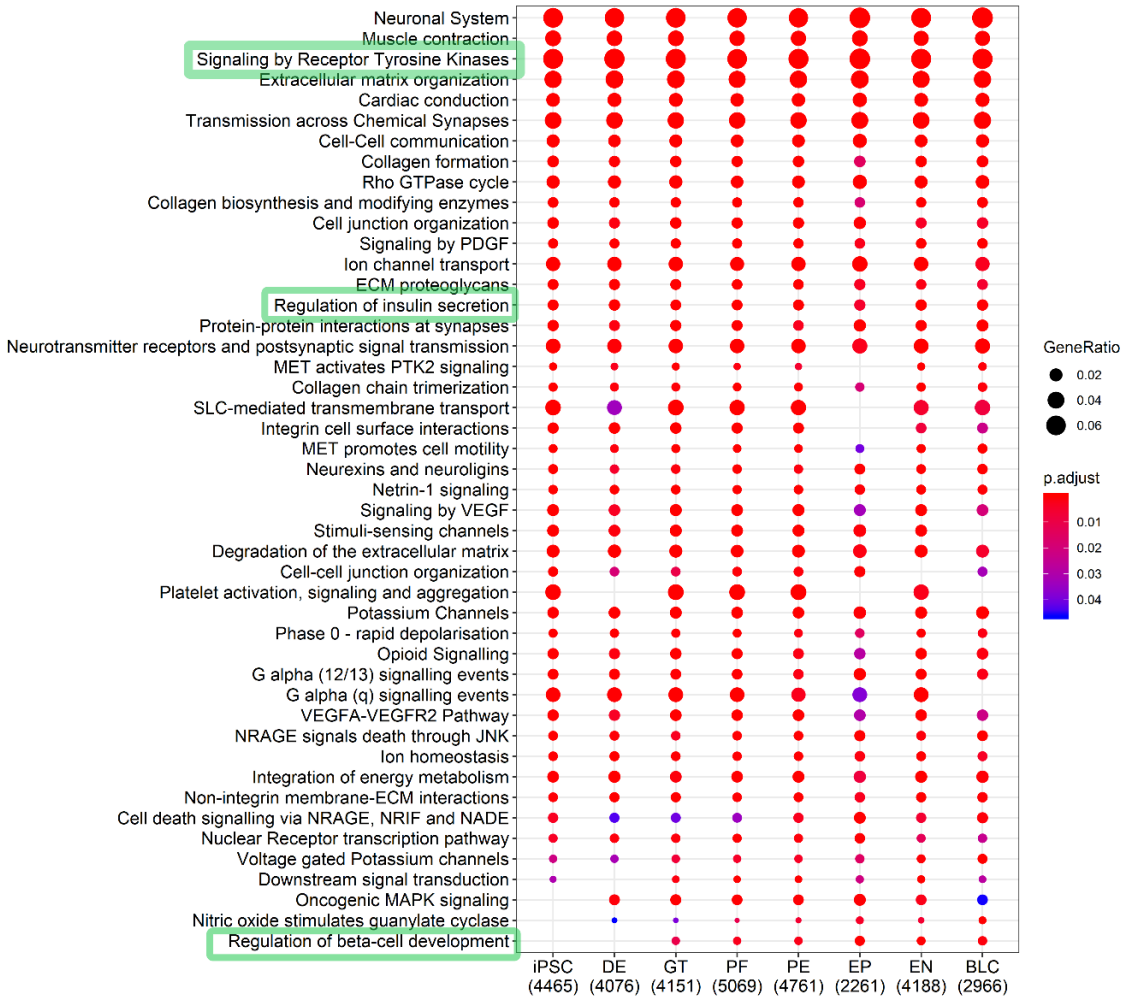
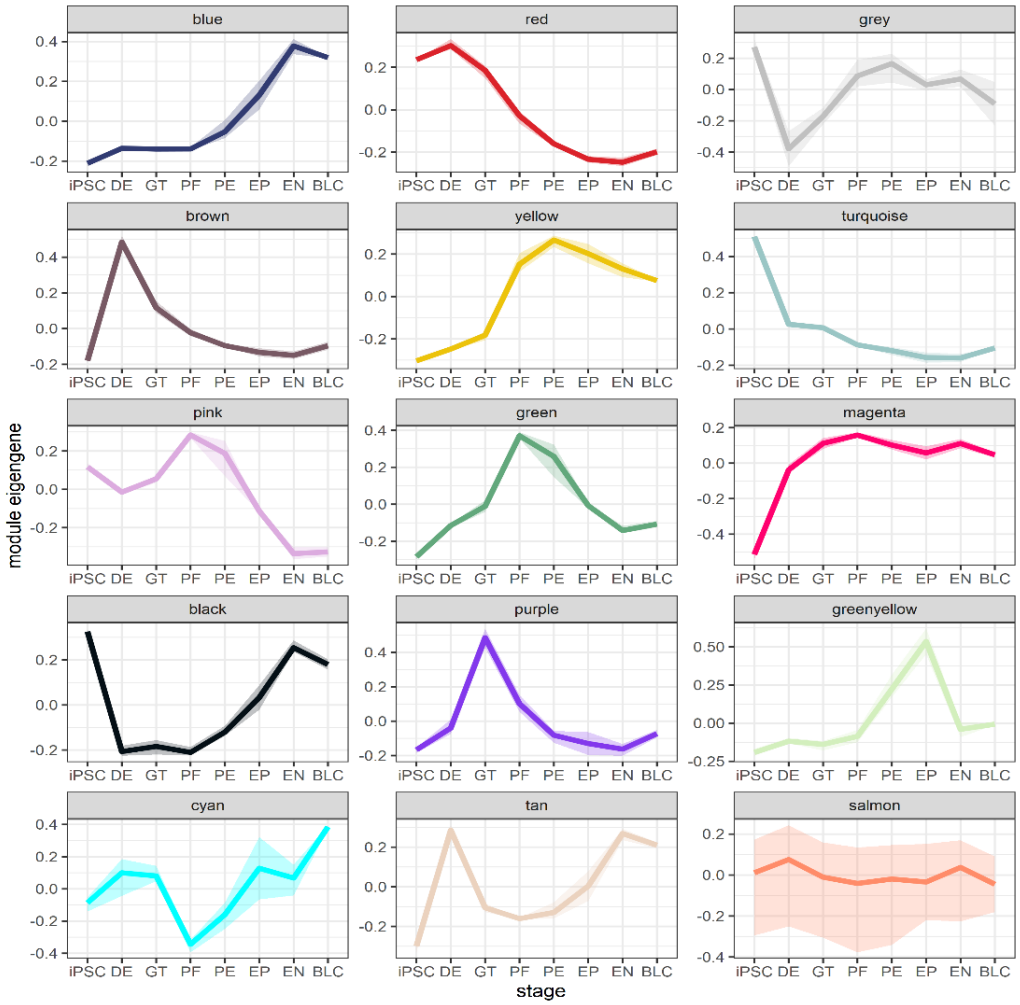


Figure 3.21. Pathway enrichment analysis of ATAC-seq peaks. 1 CPM-filtered consensus set of peaks have been mapped to the closest gene, after removing peaks shared among all stages, and enrichment of terms from the Reactome pathway (y-axis) was calculated per stage (x-axis, total genes in parentheses). In green: pathways mentioned in the main text. “GeneRatio”: genes found per term over total genes in all terms per stage. “p.adjust”: Benjamini-Hochberg adjusted p-value

a



b

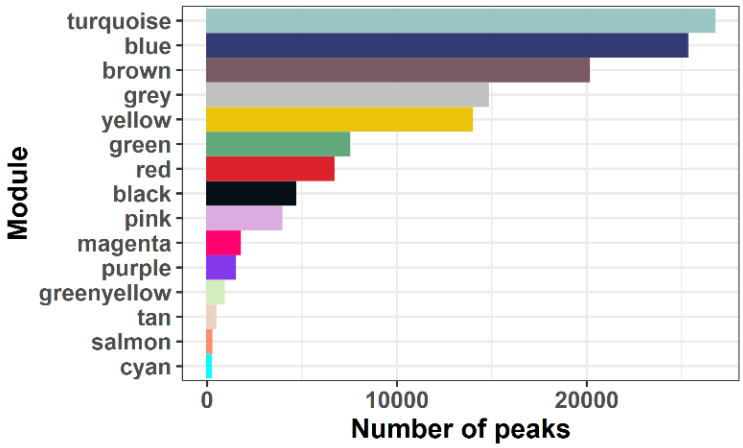


Figure 3.22. Clustering of 1 CPM-filtered consensus set of ATAC-seq peaks. ATAC-seq peaks are clustered into 15 modules. Peaks that cannot be assigned to those modules are classified as "grey". **a**) Profiles of accessibility across differentiation stages. x-axis: differentiation stages. y-axis: mean value of the eigenvector for the donor replicates (called "eigengene"). Shading: delimits the minimum and maximum value per stage. **b**) Number of peaks (x-axis) per module (y-axis).

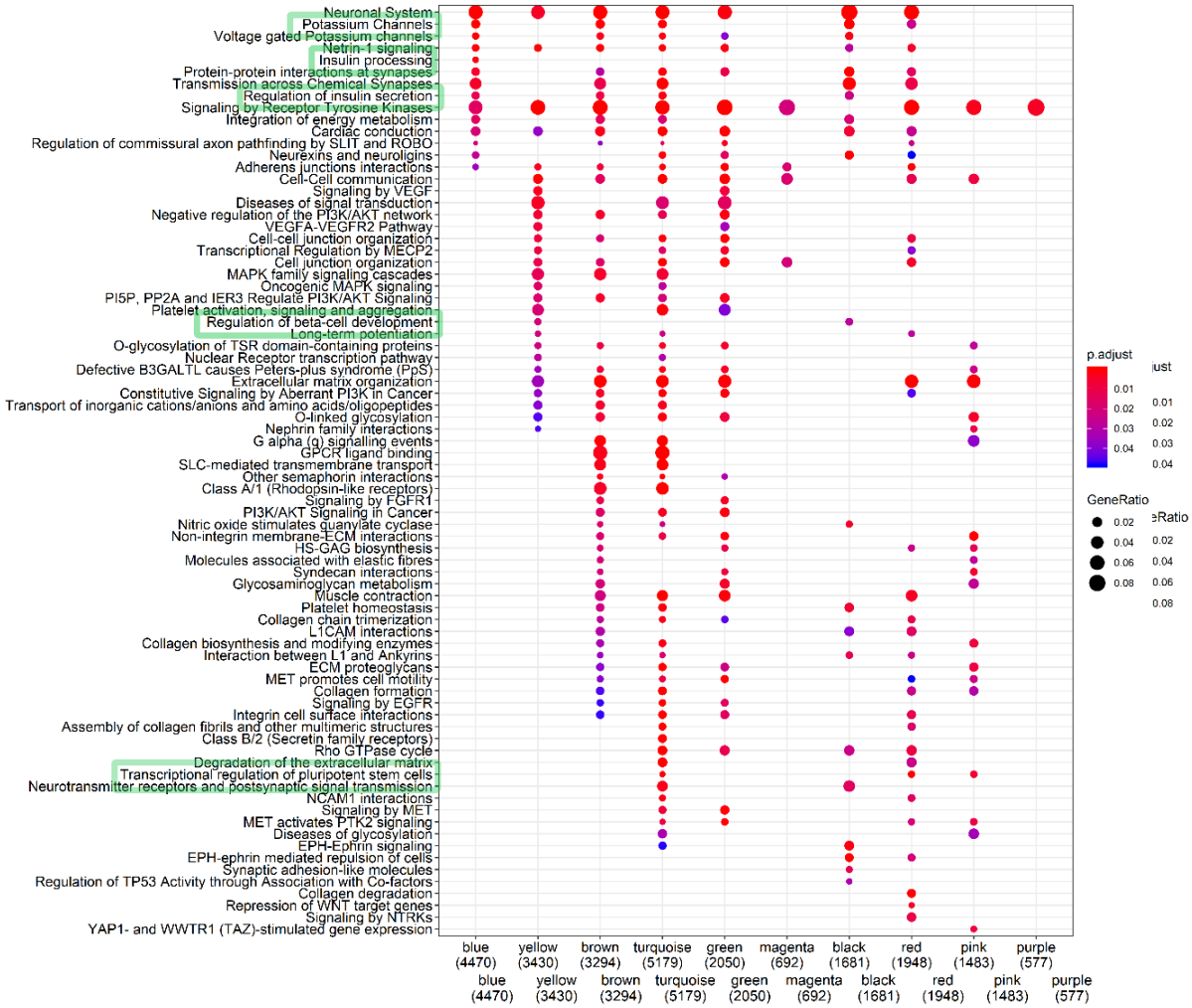


Figure 3.23. Pathway enrichment analysis of ATAC-seq WGCNA modules. Peaks within modules have been mapped to the closest gene, and enrichment of terms from the Reactome pathway (y-axis) was calculated per module (x-axis, total genes in parentheses). In green: pathways mentioned in the main text. “GeneRatio”: genes found per term over total genes in all terms per stage. “p.adjust”: Benjamini-Hochberg adjusted p-value

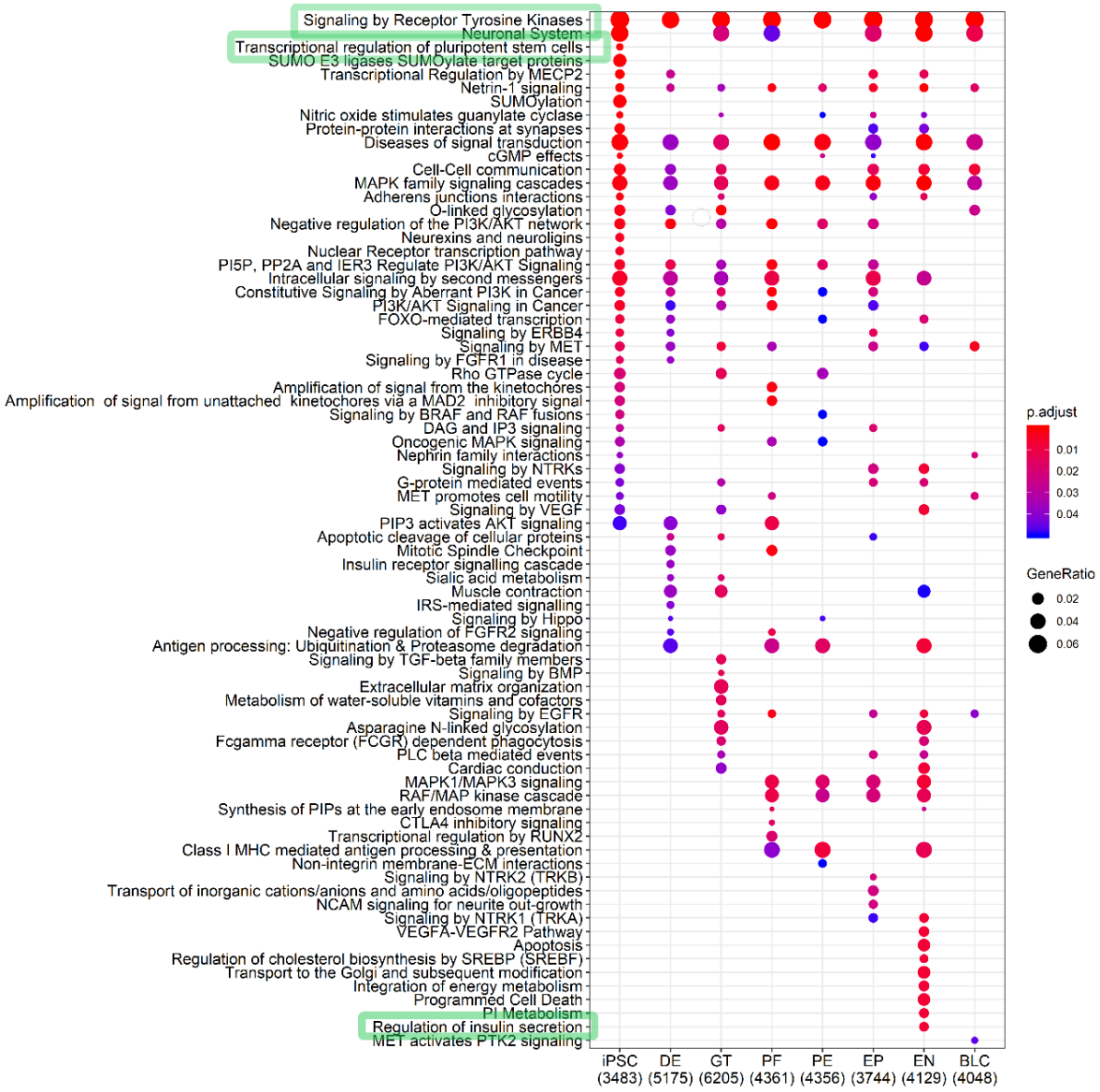


Figure 3.24. Pathway enrichment analysis of H3K27ac ChIP-seq peaks. 1 CPM-filtered consensus set of peaks have been mapped to the closest gene, after retaining peaks that are present only in one stage ("stage-specific"), and enrichment of terms from the Reactome pathway (y-axis) was calculated per stage (x-axis, total genes in parentheses). In green: pathways mentioned in the main text. "GeneRatio": genes found per term over total genes in all terms per stage. "p.adjust": Benjamini-Hochberg adjusted p-value

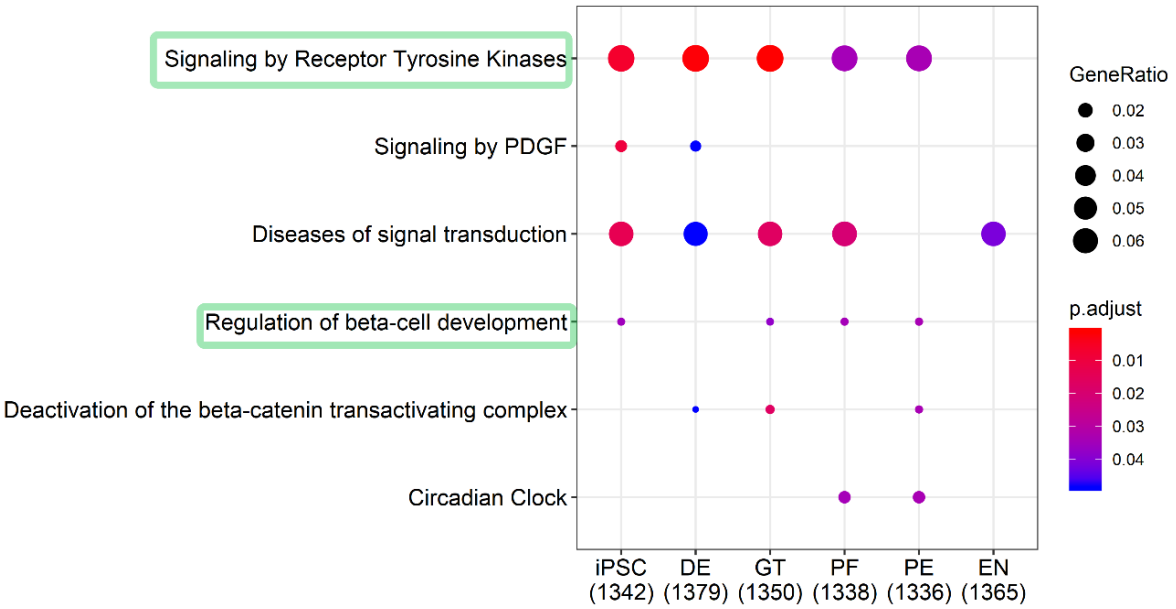


Figure 3.25. Pathway enrichment analysis of low methylated regions. Low methylated regions (LMRs) was mapped to the closest gene, and enrichment of terms from the Reactome pathway (y-axis) was calculated per stage (x-axis, total genes in parentheses). In green: pathways mentioned in the main text. “GeneRatio”: genes found per term over total genes in all terms per stage. “p.adjust”: Benjamini-Hochberg adjusted p-value

3.3.4. Transcription factors regulating islet development

After looking at the gene networks likely being regulated by these chromatin marks, I investigated the transcription factors (TFs) binding to these regions, calculating their enrichment in TF binding motifs using HOMER (see **Section 3.2.8**). The analysis of all ATAC-seq peaks (minus those shared among all stages, as explained above), ATAC-seq peak modules, stage-specific H3K27ac peaks and hypomethylated differentially methylated regions (DMRs) highlighted three TFs common to these datasets (see **Figures 3.26 to 3.29**): the essential regulator of the development of pancreatic progenitors, *PDX1*; and monogenic diabetes genes *MNX1* (marker of the dorsal endoderm) and *PAX4* (marker of endocrine progenitors). All three TFs appeared enriched as expected from intermediate to late stages of development (DE to BLC stage, and corresponding WGCNA modules in which the peaks are more open at those stages). Hypermethylated DMRs were not examined, as those are consistently associated with reduced gene expression (Suzuki and Bird, 2008; Jones, 2012; Schübeler, 2015; Korthauer and Irizarry, 2018). Low and high methylated regions (LMRs and HMRs respectively) were not examined either because the average region was much longer than recommended for motif finding (> 500 bp), covering many TF binding motifs and difficulting the interpretation of the results.

Globally, the ATAC-seq data and hypomethylated DMRs captured more known TFs involved in pancreas development, including monogenic diabetes genes, than the H3K27ac (see **Table 3.1**). TFs encoded by genes located in T2D loci also were highlighted by this analysis, including *EBF1*, a pioneer TF involved in demethylation (Boller *et al.*, 2016), and *ETS1*, which maintains the pancreatic progenitor state in mice (Kobberup *et al.*, 2007). *ONECUT1* (*HNF6)* emerges as a possible regulator of middle and late stages of pancreas development (including ATAC-seq modules green, yellow and blue, and hypomethylated DMRs at the EP stage), as observed in murine studies (Vanhorenbeeck *et al.*, 2007). The chromatin remodeller *CTCF* was also highly enriched in most ATAC-seq modules and stages and in hypomethylated DMRs, suggesting a high degree of interactions between enhancers and promoters at these positions (Ren *et al.*, 2017) that was not observed in the H3K27ac data, which could reflect the lowest quality of the latter.

To infer which proteins might be mediating the effect of the variants in disease risk during islet development, transcription factor binding sites were detected at T2D-associated SNPs within regulatory regions. ATAC-seq peaks, H3K27ac peaks and hypomethylated DMRs were intersected with the 109,400 SNPs from the T2D 99% credible sets (that is, the group of variants within each T2D-associated locus that collectively explain 99% posterior probability of association [PPA]) (Mahajan *et al.*, 2018a). All variants within each set of regulatory regions were interrogated to see if they disrupted TF binding motifs. **Figure 3.30** indicates the enrichment of motifs whose binding is disrupted by the T2D variant (see **Section 3.2.10**). E2F cell cycle regulators often had their binding disrupted in the three datasets. Conversely, chromatin remodellers ZBTB33 (Kaiso) and CTCF (Defossez *et al.*, 2005) motifs were enriched in variants within hypomethylated DMRs and ATAC-seq peaks, but as observed in the general enrichment analysis with HOMER not in H3K27ac peaks. On the other hand, these last two datasets were enriched in HINFP (MIZF) motifs, which interacts with the histone deacetylase complex to promote transcriptional repression (Li and Liu, 2012). Some well-known regulators of pancreas development also have their binding disrupted, including SMAD (El-Gohary *et al.*, 2013) proteins (in the H3K27ac subset of variants) and HNF1B (within ATAC-seq peaks). The subset of variants with PPA > 10% intersecting each dataset are shown in **Appendix Tables 3.1 to 3.3**, including the TFs whose motifs overlap each variant position (see **Section 3.2.10**).

In conclusion, TF analysis indicates an enrichment in known regulators of pancreas development at the appropriate stages in the sets of regulatory marks examined, with the ATAC-seq dataset emerging again as the one reflecting better the known pancreas development biology, particularly when the peaks are stratified into modules by their profile across stages. Binding of these TFs was altered by the T2D-associated variants, with chromatin regulators being particularly affected, which could explain changes in the expression of target gene effectors at these loci.

Regions analysed	Monogenic diabetes TF	TF genes in T2D loci
ATAC-seq peaks per stage	*RFX6, NEUROD1, PDX1, MNX1, GATA6, HNF1B, HNF1A, HNF4A, PAX4*	*ETS1, TCF12, MNX1, ONECUT1, EBF1, HNF1B, HNF1A, HNF4A, TCF7L2, ZBTB46*
ATAC-seq peaks per WGCNA module	Same as above	Same as above
H3K27ac peaks per stage	*PDX1, MNX1, PAX4*	*HMGA1, TCF4, MNX1, EBF1, BPTF, ZBTB46*
Hypomethylated DMRs	*RFX6, NEUROD1, PDX1, MNX1, GATA6, HNF1B, HNF1A, HNF4A, PAX4*	*ETS1, TCF12, MNX1, ONECUT1, EBF1, HNF1B, HNF1A, HNF4A, NFAT5*

Table 3.1. Highlight of transcription factors enriched in the DNA regulatory marks across all islet development stages. Transcription factors found to be enriched, at any stage or module, in each set of regions tested (left), that cause monogenic diabetes (centre) and that map to T2D-associated loci (right). Regions are defined as ATAC-seq peaks per stage, H3K27ac peaks per stage and hypomethylated DMRs per contrast (each differentiation stage to the iPSC baseline or, for adult islets, against the BLC stage).

Regions analysed	SNPs in regions (% of total 99% credible set SNPs)	Regions overlapping SNPs (% of total regions)	SNPs in regions with PPA > 10%	SNPs in regions with PPA > 50%
ATAC-seq peaks per stage	3,890 (3.6%)	1,975 (1.5%)	73	16
H3K27ac peaks per stage	4,399 (1.6%)	8,841 (8.0%)	90	19
Hypomethylated DMRs	2,435 (2.2%)	1,272 (2.6%)	41	4

Table 3.2. Overlaps of regions per DNA regulatory mark across all islet development stages with SNPs in the credible sets. The total number of unique SNPs in the 99% T2D credible sets (Mahajan *et al.*, 2018a) is 109,400. SNPs is Regions are defined as ATAC-seq peaks per stage, H3K27ac peaks per stage and hypomethylated DMRs per contrast (each differentiation stage to the iPSC baseline or, for adult islets, against the BLC stage).

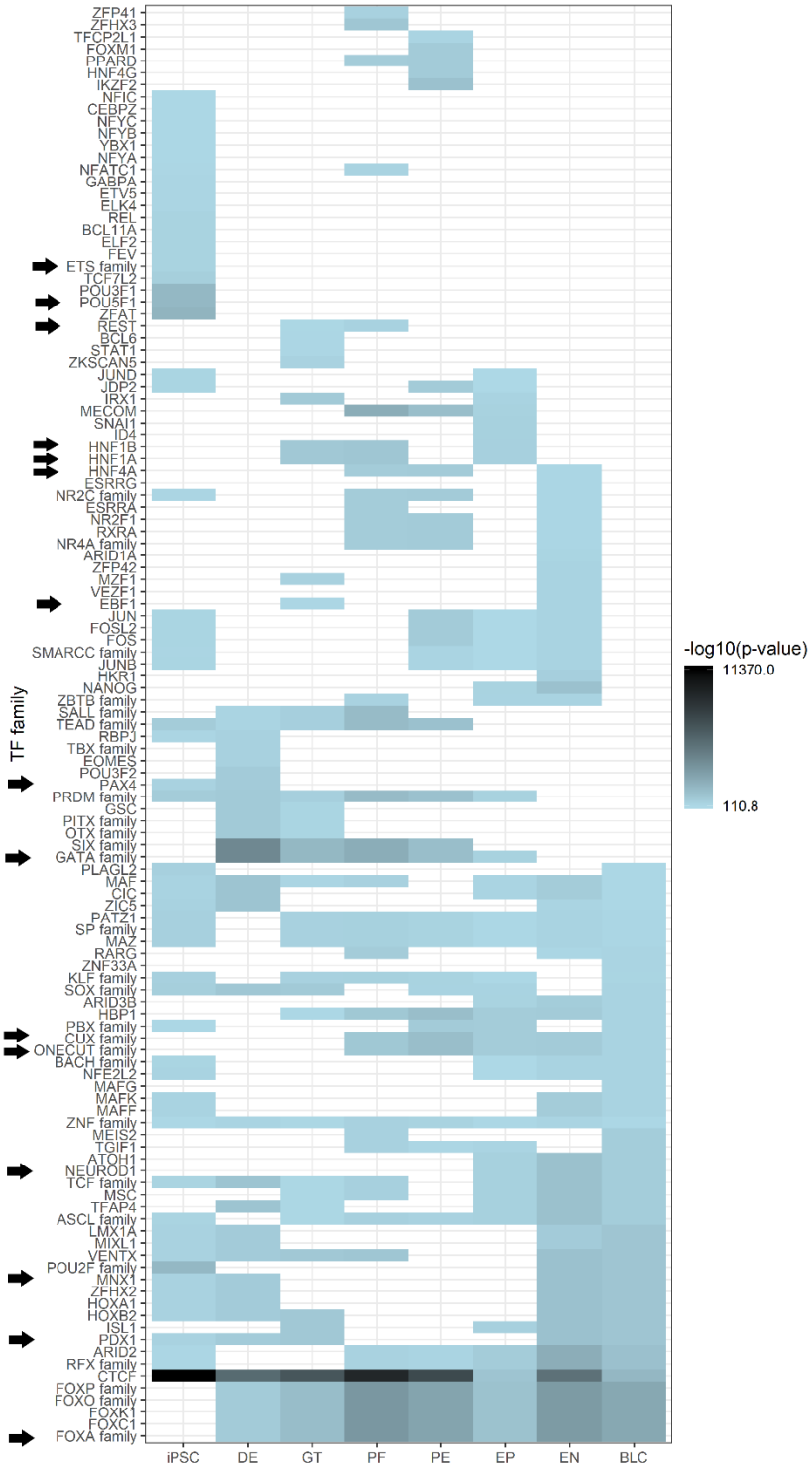


Figure 3.26. Transcription factor enrichment in ATAC-seq peaks per stage. ATAC-seq peaks shared across all stages were removed, and the remaining peaks per stage (x-axis) were tested for enrichment in TF binding motifs. Similar motifs from the same family were grouped together (y-axis). Motifs with p-value < 10^{-10} were considered significant.

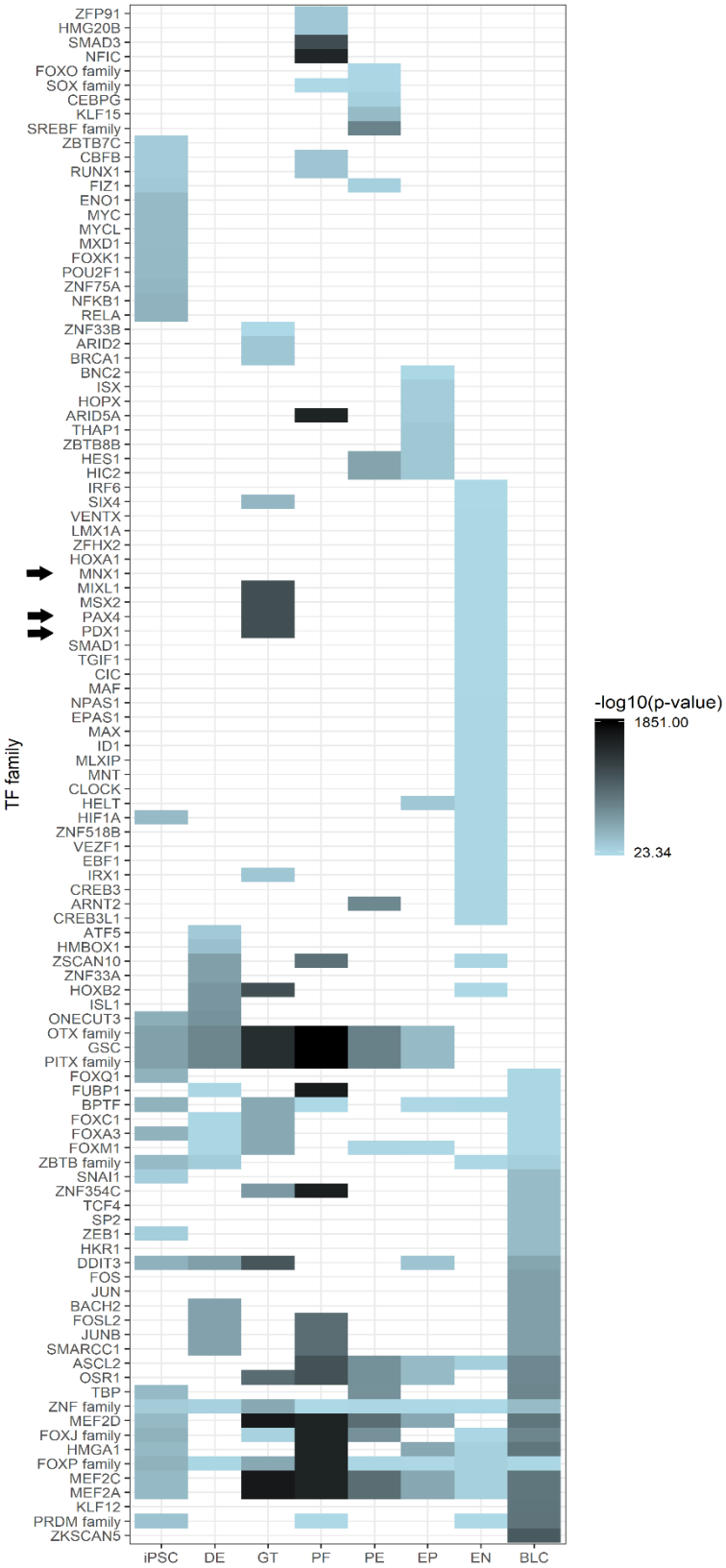


Figure 3.27. Transcription factor enrichment in H3K27ac ChIP-seq peaks per stage. TF motifs found in stage-specific (i.e. exclusive to each stage) peaks per stage (x-axis) were tested for enrichment. Similar motifs from the same family were grouped together (y-axis). Motifs with p-value < 10^{-10} were considered significant.

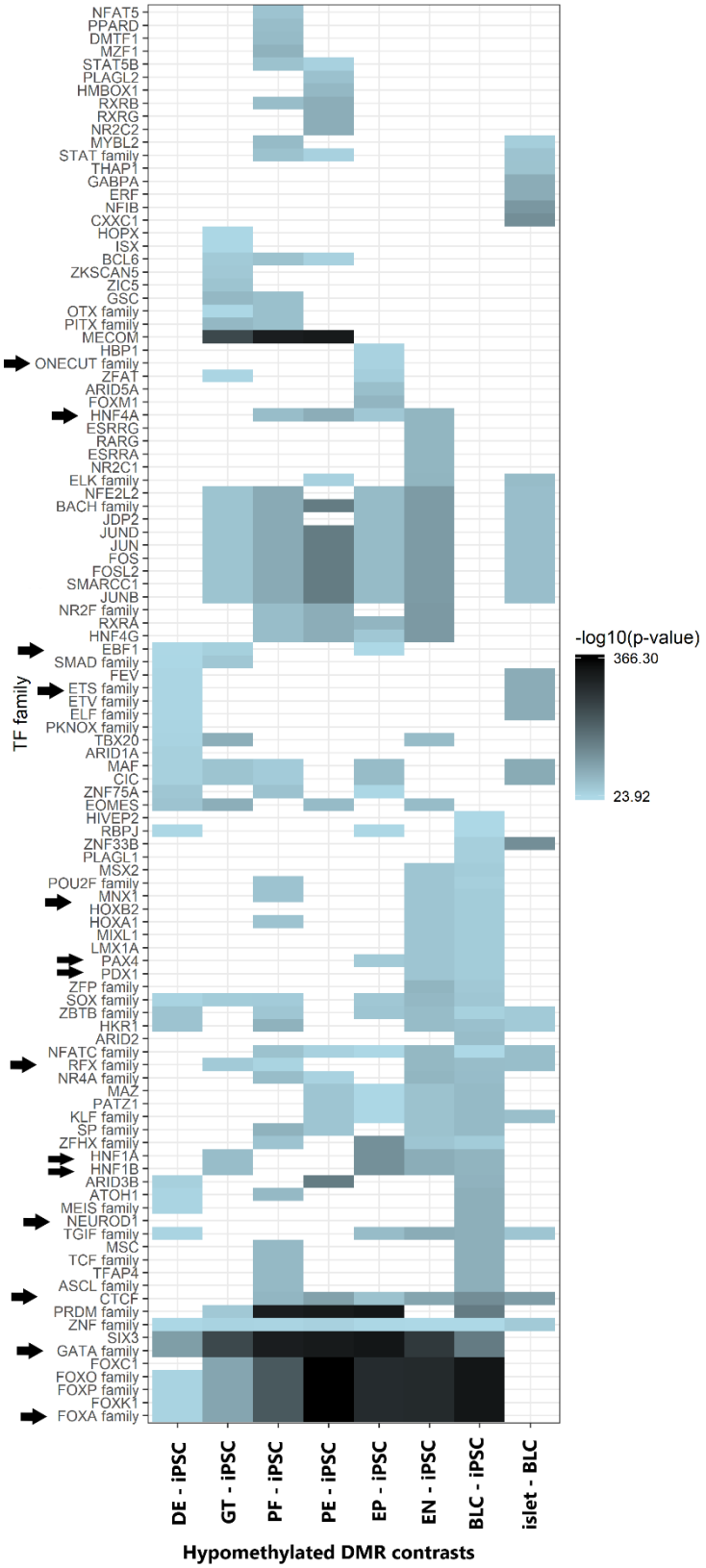


Figure 3.28. Transcription factor enrichment in hypomethylated DMRs. TF motifs found in hypomethylated DMRs, contrasting each differentiation stage to the iPSC baseline or, for adult islets, against the BLC stage (x-axis) were tested for enrichment. Similar motifs from the same family were grouped together (y-axis). Motifs with p-value < 10^{-10} were considered significant.

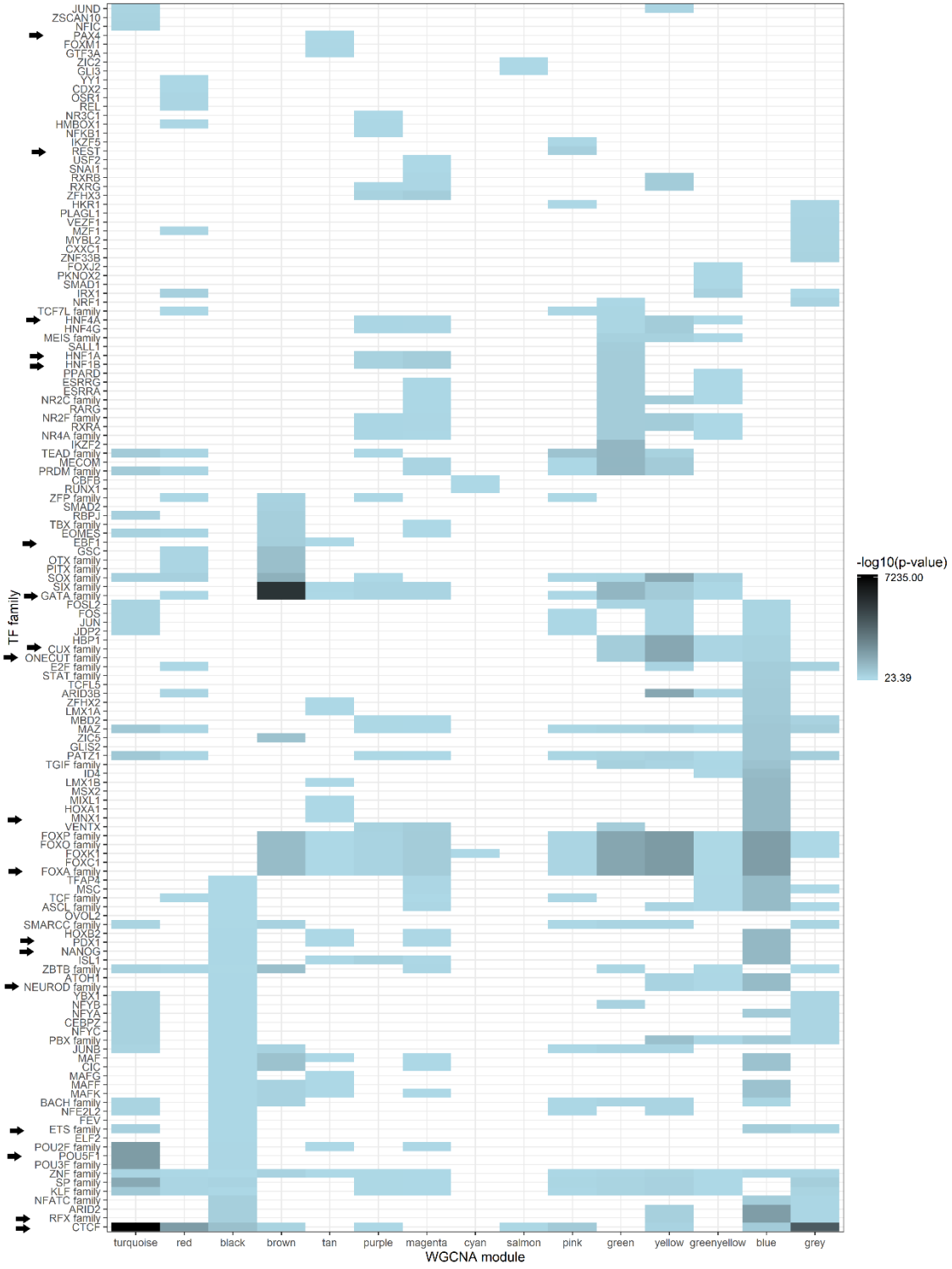


Figure 3.29. Transcription factor enrichment in ATAC-seq peaks modules. ATAC-seq peaks clustered in WGCNA modules (x-axis) were tested for enrichment in TF binding motifs. Similar motifs from the same family were grouped together (y-axis). Motifs with p-value < 10^{-10} were considered significant.

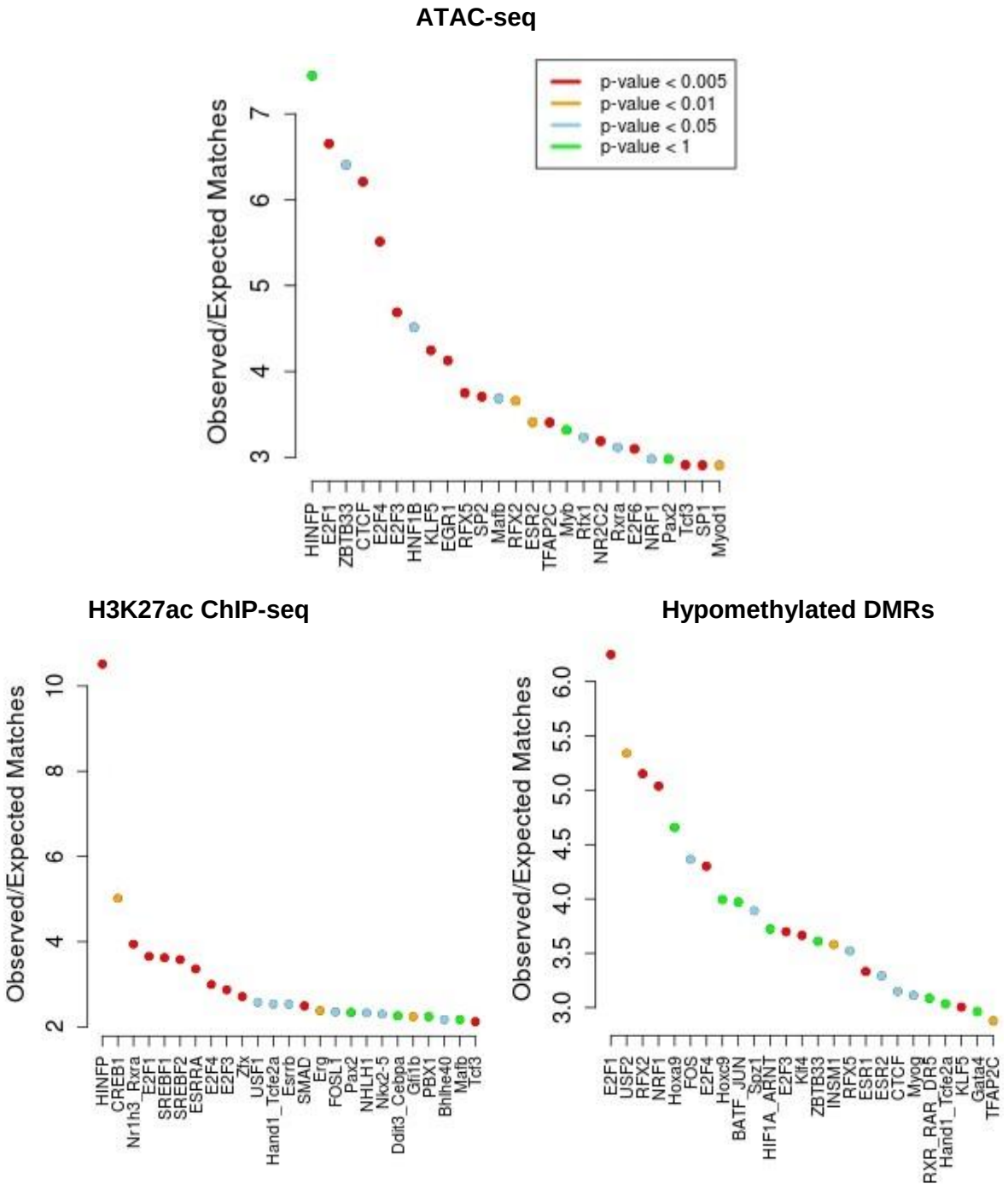


Figure 3.30. Enriched transcription factors whose binding is disrupted by T2D-associated SNPs within DNA regulatory marks. Transcription factors (x-axis) are coloured by p-value of enrichment. y-axis: ratio of observed over expected matches at SNPs in the genome per TF. **Top:** TF motifs disrupted by variants within ATAC-seq peaks. **Middle:** TF motifs disrupted by variants within H3K27ac ChIP-seq peaks. **Bottom:** TF motifs disrupted by variants within hypomethylated DMRs.

3.3.5. DNA regulatory marks are enriched in T2D-associated signals

I set out to investigate through which regulatory signals and at which stages the T2D-associated variants might be mediating changes in disease risk. Previous studies found that islet enhancer regions, defined by ATAC-seq and H3K27ac ChIP-seq histone marks (among others), are enriched in T2D-associated signals (Parker *et al.*, 2013; Pasquali *et al.*, 2014; Gaulton *et al.*, 2015; Thurner, 2018). Thurner and colleagues also found an enrichment in low methylated regions, although it was mostly driven by overlapping chromatin accessibility peaks. I therefore set out to investigate if these enrichments could be recapitulated in any of the islet development stages.

T2D-associated SNPs (from the European ancestry DIAMANTE analysis; Mahajan *et al.*, 2018a) were tested for enrichment in ATAC-seq modules, H3K27ac peaks per stage and methylated regions using fGWAS. fGWAS implements a hierarchical model considering the p-values of association with the trait of all variants (see **Section 3.2.7**). These variants were enriched in H3K27ac peaks (all peaks tested per stage) from GT onwards (lnFE = 1.29 [95% CI: 0.44— 1.86]; **Figure 3.31 red** dots), with the highest estimate in the EN stage (lnFE = 2.34 [95% CI: 1.91— 2.71]). When only stage-specific peaks were considered, just the EN stage was enriched (lnFE = 2.05 [95% CI: 1.14— 2.67]; **Figure 3.31 pink** dots).

With respect to methylated regions, I investigated their enrichment in T2D signals from a "static" point of view (low or high methylated regions, LMRs and HMRs) and a "dynamic" one (hypo- and hypermethylated differentially methylated regions [DMRs]). Hypermethylated DMRs in differentiation stages compared to the iPSC baseline did not present any enrichment (**Figure 3.32 a**), while hypomethylated DMRs did from the PF stage onwards with lnFE estimates ranging from 1.74 (95% CI: 0.31— 2.52) in BLC to 2.14 (95% CI: 0.64 — 2.95) in EP. Both hypermethylated and hypomethylated DMRs were enriched in T2D SNPs when comparing adult islets against the BLC baseline, with lnFE estimates equal to 1.68 (95% CI: 0.81— 2.30) and 1.79 (95% CI: 1.04— 2.35) respectively. Low and high methylated regions (LMRs and HMRs) were tested for enrichment in the same way, resulting in an overrepresentation in LMRs in all stages, while HMRs were significantly

depleted (**Figure 3.33 a**). Because DMRs and LMRs in general were most frequently in promoters (**Figure 3.20 b** and **c**), and it has been observed that this distance effect can drive the enrichment (Pickrell *et al.*, 2014), I tested DMRs and LMRs retaining only those at promoters and another subset removing them. For hypomethylated DMRs, EN and PE were enriched in T2D signals both at promoters and at other genomic features with potential regulatory functions, while the PF stage was enriched only in the latter (**Figure 3.32 b** and **c**). DMRs between islets and BLCs were enriched exclusively at promoters. For LMRs, all stages remained enriched when only promoter regions were considered (**Figure 3.33 b**), while those at more distal regulatory regions were not.

Finally, the enrichment in ATAC-seq peaks was tested per module, and not per stage, as the results from previous sections indicate that these clusters reflect better the biology from the differentiating islets. This approach found four enriched modules (**Figure 3.34**): blue module (lnFE = 2.79, 95%CI = 2.23—3.25); whose accessibility increases through differentiation until it peaks at the last two stages; green (lnFE= 2.58, 95%CI = 1.49—3.31) and yellow modules (lnFE= 241, 95%CI = 1.52—3.02), which peaked at the PF and PE stages respectively; and brown module (lnFE= 1.96, 95%CI = 0.65—2.72), peaking at the earlier DE stage.

These enrichments provide clues of how certain regulatory regions might be influencing disease risk and can be user to lever the evidence of association of each individual variant with the disease in each locus. To do this, I used the functional information provided by the enriched fGWAS modules (**Figure 3.34**) to re-weight the posterior probabilities of association (PPA) of individual variants within the 99% credible sets (see **Section 3.2.9**). Twenty variants within peaks from these modules at 19 loci increased their PPA over 50% (PPAf, **Table 3.3**) compared to the genetic PPA (PPAg) calculated by fine-mapping (**Figure 3.35**). Most variants overlapped peaks in the blue module, including two that presented allelic imbalance for chromatin accessibility: rs252924 in the *ANKRD55* locus and rs7667864 in the *LCORL* locus at middle and late stages of development (**Table 3.4,** see **Section 3.2.11**), none of which have been previously implicated in pancreas development. Other 99 variants presented PPAf over 50% (**Appendix Table 3.4**), without having suffered

such marked increments in PPA from the genetic fine mapping often because they were already the lead variant at each independent signal. This included two *PROX1* variants within the same credible set, hat together account for 99% PPA (brown module across the diagonal line in **Figure 3.35**).

In conclusion, both marks of active enhancers (ATAC-seq and H3K27ac) and hypomethylated regions (linked to an increase in regulatory activity and gene expression) were enriched in T2D-associated signals from middle to late stages of islet development, suggesting that they could be influencing T2D risk through regulatory processes in beta cell differentiation.

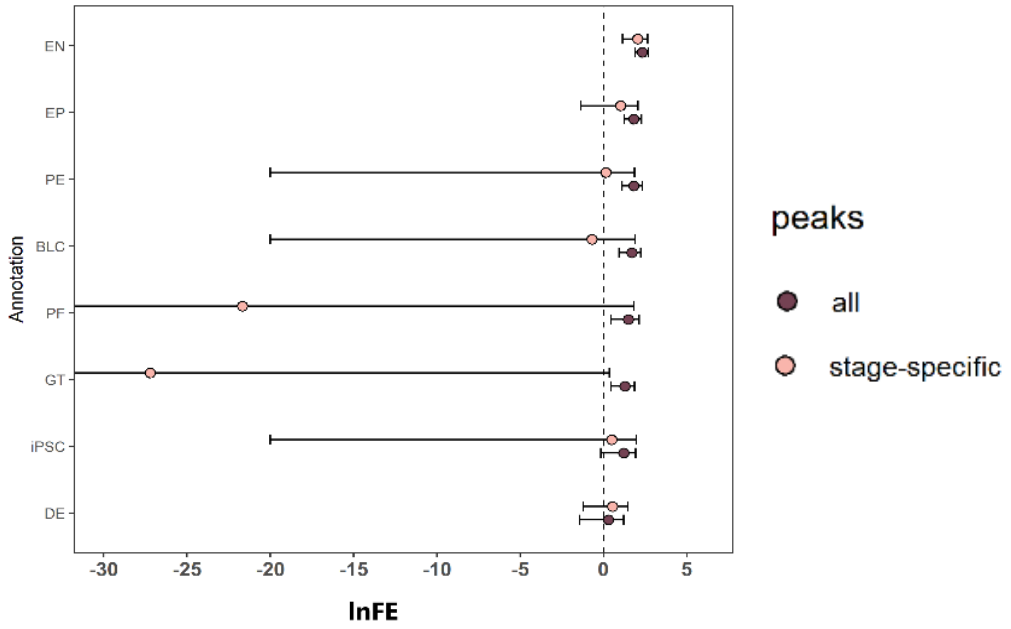


Figure 3.31. Enrichment of H3K27ac peaks in in T2D-associated SNPs. Enrichment estimates per annotation and 95% confidence intervals for the H3K27ac peaks, either all peaks per stage or only those that are stage-specific. x-axis: natural logarithm fold-enrichment (lnFE) of every annotation (y-axis) in T2D-associated loci. Annotations in descending order of lnFE estimate.

