The role of novel genetic and epigenetic mechanisms in the development of Type 2 Diabetes

written by

Matthias Thurner

University of Oxford

Wadham College

A thesis submitted for the

degree of Doctor of Philosophy

Trinity Term 2016

Contents

Acronyms

ADP	Adenosine Diphosphate.
AF	Allelle Frequency.
ATAC-seq	Assay For Transposase-Accessible Chromatin sequencing.
ATP	Adenosine Triphosphate.
BAF	B Allele Frequency.
BiPSC	Beta-cell induced Pluripotent Stem Cell.
BMI	Body Mass Index.
CAD	Coronary Artery Disease.
cAMP	cyclic Adenosine Monophosphate.
CGH	Comparative Genomic Hybridisation.
CGI	CpG island.
ChIP	Chromatin Immunoprecipitation.
chQTL	chromatin Quantitative Trait Loci.
CI	Confidence Interval.

CN	Copy Number.
CNV	Copy Number Variant.
CPD	cAMP-Response Element Binding Protein.
CRISPR	Clustered Regularly Interspaced Short Palindromic Repeats.
CT	Cycle Threshold.
CTCF	CCCTC-Binding Factor.
DGV	Database Of Genomic Variants.
DIAGRAM	Diabetes Genetics Replication And Meta-Analysis.
DMR	Differentially Methylated Region.
DOCS	Differential Open Chromatin Site.
ENCODE	The Encyclopedia Of DNA Elements.
ENGAGE	European Network For Genetic And Genomic Epidemiology.
eQTL	expression Quantitative Trait Loci.
ESC	Embryonic Stem Cell.
ExAC	Exome Aggregation Consortium.
FAIRE-seq	Formaldehyde-Assisted Isolation Of Regulatory Elements Sequencing.
FC	Fold Change.
FDR	False Discovery Rate.
FE	Fold Enrichment.
FG	Fasting Glucose.
FiPSCs	Fibroblast induced Pluripotent Stem Cells.
FPG	Fasting Plasma Glucose.
GCK	Glucokinase.

GO	Gene Ontology.
GoF	Gain of Function.
GSIS	Glucose-Stimulated Insulin Secretion.
GWAS	Genome-Wide Association Studies.
HAT	Histone Acetyltransferase.
HDAC	Histone Deacetylase.
HMM	Hidden Markov Model.
indels	Insertions and Deletions.
iPSC	Induced Pluripotent Stem Cell.
KS	Kolmogorov-Smirnov.
LD	Linkage Disequilibrium.
LDL	Low-Density Lipoprotein.
LMR	Lowly Methylated Region.
LoF	Loss of Function.
LRR	Log R Ratio.
LRT	Loglikelihood Ratio Test.
MAF	Minor Allele Frequency.
MBD	Methyl-CPG Binding Domain.
MODY	Maturity Onset Diabetes of the Young.
mQTL	methylation Quantitative Trait Loci.
NGS	Next-Generation Sequencing.
OR	Odds Ratio.

PBS	Phosphate Buffer Saline.
PCA	Principal Component Analysis.
PCR	Polymerase Chain Reaction.
PMD	Partially Methylated Domain.
PP	Posterior Probability.
PTM	Posttranslational Modification.
QC	Quality Control.
qPCR	quantitative Real-Time PCR.
RD	Read Depth.
RP	Read Pair.
SNP	Single Nucleotide Polymorphism.
SNV	Single Nucleotide Variant.
SR	Split Read.
SV	Structural Variant.
T1D	Type 1 Diabetes.
T2D	Type 2 Diabetes.
TC	Total Cholesterol.
TF	Transcription Factor.
TFBS	Transcription Factor Binding Site.
TSS	Transcription Start Site.
UMR	Unmethylated Region.
VMP	Variable Methylated Position.
WES	Whole-Exome Sequencing.

WGBS	Whole-Genome Bisulfite Sequencing.

Acknowledgements

First of all, I would like to express my sincere gratitude to both my supervisors Prof Mark McCarthy and Prof Anna Gloyn for being such great mentors throughout my DPhil project. Thank you for all the scientific advice, your time and knowledge, your trust in my ability to work on multiple complex projects, and your encouragement to make me strive for excellence as a research scientist.

I am grateful to the Wellcome Trust for funding me to pursue this project.

I would like to express my special appreciation to both the McCarthy and the Gloyn lab for their support, helpful advice and friendship. My special thanks goes to Kyle Gaulton, thanks to whom my transition from a wet-lab scientist to a computer scientist was smoother than I ever expected; to Martijn van de Bunt, whose support and continuous advice were greatly appreciated as well as to my other colleagues and friends in the office Juan, Loukas, Patrick and Anthony. I also would like to thank all colleagues and friends upstairs for their help whenever needed, in particular, Anubha, Momoko, Will and Neil. In addition, I am also very grateful for all the help and friendship provided by my colleagues and friends in OCDEM including Nicola (who has helped me a lot with all her knowledge regarding iPSCs and stem cells), Mandy, Amy and Chris as well as fellow students Soren, Natasha, Mahesh and Sara. Thank you also to everyone else who has helped and supported me throughout my DPhil.

I would like to thank my family without whom I would not be today in Oxford. Last but not least I

owe much gratitude to my newly married wife, Aga, who during the first 2 months of our marriage had to deal with a husband locked in a messy home office spending most of the day (and most nights) working on this thesis. Nevertheless, she has supported me through this time and at every stage of my DPhil and shown an infinite amount of love and patience.

Thank you to all. I have had the best DPhil experience one may imagine and I am sincerely grateful to my supervisors, colleagues, family and friends.

Abstract

Abstract

Type 2 diabetes (T2D) is a metabolic disease characterised by insulin resistance and beta-cell failure and a leading cause of mortality and morbidity. GWAS have identified pancreatic beta-cell dysfunction as a major player in T2D pathogenesis, but since the majority of signals reside in non-coding sequence the causal variant(s) and effector transcripts often remain undefined. Combining human genetic discoveries with genomic interrogation of relevant tissues was shown to provide a powerful strategy to understand the molecular disease mechanisms. Thus, the main aim of my DPhil project is to generate and integrate genetic data, such as rare coding CNVs, and epigenomic data, including human pancreatic islet DNA methylation and chromatin state data, to characterise existing T2D GWAS loci and identify novel genetic T2D susceptibility signals.

Using WGBS (n=10) and ATAC-seq (n=29), I extended the epigenomic characterisation of human pancreatic islets. Specifically, I characterised for the first time the whole-genome human islet DNA methylome and provided information about the variability in open chromatin in human islets. Enrichment analysis in T2D-related GWAS data showed that hypomethylated enhancer-like regions (n=37.1k, FE=3.2) and open chromatin (n=247.2k, FE=3.1) regulatory regions are strongly enriched in T2D GWAS data.

These annotations were integrated with existing islet ChIP-seq annotations to extend islet chromatin states. Through this approach, I identified enhancer subtypes based on DNA methylation and accessibility, which differed markedly in their enrichment levels in T2D-related GWAS signals. The refined enhancer states were then used to prioritise T2D-associated variants for testing of allelic imbalance in open chromatin. After correcting for technical artefacts, significant allelic (FDR <0.05) imbalance in open chromatin was discovered at 3 T2D GWAS loci (*ADCY5, CDC123, KLHDC5*); thus highlighting potential molecular mechanisms at these loci governing T2D risk.

I also used ATAC-seq to characterise the epigenomic landscape of human iPSC subtypes derived from beta- and fibroblast-cells that differed in their ability to differentiate into endodermal tissue. The identification of significant (FDR <0.05) differentially open chromatin sites (n=8.3k) in these iPSC, which were found to be associated with genes linked to both pancreas development and mature function, could be used to improve current *in vivo* 'iPSC to endodermal' differentiation protocols. Ultimately, this may lead to enhanced generation of human islet cells relevant for both T2D disease modelling and potential clinical use.

Finally, I determined the role of rare coding CNVs, identified from large-scale T2D-case-control exome-sequencing (n=14.5k) and exome chip array (n=11.6k) datasets, in T2D development. This analysis found at the known *C2CD4A/VPS13C* T2D GWAS locus, rare *VPS13C* gene deletions and duplications consistently associated with increased and decreased T2D-risk, respectively; hence, prioritising *VPS13C* for potential functional follow-up at this GWAS locus. These data have extended the catalogue of genetic variants (rare coding CNVs) and the epigenomic landscape of human islets and iPSC to improve understanding of islet development and regulation as well as T2D genetic susceptibility.

CHAPTER 1

Introduction

1.1 Type 2 Diabetes (T2D) and its impact on global health

Diabetes is one of the most prevalent chronic and non-communicable diseases in the world and a major global health burden. The prevalence of diabetes has increased significantly over the last years due to social and cultural changes such as ageing populations, increase in urbanisation as well as changes in diet and physical activity. Currently, 415 million people are estimated to have diabetes worldwide with about 75% of individuals living in low- and middle-income countries (Fig1.1). About 175 million of these individuals are predicted to be undiagnosed which increases their risk of dangerous long-term complications. An additional 318 million people show signs of glucose tolerance which is associated with a risk of developing the disease at a later stage. Diabetes and its related complications are one of the leading causes of mortality in the world responsible for about 5 million deaths in 2015. Apart from the effect on morbidity and mortality, diabetes and its complications are also a significant economic and financial burden. In 2015 global health spending

on diabetes was estimated at $673 billion, which is equivalent to approximately 12% of the total expenditures on global health. Over the next years the number of affected individuals is expected to rise significantly with approximately 642 million people, predominately in low- and middle-income countries, being affected by diabetes in 2040.[1]

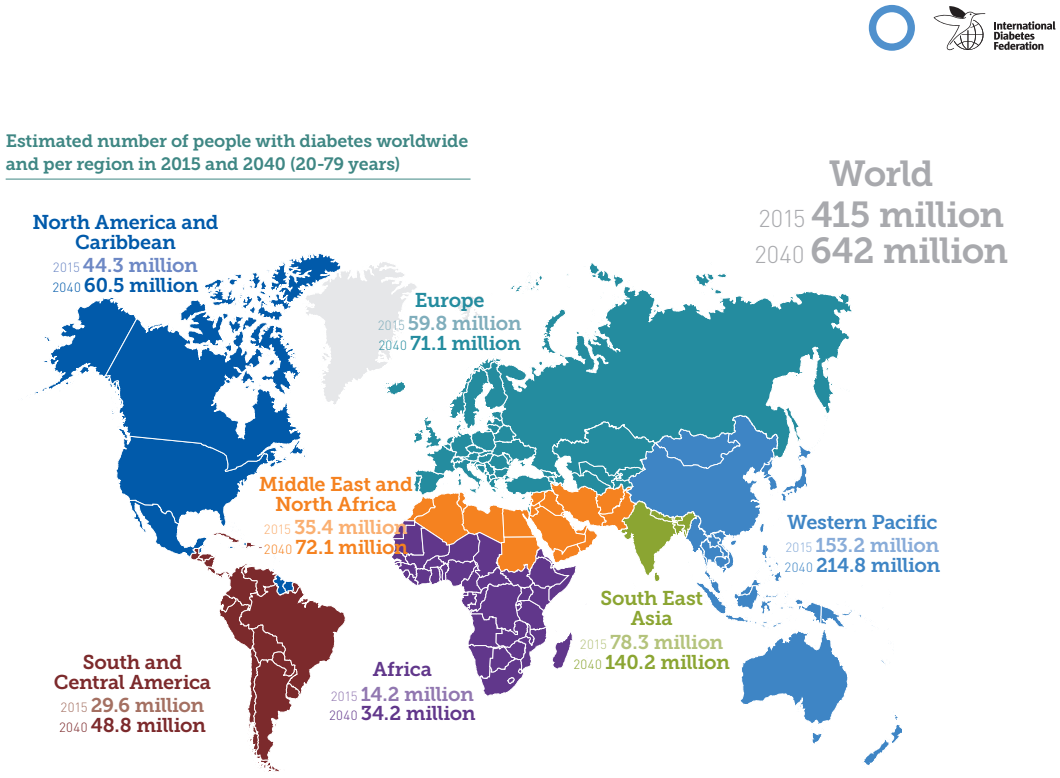


Figure 1.1: Number of individuals affected by Diabetes worldwide. Permission to reproduce this figure has been provided by the IDF which holds the copyright for this figure[1].

1.2 Pathophysiology of T2D

1.2.1 Clinical presentations

Diabetes is a group of metabolic diseases with the main clinical and diagnostic feature being hyperglycaemia. This is caused by an insufficient amount of insulin produced by the body and an inability of the body to use insulin efficiently. T2D accounts for about 87-91% of diabetes cases in high-income countries which makes it the most common form of the disease. The remaining cases represent with Type 1 Diabetes (T1D) induced by an auto-immune response against insulin producing beta-cells, Maturity Onset Diabetes of the Young (MODY) involving monogenic mutations affecting beta-cell function and other rare types of diabetes. The latter include gestational diabetes, syndromic forms of the disease and drug/chemical-induced diabetes[1].

The following diagnostic criteria have been established for the diagnosis of diabetes: Fasting Plasma Glucose (FPG) $\geq$126mg/dL (7.0 mmol/L) with fasting being defined as no intake of calories for a period of at least 8 hours or random plasma glucose levels $\geq$200mg/dL (11.1 mmol/L). In addition, 2-hour plasma glucose $\geq$ 200mg/dL after an oral glucose tolerance test and glycosylation levels of haemoglobin (Hb1AC $\geq$ 48mmol/mol or 6.5%) can be used for diagnosis. In case the obtained results for hyperglycaemia are not clear, repeated testing should be performed to determine the presence of hyperglycaemia[2].

While the classical symptoms of substantial hyperglycaemia usually comprise polyuria (excessive production of urine) and polydipsia (abnormal excessive thirst) as well as weight loss and blurred vision, T2D often remains undiagnosed for several years due to a slow progression of the hyperglycaemic state that is usually only associated with mild symptoms. Despite the lack of accurate diagnosis, patients with raised glucose levels still have an increased risk of diabetes-associated health problems. While the short-term complications of T2D are often considered as less severe, long-term effects of marked hyperglycaemia can lead to drastic and life threatening conditions. These include a large number of macrovascular complications such as atherosclerosis influencing risk for myocardial infarcts, cerebrovascular disease associated with stroke, and peripheral arterial disease often leading to amputations. In addition, raised long-term hyperglycaemia also increases

microvascular disease risk including retinopathy induced vision loss and nephropathy associated with reduced kidney function.[2,3]

1.2.2 Aetiology of T2D

T2D is typically characterised by a combination of insulin resistance observed in peripheral tissue, and decreased insulin secretion by pancreatic beta-cells[4]. The exact mechanisms involved in these processes are not clear and patients usually represent with heterogeneous disease states showing a different balance of physiological defects and different rates of disease progression and complications. This suggests that the relative contribution of insulin resistance and beta-cell failure to the development of T2D may differ across patients[5].

1.2.2.1 Insulin resistance and obesity

Insulin is an important anabolic hormone which, together with glucagon, ensures tight regulation of plasma glucose levels, that, in healthy individuals, range from 4-7mM. Shortly after food consumption, insulin acts in several ways to reduce blood glucose: Insulin supports both glucose uptake into peripheral tissue, such as skeletal muscle, and glycogenesis and glycogen storage in the liver. Insulin also acts on pancreatic alpha-cells to reduce glucagon production which suppresses hepatic glucose production through glycogenolysis and gluconeogenesis. In addition to these blood glucose lowering effects, insulin also positively regulates the synthesis of fat (lipogenesis), triglyceride storage in adipose tissue, protein production in liver and muscles and promotes cell proliferation[6,7]. Insulin sensitive tissue, in return, provides feedback, likely mediated by both neuronal and humoral mechanisms, to the beta-cells about the need for insulin secretion. In the case of insulin resistance, which often precedes T2D, cells have a reduced ability to respond to insulin. This leads to decreased effect of insulin on these tissues and, thus, glucose uptake and glucagon production is lowered while lipid levels are raised. To maintain normal glucose levels despite reduced insulin sensitivity, beta-cells increase insulin production which leads to an hyperinsulinemic state until beta-cells can no further increase insulin production. Thus, glucose levels rise leading to a prediabetic state which develops into T2D once beta-cell function deteriorates in genetically predisposed

individuals causing a more severe hyperglycaemic state defined as T2D[5,8].

Insulin sensitivity is influenced by both genetic and environmental factors. Mutations in several genes related directly to insulin signalling or adipose function have been found to cause severe monogenic forms of insulin resistance. Mutations in the insulin receptor gene (*INSR*) lead to extreme hyperinsulinemia but normal lipid profiles[9,10] while genetic defects in several other genes associated with either insulin signalling or adipose tissue function such as *AKT2*[11], *CIDEC*[12], *LMNA*[13,14], *PPARG*[15] and *ZMPSTE24*[16] cause insulin resistance and/or partial lipodystrophy. In addition, recent studies showed that insulin sensitivity also represents a complex trait which is influenced by a large number of different loci[17–19] which will be described in more detail later.

Apart from genetic factors, insulin resistance is also influenced by a number of environmental factors. Excessive calorie consumption, reduced energy expenditure, imbalanced nutrient content (including high-fat content), inflammation and ageing can affect insulin resistance as well as obesity. The latter is one of the strongest associated risk factors for T2D development. Accordingly, weight loss induced by lifestyle adjustments (dietary changes and exercise), bariatric surgery or drug treatment usually slows down or even stops T2D progression.[5,17,20].

In addition to lifestyle changes and treatment options that reduce body weight, several pharmacological agents have been developed that directly increase insulin sensitivity. For instance, thiazolidinediones are one class of insulin-sensitising drugs that act on the nuclear receptor transcription factors PPARG most highly expressed in adipose tissue. PPARG activation in fat tissue leads to removal of fatty acids in the circulation which in turn increases insulin sensitivity of muscle and liver, thus, leading to an overall decrease in plasma glucose levels. Thiazolidinediones have been suggested to increase cardiovascular risks[21,22]. However, the results from different studies were not entirely conclusive[23–25]. One of the most frequently used anti-diabetic drugs called Metformin also has a positive effect on insulin sensitivity (among other effects). Although the exact molecular mechanisms of Metformin action are not well defined, it predominantly reduces hepatic glucose production and increases insulin sensitivity of peripheral tissue[26].