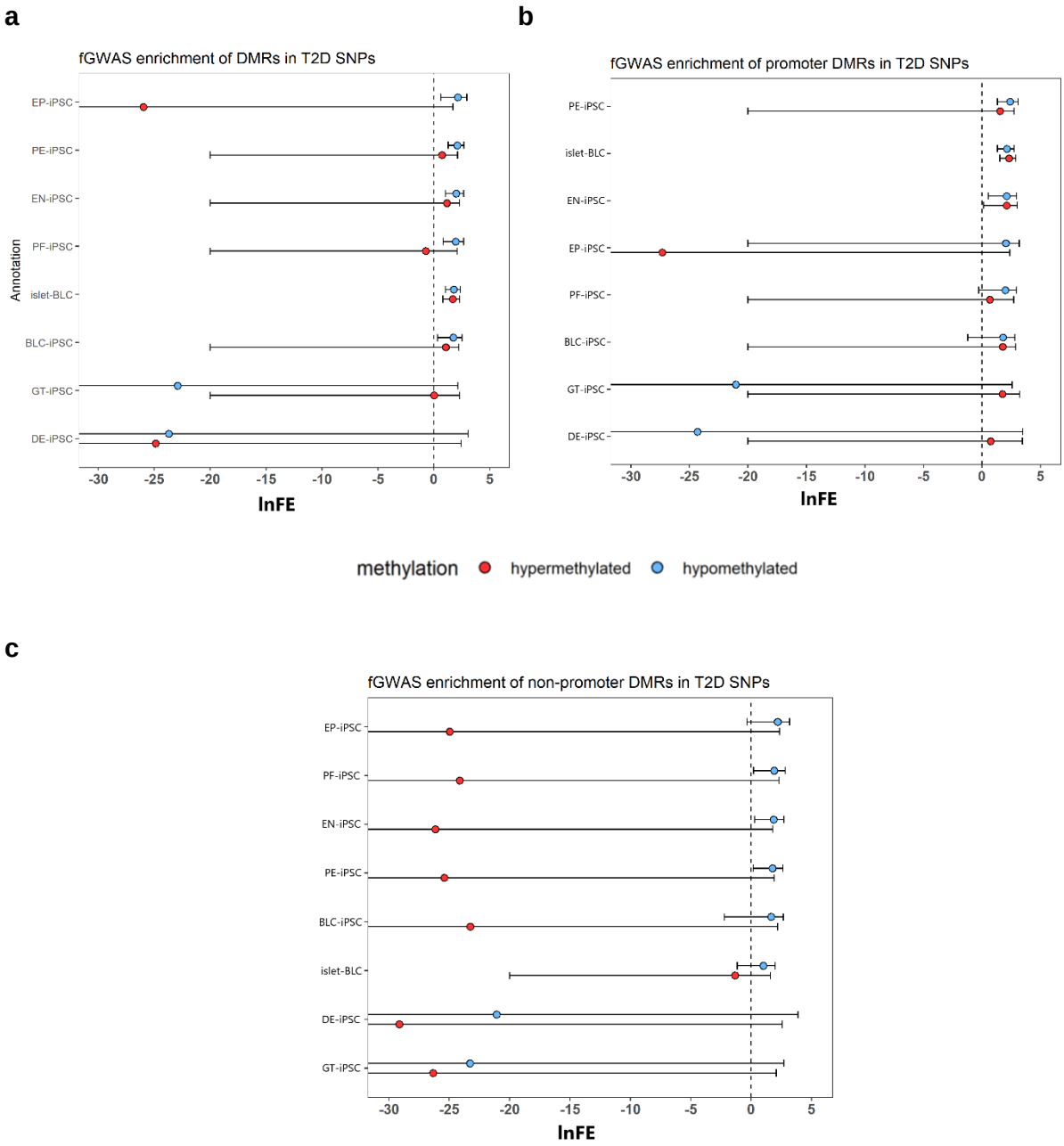


Figure 3.32. Assessing the enrichment of DMRs in T2D. Shown are the enrichment estimates per annotation +- 95% confidence intervals for the DMRs (a), DMRs in promoter regions (b) and DMRs not in promoter regions (c). x-axis: log2 fold-enrichment (lnFE) of every annotation (y-axis) in T2D-associated loci. Annotations in descending order of lnFE estimate.

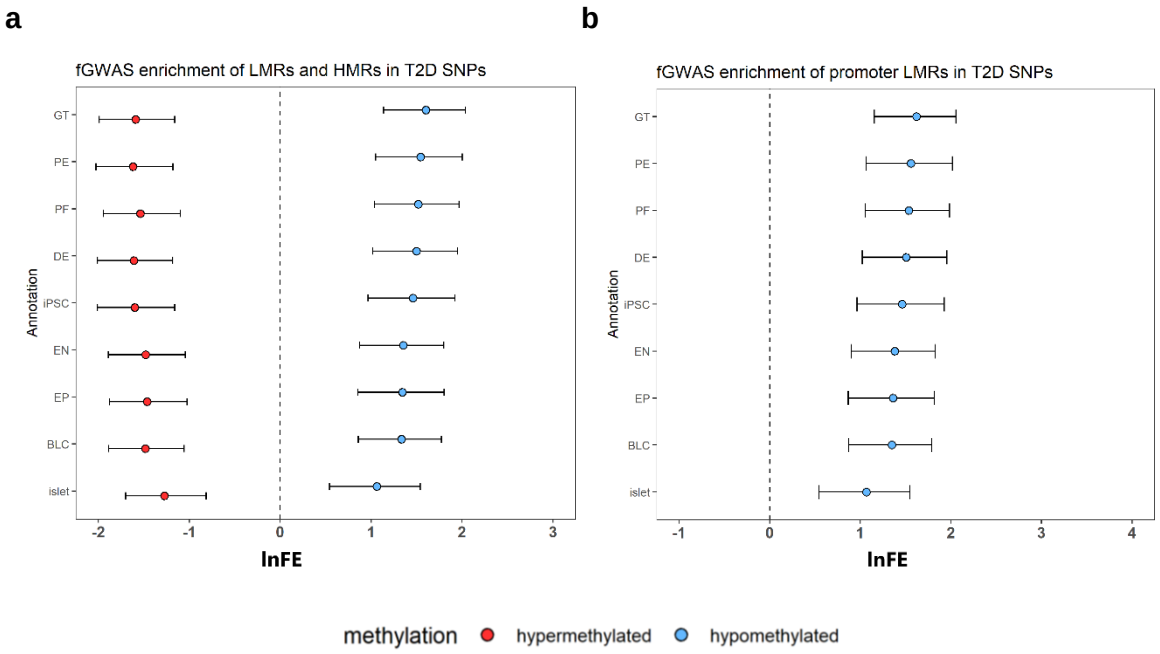


Figure 3.33. Assessing the enrichment of low and high methylated regions in T2D. Shown are the enrichment estimates per annotation +- 95% confidence intervals for the LMRs and HMRs (a), and LMRs in promoter regions (b). x-axis: natural logarithm fold-enrichment (lnFE) of every annotation (y-axis) in T2D-associated loci. Annotations in descending order of lnFE estimate.

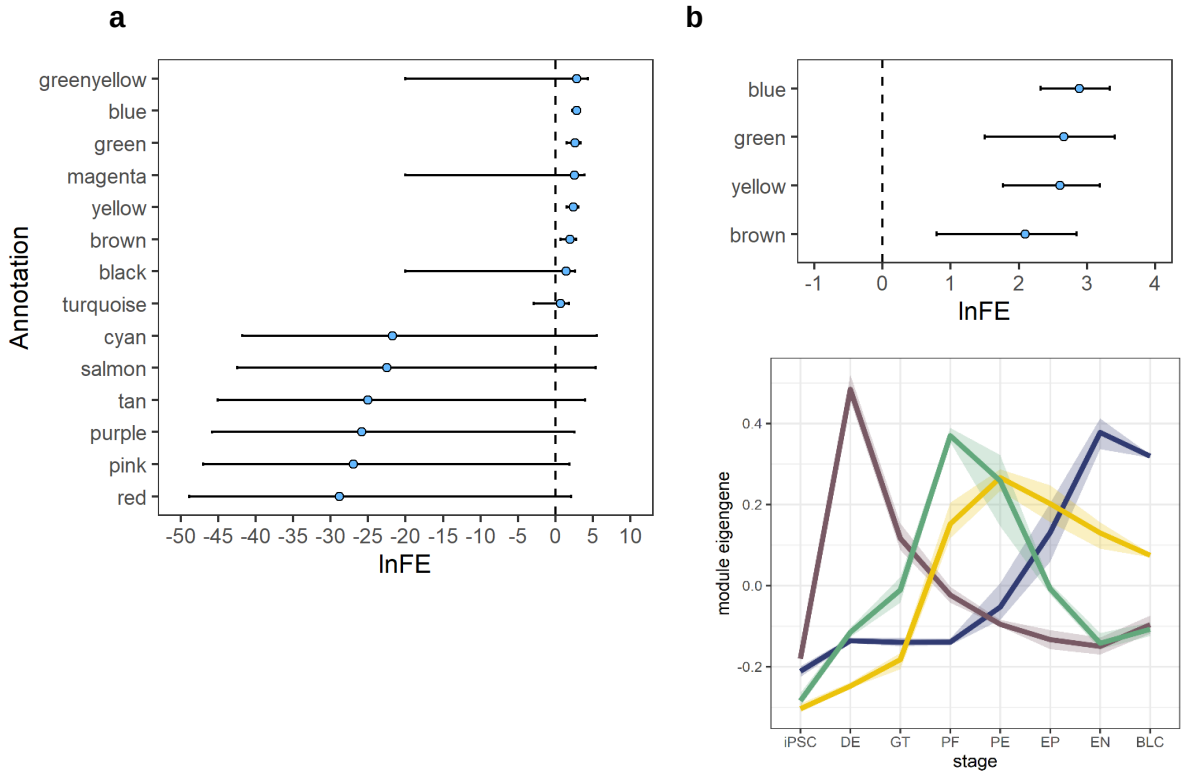


Figure 3.34. ATAC-seq peak clusters are enriched in T2D-associated SNPs. a) Enrichment estimates per module and 95% confidence intervals for the ATAC-seq peaks. x-axis: natural logarithm fold-enrichment (lnFE) of every annotation (y-axis) in T2D-associated loci. Annotations in descending order of lnFE estimate. **b**) Top: Highlight of enriched modules. Bottom: Profiles of accessibility of enriched modules across differentiation stages (x-axis). Y-axis: mean value of the eigenvector for the donor replicates (called "eigengene"). Shading: delimits the minimum and maximum value per stage.

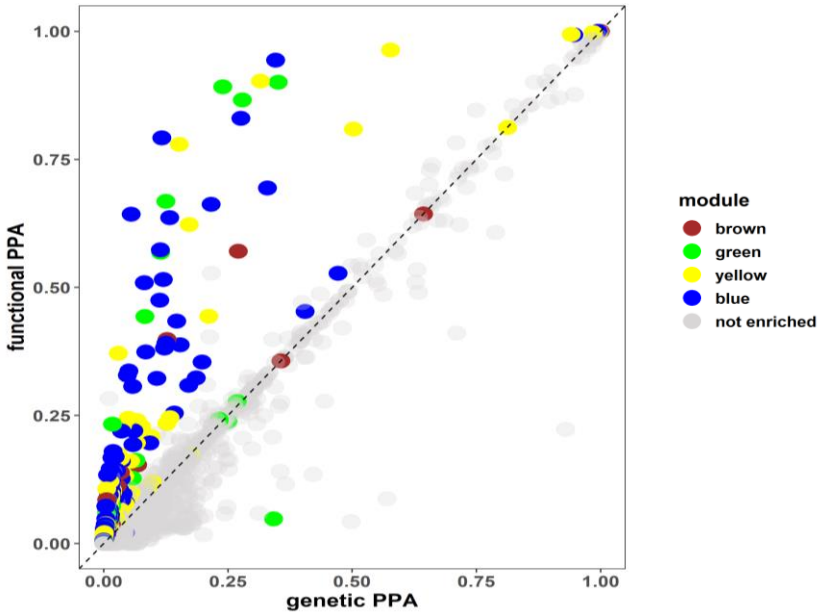


Figure 3.35. Genetic and functionally-annotated PPAs of T2D credible set variants. Functional PPAs (y axis) were calculated using the T2D-enriched modules as priors for fGWAS and compared to genetic PPAs (x-axis). Variants have been coloured according to the enriched modules they fall into. All variants from the 99% credible sets (Mahajan *et al.*, 2018a) are plotted.

Locus	ID	Position	IndexSNP	PPAg	PPAf	Module	PPAf - PPAg
CDKN2A.B	rs10757282	9:22133984	9:22134172	0.35	0.94	blue	0.60
ATP1B2	rs7503751	17:7789542	17:7740170	0.32	0.90	yellow	0.59
SLC30A8	rs13266634	8:118184783	8:118185025	0.35	0.90	green	0.55
GNG4	rs291367	1:235690800	1:235690800	0.24	0.89	green	0.65
FTO	rs1421085	16:53800954	16:53800954	0.28	0.87	green	0.59
ZBED3	rs7732130	5:76435004	5:76424949	0.28	0.83	blue	0.55
ANKRD55	**rs252924**	5:56205643	5:56196604	0.12	0.79	blue	0.67
ARL15	rs55744247	5:52777864	5:52774510	0.15	0.78	yellow	0.63
IGF2BP3	rs78840640	7:23434606	7:23434606	0.33	0.69	blue	0.36
ANKRD55	rs256904	5:55810305	5:55808475	0.12	0.67	green	0.54
LYPLAL1	rs2820446	1:219748818	1:219748818	0.22	0.66	blue	0.45
IGF2BP2	rs869399	3:185826764	3:185829891	0.06	0.64	blue	0.59
FBRSL1	rs77773608	12:133066392	12:133069698	0.13	0.64	blue	0.50
PDHX	rs286925	11:34642668	11:34642668	0.17	0.62	yellow	0.45
BCL11A	rs243018	2:60586707	2:60583665	0.11	0.57	blue	0.46
SHQ1	rs13080180	3:72793712	3:72865183	0.27	0.57	brown	0.30
VEGFA	rs6937438	6:43815364	6:43814190	0.11	0.57	green	0.45
CDKN2A.B	rs10811660	9:22134068	9:22134068	0.47	0.53	blue	0.06
LCORL	**rs7667864**	4:17811914	4:17792869	0.12	0.52	blue	0.40
HMGB1	rs1056130	13:31038651	13:31042452	0.08	0.51	blue	0.43

Table 3.3. Variants with functional PPAs over 50% and genetic PPAs under 50%. PPAs from the credible sets were re-weighted with the module estimates of the best fGWAS joint model containing the blue, yellow, green and brown modules. PPAg: genetic PPA. PPAf: functional PPA. In bold: variants and loci highlighted in the text.

Locus	SNP	PPAg	C1	C2	A1	A2	p	Stage
ANKRD55	rs252924	0.117	53	25	A	G	2.03E-03	iPSC
			94	67			4.01E-02	GT
			87	45			3.23E-04	BLC
PTPN9	rs11072553	0.136	35	8	C	T	4.19E-05	PE
SLC7A7	rs17122776	0.232	24	5	A	G	5.46E-04	PE
MAP3K11	rs12789028	0.107	28	14	G	A	4.36E-02	PE
LCORL	rs7667864	0.12	20	6	A	C	9.36E-03	EP
TP53INP1	rs10097617	0.198	45	26	T	C	3.19E-02	BLC

Table 3.4. Table of significant allelic imbalance in top T2D-associated variants from credible sets. T2D variants from 99% credible sets with PPAg over 10% that show allelic imbalance for ATAC-seq peaks over 1CPM are shown. C1 and 2: counts for variants 1 and 2 (V1 and 2) respectively. p: p-value of binomial test. In bold: variants and loci highlighted in the text. 143 variants presented allelic imbalance in at least one stage at the 5% significance threshold.

3.4. Discussion

The main results of this chapter are:

- ATAC-seq and DNA methylation data had good quality overall for quantitative analysis, while H3K27ac ChIP-seq quality was adequate for qualitative purposes.
- ATAC-seq peaks, H3K27ac peaks and low methylated regions co-localized more often than expected by chance, often at promoter regions, as expected.
- H3K27ac ChIP-seq, DNA methylation and chromatin accessibility regions were located near genes involved in beta-cell related biological pathways at the appropriate stages of development, particularly the ATAC-seq peaks.
- These three datasets were enriched in transcription factors related to beta cell development processes, particularly monogenic diabetes genes *PDX1*, *MNX1* and *PAX4*.
- Binding of chromatin regulators Kaiso, CTCF and HINFP and was affected by variants within the three DNA regulatory marks.
- Chromatin accessibility, H3K27ac peaks and hypomethylated regions were enriched in T2D-associated signals from middle to late stages of islet development.
- Functional analysis using chromatin accessibility data improved fine-mapping at 20 loci, suggesting their role in islet development.

The first step before trying to make inferences from data that cannot be directly mapped to gene function was to perform quality control, following the recommendations from the ENCODE Consortium (2012) (Davis *et al.*, 2018) and the guidelines of Illumina Infinium Arrays. Most indicators revealed the high quality of the ATAC-seq and DNA methylation data, with the exception of EP and BLC donor replicates SBAd2.1 for ATAC-seq, which appeared in **Figures 3.3, 3.6 and 3.7** as outliers, and were finally removed from the WGCNA clustering (see **Section 3.2.1.5**). H3K27ac ChIP-seq on the other hand had low measurements for almost all samples in most parameters, which affected the quality of the final counts, hence my decision to not perform differential analysis in a quantitative way.

There were some differences in the genomic classification of the three kinds of regulatory marks. A low percentage of H3K27ac peaks were at promoters compared to ATAC-seq (7% vs 31%), although the former were more in line with the ENCODE public data analysed.

This could be due to true differences in the genomic distribution of these histone marks, or a by-product of the lower quality of this data due from an incomplete sonication process in the input controls (that had more DNA) and from the known bias that causes the promoters to be sheared (and therefore sequenced) easily and more often (Teytelman *et al.*, 2009). Subsequently, at the peak calling step the input controls would present more reads mapping to the promoter regions, thus reducing the significance of the called peaks at those locations in the ChIP samples (see **Section 3.2.2.3**). Overall, promoters were hypomethylated while introns and distal intergenic regions were mostly hypermethylated. This was consistent with the observation that unmethylated promoters are not necessarily expressed, requiring other factors to intervene for gene transcription (Neiman *et al.*, 2017).

Despite the limitations highlighted by the quality control process, all datasets presented either open chromatin, low methylation or presence of H3K27ac peaks near differentiation markers and key TFs at several stages of pancreas development, as indicated by the genes driving the enrichments in the Reactome pathway analysis and also by the overrepresented TF binding motifs at every stage of differentiation. Three known regulators and monogenic diabetes genes were enriched across all the datasets tested: *PDX1, MNX1* and *PAX4* (Jennings *et al.*, 2015, Pagliuca *et al.*, 2013), although other known regulators emerged as well, particularly in the ATAC-seq peaks clustered into modules (e.g. *FOXA2, GATA4* and *GATA6*, known regulators of the posterior foregut and pancreatic progenitors [Jennings *et al.*, 2015], in modules peaking from the definitive endoderm onwards).

Some less studied candidate regulators of endocrine pancreas development emerged as well. ONECUT1 has been linked to pancreas development by regulating PDX1 (Jacquemin *et al.*, 2003), and its motifs are enriched in ATAC-seq modules peaking from PF forward and in hypomethylated DMRs. *CUX2*, which is expressed in the human dorsal pancreas but whose exact role in development has not been studied (Jennings *et al.*, 2017), presents the same patterns of TF binding motif enrichment.

From the point of view of disease risk, the results suggest that T2D-associated regions are enriched in open (measured by ATAC-seq), active (measured by H3K27ac ChIP-seq) and demethylated regions of differentiating islets, both in earlier and later stages of islet

development. This agrees with the pattern of enrichment observed in RNA-seq data (see **Chapter 2**), providing several layers of evidence that both maturing and mature cells (or as close as they can be with the current model) might be mediating T2D risk. Had I had more time it would have been interesting to see if any DNA regulatory mark in particular was driving these T2D enrichments per stage, taking advantage of the hierarchical model of fGWAS to build the best joint model.

Investigating more closely the variants from the 99% credible sets overlapping each regulatory mark I found that they often disrupted the binding of chromatin remodellers including ZBTB33 (Kaiso), CTCF and HINFP, indicating how these variants could be mediating disease risk through alteration of various regulatory processes at enhancers and promoters. Kaiso and CTCF interact in opposing ways, with the former preventing the insulator activity of the latter and incrementing enhancer activity (Defossez *et al.*, 2005), while HINFP is a positive regulator of the histone deacetylase complex, promoting deacetylation of histone H4 and transcriptional repression (Li and Liu, 2012).

Finally, re-weighting the prior probabilities of the variants at the 99% credible sets with the T2D-enriched ATAC-seq modules showed 20 independent signals mapping to peaks within those modules with posterior probabilities of association lower than 50% that had their PPA boosted over this value, which made them the most likely effector at those loci. Two of these variants showed allelic imbalance at different stages of islet development. One was rs252924 within the *ANKRD55* locus, which was implicated in mediating insulin sensitivity (Harder *et al.*, 2013). Another was rs7667864 at the LCORL locus, implicated in regulating the PRC2 H3K27 methyltransferase, whose precise activity is necessary for the correct regulation of polycomb target genes during mice embryonic development, hinting at a role in gene regulatory processes in human differentiation (Conway *et al.*, 2018). Likely more variants presenting allelic imbalance could have been found if I hadn't been limited by the lack of replicates (thus reducing the power to detect significant changes in chromatin accessibility and limiting the amount of heterozygotes at each variant position), moreover when treating each stage as a separate cell type.

4. Integration of transcriptomic and regulatory information in an islet development model

4.1. Introduction

4.1.1. Integration of *omics* data to find targets of active enhancer regions

As I presented in Chapter 1, mapping regulatory regions of the genome and studying their function is essential to understand how non-coding, intergenic variants (which constitute the majority of GWAS signals) mediate disease risk (The ENCODE Project Consortium 2007; 2012; Nicolae *et al.* 2010).

Often these regulatory regions act as “enhancers”: DNA sequences that control the transcription of genes, at a variable distance from promoters, in introns and intergenic locations. Enhancer activity is often specific to a tissue, cell type or physiological state, making these regions difficult to identify but giving clues about their role in biological processes (Pennacchio *et al.*, 2013).

H3K27ac histone marks and accessible chromatin together identify active enhancers and promoters, and they were examined in the context of islet development in the previous chapter (Thurman *et al.*, 2012; Tsompana and Buck, 2014). While the target of promoter activity is clear, linking enhancers to target genes is not trivial. Recent *in silico* approaches have incorporated chromatin conformation information using Promoter Capture Hi-C (Sexton and Cavalli, 2015; Krivega and Dean, 2012; Schoenfelder *et al.*, 2015) to ChromHMM chromatin maps (Ernst and Kellis, 2012) to link active genes to enhancers within the same chromatin loops (Greenwald *et al.*, 2019, Miguel-Escalada *et al.*, 2019).Single cell analysis also serve to create maps of co-accessible chromatin which are predicted to be acting to co-regulate the same processes at any given time, in a similar way that co-expressed genes are assumed to be functionally related (Chiou *et al.*, 2019). Other analyses rely on experimental perturbations of disease models to co-

localize expression and chromatin accessibility quantitative trait loci with disease-associated variants to find candidate gene effectors (Alasoo *et al.*, 2018).

H3K27ac annotations and chromatin conformation information is often enough to map interactions between enhancers and promoters, particularly in combination with CRISPR interference (CRISPRi) (Fulco *et al.*, 2016; Sanjana *et al.*, 2016; Mumbach *et al.*, 2017). Fulco and colleagues (2019) used one of these CRISPRi tiling experiments to validate a model of enhancer prediction, termed Activity by Contact, which could predict the activity of enhancers to genes using a combination of chromatin accessibility, H3K27ac and Hi-C.

Simple co-localization of H3K27ac and ATAC-seq data in this hiPSC differentiation model of beta cells allow the study of enhancers and promoters at any given stage of islet development, but provide no information about the target genes. The lack of Hi-C data or single cell ATAC-seq impedes the use of the techniques highlighted above and lack of replicates (3 donors per stage), limit the power to detect eQTLs that might link variants within enhancers to target genes. For all these reasons I opted to use in this chapter the ABC method (Fulco *et al.*, 2019) to find active enhancers and target genes at each stage of the islet differentiation model.

4.1.2. Multi-omics analysis in T2D susceptibility and islet development

Several studies have analysed a combination of regulatory and transcriptomic information to infer the role of islets in T2D risk. This way, islet-specific enhancers were found consistently enriched in T2D and fasting glucose (FG)-associated variants (Parker *et al.*, 2013; Pasquali *et al.*, 2014; Thurner *et al.*, 2018). Islet enhancers are also often bound by one or more known regulators of islet identity and development (e.g. MAFB, PDX1, FOXA2, NKX6-1, NKX2-2), indicating their complex interplay (Pasquali *et al.*, 2014, Khetan *et al.*, 2018). T2D variants within these enhancers have also been found to be preferentially bound by monogenic diabetes TFs such as RFX6 (Varshney *et al.*, 2017).

These enhancer annotations can also be used to help fine-mapping, reducing the number of candidate variants at each locus using functional modelling approaches such as fGWAS (Thurner *et al.*, 2018; Mahajan *et al.*, 2018a; Greenwald *et al.*, 2019), and then prioritize the experimental analysis at these loci. Mapping enhancers overlapping T2D-associated variants to gene promoters also facilitate finding causal genes. Greenwald and colleagues (2019) mapped islet enhancers to promoters in this way, using chromatin loops, highlighting loci such as *ABCC8*, *CAMK1D* and *IGF2BP2*. Experimental analysis of *IGF2BP2* in murine models proved that it is involved in insulin secretion pathways. In a similar way, deletion of candidate enhancers at T2D-associated loci reduced expression of linked genes by Hi-C maps, proving the power of CRISPR approaches to interrogate T2D biology in islets (Miguel-Escalada *et al.*, 2019).

4.1.3. Aims of this chapter

The main objective of this chapter was to integrate chromatin regulatory marks and transcriptomic data to study the role of enhancers active in islet development in T2D risk. To do that, I have:

- Used the ABC model to predict active enhancers and link them to target genes.
- Assessed if predicted enhancers are more representative of endocrine pancreas biology than either regulatory mark on its own.
- Inferred transcription factor activity that might be regulating stage-specific gene networks.
- Analysed the overlap and enrichment of the regulatory regions in T2D-associated variants.
- Used the ABC model's ability to predict targets of enhancers to link T2D-associated variants to candidate gene effectors.

4.2. Methods

Plots were created with ggplot2 (Wickham *et al.*, 2016) in R unless otherwise stated.

4.2.1. Detecting active enhancers and their target genes

Active enhancers and their predicted target genes were calculated using the Activity By Contact (ABC) model (Fulco *et al.*, 2019; https://github.com/broadinstitute/ABC-Enhancer-Gene-Prediction).

The ABC model used chromatin regulatory (ATAC-seq and H3K27ac ChIP-seq) data and was implemented in three steps. First, ATAC-seq peaks per stage (in order: iPSC, DE, GT, PF, PE, EP, EN, BLC) that passed the 10% Irreproducible Discovery Rate (IDR) (see **Section 3.2.1.3** and see **Section 3.2.1.4**) were re-centered around their summits and extended 250 bp up and downstream. Peaks within each stage that overlapped by 1 bp or more were merged with BEDTools (v. 2.24.0). These were considered candidate promoter and enhancer regions.

Then, enhancer activity was quantified per stage calculating the geometric mean of the read counts of each ATAC-seq and H3K27ac ChIP-seq replicate sample (SBAd2.1, SBAd3.1 and SBNeo1.1; see **Section 3.2.1.3** and **3.2.3.3**) at candidate enhancers, located ± 500 bp away from the transcription start sites defined by the RefSeq database (O'Leary *et al.*, 2016) mapped to reference genome GRCh37.p13.

Finally, the ABC regulatory score of each enhancer (E) to each gene (G) within 5 Mb was computed using the following formula:

$$\text{ABC score}_{\text{E-G}} = \frac{A_E \times C_{E-G}}{\sum_{e\ within\ 5\ Mb} A_e \times C_{e-G}}$$

Where A is the activity calculated in the previous step, C is the normalized Hi-C contact frequency between the gene's promoter and the enhancer (see Fulco *et al.*, 2019 for more details). This score depends on the activity of all enhancers (e) within 5 Mb on the target gene, represented in the denominator. Contact frequency data, provided by the authors, was calculated from normalized Hi-C anchored at the gene TSS from an average of the following cell types (lacking Hi-C data from the differentiation model):

GM12878, NHEK, HMEC, RPE1, THP1, IMR90, HUVEC, HCT116, K562, KBM7. This data was used because several studies indicate that chromatin loops and other measures of 3D chromatin structure are not cell-type specific (Rao *et al.*, 2014; Sanborn *et al.*, 2015)

I defined as active enhancers those linked to expressed genes (> 1 TPM) with a threshold ABC> 0.022 per islet development stage.

4.2.2. Naïve overlaps of ATAC-seq and H3K27ac ChIP-seq

Consensus sets for ATAC-seq and H3K27ac ChIP-seq peaks were created as defined in **Section 3.2.1.4.** and **3.2.3.4.** respectively. ATAC-seq peaks overlapping H3K27ac peaks by at least 1 bp per stage were merged. Positions that were located more than 2000 bp upstream and 500 bp downstream of the TSS were retained as "naïve enhancers" (16,212 in total).

4.2.3. Genomic annotations and pathway enrichment

Annotation to the closest gene per enhancer and posterior pathway enrichment was calculated as in section **3.2.6.**

4.2.4. Enrichment in transcription factor binding motifs

Enrichment of significant ABC-predicted active enhancers defined in **Section 4.2.1** in transcription factor binding motifs was calculated with HOMER and plotted as in **Section 3.2.8.**

4.2.5 Enhancer enrichment in T2D signals

Significant ABC-predicted active enhancers defined in **Section 4.2.1** that overlap in at least 1 bp across all stages were overlapped and merged at each genomic location using BEDTools (v. 2.24.0). The resulting 30,341 enhancers were analysed for enrichment in T2D-associated signals with fGWAS as in **Section 3.2.7**. The individual estimates for the enrichment were calculated per stage, and the stages that remained in the best joint model (which together present the highest likelihood for T2D-associated SNP

enrichment) were used as input for calculating the functional posterior probabilities of association.

4.2.6 Redefining 99% T2D credible sets

Functional posterior probabilities of association of variants in the 99% T2D credible set were calculated using the significant annotations from the best fGWAS joint model (see previous section) as specified in **Section 3.2.9.**

4.2.7. Publicly available H3K27ac ChIP-seq and ATAC-seq data

These data were described in **Section 3.2.5**.

4.2.8. Transcription factor binding at SNPs

T2D-associated variants within active enhancers were tested for altered binding of transcription factors (TFs) depending on their allele with SNPSelect from SNP2TFBS (https://ccg.epfl.ch/snp2tfbs/snpselect.php; Kumar *et al.*, 2017). Only significant TFs (p-value of binding < 10^{-5}) are reported. Motifs used from JASPAR Core Vertebrate 2014 PWM (Mathelier *et al.*, 2014) from MEME motif database version 12.1 (Bailey and Elkan, 1994).

4.3. Results

4.3.1. Linking enhancer activity to candidate gene effectors

To infer enhancer activity on target genes per stage, I integrated the chromatin regulatory data I have from every stage of beta cell differentiation (8 stages in total, in order: iPSC, DE, GT, PF, PE, EP, EN and BLC). I did this using the Activity By Contact (ABC) model, combining information from ATAC-seq and H3K27ac ChIP-seq, the marks of active enhancers and promoters, from 3 donors per developmental stage. ATAC-seq peaks were used to define the boundaries of the candidate enhancers, and H3K27ac read counts at those positions were included in the enhancer activity model (see **Section 4.2.1**).

I defined as active enhancers those linked to expressed genes (> 1 TPM) with a threshold ABC> 0.022 per islet development stage. Fulco and colleagues (2019) found that enhancers presenting ABC values over this threshold had 63% precision and 70% recall predicting target genes in the empirical validation, compared to 45% precision and 30% recall when considering only the closest gene. The model incorporating only chromatin conformation information had even poorer precision (<10%).

I studied the validity of combining the ATAC-seq and H3K27ac data, keeping in mind the lower quality of the latter (see quality control section of previous chapter). Overlaps of ATAC-seq and H3K27ac peaks (see **Section 4.2.2**) at promoters and enhancers showed that a median of 18% of ATAC-seq peaks overlapped H3K27ac peaks per stage (min. = 11% in DE, max. = 35% in EP; **Table 4.1**). These values were slightly lower but in the range of those taken from several tissues and cell types from ENCODE (median = 29%, range = 11—59%; **Table 4.2**). Therefore, I proceeded to integrate both datasets with the ABC model.

This analysis gave a mean of 10,152 significant (target gene expression > 1 TPM, ABC> 0.022) active enhancers per stage (min. = 6,597 in EP, max. = 12,631 in iPSC; 30,341 enhancers in total) (**Figure 4.1**), a fifth of the ~50,000 ATAC-seq peaks per stage detailed in **Section 3.3.1.1**. On average, 17% of them were exclusive of one stage (min. = 10% in EP, max. = 32% in iPSC), while ~20% of enhancers per stage were active throughout all stages of development. These active enhancers were predicted to modulate the activity of 17,802 expressed genes across all stages (min. = 14,568 in DE, max. = 15,929 in EN).

Due to the difficulty of performing empirical genome-wide assays of the regulatory activity of enhancers, particularly in human data, it is generally assumed that they act preferentially on the closest gene (Carullo and Day, 2019) (see **section 4.2.3** and **3.2.6** for details of the annotation process). An average of 70% of ABC-predicted enhancers were acting (at least) on the closest gene (range = 68% in DE to 73% in PE). Approximately 50% were genic enhancers, at a median distance of ~39 kb from their target gene (min. = 750 bp, max. = ~5 Mb). Importantly, as has been observed before

(Mohrs *et al.*, 2001) enhancers could act on more than one gene: the median number of connections was one, but had a mean of ~2 across stages (max. = 46). Similarly, each expressed gene had a median of 2 enhancers acting on them from iPSC to EP, while the last two stages had a median of one (min. = 0, max. = 15).

In summary, the ABC model predicted in these differentiating islets that each enhancer was acting on average on multiple genes, and that those genes often included the closest gene and another further away. Nevertheless, enhancer perturbation experiments similar to those performed by Fulco and colleagues (2019) would be needed to validate these targets.

Stage	Total ATAC-seq peaks	Total H3K27ac peaks	ATAC-seq peaks overlapping H3K27ac	% Overlap
iPSC	53,212	40,157	7,626	14
DE	51,584	58,060	5,759	11
GT	48,504	69,909	9,973	21
PF	62,455	43,749	7,817	13
PE	55,737	47,730	12,951	23
EP	29,473	46,676	10,283	35
EN	53,965	49,613	15,698	29
BLC	37,568	40,634	5,529	15

Table 4.1. Overlap of ATAC-seq peaks and H3K27ac peaks in beta cell development stages.

Sample	Total ATAC-seq peaks	ATAC-seq in H3K27ac	% ATAC in H3K27ac
A549 rep. 1	71,694	42,606	59
A549 rep. 2	65,058	37,345	57
A549 rep. 3	80,228	45,430	57
Adult adrenal gland	259,878	28,223	11
Adult breast epithelium	522,240	78,019	15
Adult pancreas rep. 1	279,560	85,395	31
Adult pancreas rep. 2	202,829	27,891	14
Adult pancreas rep. 3	190,100	76,199	40
Adult stomach	309,412	82,090	27
Adult thyroid	235,990	63,604	27

Table 4.2. Overlap of ATAC-seq peaks and H3K27ac peaks in other datasets. Publicly available data from ENCODE (see **Section 4.2.5**).

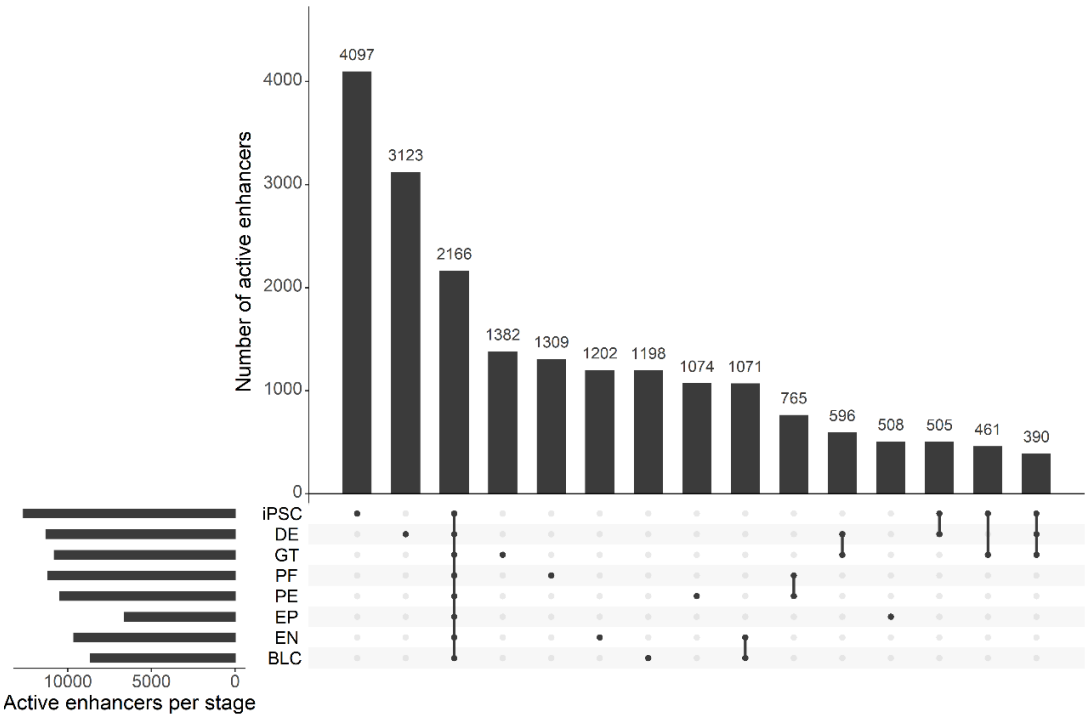


Figure 4.1. Number of ABC-predicted active enhancers per stage of beta cell development. Dotted lines indicate the number of peaks belonging to each stage combination. The number of peaks per stage is shown in the left barplot. Only the 15 most abundant peak sets are shown.

4.3.2. Active enhancer landscape of beta cell development

I first investigated if the active enhancers predicted by the ABC model ("ABC-predicted enhancers") were more representative of the known biology of beta cell development. To do this, I compared these to enhancers defined by simple ATAC-seq and H3K27ac ChIP-seq overlaps ("naïve enhancers"). **Figure 4.2 a** shows the top 10 Reactome (Fabregat *et al.*, 2018) pathway annotations per stage for the ABC-predicted target genes, showing significant enrichment for "regulation of beta cell development" from the PF (Benjamini-Hochberg adjusted p-value [p.adjust] = 0.02) to the BLC stage (p.adjust

= 6×10^{-4}). These included several key regulators of pancreas development mentioned in previous chapters such as *PDX1*, *FOXA2*, *NKX6*-1 or *NEUROG3*. In comparison, naïve enhancers (see **Section 4.2.2**) annotated to the closest gene presented higher adjusted p-values (p.adjust = 0.01 in EP and 0.03 in PE) and lacked annotations to key developmental genes such as *PDX1* or *NEUROG3* (**Figure 4.2 b**). Another relevant term, "regulation of insulin secretion" was also significantly more enriched in the ABC-predicted target genes (p.adjust = 0.01— 0.001 from PE to BLC) than in the ATAC-seq and H3K27ac ChIP-seq overlaps (p.adjust = 0.005 at EN stage).

Because the sets of naïve enhancers had different genomic locations and boundaries to the active enhancers predicted by the ABC model, the differences in biological pathways could be due just to the distinct locations of these two sets of enhancers. To assess the biological relevance of the ABC-predicted target genes themselves, I compared them to the sets of genes mapping closest to each ABC-predicted enhancer. Again, the former were more enriched in members of the "regulation of beta cell development" pathway than the closest genes (**Figure 4.3;** p.adjust values from 0.02 in PE to 0.01 in BLC). Conversely, the term "regulation of insulin secretion" was more significant when annotating the closest genes (p.adjust = 0.02— 1×10^{-5} from iPSC to BLC) than the ABC-predicted targets, despite the latter including more genes in this pathway for the last four stages (medians of 41 and 55 for the annotated closest gene and the predicted target gene respectively). This can be explained by the higher number of target genes predicted by the model that belong to Reactome pathways (indicated per stage in the x-axis of **Figure 4.2 a** and **Figure 4.3**), since each enhancer can map to multiple genes. In addition, mapping to the closest gene gave enrichment in pathways related to more adult beta cells (e.g. "regulation of insulin secretion" or "neuronal system") in earlier developmental stages, in which genes from these pathways are not expressed. Moreover, targets of ABC-predicted enhancers were significantly enriched in pathways related to the endocrine pancreas ("pancreatic bud precursor cells", "regulation of gene expression in beta cells"; not shown in figures) that are completely absent in the genes closest to the ABC-predicted enhancers or in those closest to the naïve enhancers. Comparing these results to those calculated from ATAC-seq or H3K27ac data alone

(**Figure 3.10 and 3.31** in previous chapter) also showed more significant cell-type specific enrichments of the aforementioned pathways.

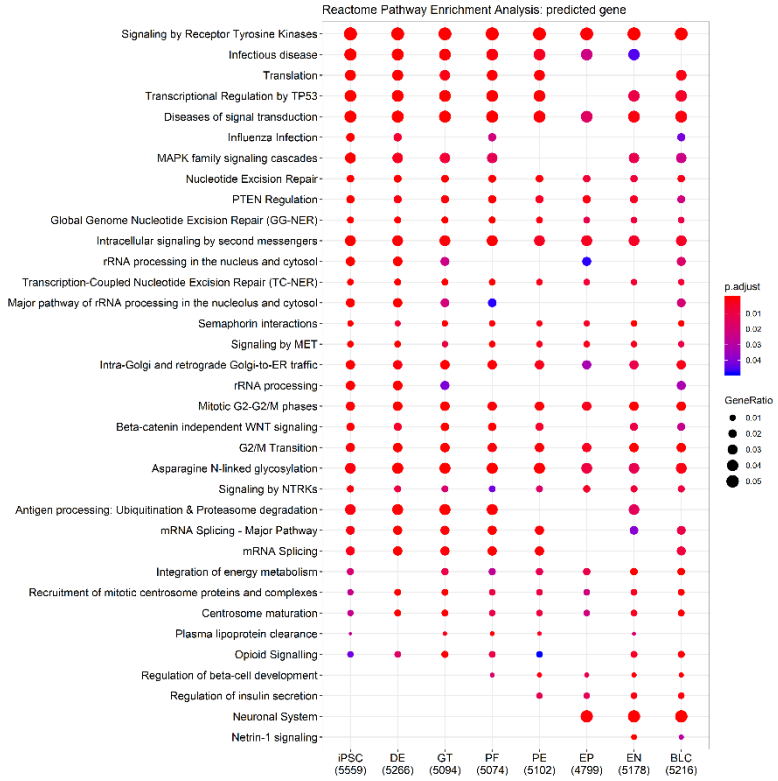
As I mentioned previously, monogenic forms of diabetes often involve defects in pancreas development. I examined the ABC-predicted target genes and found 29 out of 34 monogenic diabetes genes to be linked to an active enhancer at any stage (see **Figure 2.1** for a list of all monogenic diabetes genes). **Figure 4.4** shows their expression patterns throughout the islet development stages. In contrast, genes closest to these ABC-predicted enhancers and to the naïve enhancers contained only 22 and 16 monogenic diabetes genes respectively.

Taken together, these results suggested that the predicted targets of the ABC-predicted enhancers could reflect better the biology of the differentiating stages than the genes closest to them or to the ATAC-seq and H3K27ac peaks, either on their own or combined.

4.3.3. TFs regulating active enhancer networks during development

After establishing that the ABC-predicted target genes reflect better the biological pathways of the differentiating islets, particularly in relation to genes involved in beta cell development, I proceeded to analyse the enhancers themselves for transcription factor (TF) binding enrichment using HOMER (see **Section 4.2.4 and 3.2.8**) in the same way that I did for each individual DNA regulatory mark in the previous chapter. This way I hoped to find clues about genes regulating the identity of each stage of endocrine pancreas development, particularly those within T2D-associated loci.

a



b



Figure 4.2. Pathway enrichment analysis. Enrichments calculated from target genes of ABC-predicted enhancers (**a**) or closest genes to naïve overlaps of ATAC-seq and H3K27ac ChIP-seq at enhancer positions (> 2000bp upstream and 500 bp downstream TSS) (**b**). Enrichment of terms from the Reactome pathway (y-axis) was calculated per stage (x-axis, total genes matched to pathway terms in parentheses). “GeneRatio”: genes found per term over total genes in all terms per stage. “p.adjust”: Benjamini-Hochberg adjusted p-value. Top 10 most significant terms shown per stage.



Figure 4.3. Pathway enrichment analysis of ABC-predicted enhancers matched to their closest gene. Enrichment of terms from the Reactome pathway (y-axis) was calculated per stage (x-axis, total genes matched to pathway terms in parentheses). “GeneRatio”: genes found per term over total genes in all terms per stage. “p.adjust”: Benjamini-Hochberg adjusted p-value. Top 10 most significant terms shown per stage.

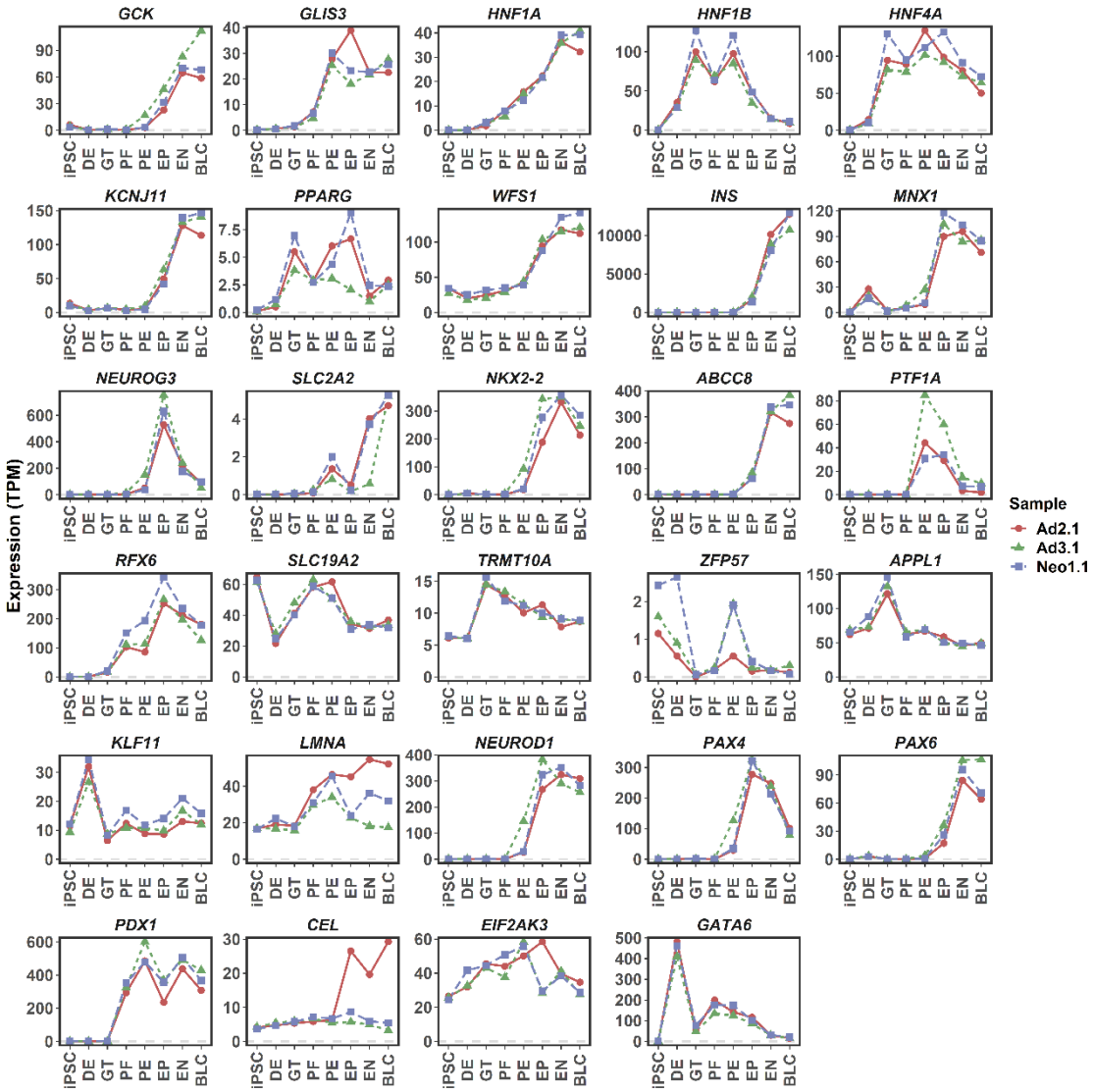


Figure 4.4. Monogenic diabetes genes linked to ABC-predicted enhancers. Gene expression in transcripts per million (TPM, y-axis) throughout he stages of islet development (x-axis) for the three donors.

Figure 4.5 shows enriched TF motifs (p-value < 10^{-10}) and families of motifs (when more than one member is found) per stage. *CTCF* emerged as the most enriched transcription factor at all stages, consistent with its role as a regulator of enhancer-promoter interactions (Ren *et al.*, 2017). ABC-predicted enhancers were significantly enriched (max. p-value at any stage = 10^{-306}) in this TF by an order of magnitude more than those defined by naïve overlaps of H3K27ac and ATAC-seq peaks (max. p-value at any stage = 10^{-42}). This provided further evidence that the ABC-predicted enhancers provide a better picture of the regulatory landscape of each stage.

As was seen in the analogous analysis of chromatin accessibility modules (**Section 3.3.4**), HOMER detected an enrichment in the iPSC stage in binding motifs of *POU5F1/OCT4* and *NANOG*, which are necessary for reprogramming into and maintaining pluripotency. Key regulators of islet development were also enriched at the appropriate stages (indicated in **Figure 4.5** with arrows). *GATA4* and *GATA6* are already active in the DE stage, while *HNF1A*, *HNF1B* and *HNF4A* start to be enriched in the GT enhancers, when the regulators of the foregut endoderm *FOXA2* and *SOX9* start to be expressed. The regulators of endocrine pancreas *PDX1* and *NEUROD1* appear enriched at PE and later, while the enhancers at later stages (EN and BLC) are bound by *MNX1*, *MAFB* and *RFX6*. In **Section 4.3.5** I will show how these TF could be implicated in T2D risk throughout islet development.

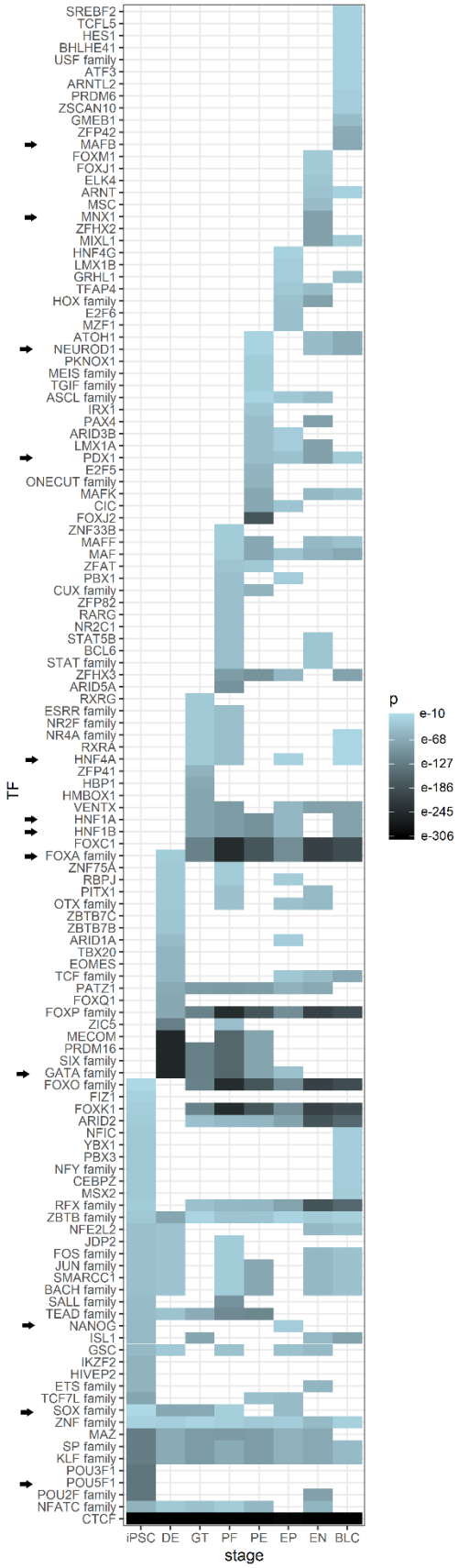


Figure 4.5. Enriched transcription factors per stage of beta cell development, in ABC-predicted enhancers. Transcription factors (y-axis) from the same family are grouped together, coloured by p-value ("p") of enrichment per stage (x-axis). Only significant TFs per stage are coloured.