1.2.2.2 Insulin secretion and pancreatic ß-cells

As described above insulin resistance contributes significantly to the development of T2D; however, pancreatic beta-cells can often compensate for the reduction in insulin sensitivity by increasing insulin production. However, when this compensatory mechanism fails (e.g. due to functional exhaustion of beta-cells), a full T2D disease state develops with pronounced hyperglycaemia. Blood glucose levels are sensed by beta-cells which in response to high glucose levels secrete insulin, a process defined as Glucose-Stimulated Insulin Secretion (GSIS). Although several other factors such as non-glucose nutrients, hormones and neuronal signals influence GSIS, the following basic model has been established (Fig1.2). Glucose enters the beta-cell via the GLUT2-transporter and is converted through phosphorylation into Glucose-6-phosphate by glucokinase (GCK). Glycolysis in the mitochondria as part of the Krebs cycle leads to the generation of ATP. The accumulation of ATP changes the ADP:ATP ratio in the cell and induces closure of ATP-sensitive potassium channels. This leads to membrane depolarization followed by calcium channel opening. The influx of calcium causes secretory granules derived from the endoplasmatic reticulum to fuse with the plasma membrane leading to exocytosis of insulin from the cell. The secreted insulin then causes a reduction in blood glucose, which, together with other mediators, leads to a feedback loop of insulin secretion[27,28].

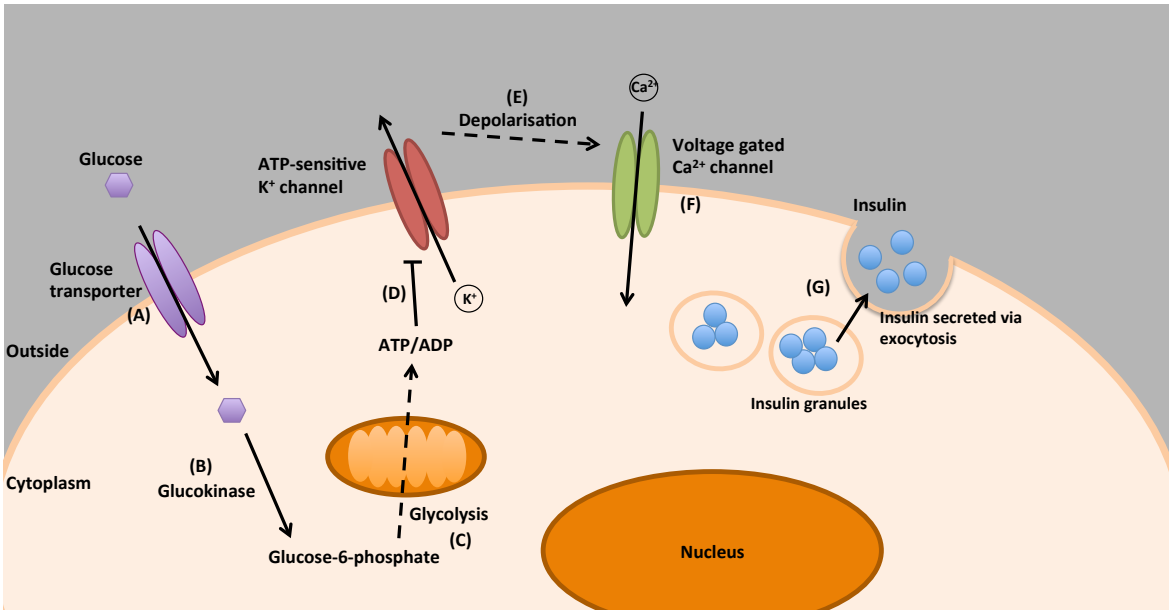


Figure 1.2: Glucose-stimulated insulin secretion. (A) Glucose enters the cell via a Glucose transporter upon which (B) it is converted into glucose-6-phosphate by Glucokinase. Subsequently, (C) glucose is broken down via glycolysis during the Krebs cycle in mitochondria which leads to the generation of ATP. This results (D) in changes in the ADP:ATP ratio which causes closure of ATP-sensitive calcium channels. (E) Closure of these channels induces membrane depolarization which in turn (F) leads to opening of voltage gated calcium channels. Calcium enters the cells which results in the fusion of secretory granules containing insulin, thus, causing exocytosis of insulin which ultimately will lower glucose level creating a feedback loop[27,28].

Similar to insulin resistance, both genetic and environmental factors influence the GSIS response of beta-cells. Valuable molecular insights were provided by a large number of monogenic forms of insulin secretion deficiency caused by genetic defects in single genes associated with beta-cell function, pancreatic development and beta-cell mass. For instance, mutations in *KCNJ11*[29–31] and *ABCC8*[32,33] often lead to hyperglycaemia shortly after birth which is defined as "neonatal diabetes". Both *KCNJ11* and *ABCC8* encode components of ATP-sensitive potassium channels (KIR6.2 and SUR1, respectively), which were known targets for drugs named Sulphonylureas. These medications increase insulin secretion by binding to SUR1 and closing the potassium channel[34]. These monogenic disorders present with very heterogeneous phenotypes that depend on both the type of genetic mutation as well as the affected gene. For instance, *GCK*[35] compound heterozygotes or homozygous loss of function leads severe hyperglycaemia due to the complete loss of sensing glucose by the beta-cell at birth. In contrast, activating mutations of *GCK* cause hypoglycaemic hyperinsulinemia[36,37] and heterozygous mutations in *GCK*[38] cause mild hyperglycaemia from early age with otherwise few symptoms. This is described as Maturity Onset Diabetes of the Young 2 (MODY2). MODY represents a group of disorders predominantly caused by mutations in *GCK* as well as *HNF4A*, and *HNF1A* although several other genes have also been identified to cause subtypes of MODY. Studying these monogenic forms of disease and associated phenotypes has substantially increased the understanding of the molecular functions of beta-cells. In addition to the identification of single gene deficiencies in insulin secretion, a large number of genetic studies has highlighted that GSIS also represents a complex trait influenced by a large number of genetic variants as described below[27,39,40].

1.3 Genetics of T2D

1.3.1 Heritability of T2D

T2D is a complex genetic disorder with both genetic and environmental factors influencing disease risk. The observed increase in T2D cases worldwide over a relatively short period of time suggests that changes in the environment are mainly responsible for this increase. These include changes

in lifestyle related to diet and energy consumption which also raised the prevalence of obesity. As described above obesity is a major risk factor for T2D development[41]. It was suggested that in the past "energy-saving" genotypes were preferentially selected and slowly evolving genetic factors did not have time to adapt to the new environment introduced by Western lifestyle, thus, providing a potential explanation for the epidemic prevalence of T2D and obesity[42]. Also, this highlights the importance of genetic factors in determining the risk of developing T2D in response to the environmental stress. Depending on the study and time of follow-up, estimates of the narrow sense heritability of T2D range from 30%-70%[43]. Also, a large number of studies have shown that T2D clusters in families and individuals with one affected parent have a 40% risk of developing T2D while individuals with both parents affected have a 70% risk of developing T2D during their lifetime. Similarly, twin studies provided high concordance estimates of 70% in monozygotic twins and 20-30% in dizygotic twins[44–47]. While these estimates vary across different studies due to differences in the population and cohort, these results clearly support the role of genetic factors in determining the lifetime risk of developing T2D under environmental conditions inducing metabolic stress[42,48].

1.3.2 Genetic studies and T2D

The earliest genetic efforts to identify T2D risk variants were family-based linkage studies and candidate gene studies. While linkage studies were successful in identifying rare high-risk variants associated with monogenic forms of diabetes such as MODY and neonatal diabetes, failure to find consistent linkage indicated that the same kind of rare high penetrant variants were not likely to be major players in common forms of diabetes. Similarly, candidate gene-based association studies showed also limited success apart from coding variants in the genes encoding known diabetic drug targets *PPARG*[49] and *KCNJ11*[50] that were consistently replicated in later studies[48].

Advances in genotyping array technology and understanding of the haplotype structure through the HapMap project gave rise to a new era of Genome-Wide Association Studies (GWAS). These studies exploited extensive Linkage Disequilibrium (LD) structures observed in the human genome, tagged by common Single-Nucleotide Polymorphisms (SNPs). This approach provided for the first time the opportunity to systematically analyse the role of common genetic variants across the

genome in disease traits without requiring any knowledge of the underlying disease biology. After the first phase of GWAS[51–54], results across different studies were combined to perform a meta-analysis in larger number of samples to increase power[18,55]. In addition, studies in non-European cohorts[56–62] were performed and the genetic data from different populations was combined for trans-ethnic analysis[63]. Most recently, a large-scale study also investigated the role of rare and low-frequency variants in the genetic susceptibility of T2D and based on this study these variants are not likely to play a major role in the genetic architecture of T2D[64] (as will be discussed in more detail later). Together these approaches were highly successful in determining T2D (or related traits) risk variants and to date, GWAS have found more than 150 lead variants at approximately 120 loci in the genome. Although large numbers of associated variants were identified, characterising the biological mechanisms acting at these GWAS signals, which is important for clinical translation, remains difficult. Low effect size estimates and the lack of obvious candidate genes are some of the main reasons hindering the efficient interpretation of these variants. Nevertheless, GWAS results have provided useful insight into the pathophysiological mechanisms of T2D by showing that the majority of variants are associated with insulin secretion while only a much smaller number contributes to insulin resistance. Also, GWAS have helped to dissect the genetic architecture of T2D in more detail[42,64].

1.3.3 The role of coding vs non-coding variation in the genetic susceptibility of T2D and other disease traits

For most complex traits, including T2D, the vast majority of identified GWAS variants is found in the non-coding genome. In contrast, protein-coding SNPs represent only a small number of GWAS variants (despite enrichment of coding variants in GWAS signals when taking into account the low exonic content of the genome). Limited functional knowledge and lack of comprehensive annotation of the non-coding genome has made the inference of biological mechanisms at GWAS loci difficult. To improve the understanding of non-coding regions, recent large-scale projects such as the Encyclopedia of DNA Elements (ENCODE) Project Consortium[65], the NIH Roadmap Epigenomics Mapping Consortium[66] and BLUEPRINT Consortium[67], have started to systematically annotate non-coding regulatory regions using various biochemical assays. These assays include

- Chromatin Immunoprecipitation (ChIP[68]) to determine the location of histone marks and transcription factor binding sites
- DNase-seq to determine DNase hypersensitivity sites indicative of open chromatin regions[69]
- bisulphite conversion approaches to map DNA methylation[70,71]
- transcriptional assays to determine gene and non-coding RNA expression[72,73].

While these efforts highlighted the complexity of non-coding regulatory elements that show variable activity across cell-types and developmental stage, they have been tremendously useful in determining relevant biology for GWAS signals.

Genomic annotations can be used for global GWAS analysis to identify relevant non-coding annotations, biological pathways and causal variants[74,75]. A number of studies have shown enrichment of GWAS regions in cell-type-specific enhancer regulatory elements which highlighted the importance of distal regulatory elements and may indicate disease-relevant tissue for future studies[76–82]. For instance, as described in more detail in Chapters 3-5, T2D GWAS variants were found to be enriched in human pancreatic islet stretch and active enhancers regions as well as islet-relevant transcription factor binding sites[83–86], thus, prioritising these variants for follow-up experiments in human islets or human islet cell models. Similarly, BMI-related GWAS regions were found to be enriched in promoter and enhancer regions of neuronal tissues[87].

The identification of functional variants through overlap of regulatory annotation is particularly useful at GWAS loci with multiple-disease-associated variants due to LD. For instance, Harismendy et al 2011 used an integrative analysis combining sequencing, bioinformatic data and biological assays to determine potential molecular mechanisms involved in coronary artery disease (CAD) at the 9p21 locus. This region harbours 8 significantly associated CAD or other cardiovascular disease loci, thus, making it difficult to determine the causal CAD variant. After sequencing the region, they found two variants rs10811656 and rs10757278 in high LD (r2>0.8) with the GWAS lead SNPs at which the risk allele disrupted STAT1 binding in an active enhancer element in HeLa cells based on chromatin signatures and the activity in luciferase reporter experiments. Additional experiments including expression analysis and chromatin capture in human vascular endothelial cells

revealed that the risk alleles of rs10811656 and rs10757278 prevent not only interferon-gamma-dependent STAT1 binding at the enhancer but also modify long-range chromatin interactions and expression of several neighbouring genes[88]. Although these results did not provide full evidence that rs10811656 and rs10757278 are the causal variants at this locus, this study clearly demonstrated how the combination of GWAS signals with non-coding annotations could be useful to indicate functional regulatory variants as well as novel biological mechanisms. Similar efforts of combining regulatory information with genetic data to highlight regulatory mechanisms have been performed for T2D relevant datasets and the analysis in Chapter 5 will describe the results of such an integrative approach in more detail.

1.3.4 Importance of GWAS data to highlight disease-relevant tissue which for T2D GWAS loci points towards pancreatic beta-cells

The results from GWAS have also provided certain useful insights into the pathophysiological mechanisms underlying the genetic susceptibility of T2D and other diseases. For instance, the aforementioned enrichment of BMI GWAS variants in non-coding regulatory elements in neuronal tissues indicates an important role of the central nervous system in governing the genetic susceptibility of obesity[87]. Also, large-scale efforts that combined T2D GWAS data with measures of insulin processing, secretion and resistance across a large number of individuals have shown that defects in insulin secretion are likely the primary driver of the genetic susceptibility of many T2D GWAS signals. In contrast, only a small number such as *PPARG, KLF14, IRS1* and *ADAMTS9* are associated with insulin resistance[18,19].

Further support for the role of the islet-cell in T2D susceptibility was provided by the fact that several T2D GWAS regions are adjacent to genes with a known role in monogenic forms of diabetes. These include genes such as *GCK, GCKR, HNF1A, HFN1B, HNF4A, INSR, KCNJ11, LMNA, SLC2A2* and *WFS1* of which the majority is associated with beta-cell function rather than insulin resistance[39,40]. Thus, T2D GWAS results clearly suggest that the pancreatic beta-cell responsible for insulin secretion is one of the key disease-relevant tissues through which genetic variants exert their effect on T2D susceptibility. Considering the importance of non-coding regulatory elements in

GWAS signals and the cell-type specific nature of these elements, knowledge about disease-relevant tissues is critical to infer biological mechanisms from GWAS data.

This is exemplified by the *FTO* locus that is the most strongly associated BMI GWAS locus and through its correlations with obesity is also associated with T2D in GWAS studies. A recent study[89] overlapped GWAS risk variants at the *FTO* locus with regulatory information from 127 different cell types and found a unique long enhancer domain in adipocyte progenitor cells with strong haplotype-dependent regulatory activity in human adipocytes. Shorter enhancer regions in other tissues such as brain did not show any genotype-dependent effects.

Further experiments involving chromatin capture techniques in preadipocytes identified downstream target genes with long-range chromatin interactions that influenced gene expression of *IRX3* and *IRX5* in a genotype-dependent manner. Consistent with the developmental regulatory role of *IRX3* and *IRX5* these genotype-correlated differences in expression were not observed in whole-adipose tissue indicating a cell-type specific effect in preadipocyte cells. Through further experiments including CRISPR genotype modifications, this study identified a causal mechanism governing obesity risk at this locus. The risk allele of the variant rs1421085 was found to disrupt the Transcription Factor (TF) motif of a highly expressed adipocyte TF ARID5B thereby inhibiting the repressive effect of ARID5B on *IRX3* expression. Overexpression of *IRX3* was found to decrease mitochondrial function and thermogenesis in human adipocytes while an anti-obesity phenotype including increased energy expenditure was observed in *Irx3* whole-body knockout mice. The effect on energy expenditure was shown to be tissue-specific to adipose cells since differences in energy regulation were only observed when knocking *Irx3* out in adipose cells but not in the hypothalamus.

In summary, this study determined a causal *FTO* GWAS enhancer variant at which the risk allele modified in a tissue-specific manner *IRX3* and *IRX5*, gene expression via modified ARIDB5 TF binding. This led to a shift in adipocyte development that affected energy expenditure via thermogenesis. Thus, these data highlighted molecular pathways at the *FTO* GWAS locus that influence obesity risk and may ultimately lead to the development of new anti-obesity drugs targeting these pathways[89]. This study clearly showed the importance of interrogating disease-relevant tissue; however, it can not rule out the existence of alternative molecular mechanisms influencing obesity

risk at the *FTO* locus. For instance, a previous study identified genotype-dependent *IRX3* expression changes in human cerebellum[90] suggesting that neuronal factors may also govern genetic susceptibility of obesity at the *FTO* locus.

1.4 Genome regulation and causal mechanisms at GWAS loci

1.4.1 Challenges of defining causal GWAS variants

GWAS were very successful in identifying a large number of regions associated with disease and quantitative traits. However, the identification of causal variants and exact molecular mechanisms governing T2D or other disease risk has proven more challenging. The lack of understanding how GWAS signals contribute to biology has hindered the translation of genetic results into clinical practice.

There are several challenges associated with the identification of causal functional mechanisms at T2D GWAS loci.

1. the LD patterns and haplotype structures exploited by GWAS to identify associations make it difficult to actually determine the causal variant. That is, the SNPs tested for association with disease traits on array platforms were selected to tag LD blocks of several kb in size containing a large number of variants. These LD patterns do not allow to statistically distinguish between multiple variants equally associated with the disease trait.

2. small effect sizes of GWAS variants may translate into small biochemical effects that could require large numbers of samples to establish statistical significance.

3. once biochemical assays indicated variants of potential upstream functional significance, the downstream gene targets regulated by the variant and the biological mechanisms causing the disease associations still remain unclear. Since the majority of GWAS hits are found in non-coding regulatory regions near genes that often do not represent obvious candidate

genes establishing the link between upstream functional variant and downstream gene target remains difficult. While GWAS loci are usually named according to the nearest gene this does not indicate that this gene plays a role in the susceptibility of a disease[74,75].

Due to the aforementioned reasons the causal variants and downstream effector transcripts for most GWAS hits are still not fully understood. However, recent research has definitely improved our understanding and highlighted potential causal evidence at a number of GWAS loci, including T2D loci[74,75].