4.3.4. T2D enrichment in active enhancers

To measure the extent by which T2D-associated signals could be acting through ABC-predicted enhancers, and in which stage of development, I calculated their enrichment in the latter using fGWAS, in an analogous way to the analyses performed in the previous chapter for ATAC-seq, H3K27ac and methylation data separately. Testing all the ABC-predicted enhancers per developmental stage showed that T2D-associated signals were enriched in all stages excluding iPSCs (**Figure 4.6**, top-left), with the highest enrichment in the EN stage (lnFE = 2.82, 95%CI = 2.18—3.31). To see if the enrichment at any stage was driven by peaks that were open in all stages (which would be less informative of stage-specific regulatory processes), I removed those from the analysis (retaining 28,175 enhancers across all stages). This left stages GT to EN as significantly enriched (**Figure 4.6**, top-right), and reduced the enrichments estimates overall (e.g. the most enriched stage was still EN, with lnFE = 2.74, 95%CI = 1.98—3.30).

As indicated before, enhancers and GWAS variants within them can regulate the expression of multiple genes, often the closest ones (van de Bunt *et al.*, 2015) but not always (Gaulton *et al.*, 2015). To see if the T2D signals could act preferentially by regulating closest genes, I tested their enrichment in all enhancers per stage that show this behaviour (21,580 enhancers across all stages), which marginally improved the enrichment estimates for every stage (largest enrichment in EP stage: lnFE = 3.14, 95%CI = 2.36—3.69), while iPSC still remained non-significant (**Figure 4.6**, bottom-left). Conversely, enhancers whose activity is not linked to the closest gene (18,051 enhancers across all stages) removed the significance of enrichment for DE and BLC stages, while keeping the EN stage as most significantly enriched (lnFE = 2.65, 95%CI = 1.75—3.27).

Looking more closely at the variants that make the 99% T2D-associated credible sets (those variants within each signal that collectively explain 99% posterior probability of association with T2D), 1,924 of them (1.8% of total variants) were located within 852 ABC-predicted active enhancers (2.80% of total enhancers) at any stage, with 29 credible set variants with PPA > 10% and 9 with PPA > 50%. 176 out of 380 T2D-

associated loci (46%) presented at least one variant located within an active enhancer. Of these, only 38% had gene targets that matched the gene naming the locus (typically the closest gene to the index variant). This estimate might be higher due to the difficulty of mapping enhancers to the closest gene in gene-dense regions, and because 70% of variant-containing enhancers targeted at least the closest gene. 51% of these enhancers overlapping T2D-associated variants were acting on multiple genes.

These results suggest that some T2D signals could be acting through enhancers regulating developmentally active genes, particularly in mid-late endocrine pancreas development.

4.3.5. Functional analysis of T2D-associated loci

I then assessed if these ABC-predicted enhancers could provide information about the stage-specific regulatory processes that the T2D-associated variants might be mediating. To do that, I used information provided by the ABC-predicted enhancers to re-calculate the posterior probabilities of association (PPA) from the 99% T2D credible sets, in the same way that was done for the chromatin accessibility clusters in the previous chapter (see **Section 4.2.6** and **3.2.9**).

To re-weight the genetic PPA (PPAg) of variants, I first calculated the enrichment estimates from stages that were retained in the best joint model of fGWAS, from all ABC-predicted enhancers (see **Section 4.2.5**). The intermediate PF stage and the more adult EN stage remained significant in this model (**Figure 4.7**)

I then performed the functional analysis itself using fGWAS, which boosted the PPA from variants overlapping active enhancers at these stages (**Figure 4.8**). 49 variants increased their PPA by at least 10%, including 12 variants that gained a functional PPA (PPAf) of over 50%, which are presented along with those that already had PPAg > 50% in **Table 4.3**. These included variants at SLC30A8, ATP1B2, FBRSL1, CENTD2 - ARAP1, MAP3K11, ZMIZ1, HMGB1, GIPR, KCNJ11 and THADA, which became the most likely causal variant for each independent signal in their locus.

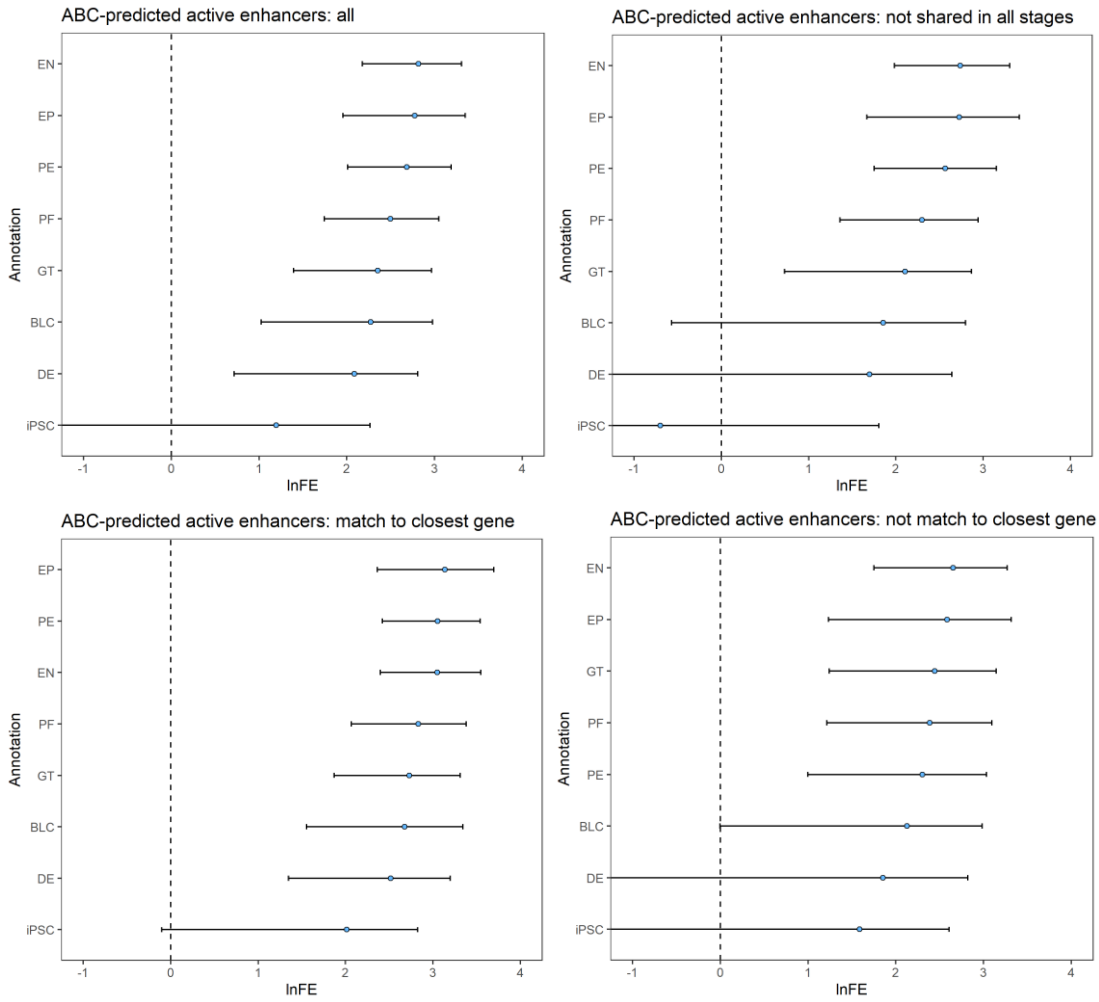


Figure 4.6. Active enhancers are enriched in T2D-associated SNPs. Enhancers in each stage (y-axis) were tested for enrichment with fGWAS (results in natural logarithm fold-enrichment [lnFE] and 95% CI, x-axis).

Most of these variants were within enhancers active both in the PF and EN stage ("both" in **Table 4.3**), which give clues about their activity in mid-late stages of islet development. **Figure 4.9 A** indicates the activity of the enhancers per stage for those overlapping variants with PPAf > 50%. In conjunction with **Figure 4.9 B**, which presents the gene expression of the target promoters, they can be used to prioritize experimental analysis of candidate T2D-associated variants and genes in islet development.

Looking globally at all the variants within active enhancers per stage gives more power to detect the regulatory processes and gene pathways they might be mediating. I analysed the target genes of ABC-predicted enhancers overlapping these signals per stage for enrichment in Reactome pathways (see **Section 4.2.3**). This analysis resulted in only one significant term, "Regulation of beta-cell development" in the PE stage (p.adjust = 0.002), driven by *HNF6* (*ONECUT1*), *HNF1B*, *NEUROG3*, *HNF1A* , *GCK*, *HNF4A*, *INS* and *NOTCH1*.

These TFs were not only ABC-predicted targets, but also were binding to ABC-predicted enhancers as was shown in **Section 4.3.3**. In addition, some of the T2D-associated variants in these enhancers were predicted to significantly alter the binding affinity of these TFs (see **Section 4.2.8**). Globally, 16 TFs were significantly more likely to bind at T2D-associated SNPs within these enhancers compared to the background of all SNPs in the genome. CTCF and the zinc finger protein ZBTB33 appear as the most enriched TFs binding at T2D-associated variants in development, which may be due their mutual interactions as chromatin remodellers in regulatory regions (Defossez *et al.*, 2005). Other significant transcription factors include RFX1 and E2F regulators of the cell cycle progression. In total, 285 signals out of the 1,924 that fell within these enhancers were in a TF binding motif, including important TF for islet development and function (*e.g. PAX6, MAFB, PDX1, PPARG, HNF1A, HNF1B*). **Appendix Table 4.1** gives details of variants with PPA > 10% whose alleles significantly disrupt TF binding.

In conclusion, these integrative analyses provide evidence that some T2D-associated variants might be mediating regulatory processes in mid-late stages of development,

highlighting specific candidate variants and genes and, through their regulatory transcription factors, the pathways they could be mediating.

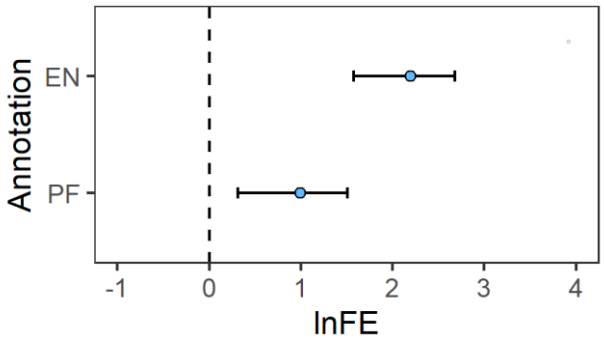


Figure 4.7. Stage annotations retained in the best joint model of fGWAS. Enhancers in each stage (y-axis) were tested for enrichment with fGWAS (results in natural logarithm fold-enrichment [lnFE] and 95% CI, x-axis).

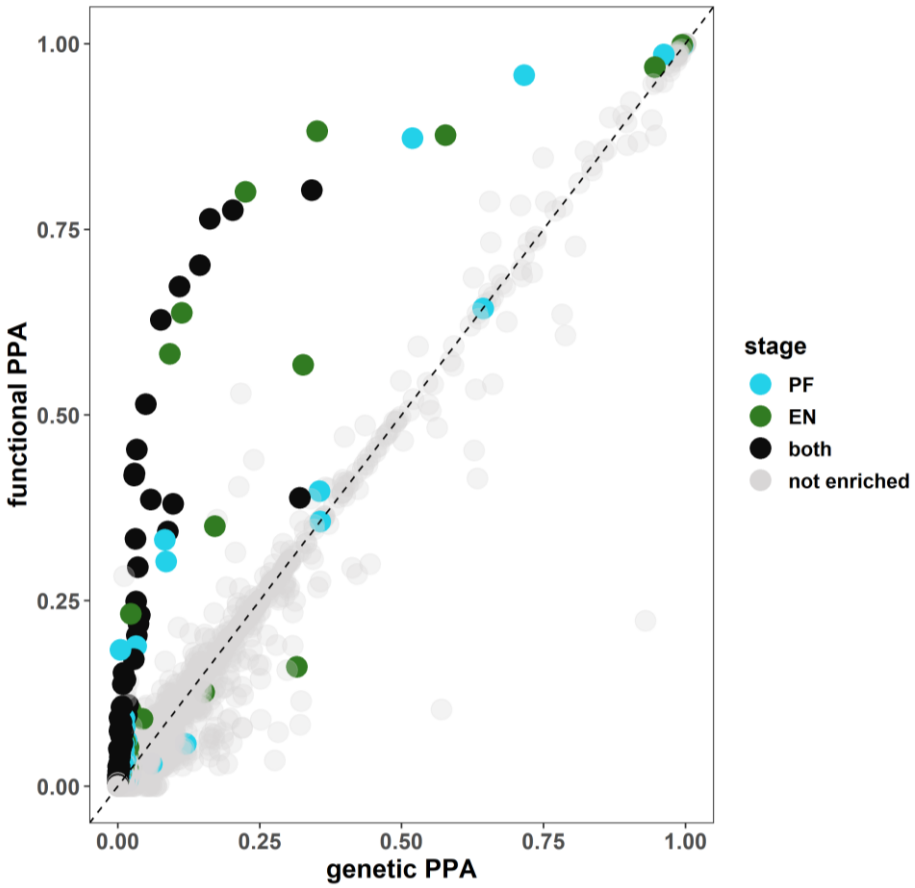


Figure 4.8. Genetic and functionally-annotated PPAs of T2D credible set variants. Functional PPAs (y axis) were calculated using the T2D-enriched enhancers retained in the best joint model as priors for fGWAS and compared to genetic PPAs (x-axis). All T2D-associated variants within 99% credible sets (Mahajan *et al.*, 2018) have been coloured according to the enriched ABC-predicted enhancers they overlap per stage ("both" if active in PF and EN).

Locus	SNP ID	Index variant	Position	PPAg	PPAf	PPAf-PPAg	Stage
KCNQ1	rs231361	chr11:2691500	chr11:2691500	1.00	1.00	0.00	EN
FAM49A	rs11680058	chr2:16574669	chr2:16574669	0.99	1.00	0.00	EN
WDR72	rs528350911	chr15:53747228	chr15:53747228	0.99	1.00	0.00	PF
CRHR2	rs917195	chr7:30728452	chr7:30728452	0.96	0.99	0.02	PF
HTT	rs362307	chr4:3241845	chr4:3241845	0.95	0.97	0.02	EN
INS-IGF2	rs11042596	chr11:2118860	chr11:2118860	0.72	0.96	0.24	PF
SLC30A8	rs13266634	chr8:118185025	chr8:118184783	0.35	0.88	0.53	EN
DGKB	rs17168486	chr7:14898282	chr7:14898282	0.58	0.88	0.30	EN
HNF4A	rs191830490	chr20:43023355	chr20:43023355	0.52	0.87	0.35	PF
ATP1B2	rs62059712	chr17:7740170	chr17:7740170	0.34	0.80	0.46	both
ATP1B2	rs1624085	chr17:7549681	chr17:7553743	0.22	0.80	0.58	EN
ARL15	rs62370480	chr5:52774510	chr5:52774510	0.20	0.78	0.57	both
FBRSL1	rs11614914	chr12:133069698	chr12:133070294	0.16	0.76	0.60	both
CENTD2 - ARAP1	rs7109575	chr11:72460398	chr11:72463435	0.14	0.70	0.56	both
MAP3K11	rs11227217	chr11:65294799	chr11:65307402	0.11	0.67	0.56	both
PROX1	rs79687284	chr1:214150821	chr1:214150821	0.64	0.64	0.00	PF
ZMIZ1	rs703977	chr10:80952826	chr10:80944230	0.11	0.64	0.52	EN
HMGB1	rs73167313	chr13:31042452	chr13:31019580	0.08	0.63	0.55	both
GIPR	rs35816837	chr19:46157019	chr19:46148903	0.09	0.58	0.49	EN
KCNJ11	rs5215	chr11:17408404	chr11:17408630	0.33	0.57	0.24	EN
THADA	rs921069	chr2:43207872	chr2:43206922	0.05	0.51	0.47	both

Table 4.3. Variants with functional PPAs over 50% annotated to stages enriched in the best joint fGWAS model. PPAs from the credible sets were re-weighted with the estimates of the best joint model from fGWAS (with enhancers mapping to stages PF and EN or both). PPAg: genetic PPA. PPAf: functional PPA. PPAf – PPAg: increase in PPA.

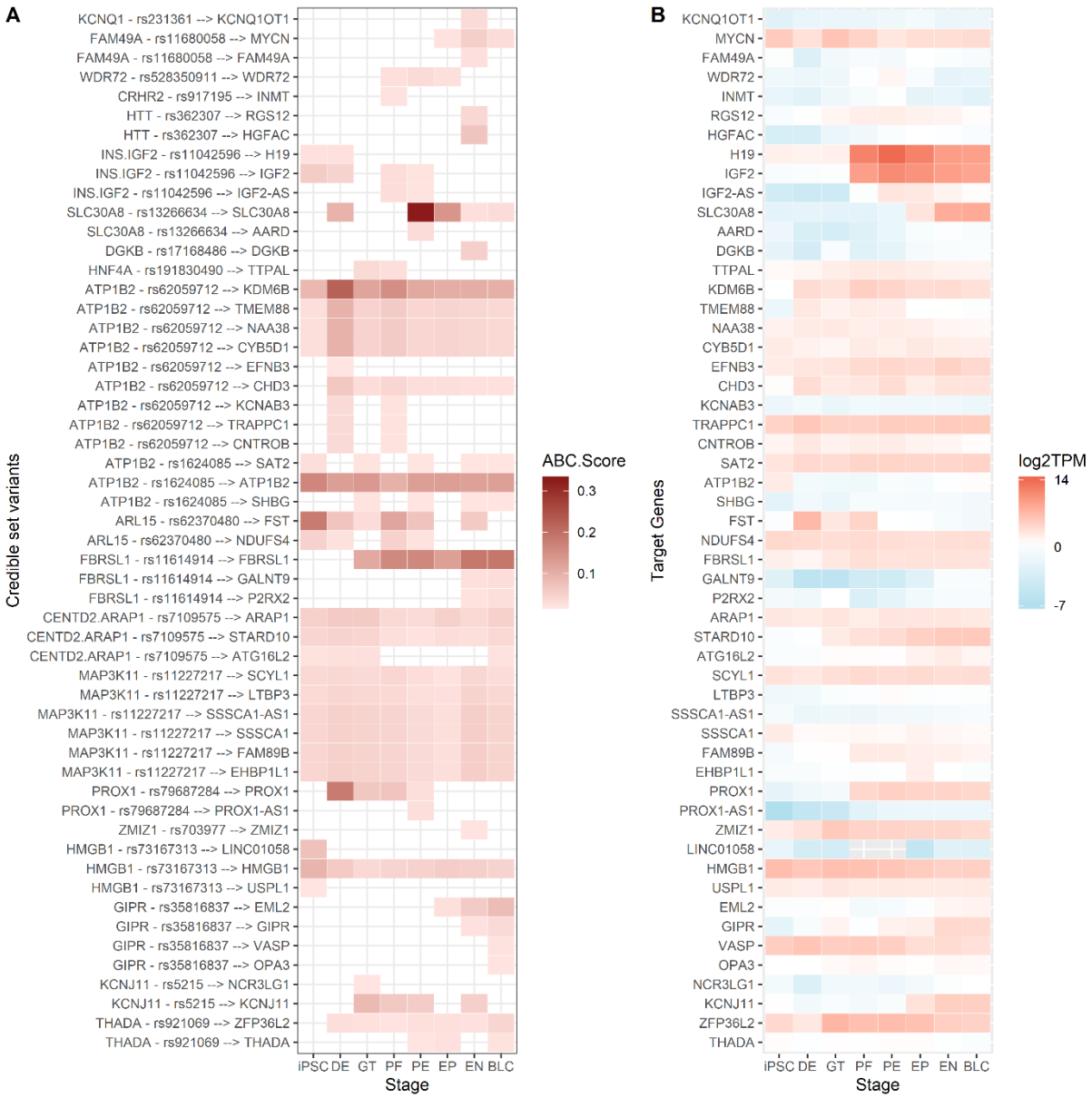


Figure 4.9. Enhancer scores and target gene expression at top PPAf T2D credible set variants. A) Enhancer-gene connection scores ("ABC.Scores") of variants with functional PPA (PPAf) > 50% (y axis, named by "locus name – SNP id → target gene"), in descending order of PPAf, per stage of beta cell development (x-axis). B) Expression of target genes (y-axis) across stages of beta cell development (x-axis), in log_2 transcripts per million ("log2TPM"). In grey: stages where expression = 0 TPM.

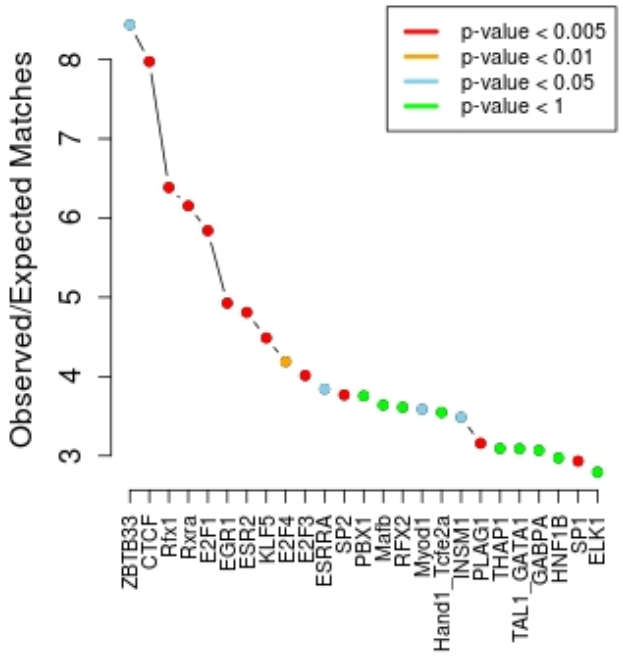


Figure 4.10. Enriched transcription factors whose binding is disrupted by T2D-associated SNPs within active enhancers. Transcription factors (x-axis) are coloured by p-value of enrichment. y-axis: ratio of observed over expected matches at SNPs in the genome per TF.

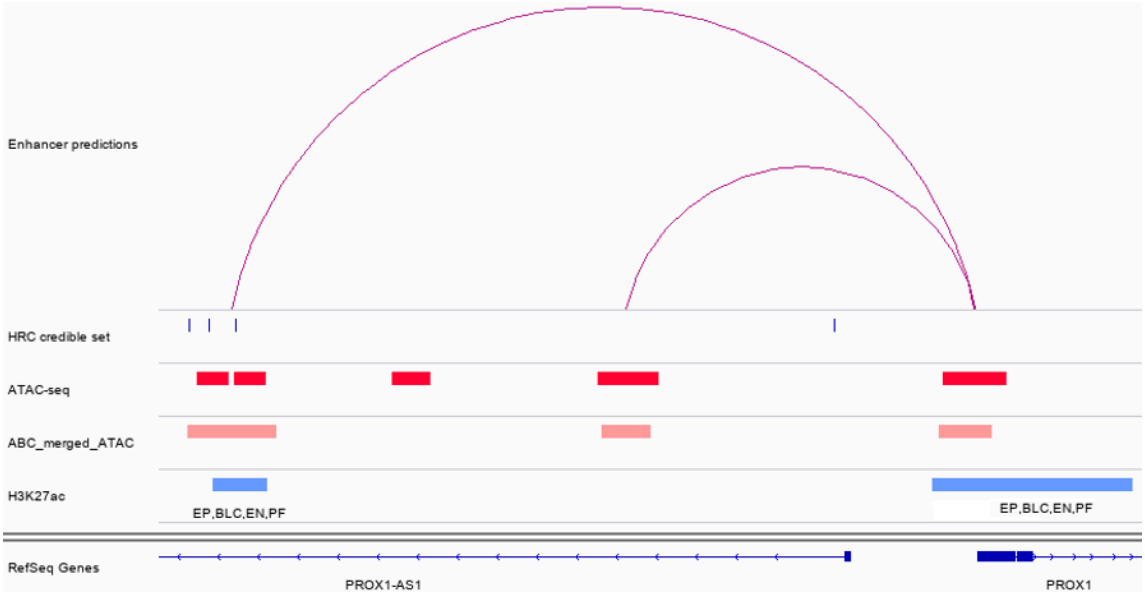
4.3.6. *PROX1* as a candidate T2D gene affecting islet development

Transcriptomic and regulatory information of the beta cell differentiation model can highlight T2D-associated variants and genes that might be involved in changes in disease risk through effects in islet development.

PROX1 is one of such candidates. This locus harbours variants present in the same T2D credible set located at a likely enhancer (**Figure 4.11 A)**, which together have 99% genetic posterior probability of association (PPAg). SNP rs79687284 (PPAg = 0.64) is within an ATAC-seq and a H3K27ac peak that are both open at the intermediate PF stage, that is when the gene starts to be expressed (**Figure 4.11 B)**. SNP rs17712208 (PPAg = 0.36) overlaps an ATAC peak directly upstream the previous one, open from DE until the PE stage, and shows altered binding of the monogenic diabetes TFs HNF1A and HNF1B (**Appendix Table 4.1**). Both are located at intronic positions of *PROX1-AS1* transcript within an ABC-predicted enhancer significantly connected to the *PROX1* promoter (ABC score at PF stage = 0.068), situated approximately 10 kb downstream.

Because of the support provided by the integration of transcriptomic, chromatin accessibility and enhancer activity histone mark, and the mounting evidence from mouse studies that this transcription factor influences pancreas differentiation (Westmoreland *et al.*, 2012; Paul *et al.*, 2016; Wang, 2005; Burke *et al.* 2002), I decided to assess the function of this gene and its candidate enhancer in beta cell differentiation, which will be described in the next chapter.

A



B

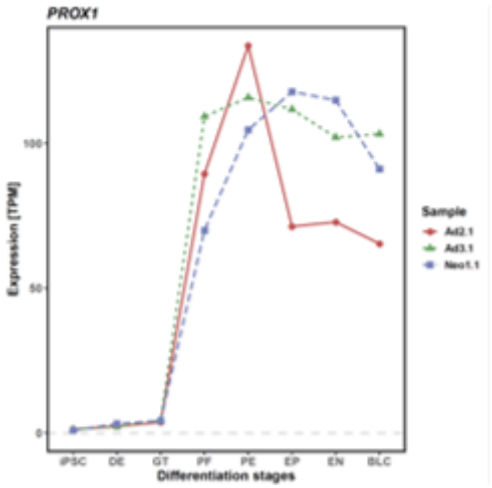


Figure 4.11. *PROX1* T2D credible set SNPs are in a candidate enhancer. A) Significant enhancer predictions (ABC scores > 0.022) of enhancers built from ATAC-seq peaks (in red and salmon, ATAC-seq peaks and merged regions for ABC) and H3K27ac ChIP-seq peaks for the PF stage upstream *PROX1*. SNPs in the T2D credible sets ("HRC credible set") from left to right: rs17712208 (PPAg = 0.64), rs79687284 (PPAg = 0.34), and rs340874 (independent signal, PPAg = 1). **B)** *PROX1* gene expression (in TPM, y-axis) across the stages (x-axis) for the three SB replicates.

4.4. Discussion

The main results of this chapter are:

- Active enhancers predicted by the ABC model act often through their closest genes while at the same time presenting additional distal connections.
- Pathway enrichment and known islet development biology suggests that ABC-predicted genes reflect better the biological landscape of differentiating islets than mapping regulatory signals to their closest genes.
- T2D-associated variants are most enriched in active enhancers at mid-late stages of development, particularly the PF and EN stage.
- T2D-associated variants within enhancers disrupted the binding of chromatin remodellers important for enhancer function, and other TFs regulating pancreas development.
- Regulatory and transcriptomic information, along with functional studies, point to the role of the T2D-associated locus *PROX1* in islet differentiation.

In line with previous empirical studies, islet development enhancers often act on multiple genes (Mohrs *et al.*, 2001) and genes on average have more than one enhancer acting on them at the same or different stages of development, contributing additively or synergistically to increase gene expression (Andrey and Mundlos, 2017). The ABC model limits predictions of enhancer target genes to 5 Mb away, missing connections to even more distal genes.

Analysis of target genes of ABC-predicted enhancers revealed that they were enriched in relevant beta cell pathways at the appropriate stages, implying that these regions recapitulate islet regulatory networks at every stage of development, particularly compared to each regulatory mark on their own or to the simple assignment of regulatory regions to the nearest gene. This was often the only option available due to the lack of modelling frameworks or means to test them experimentally. Recent advances in

CRISPR screening and the possibility of sequencing several kinds of genomic data will continue to facilitate this process (Fulco *et al.*, 2016; Fulco *et al.*, 2019; Sanjana *et al.*, 2016; Mumbach *et al.*, 2017).

Apart from monogenic diabetes genes and other known regulators of each stage of development, enriched TFs included many zinc finger proteins, in line with what was observed by Jennings and colleagues (2017) in a transcriptomic analysis of human foetal dorsal pancreas. These enhancers were also greatly enriched at all stages in binding motifs of the chromatin remodeller CTCF, a well-known regulator of the 3D conformation of the chromatin and facilitator of enhancer-promoter interactions (Ren *et al.*, 2017). CTCF binds to insulator sequences that define the boundaries of topologically associating domains (TAD), which are regions of the genome that preferentially present internal contacts (i.e. within the regions themselves, rather than outside [Pombo and Dillon, 2015]). Recent studies proved that disrupting CTCF binding at TADs affected gene expression by altering the 3D chromatin architecture of the region (Guo *et al.* 2015; Lupiáñez *et al.* 2015). CTCF binding enrichment was markedly more significant in these enhancers than in regulatory regions defined only by ATAC-seq peaks, which also supports the higher activity of these ABC-predicted enhancers in comparison.

Furthermore, T2D-associated variants were significantly disrupting CTCF and its co-binding regulator ZBTB33, also known as Kaiso (Defossez *et al.*, 2005). This suggests a possible disease mechanism of these variants in islet development through altering enhancer-promoter interactions, and therefore gene expression, at their loci. I also found the binding several other TF to be significantly disrupted at T2D-associated variants, including essential TFs in pancreas development and causal of monogenic diabetes HNF1A, HNF4A and PDX1, and RFX binding motifs that were previously observed to be greatly enriched overall in T2D-associated loci (Varshney *et al.*, 2017). Limiting the search of binding sites to a set of variants within enhancers reduces the power to detect these enrichments, but it will increase as new loci are discovered by GWAS.

As other studies have found (Parker *et al.*, 2013; Pasquali *et al.*, 2014; Thurner *et al.*, 2018), T2D-associated signals were enriched in islet enhancers, particularly those in

more adult stages (EN) and intermediate stages of development (PF). This finding agrees with that of Chapter 2, which showed that these variants were enriched also in PF and in the most adult-like EN stage, and adds more evidence to center islet biology and its defects in the pathophysiology of T2D.

Looking at the target genes, I found that most enhancers overlapping T2D-associated variants (62%) were not predicted to act on the gene naming the locus. This was a reminder to avoid the temptation to focus on those genes exclusively in functional studies, and helps to prioritize candidates. It also highlights the obvious problem of inaccurate matching of regulatory regions to their closest possible effector in gene-dense regions, implicating that the percentage of targets that included at least the closest gene (70%) could be higher.

If these predicted enhancers reflect cell-specific biology more accurately, they could provide to be more fruitful for fine-mapping than other regulatory signals. They positioned 14 signals among those with PPAf > 70%, compared to the 9 using ATAC-seq peak modules alone. Combined with the ability to predict target genes, the ABC model could prove to be useful in highlighting candidate variants and genes for experimental analysis. Nevertheless, one should keep in mind that just because a variant is present at an enhancer at any stage of development it does not necessarily mean that it is affecting T2D-risk through this process (or not only). Experimental validation is always needed to clarify the role of these variants, putative enhancers and possible gene effectors in islet development and T2D.

One of these candidate genes is the homeobox transcription factor *PROX1*, with two variants with high PPAf overlapping an active enhancer in middle and late stages of beta cell differentiation. *Prox1* is detected early in the foregut endoderm of mice and in the dorsal and ventral pancreas that arise from those, as well as the early hepatic bud (Burke and Oliver, 2002) and in later stages of endocrine progenitors and several fetal and adult islet subtypes including beta cells (Wang *et al.*, 2005; Hvratin *et al.*, 2014; Paul *et al.*, 2016).

Multiple sources point to the role of Prox1 in murine pancreas development. Mice with pancreas-specific deletions of this gene did not complete formation of the tips of pancreatic ducts, and activated prematurely the expression of acinar genes in the pancreatic progenitors (Westmoreland *et al.*, 2012). Wang and colleagues (2005) also observed significant alteration of key pancreatic genes such as *Ngn3*, *NeuroD* and *Pax4*, with decrease in insulin expression in late endocrine cells in *Prox1* double knock-outs. Conversely, overexpression in postnatal beta cells of mice reduced their volume and the expression of genes involved in islet secretion and processing pathways (Paul *et al.*, 2016), providing an example of the importance of fine-tuning the expression of transcription factors in models of adult and differentiating beta cells. Despite all this murine evidence, no studies have been performed on the specific role of *PROX1* in human pancreas development. In the next chapter I will describe how I used CRISPR regulation approaches to study the role of this gene in islet development.

5. Studying the role of a T2D-associated candidate gene in a beta cell differentiation model using CRISPR

5.1. Introduction

5.1.1. Genome engineering using CRISPR-Cas9

The CRISPR-Cas9 system has revolutionized functional analysis, with applications that include site-specific modifications of DNA, direct labelling of gene loci and modulation of gene expression (Doudna and Charpentier, 2014). In its native form, the Cas9 enzyme uses its catalytic centre to cut double-stranded DNA at specific target sites located next to a protospacer adjacent motif (PAM) sequence, using a guide-RNA (gRNA) which comprises 20 nucleotides of target complementarity. The double strand break is repaired by the cellular repair machinery. This way, single variants or larger sequences can be edited with the help of template by homology-directed repair, and knock-outs of a target sequence can be generated by indels that results from non-homologous end joining, or, if two cuts are provided in *cis*, deletion of the intervening region can occur (Komor *et al.*, 2017).

For modulating the expression of selected genes, a deactivated Cas9 protein (dCas9), that has no catalytic activity, is directed to regulatory regions of target genes (either enhancers or promoters) for their activation ("CRISPRa") or repression/interference ("CRISPRi"). To exert its regulatory activity, the dCas9 is fused to activation domains like the tetramer of the herpex simplex viral protein 16 (VP64) or to repression domains such as the Krüppel-associated box (KRAB) domain (**Figure 5.1 a**), although for repression the binding of dCas9 protein alone is enough to block transcription (Gilbert *et al.*, 2013 and 2014; Qi *et al.*, 2013).

VP64 can activate gene expression by recruiting certain parts of the preinitiation complex like SWI/SNF, PCR4 and CBP/p300 (Neely *et al.*, 1999; Kretzschmar *et al.*, 1994; Ikeda

et al., 2002; Memedula *et al.*, 2003; Hilton *et al.*, 2015). Konermann and colleagues (2015) improved the efficiency of activation with the addition of a fusion protein formed by the activation domain p65, heat shock factor 1 (HSF1) and dimerized MS2 bacteriophage coat proteins, which binds a specific MS2 aptamer RNA motif. P65-HSF1-MS2 binds to dCas9-VP64 complex through the inclusion of the MS2 aptamer sequence incorporated into the hairpin loops of the gRNA (**Figure 5.1 b**). P65 then recruits different parts of the initiation complex to VP64 (AP-1, ATF/CREB and SP1) (van Essen *et al.*, 2009) and all together they increment the expression of the gene of interest. This system, deemed synergistic activation mediator (SAM) can be over 10 times more efficient at activation than dCas9-VP64 on its own, although this varies from gene to gene (Konermann *et al.*, 2015).

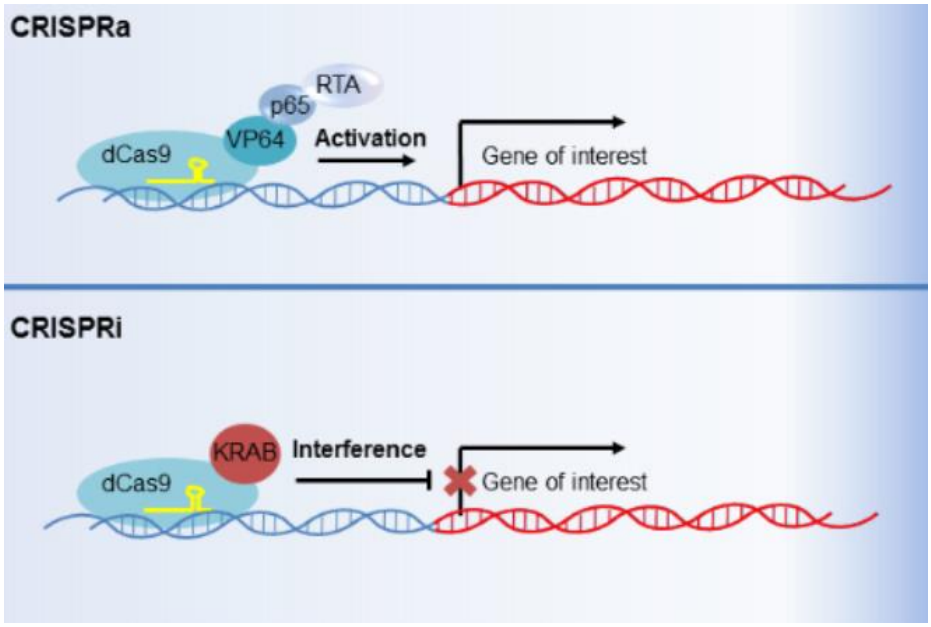
Often these strategies rely on the transfection of several large vectors, an inefficient process that reduces the efficacy of the CRISPRa/i systems. The inefficient delivery of the CRISPRa/i machinery can be circumvented by engineering the target cell line to express the dCas9 components and the p65-HSF1-MS2 activators. This can be achieved by inserting expression cassettes for the machinery into a "safe harbour" locus of the human genome, so that the cells express them constitutively (González *et al.*, 2014). The expression of target genes can then be modulated in these cell lines by simple delivery of the gRNA.

5.1.2. Using in vitro differentiation models of islet development to study T2D-associated loci

As described before, most of the follow-up work from GWAS has been performed in adult islets and the immortalized EndoC-βH1 cell line (Parker *et al.*, 2013; Fadista *et al.*, 2014; Pasquali *et al.*, 2014; van de Bunt *et al.*, 2015; Wang *et al.* 2016; Ackermann *et al.*, 2016; Varshney *et al.*, 2017; Lawlor *et al.*, 2017; Lawlor *et al.*, 2019), but islet development appears to be key in modulating T2D risk (Travers *et al.*, 2013; Fuchsberger *et al.*, 2016). hiPSC-derived models of islet differentiation have been developed to complement and cover the limitations of the murine models (Bruin *et al.*, 2014; Pagliuca *et al.*, 2014; Rezania *et al.*, 2014; Russ *et al.*, 2015; van de Bunt *et al.*, 2016). Genome editing tools,

such as CRISPR-Cas9, are being used for disease modelling and islet development characterization (Beer and Gloyn, 2016).

a



b

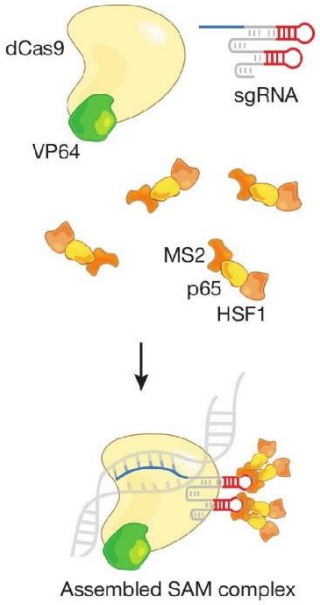


Figure 5.1. CRISPR activation and interference systems. a) Schematic of the modulation of gene expression by CRISPRa and CRISPRi, and their components. Reproduced from Gebre *et al.,* 2018. **b**) Assembly of the components of the CRISPRa SAM system. P65-HSF1-MS2 binds to the dCas9-VP64 fusion protein by the sgRNA. Reproduced from Gilbert *et al.*, (2013).

This chapter describes how to use one of these hiPSC-derived models of beta cell differentiation to characterise the role of the T2D-associated candidate gene *PROX1* in islet development. To do this, in the next pages I will:

- Design and implement CRISPR activation, repression and knock-out strategies of *PROX1* in iPSC cell lines, particularly using those that have been edited to facilitate CRISPR modulation.
- Study if these cell lines can be differentiated towards beta-like cells.
- Assess if the modulation of *PROX1* in intermediate stages of development altered the differentiation efficiency of these cells.

5.2. Materials and Methods

5.2.1. Acquisition of iPSC cell lines

Human induced pluripotent stem cells (iPSC) from SB clone Ad3.1 (from now on "SB"), belonging to an individual without diabetes, were generated as described by van de Bunt and colleagues (2016) and obtained from the StemBANCC consortium (www.stembancc.org) via the Human Biomaterials Resource Centre, University of Birmingham (http://www.birmingham.ac.uk/facilities/hbrc).

iPSCs from the cell line KOLF clone C2 (ID: HPSI0114i-kolf_2, from now on "KOLFC2"), belonging to an individual without diabetes, were obtained from HipSci (Human Induced Pluripotent Stem Cell Initiative; www.hipsci.org).

5.2.2. Generation of KOLFC2 SAM and KRAB cell lines.

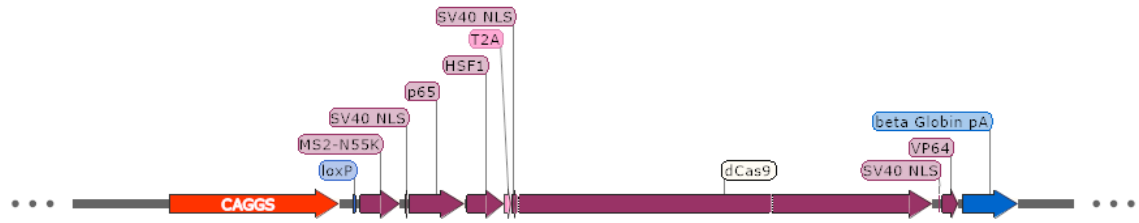
KOLFC2-SAM and KOLFC2-KRAB cell lines were generated by Dr Phalguni Rath by introducing cassettes of dCas9 from *S. pyogenes* fused to p65 activation domain (Konermann *et al.* 2015) or KRAB repression domain (Gilbert *et al.*, 2013) into the *AASV1* locus of KOLFC2 cells using an integrase mediated cassette exchange approach, respectively (Chen *et al.*, 2011). Schematics of the cloned regions are shown in **Figure 5.2**.

5.2.3. *In vitro* iPSC culture

iPSC SB cells were cultured in mTeSR1™ medium (STEMCELL Technologies, #05850) on tissue culture flasks coated with hESC-grade matrigel diluted per manufacturer's instructions (Corning #354277). All iPSC KOLFC2 cell lines were cultured in Essential 8™ (E8) medium (STEMCELL Technologies, #A1517001) on tissue culture flasks coated with hESC-grade Vitronectin diluted per manufacturer's instructions (Gibco™ #A14700).

Cells were maintained at 37°C, 5% CO_2 and passaged as single cells every 3-4 days or when confluent using TrypLE™ Select (Thermo Scientific™, #12563011) including 5uM Y-27632 (Rock Inhibitor, Sigma-Aldrich #Y0503) in the medium the first day of passaging, or using ReLeSR™ (STEMCELL Technologies, #05872).

a



b

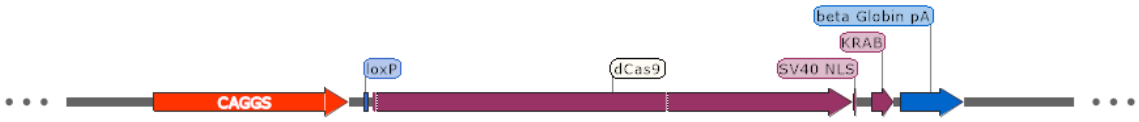


Figure 5.2. The targeted AAVS1 alleles of KOLFC2-KRAB and KOLFC2-SAM cell lines. a) Integrated dCas9-SAM cassette. **b**) Integrated dCAS9-KRAB cassette.

5.2.4. Differentiation to beta-like cells

KOLFC2 iPSC cell lines were maintained as detailed before. All lines were harvested to a single cell suspension using StemPro™ Accutase™ (Gibco™, #A1110501). They were seeded onto Corning CellBIND 12 well plates coated with Growth Factor Reduced (GFR) Matrigel diluted 1:30 (Corning, #356230). Cells were seeded at 5×10^5 cells/well for

KOLFC2 cell lines and at1×10^6 cells/well for SB (unless otherwise stated) in mTeSR1™ medium with 5uM Rock inhibitor and incubated overnight at 37°C, 5% CO_2. Then medium was aspirated, and cells were washed once in PBS (without Mg^{2+} and Ca^{2+}) before adding differentiation media for the first day.

As detailed in **section 2.2.1**, i*n vitro* differentiation involved the addition of recombinant growth factors and small molecules to iPSCs, to generate in order: definitive endoderm (DE), primitive gut tube (GT), posterior foregut (PF), pancreatic endoderm (PE), endocrine progenitors (EP), endocrine-like cells (EN), and beta-like cells (BLC). The differentiation was performed as in Perez-Alcantara and colleagues (2018), but also including an additional well from EP onwards where cells grow on an air-liquid interface. Briefly, cells were collected with Accutase™ at the end of the PE stage and 6-8 clusters of 10ul cells were grown on dry filter inserts (Corning, # 353090) directly on top of the differentiation media in 6-well plates (Corning, #355467).

5.2.5. Immunofluorescence staining

As in **section 2.2.2**, *in vitro* differentiation efficiency was evaluated by measuring the presence of important pancreas development proteins at definitive endoderm (CXCR4) and posterior foregut (NKX6-1 and PDX1). Methods were as described previously (van de Bunt *et al.*, 2016), with details of antibodies and their dilutions listed in **Table 5.1.** Cells were filtered using cell 35 µm nylon mesh strainers (Corning, #352235) before proceeding to FACS.

5.2.6. Generation of vectors

The plasmid dCas9-VP64-GFP (“VP64”; Addgene, plasmid #61422), MS2-P65-HSF1-GFP (“P65”; Addgene, plasmid #61423) and SAM guide vector (“SAM”; Addgene, plasmid #61424) have been designed by Konermann *et al.* (2015). Vector schematics are shown in **Figure 5.3**.

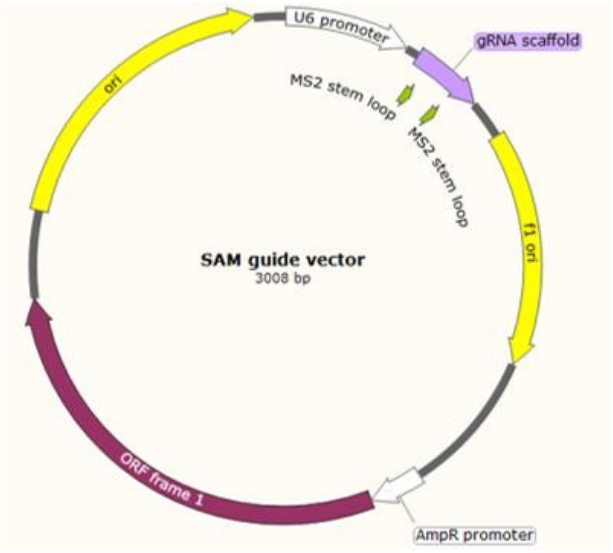
5.2.7. *PROX1* RNA design

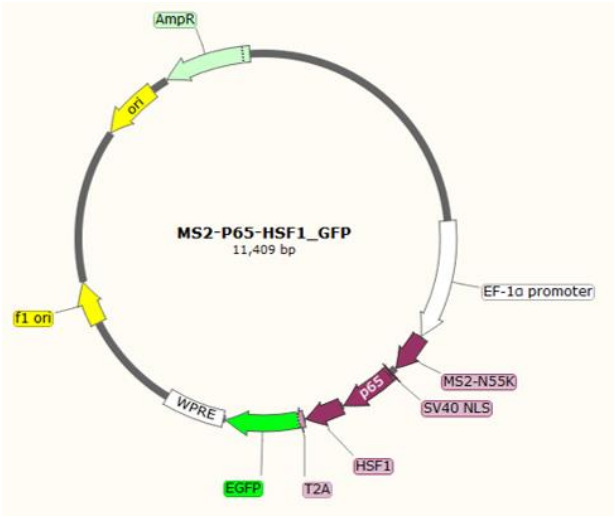
20 nucleotide target sequences for incorporation into the gRNA for CRISPR activation and repression were designed to target a putative *PROX1* enhancer using the web tools

CRISPOR (http://crispor.tefor.net/) (Haeussler *et al.*, 2016) and CRISPR-ERA (http://crispr-era.stanford.edu/) minimizing predicted off-target effects. The selected target sequences are listed in **Table 5.2,** and **Figure 5.4 a** and **b** show where the target sequences are located. *PROX1* sequence and exon annotations are based on Ensembl transcript *PROX1*-205 (ENST00000498508).

Antibody	**Isotype control**	**Dilution**	**Fluorophore**	**Manufacturer**
Anti-CXCR4 #FAB173P-100	Mouse IgG2B control # IC0041P	1:500	PE	R&D Systems
Anti-NKX6-1 #563338	AF647 IgG1 *k* Isotype #557783	1:40	AF647	R&D Systems
Anti-PDX1 #562274	AF488 IgG1 *k* Isotype #557721	1:40	AF488	R&D Systems
Anti-PDX1 #562161	PE IgG1 *k* Isotype #554680	1:40	PE	R&D Systems
Anti-PROX1 #NBP1-30045AF647	Mouse IgG1 Isotype Control #IC002R	1:40	AF647	Novus Biologicals

Table 5.1. Antibodies used for immunofluorescence staining.





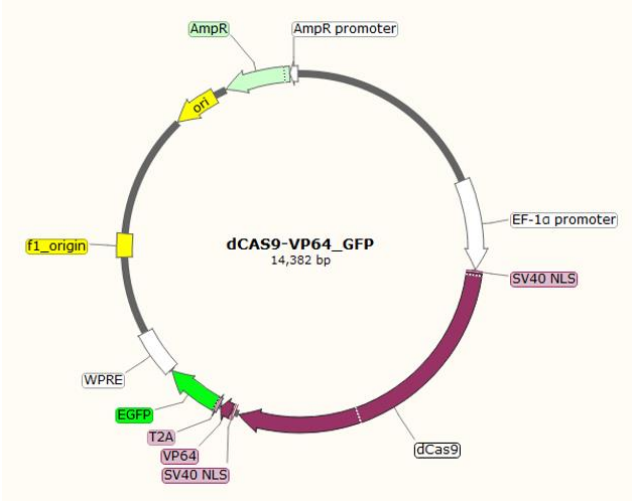


Figure 5.3. CRISPR activation vectors. Activation gRNAs were cloned into the SAM guide vector. Source: www.addgene.org

5.2.8. Cloning of gRNA oligos into SAM vector

3 activation gRNA oligos targeting the *PROX1* promoter and 3 targeting the candidate enhancer were annealed and cloned into BbsI sites of the SAM guide vector (see **section 5.2.6**) using the Golden Gate cloning approach (Engler *et al.*, 2008). The resulting ligation product was treated with Plasmid-Safe™ ATP-Dependent DNase (Lucigen, #E3101K) following manufacturer' instructions, transformed using 10ul product into chemically competent *E.coli*, and plated on LB with 100mg/ml ampicillin for overnight growth at 37ºC. Colonies were picked the following day and plasmid DNA was amplified and extracted using GeneJET Plasmid Miniprep Kit following the manufacturer´s instructions (Thermo Scientific™, #K0503). gRNA integration was confirmed by digesting the plasmids with ScaI HF (New England Biolabs, #R3122S) and BbsI HF (New England Biolabs, #R3539S) for 1h 37ºC and performing DNA electrophoresis, followed by Sanger sequencing.

5.2.9. T7 gRNA synthesis

T7 gRNAs for CRISPR interference and for the Knock-out were synthesized with EnGen sgRNA Synthesis Kit (New England Biolabs, #E3322S) with the oligonucleotides listed in **Table 5.2** according to manufacturer's instructions, followed by DNase I treatment (New England Biolabs, #M0303AA) for 15 min at 37°C. gRNAs were purified with the Monarch RNA Cleanup Kit (New England Biolabs, #T2040L) following manufacturer's instructions, and kept at -80°C until use.