1.4.2 Using Fine mapping to establish causal variants

The success of GWAS was largely dependent on the existence of extended LD patterns. These patterns allowed for the design of efficient genotyping platforms that interrogated the whole-genome by assaying common tag SNPs (Minor Allele Frequency >5%) in these LD blocks. While this approach facilitates the discovery of large GWAS regions spanning several kilobases, it is at disadvantage when trying to pinpoint causal variants and to distinguish multiple association signals at a locus. To overcome these limitations, fine-mapping approaches have been developed that aim to reduce the size of genomic intervals and variant sets at GWAS regions. In contrast to GWAS arrays, fine-mapping requires the creation of extensive variant sets to be certain the causal variant is likely genotyped or imputed. This is followed by statistical analysis that breaks down the LD pattern since functional analysis of the identified variants would often not be feasible (although recent technological advances, such as massive parallel reporter assays, have substantially improved the ability to perform high-throughput functional experiments[91]). Resequencing is one of the most accurate ways to catalogue common and rare variants in GWAS regions; however, the large sample size required to detect small effect sizes observed in T2D GWAS makes this method prohibitively expensive for T2D case-control studies. Thus, customised genotype arrays providing high-density coverage in GWAS regions that can be imputed in publicly available large-scale sequencing panels such as 1000G have been established as cost effective alternative for fine-mapping[92]. For instance, the Metabochip array[93] contains approximately 200,000 variants at GWAS loci of cardiovascular and metabolic traits including T2D and has recently been successfully used to create credible vari-

ant sets of 49 distinct association signals of GWAS loci using a Bayesian factor analysis[83]. These credible variant sets can be interpreted in a similar way to a confidence interval determined from frequentist approaches. For each variant in the set, an estimate is derived from the fine-mapping data that represents the Posterior Probability (PP) for a variant to drive the GWAS signal at a given locus (assuming a single causal variant). The 99% credible set represents the variants that collectively capture 99% of the PP at the locus. Overlap with disease-relevant tissue regulatory annotations can then be used to further refine the causal variant or to highlight the causal mechanism[92]. For instance, using metabochip, a recent study showed that at 10/49 distinct association signals (at 39 established T2D GWAS loci) the number of 99% credible set variants was less than 10. At the *MTNR1B* locus, the 99% credible set in fact contained only the single variant rs10830963 and it was possible to show that the risk allele induced NEUROD1 binding and increased enhancer activity leading to higher *MTNR1B* expression[83].

An alternative approach to customised chips is trans-ethnic fine-mapping meta-analyses that exploit differences in LD pattern across different ethnicities with the underlying assumption that the causal variant is shared across populations. The haplotype diversity in the different populations is expected to reduce the number of credible set variants to ones that are in high LD with the causal variant across all trans-ethnic groups[92]. The advantages of trans-ethnic analysis were highlighted recently when contrasting the results from European only and trans-ethnic meta-analysis. For almost all investigated loci the number of variants in the credible set was reduced through the trans-ethnic meta-analysis compared to the European analysis. The best improvement was seen at the *JAZF1* locus at which the European analysis identified 9 credible set variants that spanned 76kb while the trans-ethnic meta-analysis reduced the number to 4 SNPs within a region of 16kb[63]. One of the 4 remaining credible set SNPs mapped to an enhancer region, showed allelic imbalance in enhancer activity in human islets and was associated with *CREB5* expression[94].

1.4.3 Functional methods to link disease regions to GWAS hits

As described fine-mapping represents a powerful approach to reduce the number of potential causal variants, however, it does not provide any functional annotation. In the following sections various attempts to ascribe function to GWAS loci and highlight potential molecular mechanisms involved

in disease risk will be described.

1.4.3.1 DNA methylation

DNA methylation is an important heritable epigenetic mark that contributes to a number of biological regulatory processes including development, X-chromosome inactivation, genomic imprinting and regulation of gene expression. DNA methylation occurs in vertebrates predominantly at the 5' position of cytosines adjacent to guanines defined as CpG sites most of which show high and stable methylation levels in mammalian genomes. The self-complementary nature of CpGs allows for the heritable propagation of the DNA methylation status through cell division by the maintenance DNA methyltransferase enzyme DNMT1. In a process similar to the semi-conserved DNA replication mechanism, the enzyme recognises hemi-methylated CpG sites in the genome and restores the methylation state at the unmethylated CpG site. While the majority of the genome is highly methylated, CpG-islands (CGI) are mainly unmethylated. CGIs represent very CpG-dense regions which are predominantly found in 5' promoter regions.[95]. While DNA methylation is largely influenced by methyltransferase activity, recent studies showed that the genetic code may also influence DNA methylation levels in the form of methylation Quantitative trait loci (mQTL)[96–104]. mQTLs are defined as genetic variants at which the genotype is strongly correlated with nearby CpG-methylation levels, and these genotype-methylation associations are often tissue-specific[105].

Given the important regulatory role of DNA methylation, it is not surprising that an ever increasing number of both monogenic and complex diseases has been associated with changes in DNA methylation. Imprinted regions are highly susceptible to be involved in clinical disorders since genes in this regions only have one active copy; thus, the loss of a single functional allele due to genetic and/or epigenetic mutations can have severe consequences. Several imprinting diseases have been identified which are at least partially caused by changes in CpG-methylation pattern at imprinted regions[106]. These include Prader-Willi Syndrome[107,108] and Angelman syndrome[109] associated with deletions on chromosome 15 in the imprinted region 15q11 .

There is also accumulating evidence for a role of DNA methylation in the aetiology of complex disease phenotypes. GWAS regions associated with complex disease phenotypes are predominantly

found in non-coding regions with little knowledge regarding the precise molecular mechanisms driving disease association; it has been suggested that some GWAS loci may mediate their effect on the disease phenotype via methylation. This idea has been supported by several observations. Firstly, the addition of methyl groups to CpG sites renders them much more susceptible to DNA mutations thus, creating mutation hotspots in the genome[95]. Secondly, GWAS regions associated with complex diseases were found to be enriched in disease differentially methylated regions (DMRs), which differ between healthy and affected individuals, as well as genetic mQTL variants associated with DNA methylation of nearby CpG-sites[110]. For instance, adipose and fetal brain mQTLs were previously linked to distal regulatory elements associated with metabolic disease[100], bipolar disorder[111] and schizophrenia[96] GWAS regions. Studies also found enrichment for mQTLs and DMRs in T2D GWAS regions, and these are discussed in more detail in Chapter 3. While limitations of defining directional causality also apply to these studies, the integration of genetic and epigenetic data provides still useful information to prioritise potential genetic upstream effector variants and downstream gene targets[105]. An adipose tissue study investigating metabolic disease identified a significant mQTL SNP rs713586 that was a lead BMI GWAS variant and correlated with DNA methylation in an enhancer element near the *ADCY3* gene which was previously linked to obesity. The variant rs713586 was also in high LD with the SNP rs6749422 which was present in multiple transcription factor binding sites (TFBS), thus providing a potential molecular mechanism that may be responsible for the genetic association[100]. Despite the usefulness of combining genetic and epigenetic data to understand GWAS signals of complex disease phenotypes, these studies still remain complicated since epigenetic regulatory regions are often highly tissue-specific. For instance, Ziller et al 2013 showed that tissue DMRs, which are highly cell-type-specific methylation regions, are mutation hotspots and strongly enriched in distal regulatory elements such as enhancers and TFBS. In addition, these tissue DMRs also represent mutation hotspots and are highly over-represented near GWAS hits[112]. Based on these data it can be concluded that the interpretation of GWAS signals using epigenetic context requires knowledge about and access to disease-relevant tissue which may not be possible for all complex diseases. Further information about human pancreatic islet DNA methylation, the identification of hypomethylated regulatory elements in human islets and the role of these elements in T2D genetic susceptibility will be provided in Chapter 3.

1.4.3.2 Chromatin structure and transcription factor binding

Chromosomal DNA in eukaryotic genomes is always associated with chromatin. That is, about 147bp of DNA are periodically wrapped around histone protein subunits (H2A, H2B, H3 and H4) that form octameric structures termed nucleosomes which are separated by about 50bp of linker DNA and a linker histone named H1[113]. Both histone and nucleosomal structures influence DNA accessibility and broadly split the genome into accessible and transcriptionally active euchromatin and in inactive heterochromatin, thereby regulating cell-type specific transcription, splicing, responses to DNA damage and overall chromatin structure. Dynamic changes of histones in the form of posttranslational modifications (PTMs) of amino acid tails such as methylation, acetylation and phosphorylation influence the chemical properties of histones and their affinity for DNA or other DNA-binding proteins such as transcription factors (TFs)[114,115]. Active euchromatin is associated with high levels of acetylation at H3K27 and trimethylation of H3K4, H3K36 and H3K79, while heterochromatin is marked by low acetylation of H3K27 and increased methylation of H3K9, H3K27 and H4K20[116,117]. These PTMs are controlled by a large number of enzymes that can either add or remove certain types of PTMS and are generally described as histone writers or erasers. Histone writers include methyltransferases and histone acetyltransferase (HAT) that increase methylation and acetylation levels, respectively, while histone erasers represent histone demethylases (HDAC) reducing methylation and histone deacetylases decreasing acetylation modifications. PTMs are often altered during transcription through transcriptional co-activators with HAT domains or corepressors that have HDAC domains[113,117,118]. Recently it was also suggested that the main transcriptional machinery including RNA polymerase II could recruit HDACs and HATs to actively transcribed regions to regulate these genomic segments[119]. In addition to histone PTMs, nucleosomes also influence chromatin structure and DNA accessibility, particularly, at Transcription Start Sites (TSS) to regulate initiation of transcriptional activity by providing access to TFs and other DNA-binding proteins. Removal of nucleosomes from the TSS is associated with increased gene expression while the accumulation of nucleosomes at the TSS leads to decreased transcription[120]. This process of chromatin remodelling is controlled by large complexes which cause ATP-dependent restructuring or destabilisation of nucleosomes. Different types of remodelling complexes exist that have marginally different roles in the regulation of chromatin structure[113,121].

Similarly to DNA methylation, chromatin also has an important regulatory role and thus, is critical for both monogenic and complex diseases. For instance, altered histone modifications are associated with Rubinstein-Taybi syndrome. It is a genetically heterogeneous autosomal dominant neurodevelopmental disease characterised by intellectual disability, facial abnormalities, heart defects and increased tumour risk. In the majority of cases Rubinstein-Taybi syndrome is caused by dysfunction of the HAT activity of the cAMP-response element binding protein (CPD)[122], thus, leading to disturbed histone acetylation in the genome which in some cases can be treated by HDAC inhibitors[106]. Defects related to chromatin structure are also associated with neurodevelopmental disease. The X-linked alpha-thalassemia mental retardation (ATRX) syndrome is a clinically heterogeneous disease with affected individuals showing severe intellectual disability and a range of other symptoms. ATRX is caused by mutations in the *ATRX* gene which encodes a chromatin remodeler localised to pericentromeric heterochromatin. On the molecular level mutations in *ATRX* are associated with a range of epigenetic abnormalities including modifications in imprinting, abnormal methylation and downregulation of alpha-globin[106]. In addition to being involved in monogenic disease, chromatin accessibility and histone modifications are also useful for understanding complex disease biology by annotating the genome as described before in more detail.

Next generation sequencing techniques such as ChIP-seq and DNase-seq, which assay DNA binding proteins and open chromatin, have provided the opportunity to systematically annotate an ever increasing number of tissues for histone modifications and chromatin accessibility. However, these studies were often limited to providing a static view of the epigenomic landscape of rare primary tissues since high cell numbers (in the range of millions of cells) were required for these assays[118]. With the development of a novel more efficient approach such as ATAC-seq[123], which requires much fewer cells (about 10,000-50,000), future studies will be able to provide more detailed insight into the variability of chromatin elements across individuals and dependent on disease status (for a more in-depth comparison of biochemical chromatin assays see Chapter 4). For instance, recent studies have identified in certain tissues chromatin QTLs (chQTL) which are genetic variants at which the genotype is associated with chromatin accessibility potentially through modifying a TF binding event. A recent study showed that chQTLs identified from lymphoblastoid cell lines using DNase-seq modified the position of nucleosomes and were enriched for TFBS and GWAS SNPs[124].

Assessment of chQTL in (often rare) disease relevant primary tissue using an efficient method like ATAC-seq could provide valuable insight into disease pathophysiology. Chapter 4 will describe the analysis of human islet and iPSCs ATAC-seq open chromatin data to improve understanding of islet function and development and in Chapter 5 the islet ATAC-seq data will be used to identify highly T2D-associated variants with allelic imbalance in open chromatin to provide potential functional mechanisms at T2D GWAS loci.

1.4.3.3 Transcription and expression of genes

While DNA found in all tissues encodes the full information required to generate any type of protein, it is not directly translated into protein. DNA is initially transcribed in a tissue-specific manner into RNA, which later on is translated into protein. Hence, measuring RNA levels through either microarray technologies or RNA-sequencing provides valuable information on the expression and activity of genes in a given tissue. Since gene expression is highly heritable, it provides a powerful tool to link genetic variation to regulatory function through the identification of expression QTLs (eQTLs). eQTLs refer to genetic variants at which the genetic status is correlated with the expression of nearby genes (cis-eQTLs which are often defined to be within 1Mb of a gene) or distant genes (trans-QTLs which can be associated with the expression of any gene in the genome)[125]. Cellular context plays an important role for these associations which are often tissue-specific[126–128]. In addition to regulatory information on a given variant, eQTLs also link genetic variants to gene products and, hence, have been suggested to be extremely useful in elucidating the role of GWAS variants in complex diseases[129]. In recent years the GTEX project has been established to collect RNA-seq data on 60 tissues across 900 individuals to provide a resource database on the interaction between genetic variation and gene expression[130]. Consistent with the regulatory role of GWAS SNPs, eQTLs of disease-relevant tissue were found to be enriched in GWAS SNPs. While eQTLs are a highly useful tool to interrogate the regulatory role of GWAS variants, LD across multiple variants makes it difficult to pinpoint a causal variant. Additionally, eQTLs do not necessarily indicate a causal relationship between the genetic variant and gene expression, thus, experimental follow-up of these interactions is required to determine causality[129].

1.4.3.4 Integration of genetic, epigenomic and transcriptomic data

While the previous parts of the introduction have described a number of epigenetic (DNA methylation), epigenomic (chromatin accessibility and histone modifications) and transcriptomic annotations as individual entities, the reality is much more complex considering that none of these genomic annotations are acting in isolation. In fact, the majority of these annotations is highly interconnected and correlated in the genome. In a recent study, Ernst et al developed a software tool called chromHMM which represents a multivariate Hidden Markov Model (HMM). It was applied to 9 chromatin marks across 9 cell types to segment the genome into 15 types of regulatory states. This approach drastically reduced the possible number of combinations of investigated chromatin marks and identified poised and active promoters, weak and strong enhancers as well as transcriptionally active, insulator, repressed and inactive regions of the genome. Active regulatory states such as promoters and enhancers were found to be highly cell-type-specific and their activity correlated with gene expression of nearby genes. As described before, enhancers of disease-relevant tissue were enriched for GWAS hits confirming their relevance for complex disease biology[76]. Similarly, the Epigenome Roadmap Project investigated DNA methylation, chromatin accessibility and a large number of histone PTMs across 111 tissues and using chromHMM predicted up to 50 chromatin states by combining up to 24 histone marks in the model. Through this approach, they also investigated the co-occurrence of histone marks with open chromatin and DNA methylation. In general, active regulatory elements were found to be associated with H3K4me, H3K4me1, H3K27ac, open chromatin and either low-intermediate methylation (enhancers) or very low methylation levels (promoters). Although DNA methylation levels across different annotation varied according to the differentiation state of cell types, it was found that the activity of distal regulatory enhancer elements correlated with lower DNA methylation levels and higher levels of the histone mark H3K27ac. In accordance with previous results, this study also found that active regulatory elements and in particular, enhancers showed enrichment across GWAS traits and can be used to highlight disease-relevant tissue[82]. In addition to identifying disease-relevant regulatory states and tissues, another study investigating cardiac QT interval and QRS duration GWAS loci found that functional prioritisation of enhancer elements using a combination of histone marks, DNA methylation and open chromatin can be used to identify novel disease-relevant loci that have not passed genome-wide significance[131]. Overall these results indicate the importance of such integrative

approaches to uncover and understand biological mechanisms governing genetic susceptibility to complex diseases. The work in Chapter 5 will integrate newly generated and existing epigenomic human pancreatic islet datasets to refine islet chromatin states and identify likely causal T2D risk variants.

1.4.4 The contribution of rare strong effect size variants and copy number variants in disease

GWAS has been highly successful in identifying large numbers of loci associated with T2D, however, by design GWAS has mainly focussed on common variants with a minor allele frequency (MAF) >5%. Based on GWAS results the common-variant common-disease hypothesis has been used to describe the genetic architecture of common diseases. That is, a large number of common variants, each with small effect size contribute to common disease risk[132,133]. However, common genome-wide significant GWAS Single Nucleotide Variants (SNVs) identified to date only explain a portion of the estimated heritability and, hence, the rare variant hypothesis was proposed. Considering that rare variants collectively are very common throughout the genome, it was suggested that rare alleles with strong effect size contribute most to common disease risk[134–136]. However, recent evidence indicates that common variants with small effect size (often not passing genome-wide significance) explain most of the genetic susceptibility of common diseases or traits[137,138]. In the field of T2D, a recent genetic study investigating more than 100k T2D case-control samples using a combination of whole-genome sequencing, whole-exome sequencing and exome chip genotyping found that the genetic susceptibility of T2D is mainly explained by common variants and that low-frequency and rare variants likely do not explain a large portion of genetic T2D susceptibility. These results do not rule out though the existence of rare variants with strong effect size in the population that may provide useful insights into disease biology[64].

Protein truncating variants (PTVS) represent one type of strong effect size variants that perturb gene activity. PTVs are often also referred to as Loss of function (LoF) variants; however, this term is ambiguous since LoF on the gene level may cause Gain of function (GoF) on the protein level (e.g. when an inhibitory domain is affected). Hence, throughout this thesis, the term PTV

will be used. These variants may be used to evaluate the effect of reduced gene function *in vivo* since they are almost equivalent to human gene knockouts (in homozygous state) which may provide clinically relevant information[139]. For instance, PTVs of *CCR5* are associated with reduced susceptibility to HIV[140], reduced function in the gene *CASP12* decreases the risk of sepsis[141] and PTVs of *PCSK9* are associated with reduced Low-density lipoprotein levels and protection against atherosclerosis. Notably, mutations leading to enhanced protein function in *PCSK9* are associated with severely increased cholesterol defined as familial hypercholesterolaemia. These observations initiated drug development aiming to inhibit and replicate the positive effect of *PCSK9* PTVs which led to the development of PCSK9 inhibitor drugs that have been shown to reduce atherosclerosis risk.

This highlights that strong effect size variants provide useful insight into the physiological relevance of genes and may indicate new drug pathways. However, these variants are almost always rare and not well detected using current technologies and annotation approaches. SNVs, which only represent a subset of potential PTVs, can be detected with high accuracy nowadays; however, small insertions and deletions (indels) and larger Copy Number Variants (CNVs) or Structural Variants (SVs), which can also act as PTVs, remain difficult to detect using next-generation sequencing (NGS) technologies. Differences in read depth and the use of short reads makes it difficult to identify these indels and CNVs confidently. Once potential PTVs are identified, it remains challenging to annotate the variants correctly. In particular, due to the existence of different gene annotation datasets and tools that may provide contradicting results with respect to the location of a variant in a gene. Accurate information about the location of a variant in a gene is important for functional predictions on gene function. In addition to accurately identifying and annotating variants, determining the effect of the variant on protein function is even more challenging due to several reasons. Firstly, transcription levels of affected genes which are often simple to obtain for most tissues, may not necessarily correlate with protein levels due to post-translational regulatory mechanisms[142]. Secondly, due to the existence of multiple transcripts which are the results of alternative splicing, only some isoforms may be affected by an identified variant. Thirdly, compound heterozygous or other modifier mutations may modify the effect of a variant across individuals[143].

This thesis will mainly focus on CNVs which are one type of major impact variants. These repre-

sent regions of the genome at which gains or losses in copy number lead to duplications or deletions, respectively, and can affect several kb or even Mb in the genome. In particular, CNVs affecting the coding regions can act as PTVs (although this terminology may not be suitable for CNVs and hence, the term Protein Duplicating Variant (PDV) will be used together with PTVs to describe these variants better). In recent years it could be shown that CNVs frequently occur in the human genome and contribute significantly to human genetic variation and disease[144]. CNVs can be distinguished into two types. Recurrent CNVs generally are of similar size and often share breakpoints while non-recurrent CNVs vary in size, have different breakpoints and often show only partial overlap. Three main processes are involved in the generation of CNVs. Firstly, non-allelic homologous recombination occurs in regions of the genome with a large number of duplicated segments which can lead to misalignment and unequal crossing over during cell division and, thus, create often recurrent deletions or duplications. Secondly, non-homologous end-joining can result in deletions after a double strand break occurs in the DNA and the separated strands are ligated together. Thirdly, fork-stalling and template-switching may lead to the generation of CNVs during DNA replication when the replication fork comes to a standstill, and the lagging strand becomes disattached. When the disattached strand comes in contact with an upstream or downstream replication fork, this can lead to a deletion or duplication, respectively[145] .

While CNVs represent a form of genetic variation in healthy individuals, CNVs have also been associated with a variety of diseases, in particular, monogenic forms of disease. Syndromic disorders such as DiGeorge/Velocardiofacial syndrome[146] or Williams-Beuren syndrome[147] often involve relatively large *de-novo* microdeletions or microduplications with high penetrance and intellectual disability or congenital abnormalities as clinical phenotypes. However, in recent years it was discovered that CNVs also cause highly variable disease phenotypes. For instance, CNVs at 16p11.2 are associated with varying clinical features ranging from reduced intellectual ability, language impairment, autism, seizures and obesity or underweight depending on the CNV and affected region[148,149].

In addition to monogenic forms, CNVs have been suggested to play a role in the development of complex diseases. Studying CNVs in complex diseases may lead to the identification of novel GWAS hits that contribute to the missing heritability, and may potentially provide causal mech-

anisms at already established GWAS loci[150]. While for neuropsychological disorders such as schizophrenia and autism, CNVs may contribute substantially to genetic disease risk[151], it was shown in 2010 that common CNVs likely do not have a major impact on the genetic susceptibility of most complex diseases. However, these studies did not rule out the existence of rare CNVs, especially those in the coding region that may influence common disease risk[152]. Thus, novel approaches to study rare coding CNVs are needed to shed more light on the role of these variants in common disease pathophysiology. Chapter 6 will analyse the contribution of coding CNVs to T2D susceptibility and describe a novel pipeline for deletion detection from exome chip array data.

1.5 Aims of this thesis

- Characterising the whole-genome DNA methylome of human pancreatic islets using WGBS and determining the role of islet DNA methylation in T2D genetic susceptibility
- Exploring the open chromatin landscape of human pancreatic islets and induced Pluripotent Stem Cells (iPSCs) using ATAC-seq to refine the location of T2D risk alleles at T2D GWAS loci and improve the understanding of beta-cell development and function
- Integrating different types of genomic datasets with T2D genetic association data and testing highly associated variants overlapping enhancer states for allelic imbalance in islet open chromatin to define likely causal mechanisms at T2D GWAS loci
- Characterising the role of rare coding CNVs in T2D and obesity genetic susceptibility

CHAPTER 2

General Methods

2.1 Human islet tissue databank

Human adult pancreatic islets were isolated from deceased donors that have provided full ethical consent at either the Oxford islet isolation facility in OCDEM, University of Oxford, UK, or at the Alberta Diabetes Institute in Edmonton, Canada, through a collaborative effort with Dr Patrick McDonald. DNA was obtained from 100-150 hand-picked human pancreatic islets per sample to assay DNA methylation (Chapter 3), to interrogate the open chromatin landscape (Chapter 4) of human pancreatic islets and to integrate these datasets with other previously generated epigenomic data (Chapter 5). Sample details and characteristics are provided in the relevant Chapters 3-5.

2.1.1 DNA extraction from human pancreatic islets for bisulphite processing and DNA methylation analysis

DNA extraction was performed by one of the members of the Prof Mark McCarthy and Prof Anna Gloyn laboratory using one of the following methods.

2.1.1.1 DNA extraction based on phenol-chloroform extraction

DNA for all whole-genome bisulphite sequencing (WGBS) samples (n=10) and for a subset of the 450k samples (n=18) was extracted from a phenol-chloroform fraction that remained after RNA was extracted from human pancreatic islets (for other purposes) using a method named guanidinium thiocyanate-phenol-chloroform extraction[153]. A short description of the preceding RNA extraction follows below:

After isolation, islets were pelleted by centrifugation, washed using phosphate buffered saline (PBS, Sigma Aldrich) and lysed with Trizol. The lysate was transferred to a new tube and mixed with 200µl chloroform to separate the organic and aqueous phase. The latter containing the RNA was removed (and processed separately) while the remaining fraction (including DNA) was stored at -20°C for DNA extraction at a later point. After storage of the DNA-containing fraction (up to several weeks), the DNA was precipitated by addition of 300µl of 100% ethanol to the thawed solution. Next, the solution was incubated for at least 30minutes. In order to create a DNA pellet, the solution was centrifuged at 4000g for 5minutes at 4°C. Following removal of the supernatant, 0.1M trisodium citrate (Sigma Aldrich) in 10% ethanol was used to wash and incubate the pellet for 30 minutes at room temperature. The centrifugation, wash and incubation step was repeated and followed by another wash step using 75% ethanol. Subsequently, the pellet was air-dried for 10 minutes to allow removal of any residual ethanol and resuspended in at least 100µl 8mM NaOH (Sigma Aldrich) using a heated (50°C) plate mixer set at 400g for 1h.

2.1.1.2 DNA extraction using spin columns

Phenol-chloroform extracted DNA, as described above, often did not provide enough high-quality DNA for bisulphite conversion. Hence, additional islet DNA, used for the Illumina 450k array (Oxford islets n=13, Edmonton islets n=10, the latter were sent on dry ice), was extracted using the ReliaPrep gDNA Tissue Miniprep system (Promega) according to the instructions provided by the manufacturer. Specifically, islet tissue was initially digested with Proteinase K followed by cell lysis. Subsequently, RNase was added to the solution to remove RNA, and the samples were extracted using a spin-column.

2.1.2 Quality control (QC) of extracted DNA for bisulphite conversion

After extraction, the DNA was stored at -80°C. Before processing the DNA for bisulphite conversion, I performed DNA quality control. Specifically, the quality and quantity were assessed through spectrometry, picogreening and the use of the Agilent Bioanalyser (Agilent).

2.1.2.1 Spectrophotometer analysis

A ThermoScientific NanoDrop spectrophotometer was set to analyse dsDNA, and after blanking with nuclease-free water (1µl), 1µl of DNA from each sample was added to the pedestal. Subsequently, the absorbance was measured at wavelengths of 260nm and 280nm. The purity of the DNA was determined by the 260/280 ratio, and DNA was defined as pure if the 260/280 ratio fell between 1.8 and 2.1.

2.1.2.2 Picogreening of DNA

A DNA standard was prepared to range from 1000pg/µl to 25pg/µl by twofold serial dilution of DNA activated type XV standard (Sigma Aldrich). Picogreen dsDNA reagent (P7589 Component A, Quant-iT, Invitrogen) was diluted 1 in 200 in Tris-EDTA (TE, Sigma Aldrich) buffer. For each

DNA sample 50µl of 1 in 100 diluted DNA (in TE) were transferred into a well of a black assay plate (Optiplate 96F, from PerkinElmer) and combined with 50µl of diluted picogreen reagent. After adding picogreen reagent, the plate was covered and allowed to mix for 6 minutes. The fluorescence signal of the picogreen associated DNA was then measured using a laser implemented in the Cytofluor machine (Multi-well plate reader Series 4000, Perspective Biosystems). The final DNA concentration (representing double stranded DNA) was calculated using linear regression parameters derived from the standard DNA dilution series.

2.1.2.3 Agilent Bioanalyser run to test DNA integrity

The Agilent DNA 12000 chip was used in combination with the Agilent ABI 2100 Bioanalyser to determine the integrity of genomic DNA according to manufacturer instructions using 1µl of DNA per sample. Proprietary software of the Bioanalyser estimated concentrations and peaks of DNA using a dye that bound to the DNA and created a fluorescence signal when passing through a laser. Samples were used for the bisulphite conversion if no DNA was detected below a size of 1000bp.

2.2 Bisulphite conversion of human islet DNA

Bisulphite conversion is among the most common techniques to study DNA methylation. Treatment of DNA with bisulphite triggers the conversion of unmethylated Cytosines into Uracils while 5' methyl-Cytosines are unaffected. Following Polymerase Chain Reaction (PCR) amplification Uracils are converted into Thymine (Fig2.1) and based on the number of Cytosine to Thymine conversions in the genome the methylation levels for individual Cytosines can be detected by sequencing or microarray analysis.

2.2.1 Whole-genome bisulphite sequencing

400ng of DNA per human islet samples (n=10) were sent as part of a collaborative effort to Prof Vardhman Rakyan at the Blizard Institute, Queen Mary University, London, UK and bisulphite-

converted using the Ovation Ultralow Methyl-Seq DR Multiplex System 1-8 (Nugen). In short, the DNA was fragmented using sonication to 200-300 bp followed by end-repair to generate blunt ends and allow for adaptor ligation. The adaptor-ligated DNA was then bisulphite-converted and PCR-amplified (10 cycles) to produce the bisulphite-converted library for sequencing. The libraries were purified using Agencourt AMPure beads (Beckman Coulter) and sequenced as described in Chapter 3.

2.2.2 Illumina 450k array bisulphite conversion

18 samples were bisulphite-converted and processed as part of a collaboration with Prof Stephan Beck by members of his lab at the UCL Cancer Institute, University College London, London, UK. The remaining 21 samples were processed in OCDEM, University of Oxford, Oxford, UK, by myself. For both types of samples, the EZ DNA Methylation Kit (Zymogen) was used to perform the bisulphite conversion according to the instructions from the manufacturer. In detail, 1µg of DNA was used per sample in a volume of maximum 45µl and combined with a dilution buffer inducing DNA denaturation. Then the CT conversion buffer was added to the denatured DNA and the solution was incubated overnight for the bisulphite conversion. Following the conversion, the bisulphite-converted DNA was purified using a spin column extraction method provided by the kit. The bisulphite-converted DNA was then processed by the UCL or Oxford sequencing core using the Illumina 450k array beadchip according to the instructions provided by Illumina.

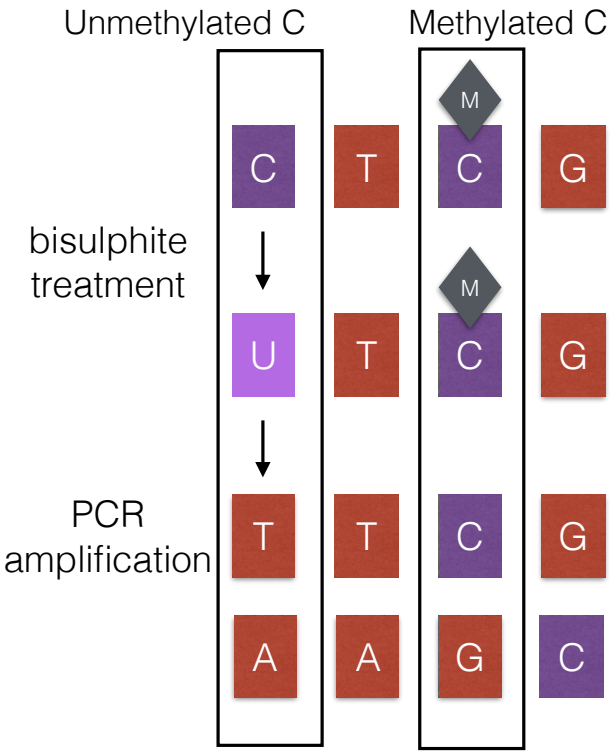


Figure 2.1: Bisulphite treatment of methylated and unmethylated Cytosines. Bisulphite treatment leads to the conversion of unmethylated Cytosines (C, purple) into Uracils (U, light purple) which upon PCR amplification are converted into Thymine (T, red). In contrast, 5' methyl-Cytosines (C purple with grey diamond) are unaffected by the bisulphite treatment and remain unchanged.

2.3 Assay for Transposase-Accessible Chromatin with high throughput sequencing (ATAC-seq)

As described in more detail in Chapter 4, ATAC-seq dramatically reduces the cell number requirements for studying open chromatin. This is based on the use of a transposase that binds preferentially to open chromatin and the basic principle is shown in Fig2.2.

ATAC-seq was performed on 50,000 cells as described in Buenrostro et al 2013 and 2015[123,154]. After picking cells by hand, they were transferred into 500µl CMRL, washed and lysed. The DNA from the lysed cells was then transposed using a Nextera Tn5 Transposase (Illumina) and purified using a PCR Minelute kit (Qiagen). Subsequently, the transposed DNA was PCR-amplified using customised Nextera primers provided by Buenrostro et al 2013 (primer 1 AD-noMX and one of AD2.1-2.6 as primer 2 to multiplex samples) for 11 cycles on the PTC-225 Peltier Thermocycler (MJ Research) machine and purified into 20µl using a PCR Minelute kit. Then another round of purification with Agencourt AMPure XP magnetic beads (Beckman Coulter) was performed according to the manufacturer's instructions to reduce primer dimers. The resulting libraries were then quantified using a library quantification kit for next-generation sequences (Kapa Biosystems) on an Applied Biosystems 7900HT system in accordance with the instructions provided by the manufacturer. The quantified libraries were then submitted for sequencing as described in Chapter 4.

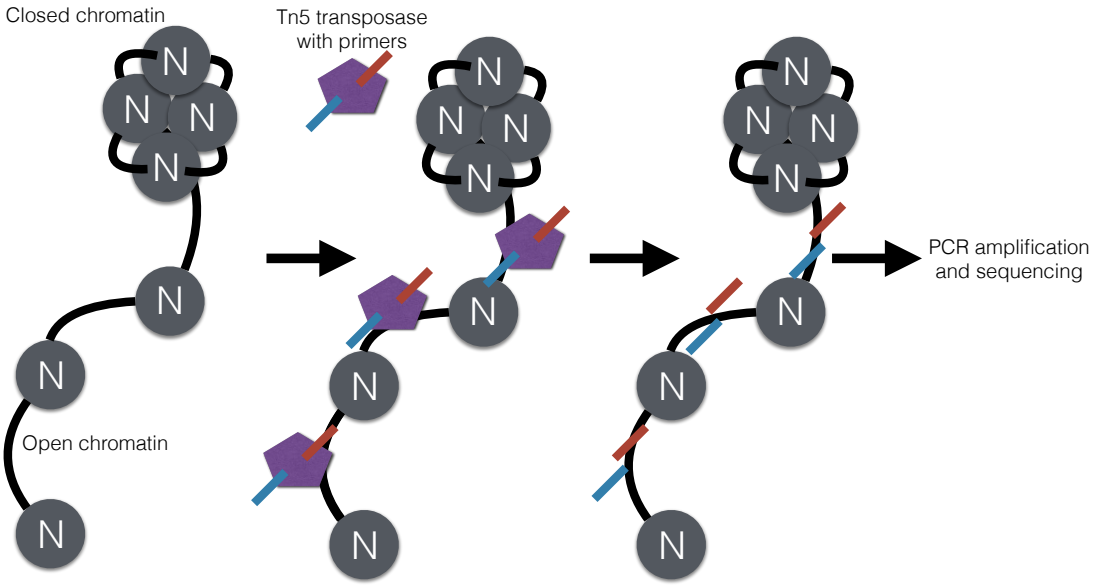


Figure 2.2: ATAC-seq interrogates open chromatin. The transposase (purple) is charged with primers (red and blue) and has a preference for open chromatin (regions with lack of nucleosomes (N) which are shown in grey) where the primers are inserted into the DNA (black). The primers are then used for PCR amplification which leads to the generation of a large number of fragments that can be sequenced[123].

2.4 CNV *in vitro* genotyping validation using qPCR

As part of the exome chip deletion analysis in Chapter 6, Quantitative real-time PCR (qPCR) using TaqMan copy number assays (according to manufacturer's instructions) was performed on up to 96 selected Oxford-based UK T2D case-control DNA samples, to determine the copy number status at selected gene deletion regions.

The experiment was conducted in either duplicates or triplicates per deletion region using a reaction volume of 10µl per replicate including:

- 5µl of TaqMan Genotyping Mastermix
- 0.5 µl of TaqMan Copy number assay (depending on gene deletion region, see Table 2.1 for probes labelled with FAM fluorescence dye)
- 0.5 µl of TaqMan Copy number Reference RNaseP assay (labelled with VIC fluorescence dye)
- 2µl of DNA (total amount 10ng)
- 2µl of nuclease-free H2O

The qPCR plate was thermocycled and fluorescence analysed on an ABI 7900 HT machine using the "Absolute Quantification" method with the following cycle conditions

- 10 minutes at 95°C
- 40 cycles of 15 seconds at 95°C and 60 seconds at 60°C.

The resulting data was exported and analysed by applying the ΔΔCt (cycle threshold) analysis method using customised R scripts. After determining the mean Ct of the multiplexed target FAM ($\bar{C}t_{iFAM}$) and reference RNaseP VIC ($\bar{C}t_{iVIC}$) assay for each sample (i) the ΔCt_i sample values were calculated. Subsequently, $\Delta\Delta Ct_i$ values per sample were calculated by subtracting the median ΔCt value of predicted diploid samples from individual sample ΔCt_i. $\Delta\Delta Ct_i$ was then used to determine the relative copy number (CN) state. Specifically, the following formulas were used:

$$\Delta Ct_i = \bar{C}t_{iFAM} - \bar{C}t_{iVIC}$$

$$\Delta\Delta Ct_i = \Delta Cti - median(\Delta Ct_{diploidsamples})$$

$$CN_i = 2^{(-\Delta\Delta Ct_i)}$$

Gene	Assay	Deletion Samples	Diploid CN sample
ANKRD35	Hs00491482-cn	52	44
DOCK8	Hs00572743-cn	21	75
KIF7	Hs01635981-cn	29	7
PTCHD3	Hs00913691-cn	22	74
TRIB3	Hs07209681-cn	37	69

Table 2.1: CNV *in vitro* genotyping TaqMan probes and samples. The gene, the TaqMan probe ID (Assay) and the number of deletion and diploid copy number (CN) samples used for the CNV genotyping are shown.