5.2.10. Transient modulation of gene expression by CRISPR activation and interference

5.2.10.1. Using SAM-integrated gRNAs in iPSC SB cells

5ul FuGENE® 6 (Promega, #E2691) were incubated with equal volumes of mTesR per well for 15 min at room temperature (RT). After that, the 10ul mixes were added to equal volumes of 1:1:1 combinations of 100ng/ul VP64, P65, and SAM gRNA vectors (containing the oligos for activator gRNAs or without any inserted, see **section 5.2.6**) or to 10 ul of 100ng/ul GFP vector (the transfection positive control) and incubated for 5 min

RT. Meanwhile, 2.5 x 10^5 cells / well were plated on matrigel-coated 24-well plate with a volume of 500 ul mTESR and 0.5ul rock inhibitor. After DNA incubation, the mixes were added to the appropriate well and incubated undisturbed for 24h at 37°C. GFP expression was checked by microscopy the following day and RNA was collected after 48h.

5.2.10.2. Using gRNAs in differentiating KOLFC2-KRAB cells

Lipofectamine RNAiMAX (Thermo Scientific™, #13778075) diluted in OptiMEM medium (Thermo Scientific™, #31985070) was used to transfect 30 uM gRNAs into the monolayer of differentiating KOLFC2-KRAB cells following the manufacturer's guidelines for 0.25-1 million cells/well. Cells were transfected on the first and second day of differentiation towards the PF stage, and collected 24h after the second transfection. A gRNA that did not target any region of the human DNA ("scrambled", see **Table 5.2**) was used as baseline for repression.

5.2.11. Generation of a PROX1 knock-out in KOLFC2 cells

5.2.11.1. Electroporation of KOLFC2 iPSC.

Two pairs of 5' ("T7-Prox1-KO-5'1" and "T7-Prox1-KO-5'2") and 3' ("T7-Prox1-KO-3'1" and "T7-Prox1-KO-3'2") gRNAs were tested for KO in iPSC KOLFC2 cells (see **Table 5.2, Figure 5.4 b**). A gRNA targeting the *FLNC* gene was used as a positive control for assessing cutting efficiency. For each KO gRNA combination, an electroporation mix was prepared as follows:

Reagent	Volume
gRNA T7-Prox1-KO-5' (20 uM/681 ng)	0.5 uL
gRNA T7-Prox1-KO-3' (20 uM/681 ng)	0.5 uL
Cas9 (20 uM) (NEB, #M0646T)	0.6 uL
Buffer R (Invitrogen, #MPK1096R)	4.4 uL

Target	Oligo sequence (5' to 3') for cloning or synthesis	Name	Type
GAGGACTCCCACGCCTTGTA	CACCGAGGACTCCCACGCCTTGTA	En-Prox1-Ac1A	*PROX1* enhancer activation
TACAAGGCGTGGGAGTCCTC	AAACTACAAGGCGTGGGAGTCCTC	En-Prox1-Ac1B	
TTGGCTGACTGGATCGTTAA	CACCGTTGGCTGACTGGATCGTTAA	En-Prox1-Ac2A	
TTAACGATCCAGTCAGCCAA	AAACTTAACGATCCAGTCAGCCAAc	En-Prox1-Ac2B	
AAAGGTCTGGTGATGCCCGT	CACCgAAAGGTCTGGTGATGCCCGT	En-Prox1-Ac5A	
ACGGGCATCACCAGACCTTT	AAACACGGGCATCACCAGACCTTTc	En-Prox1-Ac5B	
AAATAAACTGGTAGGAGTA	CACCGAAATAAACTGGTAGGAGTA	Prom-SNP-Prox1-A	*PROX1* promoter activation
TACTCCTACCAGTTTATTT	AAACTACTCCTACCAGTTTATTTc	Prom-SNP-Prox1-B	
ACCCAATCGCGAGGCGCCTT	CACCGACCCAATCGCGAGGCGCCTT	Prom-Prox1-Ac2A	
AAGGCGCCTCGCGATTGGGT	AAACAAGGCGCCTCGCGATTGGGTc	Prom-Prox1-Ac2B	
ACGACTGGATCGCCGCGCCT	CACCGACGACTGGATCGCCGCGCCT	Prom-Prox1-Ac3A	
AGGCGCGGCGATCCAGTCGT	AAACAGGCGCGGCGATCCAGTCGTc	Prom-Prox1-Ac3B	
GCGGAGCGCTGAGTTCCACGG	TTCTAATACGACTCACTATAgGCGGAGCGCTGAGTTCCACGGGTTTTAGAGCTAGA	T7-Prox1-Rep2	*PROX1* promoter repression
GAGAAGATACTTACGCGGTGC	TTCTAATACGACTCACTATAgGAGAAGATACTTACGCGGTGCGTTTTAGAGCTAGA	T7-Prox1-Rep3	
GGCGCAAAAGAGCTCTCGCC	TTCTAATACGACTCACTATAGGCGCAAAAGAGCTCTCGCCGTTTTAGAGCTAGA	T7-Prox1-Rep4	
ATCTTCAAAAGCTCGTCAGC	TTCTAATACGACTCACTATAgGATCTTCAAAAGCTCGTCAGCGTTTTAGAGCTAGA	T7-Prox1-KO-5'1	*PROX1 e*xon 2 KO
GCTGTCATGGTCAGGCATCAC	TTCTAATACGACTCACTATAgGCTGTCATGGTCAGGCATCACGTTTTAGAGCTAGA	T7-Prox1-KO-5'2	
GTCTGAAATATCACCTTATTC	TTCTAATACGACTCACTATAgGTCTGAAATATCACCTTATTCGTTTTAGAGCTAGA	T7-Prox1-KO-3'1	
ATCTTGAAGATCGCCGCACT	TTCTAATACGACTCACTATAgGATCTTGAAGATCGCCGCACTGTTTTAGAGCTAGA	T7-Prox1-KO-3'2	
GCTCCTGGGCAGACACCATC	TTCTAATACGACTCACTATAgGCTCCTGGGCAGACACCATCGTTTTAGAGCTAGA	T7-FLNC-5'	*FLNC* KO cutting control

GCACTACCAGAGC TAACTCA	TTCTAATACGACTCACTATAgGCACTAC CAGAGCTAACTCAGTTTTAGAGCTAGA	scrambled	CRISPRi negative control

Table 5.2. Target and oligo sequences from gRNAs.

The mix was incubated for 15 minutes RT. Meanwhile, KOLFC2 cells were passaged using Accutase in 0.5 ml E8 with 0.5 ul Rock inhibitor, centrifuged, and resuspended in 15ul buffer R (Neon™ Transfection System 10 µL Kit, Invitrogen, #MPK1096) to a final amount of $2x10^5$ cells/tube, which were added to their respective electroporation complex mixes and incubated for 5 min on ice. Electroporation was performed with the Neon™ Transfection System Kit detailed above, using the parameters: 1,200 volts, 30 ms, 2 pulses. Cells were seeded in one well per mix of vitronectin-coated 48 well plate, with 200 µl E8 media with Rock inhibitor, grown at 37°C. After 48h, 5,000 cells were plated on 60 mm dishes using E8 with Rock inhibitor and the remaining cells were gathered for DNA extraction to assess cutting efficiency.

5.2.11.2. DNA extraction and amplification.

DNA was extracted using the GeneJet Genomic DNA Purification kit (Thermo Scientific™, #K0721) and DNA was amplified at each cutting position for each pair of KO gRNAs tested and for the *FLNC* control, using the Apptaq DNA polymerase (Appleton, #ARP001) following the manufacturer's instructions. The primers used are listed in **Table 5.3**. **Figure 5.4 b** shows where *PROX1* primers amplify in exon 2.

The PCR product was treated with Exonuclease I (New England Biolabs, #M0293S) and Shrimp Alkaline Phosphatase (rSAP) (New England Biolabs, #M0371) following the manufacturer's instructions to remove the unincorporated primers and nucleotides, and incubated at 37°C for 15 minutes followed by enzyme inactivation (80°C for 15 minutes).

Purified products for each KO gRNA combination were sent for Sanger sequencing (Eurofins Genomics) with the Prox1-R1 primer, while the *FLNC* control was sent with Flnc-F1 primer. The chromatogram was uploaded for Tracking of Indels by Decomposition (TIDE) analysis (https://tide.deskgen.com/, Brinkman *et al.*, 2014), using amplicons from the control transfection targeting *FLNC* to benchmark cutting efficiency.

5.2.12. DNA electrophoresis

DNA products were resolved on agarose (Sigma-Aldrich, #A9539) gels alongside 5ul of 1kb and 100bp DNA ladders (Thermo Scientific™, #SM0311 and #SM0241) running at 130V in 1X TAE buffer (Sigma-Aldrich) until the fragments were resolved.

5.2.13. DNA and RNA quantification

DNA and RNA concentrations were quantified with a NanoDrop ND1000 spectrophotometer (Thermo Scientific™).

5.2.14. RNA extraction

RNA was extracted adding 1ml TRIzol™ (Invitrogen, #15596026) and 200ul chloroform per sample. After mixing and incubation 5 min RT, samples were centrifuged for 12,000xg 10 min 4°C. The upper aqueous layer was transferred to new tubes and 500ul isopropanol with 1:500 5ng/ul glycogen was added. After mixing and incubation 5min RT, samples were incubated at -20°C overnight. Samples were then centrifuged 12,000xg 30 min 4°C, followed by two washes with 1ml 75% ethanol and 15 min centrifugation. Samples were dried to remove leftover ethanol and RNA was resuspended in RNase-free autoclaved water.

5.2.15. cDNA synthesis

cDNA was retrotranscribed from RNA using the qPCRBIO cDNA synthesis kit (PCR Biosystems, #PB30.11-10), following the manufacturer's instructions.

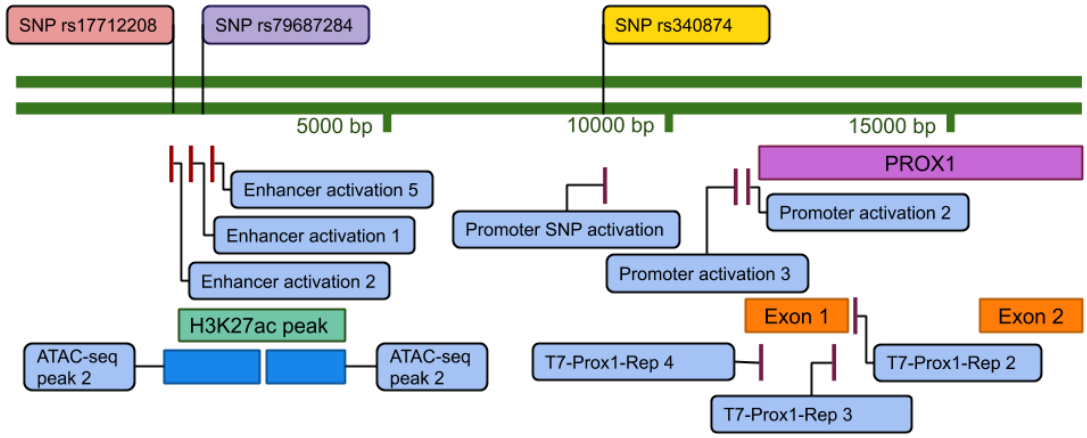
5.2.16. Quantitative PCR

Quantitative PCR was performed on 40ng/well of cDNA using TaqMan® probes (Thermo Scientific™) using TaqMan™ Fast Advanced Master Mix (Thermo Scientific™, #4444556) following the manufacturer's instructions. See **Table 5.4** for full list of probes used.

Oligo sequence for PCR amplification (5' to 3')	Name
GTGCCATAAATCCCAGAGCCTA	Prox1-F1
GTTGTCCGCAAAAACTCCTCTG	Prox1-F2
AAACCCAACCTTTTGGGAAACC	Prox1-R1
TTCATACGAGTTCGCCCTCTTC	Prox1-R2
GCTCCTGGGCAGACACCATC	Flnc-F1
CTTCACTGTGAAGGTGACCGGC	Flnc-R1

Table 5.3. Primers for testing KO in KOLFC2 cells.

a



b

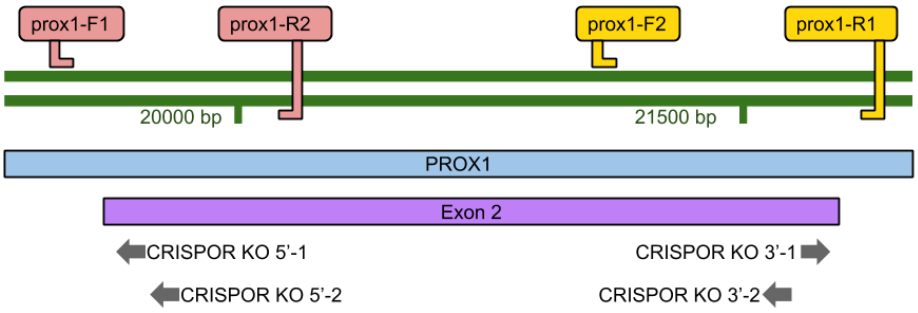


Figure 5.4. Binding of gRNAs at *PROX1* locus. a) Activation and repression gRNAs with respect to putative enhancer and *PROX1* promoter. The two SNPs of interest at the enhancer position are shown overlapping two ATAC-seq peaks (blue). H3K27ac ChIP-peak in green. Binding sites of activation ("Promoter" or "enhancer activation") and repression gRNAs ("T7-Prox1-Rep") are indicated. **b)** KO gRNAs and primers to amplify the cutting locations in exon 2.

Differentiation across the stages was assessed by the relative expression of key marker genes compared to the *TBP* baseline ($2^{-\Delta CT}$) (Livak and Schmittgen, 2001) per stage. Standard deviation of ΔCT was calculated as:

$$sd = \sqrt{\mathrm{sd}_{CT,TBP}^2 + \mathrm{sd}_{CT,target}^2}$$

For CRISPR activation and interference, changes in the expression of genes per gRNA tested were compared to the treatment with transfection reagent only (for CRISPR activation) or with lipofectamine and scrambled gRNA (for repression), normalized to *TBP* ($2^{-\Delta\Delta CT}$). Standard deviation of ΔΔCT for each plating replicate equals the standard deviation calculated above for ΔCT (Livak and Schmittgen, 2001).

t - tests were calculated on the means of ΔCT values per gene and plating replicate in R (v. 3.6.0), stablishing the p-value threshold for significance at 0.05. Plots were generated using the package ggplot2 (v. 3.2.1) in R.

5.2.17. Genotype, RNA-seq and transcript abundance

Transcript abundance and genotype data for KOLFC2 cells was gathered from http://www.hipsci.org/lines/#/lines/HPSI0114i-kolf_2

For SB cells, isoform abundance was derived from RNA-seq data (as described in **chapter 2**), calculating the amount of reads that map to hg19 Ensembl transcripts.

Target gene	Number
TBP	Hs00427620_m1
NKX6-1	Hs00232355_m1
NKX2-2	Hs00159616_m1
GCG	Hs0103136_m1
INS	Hs00355773_m1
PDX1	Hs00236830_m1
PROX1	Hs00896293_m1
SOX17	Hs0075172S_m1

Table 5.4. TaqMan™ probes used for quantitative PCR. All coupled to fluoresceine (FAM), and all but *SOX17* probe span exon-exon junctions.

5.3. Results

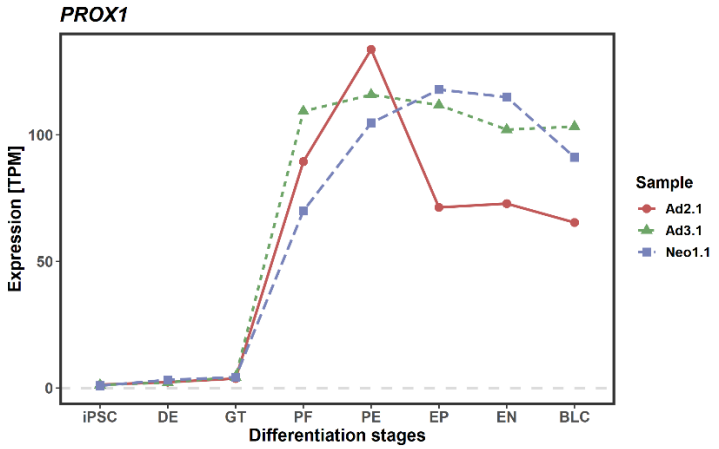
5.3.1. Strategies for modifying *PROX1* expression in differentiating beta cells

In chapter 4 the integration of transcriptomic and regulatory information of the beta cell differentiation model highlighted *PROX1* as a suitable T2D candidate gene involved in beta cell differentiation. I studied the effect of *PROX1* in this process following two approaches. Firstly, *PROX1* expression was modulated transiently and exclusively at the PF stage (when it starts to be transcribed, see **Figure 5.5 a**) using CRISPRa and CRISPRi. As described in the introduction to this chapter, these techniques use a deactivated Cas9 (dCas9) coupled to activation or repression cofactors that can be directed, with gRNAs, to the *PROX1* promoter and the candidate enhancer (defined by ATAC-seq and H3K27ac-ChIP-seq peaks overlapping two T2D associated variants in this locus, see **Table 5.5**). To deliver the minimum number of vectors, KOLFC2 cell lines were used that had the dCas9 and activation and repression cofactors integrated in the *AASV1* "safe harbour" locus in the cell: KOLFC2-SAM and KOLFC2-KRAB respectively (see **Methods**). The SB cell line was used to test the activation gRNAs. Secondly, I attempted to generate a *PROX1* KO in KOLFC2 cells that I could then differentiate towards the most adult stage (BLC). This could potentially provide the effect of the complete downregulation in case the CRISPRi did not produce a phenotype strong enough to be detectable among the high inherent variability of the differentiation.

5.3.2 Gene architecture and design of gRNAs

To design the gRNA sequences to target *PROX1* the gene architecture was studied first. Seven isoforms of *PROX1* exist (**Figure 5.5 b**), but due to the similarity of their sequences it was not possible to accurately quantify them at any stage using RNA-seq for SB cells. iPSC KOLFC2 had less than 1 transcripts per million (TPM) at any isoform, and thus it was not possible to conclude that any transcript was more abundant. To cover most possible isoforms the gRNA design was based on the largest transcript, *PROX1*-205. It codes for a 737 amino acid protein and spans 53,570 bp (8,498 bp of RNA transcript)

a



b

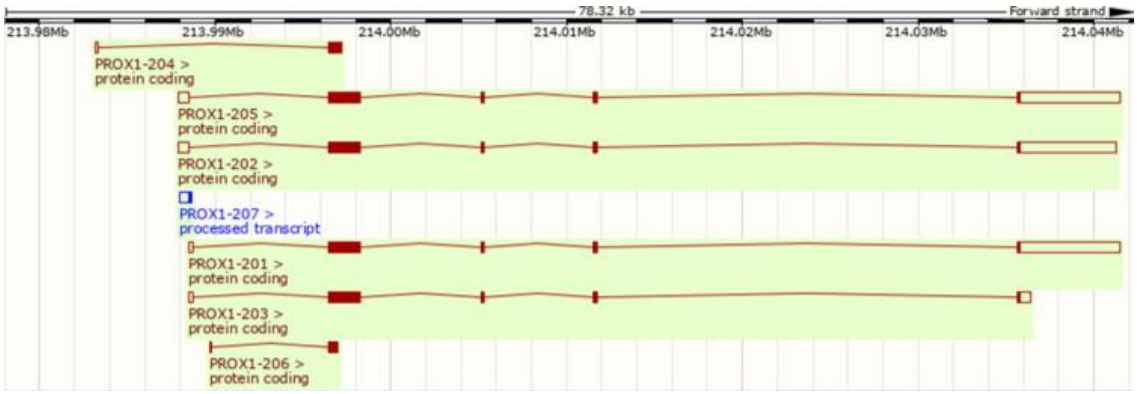


Figure 5.5. *PROX1* expression and isoforms. a) *PROX1* gene expression (in TPM, y-axis) across the stages (x-axis) for the three SB replicates. **b**) I used transcript *PROX1*-205 sequence as reference for promoter and exon annotation. From the hg19 genome assembly.

SNP id	Position	PPAg	ATAC-seq peak	H3K27ac peak
rs79687284	chr1:214150821	0.64	chr1:214150811-214151259 (DE, GT, PF)	chr1: 214150512-214151271 (PF, EP, EN, BLC)
rs17712208	chr1: 214150445	0.36	chr1:214150281-214150738 (DE, GT, PF, PE)	none

Table 5.5. Regulatory elements overlapping *PROX1* T2D credible set SNPs. PPAg: genetic posterior probability of association from the 99% T2D credible sets (from Mahajan and colleagues, 2018). ATAC-seq and H3K27ac peaks: position of the respective peaks (in parenthesis, differentiation stages when the peaks are present).

Figure 5.4 a indicates the location of CRISPR activation and repression gRNAs with respect to *PROX1* enhancer and promoter. For repression 3 gRNAs were designed, targeting the region downstream the promoter of *PROX1* (see **table 5.2**, "T7-Prox1-Rep" 2, 3 and 4). For activation, 3 gRNAs targeted the putative *PROX1* enhancer ("enhancer activation" 1, 2 and 5) and 3 targeted the regions directly upstream the *PROX1* promoter ("promoter SNP" and "promoter activation" 2 and 3). The gRNA "promoter SNP" overlapped a variant with high genetic posterior probability of association in the 99% genetic credible sets (i.e. sets of T2D-associated variants that collectively encompass 99% genetic posterior probability of association with the trait; rs340874 in chr1:214159256, PPAg = 1), independent of the association in this locus of the two SNPs located at the candidate enhancer. **Table 5.6** indicates the genotype at this SNP position and at the two other SNPs of interest from the 99% credible sets, for the three SB cell lines and for the KOLFC2 cell lines. The SB clone used (SBAd3.1) was a heterozygote for the risk allele of rs340874, while it was a homozygote for the non-risk allele at the other two SNPs. The KOLFC2 cell line was homozygous at all variants studied.

The KO gRNAs were designed to guide the Cas9 towards deleting ~1.5 kb of genomic sequence encompassing exon 2, the largest exon in the gene (**Figure 5.4 b**). This would lead to a substantial deletion of the DNA binding domain. Furthermore, the deletion region was chosen to ensure a frame-shift in the event of any aberrant splicing, creating a premature STOP codon and hopefully causing the translation of an inactive protein.

Table 5.6. Genotype of SNPs of interest at *PROX1* locus in two cell types.
Risk allele information from Mahajan and colleagues (2018).

SNP id	Reference allele	Risk allele	SB	KOLFC2
rs79687284	G	C	G/G (all samples)	G/G
rs340874	T	C	T/T (SBAd2.1) | **T/C (SBAd3.1)** | C/C (SBNeo1.1)	**C/C**
rs17712208	T	A	T/T (all samples)	T/T

5.3.3. Differentiating KOLFC2 lines towards beta-like cells

The three iPSC KOLFC2 cell lines (KOLFC2, KOLFC2-KRAB and KOLFC2-SAM) in which *PROX1* expression would be regulated were differentiated across all stages of endocrine pancreas development. These stages were identical to those described in the previous chapters: definitive endoderm (DE), primitive gut tube (GT), posterior foregut (PF), pancreatic endoderm (PE), endocrine progenitors (EP), endocrine-like cells (EN), and beta-like cells (BLC). This was performed to ensure that appropriate differentiation could be achieved at every stage, including insulin-producing BLCs, and to check that *PROX1* was expressed in an analogous way to differentiating SB cells (see **Figure 5.5 a**). Furthermore, the engineered KOLFC2 cells had never been validated for differentiation towards endodermal lineages.

5.3.3.1. Testing the optimal cell plating density.

First, a seeding test was performed to measure the best cell density to plate KOLFC2 iPSC cell lines so that they differentiated efficiently towards the DE stage. This is a critical step in the differentiation process, and empirical evidence from our laboratory (data not shown) indicated that differentiations that do not achieve 80% CXCR4-positive cells by FACS should not proceed to the next stage. A population with at least 90% CXCR4-positive cells is considered ideal.

KOLFC2 and KOLFC2-SAM iPSCs were plated in 12-well plates in triplicate for each of the following amounts: 0.5, 0.75, 1 and 1.25 million cells per well. They were differentiated to the DE stage, and abundance of the CXCR4 protein was measured against the KOLFC2 iPSC baseline. **Figure 5.6** shows that both cell lines achieved a mean of 90% and 91% efficiency respectively when plating 0.5 to 0.75 million cells per well. The efficiency dropped for KOLFC2-SAM at 0.75 million cells plated (80% $CXCR4^{+}$), and even more when plating 1.25 million cells (KOLFC2: 82%, KOLFC2-SAM: 75%). I allowed for ~3% false positives in the iPSC negative control (most likely due to autofluorescence). **Figure 5.7** shows the transcriptomic data for *CXCR4* and other markers across the differentiation stages for reference.

This indicated that the minimum optimal seeding amount for KOLFC2 cell lines in 12-well plates is 0.5 million iPSCs / well, which corresponds to a seeding density of ~1.3x10^5 cells /cm^2. KOLFC2-KRAB was not tested at the time due to limitations in cell availability but did differentiate well to DE after plating at this density, as will be shown later.

5.3.3.2. Differentiating three KOLFC2 cell lines towards BLCs

Two differentiations (KOLFC2 and KOLFC2-KRAB) were performed in parallel, while KOLFC2-SAM differentiation was done at a different time, plating the optimal density of cells as calculated before. Two well replicates were plated per stage and cell line to collect cells for FACS (to measure differentiation markers) and RNA for qPCR (to measure those markers and PROX1 expression). Three extra wells were plated to culture cells in air-liquid interface at the EP, EN and BLC stage (see **section 5.2.4**), which achieved better endocrine cell maturation and insulin secretion in previous studies (Attali et al., 2007; Rezania et al., 2014). FACS analysis at the DE stage (**Figure 5.8**) showed 91%, 88% and 96% CXCR4+ cells for KOLFC2, KOLFC2-KRAB and KOLFC2-KRAB respectively, all above the threshold of a good DE stage efficiency (80%) and thus were differentiated further.

The next stage assessed by FACS was PF, in which the pancreatic progenitor marker *PDX1* (Gu et al., 2002) should be expressed (see **Figure 5.7**). However, KOLFC2 and KOLFC2-KRAB cells did not show any staining for PDX1 (**Figure 5.8**). Technical problems with antibody staining could not be ruled out at that moment due to a lack of PDX1+ cells of similar morphology to those analysed that could be used as positive control. Furthermore, the side scatter (SSC) against forward scatter (FSC) (which give an estimation of the granularity and size of the cell, respectively) indicated few live cells (~26%) that should be present roughly in the diagonal of the plot, indicating round, single cells. Elements to the left of that area suggested abundant debris, although it was difficult to establish the gating strategy because the composition and size of differentiating cells is not well described. Due to the uncertainty surrounding the fidelity of this assay, the expression of the PROX1 PF marker was investigated at the transcript level only on KOLFC2-SAM cells.

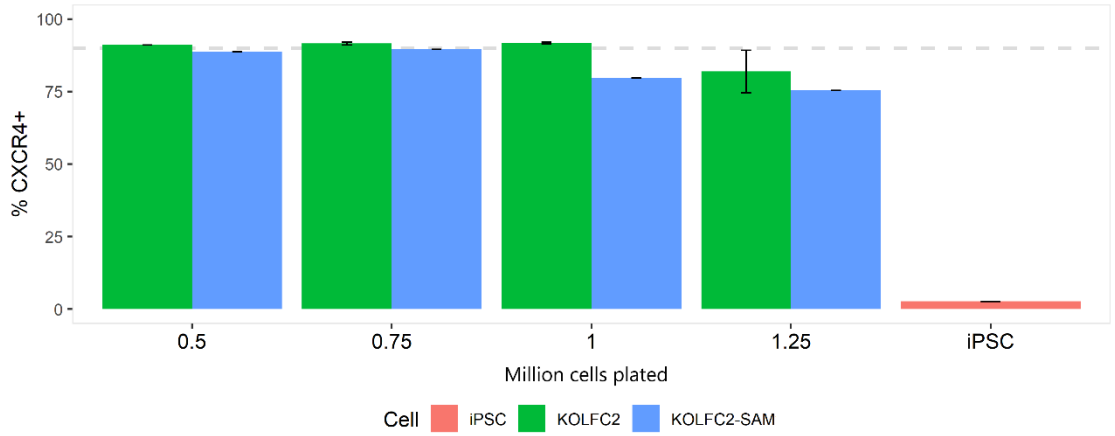


Figure 5.6. Seeding density test FACS at DE stage. Barplots represent mean ± se. of CXCR4^{+} populations (y-axis) per million cells plated (x-axis). iPSC was used as negative control, allowing for ~3% false positives. Grey line: 90% efficiency threshold of an ideal differentiation.

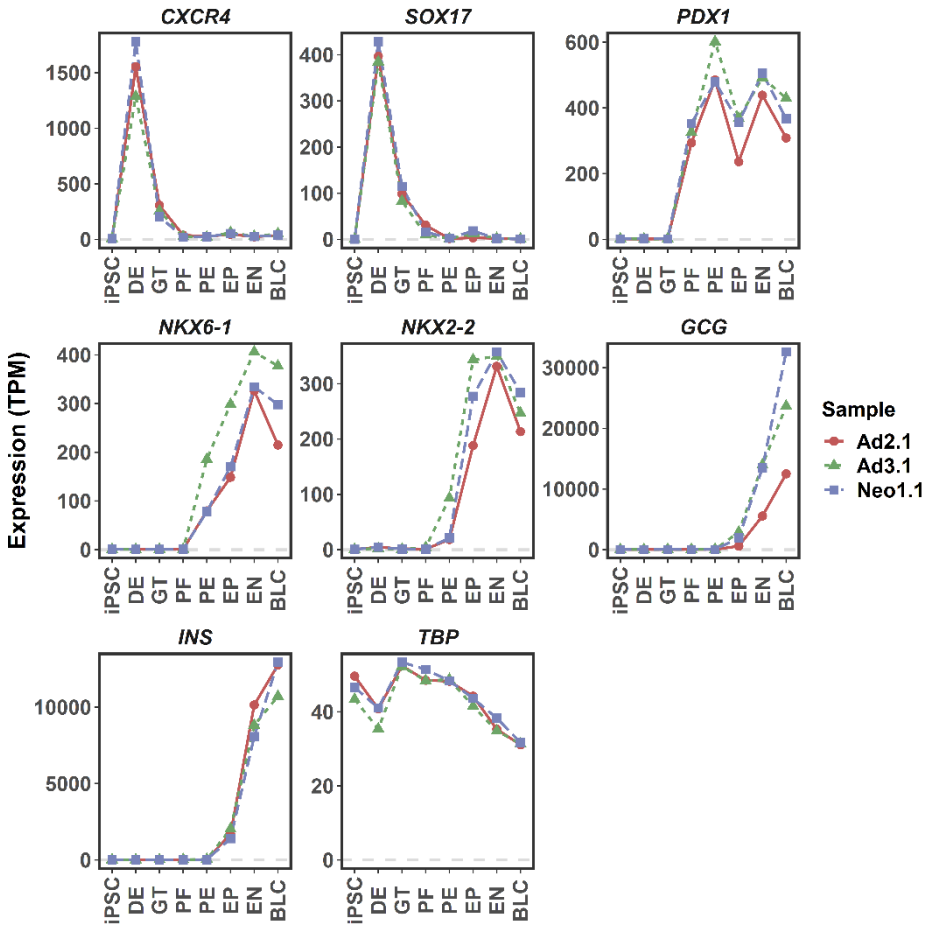


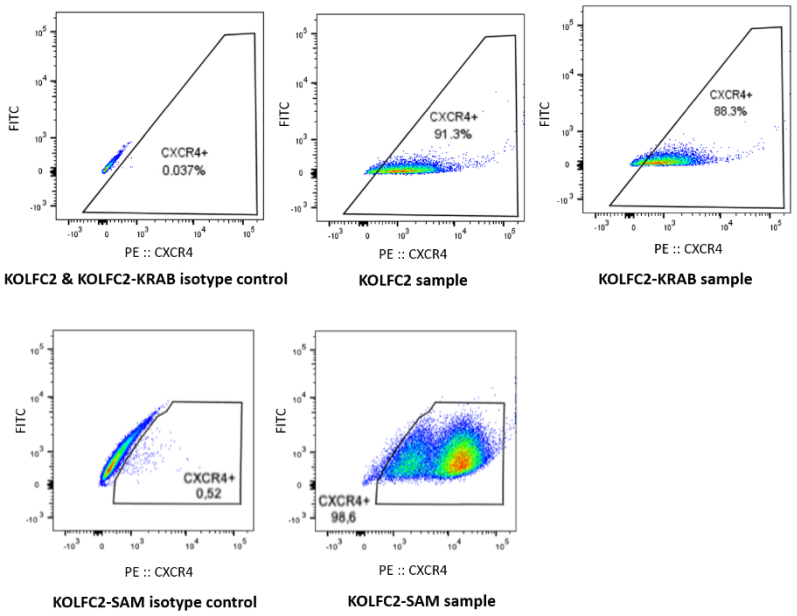
Figure 5.7. Expression of differentiation markers and *TBP* housekeeping gene in differentiating SB cell lines. y-axis: transcripts per million. x-axis: differentiation stages.

Figure 5.9 shows the expression of marker genes per stage measured by qPCR. Some stages were not measured due to the low quality or concentration of RNA for reverse-transcription, particularly at late stages where cell death was abundant. This analysis did provide evidence of *PDX1* expression from the PF stage onwards for two of the three cell lines (**Figure 5.9**), confirming that the previous FACS analysis was technically flawed. The expression of the pancreatic endocrine regulator *NKX6-1* (Sander *et al.*, 2000) was also confirmed. The regulator of endocrine programme and beta cell identity *NKX2-2* (Papizan *et al.*, 2011; Gutierrez *et al.*, 2017) was absent in KOLFC2-SAM and KOLFC2-KRAB cells but was highly expressed in KOLFC2 from EP onwards. It was not possible to gather enough high-quality RNA from the most adult BLC stage to see if these changes in the expression of important transcription factors altered the efficiency of generation of insulin-producing cells. Based on insulin expression at later stages, KOLFC2-KRAB presented better differentiation towards BLC, and KOLFC2-SAM appeared to have a cell population more akin to alpha cells based on the expression of *GCG* (Jennings *et al.*, 2015).

Finally, my candidate gene *PROX1* appeared to be expressed mostly from PF onwards on KOLFC2-KRAB and KOLFC2-SAM cells, mimicking the SB RNA-seq data, although comparison with *TBP* showed very low levels of expression (see **Figure 5.5 1**, **Figure 5.7 and Figure 5.9**). This overall low expression was also detected for other markers and for *TBP* itself, presenting CT values 4-5 cycles higher than those usually observed in the laboratory for SB cells, potentially highlighting low input RNA in the assay. Judging by the expression of several markers, it also seemed like the KOLFC2 cell line failed to differentiate properly. Despite this, the patterns of *PROX1* expression across the stages in the other two cell lines were adequate and I proceeded to use them to modulate *PROX1* expression across beta cell differentiation.

In conclusion, despite the technical problems and the inherent high variability of the differentiation process, the patterns of expression of marker genes at each stage of beta cell development suggest that the KOLFC2 cell lines can be differentiated towards the islet lineage.

DE stage



PF stage

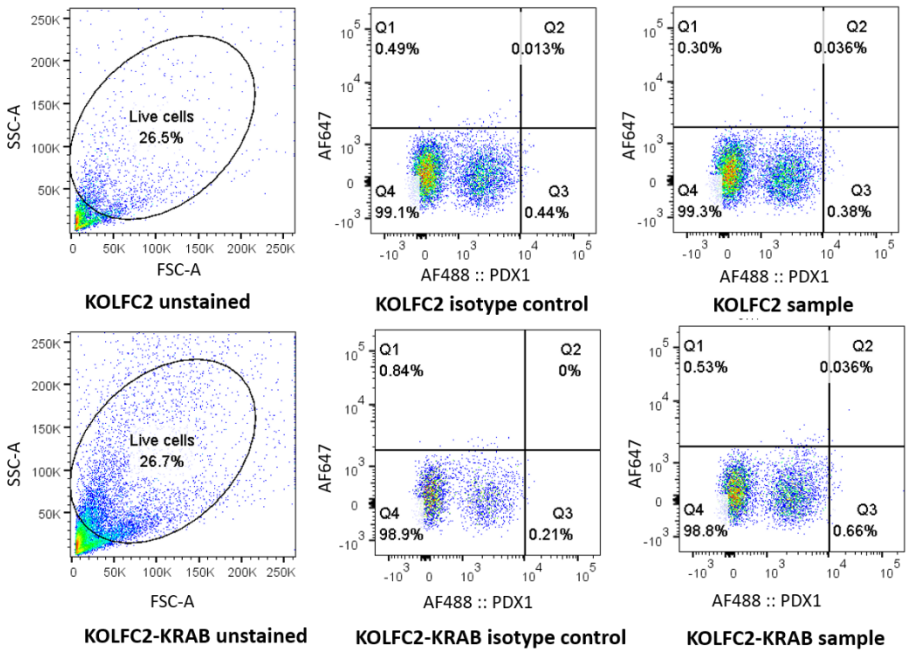


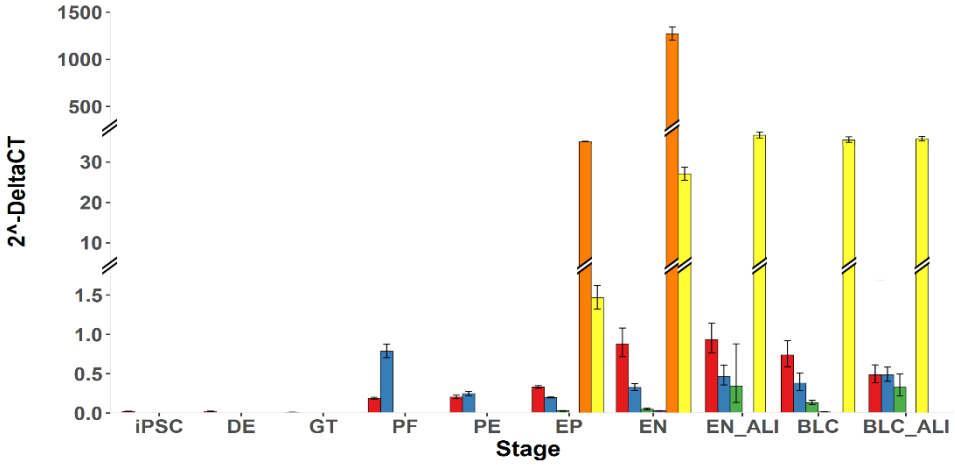
Figure 5.8. FACS staining of differentiating KOLFC2 cell lines. KOLFC2, KOLFC2-KRAB and KOLFC2-SAM were stained for markers of the DE stage (anti-CXCR4 coupled to PE fluorophore) and the PF stage (anti-PDX1 coupled to AF488). Isotype controls shown for reference. Unstained controls show abnormal debris at PF stage.

5.3.4. CRISPR activation of *PROX1* enhancer and promoter

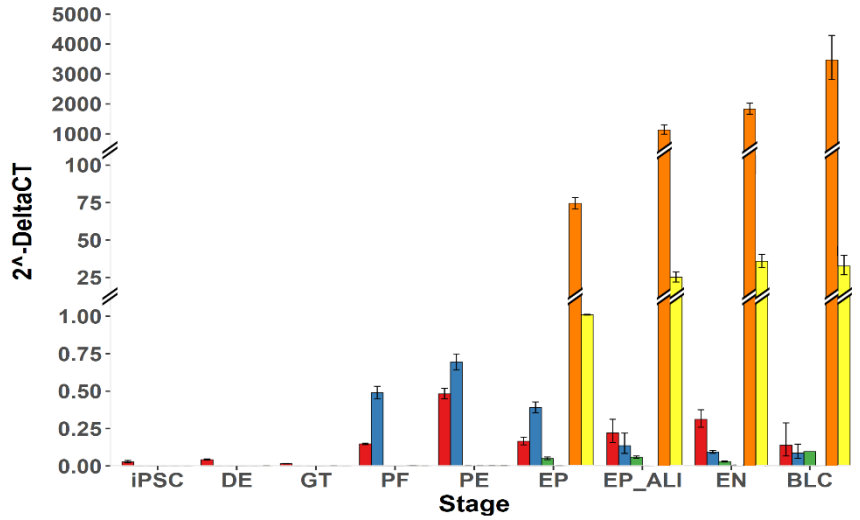
To select the best gRNA for future use during beta cell differentiation, the six activating gRNAs were tested on SB iPSCs (3 targeting the promoter and 3 targeting the candidate enhancer, see **section 5.3.2**). Because these cells do not have the activation cofactors and dCas9 integrated in the cell, the gRNAs were delivered on vectors ("SAM"), driven by the U6 promoter, together with the vectors encoding the activation machinery in the transfections (see **Methods**). This system was used for testing simply because, at the time of designing the sgRNAs, the KOLFC2-SAM iPS cell line was not yet available. After confirming the correct cloning of each gRNAs into the SAM vector, each of the six vectors were transfected in duplicate wells into iPSC SB cells (**Figure 5.10)**, including the SAM without any gRNA integrated ("empty SAM") to use as baseline, and RNA was collected after 48 h.

qPCR revealed a mean relative expression increase for each of the two plating replicates of the "promoter activation 3" gRNA compared to the first empty SAM replicate (and normalized to *TBP*, see **Methods**) of 43 fold (standard deviation [sd] of the qPCR technical triplicates = 33-56) and 25 fold (sd = 23-28). Assuming that the means of ΔCT (that is, cycle threshold of *PROX1* minus that of *TBP*) are normally distributed, the p-value of the t-test for difference in means between the replicates for this gRNA and the baseline was 0.04. From the significance and the consistent directionality of the results, I concluded that "promoter activation 3" gRNA was the best candidate to activate *PROX1* in differentiating KOLFC2-SAM cells. These experiments could not be performed due to time constraints at the end of my experimental work.

KOLFC2-KRAB



KOLFC2-SAM



KOLFC2

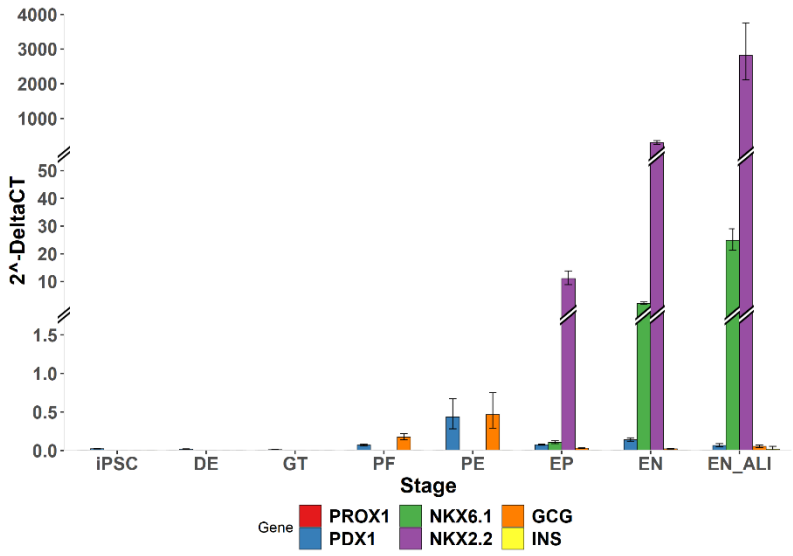


Figure 5.9. Quantitative PCR results for differentiation markers in KOLFC2, KOLFC2-SAM and ***(Continues on next page)***

KOLFC2-KRAB cell lines. y-axis: relative expression of each gene compared to the *TBP* housekeeping control, per differentiation stage. Error bars: $2^{-\Delta CT \pm sd}$ (see **Methods**). ALI: Air-liquid interface. KOLFC2-SAM did not have ALI data on the EN or BLC stage. KOLFC2-KRAB did not have ALI data on the EP stage. KOLFC2 did not have ALI data on the BLC or EP stage, or non-ALI data on the BLC stage.

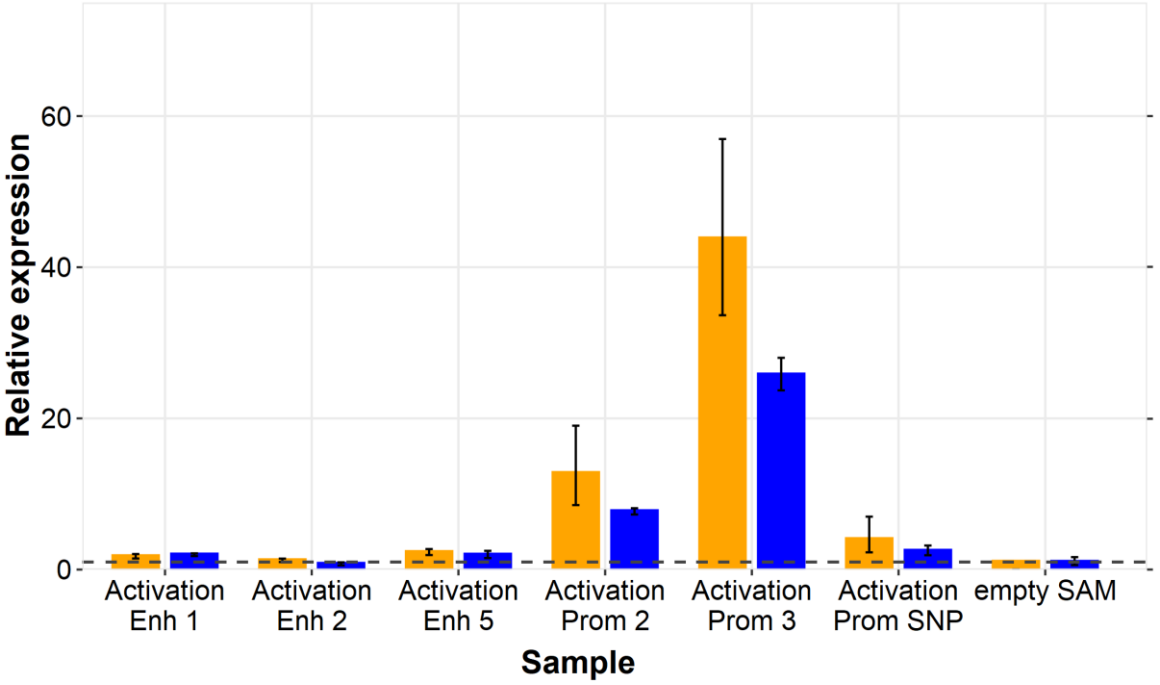


Figure 5.10. qPCR results of activation gRNAs in SB iPSCs. y-axis: relative expression compared to the first transfection replicate of the SAM vector without gRNA ("empty SAM"). Error bars: $2^{-\Delta\Delta CT \pm sd}$ (see **Methods**). x-axis: transfected gRNAs in duplicates. Two plating replicates shown (in orange and blue).

5.3.6. CRISPR repression of *PROX1*

The test of repression using gRNAs for *PROX1* required transfection in differentiating cells, due to the low expression of this gene in undifferentiated iPSCs (see **Figure 5.3 B**). Three candidate gRNAs were transfected into KOLFC2-KRAB cells directly, without the need of vectors, which would theoretically increase the efficiency of the repressive system.

Each gRNA was transfected in duplicate wells per plate in cells differentiating from GT towards PF stage (see **Methods**), and in two independent plates ("replicate 1 and 2"). There were in total four wells per repression gRNA. A "scrambled" gRNA that did not target any sequence of the human genome was used as negative control and baseline for comparisons, while two wells were left untransfected to measure differentiation in completely untreated cells.

As before, FACS analysis of the DE stage indicated good differentiation at this point (92 and 96% $CXCR4^+$ cells in replicate 1 and 2 respectively; **Figure 5.11**). At the end of the PF stage I measured the efficiency by staining for PDX1. To avoid possible technical problems with the anti-PDX1 antibody (see **Figure 5.8**), I tested a new antibody coupled to a different fluorophore, and also stained cells for PROX1.

Replicate 1 did not show any $PDX1^+$ cells (data not shown), while replicate 2 did (**Figure 5.12**). In that experiment, live cells were abundant as measured by FSC and SSC. From those, the subset of cells in the diagonal of FSC-A and FSC-H were considered single cells (not doublets), which were evaluated for PDX1 and PROX1 staining, using the GT cells as negative control for both proteins.

There was a clear shift to $PDX1^+$ populations in all PF samples for this replicate, although it is clearer in the untreated sample ("PF_nogRNA"), which indicates some successfully differentiated cells (ranging from approximately 14% to 6%). Unfortunately, untreated PF cells did not show any PROX1 staining compared to GT cells (**Figure 5.12**, blue square), indicating that these differentiations could not be used to assay PROX1 repression. However, it was noted that $PROX1^+$ cells were readily detected in KOLFC2-KRAB cells

transfected with gRNA (**Figure 5.12**, red squares). In these, PDX1^{+} and PROX1^{+} cells constitute two different populations: PROX1^{+} cells that are PDX1^{-}, and PDX1+ cells that are PROX1^{-}. This increase in PROX1 expression was also apparent for the KOLFC2-KRAB cells transfected with scrambled gRNA. Thus, treating cells with lipofection reagent seemed to alter the cellular state in favour of PROX1 expression.

At the same time, lipofectamine-treated samples have lower PDX1^{+} cells, indicating less efficient differentiation. The issues surrounding differentiation efficiency in the case of lipofection would clearly confound any conclusions which could be drawn concerning the functionality of the CRISPRi system.

Nonetheless, the samples were analysed at the RNA levels by qPCR (**Figure 5.13**). In agreement with the FACS analysis, replicate 1 could not be relied upon; although the DE stage marker *SOX17* showed the expected stage-dependent reduction (**Figure 5.13 a**), there was no discernible increase in either *PDX1* or *PROX1*. Replicate 2, in contrast, showed induction of *PDX1* and *PROX1* in the PF stage.

Assessing *PROX1* levels after CRISPRi, the qPCR results suggested no obvious repression in the samples receiving gRNAs (relative to scrambled, **Figure 5.13 b**). However, untreated cells ("PF no gRNA") showed a marked increase in *PROX1* levels. The same observations were made for *PDX1*. Overall the results suggest that the differences of target gene seen in both FACS and at the transcript level were far too variable to draw any conclusions regarding the efficiency of these differentiations and of *PROX1* repression.

5.3.4. Generating a KOLFC2 *PROX1* knock-out

The first step to generate a PROX1 KO in KOLFC2 cells was to test the gRNAs (see Methods and section 5.3.3). All possible combinations of 5' and 3' gRNAs and a positive control for cutting (at the FLNC gene) were electroporated into the iPSC cells, and after plating and recovery for two days DNA was extracted. The predicted deleted region of exon 2, and the regions around the cutting sites at each end were amplified (see **Figure 5.5 B** for gRNA and primer binding).

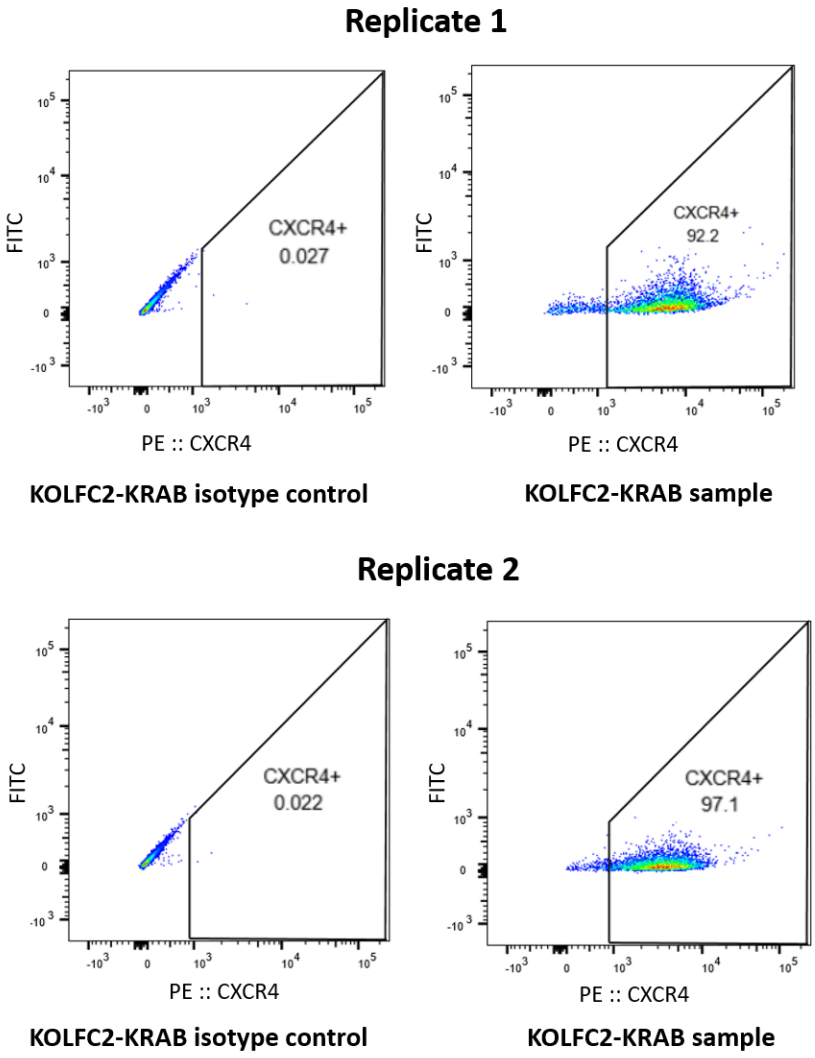


Figure 5.11. PROX1 repression in PF stage: DE differentiation. Cells successfully differentiate towards DE stage in the two biological replicates.