2.5 Genetics studies used in this thesis

2.5.1 GWAS datasets

DIAGRAM, ENGAGE and/or GIANT GWAS SNP summary level data (more information about the consortia below) were used for the FGWAS enrichment analysis of hypomethylated regulatory regions (LMRs and UMRs) in Chapter 3, ATAC-seq open chromatin peaks in Chapter 4 and refined chromatin states in Chapter 5. In Chapter 6, the summary level data was also used to determine whether genic CNVs associated with T2D or BMI were near known suggestive ($P < 5x10^{-6}$) DIAGRAM T2D or ENGAGE BMI GWAS regions, respectively.

In addition, DIAGRAM and ENGAGE GWAS SNP summary level data were used in Chapter 3 to compare the ability of 450k array and WGBS CpG sites to capture GWAS signal.

2.5.1.1 T2D DIAGRAM consortium

The DIAGRAM (DIAbetes Genetics Replication And Meta-analysis) consortium represents a group of investigators that are interested in performing large-scale studies to characterise the genetic basis of T2D primarily in individuals with European descent. Further details about the DIAGRAM consortium are provided at the website `http://diagram-consortium.org`. The GWAS SNP summary level data were generated and provided by members of the DIAGRAM consortium (Dr Robert Scott) and derived from 26,676 T2D cases and 132,532 controls of European descent that were analysed on different types of GWAS SNP arrays and imputed up to the 1000 Genomes March 2012 reference panel.

Also, DIAGRAM GWAS 99% credible set Posterior Probabilities (PP) were provided. These sets contained the variants that combined capture 99% of the PP. The PP was an estimate for a variant driving the GWAS signals and was derived from fine-mapping data.

2.5.1.2 ENGAGE consortium

The ENGAGE (European Network for Genetic and Genomic Epidemiology) consortium seeks to translate large-scale genetic and genomic epidemiology population datasets into information relevant for clinical use to ultimately understand risk factors, disease progression and differences in treatment response (more information on `http://www.euengage.org/`).

Publicly available BMI, Fasting Glucose (FG) and Total Cholesterol (TC) GWAS SNP summary level results were obtained from the ENGAGE website (`http://diagram-consortium.org/2015_ENGAGE_1KG/`). These were generated from 1000 Genomes imputed (June 2011 reference panel) GWAS meta-analysis data which were derived from a range of SNP genotyping arrays of a varying number of individuals (additional information can be found on the website and in the relevant publications):

- 87.0k individuals for BMI[155]
- 46.7k individuals for FG[155]
- 62.2k individuals for TC[156]

In addition, BMI and FG 99% credible set data were also received from ENGAGE consortium members (Dr Momoko Horikoshi).

2.5.1.3 Giant Height

The Genetic Investigation of ANthropometric Traits (GIANT) consortium aims to determine genetic loci that affect anthropometric traits such as height and obesity (see website for more information `http://portals.broadinstitute.org/collaboration/giant/index.php/Main_Page`).

Publicly available Height GWAS SNP summary level data from 253k individuals[138], which were genotyped on a number of different SNP arrays and imputed into the Phase II CEU HapMap data, was obtained from the GIANT website.

2.5.2 Exome-sequencing and exome chip datasets for CNV calling

2.5.2.1 ExAC CNV calls

The Exome Aggregation Consortium (ExAC) is a coalition of scientists with the aim to aggregate and harmonise large-scale exome sequencing data from a number of different projects and to provide summary level data for the wider scientific community. As part of the ExAC efforts, CNVs were jointly called from exome sequencing data of approximately 60k individuals by members of the ExAC consortium (Dr Douglas Ruderfer) as described in Ruderfer et al 2016[157] and Chapter 6 Section 6.2.1.2 in more detail. Exonic CNVs calls for a subset of 14.5k T2D-case-control samples were provided for the analysis presented in Chapter 6. The exome sequencing data of these 14.5k samples was generated as part of the T2D-Genes (10.2k samples), GoT2D (2.7k samples) and SIGMA (3.8k samples) consortia. The studies used in these consortia (which include additional samples not used in the CNV analysis) are summarised in Table 2.2.

2.5.3 Exome chip CNV samples

For the exome chip CNV analysis, a subset of 11,686 UK exome chip samples were obtained from Oxford-based UK T2D case-control samples (n=8178, comprising of UKT2D/GoT2D, Young Diabetics Oxford (YDX), Oxford Biobank (OBB) and TwinsUK studies) as well as the Genetics of Diabetes and Audit Research Tayside Study (GoDARTS, n=3508). For 11,414 unrelated samples, full phenotypic data for T2D-association testing was available and Table2.3 provides an overview of the samples used for association testing. These are part of a larger 82k sample exome chip analysis and more information about the samples can be found here[64] and at `http://www.type2diabetesgenetics.org`.

Consortium	N	Cases	Ctrl	Study	Ethnicity
	1026	500	526	Jackson Heart Study Candidate Gene Association Resource	African American
	1048	518	530	Wake Forest Study	African American
	1087	526	561	KARE and KNIH	East Asian
	1078	486	592	Singapore Diabetes Cohort Study and Singapore Prospective Study Program	East Asian
	861	506	355	Longevity Genes in Founder Populations (Ashkenazi)	European
T2D-GENES	982	484	498	Metabolic Syndrome in Men Study (METSIM)	European
	490	272	218	San Antonio Mexican American Family Studies, Texas	Latino
	1453	749	704	Starr County, Texas	Latino
	1069	531	538	London Life Sciences Population (LOLIPOP)	South Asian
	1148	563	585	Singapore Indian Eye Study	South Asian
Total T2D-Genes	10242	5135	5107	T2D-Genes	Mixed Ancestry
	958	472	486	Finland-United States Investigation of NIDDM Genetics (FUSION)	European
GoT2D	187	97	90	KORAgen Study Helmholtz zentrum Mu_nchen (KORA)	European
	642	322	320	UK Type 2 Diabetes Genetics Consortium (UKT2D)	European
	921	478	443	PPP-Malmo-Botnia Study	European
Total GoT2D	2708	1369	1339	GoT2D	European only
	1098	551	547	UNAM/INCMNSZ Diabetes Study (UIDS)	Hispanic
SIGMA	968	509	459	Diabetes in Mexico Study (DMS)	Hispanic
	796	270	526	Mexico City Diabetes Study (MCDS)	Hispanic
	930	487	443	Multiethnic Cohort (MEC)	Hispanic
Total SIGMA	3792	1817	1975	SIGMA	Hispanic only
Total All	16742	8321	8421	All	Mixed Ancestry

Table 2.2: Sample details about the consortia involved in ExAC. For each study, the number of samples (N), divided into the number case and control (Ctrl) individuals, is shown. Additionally, information about the contributing studies and sample ethnicity is provided. Abbreviations: KARE: Korea Association Research Project, KNIH: Korean National Institute for Health

Cohort	N	Cases	Ctrl	Ethnicity
Oxford-based UK T2D case-control	7906	2195	5711	European
The Diabetes Audit and Research in Tayside Scotland (GoDarts)	3508	1715	1793	European
Total	11414	3910	7504	European

Table 2.3: Sample details for the exome chip analysis. For each study, the number of samples (N), divided into the number case and control (Ctrl) individuals, is shown.

2.6 Computational analysis

2.6.1 FGWAS enrichment analysis

FGWAS (version 0.3.6)[158] was used with standard settings to determine enrichment of islet LMRs and UMRs (Chapter 3), ATAC-seq islet open chromatin (Chapter 4) and islet chromatin states (Chapter 5) in DIAGRAM (setting "cc" was applied for use with T2D-case-control GWAS data), ENGAGE and GIANT GWAS signal. Specifically, FGWAS applied a hierarchical model that determined shared properties of loci affecting a trait. The FGWAS model used SNP-based GWAS summary level data and divided the genome into windows (setting "k"=5000 which represents the number of SNPs per window), which are larger than the expected LD patterns in the population. The model assumed that each window either contained a single SNP that affected the trait or that there was no SNP in the window that influenced the trait. The model estimated the prior probability of a window to contain an association and the conditional prior probability that a SNP within the window was the causal variant. These prior probabilities were variable, dependent on regional annotations and estimated based on enrichment patterns of annotations across the genome using a Bayes approach[158]. For each annotation, the model provided maximum likelihood enrichment parameters and annotations were considered as significantly enriched if the parameter estimate and 95% confidence interval (CI) was above 0.

As described in Chapter 5 Section 5.2.3 the best (maximum likelihood) FGWAS model for multiple annotations was determined by combining significantly enriched annotations followed by a cross-validation approach. Based on the empirical estimates of the regional and SNP prior probabilities of the best-fitted model a Bayes factor summarising the evidence of association was determined. This information was then used to derive FGWAS Posterior Probabilities (PP) for both the window and each SNP within a window being causal. The window and SNP PP were used to prioritise likely causal variants at significant GWAS loci for testing of allelic imbalance in open chromatin in Chapter 5.

2.6.2 Enrichment analysis of islet annotations

Enrichment of hypomethylated regulatory regions (Chapter 3) and ATAC-seq open chromatin peaks (Chapter 4) in the following publicly available genomic annotation datasets (for Chapter 4 all but the islet mQTLs were used) was calculated through random permutation as described below.

- islet ChIP-seq TFBS predictions (*FOXA2, MAFB, NKX2.2, NKX6.1* and *PDX1*) provided by Dr Kyle Gaulton[85]
- significant islet eQTLs provided by Dr Martijn van de Bunt[159]
- genome-wide human CTCF sites from the CTCFBSDB 2.0 database[160]
- significant islet mQTLs and associated CpG-sites[99]

To calculate enrichment the true overlap between regulatory regions identified in this thesis (LMRs, UMRs and ATAC-seq open chromatin peaks) and the above-mentioned annotations was determined. Subsequently, the regulatory regions (LMRs, UMRs and ATAC-seq open chromatin peaks) were randomly shifted throughout the genome using bedtools shuffleBed function (1000 times) and bedtools intersectBed function was used to determine the overlap between both permuted and unmodified regulatory regions. By comparing the distribution of the permuted overlap to the true overlap, a P-value (minimum P=0.001) could be derived based on how often the number of overlapping random sites exceeded the number of overlapping true sites. The resulting P-values were multiple testing corrected using Bonferroni correction. The fold enrichment was calculated based on comparing the number of overlapping true sites to the mean number of permuted sites that overlapped.

2.6.3 Overlap analysis of genomic annotations and genetic variants

For all chapters, the overlap of regulatory regions (CpG sites, LMRs and UMRs in Chapter 3, ATAC-seq open chromatin Chapter 4 and islet chromatin states Chapter 5) and genetic variants (CNVs in Chapter 6) identified in this thesis with previously generated genomic annotations (see

Chapter 2 Section 2.6.2, Chapter 3 Section 3.2.4 and Chapter 4 Section 4.2.7) and/or GWAS summary level data (see Chapter 2 Section 2.5.1) was determined using the intersectBed function as implemented in bedtools (version v2.21.0)[161].

2.6.4 Visualisation of genomic annotations and genetic variants

For the visualisation of the location of regulatory elements (ATAC-seq open chromatin in Chapter 4 and islet chromatin states, LMRs and open chromatin in Chapter 5) and CNVs (Chapter 6) in the genome, the UCSC genome browser[162] was used.

2.6.5 Statistical analysis and figure generation

Statistical analysis including Wilcox-test, Spearman's correlation, binomial test, Principal Component Analysis (PCA), hierarchical clustering and multiple testing corrections (FDR and Bonferroni) were all performed using R (version 3.0.2)[163]. To facilitate the analysis of large-scale data the R package data.table (version 1.9.2)[164] was used. Statistical significance was defined as $P < 0.05$ unless stated otherwise.

The majority of figures in this thesis were generated using the R package ggplot2 (version 1.0.0)[165].

CHAPTER 3

DNA Methylation in human islets

3.1 Introduction to DNA Methylation

3.1.1 The regulatory role of DNA methylation

As described in Chapter 1, DNA methylation is an important epigenetic modification that is involved in the regulation of the genome. Throughout life, most of the DNA methylome is stable apart from a genome-wide epigenetic reprogramming phase during early development when pluripotency is established. After global demethylation through both active (enzyme mediated mechanisms such as oxidation of DNA methylation by TET enzymes) and passive (lack of DNA maintenance) demethylation mechanisms, *de novo* methyltransferase including DNMT3A, DNMT3B and DNMT3L re-establish high methylation levels across the vast majority of CpG sites (70-80%). CpG-dense regions, defined as CpG-islands (CGI), are a notable exception and remain largely unmethylated throughout the whole life by a set of proteins containing a CXXC DNA binding domain

that prevent methyltransferases from adding methyl-groups to these sites[95,166].

CpG-islands (CGIs) are CpG-rich short DNA regions (approximately 1kb in size) that are predominantly found in promoter and 5' regions of genes. At these CGI promoter regions, DNA methylation is inversely correlated with gene expression and associated with silencing of transcriptional activity as part of X-chromosome inactivation[167] and imprinting[168]. The regulatory effects of DNA methylation on gene expression have been linked to both active and passive mechanisms: on the one hand, the presence of DNA methylation was shown to actively influence the binding of proteins that are sensitive to methylation levels including transcription factors[169] or Methyl-CpG Binding Domain (MBD) proteins that mediate a repressive function[170,171]. On the other hand, transcription factor binding or activity of the transcriptional machinery may influence local DNA methylation levels suggesting a passive involvement of DNA methylation in the regulation of gene expression. Such passive mechanisms have been proposed to act at distant enhancer regulatory elements which are defined as low-methylated regions (LMR). These LMRs are created through transcription factor binding which modifies local DNA methylation levels[172]. In contrast to promoter methylation, gene body methylation was positively associated with transcriptional activity and may mark alternative intragenic promoters and affect splicing (Fig3.1A). Similarly, the regions directly next to CGIs defined as CpG-shores (up to 2kb away from CGIs) and shelves (located 2-4kb away from CGIs) have also been suggested to be involved in tissue-specific gene expression and genome regulation[173], however, their exact role is not well understood yet.[95,174].

Recent advances in array technology including the Illumina 450k HumanBead chip (450k array)[175] and Infinium MethylationEPIC chip (850k array)[176] as well as next-generation sequencing such as Whole-genome bisulphite sequencing (WGBS)[70,71] have allowed to study methylation levels of single CpG sites compared to earlier methods that only provided regional level information on methylation status. While the 450k and 850k arrays interrogate with (likely) higher accuracy only about 485k and 850k CpG sites, respectively, the vast majority of which are found in CpG island (CGI) and promoter regions, WGBS can, depending on read depth, provide comprehensive methylation information on all CpG-sites in the genome (about 27 million in hg19).

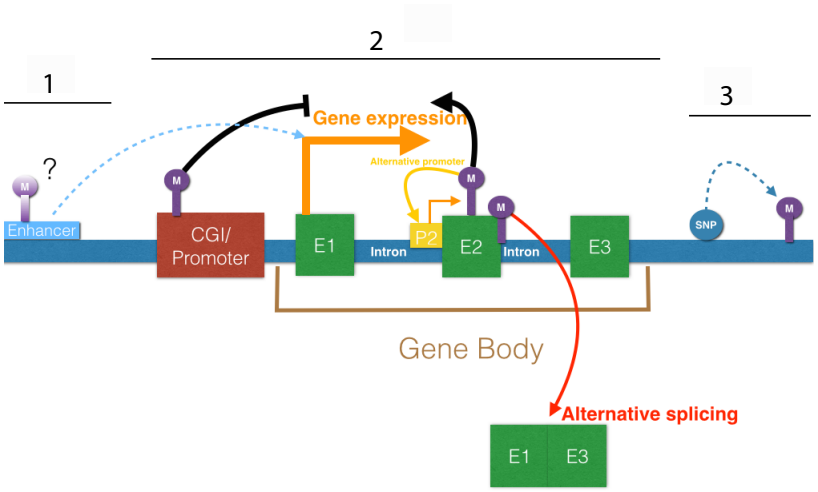
The improvements in technology have allowed for the identification of methylation Quantitative Trait Loci (mQTLs), which represent associations between genotype and methylation levels. Due to

observed associations between transcriptional activity and both DNA methylation and genetic variants with the latter defined as expression QTLs, the following model assuming causal directionality between these factors was proposed: genetic variants considered as (almost) unmodifiable over time act as causal triggers inducing changes in DNA methylation that mediate changes in transcription. However, the causal directionality between an epigenetic factor and a correlated phenotype can not be derived without additional information since epigenetic information is not stable over time. Thus, changes in DNA methylation could be the consequence of differences in cellular activity influenced by either internal changes such as transcription factor binding or even external environmental factors (Fig3.1B). Recent studies on the population level across three distinct tissues have provided substantial evidence for more complex relationships including causal, consequential and independent interactions between DNA methylation and gene expression [101,105].

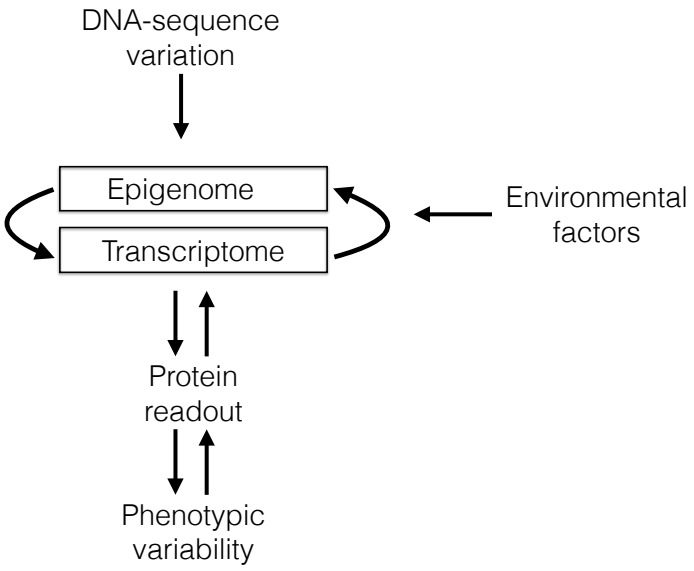
3.1.2 Role of DNA methylation in T2D diabetes and islet function

To date, a large number of studies have linked changes in DNA methylation to both monogenic and complex forms of diabetes. For instance, imprinting defects associated with abnormal pancreatic or diabetic phenotypes such as Beckwith-Wiedemann Syndrome near the *IGF2* and *KCNQ1* region on chromosome 11 [178] or Paternal Uniparental disomy 6 at 6q24 [179] represent single gene disorders at which changes in methylation and imprinting are associated with severe phenotypes including early-onset syndromic forms of diabetes [106].

In addition to being associated with monogenic diabetes forms, DNA methylation was also proposed to influence T2D pathogenesis. Based on epidemiological observations it was found that poor early growth is associated with increased T2D risk later in life [180–182]. For instance, low birth weight due to poor nutrition levels as observed in the Dutch Hunger Winter cohort [183] as well as high birth weight caused by over-nutrition or hyperglycaemia during pregnancy is associated with increased risk of T2D in later life [184]. DNA methylation was proposed to be influenced by an adverse uterine environment and thus, the mechanism that transmits these metabolic memory effects, which lead to a higher risk of T2D development in adult life [185]. Follow-up studies involving the Dutch Hunger cohort have shown that six decades later the group with an adverse uterine environment due to poor nutrition levels displayed differential blood methylation levels at the *IGF2*



(A) DNA methylation and gene expression



(B) DNA methylation and gene expression

Figure 3.1: The complex role of DNA methylation in regulating gene expression. (A1) Dynamic changes in DNA methylation have been suggested to influence chromatin regulatory states, such as enhancers (light blue), that may regulate gene expression. (A2). DNA methylation in CpG islands (CGI) in or around promoter regions (dark red) is mostly associated with silencing of gene expression. In contrast, gene body methylation (green) is positively associated with gene expression and may also be responsible for alternative splicing (red). Additionally, gene body methylation may also mark alternative intragenic promoters (yellow). (A3) Genetic variants (blue), defined as mQTLs, have been shown to correlate with the DNA methylation state of nearby CpG sites [174]. (B) In contrast to interactions between SNP and phenotype, the directionality of interactions between the epigenome, the transcriptome and non-molecular phenotypes can not be directly established without additional evidence [177].

imprinted gene as well as several other genes linked to metabolic and cardiovascular function compared to the group exposed to a normal uterine environment[186,187]. These results provided additional evidence that environmental conditions experienced early in life may have an impact on adult life and influence metabolic disease risk.

Further evidence for the role of DNA methylation in T2D pathogenesis was provided by both T2D case-control studies that identified T2D-associated differentially methylated regions. Since blood is readily available, several studies linking DNA methylation to T2D risk have been performed in blood cells[188–192]. For instance, Chambers et al 2015 performed one of the largest T2D Epigenome-wide association studies to date including 2500 Indian Asian T2D-case-controls individuals and a replication cohort of approximately 1500 Europeans. Using the Illumina 450k array, the study found CpG methylation levels near five genes (*ABCG1, PHOSPHO1, SOCS3, SREBF1* and *TXNP1*) associated with future T2D risk. While these genes are plausible candidates involved in pathways that may explain the risk of T2D or related metabolic defects, it remains unclear how the observed differences in methylation are related to T2D onset[191]. Similarly, Yuan et al 2014 identified T2D-associated DMRs in whole blood in monozygotic twins discordant for T2D. The most significant signal was in the promoter of *MALT1* which is involved in glycaemia and insulin pathways. The identified T2D DMRs also showed enrichment in T2D GWAS regions thus, highlighting the importance of integrating genetic and epigenetic mechanisms[190].

While these studies support the role of DNA methylation in T2D development, the direction of causality remains questionable, and these methylation changes may reflect only biomarkers that are a consequence of T2D pathogenesis. In particular, since epigenetic effects are often tissue-specific and hence, it is not clear how these T2D-associated methylation changes in blood could be driving T2D disease risk. Since T2D is characterised by insulin resistance and beta-cell failure, studies investigating DNA methylation in disease-relevant tissue involved in the response to insulin (e.g. skeletal muscle and/or adipose tissue)[100,193–195] as well as insulin secretion (pancreatic islets[99,196–198]) are critical. The majority of T2D GWAS susceptibility loci is associated with reduced beta-cell function; thus, this part of the introduction will focus on the role of DNA methylation in human islet function and how this is related to T2D loci.

Genome-wide approaches, which studied human islet DNA methylation using the 450k array, found

methylation changes at 1649 CpG sites and 853 genes associated with T2D status some of which overlapped with genes near T2D and obesity GWAS loci. For instance, *TCF7L2*, *KCNQ1* and *FTO* showed T2D-associated methylation changes in human islets. While both *TCF7L2* and *KCNQ1* loci were previously associated with islet regulation or function, the role of islet DNA methylation at the *FTO* locus in genetic susceptibility of obesity remains unclear. At 102 genes these disease-associated methylation changes correlated with differential gene expression *in vitro*. Follow-up of selected candidate genes in murine alpha- and beta-cells showed that differentially methylated and expressed genes affect both glucagon and insulin secretion [198]. Using the same human islet dataset, a later study found correlations between ageing and the islet methylation levels of 241 CpG sites of which some were associated with islet gene expression and beta-cell function in *in vitro* assays. For a subset of 139 CpG sites, age-associated methylation level changes were previously found in white blood cells suggesting that blood, which is easier accessible compared to islets, may act as a surrogate biomarker for these age-associated changes [199]. However, changes in cell-type composition may affect blood-based biomarker methylation levels and bias the results of such a test. In addition to islet methylation changes associated with T2D development, a recent study also characterised the interactions between genetic variation, DNA methylation and transcription in human islets. Olsson et al 2014 identified 36.8k human islet mQTLs at which the genotype is associated with the methylation status of 11.7k CpG sites. A number of mQTLs have been previously implicated in T1D and T2D GWAS loci including *ADCY5*, *KCNJ11*, *HLADQA1*, *INS*, *PDX1* and *GRB10*. In combination with gene expression and insulin secretion data, a subset of mQTLs was identified that was associated with gene expression or insulin secretion; however, statistical causality using a causal inference test could only be established at very few sites. Knockdown of selected genes such as *Gpx7*, *Gstt1* and *Snx19* in murine INS1 beta-cells, at which expression changes were associated with methylation status, had an effect on proliferation and apoptosis, thus, potentially highlighting how changes in DNA methylation may affect gene expression and beta-cell function [99].

While these studies provided useful insight into the role of DNA methylation in human islet biology, there were also several limitations. Firstly, the studies were restricted to the 485,000 CpG sites on the 450k array that are primarily located in promoter and CpG islet regions. These CpG-dense regions were found to be depleted of T2D-associated islet methylation changes and mQTLs and show limited overlap with T2D GWAS signals enriched for distant non-coding regions. Sec-

ondly, mQTL overlap with GWAS signals was reported mainly at the locus level without precisely defining the location of the risk variant or the molecular mechanisms governing disease risk. In particular, detailed information related to the haplotype structure at the GWAS locus as well as information regarding the directionality of the methylation effect on disease risk and gene expression were missing. Thirdly, non-coding regulatory sites are often marked by multiple variable and highly inter-dependent epigenomic factors (including chromatin, histone modifications, DNA methylation and non-coding RNA). Thus, integration of different types of genomic annotations is critical to fully understand molecular regulatory interactions. Fourthly, although the previous studies identified associations between genetic, epigenetic and expression variation, inference of causal directionality of epigenetic factors on phenotypic variation remains extremely challenging. In particular, since a recent study showed that both active and passive relationships between DNA methylation and gene expression exist[101].

The analysis in this chapter aims to characterise the human islet methylome by generating WGBS and 450k array methylation data and to determine the role of methylation in islet function and T2D development. The main objectives of the chapter are:

- to compare the WGBS and 450k array methylation data to determine the ability of the two methods to interrogate the islet methylome, in particular near T2D GWAS regions
- to identify hypomethylated islet regulatory regions using the WGBS islet methylation data
- to determine the involvement of these hypomethylated regulatory elements in the genetic susceptibility of T2D

3.2 Methods

3.2.1 Human pancreatic islets

Human islets were retrieved from diseased donors from the Oxford DRWF Human Islet Isolation Facility (n=34) and through collaborative work with Dr. Patrick McDonald at the Alberta Diabetes

Institute in Edmonton in Canada (n=10). The purity of islet endocrine content was determined by dithizone labelling, which is part of the quality control process performed by the islet isolation facility. The obtained purity estimates were used to select only high purity islet samples (purity >70%) for the WGBS and DNA methylation array.

3.2.2 Generation of WGBS data

Ten high purity (purity >70%) endocrine islet samples were used for whole-genome bisulphite sequencing. DNA was extracted, the quality was assessed and the DNA was bisulphite converted as described in Chapter 2 Section 2.1. Following the bisulphite conversion, libraries were sequenced at the High-Throughput Genomics group which is part of the Wellcome Trust Centre for Human Genetics, University of Oxford, Oxford, UK. Samples were sequenced as multiplex libraries across 3 HiSeq2000 lanes with 100bp paired-end read length to obtain high-coverage read data. To deal with the low complexity libraries were sequenced together with a control lane and a PhIX spike-in of 5%.

3.2.2.1 Processing of raw reads

Raw reads were obtained in FASTQ format and basic quality control regarding base quality scores and duplication rate was performed using FASTQC[200]. To enhance mapping efficiency paired reads were cut 10bp at the start and the second read in the pair was also cut 15bp at the end using a customised python script (version 3.4.1 `http://www.python.org`).

3.2.2.2 Read mapping and estimating bisulphite conversion efficiency using Bismark

The trimmed reads were aligned to hg19 using the software Bismark version 0.12.5[201]. In short, Bismark determines the best alignment for bisulphite converted reads by performing C-T and G-A conversion on bisulphite converted reads which are then aligned by bowtie2[202] (version 2.2.0) to an equally converted reference genome. Initially single-end alignment of the reads was performed

with a reduced scoring function (L,0,-0.6) but otherwise standard settings to estimate several parameters. Firstly, mapping efficiency was calculated by determining the fraction of reads that could be mapped to the genome for each pair separately. Secondly, insert and fragment size distribution were estimated from the generated bam files by using samtools[203] (version 0.1.19) and customised python scripts. Thirdly, bisulphite conversion efficiency was determined by evaluating the percentage of CpG methylation (expected to be high >70%) and non-CpG methylation (expected to be low <10%).

Subsequently, paired-end alignment of trimmed reads was performed and unmapped reads from read 1 were realigned using Bismark and merged with the paired-end alignment using samtools in order to increase mapping efficiency.

Coverage for the merged paired-end and realigned HiSeq read alignments was estimated for the human mappable genome (NCBI hg19 2.8 billion base pairs excluding gaps and unmappable and blacklisted regions according to UCSC and Encode) using bedtools[161] (version v2.21.0) genomecov function (see Fig3.2)

3.2.2.3 Determining individual CpG methylation levels and variability across samples

CpG methylation levels, defined as beta, which is the ratio of unmodified C (methylated) and bisulphite converted T (unmethylated) alleles, were calculated using BiFAST (mentioned in Lowe et al 2013[204]) by determining the number of reads at which either a C or T allele overlaps the C on either strand. Let C be M and T be U and beta be ß then $\beta = M/(M+U)$. This analysis provided methylation estimates and read counts for each C in a CpG-pair for individual samples. High-pass pooled WGBS data was generated by adding methylated and unmethylated read counts across individual low-pass samples to then estimate the average beta methylation levels. All CpG sites with a coverage of <10X were removed from the analysis.

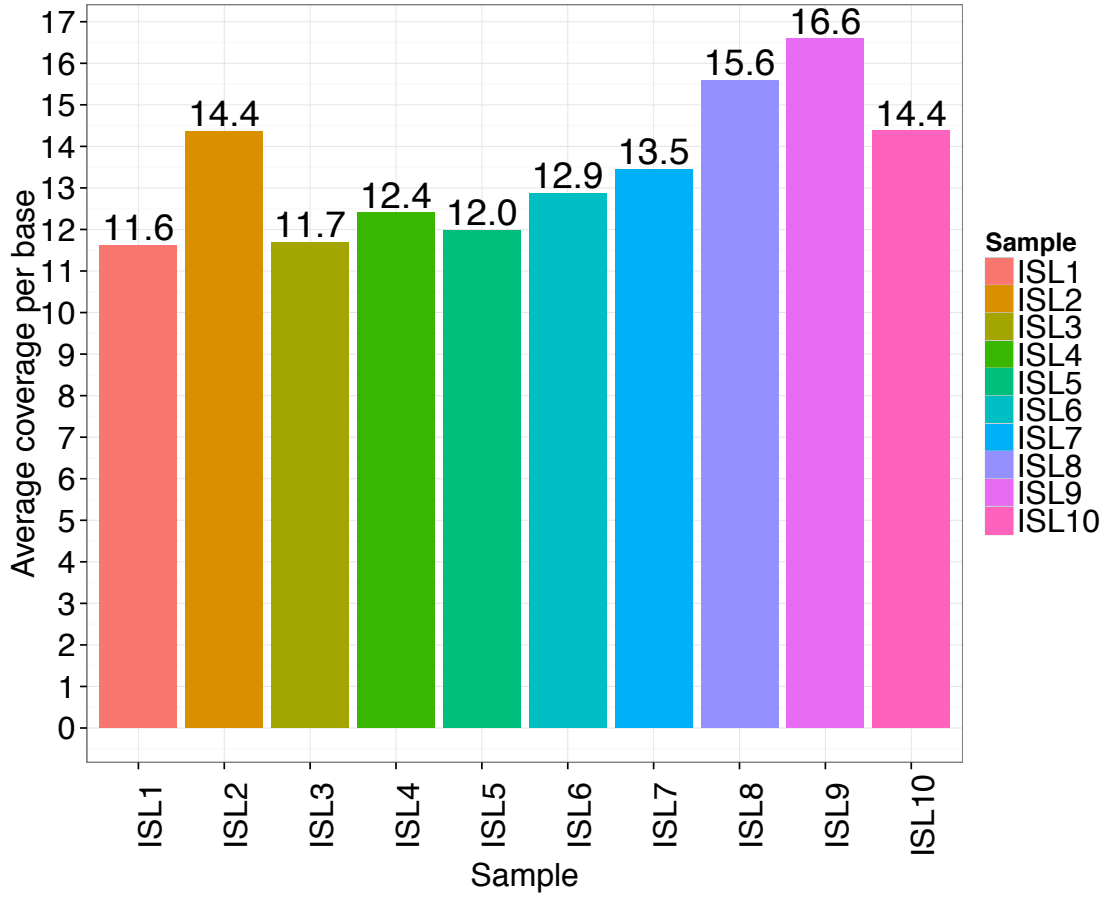


Figure 3.2: Average genome coverage per sample. Barplot of the mean coverage per base (y-axis) across the mappable human genome (2.85x10^9bp) for each sample (x-axis, different colours are used per sample). The actual coverage is also shown on top of each bar.

3.2.2.4 Identifying of hypomethylated regulatory regions using WGBS data and enrichment analysis in GWAS regions

Islet regulatory regions were identified from the pooled WGBS data using the R package methylseek. After removing CpG sites overlapping SNP markers and partially methylated domains (PMDs), which represent highly heterogeneous methylation states determined by DNA sequence features[205], from the methylome, regions with low methylation (<50%) were predicted at an FDR of 5%. The optimal methylation and FDR parameters were inferred from the WGBS data as suggested in the methylseek workflow. The regions with low methylation levels were then separated into CpG-rich UnMethylated Regions (UMRs) and CpG-poor Low-Methylated Regions (LMRs) based on CpG number (threshold 30 CpGs). These sites were previously shown to correspond to promoters (UMRs) and enhancer regulatory regions (LMRs)[206]. Both LMRs and UMRs were tested for global enrichment in BMI, FG, T2D and Total Cholesterol (TC) GWAS signals by applying FGWAS as described in Chapter 2 Section 2.6.1. Additionally, liver, adipose and pancreas WGBS data were obtained from the Epigenome Roadmap Project to identify tissue-specific LMRs and UMRs as described above and repeat the FGWAS enrichment analysis with non-islet regulatory regions to determine tissue-specific enrichment in GWAS signal.

3.2.2.5 Identification of islet Variable Methylated Positions (VMPs)

To determine the most variable sites in the genome, the analysis was focussed on about 10 million CpGs with a minimum coverage of 20X across at least 5 samples (coverage per sample at least 5X). For each of these CpG sites the variance was estimated based on individual sample methylation levels and the top 5% most variable sites were identified.

3.2.3 Illumina 450k array analysis

3.2.3.1 Bisulphite conversion and data generation

41 islet samples were bisulphite converted using the EZ DNA Methylation Kit from Zymogen as described in Chapter 2 Section 2.2.2. Twenty-three of the bisulphite converted samples were then processed using the Illumina 450k assay by the WTCHG High-Throughput Genomics core while the other 18 samples were processed as part of a collaborative effort with Prof Stephan Beck by the sequencing core at University College London, London, UK.

3.2.3.2 Quality Control, Normalisation and extraction of methylation levels

The data was obtained as idat file and analysed using the R package minfi[207] and custom R-scripts (R version 3.0.2)[163]. To ensure high-quality data, CpG sites with a detection P-value >0.01 which is indicative of a high background signal were removed from the analysis and samples with >1% of CpG sites failing this threshold were also removed from the analysis. Also, samples were filtered based on the results of the bisulphite conversion controls on the array.

Following separate quantile normalisation of signal intensities derived from methylated and unmethylated Type I probes and Type II probes, methylation levels (ß) were estimated, based on the intensities of the methylated (M) and unmethylated (U) signal in the following way: $\beta = M/(M + U * 100)$

3.2.4 Annotating WGBS and 450k array methylation data and enrichment analysis across annotations

CpG sites or regions identified from the WGBS and/or 450k array data were annotated with CpG-islands derived from the UCSC genome browser[162], islet chromatin states[86], islet transcription factor binding sites[85], genome-wide human CTCF sites[160], islet mQTLs[99] and eQTLs[159] from various sources.

Enrichment analysis of LMRs and UMRs in these genomic annotations was performed by permutation as described in more detail in Chapter 2 Section 2.6.2. In short, enrichment was calculated for both LMRs and UMRs, separately, based on comparing the true number of overlapping sites with most of the above-mentioned annotations (all but CpG-islands and islet chromatin states) compared to the mean number of overlapping sites derived from 1000 permutations (UMRs and LMRs were randomly shuffled using shuffleBed function implemented in bedtools[161]). P-values were determined based on how often the random number of overlapping sites exceeded the true number of overlapping sites (min P-value 0.001). The resulting P-values were Bonferroni-corrected to adjust for multiple testing.