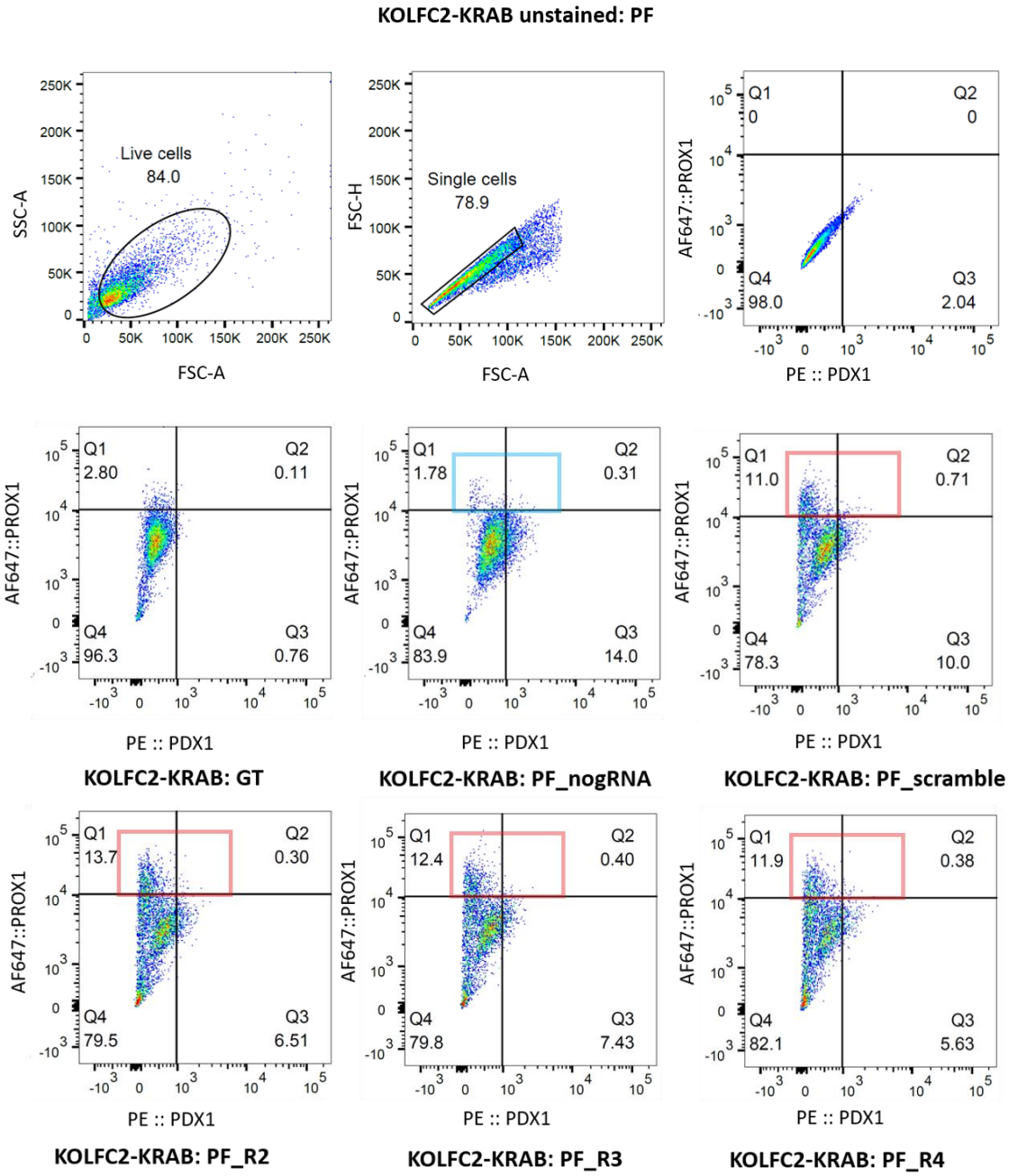


Figure 5.12. *PROX1* repression at the PF stage: FACS of PROX1 and PDX1. GT cells were used as negative staining for gating. Only replicate two showed PDX1 and PROX1 staining, and only results from one well are shown.

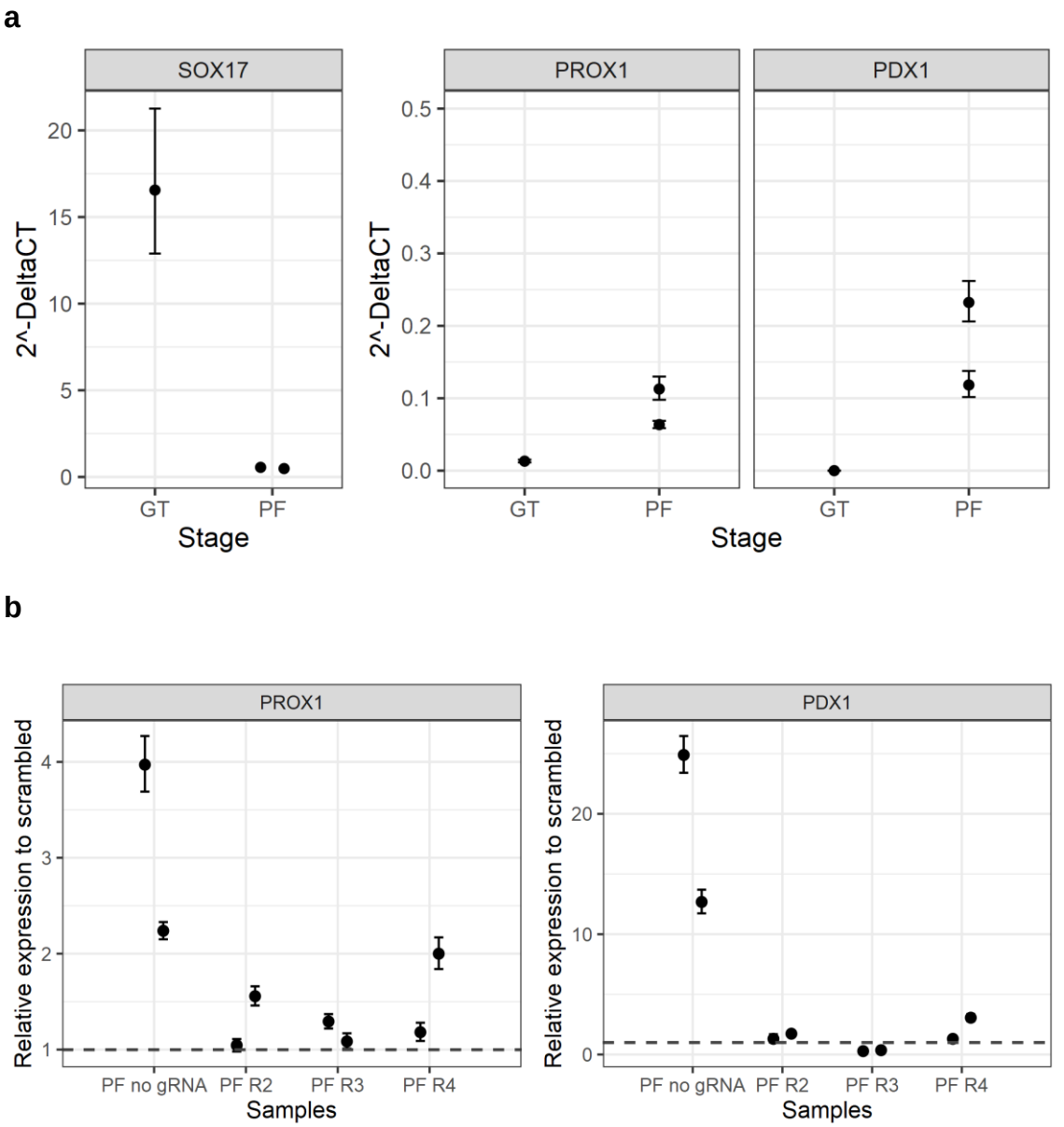
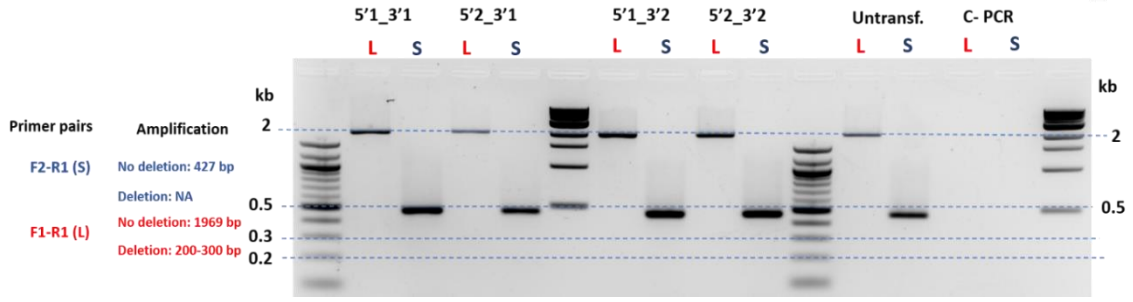


Figure 5.13. *PROX1* repression at the PF stage: qPCR of *PROX1* and *PDX1*. Only biological replicate two shown (replicate one had no expression of *PROX1* or *PDX1*). Two plating replicates represented (except for the GT stage, that had one replicate only). **a)** Differentiation assessment at the DE, GT and PF stages (represented by untreated PF sample) by *SOX17*, *PDX1* and *PROX1* expression. y-axis: relative expression compared to the *TBP* housekeeping control. **b)** Comparison of PF samples (x-axis) against the scrambled baseline (see **Methods**), normalized to *TBP* (y-axis). Error bars: $2^{-\Delta CT \pm sd}$ (**a**) or $2^{-\Delta\Delta CT \pm sd}$ (**b**).

Figure 5.14 a shows the results for the 4 combinations, and the uncut exon 2 for reference. As shown in the amplification size guide, no exon 2 deletion was observed for any pair of gRNAs (which should give a 200-300 bp product after repair of the two double strand breaks and the resulting deletion of the intervening sequence). As no *cis* deletion event could be detected, the activity of the individual gRNAs was tested by amplifying the corresponding region and assessing indel frequency by Sanger sequence and TIDE analysis. **Figure 5.4 b** shows the position of the primers used for this analysis. Indel assessment (see **Methods**) indicated cutting efficiencies of 5% or under for each *PROX1* gRNA tested. However, the electroporation was successful as an indel frequency of 33% could be detected using the *FLNC* gRNA control (**Figure 5.14 b**). The observed low frequency of mutagenesis is consistent with the inability to detect the desired *cis* deletion event. It was concluded that the KOLFC2 *PROX1* KO cell line could therefore not be generated with the designed gRNAs, and time constraints prevented further experimentation.

a



b

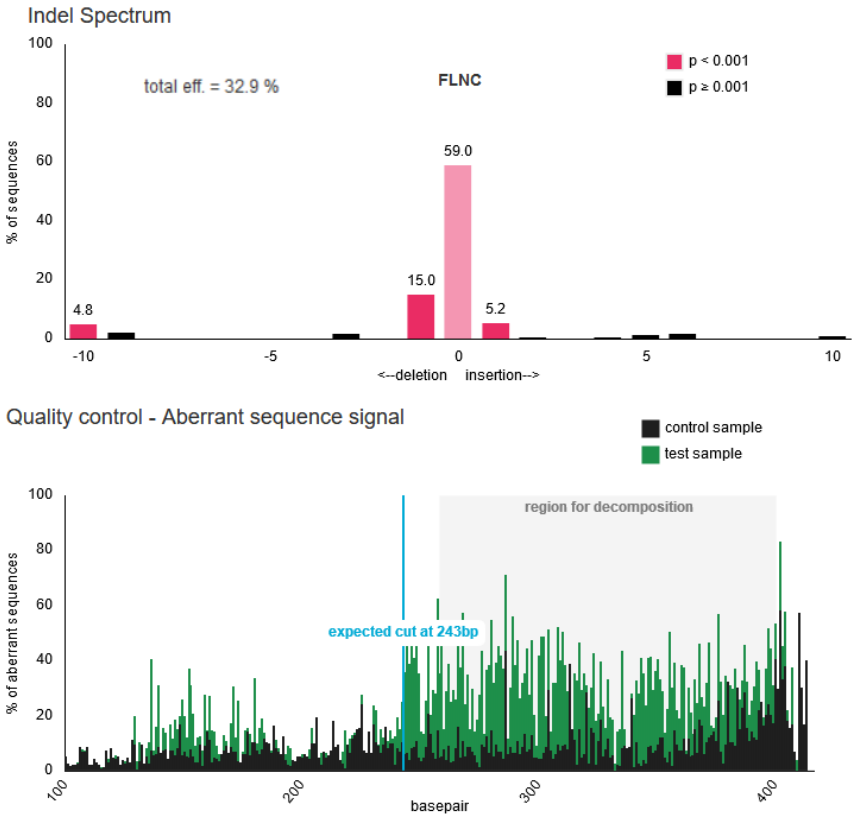


Figure 5.14. Generating a knock-out of *PROX1* in KOLFC2 cells. a) DNA gel of amplification products of the region of deletion for every gRNA combination and the negative control. Dotted blue lines show fragments of equal size. For reference of where the primer pairs anneal, see **Figure 5.4 b**. b) *FLNC* positive control of cutting: indel spectrum (above) and chromatogram of cut sample versus control (below).

5.4. Discussion

The main results of this chapter are:

- *PROX1* is expressed in the KOLFC2-SAM and KOLFC2-KRAB cell lines in the same islet development stages than in the SB cell lines. No conclusion could be drawn about the expression of the gene in KOLFC2 cells.
- *PROX1* activation was successful in iPSC SB cells, using gRNAs targeting the promoter of the gene.
- It was not possible to assess *PROX1* repression in differentiating KOLFC2-KRAB cells.
- *PROX1* knock-out in KOLFC2 cells was not achieved due to the inability to detect any cutting activity for the selected gRNAs.

In the previous chapter I argued how integration of transcriptomic and DNA regulatory marks, and murine studies on islet development (Wang *et al.*, 2005; Westmoreland *et al.*, 2012), point to *PROX1* as a candidate effector gene within a T2D-associated locus possibly involved in the differentiation of the pancreas. I have shown how the *PROX1* candidate gene is expressed from the posterior foregut (PF) to beta-like cells (BLC), both in the SB cell lines (as demonstrated by the RNA-seq data) and in the differentiating cells derived from KOLFC2 iPSCs. KOLFC2-SAM and KOLFC2-KRAB expressed insulin and glucagon in the two most adult stages (EN and BLC), which shows that they can be differentiated at least to a multihormonal state. Meanwhile, KOLFC2 presented a different profile of gene expression markers, but without a biological replicate a bad differentiation (instead of general absence of *PROX1* expression) could not be ruled out. It would have been interesting to measure *NEUROG3* to see if differences in expression of this gene could explain the absence of insulin.

Nevertheless, these results demonstrate that KOLFC2 cell lines can be used as models to facilitate the study of *PROX1* in islet development. The test of the CRISPRa gRNAs targeting the promoter was successful, demonstrating that *PROX1* expression can be activated in the SB cell line. The next steps would be to transfect the most effective gRNA

("Promoter activation 3") into KOLFC2-SAM cells during differentiation towards the PF stage, and assess if this produces any effect on the efficiency of differentiation. Conversely, the activation gRNAs targeting the putative *PROX1* enhancer did not alter significantly the expression of this gene, which could be due to the enhancer being inactive in iPSCs. Delivery of the enhancer activation gRNAs into KOLFC2-SAM cells during differentiation may prove more effective.

Results of the effect of *PROX1* repression at the PF stage on endocrine pancreas development were inconclusive, probably due to several factors. Firstly, technical problems with immunofluorescence staining and qPCR made it difficult to reliably detect PROX1 at the protein or gene expression level. Secondly, toxicity of transfection reagents itself altered gene expression. Thirdly, efficiency of hiPSC differentiation towards beta cells was highly variable, even between wells of the same batch, which masked small phenotypic effects of induced gene regulation. Lastly, there was no positive control showing the repression of a control gene at the PF stage, which would have provided a control for successful delivery of the gRNA to this differentiated cell type by lipofection. Sadly, such control did not exist at that moment. It is likely that the gRNA delivery system for differentiating cells needs to be extensively optimized before conclusive results can be achieved. It is well recognized that certain cell types are very difficult to transfect with lipofection reagents. Subsequently, alternative methods of sgRNA delivery such as electroporation or viral delivery could be explored.

It was difficult to troubleshoot FACS, as there are no positive controls available that match the protein expression and morphology of the differentiating cells. Attempts to use immortalized adult beta cells (EndoC-βH1, Ravassard *et al.*, 2011) failed due to the sensitivity of the cells to permeabilization. Also, cells in the PF stage present a "sticky" morphology and are resistant to resuspension into single cells. FACS forward and side scatter indicate an abundance of clumps of cells or other debris, which can mean that a huge part of the $PDX1^+$ and/or $PROX1^+$ cells were filtered out or otherwise lost during staining. Optimization is needed to improve FACS of transcription factors so I do not observe false positives or negatives of PROX1 and PDX1 staining in the future.

Expression of *PROX1* and of other genes including the housekeeping *TBP* measured by qPCR was abnormally low, based on comparisons of CT values obtained by other members of the laboratory when amplifying the same amount of cDNA from other cell lines. Because cDNA concentration cannot be measured directly, I could not pinpoint the problem to cDNA synthesis. The RNA quality measured by 260/230 and 260/280 was commonly over the thresholds of purity (>2 in both cases). Therefore, it is possible that KOLFC2 cell lines express several genes, including the housekeeping gene *TBP*, at a reduced level, however a more likely explanation is a systematic problem with either cDNA synthesis or in the qPCR.

FACS and qPCR indicated that treatment with the transfection reagent (PF cells transfected with either the *PROX1*-repressing or scramble gRNAs) could be triggering pathways that increase *PROX1* expression, surpassing the effect of the repressor gRNA. In this context it is unlikely that any effects of *PROX1* repression would be visible against this apparent lipofection-induced overexpression of the gene. On the other hand, *PROX1* activation in iPSC SB cells could be distinguished from the effect of the transfection without gRNAs, even with the limited replicates and subsequent inference limitations (replicate wells in same plate are not really independent). This could be due to the transcriptomic landscape of the iPSCs or to differences in the effects of the transfection reagents used on the target cell type (see **Methods**). These problems are exacerbated by the variability in the differentiation process itself, highlighted by the variable *PDX1* and *PROX1* expression in the two replicates of KOLFC2-KRAB differentiations (**Figure 5.13 a**). This variability has been observed consistently by other members of the laboratory, and is due to unknown causes that might include deficient handling, preparation and storage of differentiation media, genetic heterogeneity of iPSC cell populations due to defective maintenance, and selection pressures during differentiation.

In the future, I would like to keep exploring the role of downregulating the expression of this candidate gene in beta cell differentiation, attempting again the KO cell line with gRNAs stabilised and protected against exonucleases (Basila *et al.*, 2017), that other members of the laboratory have successfully used to produce knock-outs of other genes.

This is more likely to produce a phenotypic effect that can be measured among the expression noise of the differentiation model.

6. General discussion

The work presented in this thesis led to the following conclusions:

- PCA, pathway analysis and TFs regulating the transcriptome and the DNA regulatory marks analysed indicated that the hiPSC-derived differentiation model from which the data originated recapitulates key aspects of beta cell development *in vivo*, and achieved more adult beta cells (though still immature) than previously-published models.
- Analysis of differentially expressed genes and regulatory marks within each stage indicated that both the most adult stage (beta-like cells) and less differentiated stages were enriched in T2D-associated signals, pointing to the contribution of both developing and adult islets to T2D risk.
- Novel potential regulators of islet development were identified (e.g. *REST* and *CUX2*), including genes within loci associated with T2D by GWAS (e.g. *PROX1*, *ONECUT1/HNF6*) and TFs whose binding was disrupted by T2D-associated variants (e.g. CTCF, ZBTB33)
- Integration of DNA regulatory marks with transcriptomic information facilitated the identification of active enhancers in islet development and their possible target genes, including candidate gene effectors of T2D-associated variants (e.g. *PROX1*)
- It was possible to replicate the differentiation model in different cell lines than those that originated the *omics* data, providing a tool to study the effect of candidate genes in islet development.
- CRISPRa technique to upregulate *PROX1* expression was functionally validated in the undifferentiated hiPSC system.

This thesis presents a multi-*omics* analysis of an hiPSC-derived islet development model, capitalizing and expanding on the recent advances in protocols to generate functional beta cells *in vitro* (D'Amour *et al.*, 2005; D'Amour *et al.*, 2006; Bruin *et al.*, 2014; Pagliuca *et al.*, 2014; Rezania *et al.*, 2014; Xie *et al.*, 2013; Russ *et al.*, 2015; Nostro *et al.*, 2015; van de Bunt *et al.*, 2016; Perez-Alcantara *et al.*, 2018, Nair *et al.*,

2019). The transcriptomic evaluation of the model in **Chapter 2** confirmed that it recapitulates the developmental steps observed *in vivo*, measured by the adequate expression at pertinent stages of markers of the endocrine pancreas previously characterized in mice and humans (e.g. *PDX1* was expressed from the dorsal pancreatic bud onwards, approximated to the posterior foregut [PF] stage in our model) (Jennings *et al.*, 2015).

These markers are often transcription factors that play essential roles in the networks of gene expression that maintain cellular identity or trigger differentiation. Indeed, **Chapter 3** demonstrated the enrichment of the binding motifs of several of them within regulatory regions defined by accessible chromatin, DNA hypomethylation or the presence of H3K27ac histone marks at the pertinent stages of development, including binding sites for several transcription factors with known pioneer roles in opening and activating closed chromatin during pancreas development, such as PDX1 and members of the FOXA and RFX family (Wang *et al.*, 2015; Atlasi and Stunnenberg, 2017). Gene ontology and pathway enrichment analyses, and the expression and presence of chromatin marks near monogenic diabetes genes (which often affect pancreas development [Schwitzgebel, 2014]), also confirmed the likely endocrine fate of these cells. Furthermore, principal component analysis of the transcriptome and methylome showed that beta-like cells (BLC) were more similar to beta cells matured *in* vivo and adult islets respectively than the best endpoint of differentiation that could be achieved only three years ago (the EN stage) (van de Bunt *et al.*, 2016; Xie *et al.*, 2013). Therefore, both transcriptomic information and DNA regulatory marks gave confidence that these differentiating hiPSCs can be used as model system to study islet development and how it could contribute to T2D risk, bypassing the need to use mice models which, as indicated before, do not completely recapitulate human pancreas development (Steiner *et al.*, 2011; Benner *et al.*, 2014; Jennings *et al.*, 2013).

All models are simplifications of reality and thus imperfect, and the limitations of the one used in this thesis were highlighted by the *omics* datasets. For example, markers of islet

subpopulations other than beta cells were expressed in the most adult differentiation stage (BLC), including glucagon (*GCG*, from alpha cells) and somatostatin (*SST*, from delta cells) (Pagliuca *et al.*, 2013; Ackermann *et al.*, 2016). This indicates that at each stage of differentiation there were subpopulations of cells with different genetic profiles, which in turn affect the development towards the following stages through mechanisms that might include the lack of adequate enhancer priming for the beta cell lineage and the absence of correct spatial cues (Wang *et al.*, 2015; Larsen *et al.*,2017). Single cell analysis is already being used to overcome these limitations and investigate what causes the different subpopulations to commit to one lineage or another (Borup *et al.*, 2017).

Since disease-associated GWAS variants are often found in intergenic regions with presumed regulatory activity, it follows that studying them in islet differentiation models could be useful for understanding the aetiology of T2D, especially from a developmental perspective. Enhancers present in our differentiation model and defined by accessible chromatin and H3K27ac peaks distal to transcription start sites (Thurman *et al.*, 2012; Tsompana and Buck, 2014), were mostly exclusive to each differentiation stage or shared only between neighbouring stages. This agrees with previous evidence about the specificity of regulatory annotations in development and across tissues or cell types (Parker *et al.*, 2013; Pasquali *et al.*, 2014; Wang *et al.*, 2015; Roadmap Epigenomics Consortium, 2015). **Chapter 3** and **4** demonstrated that islet enhancers are enriched in T2D-associated variants, as has been consistently observed before (Parker *et al.*, 2013; Pasquali *et al.*, 2014; Gaulton *et al.*, 2015; Thurner *et al.*, 2018). Significantly, not only late stages of islet development were associated with T2D, as would be expected from the more adult gene expression profile of BLCs (demonstrated by the expression of *MAFA*, *INS* and other genes encoding elements of the insulin processing and secretory pathway), but in intermediate stages of pancreas development as well. This adds to the existing evidence that part of the contribution to T2D risk comes from dysregulation of islet development, provided mostly by the overlap between monogenic diabetes genes and T2D-associated loci.

It is obvious that the ability to interrogate the contribution of DNA regulatory activity to T2D risk depends on two factors: the quality and quantity of the regulatory region annotations employed in the analysis, and how many variants have been significantly associated with the disease. The former is highlighted by the limitations of the EPIC DNA methylation array, that unlike whole genome bisulfite sequencing is only able to examine around 850k positions in the genome. A recent quantification using an older 99% credible set estimate of 96 T2D-associated loci from DIAGRAM estimated that the median posterior probability of association covered per locus was just 39% (Thurner *et al.*, 2018). Therefore, the DNA methylation information would have to be assessed using whole genome bisulfite data to be comparable with the ATAC-seq and H3K27ac ChIP-seq.

The number of variants associated with T2D, the strength of associations and how precisely they have been fine mapped in each locus are also important variables to understand the results and limitations of the enrichment analyses. For example, in the study published in 2018 on the RNA-seq data of this differentiation model (Perez-Alcantara *et al.*, 2018) only 96 loci associated with the disease (Scott *et al.*, 2017) were used in the analysis, and presented different credible set boundaries to the 243 loci used in the updated analysis presented in **Chapter 2.** As a result, in the re-analysis more genes were naively associated (by proximity to the variants) with the disease, changing the p-values assigned to each one by MAGENTA and altering the significance and number of enriched stages (e.g. the significantly enriched stage by hypergeometric test changed from BLC in Perez-Alcantara *et al.* [2018] to EN). It is also important to keep in mind that in the differential expression analysis from **Chapter 2** the genes were assigned exclusively to the stage in which their expression peaked, complicating further the interpretation of the results. Despite all these limitations, the transcriptomic analysis of the model also pointed to the role of developing and adult islets in T2D.

Enhancers can bypass the nearest gene or act on multiple genes (Pennacchio *et al.*, 2013; Javierre *et al.*, 2016), and therefore a naïve assignation of effector genes to the closest promoter is bound to cause false positives. This has important implications for

the study of how noncoding variants affect disease risk within these regulatory regions. Recent advances in the study of the 3D genome allow the examination of physical interactions between promoters and distal regulatory regions using approaches such as Promoter Capture HiC (Schoenfelder *et al.*, 2015), which has been applied to study GWAS signals in adult islets (Miguel-Escalada *et al.*, 2019; Greenwald *et al.*, 2019). Single cell analysis also allows maps of co-accessible chromatin to be built, reasoning that these locations are likely to be functionally related in the same way that is assumed for co-expressed genes (Chiou *et al.*, 2019). In fact, recent analysis of histone regulatory marks found regions of correlated regulatory activity within the boundaries of contacting chromatin domains (Delaneau *et al.*, 2019).

Lacking chromatin conformation data from the hiPSC-derived differentiating cells the Activity By Contact (ABC) model was used to predict enhancer activity on target genes (Fulco *et al.*, 2019). The data shown in **Chapter 4** is promising because it indicates that it could recapitulate known islet biology more comprehensively than enhancers defined by just the overlaps between H3K27ac and ATAC-seq peaks or either of those datasets on their own. I also discovered that most of the GWAS signals overlapping active enhancers were not linked by the ABC model to the gene naming the locus. This highlights the importance of characterizing the activity of these enhancer regions experimentally. Limitations of the ABC model include distance restrictions and the fact that is that it was benchmarked using CRISPR interference on enhancers of cell types different than differentiating islets, and it is arguable how much of the assumptions of the model would be transferrable to other cell types (e.g. that the strength of each individual enhancer's activity is linearly dependent on the contact frequency, as opposed to non-linear relationship products of unknown underlying synergies). Recent advances in CRISPR screening and the possibility of sequencing several kinds of genomic data will continue to facilitate the mapping of enhancers to target genes (Fulco *et al.*, 2016; Fulco *et al.*, 2019; Hsu *et al.*, 2018; Sanjana *et al.*, 2016; Mumbach *et al.*, 2017). Accurately predicting enhancer positions also helps with fine-mapping efforts, reducing even further the credible sets by giving the variants more accurate prior probabilities based on the

tissue or cell type-specific enhancer information. A logical next step in these analyses would have been to quantify how the credible sets were reduced after recalculating the posterior probabilities of association.

Chapter 3 and **4** indicated that chromatin architecture remodellers such as CTCF, ZBTB33 (Kaiso) and HINFP (Defossez *et al.*, 2005; Li and Liu 2012) were enriched in the DNA regulatory marks, particularly the enhancers defined by the ABC model, and their binding was consistently disrupted by the T2D associated variants. Disrupting CTCF binding at specific locations deeply alters chromatin-chromatin interactions, modifying the delimitations of topologically associated domains and causing significant changes in gene expression within it (Guo *et al.*, 2015; Lupiáñez *et al.* 2015). This provides a plausible mechanism of how gene variation at a subset of T2D-associated SNPs could be influencing disease risk by altering gene regulation at enhancers during islet development. It would be interesting to confirm in the laboratory if binding of these TFs to some of these variants is altered using ChIP-qPCR, DNase footprinting or electrophoretic mobility shift assays in our islet development model.

The integration of different sets of DNA regulatory marks led me to study the role of the homeobox transcription factor PROX1 in this process. Two variants from the same credible set, together presenting 99% posterior probability of association with T2D, were located in an enhancer predicted by the ABC model to act on the *PROX1* gene at intermediate to late stages of endocrine pancreas development, when it is expressed. Transcription factor binding at the two credible set variants, presented in the **Appendix Tables**, suggest that the regulator of the posterior foregut lineage, HES1 and the transcription factors HNF1A and HNF1B (causal of monogenic diabetes) could be controlling the progression towards the endocrine pancreas lineage at this locus (Koike *et al.*, 2019). This adds onto a body of evidence of *PROX1* expression in adult and developing islets from humans, and its demonstrated roles in regulating pancreatic gene networks in mice (Burke and Oliver, 2002; Wang *et al.*, 2005; Westmoreland *et al.*, 2012; Hvratin *et al.*, 2014; Paul *et al.*, 2016).

I moved forward with preliminary experiments, aimed at addressing the role of *PROX1* in beta cell development using CRISPR activation and interference in differentiating hiPSC cell lines, engineered with the activation or repression CRISPR machinery integrated and ubiquitously expressed from a safe harbour locus (González *et al.*, 2014). This would theoretically improve the efficiency of the modulation and potentially allow experimental manipulation at specific stages, by simple sgRNA delivery. sgRNAs that robustly upregulate the *PROX1* gene were identified and validated in undifferentiated cells. Sadly the likely inefficiency of transfections, combined with the inherent variability of the differentiations, possibly exacerbated by the negative effects of lipofectamine, made this study inconclusive. This study, however did permit a demonstration that different cell lines to the one from which the *omics* data came from can be successfully differentiated towards the most adult beta-like cell stage, as seen from the high expression of *INS* and *GCG* at the EN and BLC stages. In the future, I would like to reattempt to create a *PROX1* knock-out cell line using stabilised gRNAs, protected against exonucleases (Basila *et al.*, 2017). This would reduce problems associated with the sgRNA delivery and produce a stronger phenotype that can hopefully be detected amidst the inherent variability of the differentiation process. In addition, as the CRISPR activation experiment was successful in iPSCs, this may also cause a measurable effect if I were to repeat the experiment in differentiating gut tube cells.

Finally, the catalogue of DNA regulatory data of islet development presented in this thesis will provide a useful resource for the scientific community, particularly for prioritizing the study of T2D-associated loci in islet development, both at the level of variants and candidate gene effectors.

Appendices

Appendix Table 2.1: Significant pathway enrichment terms from genes upregulated in the EN stage from the current protocol versus the previous protocol. Genes significantly upregulated (q<0.01, log2FC > 1) in the EN stage from the current protocol compared to the previous protocol were tested for enrichment in Reactome pathway terms. Only significant (Benjamini-Hochberg adjusted p-values < 0.05) terms are shown. Beta-cell related terms in bold.

Reactome Pathway ID	Description	BH-adjusted p-value	Gene
R-HSA-112316	Neuronal System	2.50E-24	*RASGRF1, CAMK2A, NSF, DLGAP4, NRXN2, LIN7A, RASGRF2, GLS, GABBR2, HCN4, SYT2, KCNMA1, LRFN2, CACNA2D3, GRIK3, HCN1, ADCY1, KCNAB1, KCNK9, KCNA3, KCNA2, KCNB2, UNC13B, GRIN3A, GNAT3, TUBB3, CACNA2D2, MAPK3, GRIK5, STX1A, PRKAG2, PRKAR2A, CHRNB4, KCNJ2, GNAI1, GAD1, KCNC3, STXBP1, TUBB2A, PPFIA2, KCNH1, CACNA1B, LIN7C, DLG2, CACNA2D1, KCNJ3, SLC18A2, GABRB3, KCNH6, PPM1E, KCNJ11, CACNA1E, ABCC8, SYT7, KCNH2, ADCY2, CHRNA3, EPB41L3, KCNH4, TUBB1, CHRNA4, APBA1, KCNC4, DLGAP3, GRIA2, KCNK17, PRKCG, TUBA4A, DLG4, SNAP25, CHRNA1, KCNH5, CACNA1A, KCNN3, HCN3, KCNK13, GRIA4, CHRNB2, ADCY9, PRKCB, GNG4, CHRNA5, KCNG3, SYN3, GNG2, LRRTM3*
R-HSA-112315	Transmission across Chemical Synapses	2.50E-15	*RASGRF1, CAMK2A, NSF, LIN7A, RASGRF2, GLS, GABBR2, CACNA2D3, GRIK3, ADCY1, UNC13B, GRIN3A, GNAT3, TUBB3, CACNA2D2, MAPK3, GRIK5, STX1A, PRKAG2, PRKAR2A, CHRNB4, KCNJ2, GNAI1, GAD1, STXBP1, TUBB2A, PPFIA2, CACNA1B, LIN7C, DLG2, CACNA2D1, KCNJ3, SLC18A2, GABRB3, PPM1E, CACNA1E, ADCY2, CHRNA3, TUBB1, CHRNA4, APBA1, GRIA2, PRKCG, TUBA4A, DLG4, SNAP25, CHRNA1, CACNA1A, GRIA4, CHRNB2, ADCY9, PRKCB, GNG4, CHRNA5, SYN3, GNG2*
R-HSA-112314	Neurotransmitter receptors and postsynaptic signal transmission	6.55E-11	*RASGRF1, CAMK2A, NSF, LIN7A, RASGRF2, GABBR2, GRIK3, ADCY1, GRIN3A, GNAT3, TUBB3, MAPK3, GRIK5, PRKAG2, PRKAR2A, CHRNB4, KCNJ2, GNAI1, TUBB2A, LIN7C, DLG2, KCNJ3, GABRB3, PPM1E, ADCY2, CHRNA3, TUBB1, CHRNA4, APBA1, GRIA2, PRKCG, TUBA4A, DLG4, CHRNA1, GRIA4, CHRNB2, ADCY9, PRKCB, GNG4, CHRNA5, GNG2*
R-HSA-1296071	Potassium Channels	1.68E-09	*GABBR2, HCN4, KCNMA1, HCN1, KCNAB1, KCNK9, KCNA3, KCNA2, KCNB2, KCNJ2, KCNC3, KCNH1, KCNJ3, KCNH6, KCNJ11, ABCC8, KCNH2, KCNH4, KCNC4, KCNK17, KCNH5, KCNN3, HCN3, KCNK13, GNG4, KCNG3, GNG2*
R-HSA-163685	Integration of energy metabolism	1.64E-07	*GNAS, GLP1R, ACLY, ITPR1, CACNA1D, ADCY1, CACNA2D2, MLXIPL, STX1A, PRKAG2, PRKAR2A, GNAI1, STXBP1, KCNJ11, CACNA1E, ABCC8, ADCY2, SYT5, SNAP25, CACNA1A, CACNA1C, ADCY9, GNG4, GNG2, GCGR, INS*

R-HSA-5576891	Cardiac conduction	1.64E-05	*HIPK2, CAMK2A, KAT2B, FGF12, ITPR1, CACNA1D, CACNA2D3, KCNK9, ATP2A2, RYR2, CACNA2D2, ATP2A3, KCNJ2, SCN7A, CACNA2D1, ATP2B2, SCN3B, KCNJ11, SCN8A, KCNH2, FGF14, ATP1A3, KCNK17, CACNA1C, KCNK13, NPPC*
R-HSA-422356	**Regulation of insulin secretion**	1.64E-05	*GNAS, GLP1R, ITPR1, CACNA1D, CACNA2D2, STX1A, PRKAR2A, GNAI1, STXBP1, KCNJ11, CACNA1E, ABCC8, SYT5, SNAP25, CACNA1A, CACNA1C, GNG4, GNG2, INS*
R-HSA-442755	Activation of NMDA receptors and postsynaptic events	1.64E-05	*RASGRF1, CAMK2A, LIN7A, RASGRF2, ADCY1, GRIN3A, TUBB3, MAPK3, PRKAG2, PRKAR2A, TUBB2A, LIN7C, DLG2, PPM1E, TUBB1, APBA1, GRIA2, TUBA4A, DLG4, GRIA4*
R-HSA-212676	Dopamine Neurotransmitter Release Cycle	3.97E-05	*LIN7A, UNC13B, STX1A, STXBP1, PPFIA2, LIN7C, SLC18A2, APBA1, SNAP25, SYN3*
R-HSA-418597	G alpha (z) signalling events	5.03E-05	*GNAS, GNAZ, ADCY1, GNAT3, GNAI1, PRKCH, ADCY2, RGS17, PRKCG, ADCY9, PRKCB, GNG4, PRKCE, GNG2*
R-HSA-418555	G alpha (s) signalling events	8.10E-05	*REEP1, GNAS, GLP1R, GNAZ, PDE3B, GIP, ADCY1, REEP3, GNAT3, PTHLH, PTGER2, GNAI1, GIPR, ADCY2, PDE4C, IAPP, REEP2, ADCY9, TSHR, GNG4, GPBAR1, GNG2, PDE2A, GCGR*
R-HSA-6794361	Neurexins and neuroligins	2.28E-04	*DLGAP4, NRXN2, LIN7A, SYT2, STX1A, STXBP1, LIN7C, DLG2, SYT7, EPB41L3, APBA1, DLGAP3, DLG4, LRRTM3*
R-HSA-9609736	Assembly and cell surface presentation of NMDA receptors	2.28E-04	*CAMK2A, LIN7A, GRIN3A, TUBB3, TUBB2A, LIN7C, DLG2, TUBB1, APBA1, TUBA4A, DLG4*
R-HSA-373760	L1CAM interactions	2.28E-04	*SH3GL2, ALCAM, GAP43, SPTBN2, TUBB3, CNTN1, MAPK3, SCN7A, TUBB2A, ANK2, NFASC, SCN3B, SCN8A, ITGA1, SPTB, DPYSL2, TUBB1, TUBA4A, NCAN, DLG4, NCAM1*
R-HSA-1296072	Voltage gated Potassium channels	2.28E-04	*KCNAB1, KCNA3, KCNA2, KCNB2, KCNC3, KCNH1, KCNH6, KCNH2, KCNH4, KCNC4, KCNH5, KCNG3*
R-HSA-397014	Muscle contraction	2.28E-04	*HIPK2, CAMK2A, KAT2B, FGF12, TMOD2, ITPR1, CACNA1D, CACNA2D3, KCNK9, ATP2A2, RYR2, CACNA2D2, ATP2A3, KCNJ2, MYH8, SCN7A, CACNA2D1, ATP2B2, SCN3B, KCNJ11, SCN8A, ITGA1, KCNH2, FGF14, ATP1A3, KCNK17, TMOD1, CACNA1C, KCNK13, NPPC*

R-HSA-983712	Ion channel transport	3.99E-04	*CAMK2A, TRPA1, ATP6V1A, WNK4, ATP6V1C1, ATP6V0D1, ATP6V0E2, ATP2A2, CLCN1, RYR2, ATP6V0A1, ATP9A, ATP2A3, TRPM3, ASIC2, ASIC1, ATP8A1, UNC80, ATP6V1B2, ATP2B2, MCOLN1, ATP1A3, UNC79, ANO4, SLC9B2, ANO5, ATP6V1G2*
R-HSA-6794362	Protein-protein interactions at synapses	4.78E-04	*DLGAP4, NRXN2, LIN7A, SYT2, LRFN2, STX1A, STXBP1, PPFIA2, LIN7C, DLG2, SYT7, EPB41L3, APBA1, DLGAP3, DLG4, GRIA4, LRRTM3*
R-HSA-422475	Axon guidance	7.38E-04	*CACNA1G, GAB2, PLXNA2, MET, SH3GL2, PSMF1, PLXNC1, CACNA1D, KALRN, RET, ALCAM, GAP43, DAB1, SPTBN2, CDK5R1, SRGAP3, CACNA1H, COL5A2, DOK6, TUBB3, CNTN1, CLTCL1, CRMP1, MAPK3, GRB10, PRKAR2A, VAV3, SCN7A, TUBB2A, EPHA5, ANK2, UNC5D, NFASC, SCN3B, RGMA, MAPK11, SCN8A, ITGA1, ZSWIM8, ISL1, SPTB, RAP1GAP, DPYSL2, TUBB1, SEMA5A, ST8SIA4, TUBA4A, NCAN, DLG4, VLDLR, SHC3, NCAM1, CACNA1C, CLASP2, ABLIM3, ROBO2, DCC, COL6A6, ARHGEF28*
R-HSA-388396	GPCR downstream signalling	8.44E-04	*CASR, REEP1, CNGB1, CAMK2A, GNAS, PLCB4, GLP1R, RASGRF2, MCF2L, GNAZ, DAGLA, GABBR2, ITPR1, PDE3B, GIP, KALRN, ADCY1, REEP3, RASGRP1, GNAT3, GNAO1, PTHLH, MAPK3, APOC3, PRKAR2A, PTGER2, GRK4, GNAI1, VAV3, MYO7A, FGD4, LRP8, APOA2, TAC1, GIPR, PRKCH, ADCY2, RGS17, PDE4C, GPR68, IAPP, FFAR2, PRKCG, RGS22, REEP2, ARHGEF4, RDH12, ADCY9, TSHR, PRKCB, GNG4, PRKCE, GPBAR1, PTGDR2, GNG2, PDE2A, AKR1C1, GPC2, GCGR*
R-HSA-112310	Neurotransmitter release cycle	1.20E-03	*LIN7A, GLS, UNC13B, STX1A, GAD1, STXBP1, PPFIA2, LIN7C, SLC18A2, APBA1, SNAP25, SYN3*
R-HSA-264876	**Insulin processing**	1.38E-03	*SLC30A8, KIF5C, RAB27A, ERO1B, STX1A, PCSK2, MYO5A, KIF5A, INS*
R-HSA-181431	Acetylcholine binding and downstream events	1.38E-03	*CHRNB4, CHRNA3, CHRNA4, CHRNA1, CHRNB2, CHRNA5*
R-HSA-622327	Postsynaptic nicotinic acetylcholine receptors	1.38E-03	*CHRNB4, CHRNA3, CHRNA4, CHRNA1, CHRNB2, CHRNA5*
R-HSA-5576892	Phase 0 - rapid depolarisation	1.56E-03	*CAMK2A, FGF12, CACNA1D, CACNA2D3, CACNA2D2, SCN7A, CACNA2D1, SCN3B, SCN8A, FGF14, CACNA1C*

R-HSA-210745	**Regulation of gene expression in beta cells**	1.56E-03	*HNF1A, FOXO1, MAFA, NEUROD1, PAX6, IAPP, RFX6, INS*
R-HSA-186712	**Regulation of beta-cell development**	1.87E-03	*PAX4, KAT2B, HNF1A, FOXO1, MAFA, NEUROD1, PAX6, IAPP, INSM1, RFX6, INS*
R-HSA-112308	Presynaptic depolarization and calcium channel opening	2.23E-03	*CACNA2D3, CACNA2D2, CACNA1B, CACNA2D1, CACNA1E, CACNA1A*
R-HSA-181429	Serotonin Neurotransmitter Release Cycle	2.39E-03	*UNC13B, STX1A, STXBP1, PPFIA2, SLC18A2, SNAP25, SYN3*
R-HSA-112043	PLC beta mediated events	3.15E-03	*CAMK2A, PLCB4, GNAZ, ITPR1, ADCY1, GNAT3, GNAO1, PRKAR2A, GNAI1, ADCY2, PRKCG, ADCY9*
R-HSA-977444	GABA B receptor activation	3.15E-03	*GABBR2, ADCY1, GNAT3, KCNJ2, GNAI1, KCNJ3, ADCY2, ADCY9, GNG4, GNG2*
R-HSA-991365	Activation of GABAB receptors	3.15E-03	*GABBR2, ADCY1, GNAT3, KCNJ2, GNAI1, KCNJ3, ADCY2, ADCY9, GNG4, GNG2*
R-HSA-73728	RNA Polymerase I Promoter Opening	3.24E-03	*HIST1H2AC, HIST1H2BK, MAPK3, HIST1H2BJ, HIST1H2BD, HIST4H4, HIST2H2BE*
R-HSA-74752	Signaling by Insulin receptor	3.36E-03	*GAB2, ATP6V1A, PDE3B, ATP6V1C1, FGF17, ATP6V0D1, ATP6V0E2, ATP6V0A1, MAPK3, GRB10, FGF5, ATP6V1B2, ATP6V1G2, INS*
R-HSA-77387	Insulin receptor recycling	3.40E-03	*ATP6V1A, ATP6V1C1, ATP6V0D1, ATP6V0E2, ATP6V0A1, ATP6V1B2, ATP6V1G2, INS*
R-HSA-112040	G-protein mediated events	3.40E-03	*CAMK2A, PLCB4, GNAZ, ITPR1, ADCY1, GNAT3, GNAO1, PRKAR2A, GNAI1, ADCY2, PRKCG, ADCY9*
R-HSA-438064	Post NMDA receptor activation events	3.66E-03	*RASGRF1, CAMK2A, RASGRF2, ADCY1, TUBB3, MAPK3, PRKAG2, PRKAR2A, TUBB2A, DLG2, TUBB1, GRIA2, TUBA4A, DLG4*
R-HSA-171306	Packaging Of Telomere Ends	4.11E-03	*HIST1H2AC, HIST1H2BK, HIST1H2BJ, HIST1H2BD, HIST4H4, TERF2IP, HIST2H2BE*

R-HSA-2980736	Peptide hormone metabolism	4.11E-03	*GIP, SLC30A8, KIF5C, GNAT3, RAB27A, ERO1B, STX1A, PCSK2, MYO5A, PAX6, ISL1, KIF5A, MME, INS*
R-HSA-977443	GABA receptor activation	6.12E-03	*GABBR2, ADCY1, GNAT3, KCNJ2, GNAI1, KCNJ3, GABRB3, ADCY2, ADCY9, GNG4, GNG2*
R-HSA-442660	Na+/Cl- dependent neurotransmitter transporters	6.90E-03	*SLC6A15, SLC6A19, SLC6A18, SLC18A2, SLC18A1, SLC6A6*
R-HSA-983189	Kinesins	8.60E-03	*KIF3C, KIF3A, TUBB3, KIF1A, KIF12, TUBB2A, KIF21A, KIFAP3, TUBB1, TUBA4A, KIF5A*
R-HSA-181430	Norepinephrine Neurotransmitter Release Cycle	9.71E-03	*UNC13B, STX1A, STXBP1, PPFIA2, SLC18A2, SNAP25*
R-HSA-111885	Opioid Signalling	9.71E-03	*CAMK2A, PLCB4, GNAZ, ITPR1, ADCY1, GNAT3, GNAO1, PRKAR2A, GNAI1, ADCY2, PDE4C, PRKCG, ADCY9, GNG4, GNG2*
R-HSA-375165	NCAM signaling for neurite out-growth	9.82E-03	*CACNA1G, CACNA1D, SPTBN2, CACNA1H, COL5A2, MAPK3, SPTB, ST8SIA4, NCAN, NCAM1, CACNA1C, COL6A6*
R-HSA-917977	Transferrin endocytosis and recycling	1.14E-02	*ATP6V1A, ATP6V1C1, ATP6V0D1, ATP6V0E2, ATP6V0A1, ATP6V1B2, MCOLN1, ATP6V1G2*
R-HSA-110330	Recognition and association of DNA glycosylase with site containing an affected purine	1.24E-02	*HIST1H2AC, HIST1H2BK, HIST1H2BJ, HIST1H2BD, HIST4H4, TERF2IP, HIST2H2BE*
R-HSA-110331	Cleavage of the damaged purine	1.24E-02	*HIST1H2AC, HIST1H2BK, HIST1H2BJ, HIST1H2BD, HIST4H4, TERF2IP, HIST2H2BE*
R-HSA-73927	Depurination	1.24E-02	*HIST1H2AC, HIST1H2BK, HIST1H2BJ, HIST1H2BD, HIST4H4, TERF2IP, HIST2H2BE*
R-HSA-352230	Amino acid transport across	1.34E-02	*SLC7A2, SLC6A15, SLC7A8, SLC6A19, SLC6A18, SLC36A4, SLC6A6, SLC43A2*

	the plasma membrane		
R-HSA-1855183	Synthesis of IP2, IP, and Ins in the cytosol	1.47E-02	*IMPA1, SYNJ1, MTMR7, INPP1, INPP5J*
R-HSA-419037	NCAM1 interactions	1.55E-02	*CACNA1G, CACNA1D, CACNA1H, COL5A2, ST8SIA4, NCAN, NCAM1, CACNA1C, COL6A6*
R-HSA-163359	Glucagon signaling in metabolic regulation	1.61E-02	*GNAS, ADCY1, PRKAR2A, ADCY2, ADCY9, GNG4, GNG2, GCGR*
R-HSA-5334118	DNA methylation	1.63E-02	*HIST1H2AC, HIST1H2BK, HIST1H2BJ, HIST1H2BD, HIST4H4, HIST2H2BE*
R-HSA-5625886	Activated PKN1 stimulates transcription of AR (androgen receptor) regulated genes KLK2 and KLK3	2.20E-02	*HIST1H2AC, HIST1H2BK, HIST1H2BJ, HIST1H2BD, HIST4H4, HIST2H2BE*
R-HSA-425393	Transport of inorganic cations/anions and amino acids/oligopeptides	2.59E-02	*SLC7A2, SLC4A8, SLC6A15, SLC17A6, SLC7A8, AHCYL2, SLC6A19, SLC24A2, SLC6A18, SLC17A8, SLC36A4, SLC6A6, SLC24A4, SLC4A10, SLC43A2*
R-HSA-1369062	ABC transporters in lipid homeostasis	2.80E-02	*ABCG1, ABCA7, ABCA9, ABCA3, ABCA10*
R-HSA-170670	Adenylate cyclase inhibitory pathway	2.80E-02	*ADCY1, GNAT3, GNAI1, ADCY2, ADCY9*
R-HSA-1296065	Inwardly rectifying K+ channels	2.80E-02	*GABBR2, KCNJ2, KCNJ3, KCNJ11, ABCC8, GNG4, GNG2*

R-HSA-2299718	Condensation of Prophase Chromosomes	2.80E-02	*SMC2, HIST1H2AC, HIST1H2BK, HIST1H2BJ, HIST1H2BD, HIST4H4, HIST2H2BE*
R-HSA-399719	Trafficking of AMPA receptors	2.80E-02	*CAMK2A, NSF, GRIA2, PRKCG, DLG4, GRIA4, PRKCB*
R-HSA-399721	Glutamate binding, activation of AMPA receptors and synaptic plasticity	2.80E-02	*CAMK2A, NSF, GRIA2, PRKCG, DLG4, GRIA4, PRKCB*
R-HSA-6811434	COPI-dependent Golgi-to-ER retrograde traffic	3.21E-02	*NSF, KIF3C, KIF3A, TUBB3, NAPB, KIF1A, KIF12, TUBB2A, KIF21A, SURF4, KIFAP3, TUBB1, TUBA4A, KIF5A*
R-HSA-110328	Recognition and association of DNA glycosylase with site containing an affected pyrimidine	3.21E-02	*HIST1H2AC, HIST1H2BK, HIST1H2BJ, HIST1H2BD, HIST4H4, TERF2IP, HIST2H2BE*
R-HSA-110329	Cleavage of the damaged pyrimidine	3.21E-02	*HIST1H2AC, HIST1H2BK, HIST1H2BJ, HIST1H2BD, HIST4H4, TERF2IP, HIST2H2BE*
R-HSA-420092	Glucagon-type ligand receptors	3.21E-02	*GNAS, GLP1R, GIP, GIPR, GNG4, GNG2, GCGR*
R-HSA-445095	Interaction between L1 and Ankyrins	3.21E-02	*SPTBN2, SCN7A, ANK2, NFASC, SCN3B, SCN8A, SPTB*
R-HSA-73928	Depyrimidination	3.21E-02	*HIST1H2AC, HIST1H2BK, HIST1H2BJ, HIST1H2BD, HIST4H4, TERF2IP, HIST2H2BE*
R-HSA-416993	Trafficking of GluR2-containing AMPA receptors	3.47E-02	*NSF, GRIA2, PRKCG, GRIA4, PRKCB*

R-HSA-1222556	ROS and RNS production in phagocytes	3.87E-02	*ATP6V1A, ATP6V1C1, ATP6V0D1, ATP6V0E2, ATP6V0A1, ATP6V1B2, ATP6V1G2*
R-HSA-202040	G-protein activation	3.88E-02	*GNAZ, GNAT3, GNAO1, GNAI1, GNG4, GNG2*
R-HSA-427359	SIRT1 negatively regulates rRNA expression	3.88E-02	*HIST1H2AC, HIST1H2BK, HIST1H2BJ, HIST1H2BD, HIST4H4, HIST2H2BE*
R-HSA-168799	Neurotoxicity of clostridium toxins	4.30E-02	*SYT2, STX1A, SNAP25, SV2A*
R-HSA-164378	PKA activation in glucagon signalling	4.38E-02	*GNAS, ADCY1, PRKAR2A, ADCY2, ADCY9*
R-HSA-264642	Acetylcholine Neurotransmitter Release Cycle	4.38E-02	*UNC13B, STX1A, STXBP1, PPFIA2, SNAP25*
R-HSA-73929	Base-Excision Repair, AP Site Formation	4.38E-02	*HIST1H2AC, HIST1H2BK, HIST1H2BJ, HIST1H2BD, HIST4H4, TERF2IP, HIST2H2BE*
R-HSA-210500	Glutamate Neurotransmitter Release Cycle	4.63E-02	*GLS, UNC13B, STX1A, STXBP1, PPFIA2, SNAP25*

Appendix Table 2.2: Selected transcription factors enriched in differentially expressed genes assigned to each stage. Differentially expressed genes (q<0.01, log2FC > 1) assigned to the stage in which they presented the maximum log2FC compared to the iPSC baseline were tested for enrichment in transcription factor binding motifs and binding (by ChIP-seq) using iRegulon. Only significant (NES > 3) terms are shown. Transcription factors mentioned in the main text in bold.