3.3 DNA methylation results

3.3.1 Characterising the DNA methylation landscape of human pancreatic islets using WGBS and the 450k array

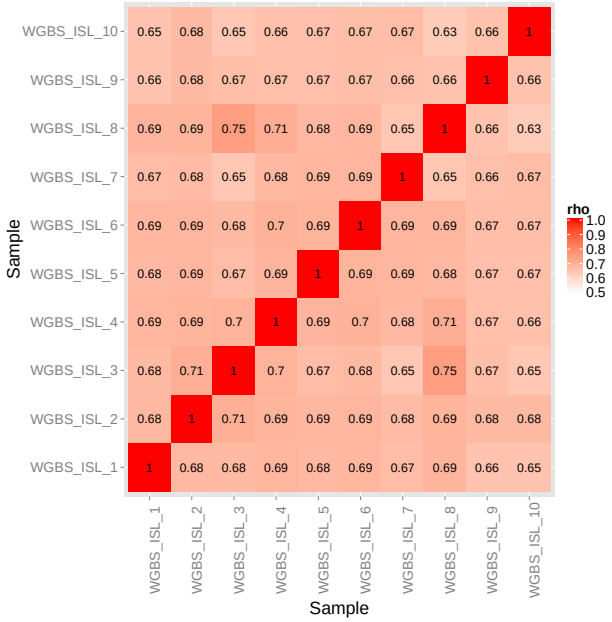
3.3.1.1 The global methylome of human pancreatic islets

WGBS data was generated for 10 human pancreatic islet samples with a mean number of 830 million reads per sample. On average 475 million of the reads per sample were successfully mapped to the genome (hg19) which represents a median mapping efficiency of 58%. The mean coverage of the samples across the mappable human genome was calculated at 13X and per sample between 5.1 million to 10.4 million CpG sites were detected with a minimum coverage of 10X. When comparing the CpG methylation beta levels across samples, high-correlation (mean Spearman's rho=0.71) between the individual samples was observed (Fig3.3A). Such high levels of correlation are expected from DNA methylation data considering that most CpG sites are either fully methylated or unmethylated with little variability across individuals[112]. Since the individual WGBS sample data was low-pass, the data was pooled across samples to obtain a high-pass WGBS islet methylome representing the average islet CpG methylation levels. In total, 23.3 million CpG sites (minimum coverage 10X) representing the human islet methylome were identified.

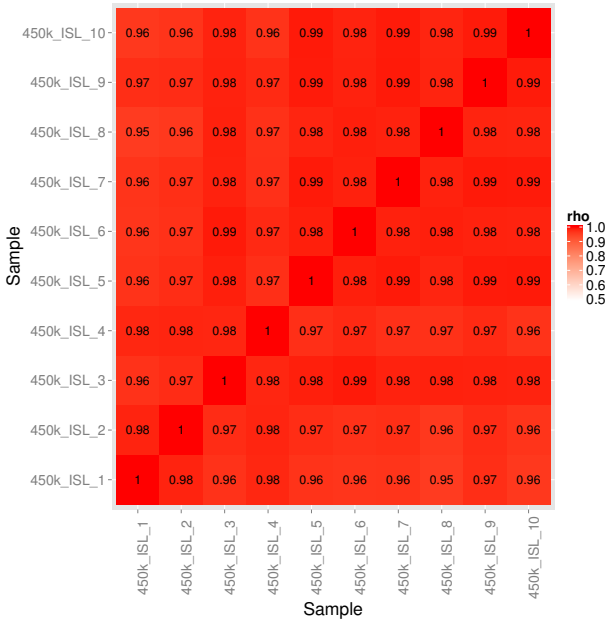
In addition to the WGBS data, methylation data was obtained for 41 human pancreatic islet samples at approximately 485,000 CpG represented on the 450k array. After performing quality control by removing probes with detection P-value >0.01 and samples with the number of failed probes >1%, 450k array methylation data for a mean number of 480,000 CpG-sites and for 32 human islet samples (these include 5 samples that were also processed for WGBS) remained. Across 450k array samples, high correlation in islet CpG methylation levels was observed (mean rho=0.97, Fig3.3B). Similar to the WGBS data, the 450k data was pooled across samples to obtain average islet 450k methylation levels and the high-pass-pooled 450k and WGBS data at overlapping CpG sites were compared.

High correlation between the 450k array and the WGBS methylation data at overlapping CpG sites across all pooled samples was observed (Spearman's rho=0.89 Fig3.4A) despite the fact that the overlapping CpG sites had lower and more variable coverage in the WGBS data compared to the 450k array data (due to the array technology). As expected from previous studies the global DNA methylome of human islets was bimodally distributed, and the 450k data was significantly more hypomethylated (Kolmogorov-Smirnov (KS) test P $<2.2x10^{16}$) compared to the WGBS methylation data (Figure3.4B) due to the focus on lowly-methylated CpG regions on the array.To further delineate these differences between the two methods, CpG methylation status and CpG number of the 450k array and WGBS data across different functional genomic domains based on islet chromatin state (Fig3.5A) and CpG island annotations (Fig3.5B) were evaluated. Significant differences were observed in the CpG methylation distributions across most states with the largest differences being found across weak promoter (KS test D=0.56), transcriptional elongation (D=0.52), transcriptional transition (D=0.39) and weak enhancer (D=0.39) chromatin states (Fig3.6A). Similarly significant and substantial differences were also observed in CpG methylation distributions across CpG shelf (D=0.39), open sea (D=0.35), shore (D=0.30) and even CpG island (D=0.21) regions (Fig3.6B). Although a number of chromatin states only displayed small differences in the overall methylation distribution between the 450k array and the WGBS data, for most chromatin states the CpG sites captured by the 450k array only represented a minor fraction of the total number of WGBS CpG. For instance, 450k array CpG sites only represent 10.4% (range 5.2-17.0%) of CpG sites in strong enhancer states that were previously shown to be enriched for GWAS signals (Fig3.7). Overall these findings show that the 450k array has only very partial and selective coverage outside of pro-

moter and CGI regions, suggesting that the array is not well-suited to investigate the methylation status near GWAS or distant regulatory regions.

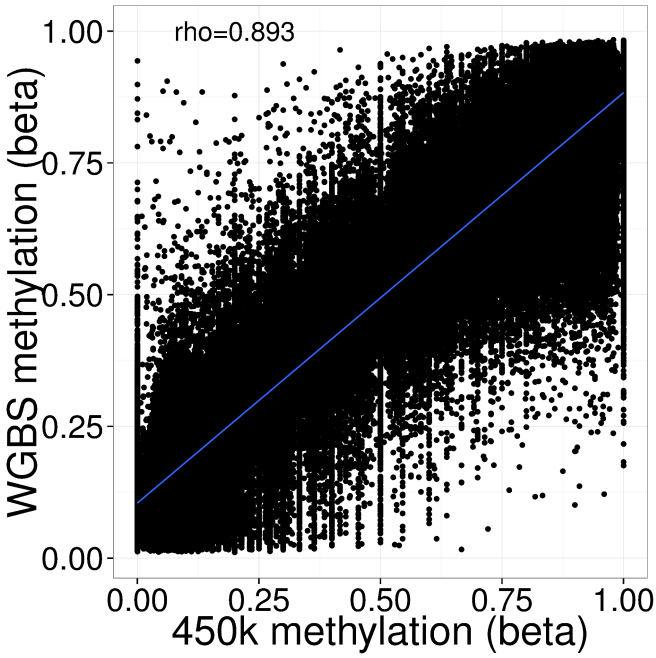


(A) WGBS correlation

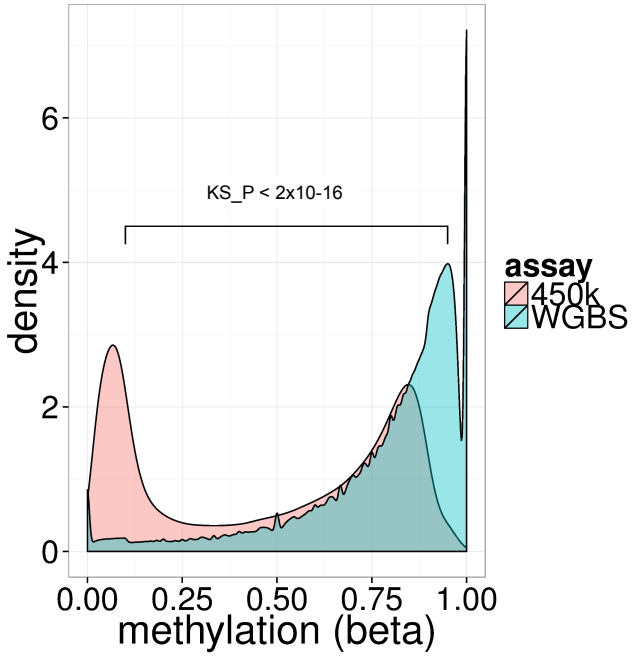


(B) Correlation 450k

Figure 3.3: Correlation heatmap across individual WGBS and 450k samples. Spearman's rho correlation of DNA methylation across individual (A) WGBS and (B) selected 450k samples on the x-axis and y-axis.

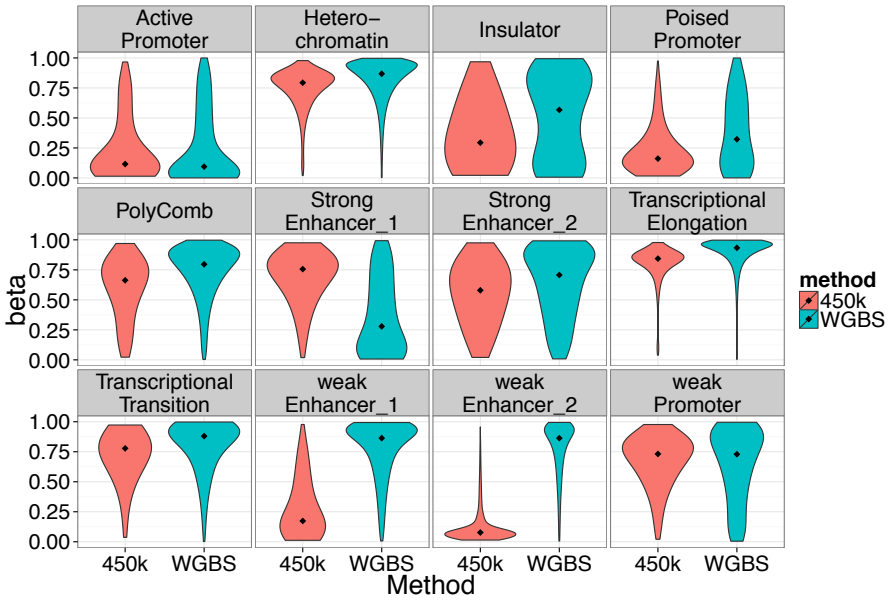


(A) Correlation 450k VS WGBS methylation

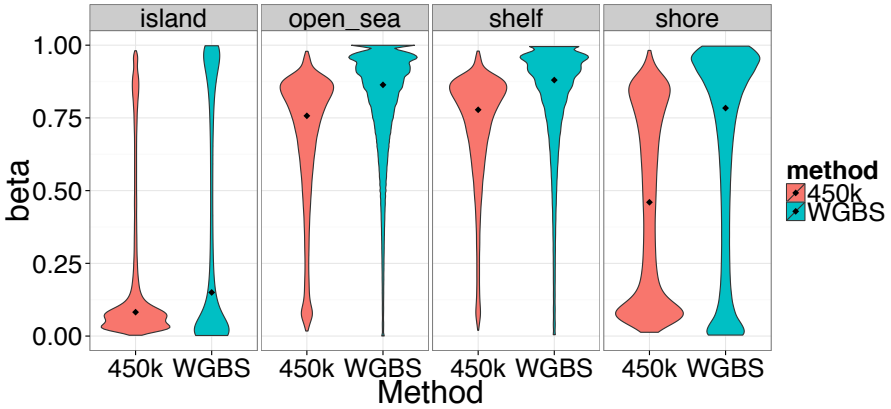


(B) Genome-wide Methylation

Figure 3.4: Distribution of genome-wide methylation levels.(A) Spearman's rho correlation between the 450k array (x-axis) and WGBS (y-axis) at overlapping sites. (B) Comparing the 450k array (red) and WGBS (blue) methylation levels (x-axis) of all CpGs genome-wide assayed by either method (y-axis shows density). The P-value shown is derived using a Kolmogorov-Smirnov (KS) test.

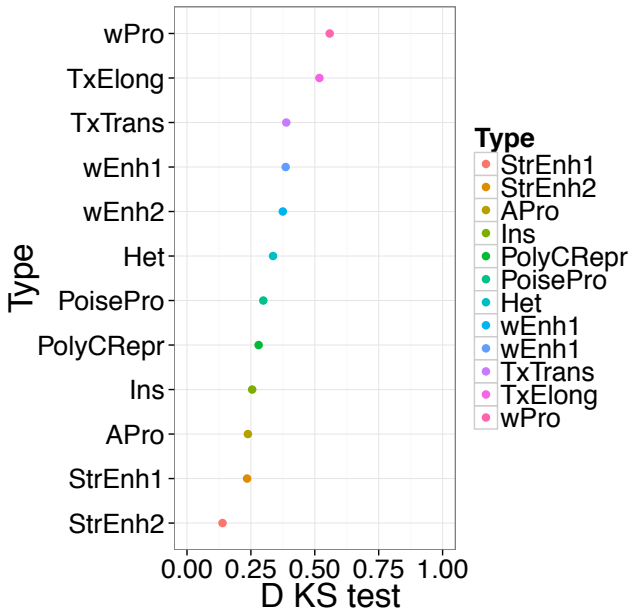


(A) Chromatin state

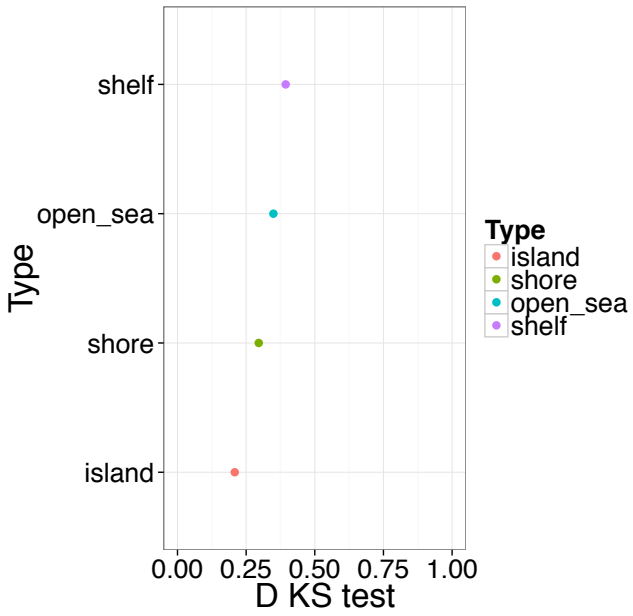


(B) CpG island annotation

Figure 3.5: WGBS and 450k methylation levels in regulatory domains (A)-(B) Distribution of 450k (red, x-axis) and WGBS (blue, x-axis) methylation levels (y-axis) across (A) islet chromatin state (obtained from Parker et al 2014[86]) and (B) CpG-island annotations.



(A) Differences in methylation across islet chromatin states



(B) Differences in methylation across CGI annotations

Figure 3.6: Differences in the methylation distribution in CGI annotation and islet chromatin states. The differences in the 450k and WGBS methylation distribution measured as D statistic from the Kolmogorov-Smirnov test is shown for (A) chromatin states and (B) CpG island annotations. Abbreviations: weak Promoter (wPro), Transcriptional Elongation (TxElong), Transcriptional Transition (TxTrans), weak Enhancer (wEnh), Heterochromatin (Het), Poised Promoter (PoisePro), Polycomb Repressed (PolyCRepr), Insulator (Ins), Active Promoter (APro), Strong Enhancer (Strong Enhancer).

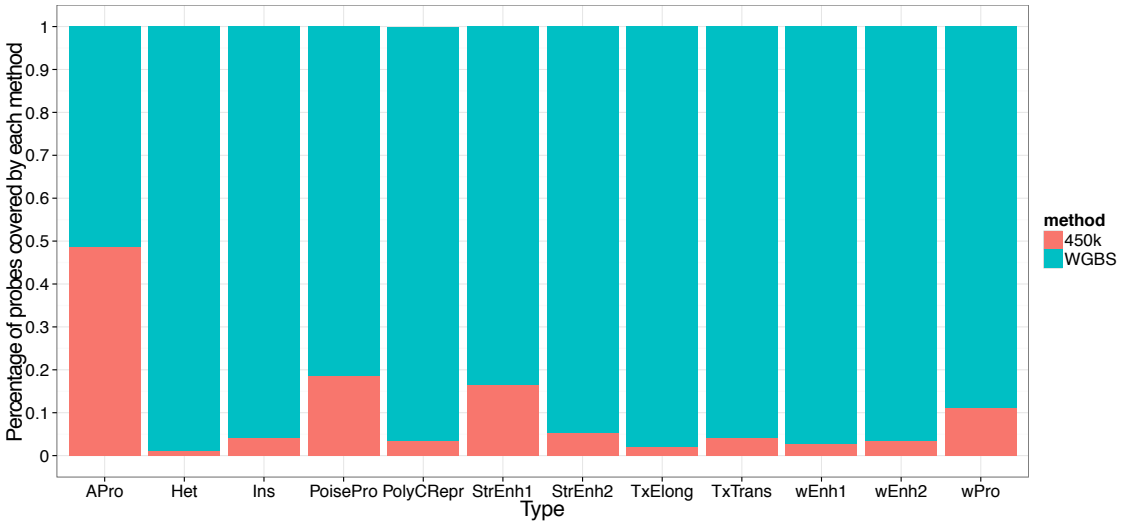
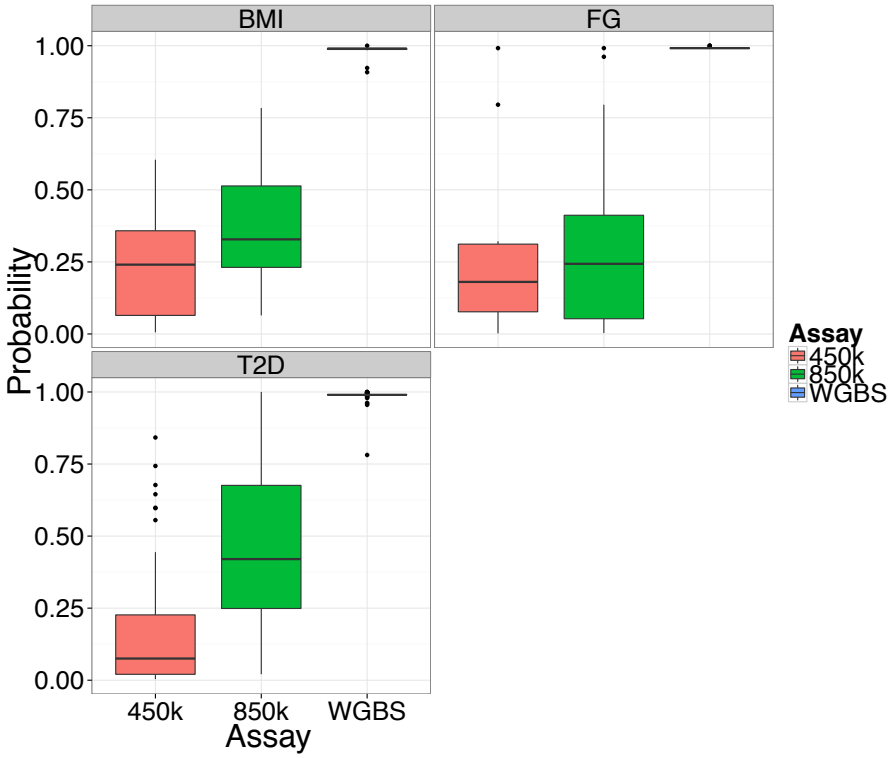


Figure 3.7: 450k and WGBS CpG coverage per islet chromatin state. For each chromatin state (x-axis) the fraction of CpGs for the 450k array (red) in comparison to the WGBS (blue) data is shown (WGBS is considered as the full set containing all CpG sites). Abbreviations: weak Promoter (wPro), Transcriptional Elongation (TxElong), Transcriptional Transition (TxTrans), weak Enhancer (wEnh), Heterochromatin (Het), Poised Promoter (PoisePro), Polycomb Repressed (PolyCRepr), Insulator (Ins), Active Promoter (APro), Strong Enhancer (Strong Enhancer).