Transcription factor binding motifs			
Stage	**Transcription factor**	**NES**	**Target genes**
GT	**HNF1B**, **HNF1A**, CRX, NFYC, NFYA, RUNX1, PPARG, IRX4, POU3F2, PDX1, PAX4, FOXO1, HMBOX1, MYB, BPTF, ISL1, POU3F1, NOBOX, RAX, CDX2, TLX2, NR1H4, NR1I3, PBX1, NR1I2, POU2F1, POU1F1, HLTF, DBX2, NR2F1, NR2F2, RHOXF1	5.7	*SLC40A1, GRB7, CDC42EP4, TBX2, PAX2, CA4, FGFR4, SMAD7, RORA, ID2, SPDEF, VEGFA, FBN2, LAMB1, TCEA3, APOM, ZMIZ1, SLC25A45, ARHGAP24, TSPAN9, KCP, RAB3IL1, GATA2, MSX2, GAS2, PIK3R1, TRPV4, GATS, IHH, COL4A1, SEMA4G, LGI2, NRG2, RHBG, SLC39A5, PCDH17, VASN, DPYD, SLC23A1, PAMR1, CUBN, VCAN, ATP13A4, TMPRSS2, RFFL, FLRT3, GGCX, NRCAM, PEG10, GRK5, SLC22A2, ACPP, FAM20C, KLHL24, ANO10, VAX1, PCBD1, C2orf88, SLA2, PLA1A, SMAD6, RASSF10, CLDN19, TMEFF1, PMM1, ID1, ACSF2, BIN1, ANK3, MGAT3, FOLR1, TBX3, DKK1, SCN7A, LGALS2, LOXL1, RNF24, RALGPS2, ADD3, TBKBP1, RGS5, KCNC4, NTF3, ZNRF2, AGRN, PPARG, BMP6, GLIS1, SALL1, MARCH1, RDX, HEY2, KYNU, BIRC2, ATP10B, BAMBI, SSBP3, STK38, GJB4, ACADSB, C1orf131, HGF, SMAD1, SETD1B, KANK1, GPR107, FOXC1, NYNRIN, RAP1A*
EN	**REST**	5.3	*LHX5, RXFP3, TRPC7, SRRM4, NXPH1, GPRIN1, RESP18, STXBP5L, CDK5R2, HCN3, TUBB3, PRSS27, KCNN4, TNR, NCAN, MAPK8IP1, ELAVL3, MMP24, VGF, SLC12A5, FGF14, PENK, SCAMP5, PLXNA2, CDKN2B, MYT1, SULT4A1, FAM57B, UNC5A, CDK5R1, ELAVL2, CELF3, KCNJ6, MADD, NEUROD1, MAPK11, DLL4, APC2, PDE2A, FBXL16, CELSR3, ACAP3, CLVS1, SLC46A1, FOS*
BLC	**REST**	7.9	*SNAP25, GOLGA7B, SST, GRIK2, VWC2L, CHRNB2, STMN2, SCN2A, SLC7A14, KCNH2, INA, CABP7, NRSN1, SLC17A6, CHGB, PLCH2, FBLL1, GABRG2, NEFM, RPH3A, PTPRN, GRM1, CACNA1B, SEZ6, OTP, HCN1, PHYHIPL, CHRM4, EPHA10, NFASC, POU4F2, CHGA, SLC6A5, CTNNA2, GDAP1L1, SCN3B, HCN2,*

			SLC18A3, NRXN1, C1QL2, ATP6V0A1, CADPS, GLRA3, RAP1GAP, TMEM179, CRYBA2, GALNT13, SARM1, CBLN1, SYBU, DOCK10, HNF1A, SLC6A17, DPP6, ACSL6, OLFM3, CACNA1E, LHFPL5, OMG, HPCA, APBB1, NPTX2, PCLO, GRIA2, SPNS2, GRIN3A, NKAIN4, ACSS1, DLGAP3, CA10, NRXN2, CRIP2, SPTBN4, C3orf14, HES2, DNER, KY, GAD2, CPNE4, RIPPLY2, RIMBP2, VSTM2A, LRRC24, GABBR2, CPLX2, IGSF21, MYO1A, ENOX1, PURA, RASGRF1, CYP4F3, DDC, MTUS2, BCL2L2, SYNJ1, KCNMA1, KIAA1644, VPS54, PCSK2, KCNC3, GRID1, ELFN2, SERPINA10, KCNIP1, MAST4, SLC29A4, LRFN2, ZNF536, CRTAC1, RGS17, PRUNE2, ATP8A2, DENND3, PCSK1, HES5, PRKAG2, GRIA1, CELF4, CHRNA4, GRM7, SV2B, OLFM4, SGK2, RUNDC3A, GNAS, FHIT, LRP5L, MEGF9, ATP2B2, VWA1, ELOVL4, ARHGEF10L
ChIP-seq tracks			
Stage	**Transcription factor**	**NES**	**Target genes**
PF	**TCF7L2**	3.4	*AUTS2, ZNF703, ZSWIM6, AMDHD1, ZFHX3, TCF7, NR4A1, ZNF443, TIGD5, TCF7L2, ZNF799, SSH1, SLC35E2, PIK3C2B, MIDN, ZNF823, HS3ST3B1, PRSS16, ROBO2, TLE3, PRTG, TSHZ2, SLC35E2B, FOXP1, SH3RF2, HOXA1, GUCY1A3, HES1, FN1, SH3BP2, MECOM, ZNF764, TES, PLXNA1, NR2F2, MNT, PHLDB2, SIAH1, MEIS2, RARB, TIAM1, ARHGAP10, ZNF570, DSP, IDH2, BCL9, H1F0, SERPINB1, CDKN2AIP, MED13L, UTRN, BMF, OCLN, PLEKHH1, RANBP9, SOX4, PLCG1*
EN	**REST**	4.9	*TMEM196, GJD2, RXFP3, TMEM151A, SEC14L5, SPIRE2, SYT5, PPFIA3, LHX5, MYT1, KCNJ6, GPR158, GDAP1, ACAP3, XKR7, CDK5R2, SCAMP5, NEUROD1, SH2D3C, TUBB3, SULT4A1, HCN3, AP3B2, FGF14, MMP24, ACTL6B, STXBP5L, RTN2, PSD, HEXDC, SMPD3, SLC12A5, MAPK11, CLVS1, FBXL16, CNTNAP5, MARCH4, VGF, NCAN, FAM57B, XKR4, RESP18, NXPH1, APC2, MAPK8IP1, RIMS2, SEZ6L, SRRM4, UNC5A, TRPC7, TMEM145, MPPED1, KIAA0895L, DISP2, GPRIN1, CLPS, GPS2, KCNN4, GNG4, PDE2A, VWA5B2, PPM1L, WNT2B,*

			PRSS27, RGS7, CELF3, PODXL2, FGF18, NLRP1, ABHD2, GPR149, KCNK3, ELAVL3, TNR, CRTC1, PENK, FAM78B, DPYSL5, CDK5R1, CLCN1, CELSR3
BLC	**REST**	4.9	*RPH3A, TPH2, PTPRN, HPCA, SEZ6, CHRM4, POU4F2, SVOP, SLC6A5, PCSK1, HAND2, SST, GRM1, VWC2L, CHGB, C1QL2, CPLX2, IQSEC3, GABRG2, EPHA10, SLC6A17, SNTG1, ALK, RAP1GAP, SNAP25, SCGN, NFASC, CHGA, SEZ6L2, ASPHD1, KCNH2, SLC7A14, LHFPL4, CELF4, PDE4D, GRIA2, CABP7, SCG3, LHFPL5, NEFM, GLRA3, HCN2, OLFM3, EGR4, NRSN1, SLITRK1, CPNE4, MANEAL, HMP19, CA10, PLCH2, KCNB1, ADCYAP1, NRXN1, CNGA3, UNC80, NMNAT2, MAPK8IP2, INA, DDC, GDAP1L1, CACNA1B, PEX5L, MYT1L, KCNC3, SCRT2, RUNDC3A, RIMBP2, GPR26, SHISA7, TMEM179, GABBR2, GRIK2, CHRNB2, PHYHIPL, NPDC1, SPNS2, DNER, CTNNA2, TRIM67, KCNC1, KCNC2, ACSL6, VSTM2A, CADPS, TBC1D30, TTC9B, SLC17A6, RAB3C, SYNJ1, TMEM198, SCN3B, FBXL15, GOLGA7B, DPP6, DLGAP3, SLC18A3, SARM1, GNAO1, NPTX2, SLC8A2, SPTBN4, ATP6V1B2, PRUNE2, SCN2A, HIC1, OTP, C3orf14, TWIST1, BSN, GALNT13, FBLL1, SLC6A2, CRYBA2, CPLX1, SYBU, STMN2, CBLN4, HIST1H4H, MUC20, RASGRF1, FSTL4, CYP4F3, PTPRN2, NEUROD4, MYO1A, OTOP3, HIST2H2BE, LRRC24, PCLO, ATL1, GRIN3A, HIST1H2AC, NRXN2, HIST1H2BC, HCN1, PBLD, PRSS22, PTEN, SLC29A4, RIC3, HIST1H2BN, GPBAR1, BTG1, GABRA1, NFIC, ATP6V0A1, RFX4, ENOX1, CYB561*

Appendix Table 3.1: ATAC-seq peaks overlapping credible set variants > 10% PPAg. PPAg: Genetic posterior probability of association. ATAC-seq peak presence (1) or absence (0) at the SNP position per stage. TF: transcription factor binding motif at SNP position.

			ATAC-seq peak presence								
rsID	**locus**	**PPAg**	**iPSC**	**DE**	**GT**	**PF**	**PE**	**EP**	**EN**	**BLC**	**TF**
rs231361	*KCNQ1*	1.00	0	0	0	0	0	0	1	0	MECOM, RFX5, ZSCAN9
rs17772814	*CASC11*	1.00	0	1	0	0	0	0	0	0	HNF1A, HNF1B, HOXA3, MEF2C, ZBTB41, ZFP62, ZNF195, ZNF236, ZNF514
rs528350911	*WDR72*	0.99	0	0	0	1	1	1	0	0	ARID5A, FOXJ2, HMGA1, MECOM, NR2F2, PPARD, PRDM16, RXRA, ZNF761
rs11680058	*FAM49A*	0.99	0	0	0	0	0	0	1	0	ARID5A, GTF3A, STAT3, STAT5B, ZNF467, ZNF764
rs1800574	*HNF1A*	0.98	0	0	0	1	1	1	1	1	CTCF, GLIS3, PLAGL2
rs917195	*CRHR2*	0.96	0	0	0	1	0	0	0	0	ID4, MEIS3, ASCL5, ASCL1, MYF5, RFX1, RFX2, RFX3, RFX5, RFX6, S I1, TBX2, TCF12, TCF3, TCF4, ZEB1, ZEB2, TSHZ2, TSHZ1, ZNF585B, ZNF91
rs362307	*HTT*	0.95	0	0	0	0	0	0	1	0	
rs150111048	*IGF2BP2*	0.94	0	0	0	0	1	0	0	0	BPTF, FOXJ2, FOXJ1, FOXJ3, FOXN3, FOXN2, FOXP4, FUBP1, IRF2, MECOM, MEF2A, STAT2, ZIK1, ZNF117, ZNF22, ZNF239, ZNF287, ZNF362, ZNF594
rs12001437	*UBAP2*	0.92	1	0	1	0	1	0	1	1	CTCF, NFATC2, NFATC3, NFATC4, SRF, ZNF582
rs2237897	*KCNQ1*	0.81	0	0	1	1	1	1	1	0	PURA, ZNF263, ZNF513, ZSCAN25
rs11042596	*INS/IGF2*	0.72	1	1	0	1	1	0	0	0	EPAS1, FOXO1, FOXO3, GLI1, HINFP, KLF3, KLF5, KLF7, MAZ, MZF1, SP1, SP3, ZIC2, ZNF148, ZNF740, ZNF746
rs199795270	*FAM57B*	0.66	1	1	1	1	1	1	1	1	PROX2, PROX1, ZNF510, ZNF594, ZNF721
rs79687284	*PROX1*	0.64	0	1	1	1	0	0	0	0	HES1, HES6, HES3, SREBF1, ZNF107, ZNF836, ZSCAN30
rs17168486	*DGKB*	0.58	0	1	0	1	0	0	1	1	NFIB, RFX1, RFX2, RFX3, RFX6, TBX2
rs191830490	*HNF4A*	0.52	0	0	1	1	1	0	1	0	CTCF, NFIB, ZNF554, ZNF594
rs963740	*DLEU1*	0.50	0	0	0	1	1	0	1	1	CEBPG, MAFB, ZBTB38, ZNF627
rs10811660	*CDKN2A/B*	0.47	0	0	0	0	0	0	1	0	MAFF, MAFG, MAFK, THRB, ZNF440
rs10811661	*CDKN2A/B*	0.41	0	0	0	0	0	0	1	0	STAT2
rs17712208	*PROX1*	0.36	0	1	1	1	1	0	0	0	HNF1A, ISX, ZNF101, ZNF526, ZNF700, ZNF730

rs13266634	*SLC30A8*	0.35	0	1	1	1	1	1	1	1	ASCL2, ATOH1, BRCA1
rs10757282	*CDKN2A/B*	0.35	0	0	0	0	0	0	1	0	NKX2-2, NKX6-1, TCF7, TCF7L2, ZNF195, ZNF607
rs62059712	*ATP1B2*	0.34	1	1	1	1	1	1	1	1	NR2C2, ZBTB22, ZBTB43, ZNF236, ZNF347, ZNF467, ZNF529, ZNF69
rs78840640	*IGF2BP3*	0.33	0	0	0	0	0	0	1	1	HIC2, MAFG, MAFK, STAT3, STAT4, ZNF248, ZNF860
rs5213	*KCNJ11*	0.32	0	0	1	1	1	0	1	0	ID4, KLF6
rs7503751	*ATP1B2*	0.32	0	0	0	0	0	0	1	0	CTCF, HEY2, NR2C2, ZIC5, ZNF213, ZNF407, ZNF655
rs2115107	*MAP2K7*	0.31	1	1	1	1	1	1	1	1	INSM1, MAZ, MZF1, NR2F6, PATZ1
rs1421085	*FTO*	0.28	0	0	0	1	0	0	0	0	ZGLP1, GATA2, HNF1A, ONECUT2, PBX2, PBX4, ZNF205, ZNF287, ZNF562
rs7732130	*ZBED3*	0.28	0	0	1	0	0	0	0	0	RFX2, RFX5, SMAD4, SMAD6, SREBF1, SREBF2, ZNF302, ZNF526, ZNF853, ZSCAN25
rs13080180	*SHQ1*	0.27	0	1	1	1	1	1	0	0	BPTF, FOXA2, FOXQ1, FUBP1, SOX4, ZBTB11, ZBTB43, ZNF510
rs17122772	*SLC7A7*	0.27	0	0	1	1	1	1	1	0	ZNF276, ZNF853
rs17122779	*SLC7A7*	0.25	0	0	1	1	1	1	1	0	CTCF, MAZ, PRDM10, SP1, ZNF263, ZNF281, ZNF467, ZNF519, ZNF730, ZNF746
rs291367	*GNG4*	0.24	0	0	1	1	1	0	1	0	RFX1, ZNF334, ZNF584, ZSCAN30
rs17122776	*SLC7A7*	0.23	0	0	1	1	1	1	1	0	ATF4, MAFF, MAFG, MAFK, ZFP41, ZNF700
rs17122782	*SLC7A7*	0.23	0	0	1	1	1	1	1	0	KLF15, MAZ, MZF1, RARA, RARG, RREB1, SP1, ZNF148, ZNF263, ZNF784
rs1624085	*ATP1B2*	0.22	1	1	1	1	1	1	1	1	EBF1, EBF2, EBF4, GLI3, NR2F1, PPARD, ZNF586, ZNF746
rs2820446	*LYPLAL1*	0.22	0	0	0	0	0	1	1	1	HLTF, ZNF236, ZNF592, ZNF7
rs4417225	*INS/IGF2*	0.21	0	0	0	0	1	0	0	0	HNF4A, IRF6, NR2F2, PPARD, RARG, STAT2, ZNF273
rs62370480	*ARL15*	0.20	1	1	1	1	1	0	1	0	E2F6, GTF3A, YBX3, YBX1, ZNF287, ZNF347, ZNF710, ZNF721, ZNF829
rs55911137	*CCND1*	0.20	1	1	1	1	1	1	0	0	CTCF, E2F1, E2F6
rs10097617	*TP53INP1*	0.20	1	1	1	1	1	1	1	1	IKZF5
rs329118	*PHF15*	0.19	0	0	0	0	1	0	0	0	ZSCAN25
rs13737	*PTPN9*	0.19	1	1	1	1	1	1	1	1	ATF6, CREBZF, CREB3, NR4A2, SMAD1, SMAD9, SMAD5, XBP1, ZNF354C

rs1575972	*CDKN2A/B*	0.19	0	0	0	0	0	1	1	1	FOXA1, FOXA3, FOXJ2, ZGLP1, GATA2, RFX5, SOX17, ZFP64, ZNF101, ZNF616, ZNF787, ZNF841
rs74046911	*KCNQ1*	0.17	0	0	1	1	1	1	1	0	EGR1, MAZ
rs286925	*PDHX*	0.17	0	0	0	0	1	0	1	0	ZIC5, ZNF143
rs144155527	*TCF7L2*	0.17	0	0	0	0	0	0	1	1	TGIF1, ZIC5
rs12593201	*RASGRP1*	0.17	0	0	0	0	0	0	1	0	BPTF
rs329120	*PHF15*	0.16	0	0	0	0	1	0	0	0	NFIB, SOX17, ZNF236
rs13173241	*ANKRD55*	0.15	0	0	0	0	0	0	1	1	HIVEP2, PLAGL2, SREBF2, ZNF200
rs55744247	*ARL15*	0.15	0	0	0	0	1	0	1	0	ELK3, HIF1A, HIF3A, MAFB, ZBTB4, ZNF263, ZNF467
rs1848068	*SCD5*	0.15	0	0	0	0	0	0	1	1	BCL11A, MYBL2, RCOR1, NKX2-2, PAX6, PPARA, TBX2, TGIF1, ZNF507, ZNF589
rs7109575	*CENTD2/ARAP1*	0.14	1	1	1	1	1	1	1	1	ARNT2, EGR1
rs896853	*TP53INP1*	0.14	1	1	1	1	1	1	1	1	HIVEP2, PURA
rs2879813	*TP53INP1*	0.14	1	1	1	1	1	1	1	1	CTCF, PLAG1
rs11072553	*PTPN9*	0.14	1	1	1	1	1	1	1	1	RCOR2, MYB, S PC4, RCOR3
rs11263763	*HNF1B*	0.13	0	0	0	0	1	0	0	0	KLF6, ZNF213
rs77773608	*FBRSL1*	0.13	1	1	1	1	1	1	1	1	MBD2, MYBL2, RCOR1, ZNF213, ZNF749
rs17728127	*PCGF3*	0.13	1	0	0	1	1	0	0	0	ELK3, ERF, ETS1, ETV1, ETV4, ETV5, FEV, FOXR1, FOXK1, FOXO1, FOXO3, FOXP2, HIVEP2, IRF3, NFATC3, PRDM6, STAT2, ZNF213, ZNF296, ZNF721
rs11065280	*HNF1A*	0.13	0	1	1	1	1	0	0	0	ZFP41, ZNF335
rs17036160	*PPARG*	0.13	1	1	1	1	1	1	1	1	ZNF550, ZNF641, ZNF682
rs256904	*ANKRD55*	0.12	0	0	0	1	1	0	0	0	ZNF764
rs7943101	*QSER1*	0.12	0	0	0	1	1	0	1	0	ARNT2, FOXO1, SP1, ZNF281
rs73322872	*TCEA2*	0.12	1	1	1	1	1	1	1	1	OSR1, THRA
rs6817471	*SCD5*	0.12	0	0	0	0	0	0	1	1	CREB3L2, KLF12, MEIS3, PLAGL2, TGIF2, ZNF730
rs7667864	*LCORL*	0.12	1	1	1	1	1	1	1	1	FOXR1, FOXK1, FOXP1, FOXP2, KLF3
rs252924	*ANKRD55*	0.12	1	1	1	1	1	1	1	1	E2F1, E2F4, E2F6
rs11759956	*RREB1*	0.12	0	1	1	1	1	1	1	0	

rs6937438	*VEGFA*	0.11	0	0	0	1	0	0	0	0	ZNF114, ZNF426
rs243018	*BCL11A*	0.11	0	1	1	1	1	1	1	1	ARNT2, FIZ1, FOXJ1, FOXN2, FOXN3, FOXJ3, SMAD2, ZNF574
rs703977	*ZMIZ1*	0.11	0	0	0	0	0	0	1	0	ASCL2, ATOH1, ID4, ASCL5, ASCL1, MYF5, NFATC3, NKX6-1, TCF3, ZBTB40, ZNF121, ZNF287, ZNF775
rs11227217	*MAP3K11*	0.11	1	1	1	1	1	1	1	1	E2F6, EGR1, HLTF, KLF15, MAZ, SP1, SP2, ZNF219
rs12789028	*MAP3K11*	0.11	0	0	1	1	1	0	0	0	FIZ1, HIVEP2, ZNF141, ZNF263, ZNF467, ZNF480
rs13428381	*THADA*	0.11	0	0	0	0	0	0	0	1	EGR1, RREB1, RUNX1, TBX20

Appendix Table 3.2: H3K27ac peaks overlapping credible set variants > 10% PPAg. PPAg: Genetic posterior probability of association. H3K27ac peak presence (1) or absence (0) at the SNP position per stage. TF: transcription factor binding motif at SNP position.

			H3K27ac peak presence								
rsID	**Locus**	**PPAg**	**iPSC**	**DE**	**GT**	**PF**	**PE**	**EP**	**EN**	**BLC**	**TF**
rs231361	*KCNQ1*	1.00	0	0	0	0	0	1	1	0	MECOM, RFX5, ZSCAN9
rs76895963	*CCND2*	1.00	1	0	1	1	1	1	1	0	
rs3217792	*CCND2*	1.00	1	0	1	1	1	1	1	0	GLI2
rs796464202	*CCND2*	1.00	1	0	1	1	1	1	1	0	GLI2
rs34298980	*LRFN2*	0.99	0	0	0	0	1	0	1	1	ZFP90
rs3217860	*CCND2*	0.99	0	0	1	0	0	0	0	0	GLIS2, GLIS3, HIVEP1, HSF1, INSM1, KLF15, MAZ, NFAT5, NR1H3, RXRA
rs3887925	*ST6GAL1*	0.99	0	0	0	0	0	0	1	1	KLF12, KLF11, KLF13, KLF16, KLF7, SP1, SP3, SP5
rs1800574	*HNF1A*	0.98	0	0	1	1	1	1	1	1	CTCF, GLIS3, PLAGL2
rs2066827	*CDKN1B*	0.98	0	0	0	1	0	1	1	0	CREM, CREBL2, ELF1, ELK3, ETS1, ETV5, FEV, ZNF594
rs362307	*HTT*	0.95	0	0	0	0	0	0	1	0	
rs41277236	*NEUROG3*	0.74	0	0	0	0	0	1	1	0	EBF1, EBF2, EBF4, GLI1, NR3C1, RFX1, RFX2, RFX5, ZBTB7A, ZBTB7B, ZBTB7C
rs199795270	*FAM57B*	0.66	0	0	0	0	1	0	0	0	PROX2, PROX1, ZNF510, ZNF594, ZNF721
rs35895680	*TTLL6*	0.66	0	0	1	0	0	0	0	1	CEBPB, DDIT3, KLF15, ZNF730, ZSCAN10
rs72802342	*BCAR1*	0.66	0	1	0	0	0	0	0	0	ARNT2, GLIS2, GLIS3, PATZ1, YY1, ZNF219
rs79687284	*PROX1*	0.64	0	0	0	1	0	1	1	1	HES1, HES6, HES3, SREBF1, ZNF107, ZNF836, ZSCAN30
rs145678014	*QSER1*	0.64	0	1	0	0	0	0	0	0	IKZF2, TFCP2, ZNF710
rs113414093	*CRYBA2*	0.63	0	0	0	0	0	1	1	1	MBD2
rs3802177	*SLC30A8*	0.57	0	0	0	0	0	1	0	0	TEAD1, TEAD2, ZNF287, ZNF710
rs4925109	*RAI1*	0.56	0	0	0	0	0	0	1	0	ELF2, PURA, STAT4, ZNF189, ZNF263, ZNF3, ZNF397, ZNF577
rs191830490	*HNF4A*	0.52	0	0	1	0	0	1	1	1	CTCF, NFIB, ZNF554, ZNF594
rs10757283	*CDKN2A/B*	0.50	0	0	0	0	0	1	1	1	PBX2, PBX4, ZNF384, ZNF426, ZNF836

rs115505614	*PAM*	0.46	0	0	0	0	0	0	1	0	KLF3, MSC, ASCL5, ASCL1, MYF5, ZNF746
rs10811661	*CDKN2A/B*	0.41	0	0	0	0	0	1	1	1	STAT2
rs560716466	*EYA2*	0.40	1	0	1	1	1	1	1	0	IRF2, STAT2, ZNF107, ZNF195, ZNF263, ZNF347, ZNF701, ZNF775, ZSCAN9
rs9430095	*SRGAP2*	0.35	0	0	0	0	0	0	1	0	ESRRA, NR3C1, NR5A2, ZNF697
rs13266634	*SLC30A8*	0.35	0	0	0	0	0	1	0	0	ASCL2, ATOH1, BRCA1
rs62059712	*ATP1B2*	0.34	0	1	1	1	1	1	1	1	NR2C2, ZBTB22, ZBTB43, ZNF236, ZNF347, ZNF467, ZNF529, ZNF69
rs2115107	*MAP2K7*	0.31	0	0	1	0	0	1	1	0	INSM1, MAZ, MZF1, NR2F6, PATZ1
rs9873618	*SLC2A2*	0.29	0	0	0	0	0	1	0	0	
rs7645517	*ST6GAL1*	0.29	0	0	0	0	0	1	0	0	BCL6, PRDM2, STAT4, ZNF446, ZNF700
rs17122772	*SLC7A7*	0.27	0	1	0	0	0	0	0	0	ZNF276, ZNF853
rs17122779	*SLC7A7*	0.25	0	1	0	0	0	0	0	0	CTCF, MAZ, PRDM10, SP1, ZNF263, ZNF281, ZNF467, ZNF519, ZNF730, ZNF746
rs1977833	*HHEX/IDE*	0.24	0	0	1	0	0	0	0	0	ESRRA, OTX1, SMAD2, ZNF581, ZNF836
rs998584	*VEGFA*	0.24	0	0	1	0	0	0	0	0	ZNF707
rs8071043	*ZZEF1*	0.24	1	0	0	0	0	0	0	0	HMGA2, ZNF500, ZNF532, ZNF793
rs17122776	*SLC7A7*	0.23	0	1	0	0	0	0	0	0	ATF4, MAFF, MAFG, MAFK, ZFP41, ZNF700
rs17122782	*SLC7A7*	0.23	0	1	0	0	0	0	0	0	KLF15, MAZ, MZF1, RARA, RARG, RREB1, SP1, ZNF148, ZNF263, ZNF784
rs739846	*TM6SF2*	0.22	0	0	1	0	0	0	0	0	NR6A1, REL, ZNF18, ZNF672
rs11699802	*CEBPB*	0.22	0	0	1	0	1	1	0	0	RCOR2, MYB, S PC4, RCOR3, PLAGL1
rs10838787	*INS/IGF2*	0.22	0	0	0	0	0	1	1	1	RBPJL, RBPJ, ZFP30, ZNF467
rs2982521	*MIR3668*	0.22	0	1	0	0	0	0	0	0	E2F6, SREBF2, ZFP42, ZNF550, ZSCAN12
rs7847880	*FOCAD*	0.22	1	0	0	0	0	0	0	1	
rs62370480	*ARL15*	0.20	0	1	0	0	0	0	0	0	E2F6, GTF3A, YBX3, YBX1, ZNF287, ZNF347, ZNF710, ZNF721, ZNF829
rs10281892	*DGKB*	0.20	0	0	0	0	0	1	0	0	MEF2C, NR5A2, ZNF526, ZNF730
rs62578804	*ZNF169*	0.20	0	0	0	0	0	0	0	1	FOXO1, FOXO3, PAX6, ZNF75A
rs867489	*CEBPB*	0.20	0	0	0	0	0	1	0	0	ZNF549
rs10407429	*GIPR*	0.20	0	0	1	0	0	1	0	0	RLF

rs10406431	*GIPR*	0.19	0	0	1	0	0	1	0	0	EBF1, EBF2, EBF4, PLAG1, PRDM10, PURA, ZNF480, ZNF766
rs329118	*PHF15*	0.19	1	0	1	0	0	0	1	0	ZSCAN25
rs13737	*PTPN9*	0.19	1	0	1	0	1	1	1	0	ATF6, CREBZF, CREB3, NR4A2, SMAD1, SMAD9, SMAD5, XBP1, ZNF354C
rs11155072	*MIR3668*	0.19	0	1	0	0	0	0	0	0	HNF4A
rs10408179	*GIPR*	0.17	0	0	1	0	0	1	0	0	CTCF, EBF1, EBF2, EBF4, ZNF319, ZNF443, ZNF697
rs144155527	*TCF7L2*	0.17	0	0	0	0	0	1	0	0	TGIF1, ZIC5
rs191304120	*HNF4A*	0.17	0	0	0	0	1	0	0	0	IKZF2, TP53
rs11885451	*THADA*	0.16	0	0	0	0	0	0	1	0	NR2F1
rs11614914	*FBRSL1*	0.16	0	0	0	0	0	0	1	0	EGR1, KLF12, KLF15, KLF16, KLF5, KLF6, KLF7, MAZ, NFKB1, PATZ1, SOX2, SP1, SP2, SP3, SP4, SP5, ZNF148, ZNF219, ZNF281
rs329120	*PHF15*	0.16	1	0	1	0	0	0	1	0	NFIB, SOX17, ZNF236
rs33320	*ANKRD55*	0.16	1	0	0	1	1	1	1	1	FIZ1, SIX4, ZNF248, ZNF836, ZNF841
rs7109575	*CENTD2/ARAP1*	0.14	0	0	1	0	0	0	0	0	ARNT2, EGR1
rs896853	*TP53INP1*	0.14	0	0	0	1	1	1	1	0	HIVEP2, PURA
rs2879813	*TP53INP1*	0.14	0	0	0	1	1	1	1	0	CTCF, PLAG1
rs7679066	*LCORL*	0.14	0	0	0	0	1	0	0	1	FOXO1, RREB1, ZNF22, ZNF467
rs11717959	*IGF2BP2*	0.14	1	1	1	1	1	1	1	1	ZNF550, ZNF594
rs11072553	*PTPN9*	0.14	0	0	0	0	1	1	1	0	RCOR2, MYB, S PC4, RCOR3
rs11263763	*HNF1B*	0.13	0	0	1	1	1	1	1	1	KLF6, ZNF213
rs13257021	*TP53INP1*	0.13	0	0	0	0	0	0	1	0	
rs569255	*SLC12A8*	0.13	0	1	0	0	0	0	0	0	ZBTB16, ZNF189, ZNF513
rs11065280	*HNF1A*	0.13	0	0	0	0	0	1	0	0	ZFP41, ZNF335
rs11981126	*IGF2BP3*	0.13	0	0	0	0	0	0	1	0	HIVEP2, NFATC4, ZNF148, ZNF581
rs150318646	*PCGF3*	0.13	0	0	1	0	0	0	0	0	EBF1, EBF2, EBF4, GTF3A, ZIC5
rs73322872	*TCEA2*	0.12	0	0	1	1	1	0	0	0	OSR1, THRA
rs6817471	*SCD5*	0.12	0	1	0	0	0	0	1	0	CREB3L2, KLF12, MEIS3, PLAGL2, TGIF2, ZNF730
rs12428731	*SPRY2*	0.12	0	1	1	0	0	0	0	0	ESRRA, MEIS1, NFIB, ZNF467

rs10750397	*ETS1*	0.12	0	0	0	0	0	0	1	0	POU3F1, ZFP30, ZNF189, ZNF287, ZNF394, ZNF610, ZNF7, ZNF700, ZNF787, ZXDC
rs2900660	*FARSA*	0.12	1	1	1	0	1	1	1	0	BACH1, MAFG, ZNF554, ZNF672, ZNF736
rs7667864	*LCORL*	0.12	1	0	1	1	1	1	1	0	FOXR1, FOXK1, FOXP1, FOXP2, KLF3
rs252924	*ANKRD55*	0.12	0	0	1	1	1	1	1	0	E2F1, E2F4, E2F6
rs11759956	*RREB1*	0.12	0	0	0	0	1	0	0	0	
rs703976	*ZMIZ1*	0.12	0	0	0	0	1	0	1	0	MAZ, MECOM, SP1, SP2, YBX3, YBX1, ZNF263, ZNF281
rs6937438	*VEGFA*	0.11	0	0	0	0	0	1	1	1	ZNF114, ZNF426
rs703977	*ZMIZ1*	0.11	0	0	0	0	0	0	1	1	ASCL2, ATOH1, ID4, ASCL5, ASCL1, MYF5, NFATC3, NKX6-1, TCF3, ZBTB40, ZNF121, ZNF287, ZNF775
rs35845603	*GIPR*	0.11	0	1	1	1	1	1	1	1	TP53, ZNF577
rs9860221	*ADAMTS9*	0.11	0	1	0	0	0	0	0	0	BACH1, JUND, MECOM, NFE2L2, SIX4, ZNF273, ZNF660, ZNF730
rs557408739	*RASGRP1*	0.11	1	0	0	0	0	0	0	0	FOXJ2, FOXJ1, FOXN2, FOXN3, FOXJ3, FOXP1, HNF1A, HNF1B, LMX1B, MECOM, PBX2, PBX4, HOPX, PROP1, ZFHX4, ZFHX3, ZNF525
rs28536554	*RASGRP1*	0.11	1	0	0	0	0	0	0	0	FOXJ2, FOXJ1, FOXN2, FOXN3, FOXJ3, FOXP1, HNF1A, HNF1B, LMX1B, MECOM, PBX2, PBX4, HOPX, PROP1, ZFHX4, ZFHX3, ZNF525
rs1415287	*LYPLAL1*	0.11	0	0	0	0	0	1	0	0	MAFF, NFE2L2, ZNF214
rs1415288	*LYPLAL1*	0.11	0	0	0	0	0	1	0	0	MAFF, MAFK, NOG, ZNF768
rs11856307	*C2CD4A/B*	0.11	0	0	1	0	0	0	0	0	PATZ1, ZNF22, ZNF354A, ZNF354B, ZSCAN23
rs11196183	*TCF7L2*	0.11	1	0	0	0	0	0	0	0	UBP1, TFCP2L1, ZNF200
rs73224254	*HNF1A*	0.10	0	0	1	0	0	0	0	0	KLF12, SMAD2, ZIC5, ZNF260
rs111775768	*HNF1A*	0.10	1	0	1	0	1	1	1	1	
rs4476672	*WFS1*	0.10	0	0	0	0	0	1	1	1	ELF1, ZNF749

Appendix Table 3.3: Hypomethylated DMRs overlapping credible set variants > 10% PPAg. PPAg: Genetic posterior probability of association. DMR presence (1) or absence (0) at the SNP position per contrast against baseline (iPSC for differentiating cells, BLC for adult islets). Notice that DMRs with different limits are not merged across contrasts. TF: transcription factor binding motif at SNP position.

			Hypomethylated DMR contrasts								
rsID	**locus**	**PPAg**	**DE - iPSC**	**GT - iPSC**	**PF - iPSC**	**PE - iPSC**	**EP - iPSC**	**EN - iPSC**	**BLC - iPSC**	**islet - BLC**	**TF**
rs2066827	*CDKN1B*	0.98	0	0	1	1	0	1	0	0	CREM, CREBL2, ELF1, ELK3, ETS1, ETV5, FEV, ZNF594
rs555759341	*INS/IGF2*	0.94	0	0	0	0	0	0	0	1	PLAGL2, ZIC5, ZNF140, ZNF398, ZNF432, ZNF692
rs72631105	*WDR11*	0.81	0	0	0	0	0	0	0	1	FOXO1, FOXO3, HIVEP2, RLF, STAT2, ZNF200, ZNF22, ZNF263, ZNF287, ZNF347, ZNF721
rs113414093	*CRYBA2*	0.63	0	0	0	0	0	0	0	1	MBD2
rs56158042	*HNF1A*	0.40	0	0	0	1	0	0	0	0	STAT5B, ZNF574
rs2233632	*HMGA1*	0.33	0	1	1	1	1	1	1	0	HIVEP2, NFATC4
rs2233631	*HMGA1*	0.33	0	1	1	1	1	1	1	0	HLTF, RFX1, SMAD1, SMAD9, SMAD5, ZNF700, ZNF730
rs78840640	*IGF2BP3*	0.33	0	0	0	0	0	0	1	1	HIC2, MAFG, MAFK, STAT3, STAT4, ZNF248, ZNF860
rs78840640	*IGF2BP3*	0.33	0	0	0	0	0	1	0	0	HIC2, MAFG, MAFK, STAT3, STAT4, ZNF248, ZNF860
rs5215	*KCNJ11*	0.33	0	0	0	0	0	0	0	1	FOXO3, ZSCAN10
rs1642762	*ATP1B2*	0.32	0	0	0	0	0	1	0	0	EGR1, ZNF766
rs1642762	*ATP1B2*	0.32	0	0	0	1	0	0	0	0	EGR1, ZNF766
rs2115107	*MAP2K7*	0.31	0	0	0	0	0	0	0	1	INSM1, MAZ, MZF1, NR2F6, PATZ1
rs5219	*KCNJ11*	0.28	0	0	0	0	0	0	0	1	ZNF467
rs7732130	*ZBED3*	0.28	0	0	1	0	0	0	0	0	RFX2, RFX5, SMAD4, SMAD6, SREBF1, SREBF2, ZNF302, ZNF526, ZNF853, ZSCAN25
rs7732130	*ZBED3*	0.28	0	0	0	1	0	1	1	0	RFX2, RFX5, SMAD4, SMAD6, SREBF1, SREBF2, ZNF302, ZNF526, ZNF853, ZSCAN25

rs1624085	*ATP1B2*	0.22	0	0	0	0	0	1	0	0	EBF1, EBF2, EBF4, GLI3, NR2F1, PPARD, ZNF586, ZNF746
rs1624085	*ATP1B2*	0.22	0	0	0	0	0	0	1	0	EBF1, EBF2, EBF4, GLI3, NR2F1, PPARD, ZNF586, ZNF746
rs4417225	*INS/IGF2*	0.21	0	0	0	0	0	0	0	1	HNF4A, IRF6, NR2F2, PPARD, RARG, STAT2, ZNF273
rs4417225	*INS/IGF2*	0.21	0	0	0	0	1	0	0	0	HNF4A, IRF6, NR2F2, PPARD, RARG, STAT2, ZNF273
rs4417225	*INS/IGF2*	0.21	0	0	1	1	0	1	1	0	HNF4A, IRF6, NR2F2, PPARD, RARG, STAT2, ZNF273
rs10305745	*FAM63A*	0.18	0	0	1	1	1	1	1	0	HOXA3, POU3F2, ZKSCAN2, ZNF518B, ZNF718
rs286925	*PDHX*	0.17	0	0	1	1	0	0	0	1	ZIC5, ZNF143
rs286925	*PDHX*	0.17	0	0	0	0	1	1	1	0	ZIC5, ZNF143
rs33320	*ANKRD55*	0.16	0	0	0	0	0	0	0	1	FIZ1, SIX4, ZNF248, ZNF836, ZNF841
rs59944054	*TCEA2*	0.15	0	0	0	0	0	0	0	1	HINFP, KLF5, SP2, ZNF316
rs57642905	*TCEA2*	0.14	0	0	0	0	0	0	0	1	ZNF667, ZNF732
rs11717959	*IGF2BP2*	0.14	0	0	0	0	0	1	0	0	ZNF550, ZNF594
rs11717959	*IGF2BP2*	0.14	0	0	0	1	0	0	0	0	ZNF550, ZNF594
rs17036160	*PPARG*	0.13	0	0	0	1	0	1	0	0	ZNF550, ZNF641, ZNF682
rs150318646	*PCGF3*	0.13	0	0	0	0	0	0	0	1	EBF1, EBF2, EBF4, GTF3A, ZIC5
rs73322872	*TCEA2*	0.12	0	0	0	0	0	0	0	1	OSR1, THRA
rs1882297	*FBRSL1*	0.12	0	0	0	0	0	0	0	1	FOXO3, PURA, SIX4, ZNF200, ZNF616
rs7667864	*LCORL*	0.12	0	0	0	0	0	0	0	1	FOXR1, FOXK1, FOXP1, FOXP2, KLF3
rs112314659	*FARSA*	0.12	0	0	0	0	0	1	0	0	GLI3, GTF3A, PURA, ZNF468, ZNF669, ZNF697
rs2242517	*FARSA*	0.12	0	0	0	0	0	1	0	0	GLI3, GTF3A, PURA, ZNF468, ZNF669, ZNF697
rs112314659	*FARSA*	0.12	0	0	0	1	0	0	0	0	GLI3, GTF3A, PURA, ZNF468, ZNF669, ZNF697
rs2242517	*FARSA*	0.12	0	0	0	1	0	0	0	0	GLI3, GTF3A, PURA, ZNF468, ZNF669, ZNF697
rs112314659	*FARSA*	0.12	0	0	0	0	0	0	1	0	GLI3, GTF3A, PURA, ZNF468, ZNF669, ZNF697
rs2242517	*FARSA*	0.12	0	0	0	0	0	0	1	0	GLI3, GTF3A, PURA, ZNF468, ZNF669, ZNF697

rs112314659	*FARSA*	0.12	0	0	0	0	1	0	0	1	GLI3, GTF3A, PURA, ZNF468, ZNF669, ZNF697
rs2242517	*FARSA*	0.12	0	0	0	0	1	0	0	1	GLI3, GTF3A, PURA, ZNF468, ZNF669, ZNF697
rs112314659	*FARSA*	0.12	0	0	1	0	0	0	0	0	GLI3, GTF3A, PURA, ZNF468, ZNF669, ZNF697
rs2242517	*FARSA*	0.12	0	0	1	0	0	0	0	0	GLI3, GTF3A, PURA, ZNF468, ZNF669, ZNF697
rs12789028	*MAP3K11*	0.11	0	0	0	0	0	0	0	1	FIZ1, HIVEP2, ZNF141, ZNF263, ZNF467, ZNF480
rs111775768	*HNF1A*	0.10	0	0	1	1	0	1	1	0	NA

Appendix Table 3.4: Variants with functional PPAs over 50%. PPAs from the credible sets were re-weighted with the module estimates of the best fGWAS joint model containing the blue, yellow, green and brown modules. PPAg: genetic PPA. PPAf: functional PPA. In bold: variants and loci highlighted in the text.

locus	ID	PPAg	PPAf	module	PPAf-PPAg
KCNQ1	rs231361	1	1	blue	0.00
MTNR1B	rs10830963	1	1	not enriched	0.00
CCND2	rs76895963	1	1	not enriched	0.00
HNF4A	rs1800961	1	1	not enriched	0.00
PAM	rs78408340	1	1	not enriched	0.00
BCL2A	rs12454712	1	1	not enriched	0.00
CDC123.CAMK1D	rs11257655	1	1	not enriched	0.00
ABCB10	rs348330	1	1	not enriched	0.00
CASC11	rs17772814	1	1	brown	0.00
ANKH	rs76549217	1	1	not enriched	0.00
FAM49A	rs11680058	0.99	1	blue	0.01
PROX1	rs340874	0.99	0.99	not enriched	0.00
KCNQ1	rs2237895	0.99	0.99	not enriched	0.00
CCND2	rs3217792	0.99	0.99	not enriched	0.00
CCND2	rs796464202	0.99	0.99	not enriched	0.00
PATJ	rs12140153	0.99	0.99	not enriched	0.00
HNF1A	rs1800574	0.98	0.99	yellow	0.01
WDR72	rs528350911	0.99	0.99	yellow	0.00
IGF2BP2	rs150111048	0.94	0.99	yellow	0.05
HTT	rs362307	0.94	0.99	blue	0.05
RREB1	rs9379084	0.99	0.99	not enriched	0.00
TCF7L2	rs140242150	0.99	0.99	not enriched	0.00
CCND2	rs3217860	0.99	0.99	not enriched	0.00
CENPW	rs11759026	0.99	0.99	not enriched	0.00

GLI2	rs11688682	0.98	0.98	not enriched	0.00
ST6GAL1	rs3887925	0.98	0.98	not enriched	0.00
LRFN2	rs34298980	0.99	0.98	not enriched	-0.01
HNF1B	rs10962	0.98	0.98	not enriched	0.00
GCK	rs878521	0.98	0.98	not enriched	0.00
HNF4A	rs76811102	0.96	0.97	not enriched	0.01
CRHR2	rs917195	0.96	0.97	not enriched	0.01
CDKN1B	rs2066827	0.98	0.97	not enriched	-0.01
GLIS3	rs10974438	0.97	0.97	not enriched	0.00
ANK1	rs13262861	0.97	0.97	not enriched	0.00
DGKB	rs17168486	0.57	0.96	yellow	0.39
PTGFRN	rs1127215	0.97	0.96	not enriched	-0.01
CDKN2A.B	rs76011118	0.95	0.95	not enriched	0.00
CMIP	rs2925979	0.96	0.95	not enriched	-0.01
INS.IGF2	rs555759341	0.95	0.95	not enriched	0.00
CDKN2A.B	rs10757282	0.34	0.94	blue	0.60
CLUAP1	rs3751837	0.90	0.92	not enriched	0.02
ATP1B2	rs7503751	0.31	0.90	yellow	0.59
KCNQ1	rs2283164	0.89	0.90	not enriched	0.01
SLC30A8	rs13266634	0.35	0.90	green	0.55
CPQ	rs149364428	0.87	0.90	not enriched	0.03
TOMM40.APOE	rs429358	0.89	0.89	not enriched	0.00
GNG4	rs291367	0.24	0.89	green	0.65
TSC22D2	rs62271373	0.95	0.88	not enriched	-0.07
UBAP2	rs12001437	0.91	0.87	not enriched	-0.05
FTO	rs1421085	0.28	0.87	green	0.59
GCKR	rs1260326	0.89	0.86	not enriched	-0.03
CACNA2D3	rs76263492	0.86	0.86	not enriched	0.00

TLE1	rs2796441	0.82	0.85	not enriched	0.03
NEUROG3	rs177045	0.85	0.85	not enriched	0.00
MAP2K5	rs4776970	0.75	0.85	not enriched	0.10
ANKH	rs17250977	0.84	0.84	not enriched	0.00
ZBED3	rs7732130	0.29	0.83	blue	0.55
FAM63A	rs145904381	0.84	0.83	not enriched	-0.01
KCNQ1	rs2237897	0.81	0.81	yellow	0.00
DLEU1	rs963740	0.50	0.81	yellow	0.31
ANKRD55	rs252924	0.12	0.79	blue	0.67
TLE4	rs17791513	0.71	0.78	not enriched	0.07
TCF7L2	rs34855922	0.78	0.78	not enriched	0.00
ARL15	rs55744247	0.15	0.78	yellow	0.63
ANKH	rs146886108	0.77	0.77	not enriched	0.00
GIPR	rs2238689	0.78	0.76	not enriched	-0.02
TTLL6	rs35895680	0.66	0.74	not enriched	0.08
AOC1	rs62492368	0.75	0.73	not enriched	-0.02
RASGRP1	rs34715063	0.73	0.73	not enriched	0.00
MAEA	rs56337234	0.72	0.73	not enriched	0.01
NEUROG3	rs41277236	0.74	0.73	not enriched	-0.01
MACF1	rs3768321	0.66	0.73	not enriched	0.07
WDR11	rs72631105	0.80	0.72	not enriched	-0.08
INS.IGF2	rs11042596	0.72	0.72	not enriched	0.00
TCF7L2	rs7918400	0.71	0.71	not enriched	0.00
CTTNBP2	rs6976111	0.65	0.70	not enriched	0.05
SPG7	rs12920022	0.73	0.70	not enriched	-0.03
IGF2BP3	rs78840640	0.33	0.69	blue	0.36
ROBO2	rs2272163	0.62	0.68	not enriched	0.06
KCNQ1	rs80102379	0.68	0.68	not enriched	0.00

ANKH	rs6885132	0.68	0.67	not enriched	-0.01
ANKRD55	rs256904	0.13	0.67	green	0.54
FOCAD	rs7867635	0.65	0.66	not enriched	0.01
FAM57B	rs199795270	0.66	0.66	not enriched	0.00
LYPLAL1	rs2820446	0.21	0.66	blue	0.45
PSMD6	rs74368513	0.65	0.66	not enriched	0.01
BCAR1	rs72802342	0.66	0.66	not enriched	0.00
PROX1	rs79687284	0.64	0.64	brown	0.00
IGF2BP2	rs869399	0.05	0.64	blue	0.59
VEGFA	rs11967262	0.64	0.64	not enriched	0.00
FBRSL1	rs77773608	0.13	0.63	blue	0.50
RNF6	rs34584161	0.67	0.63	not enriched	-0.04
RBM6	rs4688760	0.69	0.63	not enriched	-0.06
QSER1	rs145678014	0.64	0.63	not enriched	-0.01
HNF1B	rs10908278	0.71	0.62	not enriched	-0.09
PDHX	rs286925	0.17	0.62	yellow	0.45
TMEM18	rs62107261	0.62	0.62	not enriched	0.00
ADCY5	rs11708067	0.79	0.61	not enriched	-0.18
TCF7L2	rs7903146	0.59	0.59	not enriched	0.00
MAP2K7	rs4804833	0.53	0.59	not enriched	0.06
BCL11A	rs243018	0.11	0.57	blue	0.46
GRB14.COBLL1	rs13024606	0.58	0.57	not enriched	-0.01
SHQ1	rs13080180	0.27	0.57	brown	0.30
VEGFA	rs6937438	0.12	0.57	green	0.45
ZNF169	rs12236906	0.60	0.57	not enriched	-0.03
ZZEF1	rs1043246	0.56	0.57	not enriched	0.01
CCND2	rs4238013	0.54	0.54	not enriched	0.00
PIK3R1	rs4976033	0.51	0.54	not enriched	0.03
SLC30A8	rs80244329	0.49	0.54	not enriched	0.04

HNF4A	rs191830490	0.52	0.54	not enriched	0.02
SLCO6A1	rs114546213	0.63	0.53	not enriched	-0.10
CDKN2A.B	rs10811660	0.47	0.53	blue	0.06
NF1	rs71372253	0.22	0.53	not enriched	0.31
IGF2BP3	rs4279506	0.52	0.52	not enriched	0.00
LCORL	rs7667864	0.12	0.52	blue	0.40
WSCD2	rs1426371	0.54	0.51	not enriched	-0.03
POC5	rs2307111	0.63	0.51	not enriched	-0.12
HMGB1	rs1056130	0.08	0.51	blue	0.43
PURG	rs10954772	0.56	0.51	not enriched	-0.05

Appendix Table 4.1. Annotated T2D-associated variants with functional PPA > 10% in ABC-predicted enhancers. Each row represents a T2D-associated variant within an ABC-predicted active enhancer. Columns represent (left to right) the variant position, SNP ID, credible set locus, index variant position at that locus, stages of beta cell development (iPSC to BLC) in which the overlapping enhancer is active (1) or not (0), genetic posterior probability of association (PPAg), functional (re-weighted by fGWAS) posterior probability of association (PPAf), difference between them, predicted target genes by the ABC model for the enhancers overlapping the variants, closest genes to those enhancers, and transcription factor whose binding motif is disrupted by a variant allele ("NA" if no TF binding motif tested is disrupted.

Position	SNP ID	Locus	iPSC	DE	GT	PF	PE	EP	EN	BLC	PPAg	PPAf	PPAf-PPAg	Target Genes	Closest Gene	TF disrupted
11:26915 00	rs23136 1	*KCNQ1*	0	0	0	0	0	0	1	0	1	1	0	KCNQ1OT1	KCNQ1OT 1	NA
2:165746 69	rs11680 058	*FAM49A*	0	0	0	0	0	1	1	1	1	1	0	MYCN, FAM49A	FAM49A	NA
15:53747 228	rs52835 0911	*WDR72*	0	0	0	1	1	1	0	0	0.99	0.99	0.00	WDR72	WDR72	NA
7:307284 52	rs91719 5	*CRHR2*	0	0	0	1	0	0	0	0	0.96	0.98	0.02	INMT	CRHR2	Rfx1,RFX2
4:324184 5	rs36230 7	*HTT*	0	0	0	0	0	0	1	0	0.95	0.97	0.02	RGS12, HGFAC	MSANTD1	NA
11:21188 60	rs11042 596	*INS.IGF2*	1	1	0	1	1	0	0	0	0.71	0.95	0.24	H19, IGF2,IGF2-AS	MIR483	NA
8:118184 783	rs13266 634	*SLC30A8*	0	1	0	0	1	1	1	1	0.35	0.88	0.53	AARD, SLC30A8	SLC30A8	PAX5
7:148982 82	rs17168 486	*DGKB*	0	0	0	0	0	0	1	0	0.58	0.88	0.30	DGKB	DGKB	NFE2L1_MafG
20:43023 355	rs19183 0490	*HNF4A*	0	0	1	1	0	0	0	0	0.52	0.87	0.35	TTPAL	HNF4A-AS1	NA
17:77401 70	rs62059 712	*ATP1B2*	1	1	1	1	1	1	1	1	0.34	0.80	0.46	EFNB3, KDM6B, TMEM88, NAA38, CYB5D1, CHD3, KCNAB3, TRAPPC1, CNTROB	KDM6B	NA
17:75537 43	rs16240 85	*ATP1B2*	1	1	1	1	1	1	1	1	0.22	0.80	0.58	SAT2,ATP1B2,SHBG	ATP1B2	NA
5:527745 10	rs62370 480	*ARL15*	1	1	1	1	1	0	1	0	0.20	0.77	0.57	FST,NDUFS4	FST	NA

12:133070294	rs11614914	*FBRSL1*	0	0	1	1	1	1	1	1	0.16	0.76	0.60	FBRSL1,GALNT9,P2RX2	FBRSL1	SP1, SP2, EGR1, KLF5
11:72463435	rs7109575	*CENTD2. ARAP1*	1	1	1	1	1	1	1	1	0.14	0.70	0.56	ARAP1,STARD10,ATG16L2	ARAP1	NA
11:65307402	rs11227217	*MAP3K11*	1	1	1	1	1	1	1	1	0.11	0.67	0.56	SCYL1, SSSCA1-AS1, SSSCA1, FAM89B, EHBP1L1, LTBP3	SCYL1	EGR1,SP2
1:214150821	rs79687284	*PROX1*	0	1	1	1	1	0	0	0	0.64	0.64	0.00	PROX1,PROX1-AS1	PROX1-AS1	NA
10:80944230	rs703977	*ZMIZ1*	0	0	0	0	0	0	1	0	0.11	0.63	0.52	ZMIZ1	ZMIZ1	Pdx1,Prrx2
13:31019580	rs73167313	*HMGB1*	1	1	1	1	1	1	1	1	0.08	0.63	0.55	HMGB1,LINC01058,USPL1	HMGB1	NA
19:46148903	rs35816837	*GIPR*	0	0	0	0	0	1	1	1	0.09	0.58	0.49	VASP,OPA3,EML2,GIPR	EML2	ZNF263
11:17408630	rs5215	*KCNJ11*	0	0	1	1	1	0	1	0	0.33	0.57	0.24	NCR3LG1,KCNJ11	KCNJ11	NA
2:43206922	rs921069	*THADA*	0	1	1	1	1	1	1	1	0.06	0.51	0.47	ZFP36L2,THADA	LINC01819	NA
16:53544993	rs3095571	*FTO*	0	1	1	1	1	1	1	1	0.03	0.45	0.42	RBL2, AKTIP, FTO-IT1, RPGRIP1L, FTO, IRX3	AKTIP	NA
10:93999850	rs3824734	*HHEX.IDE*	1	0	1	1	1	1	1	1	0.03	0.42	0.39	CPEB3,MARCH5	CPEB3	NA
7:130013638	rs1465370	*KLF14*	0	0	0	1	1	1	1	1	0.03	0.42	0.39	KLHDC10, TMEM209, CEP41, MEST, ZC3HC1, CPA2, CPA4, CPA1, COPG2	LOC105375504, CPA1	NA
15:53091553	rs2456530	*ONECUT1*	1	0	0	1	1	1	0	0	0.36	0.40	0.04	ONECUT1,WDR72	ONECUT1	ZNF263
11:17408404	rs5213	*KCNJ11*	0	0	1	1	1	0	1	0	0.32	0.39	0.07	NCR3LG1,KCNJ11	KCNJ11	Zfx
3:63897915	rs6785040	*PSMD6*	1	1	1	1	1	1	1	1	0.06	0.39	0.33	SYNPR-AS1, THOC7, ATXN7, PSMD6-AS2, PSMD6, PRICKLE2, SYNPR, LINC00698, LINC00994	SCAANT1	NA
11:128231566	rs1125251	*ETS1*	0	0	0	1	1	1	1	1	0.10	0.38	0.28	ETS1	LINC02098	Bach1_Mafk
1:214150445	rs17712208	*PROX1*	0	1	1	1	1	0	0	0	0.36	0.36	0.00	PROX1,PROX1-AS1	PROX1-AS1	HNF1A, HNF1B, Prrx2
10:114737633	rs144155527	*TCF7L2*	0	0	0	0	0	0	1	1	0.17	0.35	0.18	TCF7L2	TCF7L2	NA

11:128231469	rs1125253	*ETS1*	0	0	0	1	1	1	1	1	0.09	0.34	0.25	ETS1	LINC02098	NA
2:161126732	rs764729	*RBMS1*	1	1	1	1	1	1	1	1	0.03	0.33	0.30	PLA2R1,TANK,ITGB6,RBMS1	LINC02478	NA
19:33894615	rs10415756	*PEPD*	0	0	0	1	1	0	0	0	0.08	0.33	0.25	SLC7A10,CEBPA-AS1,CEBPG	CEBPG	NA
9:20692438	rs10738564	*FOCAD*	0	0	1	1	0	0	0	0	0.08	0.30	0.22	FOCAD	FOCAD-AS1	NA
2:65663167	rs1807337	*CEP68*	1	1	0	1	1	1	1	1	0.04	0.30	0.26	CEP68,RAB1A,ACTR2,SPRED2,SLC1A4	SPRED2	NA
3:152025796	rs6807753	*MBNL1*	0	0	0	1	1	0	1	0	0.03	0.25	0.22	MBNL1	MBNL1	NA
11:69454276	rs36225073	*CCND1*	1	1	1	1	1	1	1	0	0.02	0.23	0.21	CCND1,ORAOV1,FGF19	CCND1	NA
6:34192036	rs114760566	*HMGA1*	1	1	1	1	1	1	1	1	0.04	0.23	0.19	HMGA1,C6orf1,NUDT3,RPS10,GRM4,PACSIN1	HMGA1	NA
12:118407351	rs55834317	*KSR2*	1	1	1	1	1	1	1	1	0.04	0.22	0.18	KSR2, RFC5, WSB2, VSIG10, TESC, NOS1, PEBP1, MED13L	KSR2	NA
11:1703564	rs4417225	*INS.IGF2*	0	0	0	0	1	0	0	0	0.21	0.21	0.00	FAM99B	FAM99B	NA
12:118406696	rs79310463	*KSR2*	1	1	1	1	1	1	1	1	0.03	0.20	0.17	KSR2, RFC5, WSB2, VSIG10, TESC, NOS1, PEBP1, MED13L	KSR2	NA
2:25533568	rs55928417	*DTNB*	1	0	1	1	0	0	0	0	0.03	0.19	0.16	DNMT3A	MIR1301	REST,MZF1_1-4
2:220265302	rs146452383	*CRYBA2*	0	0	1	1	0	0	0	0	0.00	0.18	0.18	DNPEP	DNPEP	NA
11:45870177	rs10838524	*CRY2*	0	0	0	0	1	0	0	0	0.16	0.18	0.02	CRY2	CRY2	FOXO3
17:47039302	rs8081301	*TTLL6*	1	1	1	1	1	0	1	0	0.03	0.17	0.14	SNF8, IGF2BP1, UBE2Z, GIP, B4GALNT2	GIP	NA
17:7789542	rs7503751	*ATP1B2*	0	0	0	0	0	0	1	0	0.31	0.16	-0.15	CHD3	NAA38	NA
20:62372148	rs6062510	*ZBTB46*	1	0	1	1	1	1	1	1	0.01	0.15	0.14	ARFRP1, ZGPAT, LIME1, SLC2A4RG, ZBTB46	SLC2A4RG	NA
7:156944460	rs3808317	*MNX1*	0	0	0	1	0	0	1	1	0.01	0.14	0.13	UBE3C	UBE3C	NA
20:62372706	rs4809221	*ZBTB46*	1	0	1	1	1	1	1	1	0.01	0.14	0.13	ARFRP1, ZGPAT, LIME1, SLC2A4RG, ZBTB46	SLC2A4RG	SP2,SP1,EGR1,ZNF263,Klf4,KLF5

17:36103565	rs11263763	*HNF1B*	0	0	0	0	1	0	0	0	0.13	0.13	0.00	DDX52, HNF1B	HNF1B	NA
5:52777864	rs55744247	*ARL15*	0	0	0	0	1	0	1	0	0.15	0.12	-0.03	FST	FST	SPIB
12:121260719	rs11065280	*HNF1A*	0	1	0	0	0	0	0	0	0.13	0.12	-0.01	SPPL3	SPPL3	NA
6:34202879	rs55692303	*HMGA1*	1	1	1	1	1	1	1	0	0.02	0.11	0.09	LEMD2, GRM4, HMGA1, C6orf1, NUDT3, RPS10, PACSIN1, C6orf106, SNRPC, UHRF1BP1, TAF11, ANKS1A, ZBTB9, ITPR3, UQCC2, SPDEF, IP6K3, MLN	HMGA1	E2F1,E2F4
15:53082481	rs2456523	*ONECUT1*	1	1	1	1	1	1	1	1	0.01	0.11	0.10	ONECUT1, WDR72, UNC13C	ONECUT1	NA
22:32366415	rs4820060	*YWHAH*	1	1	1	1	1	1	1	1	0.01	0.11	0.10	PISD, C22orf24, YWHAH, C22orf42, RFPL2, SLC5A4, RFPL3S, RTCB	LINC02558	Zfx
15:38846678	rs7170151	*RASGRP1*	0	0	0	0	0	0	1	0	0.02	0.10	0.08	RASGRP1	RASGRP1	NA

Bibliography

Ackermann AM, Wang Z, Schug J, et al (2016) Integration of ATAC-seq and RNA-seq identifies human alpha cell and beta cell signature genes. Mol Metab 5:233–244.

Ahsendorf, T., Müller, F.-J., Topkar, V., Gunawardena, J., & Eils, R. (2017). Transcription factors, coregulators, and epigenetic marks are linearly correlated and high redundant. PLOS ONE, 12(12), e0186324.

Akalin, A., Kormaksson, M., Li, S., Garrett-Bakelman, F. E., Figueroa, M. E., Melnick, A., & Mason, C. E. (2012). MethylKit: a comprehensive R package for the analysis of genome-wide DNA methylation profiles. Genome Biology, 13(10).

Aken BL, Ayling S, Barrell D, et al (2016) The Ensembl gene annotation system. Database 2016:baw093.

Alasoo, K., Rodrigues, J., Mukhopadhyay, S., Knights, A. J., Mann, A. L., Kundu, K., … Gaffney, D. J. (2018). Shared genetic effects on chromatin and gene expression indicate a role for enhancer priming in immune response. *Nature Genetics*, *50*(3), 424–431.

Aloia, L. et al. (2013) Polycomb complexes in stem cells and embryonic development. Development 140, 2525–2534

Alvarez-Dominguez JR, Donaghey J, Kenty JHR, Rasouli N (2019) Dissecting mechanisms of human islet differentiation and maturation through epigenome profiling.

Andrey, G., & Mundlos, S. (2017). The three-dimensional genome: Regulating gene expression during pluripotency and development. *Development (Cambridge)*, *144*(20), 3646–3658. https://doi.org/10.1242/dev.148304

Antequera, F. and Bird, A. 1993. Number of CpG islands and genes in human and mouse. Proc. Natl. Acad. Sci. 90: 11995–11999.

Aronoff, S. L., Berkowitz, K., Shreiner, B., & Want, L. (2004). Glucose Metabolism and Regulation: Beyond Insulin and Glucagon. Diabetes Spectrum, 17(3), 183–190.

Arterbery AS, Bogue CW (2016) Hhex is necessary for the hepatic differentiation of mouse ES cells and acts via Vegf signaling. PLoS One 11:e0146806.

Aryee MJ, Jaffe AE, Corrada-Bravo H, Ladd-Acosta C, Feinberg AP, Hansen KD, Irizarry RA (2014). Minfi: A flexible and comprehensive Bioconductor package for the analysis of Infinium DNA Methylation microarrays.Bioinformatics, 30(10), 1363-1369.

Ashcroft, F.M. & Rorsman, P. Diabetes mellitus and the β cell: the last ten years. Cell 148, 1160–1171 (2012).

Association, A. D. (2018). Classification and diagnosis of diabetes: Standards of medical care in Diabetes. Diabetes Care, 41(January), S13–S27.

Atlasi, Y., & Stunnenberg, H. G. (2017). The interplay of epigenetic marks during stem cell differentiation and development. *Nature Reviews Genetics*, *18*(11), 643–658.

Atouf F, Czernichow P, Scharfmann R (1997) Expression of neuronal traits in

pancreatic beta cells. J Biol Chem 272:1929. doi: 10.1074/jbc.272.3.1929

Attali, M., Stetsyuk, V., Basmaciogullari, A., Aiello, V., Zanta-Boussif, M. A., Duvillie, B., & Scharfmann, R. (2007). Control of beta-cell differentiation by the pancreatic mesenchyme. Diabetes, 56(5), 1248–1258.

Axel R (1975) Cleavage of DNA in nuclei and chromatin with staphylococcal nuclease. Biochemistry 14:2921–2925

Bannister, A. J., & Kouzarides, T. (2011). Regulation of chromatin by histone modifications. Nature Publishing Group, 21(3), 381–395.

Banovich NE, Lan X, Mcvicker G, et al (2014) Methylation QTLs Are Associated with Coordinated Changes in Transcription Factor Binding , Histone Modifications , and Gene Expression Levels. 10:1–12.

Bar-Nur, O., Russ, H. A., Efrat, S., & Benvenisty, N. (2011). Epigenetic Memory and Preferential Lineage-Specific Differentiation in Induced Pluripotent Stem Cells Derived from Human Pancreatic Islet Beta Cells. Cell Stem Cell, 9(1), 17–23.

Basila, M., Kelley, M. L., & Smith, A. van B. (2017). Minimal 2'-O-methyl phosphorothioate linkage modification pattern of synthetic guide RNAs for increased stability and efficient CRISPR-Cas9 gene editing avoiding cellular toxicity. *PLOS ONE*, *12*(11), e0188593.

Baubec, T., Ivanek, R., Lienert, F. & Schubeler, D. Methylation-dependent and -independent genomic targeting principles of the MBD protein family. Cell 153, 480–492 (2013).

Baylin, S.B. and Herman, J.G. 2000. DNA hypermethylation in tumorigenesis: Epigenetics joins genetics. Trends Genet. 16: 168–174.

Beer NL, Gloyn AL (2016) Genome-edited human stem cell-derived beta cells: a powerful tool for drilling down on type 2 diabetes GWAS biology. F1000Research.

Beer NL, Tribble ND, McCulloch LJ, et al. The P446L variant in GCKR associated with fasting plasma glucose and triglyceride levels exerts its effect through increased glucokinase activity in liver. Hum Mol Genet. 2009;18:4081–8.

Bell, C. G., Xia, Y., Yuan, W., Gao, F., Ward, K., Roos, L., Spector, T. D. (2016). Novel regional age-associated DNA methylation changes within human common disease-associated loci. Genome Biology, 17(1), 193.

Bell, O., Tiwari, V. K., Thomä, N. H., & Schübeler, D. (2011). Determinants and dynamics of genome accessibility. *Nature Reviews Genetics*, *12*(8), 554–564.

Benjamini Y, Hochberg Y (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing. Source J R Stat Soc Ser B 57:289–300.

Benner, C., van der Meulen, T., Cacéres, E., Tigyi, K., Donaldson, C. J., Huising, M. O., … Pachter, L. (2014). The transcriptional landscape of mouse beta cells compared to human beta cells reveals notable species differences in long non-coding RNA and protein-coding gene expression. *BMC Genomics*, *15*(1), 620.

Berger SL, Kouzarides T, Shiekhattar R, Shilatifard A (2009) An operational definition of epigenetics. Genes Dev 23(7):781– 783.

Bernstein, B.E., Stamatoyannopoulos, J.A., Costello, J.F., et al. (2010) The NIH

Roadmap Epigenomics Mapping Consortium. Nature Biotechnology 28: 1045.

Bersanelli, M., Mosca, E., Remondini, D., Giampieri, E., Sala, C., Castellani, G., & Milanesi, L. (2016). Methods for the integration of multi-omics data: mathematical aspects. BMC Bioinformatics, 17(2).

Bird, A. P. CpG-rich islands and the function of DNA methylation. Nature 321, 209–213 (1986).

Bittel, D. C., & Butler, M. G. (2005). Prader–Willi syndrome: clinical genetics, cytogenetics and molecular biology. Expert Reviews in Molecular Medicine, 7(14).

Boj SF, Van Es JH, Huch M, et al (2012) Diabetes Risk Gene and Wnt Effector Tcf7l2/TCF4 Controls Hepatic Response to Perinatal and Adult Metabolic Demand.

Boller, S., Ramamoorthy, S., Akbas, D., Nechanitzky, R., Burger, L., Murr, R., … Grosschedl, R. (2016). Pioneering Activity of the C-Terminal Domain of EBF1 Shapes the Chromatin Landscape for B Cell Programming. *Immunity*, *44*(3), 527–541.

Bonnefond, A., Sand, O., Guerin, B., Durand, E., & Graeve, F. De. (2012). GATA6 inactivating mutations are associated with heart defects and, inconsistently, with pancreatic agenesis and diabetes. 2845–2847.

Borgel, J. et al. Targets and dynamics of promoter DNA methylation during early mouse development. Nature Genet. 42, 1093–1100 (2010)

Bort, R., Martinez-Barbera, J. P., Beddington, R. S. P., & Zaret, K. S. (2004). Hex homeobox gene-dependent tissue positioning is required for organogenesis of the ventral pancreas. *Development (Cambridge, England)*, *131*, 797–806.

Borup, M., Petersen, K., Azad, A., Ingvorsen, C., Hess, K., Hansson, M., … Honoré, C. (2017). Single-Cell Gene Expression Analysis of a Human ESC Model of Pancreatic Endocrine Development Reveals Different Paths to b-Cell Differentiation. *Stem Cell Reports*.

Boschmann, M., Engeli, S., Moro, C., Luedtke, A., Adams, F., Gorzelniak, K., … Jordan J (2010) LMNA Mutations, Skeletal Muscle Lipid Metabolism, and Insulin Resistance. J Clin Endocrinol Metab 95:1634–1643.

Bourc'his, D., Xu, G. L., Lin, C. S., Bollman, B. & Bestor, T. H. Dnmt3L and the establishment of maternal genomic imprints. Science 294, 2536–2539 (2001).

Boutant M, Henrique Pereira Ramos O, cile Lecoeur C, et al Glucose-Dependent Regulation of NR2F2 Promoter and Influence of SNP-rs3743462 on Whole Body Insulin Sensitivity.

Bruin JE, Erener S, Vela J, et al (2014) Characterization of polyhormonal insulin-producing cells derived in vitro from human embryonic stem cells. Stem Cell Res 12:194–208.

Brunner, A. L., D. S. Johnson, S. W. Kim, et al. (2009). Distinct DNA methylation patterns characterize differentiated human embryonic stem cells and developing human fetal liver. Genome Res. 19:1044–1056.

Buenrostro JD, Giresi PG, Zaba LC, Chang HY, Greenleaf WJ (2013). Transposition of

native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position Nat Methods. 10(12):1213-8.

Buescher JM & Driggers EM 2016 Integration of omics: more than the sum of its parts. Cancer and Metabolism 4 1.

Burger, L., Gaidatzis, D., Schubeler, D. & Stadler, M. B. Identification of active regulatory regions from DNA methylation data. Nucleic Acids Res. 41, e155 (2013).

Burke, Z., & Oliver, G. (2002). Prox1 is an early specific marker for the developing liver and pancreas in the mammalian foregut endoderm. *Mechanisms of Development*, *118*(1–2), 147–155.

Burlison, J.S., Long, Q., Fujitani, Y., Wright, C.V., and Magnuson, M.A. (2008). Pdx-1 and Ptf1a concurrently determine fate specifi- cation of pancreatic multipotent progenitor cells. Dev. Biol. 316, 74–86.

Cadigan KM, Waterman ML (2012) TCF/LEFs and Wnt Signaling in the Nucleus. doi: 10.1101/cshperspect.a007906

Calviello, AK., Hirsekorn, A., Wurmus, R., Yusuf, D., & Ohler, U. (2019). Reproducible inference of transcription factor footprints in ATAC-seq and DNase-seq datasets using protocol-specific bias modeling. Genome Biology, 20(1), 42.

Campbell SA, Hoffman BG (2016) Chromatin Regulators in Pancreas Development and Diabetes. Trends Endocrinol Metab 27:142–152.

Cao J, Mu Q, Huang H (2018) The Roles of Insulin-Like Growth Factor 2 mRNA-Binding Protein 2 in Cancer and Cancer Stem Cells. Stem Cells Int 2018:4217259.

Carrasco M, Delgado I, Soria B, Rojas A (2012) GATA4 and GATA6 control mouse pancreas organogenesis. J Clin Invest 122:3504–3515.

Carrat, G. R. et al. Decreased STARD10 Expression Is Associated with Defective Insulin Secretion in Humans and Mice. Am. J. Hum. Genet. 100, 238–256 (2017).

Carter, Scott L., Brechbhler, Christian M., Griffin, Michael and Bond, An- drew T. (2004). Gene co-expression network topology provides a frame- work for molecular characterization of cellular state. Bioinformatics, 20(14), 2242–2250.

Carullo, N. V. N., & Day, J. J. (2019). Genomic Enhancers in Brain Health and Disease. *Genes*, *10*(1).

Cauchi, S., El Achhab, Y., Choquet, H., Dina, C., Krempler, F., Weitgasser, R., Nejjari, C., Patsch, W., Chikri, M., Meyre, D. et al. (2007) TCF7L2 is reproducibly associated with type 2 diabetes in various ethnic groups: a global meta-analysis. J. Mol. Med. (Berl), 85, 777–782.

Cavalcante RG, Sartor MA (2017). “annotatr: genomic regions in context.” Bioinformatics. R package version 1.10.0.

Cebola I, Rodríguez-Seguí SA, Cho CH-H, et al (2015) TEAD and YAP regulate the enhancer network of human embryonic pancreatic progenitors. Nat Cell Biol 17:615–626.

Chakrabarti, S. K., James, J. C., & Mirmira, R. G. (2002). Quantitative assessment of gene targeting in vitro and in vivo by the pancreatic transcription factor, Pdx1.

Importance of chromatin structure in directing promoter binding. The Journal of Biological Chemistry, 277(15), 13286–13293.

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

Chen C, Krohn J, Bhattacharya S, Davies B (2011) A Comparison of Exogenous Promoter Activity at the ROSA26 Locus Using a PhiC31 Integrase Mediated Cassette Exchange Approach in Mouse ES Cells. PLoS One 6:e23376.

Chen H, Gu X, Su I, et al (2009) Polycomb protein Ezh2 regulates pancreatic beta-cell Ink4a/Arf expression and regeneration in diabetes mellitus. Genes Dev 23:975–85.

Chen L, Ge B, Casale FP, Vasquez L, Kwan T, Garrido-Martín D, Watt S, Yan Y, Kundu K, Ecker S (2016) Genetic drivers of epigenetic and transcriptional variation in human immune cells. Cell 167: 1398 – 1414.e1324

Chen, T., Ueda, Y., Dodge, J. E., Wang, Z. & Li, E. Establishment and maintenance of genomic methylation patterns in mouse embryonic stem cells by Dnmt3a and Dnmt3b. Mol. Cell. Biol. 23, 5594–5605 (2003).

Cheryle A. Séguin, Jonathan S. Draper, Andras Nagy, and Janet Rossant. Establishment of Endoderm Progenitors by SOX Transcription Factor Expression in Human Embryonic Stem Cells. Cell Stem Cell, 3(2):182–195, 2008.

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

Cho NH, Shaw JE, Karunanga S, et al. (2018). IDF Diabetes Atlas: Global estimates of diabetes prevalence for 2017 and projections for 2045. Diabetes Research and Clinical Practice, 138, 271–281.

Chou WC, Cheng AL, Brotto M, et al. Visual gene-network analysis reveals the cancer gene co-expression in human endometrial cancer. BMC Genomics 2014;15:300.

Christodoulou C., Tyler A. Longmire, Steven S. Shen, et al. Mouse ES and iPS cells can form similar definitive endoderm despite differences in imprinted genes. Journal of Clinical Investigation, 121(6):2313– 2325, 2011.

Cnop M, Abdulkarim B, Bottu G, et al (2014) RNA sequencing identifies dysregulation of the human pancreatic islet transcriptome by the saturated fatty acid palmitate. Diabetes 63:1978–93.

Cnop M, Hannaert JC, Hoorens A, Eizirik DL, Pipeleers DG. Inverse relationship between cytotoxicity of free fatty acids in pancreatic islet cells and cellular triglyceride accumulation. Diabetes 2001;50:1771–1777

Cnop M. Fatty acids and glucolipotoxicity in the pathogenesis of Type 2 diabetes. Biochem Soc Trans 2008;36:348–352

Conway, E., Jerman, E., Healy, E., Ito, S., Holoch, D., Oliviero, G., … Bracken, A. P. (2018). A Family of Vertebrate-Specific Polycombs Encoded by the LCOR/LCORL Genes Balance PRC2 Subtype Activities. Molecular Cell, 70(3), 408-421.e8.

Cookson W, Liang L, Abecasis G, Moffatt M, Lathrop M. (2009) Mapping complex disease traits with global gene expression. Nat Rev Genet. 10(3):184–194.

Creyghton, M. P., Cheng, A. W., Welstead, G. G., Kooistra, T., Carey, B. W., Steine, E. J., Jaenisch, R. (2010). Histone H3K27ac separates active from poised enhancers and predicts developmental state. Proceedings of the National Academy of Sciences of the United States of America, 107(50), 21931–21936.

D'Amato E, Giacopelli F, Giannattasio A, et al (2010) Genetic investigation in an Italian child with an unusual association of atrial septal defect, attributable to a new familial GATA4 gene mutation, and neonatal diabetes due to pancreatic agenesis. Diabet Med 27:1195–1200. doi: 10.1111/j.1464-5491.2010.03046.x

D'Amour, K. A., Agulnick, A. D., Eliazer, S., Kelly, O. G., Kroon, E., & Baetge, E. E. (2005). Efficient differentiation of human embryonic stem cells to definitive endoderm. Nature Biotechnology, 23(12), 1534–1541.

D'Amour KA, Bang AG, Eliazer S, et al.(2006) Production of pancreatic hormone-expressing endocrine cells from human embryonic stem cells. Nat Biotechnol. 24(11): 1392–401.

Davis CA, Hitz BC, Sloan CA, et al (2018) The Encyclopedia of DNA elements (ENCODE): data portal update. Nucleic Acids Res 46:D794–D801.

Dayeh TA, Olsson AH, Volkov P, Almgren P, Ronn T, Ling C (2013) Identification of CpG-SNPs associated with type 2 diabetes and differential DNA methylation in human pancreatic islets. Diabetologia 56(5):1036– 1046.

Dayeh, T., Volkov, P., Salö, S., Hall, E., Nilsson, E., Olsson, A. H., Ling, C. (2014). Genome-wide DNA methylation analysis of human pancreatic islets from type 2 diabetic and non-diabetic donors identifies candidate genes that influence insulin secretion. *PLoS Genetics*, *10*(3), e1004160.

De Franco E, Shaw-Smith C, Flanagan SE, et al (2013) GATA6 Mutations Cause a Broad Phenotypic Spectrum of Diabetes From Pancreatic Agenesis to Adult-Onset Diabetes Without Exocrine Insufficiency. Diabetes 62:

Defossez, P.-A., Kelly, K. F., Filion, G. J. P., Pérez-Torrado, R., Magdinier, F., Menoni, H., … Gilson, E. (2005). The human enhancer blocker CTC-binding factor interacts with the transcription factor Kaiso. The Journal of Biological Chemistry, 280(52), 43017–43023.

Delaneau, O., Zazhytska, M., Borel, C., Giannuzzi, G., Rey, G., Howald, C., … Dermitzakis, E. T. (2019). Chromatin three-dimensional interactions mediate genetic effects on gene expression. *Science*, *364*(6439).

Dimas AS, Lagou V, Barker A, et al (2014) Impact of type 2 diabetes susceptibility variants on quantitative glycemic traits reveals mechanistic heterogeneity. Diabetes. doi: 10.2337/db13-0949

Dobin A, Gingeras TR (2015) Mapping RNA-seq reads with STAR. Curr Protoc Bioinforma 51:11.14.1-11.14.19.

Doi, A. et al. Differential methylation of tissue- and cancer-specific CpG island shores distinguishes human induced pluripotent stem cells, embryonic stem cells and fibro- blasts. Nat. Genet. 41, 1350–1353 (2009).

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

Dominic van Essen, Bettina Engist, Gioacchino Natoli, and Simona Saccani. Two modes of transcriptional activation at native promoters by NFkB p65. PLoS Biology, 7(3):0549–0562, 2009.

Doudna JA and Charpentier E. (2014) The new frontier of genome engineering with CRISPR-Cas9. Science, 346(6213):1258096–1258096.

Down, T. A., V. K. Rakyan, D. J. Turner, et al. 2008. A Bayesian deconvolution strategy for immunoprecipitation-based DNA methylome analysis. Nat. Biotechnol. 26:779–785.

El-Gohary, Y., Tulachan, S., Guo, P., Welsh, C., Wiersch, J., Prasadan, K., … Gittes, G. (2013). Smad signaling pathways regulate pancreatic endocrine development. Developmental Biology, 378(2), 83–93.

Elliott, H. R., Shihab, H. A., Lockett, G. A., Holloway, J. W., McRae, A. F., Smith, G. D., … Relton, C. L. (2017). Role of DNA Methylation in Type 2 Diabetes Etiology: Using Genotype as a Causal Anchor. Diabetes, 66(6), 1713–1722.

ENCODE Project Consortium TEP (2012) An integrated encyclopedia of DNA elements in the human genome. Nature 489:57–74.

Engler, C., Kandzia, R., & Marillonnet, S. (2008). A one pot, one step, precision cloning method with high throughput capability. PloS One, 3(11), e3647.

Ernst J, Kellis M. (2012). ChromHMM: automating chromatin-state discovery and characterization. Nature Methods 9:215–216.

Fabregat A, Jupe S, Matthews L, et al. The Reactome Pathway Knowledgebase. Nucleic Acids Res. 2018;46(D1): D649–D655.

Fadista, J., Vikman, P., Laakso, E. O., Mollet, I. G., Esguerra, J. L., Taneera, J., … Groop, L. (2014). Global genomic and transcriptomic analysis of human pancreatic islets reveals novel genes influencing glucose metabolism. Proceedings of the National Academy of Sciences, 111(38), 13924–13929.

Falcon S, Gentleman R (2007) Using GOstats to test gene lists for GO term association. Bioinformatics 23:257–258. doi: 10.1093/bioinformatics/btl567

Feng, S. et al. (2010) Conservation and divergence of methylation patterning in plants and animals. Proc. Natl Acad. Sci. USA 107, 8689–8694.

Ferron M, Lacombe J, Germain A, et al (2015) GGCX and VKORC1 inhibit osteocalcin endocrine functions. J Cell Biol. doi: 10.1083/jcb.201409111

Florez, J. (Ed.). (2016). The genetics of type 2 diabetes and related traits : Biology, physiology and translation. Retrieved from https://ebookcentral.proquest.com

Fraga, M.F. et al. (2005) Epigenetic differences arise during the lifetime of monozygotic twins. Proc. Natl. Acad. Sci. USA 102, 10604–10609.

Frayling TM, Timpson NJ, Weedon MN, Zeggini E, Freathy RM, et al. (2007) A common variant in the FTO gene is associated with body mass index and predisposes to childhood and adult obesity. Science 316: 889–894

Frommer M, McDonald LE, Millar DS, Collis CM, Watt F, Grigg GW, et al. (1992) A genomic sequencing protocol that yields a positive display of 5-methylcytosine residues in individual DNA strands. Proc Natl Acad Sci U S A.;89(5):1827–31.

Fuchsberger C, Flannick J, Teslovich TM, et al (2016) The genetic architecture of type 2 diabetes. Nat Publ Gr.

Fukaya T., Lim B., Levine M. Enhancer Control of Transcriptional Bursting. (2016) Cell. 166:358–368. doi: 10.1016/j.cell.2016.05.025

Fulco, A. C. P., Nasser, J., Jones, T. R., Munson, G., & Drew, T. (2019). Activity-by-Contact model of enhancer specificity from thousands of CRISPR perturbations. BioRxiv, 1–60.

Fulco, Perez, E. M., Kane, M., Cleary, B., & Lander, E. S. (2016). *Systematic mapping of functional enhancer-promoter connections with CRISPR interference*. *354*(6313), 769–773.

Galas D, Schmitz A. 1978. DNAse footprinting a simple method for the detection of protein-DNA binding specificity. Nucleic Acids Res 5: 3157– 3170.

Gamazon, E. R., Badner, J. A., Cheng, L., Zhang, C., Zhang, D., Cox, N. J., … Liu, C. (2013). Enrichment of cis-regulatory gene expression SNPs and methylation quantitative trait loci among bipolar disorder susceptibility variants. *Molecular Psychiatry*, *18*(3), 340–346.

Gaulton KJ, Ferreira T, Lee Y, et al (2015) Genetic fine mapping and genomic annotation defines causal mechanisms at type 2 diabetes susceptibility loci. Nat Genet. doi: 10.1038/ng.3437

Gaulton KJ, Nammo T, Pasquali L, et al (2010) A map of open chromatin in human pancreatic islets. Nat Genet. 42(3): 255–259

Gebre, M., Nomburg, J. L., & Gewurz, B. E. (2018). CRISPR-Cas9 Genetic Analysis of Virus-Host Interactions. Viruses, 10(2).

Gel B, Diez-Villanueva A, Serra E, Buschbeck M, Peinado MA, Malinverni R (2016). “regioneR: an R/Bioconductor package for the association analysis of genomic regions based on permutation tests.” Bioinformatics, 32(2), 289-291.

Gilbert LA, Larson MH, Morsut L, et al. (2014) CRISPR-Mediated Modular RNA-Guided Regulation of Transcription in Eukaryotes. Cell, 154(2):442–451.

Gilbert, L. A., Horlbeck, M. A., Adamson, B., Villalta, J. E., Chen, Y., Whitehead, E. H., Weissman, J. S. (2014). Genome-Scale CRISPR-Mediated Control of Gene Repression and Activation. Cell, 159(3), 647–661.

Gilbert, L. A., Larson, M. H., Morsut, L., Liu, Z., Brar, G. A., Torres, S. E., Stern-Ginossar, N., Brandman, O., Whitehead, E. H., Doudna, J. A., et al. (2013). CRISPR-Mediated Modular RNA-Guided Regulation of Transcription in Eukaryotes. Cell 154, 442–451.

Giresi, PG; Kim, J; McDaniell, RM; Iyer, VR; Lieb, JD (2007). "FAIRE (Formaldehyde-Assisted Isolation of Regulatory Elements) isolates active regulatory elements from human chromatin". Genome Research. 17 (6): 877–85.

Gittes GK (2009) Developmental biology of the pancreas: A comprehensive review. Dev Biol 326:4–35.

Goelz, S.E., vogelstein, B., Hamilton, S.R. & Feinberg, A.P. (1985) Hypomethylation of DNA from benign and malignant human colon neoplasms. Science 228, 187–190.

González, F., Zhu, Z., Shi, Z.-D., Lelli, K., Verma, N., Li, Q. V. and Huangfu, D. (2014). An iCRISPR Platform for Rapid, Multiplexable, and Inducible Genome Editing in Human Pluripotent Stem Cells. Cell Stem Cell 15, 215 –2

Gordân R, Hartemink A, Bulyk M. (2009). Distinguishing direct versus indirect transcription factor–DNA interactions. Genome Res 19: 2090–2100.

Grant C.E., Bailey T.L., Noble W.S., (2011) FIMO: scanning for occurrences of a given motif, Bioinformatics, 27 1017-1018.

Grant, S.F., Thorleifsson, G., Reynisdottir, I., Benediktsson, R., Manolescu, A., Sainz, J., Helgason, A., Stefansson, H., Emilsson, V., Helgadottir, A. et al. (2006) Variant oftranscription factor 7-like 2 (TCF7L2) gene confers risk oftype 2 diabetes. Nat. Genet., 38, 320–323.

Greenwald, W. W., Chiou, J., Yan, J., Qiu, Y., Dai, N., Wang, A., … Gaulton, K. J. (2019). Pancreatic islet chromatin accessibility and conformation reveals distal enhancer networks of type 2 diabetes risk. *Nature Communications*, *10*(1), 2078.

Grud, J., Madsen, S., Rauch, A., Hauwaert, E. L. Van, Schmidt, S. F., Winnefeld, M., & Mandrup, S. (2018). Integrated analysis of motif activity and gene expression changes of transcription factors. *Genome Research*, 1–13.

GTEx Consortium (2015) The genotype-tissue expression (GTEx) pilot analysis: multitissue gene regulation in humans. Science 348: 648 – 660

GTEx Consortium. (2017) Genetic effects on gene expression across human tissues. Nature 550(7675): 204_213.

Gu, G., Dubauskaite, J. and Melton, D. A. (2002). Direct evidence for the pancreatic lineage: NGN3+ cells are islet progenitors and are distinct from duct progenitors. Development 129, 2447-2457.

Guangchuang Yu, Li-Gen Wang, and Qing-Yu He. (2015) ChIPseeker: an R/Bioconductor package for ChIP peak annotation, comparison and visualization. Bioinformatics, 31(14):2382-2383

Guangchuang Yu, Li-Gen Wang, Yanyan Han and Qing-Yu He. (2012) clusterProfiler: an R package for comparing biological themes among gene clusters. OMICS: A Journal of Integrative Biology, 16(5):284-287

Guangchuang Yu, Qing-Yu He. (2016) ReactomePA: an R/Bioconductor package for reactome pathway analysis and visualization. Molecular BioSystems, 12(2):477-479

Guo Y, Xu Q, Canzio D, Shou J, Li J, Gorkin DU, Jung I, Wu H, Zhai Y, Tang Y, et al. (2015). CRISPR inversion of CTCF sites alters genome topology and enhancer/promoter function. Cell 162: 900–910.

Gutiérrez, G. D., Bender, A. S., Cirulli, V., Mastracci, T. L., Kelly, S. M., Tsirigos, A., … Sussel, L. (2017). Pancreatic β cell identity requires continual repression of non-β cell programs. *The Journal of Clinical Investigation*, *127*(1), 244–259.

Guttman, M., and Rinn, J.L. (2012). Modular regulatory principles of large non- coding RNAs. Nature 482, 339–346.

Guttman, M., Donaghey, J., Carey, B.W., Garber, M., Grenier, J.K., Munson, G., Young, G., Lucas, A.B., Ach, R., Bruhn, L., et al. (2011). lincRNAs act in the circuitry controlling pluripotency and differentiation. Nature 477, 295–300.

H. Wickham (2016) ggplot2: Elegant Graphics for Data Analysis., Springer-Verlag New York.

Haedicke, W., and R. Lesche. (2006). DNA methylation: potential for diagnosis, prognosis and therapy – prediction in oncology. Verh. Dtsch. Ges. Pathol. 90:39–45

Haffner, M. C., A. Chaux, A. K. Meeker, et al. (2011). Global 5-hydroxymethylcytosine content is significantly reduced in tissue stem/progenitor cell compartments and in human cancers. Oncotarget 2:627–637.

Hall, E., Dayeh, T., Kirkpatrick, C. L., Wollheim, C. B., Dekker Nitert, M., & Ling, C. (2013). DNA methylation of the glucagon-like peptide 1 receptor (GLP1R) in human pancreatic islets. BMC Medical Genetics, 14(1), 76.

Han, H., Shim, H., Shin, D., Shim, J. E., Ko, Y., Shin, J., Kim, H., Cho, A., Kim, E., Lee, T., Kim, H., Kim, K., Yang, S., Bae, D., Yun, A., Kim, S., Kim, C. Y., Cho, H. J., Kang, B., Shin, S. & Lee, I. (2015) TRRUST: a reference database of human transcriptional regulatory interactions. Sci. Rep. 5, 11432

Han, L., Lin, I. G. & Hsieh, C. L. (2001) Protein binding protects sites on stable episomes and in the chromosome from de novo methylation. Mol. Cell. Biol. 21, 3416–3424

Hanel, M.L. and Wevrick, R. (2001) The role of genomic imprinting in human developmental disorders: lessons from Prader-Willi syndrome. Clin Genet 59, 156-164, PubMed: 11260224

Hang Y, Stein R (2011) MafA and MafB activity in pancreatic β cells. Trends Endocrinol Metab 22:364–73.

Hang Y, Yamamoto T, Benninger RKP, et al (2014) The MafA transcription factor becomes essential to islet B-cells soon after birth. Diabetes. doi: 10.2337/db13-1001

Hannum, G., Guinney, J., Zhao, L., Zhang, L., Hughes, G., Sadda, S., … Zhang, K. (2013). Genome-wide methylation profiles reveal quantitative views of human aging rates. Molecular Cell, 49(2), 359–367.

Hansen KD (2016). lluminaHumanMethylationEPICanno.ilm10b2.hg19: Annotation for Illumina's EPIC methylation arrays. R package version 0.6.0.

Harder, M. N., Ribel-Madsen, R., Justesen, J. M., Sparsø, T., Andersson, E. A., Grarup, N., … Pedersen, O. (2013). Type 2 Diabetes Risk Alleles Near BCAR1 and in ANK1 Associate With Decreased β-Cell Function Whereas Risk Alleles Near ANKRD55 and GRB14 Associate With Decreased Insulin Sensitivity in the Danish Inter99 Cohort. The Journal of Clinical Endocrinology & Metabolism, 98(4), E801–E806.

Hasin, Y., Seldin, M., & Lusis, A. (2017). Multi-omics approaches to disease. Genome Biology, 18(1), 83.

Hatada, I., Y. Hayashizaki, S. Hirotsune, H. Komatsubara, and T. Mukai. 1991. A genomic scanning method for higher organisms using restriction sites as landmarks. Proc. Natl. Acad. Sci. USA 88:9523–9527

Haumaitre, C., Lenoir, O., & Scharfmann, R. (2008). Histone deacetylase inhibitors modify pancreatic cell fate determination and amplify endocrine progenitors. *Molecular and Cellular Biology*, *28*(20), 6373–6383.

Hayatsu, H., Y. Wataya, K. Kai, and S. Iida. (1970). Reaction of sodium bisulfite with uracil, cytosine, and their derivatives. Biochemistry 9:2858–2865.

Heard, E. & Martienssen, R. A (2014). Transgenerational epigenetic inheritance: myths and mechanisms. Cell 157, 95–109.

Heinz S, Benner C, Spann N, Bertolino E et al. (2010) Simple Combinations of Lineage-Determining Transcription Factors Prime cis-Regulatory Elements Required for Macrophage and B Cell Identities. Mol Cell May 28;38(4):576-589.

Hodges, E. et al. (2011) Directional DNA methylation changes and complex intermediate states accompany lineage specificity in the adult hematopoietic compartment. Mol. Cell 44, 17–28

Hon, G. C. et al. (2014). 5mC Oxidation by Tet2 modulates enhancer activity and timing of transcriptome reprogramming during differentiation. Mol. Cell 56, 286–297

Horikoshi M, Mägi R, van de Bunt M, et al (2015) Discovery and Fine-Mapping of Glycaemic and Obesity-Related Trait Loci Using High-Density Imputation. PLOS Genet 11:e1005230. doi: 10.1371/journal.pgen.1005230

Horvath, S. (2013). DNA methylation age of human tissues and cell types. Genome Biology, 14(10), R115.

Hrvatin, S., O'Donnell, C. W., Deng, F., Millman, J. R., Pagliuca, F. W., DiIorio, P., … Melton, D. A. (2014). Differentiated human stem cells resemble fetal, not adult, β cells. *Proceedings of the National Academy of Sciences*, *111*(8), 3038–3043.

Hsu C-C, Lee C-H, Wahlqvist ML, et al. (2012) Poverty Increases Type 2 Diabetes Incidence and Inequality of Care Despite Universal Health Coverage. Diabetes Care Nov, 35 (11) 2286-2292

Hsu, J. Y., Fulco, C. P., Cole, M. A., Canver, M. C., Pellin, D., Sher, F., … Pinello, L. (2018). CRISPR-SURF: discovering regulatory elements by deconvolution of CRISPR tiling screen data. *Nature Methods*, *15*(12), 992–993. https://doi.org/10.1038/s41592-018-0225-6

Hua, H., Zhang, Y.-Q., Dabernat, S., Kritzik, M., Dietz, D., Sterling, L., & Sarvetnick, N. (2006). BMP4 regulates pancreatic progenitor cell expansion through Id2. *The Journal of Biological Chemistry*, *281*(19), 13574–13580.

Huntzinger E, Izaurralde E (2011) Gene silencing by microRNAs: contributions of translational repression and mRNA decay. Nat Rev Genet 12: 99–110.

Irizarry RA, Ladd-Acosta C, Wen B et al. (2009) The human colon cancer methylome shows similar hypo-and hyper-methylation at conserved tissue-specific CpG island shores. Nat Genet;41:178–86.

Isaac B Hilton, Christopher M Vockley, I Pratiksha, et al. (2015) Epigenome editing by a CRISPR/Cas9-based acetyltransferase activates genes from promoters and enhancers. Nat Biotechnol, 33(5):510–517.

J. A. Doudna and E. Charpentier. (2014) The new frontier of genome engineering with CRISPR-Cas9. Science, 346(6213):1258096–1258096.

Jacquemin, P., Lemaigre, F. P., & Rousseau, G. G. (2003). The Onecut transcription factor HNF-6 (OC-1) is required for timely specification of the pancreas and acts upstream of Pdx-1 in the specification cascade. Developmental Biology, 258(1), 105–116.

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

Jaenisch, R., Harbers, K., Jahner, D., Stewart, C., and Stuhl- mann, H. 1982. DNA methylation, retroviruses, and embryogenesis. J. Cell Biochem. 20: 331–336.

Jaffe AE, Murakami P, Lee H, et al (2012) Bump hunting to identify differentially methylated regions in epigenetic epidemiology studies. Int J Epidemiol 41:200–9.

Javierre, B. M., Sewitz, S., Cairns, J., Wingett, S. W., Várnai, C., Thiecke, M. J., … Flicek, P. (2016). Lineage-Specific Genome Architecture Links Enhancers and Non-coding Disease Variants to Target Gene Promoters. *Cell*, *167*(5), 1369-1384.e19.

Jennings, R. E., Berry, A. A., Gerrard, D. T., Wearne, S. J., Strutt, J., Withey, S., Hanley, N. A. (2017). Laser Capture and Deep Sequencing Reveals the Transcriptomic Programmes Regulating the Onset of Pancreas and Liver Differentiation in Human Embryos. Stem Cell Reports, 9, 1387–1394.

Jennings, R. E., Berry, A. A., Kirkwood-Wilson, R., Roberts, N. A., Hearn, T., Salisbury, R. J., … Hanley, N. A. (2013). Development of the Human Pancreas From Foregut to Endocrine Commitment. *Diabetes*, *62*, 3514–3522.

Jennings, R. E., Berry, A. A., Strutt, J. P., Gerrard, D. T., & Hanley, N. A. (2015). Human pancreas development. Development (Cambridge, England), 142(18), 3126–3137.

Jensen J (2004) Gene Regulatory Factors in Pancreatic Development. Dev Dyn 229:176–200. doi: 10.1002/dvdy.10460

Ji, H. et al. (2010). Comprehensive methylome map of lineage commitment from haematopoietic progenitors. Nature 467, 338–342

Johnson, D. S., Mortazavi, A., & Myers, R. M. (2007). Genome-Wide Mapping of in Vivo Protein-DNA Interactions. Science, 316(June), 1497–1503.

Jones, P. A. (2012). Functions of DNA methylation: islands, start sites, gene bodies and beyond. Nature Rev. Genet. 13, 484–492

Jonsson A, Isomaa B, Tuomi T, Taneera J, Salehi A, Nilsson P, et al. (2009) A variant in the KCNQ1 gene preicts future type 2 diabetes and mediates impaired insulin secretion. Diabetes.;58(10):2409–13.

Jothi R, Cuddapah S, Barski A, Cui K, Zhao K. (2008). Genome-wide identification of in vivo protein–DNA binding sites from ChIP-Seq data. Nucleic Acids Res 36: 5221–5231.

Kahn, S.E. (2001) Clinical review 135: the importance of β-cell failure in the development and progression of type 2 diabetes. J. Clin. Endocrinol. Metab. 86, 4047–4058

Kajiwara M, Aoi T, Okita K, et al. (2012) Donor-dependent variations in hepatic differentiation from human-induced pluripotent stem cells. Proc Natl Acad Sci U S A. 109(31): 12538–43.

Kaminskas E, Farrell A, Abraham S et al. (2005) Approval summary: azacitidine for treatment of myelodysplastic syndrome subtypes. Clin Cancer Res;11:3604–08.

Kaminsky, Z. A. et al. (2009) DNA methylation profiles in monozygotic and dizygotic twins. Nat. Genet. 41, 240–245

Keiko Ikeda, Thomas Stuehler, and Michael Meisterernst. (2002) The H1 and H2 regions of the activation domain of herpes simplex virion protein 16 stimulate transcription through distinct molecular mechanisms. Genes to Cells, 7(1):49–58.

Keller MP, Choi Y, Wang P, et al. (2008) A gene expression network model of type 2 diabetes links cell cycle regulation in islets with diabetes susceptibility. Genome Res.;18(5):706–716.

Khetan, S., Kursawe, R., Youn, A., Lawlor, N., Jillette, A., Marquez, E. J., … Stitzel, M. L. (2018). Type 2 Diabetes-Associated Genetic Variants Regulate Chromatin Accessibility in Human Islets. Diabetes, 67(11), 2466–2477.

Kilpinen, H., & Dermitzakis, E. T. (2012). Genetic and epigenetic contribution to complex traits. Human Molecular Genetics, 21(R1), R24–R28.

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

Kim, K., Zhao, R., Doi, A., Ng, K., Unternaehrer, J., Cahan, P., Daley, G. Q. (2011). Donor cell type can influence the epigenome and differentiation potential of human induced pluripotent stem cells. Nature Biotechnology, 29(12), 1117–1119.

Kircher M, Witten DM, Jain P, O'Roak BJ, Cooper GM, Shendure J. (2014) A general framework for estimating the relative pathogenicity of human genetic variants. Nat Genet. 46: 310–315

Kiyokawa, H., Kineman, R.D., Manova-Todorova, K.O., Soares, V.C., Hoffman, E.S., Ono, M., Khanam, D., Hayday, A.C., Frohman, L.A., Koff, A., (1996). Enhanced growth of mice lacking the cyclin-dependent kinase inhibitor function of p27Kip1. Cell 85, 721–732

Kobberup, S., Nyeng, P., Juhl, K., Hutton, J., & Jensen, J. (2007). ETS-family genes in pancreatic development. *Developmental Dynamics*, *236*(11), 3100–3110.

Koike, H., Iwasawa, K., Ouchi, R., Maezawa, M., Giesbrecht, K., Saiki, N., … Takebe, T. (2019). Modelling human hepato-biliary-pancreatic organogenesis from the foregut-midgut boundary. *Nature*.

Kolb, H., & Martin, S. (2017). Environmental/lifestyle factors in the pathogenesis and prevention of type 2 diabetes. BMC Medicine, 15(1), 131.

Komor, A. C., Badran, A. H., & Liu, D. R. (2017). CRISPR-Based Technologies for the Manipulation of Eukaryotic Genomes. *Cell*, *168*(1–2), 20–36.

Konermann, S., Brigham, M. D., Trevino, A. E., Joung, J., Abudayyeh, O. O., Barcena, C., … Zhang, F. (2014). Genome-scale transcriptional activation by an engineered CRISPR-Cas9 complex. Nature, 517(7536), 583–588.

Kong A, Steinthorsdottir V, Masson G, et al.; DIAGRAM Consortium. Parental origin of sequence variants associated with complex diseases. Nature 2009;462:868–874

Kornberg, R. D. & Lorch, Y. Twenty-five years of the nucleosome, fundamental particle of the eukaryotic chromosome. Cell 98, 285-294 (1999).

Korthauer K, Irizarry RA (2018) Genome-wide repressive capacity of promoter DNA methylation is revealed through epigenomic manipulation. bioRxiv 381145.

Kotler E, Shani O, Goldfeld G, Lotan-Pompan M, Tarcic O, Gershoni A, Hopf TA, Marks DS, Oren M, Segal E (2018) A Systematic p53 Mutation Library Links Differential Functional Impact to Cancer Mutation Pattern and Evolutionary Conservation. Mol Cell 71 (1):178-190.e8,

Kowluru A (2000) Evidence for the Carboxyl Methylation of Nuclear Lamin-B in the Pancreatic b Cell. Biochem Biophys Res Commun 268:249–254.

Kristen E. Neely, Ahmed H. Hassan, Annika E. Wallberg, et al. Activation domain-mediated targeting of the SWI/SNF complex to promoters stimulates transcription from nucleosome arrays. Molecular Cell, 4(4):649–655, 1999

Krivega, A. Dean, Enhancer and promoter interactions-long distance calls. Curr Opin Genet Dev 22, 79-85 (2012).

Krol J, Loedige I, Filipowicz W (2010) The widespread regulation of microRNA biogenesis, function and decay. Nat Rev Genet 11: 597–610.

Kroon E, Martinson LA, Kadoya K, Bang AG, Kelly OG, Eliazer S, Young H, Richardson M, Smart NG, et al. Pancreatic endoderm derived from human embryonic stem cells generates glucose- responsive insulin-secreting cells in vivo. Nat Biotechnol. 2008; 26:443–452. [PubMed: 18288110]

Kulakovskiy, I. V., Vorontsov, I. E., Yevshin, I. S., Sharipov, R. N., Fedorova, A. D., Rumynskiy, E. I., … Makeev, V. J. (2018). HOCOMOCO: Towards a complete collection of transcription factor binding models for human and mouse via large-scale ChIP-Seq analysis. Nucleic Acids Research, 46(D1), D252–D259.

Kumar S., Ambrosini G., Bucher P., SNP2TFBS - a database of regulatory SNPs affecting predicted transcription factor binding site affinity, NUcleic Acids Res 2017 Jan4. PMC5210548

Kuroda A, Rauch TA, Todorov I, Ku HT, Al-Abdullah IH, Kandeel F, et al. Insulin gene expression is regulated by DNA methylation. PLoS One. 2009; 4(9):e6953.

Kycia, I. et al. A Common Type 2 Diabetes Risk Variant Potentiates Activity of an Evolutionarily Conserved Islet Stretch Enhancer and Increases C2CD4A and C2CD4B Expression. Am. J. Hum. Genet. 102, 620–635 (2018).

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.

Lander, ES. (1996) The new genomics: global views of biology. Science 274, 536–539

Langfelder P and Horvath S (2008). WGCNA: an R package for weighted correlation network analysis. BMC Bioinformatics, 9:559.

Lango Allen, H., Flanagan, S. E., Shaw-Smith, C., De Franco, E., Akerman, I., Caswell, R., … Ellard, S. (2011). GATA6 haploinsufficiency causes pancreatic agenesis in humans. *Nature Genetics*, *44*(1), 20–22. https://doi.org/10.1038/ng.1035

Larsen, H. L., Martín-Coll, L., Nielsen, A. V., Wright, C. V. E., Trusina, A., Kim, Y. H., & Grapin-Botton, A. (2017). Stochastic priming and spatial cues orchestrate heterogeneous clonal contribution to mouse pancreas organogenesis. *Nature Communications*, *8*(1), 605. https://doi.org/10.1038/s41467-017-00258-4

Latchman, D. S. (1997). Transcription factors: An overview. The International Journal of Biochemistry & Cell Biology, 29(12), 1305–1312.

Law CW, Chen Y, Shi W, Smyth GK (2014) voom: Precision weights unlock linear model analysis tools for RNA-seq read counts. Genome Biol 15:R29. doi: 10.1186/gb-2014-15-2-r29

Lawlor, N., George, J., Bolisetty, M., Kursawe, R., Sun, L., Sivakamasundari, V., Stitzel, M. L. (2017). Single-cell transcriptomes identify human islet cell signatures and reveal cell-type – specific expression changes in type 2 diabetes. Genome Research. 27: 1-15

Lawlor, N., Márquez, E. J., Orchard, P., Narisu, N., Shamim, M. S., Thibodeau, A., Stitzel, M. L. (2019). Multiomic Profiling Identifies cis-Regulatory Networks Underlying Human Pancreatic β Cell Identity and Function. Cell Reports, 26(3), 788-801.e6.

Le Bacquer O, Shu L, Marchand M, et al (2011) TCF7L2 splice variants have distinct effects on b-cell turnover and function. doi: 10.1093/hmg/ddr072

Lee, T. I., Jenner, R. G., Boyer, L. A., Guenther, M. G., Levine, S. S., Kumar, R. M., … Young, R. A. (2006). Control of developmental regulators by Polycomb in human embryonic stem cells. *Cell*, *125*(2), 301–313. https://doi.org/10.1016/j.cell.2006.02.043

Lei S. Qi, Matthew H. Larson, Luke A. Gilbert, et al. (2013) Repurposing CRISPR as an RNA-γuided platform for sequence-specific control of gene expression. Cell, 152(5):1173–1183.

Leti, F., Llaci, L., Malenica, I., & DiStefano, J. K. (2018). Methods for CpG Methylation Array Profiling Via Bisulfite Conversion. Methods in Molecular Biology (Clifton, N.J.), 1706, 233–254.

Li J, Witten D, Johnstone I, Tibshirani R (2012). Normalization, testing, and false discovery rate estimation for RNA-sequencing data. Biostatistics, 13:523–538.

Li, E., Beard, C. & Jaenisch, R. (1993). Role for DNA methylation in genomic imprinting. Nature 366, 362–365.

Li, H., & Liu, J. (2012). The novel function of HINFP as a co-activator in sterol-regulated transcription ofPCSK9in HepG2 cells. Biochemical Journal, 443(3), 757–768.

Li, Z., Schulz, M. H., Zenke, M., & Filho, I. G. C. (2018). Identification of Transcription Factor Binding Sites using ATAC-seq. BioRxiv, 362863.

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

Lin, K. C., Park, H. W., & Guan, K.-L. (2017). Regulation of the Hippo Pathway Transcription Factor TEAD. *Trends in Biochemical Sciences*, *42*(11), 862–872.

Lister, R. et al. (2013) Global epigenomic reconfiguration during mammalian brain development. Science 341, 1237905

Lister, R., M. Pelizzola, R. H. Dowen, et al. (2009). Human DNA methylomes at base resolution show widespread epigenomic differences. Nature 462:315–322.

Liu, Y., Zhang, X., Blumenthal, R.M. & Cheng, X. (2013) A common mode of

recognition for methylated CpG.Trends Biochem. Sci. 38, 177–183

López-Maury, L., Marguerat, S., & Bähler, J. (2008). Tuning gene expression to changing environments: from rapid responses to evolutionary adaptation. Nature Reviews Genetics, 9(8), 583–593.

Love, M.I., Huber, W., Anders, S. (2014) Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biology 15(12):550

Lu, F., Liu, Y., Jiang, L., Yamaguchi, S. & Zhang, Y. (2014) Role of Tet proteins in enhancer activity and telomere elongation. Genes Dev. 28, 2103–2119

Lupiáñez DG, Kraft K, Heinrich V, Krawitz P, Brancati F, Klopocki E, Horn D, Kayserili H, Opitz JM, Laxova R, et al. (2015). Disruptions of topological chromatin domains cause pathogenic rewiring of gene-enhancer inter- actions. Cell 161: 1012–1025.

Lupski JR, Belmont JW, Boerwinkle E, Gibbs RA (2011) Clan genomics and the complex architecture of human disease. Cell 147:32–43.

Lyssenko, V. *et al.* (2008) Clinical risk factors, DNA variants, and the development of type 2 diabetes. N. Engl. J. Med. 359, 2220–2232

Ma, H. et al. (2014) Abnormalities in human pluripotent cells due to reprogramming mechanisms. Nature 511, 177–183

MacDonald, P. E., & Wheeler, M. B. (2003). Voltage-dependent K + channels in pancreatic beta cells: Role, regulation and potential as therapeutic targets. Diabetologia, 46(8), 1046–1062.

Mahajan A, Taliun D, Thurner M, 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., Wessel, J., Willems, S. M., Zhao, W., Robertson, N. R., Chu, A. Y., … McCarthy, M. I. (2018b). Refining the accuracy of validated target identification through coding variant fine-mapping in type 2 diabetes article. *Nature Genetics*, *50*(4), 559–571.

Maller JB, McVean G, Byrnes J, et al (2012) Bayesian refinement of association signals for 14 loci in 3 common diseases. Nat Genet.

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.

Marcus Kretzschmar, Klaus Kaiser, Friedrich Lottspeich, and Michael Meisterernst (1994) A novel mediator of class II gene transcription with homology to viral immediate-early transcriptional regulators. Cell, 78(3):525– 534.

Marioni, R. E., Shah, S., McRae, A. F., Chen, B. H., Colicino, E., Harris, S. E., … Deary, I. J. (2015). DNA methylation age of blood predicts all-cause mortality in later life. Genome Biology, 16(1), 25.

Martin D, Kim Y-H, Sever D, Mao C-A, Haefliger J-A, Grapin-Botton A (2015) REST represses a subset of the pancreatic endocrine differentiation program. Dev Biol 405:316–327.

Martin, M. (2011) Cutadapt removes adapter sequences from high-throughput sequencing reads. EMBnet.journal 17, 10–12

Mathelier, A., Zhao, X., Zhang, A. W., Parcy, F., Worsley-Hunt, R., Arenillas, D. J., … Wasserman, W. W. (2014). JASPAR 2014: an extensively expanded and updated open-access database of transcription factor binding profiles. Nucleic Acids Research, 42(Database issue), D142-7.

Mattick, J.S., and Makunin, I.V. (2006). Non-coding RNA. Hum. Mol. Genet. 15 (Spec No 1), R17–R29.

Maunakea, A. K., Chepelev, I., Cui, K. & Zhao, K. (2013) Intragenic DNA methylation modulates alternative splicing by recruiting MeCP2 to promote exon recognition. Cell Res. 23, 1256–1269

Maxwell R Mumbach, 1, 2, 12, Ansuman T Satpathy, 1, … & Howard Y Chang. (2017). Enhancer connectome in primary human cells identifies target genes of disease-associated DNA elements. Nature Genetics.

McCarthy DJ, Chen Y, Smyth GK (2012). "Differential expression analysis of multifactor RNA-Seq experiments with respect to biological variation." *Nucleic Acids Research*, **40**(10), 4288-4297.

Meda, P., & Schuit, F. (2013). Glucose-stimulated insulin secretion: the hierarchy of its multiple cellular and subcellular mechanisms. Diabetologia, 56(12), 2552–2555.

Meehan, R. R., Lewis, J. D., McKay, S., Kleiner, E. L. & Bird, A. P. (1989) Identification of a mammalian protein that binds specifically to DNA containing methylated CpGs. Cell 58, 499–507

Memedula S, Belmont A.S. (2003) Sequential recruitment of HAT and SWI/SNF components to condensed chromatin by VP16. Current Biology, 13(3):241–246,.

Mifsud, B., Tavares-Cadete, F., Young, A.N., Sugar, R., Schoenfelder, S., Fer- reira, L., Wingett, S.W., Andrews, S., Grey, W., Ewels, P.A., et al. (2015). Mapping long-range promoter contacts in human cells with high-resolution capture Hi-C. Nat. Genet. 47, 598–606.

Miguel-Escalada, I., Bonàs-Guarch, S., Cebola, I., Ponsa-Cobas, J., Mendieta-Esteban, J., Atla, G., … Ferrer, J. (2019). Human pancreatic islet three-dimensional chromatin architecture provides insights into the genetics of type 2 diabetes. *Nature Genetics*, 1.

Mikkelsen TS, KuM, Jaffe DB, Issac B, Lieberman E, GiannoukosG, Alvarez P, Brockman W, Kim TK, Koche RP, et al. (2007). Genome-wide maps of chromatin state in pluripotent and lineage-committed cells. Nature 448: 553–560

Millman JR, Xie C, Van Dervort A, et al (2016) Generation of stem cell-derived β-cells from patients with type 1 diabetes. Nat Commun 7:11463. doi: 10.1038/ncomms11463

Misra, B. B., Langefeld, C., Olivier, M., & Cox, L. A. (2018). Integrated omics: tools, advances and future approaches. Journal of Molecular Endocrinology, (2016), R21–R45.

Mohrs M, et al. Deletion of a coordinate regulator of type 2 cytokine expression in mice. Nature Immunol. (2001); 2:842–847.

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

Mootha, V. K., Lindgren, C. M., Eriksson, K.-F., Subramanian, A., Sihag, S., Lehar, J., … Nicholas Houstis, Mark J Daly, Nick Patterson, L. C. G. (2003). PGC-1α-

responsive genes involved in oxidative phosphorylation are coordinately downregulated in human diabetes. *Nature Genetics*, *34*(3), 267–273. Retrieved from https://www.nature.com/articles/ng1180.pdf

Morán, I., Akerman, I., van de Bunt, M., Xie, R., Benazra, M., Nammo, T., … Melton, D. A. (2012). Human β cell transcriptome analysis uncovers lncRNAs that are tissue-specific, dynamically regulated, and abnormally expressed in type 2 diabetes. Cell Metabolism, 16(4), 435–448.

Morley M, Molony CM, Weber TM, Devlin JL, Ewens KG, Spielman RS, Cheung VG (2004) Genetic analysis of genome-wide variation in human gene expression. Nature. 430(7001):743-7.

Morloy, M., Molony, C. M., Weber, T. M., Devlin, J. L., Ewens, K. G., Spielman, R. S., & Cheung, V. G. (2004). Genetic analysis of genome-wide variation in human gene expression. Nature, 430(7001), 743–747.

Morris, J. a, & Barrett, J. C. (2012). Olorin: combining gene flow with exome sequencing in large family studies of complex disease. *Bioinformatics (Oxford, England)*, *28*(24), 3320–3321.

Morrisey EE, Tang Z, Sigrist K, *et al.* (1998) GATA6 regulates HNF4 and is required for differentiation of visceral endoderm in the mouse embryo. Genes Dev.; 12(22): 3579–90.

Mosedale M, Egodage S, Calma RC, Chi N-W, Chessler SD (2012) Neurexin-1α contributes to insulin-containing secretory granule docking. J Biol Chem 287:6350–6361.

Muraro MJ, Dharmadhikari G, Grün D, et al (2016) A Single-Cell Transcriptome Atlas of the Human Pancreas.

Nair, G. G., Liu, J. S., Russ, H. A., Tran, S., Saxton, M. S., Chen, R., … Hebrok, M. (2019). Recapitulating endocrine cell clustering in culture promotes maturation of human stem-cell-derived β cells. Nature Cell Biology, 21(2), 263–274.

Najafabadi, M. M., Villanustre, F., Khoshgoftaar, T. M., Seliya, N., Wald, R., & Muharemagic, E. (2015). Deep learning applications and challenges in big data analytics. *Journal of Big Data*, *2*(1), 1.

Neiman D, Moss J, Hecht M, et al (2017) Islet cells share promoter hypomethylation independently of expression, but exhibit cell-type-specific methylation in enhancers. Proc Natl Acad Sci U S A 114:13525–13530.

Nestor, C. E. et al. Rapid reprogramming of epigenetic and transcriptional profiles in mammalian culture systems. Genome Biol. 16, 11 (2015).

Neta Lavon, Ofra Yanuka, and Nissim Benvenisty. The effect of overexpression of Pdx1 and Foxa2 on the differentiation of human embryonic stem cells into pancreatic cells. Stem cells, 24:1923–1930, 2006.

Nica, A. C., Ongen, H., Irminger, J. C., Bosco, D., Berney, T., Antonarakis, S. E., … Dermitzakis, E. T. (2013). Cell-type, allelic, and genetic signatures in the human pancreatic beta cell transcriptome. Genome Research, 23(9), 1554–1562.

Nicolae DL, Gamazon E, Zhang W, Duan S, Dolan ME, Cox NJ. 2010. Trait-associated SNPs are more likely to be eQTLs: Annotation to enhance discovery

from GWAS. PLoS Genet 6: e1000888.

Nicolson TJ, Bellomo EA, Wijesekara N, et al. Insulin storage and glucose homeostasis in mice null for the granule zinc transporter ZnT8 and studies of the type 2 diabetes-associated variants. Diabetes. 2009;58:2070–83.

Njølstad TS, Ellard S, Njølstad PR. Chapter 10: Insights from monogenic diabetes. In Florez, JC (ed) The genetics of type 2 diabetes and related traits. Springer. 223-240

Nostro MC, Sarangi F, Yang C, et al.: Efficient generation of NKX6–1+ pancreatic progenitors from multiple human pluripotent stem cell lines. Stem Cell Reports. 2015; 4(4): 591–604

O 'Leary NA, Wright MW, Brister JR, Ciufo S, Haddad D, … Pruitt KD. Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. Nucleic Acids Res. 2016 Jan 4;44(D1):D733-45

Ohta, Y., Kosaka, Y., Kishimoto, N., Wang, J., Smith, S. B., Honig, G., … German, M. S. (2011). Convergence of the insulin and serotonin programs in the pancreatic β-cell. Diabetes, 60(12), 3208–3216.

Okano, M., Bell, D. W., Haber, D. A. & Li, E. DNA methyltransferases Dnmt3a and Dnmt3b are essential for de novo methylation and mammalian development. Cell 99, 247–257 (1999).

Olkhov-Mitsel, E., & Bapat, B. (2012). Strategies for discovery and validation of methylated and hydroxymethylated DNA biomarkers. Cancer Medicine, 1(2), 237–260.

Olsson, A. H., Volkov, P., Bacos, K., Dayeh, T., Hall, E., Nilsson, E. A., … Ling, C. (2014). Genome-Wide Associations between Genetic and Epigenetic Variation Influence mRNA Expression and Insulin Secretion in Human Pancreatic Islets. PLoS Genetics, 10(11), e1004735.

Ouaamari A El, Dirice E, Gedeon N, et al (2016) SerpinB1 Promotes Pancreatic β-Cell Proliferation HHS Public Access. Cell Metab January 12:194–205. doi: 10.1016/j.cmet.2015.12.001

Pagliuca FW, Millman JR, Gürtler M, et al (2014) Generation of functional human pancreatic β cells in vitro. Cell 159:428–39.

Pagliuca, F. W., Melton, D. A., Ang, S. L., Rossant, J., Annes, J. P., Ryu, J. H., Melton, D. A. (2013). How to make a functional β-cell. Development (Cambridge, England), 140(12), 2472–2483.

Pan D, Xiao Zhang, Chiang-Ching Huang, Nadereh Jafari, Warren a Kibbe, Lifang Hou, and Simon M Lin. (2010). “Comparison of Beta-value and M-value methods for quantifying methylation levels by microarray analysis.” BMC Bioinformatics 11 (1). BioMed Central Ltd:587.

Pan, F. C., & Brissova, M. (2014). Pancreas development in humans. *Current Opinion in Endocrinology, Diabetes, and Obesity*, *21*(2), 77.

Pan, F. C., & Wright, C. (2011). Pancreas organogenesis: From bud to plexus to gland. *Developmental Dynamics*, *240*(3), 530–565.

Pan, G. J., Chang, Z. Y., Schöler, H. R., & Pei, D. (2002). Stem cell pluripotency and transcription factor Oct4. Cell Research, 12(5–6), 321–329.

Parker, S. C. J., Stitzel, M. L., Taylor, D. L., Orozco, J. M., Erdos, M. R., Akiyama, J. A., … Young, A. (2013). Chromatin stretch enhancer states drive cell-specific gene regulation and harbor human disease risk variants. Proceedings of the National Academy of Sciences, 110(44), 17921–17926.

Pasquali L, Gaulton KJ, Mularoni L, et al (2014) Pancreatic islet enhancer clusters enriched in type 2 diabetes risk-associated variants. Nat Genet 46:136–143. doi: 10.1038/ng.2870

Paul, L., Walker, E. M., Drosos, Y., Cyphert, H. A., Neale, G., Stein, R., … Sosa-Pineda, B. (2016). Lack of Prox1 Downregulation Disrupts the Expansion and Maturation of Postnatal Murine β-Cells. *Diabetes*, *65*(3), 687–698.

Pennacchio, L. A., Bickmore, W., Dean, A., Nobrega, M. A., & Bejerano, G. (2013). Enhancers: five essential questions. *Nature Reviews. Genetics*, *14*(4), 288–295. https://doi.org/10.1038/nrg3458

Phipson, B., Maksimovic, J., and Oshlack, A. (2015). missMethyl: an R package for analysing methylation data from Illuminas HumanMethylation450 platform. Bioinformatics, btv560.

Piccand, J., Strasser, P., Hodson, D. J., Meunier, A., Ye, T., Keime, C., … Gradwohl, G. (2014). Rfx6 Maintains the Functional Identity of Adult Pancreatic β Cells. *Cell Reports*, *9*(6), 2219–2232.

Pickrell J. K. (2014). Joint analysis of functional genomic data and genome-wide association studies of 18 human traits. American journal of human genetics, 94(4), 559–573.

Pidsley, R., Zotenko, E., Peters, T. J., Lawrence, M. G., Risbridger, G. P., Molloy, P., … Clark, S. J. (2016). Critical evaluation of the Illumina MethylationEPIC BeadChip microarray for whole-genome DNA methylation profiling. Genome Biology, 17(1), 208.

Portela, A., & Esteller, M. (2010). Epigenetic modifications and human disease. Nature Biotechnology, 28(10), 1057–1068.

Prashant, M., Kevin, E., & George, C. (2014). Cas9 as a versatile tool for engeneering biology. *NIH Public Access*, *10*(10), 957–963.

Qi LS, Larson MH, Gilbert LA, et al. (2013) Repurposing CRISPR as an RNA-γuided platform for sequence-specific control of gene expression. Cell, 152(5):1173–1183.

Quinlan, A.R. & Hall, I.M. BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics 26, 841–842 (2010).

R Development Core Team (2008) R: A language and environment for statistical computing.

Raj A, Shim H, Gilad Y, Pritchard JK, Stephens M. 2015. mscentipede: modeling heterogeneity across genomic sites and replicates improves accuracy in the inference of transcription factor binding. PloS one 10(9):e0138030.

Rasmussen KD and Helin K. Role of TET enzymes in DNA methylation, development and cancer. Genes and Development. 30:733–750 (2016)

Rauch, T., and G. P. Pfeifer. 2005. Methylated-CpG island recovery assay: a new

technique for the rapid detection of methylated-CpG islands in cancer. Lab. Invest. 85:1172–1180.

Ravassard, P., Hazhouz, Y., Pechberty, S., Bricout-Neveu, E., Armanet, M., Czernichow, P., … Scharfmann, R. (2011). A genetically engineered human pancreatic β cell line exhibiting glucose-inducible insulin secretion. Journal of Clinical Investigation, 121(9), 3589–3597.

Reich DE, Lander ES (2001). On the allelic spectrum of human disease. TRENDS in Genetics 17(9). 502-510

Ren, G., Jin, W., Cui, K., Rodrigez, J., Hu, G., Zhang, Z., … Zhao, K. (2017). CTCF-Mediated Enhancer-Promoter Interaction Is a Critical Regulator of Cell-to-Cell Variation of Gene Expression. Molecular Cell, 67(6), 1049-1058.e6.

Rena, G., Hardie, D. G., & Pearson, E. R. (2017). The mechanisms of action of metformin. Diabetologia, 60(9), 1577–1585.

Rentzsch P, Witten D, Cooper GM, Shendure J, Kircher M. (2018) CADD: predicting the deleteriousness of variants throughout the human genome. Nucleic Acids Res. 47 (D1) : D886-D894.

Rezania A, Bruin JE, Arora P, et al (2014) Reversal of diabetes with insulin-producing cells derived in vitro from human pluripotent stem cells. Nat Biotechnol 32:1121–1133.

Rezania A, Bruin JE, Xu J, Narayan K, Fox JK, O'Neil JJ, Kieffer TJ. (2013) Enrichment of human embryonic stem cell-derived NKX6.1-expressing pancreatic progenitor cells accelerates the maturation of insulin-secreting cells in vivo. Stem Cells.; 31:2432–442.

Ritchie ME, Phipson B, Wu D, et al (2015) limma powers differential expression analyses for RNA-sequencing and microarray studies. Nucleic Acids Res 43:e47. doi: 10.1093/nar/gkv007

Ritchie, M. D., Holzinger, E. R., Li, R., Pendergrass, S. A., & Kim, D. (2015). Methods of integrating data to uncover genotype–phenotype interactions. Nature Reviews Genetics, 16(2), 85–97. https://doi.org/10.1038/nrg3868

Rizzo JM, Sinha S: Analyzing the global chromatin structure of keratinocytes by MNase-Seq. Methods Mol Biol 2014, 1195:49–59

Roadmap Epigenomics Consortium, Kundaje, A., Meuleman, W., Ernst, J., Bilenky, M., Yen, A., … Kellis, M. (2015). Integrative analysis of 111 reference human epigenomes. Nature, 518(7539), 317–329.

Robertson, J., A. B. Robertson, and A. Klungland. 2011. The presence of5-hydroxymethylcytosine at the gene promoter and not in the gene body negatively regulates gene expression. Biochem. Biophys. Res. Commun. 411:40–43.

Robinson MD, McCarthy DJ, Smyth GK (2010). "edgeR: a Bioconductor package for differential expression analysis of digital gene expression data." *Bioinformatics*, **26**(1), 139-140

Robinson PJ, Fairall L, Huynh VA, Rhodes D. (2006) EM measurements define the dimensions of the "30-nm" chromatin fiber: evidence for a compact, interdigitated structure. Proc Natl Acad Sci U S A. 103(17):6506–6511.

Roman, T. S. et al. A Type 2 Diabetes-Associated Functional Regulatory Variant in a

Pancreatic Islet Enhancer at the Adcy5 Locus. Diabetes db170464 (2017)

Rouhani F, Kumasaka N, de Brito MC, et al. (2014) Genetic background drives transcriptional variation in human induced pluripotent stem cells. PLoS Genet. 10(6): e1004432.

Rubin AJ, Parker KR, Satpathy AT, et al (2019) Coupled Single-Cell CRISPR Screening and Epigenomic Profiling Reveals Causal Gene Regulatory Networks. Cell 1–16.

Rukstalis JM, Habener JF (2009) Neurogenin3: a master regulator of pancreatic islet differentiation and regeneration. Islets.

Russ A H, Audrey V Parent, Ringler JJ, et al (2015) Controlled induction of human pancreatic progenitors produces functional beta-like cells in vitro. EMBO J 34:1759 – 1772.

Rusu, V. et al. Type 2 Diabetes Variants Disrupt Function of SLC16A11 through Two Distinct Mechanisms. Cell 170, 199–212.e20 (2017).

Sakhneny L, Rachi E, Epshtein A et al (2018) Pancreatic pericytes support beta-cell function in a Tcf7l2-dependent manner. Diabetes 67: 437–447.

Sako Y, Grill VE. (1990) A 48-hour lipid infusion in the rat time-dependently inhibits glucose-induced insulin secretion and B cell oxidation through a process likely coupled to fatty acid oxidation. Endocrinology;127:1580–1589

Sander M, et al. Homeobox gene Nkx6.1 lies downstream of Nkx2.2 in the major pathway of beta- cell formation in the pancreas. Development. 2000; 127:5533–5540.

Sandoval J, Heyn H, Moran S, Serra-Musach J, Pujana MA, Bibikova M, Esteller M. 2011. Validation of a DNA methylation microarray for 450,000 CpG sites in the human genome. Epigenetics 6:692–702

Sanjana, N. E., Wright, J., Zheng, K., Shalem, O., Fontanillas, P., Joung, J., … Zhang, F. (2016). High-resolution interrogation of functional elements in the noncoding genome. Science, 353(6307), 1545–1549.

Saxena, R., Gianniny, L., Burtt, N.P., Lyssenko, V., Giuducci, C., Sjogren, M., Florez, J.C., Almgren, P., Isomaa, B., Orho-Melander, M. et al. (2006) Common single nucleotide polymorphisms in TCF7L2 are reproducibly associated with type 2 diabetes and reduce the insulin response to glucose in nondiabetic individuals. Diabetes, 55, 2890–2895.

Schaffer, A.E., Freude, K.K., Nelson, S.B., and Sander, M. (2010). Nkx6 transcription factors and Ptf1a function as antagonistic line- age determinants in multipotent pancreatic progenitors. Dev. Cell 18, 1022–1029.

Schena M, Shalon D, Davis RW, Brown PO. (1995) Quantitative monitoring of gene expression patterns with a complementary DNA microarray. Science. 270(5235):467-70.

Schmittgen, T. D., & Livak, K. J. (2008). Analyzing real-time PCR data by the comparative CT method. *Nature Protocols*, *3*(6), 1101–1108. https://doi.org/10.1038/nprot.2008.73

Schoenfelder, S., Furlan-Magaril, M., Mifsud, B., Tavares-Cadete, F., Sugar, R., Javierre, B.M., Nagano, T., Katsman, Y., Sakthidevi, M., Wingett, S.W., et al. (2015). The pluripotent regulatory circuitry connecting promoters to their long-

range interacting elements. Genome Res. 25, 582–597.

Schübeler D (2015) Function and information content of DNA methylation. Nature 517:321–326.

Schuit F, Van Lommel L, Granvik M, Goyvaerts L, de Faudeur G, Schraenen A, et al. β-cell- specific gene repression: a mechanism to protect against inappropriate or maladjusted insulin secretion? Diabetes. May; 2012 61(5):969–75.

Schwartz, M. W., Guyenet, S. J., & Cirulli, V. (2010). The hypothalamus and ß-cell connection in the gene-targeting era. Diabetes, 59(12), 2991–2993.

Schwitzgebel VM (2014) Many faces of monogenic diabetes. J Diabetes Invest 5:121–133.

Scott LJ, Mohlke KL, Bonnycastle LL et al (2007) A genome-wide association study of type 2 diabetes in Finns detects multiple susceptibility variants. Science 316:1341– 1345

Scott RA, Scott LJ, Mägi R et al (2017) An expanded genome-wide association study of type 2 diabetes in Europeans. Diabetes 66: 2888–2902

Segrè, A. V., Groop, L., Mootha, V. K., Daly, M. J., Altshuler, D., Daly, M. J., … Ennis, S. (2010). Common inherited variation in mitochondrial genes is not enriched for associations with type 2 diabetes or related glycemic traits. *PLoS Genetics*, *6*(8), e1001058.

Sexton,T. & Cavalli, G. (2015) The role of chromosome domains in shaping the functional genome. Cell 160, 1049-1059

Shannon P, Markiel A, Ozier O, et al (2003) Cytoscape: A Software Environment for Integrated Models of Biomolecular Interaction Networks.

Sharif, J. et al. (2007) The SRA protein Np95 mediates epigenetic inheritance by recruiting Dnmt1 to methylated DNA. Nature 450, 908–912

Shaw-Smith C, Elisa De Franco, Hana Lango Allen, et al (2014) GATA4 Mutations Are a Cause of Neonatal and Childhood-Onset Diabetes. Diabetes 63:2888 – 2894.

Sherwood, R.I.; Hashimoto, T.; O'Donnell, C.W.; Lewis, S.; Barkal, A.A.; van Hoff, J.P.; Karun, V.; Jaakkola, T.; Gifford, D.K. (2014) Discovery of directional and nondirectional pioneer transcription factors by modeling DNase profile magnitude and shape. Nature Biotechnol., 32, 171–178.

Shi Z, Lee K, Yang D, et al (2017) Genome Editing in hPSCs Reveals GATA6 Haploinsufficiency and a Genetic Interaction with GATA4 in Human Pancreatic Development Article Genome Editing in hPSCs Reveals GATA6 Haploinsufficiency and a Genetic Interaction with GATA4 in Human Pancreatic Develop. Stem Cell 20:1–14.

Shih, H. P., Seymour, P. A., Patel, N. A., Xie, R., Wang, A., Liu, P. P., … Zaret, K. S. (2015). A Gene Regulatory Network Cooperatively Controlled by Pdx1 and Sox9 Governs Lineage Allocation of Foregut Progenitor Cells. *Cell Reports*, *13*(2), 326–336. https://doi.org/10.1016/j.celrep.2015.08.082

Shukla, S. et al. (2011). CTCF-promoted RNA polymerase II pausing links DNA methylation to splicing. Nature 479, 74–79

Siggens, L., & Ekwall, K. (2014). Epigenetics, chromatin and genome organization:

recent advances from the ENCODE project. Journal of Internal Medicine, 276(3), 201–214

Skyler, J. S., Bakris, G. L., Bonifacio, E., Darsow, T., Eckel, R. H., Groop, L., … Ratner, R. E. (2017). Differentiation of Diabetes by Pathophysiology, Natural History, and Prognosis. Diabetes, 66(2), 241–255.

Smallwood, S. A. et al. (2011) Dynamic CpG island methylation landscape in oocytes and preimplantation embryos. Nature Genet. 43, 811–814

Smart, N. G., Apelqvist, S. A., Gu, X., Harmon, E. B., Topper, J. N., Macdonald, R. J., & Kim, S. K. (2006). Conditional expression of Smad7 in pancreatic beta cells disrupts TGF-beta signaling and induces reversible diabetes mellitus. *PLoS Biology*, *4*(2), e39.

Smyth G: (2005) Limma: linear models for microarray data. In Bioinformatics and Computational Biology Solutions using R and Bioconductor. Edited by Gentleman R, Carey V, Dudoit S, Irizarry R, Huber W. New York: Springer;:397–420.

Smyth, G. K. (2004). Linear Models and Empirical Bayes Methods for Assessing Differential Expression in Microarray Experiments. Statistical Applications in Genetics and Molecular Biology, 3(1), 1–25.

Solimena M, Schulte AM, Marselli L et al (2018) Systems biology of the IMIDIA biobank from organ donors and pancreatectomised patients defines a novel transcriptomic signature of islets from individuals with type 2 diabetes. Diabetologia 61: 641–657.

Song L, Zhang Z, Grasfeder LL, et al. (2011) Open chromatin defined by DNaseI and FAIRE identifies regulatory elements that shape cell-type identity. Genome Res. 21(10):1757–1767.

Song., Langfelder, S., & Horvath S. (2012). Comparison of co-expression measures: Mutual information, correlation, and model based indices. BMC Bioinformatics, 13(1).

Spaeth JM, Hunter CS, Bonatakis L, et al (2015) The FOXP1, FOXP2 and FOXP4 transcription factors are required for islet alpha cell proliferation and function in mice HHS Public Access. Diabetologia 58:1836–1844. doi: 10.1007/s00125-015-3635-3

Stadler, M. B. et al. (2011) DNA-binding factors shape the mouse methylome at distal regulatory regions. Nature 480, 490–495

Stefflova, K. et al. (2013). Cooperativity and rapid evolution of cobound transcription factors in closely related mammals. Cell 154, 530–540

Steiner DJ, Kim A, Miller K, *et al.*(2010) Pancreatic islet plasticity: interspecies comparison of islet architecture and composition. Islets.; 2(3): 135–45

Stitzel, M. L., Kycia, I., Kursawe, R., & Ucar, D. (2015). Transcriptional Regulation of the Pancreatic Islet: Implications for Islet Function. *Current Diabetes Reports*, *15*(9), 66. https://doi.org/10.1007/s11892-015-0635-0

Stitzel, M. L., Sethupathy, P., Pearson, D. S., Chines, P. S., Song, L., Erdos, M. R., … Collins, F. S. (2010). Global Epigenomic Analysis of Primary Human Pancreatic

Islets Provides Insights into Type 2 Diabetes Susceptibility Loci. *Cell Metabolism*, *12*(5), 443–455. https://doi.org/10.1016/J.CMET.2010.09.012

Strahl, B. D., & Allis, C. D. (2000). The language of covalent histone modifications. Nature, 403(6765), 41–45.

Straussman, R. et al. Developmental programming of CpG island methylation profiles in the human genome. Nat. Struct. Mol. Biol. 16, 564–571 (2009).

Stuart JM, Segal E, Koller D and Kim SK. (2003). A Gene-Coexpression Network for Global Discovery of Conserved Genetic Modules. Science, 302(5643), 249–255.

Subramanian, A., Tamayo, P., Mootha, V. K., Mukherjee, S., Ebert, B. L., Gillette, M. A., … Mesirov, J. P. (2005). Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proceedings of the National Academy of Sciences of the United States of America*, *102*(43), 15545–15550.

Suckow AT, Comoletti D, Waldrop MA, et al (2008) Expression of Neurexin, Neuroligin, and Their Cytoplasmic Binding Partners in the Pancreatic □-Cells and the Involvement of Neuroligin in Insulin Secretion. Endocrinology 149:6006–6017. doi: 10.1210/en.2008-0274

Suzuki, M. M. & Bird, A. (2008) DNA methylation landscapes: provocative insights from epigenomics. Nature Rev. Genet. 9, 465–476

Swift GH, Liu Y, Rose SD, et al (1998) An Endocrine-Exocrine Switch in the Activity of the Pancreatic Homeodomain Protein PDX1 through Formation of a Trimeric Complex with PBX1b and MRG1 (MEIS2). 18:5109–5120.

Tahiliani, M. et al. (2009) Conversion of 5-methylcytosine to 5-hydroxymethylcytosine in mammalian DNA by MLL partner TET1. Science 324, 930–935

Takahashi K, Tanabe K, Ohnuki M, et al. (2007) Induction of pluripotent stem cells from adult human fibroblasts by defined factors. Cell. 131(5): 861–72.

Tan JT, Nurbaya S, Gardner D, Ye S, Tai ES, Ng Daniel PK. (2009) Genetic variation in KCNQ1 associates with fasting glucose and beta-cell function: a study of 3734 subjects comprising three ethnicities living in Singapore. Diabetes,58(6):1445–9.

Tanay, A., O'Donnell, A. H., Damelin, M. & Bestor, T. H. (2007) Hyperconserved CpG domains underlie Polycomb-binding sites. Proc. Natl Acad. Sci. USA 104, 5521–5526

Tavassoly, I., Goldfarb, J., & Iyengar, R. (2018). Systems biology primer: the basic methods and approaches. Essays in Biochemistry, 62(4), 487–500.

Taylor R. (2013) Type 2 Diabetes. Etiology and reversibility 36(4): 1047-1055.

Teytelman, L., Ozaydin, B., Zill, O., Lefrançois, P., Snyder, M., Rine, J., & Eisen, M. B. (2009). Impact of chromatin structures on DNA processing for genomic analyses. PloS One, 4(8), e6700.

The ENCODE Project Consortium. (2007). Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. Nature 447: 799–816

The ENCODE Project Consortium. (2012) An integrated encyclopedia of DNA elements in the human genome. Nature 489: 57.

Thomas, J. O., & Kornberg, R. D (1975) Octamer of histones in chromatin and free in solution. Proceedings of the National Academy of Sciences 72, 2626–2630

Thurman RE, Rynes E, Humbert R, et al (2012) The accessible chromatin landscape of the human genome. Nature 489:75–82.

Thurner, M., Shenhav, L., Wesolowska-Andersen, A., Bennett, A. J., Barrett, A., Gloyn, A. L., Efrat, S. (2017). Genes Associated with Pancreas Development and Function Maintain Open Chromatin in iPSCs Generated from Human Pancreatic Beta Cells. Stem Cell Reports, 9(5), 1395–1405.

Thurner, M., van de Bunt, M., Torres, J. M., Mahajan, A., Nylander, V., Bennett, A. J., … McCarthy, M. I. (2018). Integration of human pancreatic islet genomic data refines regulatory mechanisms at type 2 diabetes susceptibility loci. ELife, 7, 1–30. https://doi.org/10.7554/eLife.31977

Timothy J. Peters, Michael J. Buckley, Aaron L. Statham, Ruth Pidsley, Katherine Samaras, Reginald V. Lord, Susan J. Clark and Peter L. Molloy (2015) De novo identification of differentially methylated regions in the human genome, Epigenetics & Chromatin, 8:6.

Timothy L. Bailey and Charles Elkan (1994) .Fitting a mixture model by expectation maximization to discover motifs in biopolymers. Proceedings of the Second International Conference on Intelligent Systems for Molecular Biology, pp. 28-36, AAAI Press, Menlo Park, California.

Timothy L. Bailey, Mikael Bodén, Fabian A. Buske, Martin Frith, Charles E. Grant, Luca Clementi, Jingyuan Ren, Wilfred W. Li, William S. Noble, (2009) MEME SUITE: tools for motif discovery and searching. Nucleic Acids Research, 37:W202-W208,

Tiyaboonchai A, Cardenas-Diaz FL, Ying L, et al (2017) GATA6 Plays an Important Role in the Induction of Human Definitive Endoderm, Development of the Pancreas, and Functionality of Pancreatic Cells. Stem Cell Reports 8:589–604.

Trapnell C. 2015. Defining cell types and states with single-cell genomics. Genome Res 25: 1491–1498.

Travers ME, Mackay DJG, Nitert MD, et al (2013) Insights Into the Molecular Mechanism for Type 2 Diabetes Susceptibility at the KCNQ1 Locus From Temporal Changes in Imprinting Status in Human Islets. Diabetes 62:987–992.

Tsompana, M., & Buck, M. J. (2014). Chromatin accessibility: a window into the genome. Epigenetics & Chromatin, 7(1), 33.

van Arensbergen J, García-Hurtado J, Moran I, et al (2010) Derepression of Polycomb targets during pancreatic organogenesis allows insulin-producing beta-cells to adopt a neural gene activity program. Genome Res 20:722–32.

van Dam S, Võsa U, van der Graaf A, et al (2018) Gene co-expression analysis for functional classification and gene-disease predictions. Brief Bioinform 19:575–592.

van de Bunt M, Lako M, Barrett A, et al (2016) Insights into islet development and biology through characterization of a human iPSC-derived endocrine pancreas model. Islets 8:83–95. doi: 10.1080/19382014.2016.1182276

van De Bunt M, Manning Fox JE, Dai X 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

Vanhorenbeeck V, Jenny M, Cornut JF, et al. (2007) Role of the Onecut transcription factors in pancreas morphogenesis and in pancreatic and enteric endocrine differentiation. Dev Biol.305:685–694.

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

Vardimon, L., Kressmann, A., Cedar, H., Maechler, M. & Doerfler, W. (1982) Expression of a cloned adenovirus gene is inhibited by in vitro methylation. Proc. Natl Acad. Sci. USA 79, 1073–1077

Varshney A, Scott LJ, Welch RP, et al (2017) Genetic regulatory signatures underlying islet gene expression and type 2 diabetes. PNAS 114(9):2301-2306

Velasco, G. et al. (2010) Dnmt3b recruitment through E2F6 transcriptional repressor mediates germ-line gene silencing in murine somatic tissues. Proc. Natl Acad. Sci. USA 107, 9281–9286

Venkatesh, S., & Workman, J. L. (2015). Histone exchange, chromatin structure and the regulation of transcription. Nature Publishing Group, 16.

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. Sci- ence 291: 1304–1351.

Verfaillie A, Imrichová H, Van de Sande B, et al (2014) iRegulon: From a Gene List to a Gene Regulatory Network Using Large Motif and Track Collections.

Viacyte: A Safety, Tolerability, and Efficacy Study of VC-01^{TM} Combination Product in Subjects With Type I Diabetes Mellitus.

Vogel, C. and Marcotte, E.M. (2012) Insights into the regulation of protein abundance from pro- teomic and transcriptomic analyses. Nat. Rev. Genet. 13.4 , pp. 227–32.

Voight, B. F., Scott, L. J., Steinthorsdottir, V., Andrew, P., Aulchenko, Y. S., Thorleifsson, G., … Ferreira, T. (2010). Twelve type 2 diabetes susceptibility loci identified through large-scale association analysis. *Nat Genet*, *42*(7), 579–589.

Voineagu I, Wang XC, Johnston P, et al. (2011) Transcriptomic ana- lysis of autistic brain reveals convergent molecular pathology. Nature;474:380

Volkmar M, et al. (2012) DNA methylation profiling identifies epigenetic dysregulation in pancreatic islets from type 2 diabetic patients. EMBO J. 31:1405–1426.

Wagner JR, Busche S, Ge B, et al (2014) The relationship between DNA methylation , genetic and expression inter-individual variation in untransformed human fibroblasts. 15:1–17.

Wagner M, Kunsch S, Duerschmied D, et al (2003) Transgenic overexpression of the oncofetal RNA binding protein KOC leads to remodeling of the exocrine pancreas. Gastroenterology 124:1901–1914.

Wainschtein P, Jain DP, Yengo L, et al. (2019) Recovery of trait heritability from whole genome sequence data. bioRxiv 588020;

Wang H, Brun T, Wollheim CB, et al (2007) MafA controls Genes implicated in Insulin

Biosynthesis and Secretion. Diabetologia

Wang X, Tucker NR, Rizki G, et al. (2016). Discovery and validation of sub-threshold genome-wide association study loci using epigenomic signatures. eLife 5:e10557.

Wang, A., Ren, B., & Correspondence, M. S. (2015). Epigenetic Priming of Enhancers Predicts Developmental Competence of hESC-Derived Endodermal Lineage Intermediates. *Cell Stem Cell*, *16*, 386–399.

Wang, J., Kilic, G., Aydin, M., Burke, Z., Oliver, G., & Sosa-pineda, B. (2005). Prox1 activity controls pancreas morphogenesis and participates in the production of "' secondary transition "' pancreatic endocrine cells. 286, 182–194. https://doi.org/10.1016/j.ydbio.2005.07.021

Wang, Y.J., Schug, J., Won, K.-J., Liu, C., Naji, A., Avrahami, D., Golson, M.L., and Kaestner, K.H. (2016). Single-Cell Transcriptomics of the Human Endocrine Pancreas. Diabetes 65, 3028–3038.

Wang, Z., Gerstein, M., & Snyder, M. (2009). RNA-Seq: a revolutionary tool for transcriptomics. *Nature Reviews. Genetics*, *10*(1). https://doi.org/10.1038/nrg2484

Wang, Z., Zang, C., Rosenfeld, J. A., Schones, D. E., Barski, A., Cuddapah, S., … Zhao, K. (2008). Combinatorial patterns of histone acetylations and methylations in the human genome. *Nature Genetics*, *40*(7), 897–903.

Wawryk-Gawda, E., Chylińska-Wrzos, P., Lis-Sochocka, M., Chłapek, K., Bulak, K., Jędrych, M., & Jodłowska-Jędrych, B. (2014). P53 protein in proliferation, repair and apoptosis of cells. *Protoplasma*, *251*(3), 525–533.

Weber, AP. (2015) Discovering New Biology through Sequencing of RNA. Plant Physiology 169(3): 1524-1531.

Weirauch, M. T., Yang, A., Albu, M., Cote, A. G., Montenegro-Montero, A., Drewe, P., … Hughes, T. R. (2014). Determination and inference of eukaryotic transcription factor sequence specificity. *Cell*, *158*(6), 1431–1443.

Weisenberger, D. J., D. Van Den Berg, F. Pan, B. P. Berman, and P. W. Laird. (2008). Comprehensive DNA methylation analysis on the Illumina® Infinium® assay platform. [Technical report], Illumina, Inc., San Diego, CA.

Westmoreland, J. J., Kilic, G., Sartain, C., Sirma, S., Blain, J., Rehg, J., … Pineda, –. (2012). Pancreas-Specific Deletion of Prox1 Affects Development and Disrupts Homeostasis of the Exocrine Pancreas. Gastroenterology, 124(4), 999–1009.

Wood, A. R., Jonsson, A., Jackson, A. U., Wang, N., & van Leewen, N. (2017). A genome-wide association study of IVGTT-based measures of first phase insulin secretion refines the underlying physiology of type 2 diabetes variants. Diabetes, 66(8), 2296–2309.

Wright, GW, & Simon RM. (2003). A random variance model for detection of differential gene expression in small microarray experiments. Bioinformatics, 19(18), 2448–2455.

Wu Z-Y, Zhu L-J, Zou N, et al (2012) AMPA Receptors Regulate Exocytosis and Insulin Release in Pancreatic β Cells. Traffic 13:1124–1139. doi: 10.1111/j.1600-0854.2012.01373.x

Xie R, Carrano AC, Sander M (2015) A systems view of epigenetic networks regulating pancreas development and β-cell function. Wiley Interdiscip Rev Syst Biol Med 7:1–11.

Xie R, Everett LJ, Lim H-W, et al (2013) Dynamic chromatin remodeling mediated by Polycomb proteins orchestrates pancreatic differentiation of human embryonic stem cells. Cell Stem Cell 12:224–237. doi: 10.1016/j.stem.2012.11.023

Xin, Y., Kim, J., Okamoto, H., Ni, M., Wei, Y., Adler, C., Gromada, J. (2016). RNA Sequencing of Single Human Islet Cells Reveals Type 2 Diabetes Genes. Cell Metabolism, 24(4), 608–615.

Xu C-R, Cole PA, Meyers DJ, et al (2011) Chromatin "Prepattern" and Histone Modifiers in a Fate Choice for Liver and Pancreas. Science 332:963–6.

Xu C-R, Li L-C, Donahue G, et al (2014) Dynamics of genomic H3K27me3 domains and role of EZH2 during pancreatic endocrine specification. EMBO J 33:2157–2170.

Xu Z, Niu L, Li L, Taylor J (2016). "ENmix: a novel background correction method for Illumina HumanMethylation450 BeadChip." Nucleic Acids Research.

Yang BT, Dayeh TA, Kirkpatrick CL, Taneera J, Kumar R, Groop L, et al. (2011) Insulin promoter DNA methylation correlates negatively with insulin gene expression and positively with HbA(1c) levels in human pancreatic islets. Diabetologia, 54(2):360–7.

Yang BT, Dayeh TA, Volkov PA, Kirkpatrick CL, Malmgren S, Jing X, et al. (2012) Increased DNA methylation and decreased expression of PDX-1 in pancreatic islets from patients with type 2 diabetes. Mol Endocrinol.; 26(7):1203–12.

Yang C, Xu Y, mo M, et al (2017) Induced pluripotent stem cell modelling of HLHS underlines the contribution of dysfunctional NOTCH signalling to impaired cardiogenesis. Hum. Mol. Genet.

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

Zeggini E, Weedon MN, Lindgren CM et al (2007) Replication of genome-wide association signals in U.K. samples reveals risk loci for type 2 diabetes. Science 316:1336– 1341

Zhang B and Horvath S (2005) A General Framework for Weighted Gene Co-Expression Network Analysis, Statistical Applications in Genetics and Molecular Biology: Vol. 4: No. 1, Article 17 PMID: 16646834

Zhang C, Moriguchi T, Kajihara M, et al (2005) MafA Is a Key Regulator of Glucose-Stimulated Insulin Secretion. Mol Cell Biol 25:4969–4976. doi: 10.1128/MCB.25.12.4969–4976.2005

Zhang J, McKenna LB, Bogue CW, Kaestner KH (2014) The diabetes gene Hhex maintains d-cell differentiation and islet function. Genes Dev 28:829–834. doi: 10.1101/gad.235499.113

Zhao W, Langfelder P, Fuller T, et al. Weighted gene coex- pression network analysis: state of the art. J Biopharm Stat 2010;20:281–300.

Zhao, Y., & Garcia, B. A. (2015). Comprehensive Catalog of Currently Documented Histone Modifications. *Cold Spring Harbor Perspectives in Biology*, *7*(9), a025064. https://doi.org/10.1101/cshperspect.a025064

Zheng, Y., Ley, S. H., & Hu, F. B. (2018). Global aetiology and epidemiology of type 2 diabetes mellitus and its complications. Nature Reviews Endocrinology, 14(2), 88–98.

Zhou, Q., & Melton, D. A. (2018). Pancreas regeneration. Nature, 557(7705), 351–358.

Zhu Y, Tonne JM, Liu Q, et al (2019) Targeted Derivation of Organotypic Glucose- and GLP-1-Responsive β Cells Prior to Transplantation into Diabetic Recipients. Stem Cell Reports 0:

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

Ziller, M. J. et al. (2013) Charting a dynamic DNA methylation landscape of the human genome. Nature 500, 477–481

Zoungas S, Woodward M, Li Q, et al. (2014) Impact of age, age at diagnosis and duration of diabetes on the risk of macrovascular and microvascular complications and death in type 2 diabetes. Diabetologia 57(12): 2465-2474.

Zuela N, Daniel Z. Bar, Yosef Gruenbaum (2012) Lamins in development, tissue maintenance and stress. EMBO Rep 13:1070–1078.