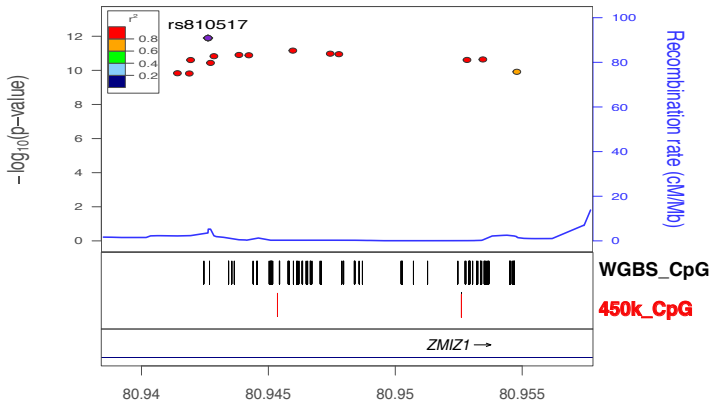
3.3.2 CpG coverage of WGBS and the 450k array in GWAS regions

To quantify the ability of the 450k array to interrogate methylation in GWAS regions, 99% credible set variant data from the DIAGRAM consortium was used, which is one of the largest T2D GWAS datasets to date. Additionally, FG and BMI GWAS credible sets were obtained from the ENGAGE consortium (see Chapter 2 Section 2.5.1 for details). The credible sets were derived using fine-mapping approaches that provide for each variant in the credible set a posterior probability (PP) estimate for driving the GWAS signals. The credible sets contain the variants that collectively capture 99% of the PP at each locus. To estimate the PP captured by the 450k array and WGBS CpG sites, the PP of all 99% credible set variants within 1000bp of a CpG site per locus was determined considering that CpG methylation is highly correlated across short distances. For all investigated GWAS traits, the coverage of the 450k array of likely causal GWAS variants was much lower compared to the WGBS approach (450k array median PP per locus=0.14 VS WGSB median PP per locus=0.99, Fig3.8A). The limited coverage of 450k array CpG sites in T2D GWAS regions is particularly represented by the *ZMIZ1* T2D locus at which the 450k array CpG-sites captures only a PP of 0.10 (Fig3.8B) compared to PP of 0.99 captured by WGBS CpG-sites.

Since Illumina recently developed a new methylation array named Infinium MethylationEPIC 850k array with 400,000 additional CpG sites compared to the 450k array, the PP captured by the 850k array was also determined. While the coverage of causal GWAS variants by the 850k array improved compared to the 450k array, the coverage is still substantially lower compared to WGBS (850k median PP per locus=0.36 vs WGBS median PP per locus=0.99). These results show that WGBS interrogates the methylation status more comprehensively, particularly around GWAS loci, compared to the currently available 450k and 850k methylation array technologies. This observation is driven by the selection bias of the array CpG probes, in particular of the original 450k probes found on both arrays, towards promoter and CGI regions.



(A) GWAS PP captured by different methylation assays



(B) PP at *ZMIZ1* T2D GWAS locus

Figure 3.8: GWAS Posterior Probability distribution captured by WGBS and 450k CpG-sites. (A) Distribution of GWAS Posterior Probabilities (BMI, FG and T2D) captured by CpG sites on the 450k array (read), 850k array (green) and WGBS (blue and black narrow boxplot). (B) LocusZoom[208] plot showing CpG density and credible set SNPs. SNPs are shown with P-values (dots, y-axis left), recombination rate (line, y-axis right) and chromosome positions (x-axis) while CpG and gene annotations are shown below. These annotations include CpGs identified from WGBS (black stripe) and 450k CpG density (red stripes) and gene overlap (*ZMIZ1* label). At the very bottom is the region size shown in Megabases (Mb).

3.3.3 Identification of islet regulatory regions using WGBS and their role in T2D and FG GWAS regions

3.3.3.1 Identification of hypomethylated regulatory regions from WGBS data

To explore the role of DNA methylation in human islet function and T2D development, hypomethylated regulatory elements were identified from the pooled high-pass WGBS data and classified into CpG-poor LMRs and CpG-rich UMR regions. LMRs and UMRs were previously shown to correspond to distal regulatory elements such as enhancers and proximal regulatory elements including promoters, respectively[206]. In total, the analysis detected 37.1k LMRs (median CpG number 5 per region with a region being defined as a continuous stretch of hypomethylated CpGs, see Section 3.2.2.4 for further details) and 13.6k UMRs (median CpG number 106 per region, Fig3.9A). Consistent with previous observations, the identified LMR and UMR regions demonstrated substantial overlap with previously described chromatin state enhancer (82% of LMRs) and promoter (96% of UMRs) regions (derived from Parker et al 2012 and Pasquali et al 2014[85,86]), respectively.

These regulatory regions were characterised further by overlapping them with different islet annotations that have been identified in previous studies. LMRs and UMRs were tested for enrichment in various islet annotations including previously described islet expression QTLs[159], genome-wide CTCF binding sites[160] and relevant human islet ChIP-seq transcription factor binding sites (TFBS) including FOXA2, MAFB, NKX2.2, NKX6.1 and PDX1[85]. These islet TFBS were shown to represent important cis-regulatory regions in islets enriched for T2D GWAS signals. In addition, publicly available islet mQTLs and associated CpG-sites with correlated methylation status were obtained from a study that used 450k methylation data from 89 human pancreatic islet samples for the mQTL detection[99]. Since these mQTL SNPs and associated CpG-sites were derived from 450k array methylation data they likely represent only a subset of all islet mQTL SNPs and correlated CpGs. This is due to the limited number of CpG sites represented on the 450k array and the depletion of mQTLs in CpG-dense region primarily interrogated by the 450k array.

Islet LMRs and UMRs were significantly enriched (adjusted P <0.05) in islet eQTL (LMR FE=2.2, UMR FE=6.4), mQTL SNPs (LMR FE=2.6, UMR FE=3.9) and CpG sites (LMR FE=4.3, UMR

FE=14.1) as well as islet TFBS (LMR FE=17.2-24.3, UMR FE=5.4-14.8) compared to randomised sites (Fig3.9B). In addition, LMR and UMR enrichment was also determined for Variable Methylated Positions (VMPs) defined here as showing high deviations from the mean CpG methylation across at least 5 individual WGBS islet samples (see Section 3.2.2.5). Significant enrichment of LMRs and depletion of UMRs was found in the top 5% VMPs (LMR FE=2.3, UMR FE=0.6) representing the most variable sites in the genome (Fig3.9B). Although technical biases related to coverage may increase variability at sites of intermediate methylation, these findings, which are consistent with previous results, suggest that LMRs are more variable than UMRs. In summary, these results indicate that WGBS methylation data can be used to successfully identify LMRs and UMRs that represent important distal and proximal regulatory domains, respectively, and are strongly enriched for several types of islet annotations.

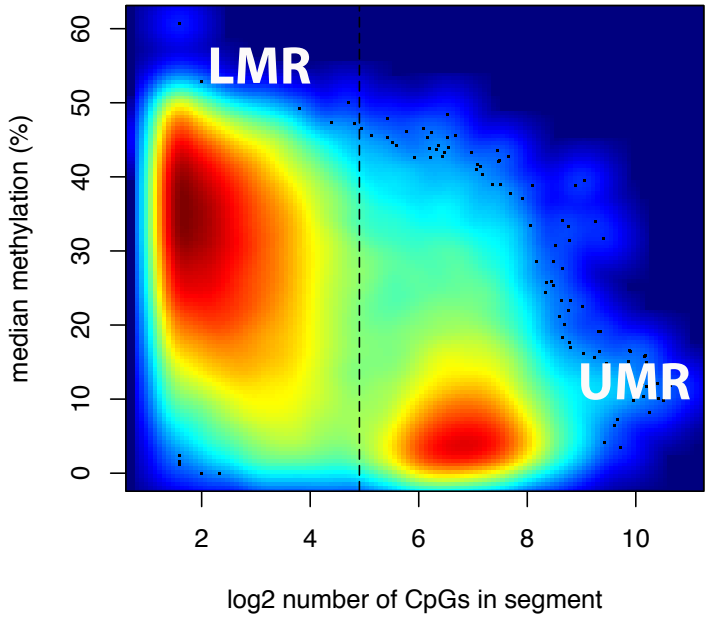
3.3.3.2 Enrichment of WGBS hypomethylated regulatory region in T2D and FG GWAS signal

Considering the regulatory role of these hypomethylated LMRs and UMRs, these regions were tested for enrichment in T2D and FG GWAS regions. To determine whether the enrichment was islet-specific, WGBS Epigenome Roadmap data from different tissues including adipose and liver, which have also been proposed to play a role in T2D susceptibility, were obtained. Also, Epigenome Roadmap WGBS whole pancreas data of which only a small fraction (1%) represents endocrine islet cells, was accessed. For each Epigenome Roadmap tissue LMRs and UMRs were determined separately using the R package methylseek and the identified LMRs and UMRs from all tissues were tested independently for enrichment in GWAS regions by applying a hierarchical modelling approach described as FGWAS. In contrast to using PP, FGWAS determines enrichment based on the association statistics of all genetic variants independent of genome-wide significance. Islet LMRs showed significant enrichment in T2D (log2FE=3.2, Confidence Interval (CI)=2.3 to 3.9) and FG (log2FE=3.9, CI=2.3 to 5.1) GWAS regions while no enrichment was found across other likely non-islet mediated GWAS traits such as BMI (log2FE=-24.4, CI=-44.4 to 3.2) and TC (log2FE=-24.4, CI=-44.4 to 2.0, Fig3.10A). While the enrichment of islet LMRs (log2FE=3.2, CI=2.3-3.9) was only slightly stronger compared to enhancer regions predicted from ChIP-seq

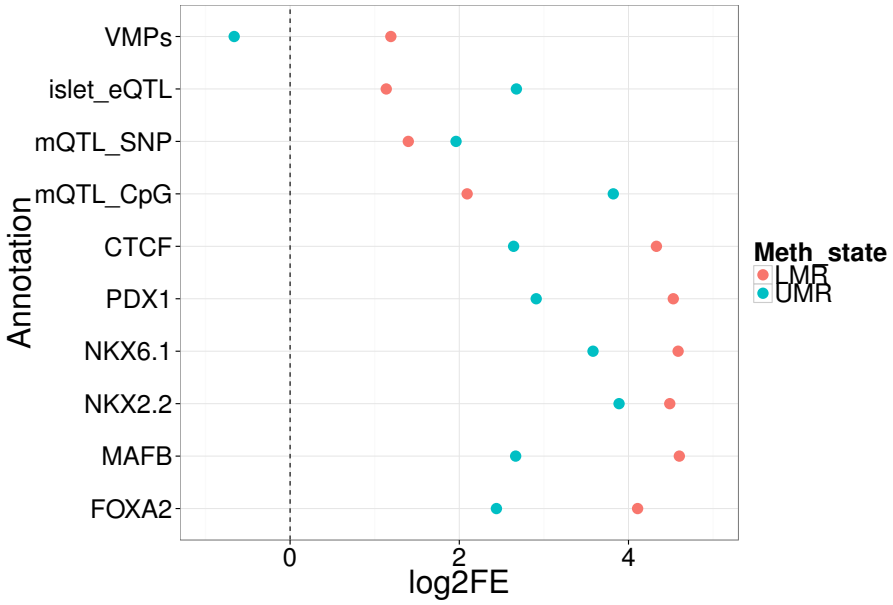
histone marks (Strong Enhancer1 log2FE=2.8 CI=1.8-3.5, Weak Enhancer2 log2FE=2.5, CI=1.8-3.2, only T2D results shown Fig3.10B), this highlights the importance of DNA methylation in identifying T2D- and FG-associated regulatory regions.

When testing the non-islet LMR regions for enrichment across GWAS traits (Fig3.10A), enrichment signals for T2D and FG GWAS enrichment in Epigenome Roadmap whole pancreas LMRs was detected (T2D log2FE=2.4, CI=1.1-3.2 and FG log2FE=3.0, CI=1.1-4.3), while adipose and liver LMRs were not enriched in T2D or FG GWAS regions. The shared enrichment of whole pancreas and islet LMRs is likely the result of shared methylation marks during development considering that whole pancreas and islets are derived from the same embryological root. Liver LMRs were also found to be enriched in TC GWAS regions (log2FE=2.1, CI=0.4-3.7) which is in accordance with the fact that the liver is crucial for controlling lipid levels considering its role in cholesterol, phospholipids and lipoproteins synthesis[209].

In contrast to LMRs and consistent with previous studies, we did not find any enrichment of promoter-like UMRs in T2D or FG GWAS signal across any tissue; however, UMRs across all tissues were enriched for BMI GWAS signal (log2FE=3.6-3.8). This is consistent with previous observations that regions near promoters are enriched for BMI GWAS signal in certain cell-types[158] and that CpG-methylation levels at promoters are associated with BMI GWAS variants[210]. Overall these finding highlight the importance of islet LMRs to understand islet-mediated GWAS traits such as T2D and FG.

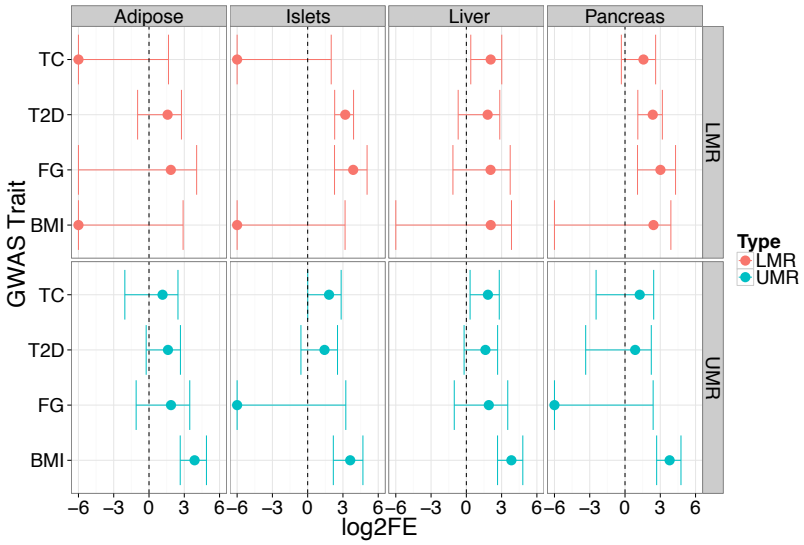


(A) LMR and UMR methylation level and CpG number

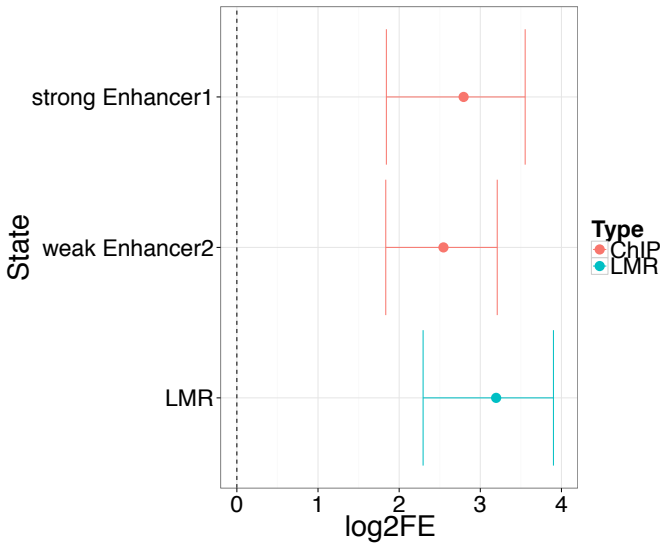


(B) Chromatin state

Figure 3.9: LMR and UMR methylation status and enrichment in regulatory regions (A) Methylation levels in percent (y-axis) and log2 CpG density (x-axis) of UMR and LMR regulatory regions with the dashed line indicating the CpG-number (30 CpGs) that distinguishes LMRs and UMRs.(B) Enrichment of LMRs and UMRs in different types of regulatory regions including islet relevant TFBS (PDX1, NKX6.1 NKX2.2, MAFB, FOXA2) from[85], significant expression QTLs[159] and significant islet mQTL SNPs and associated CpG sites obtained from[99] as well as regions near (within 1000bp) Variable Methylated Positions (VMPs) derived from the islet WGBS methylation data.



(A) LMR UMR FGWAS enrichment



(B) LMR enrichment compared to ChIP chromatin states

Figure 3.10: Enrichment of hypomethylated regions in T2D GWAS regions (A) FGWAS log2 Fold Enrichment (log2FE, x-axis) of LMRs and UMRs identified from adipose, islet, liver and pancreas tissue in various GWAS traits (BMI, FG, T2D, TC). (B) Comparing the FGWAS log2FE (x-axis) of islet LMRs to previously identified ChIP-seq enhancer chromatin states (y-axis) from[86].

3.4 Discussion

The main findings of this chapter were:

- the comparison between the 450k and WGBS islet methylation data showed that WGBS is important to fully characterise the islet methylome considering the selective and partial coverage of the 450k array, in particular, around GWAS regions
- based on the overlap with known islet regulatory regions the results demonstrated that WGBS methylation data can be used to identify hypomethylated regulatory regions corresponding to promoters (UMR) and enhancers (LMR)
- the FGWAS analysis determined enrichment specifically for islet enhancer-like LMRs in T2D and FG GWAS signal while no enrichment was observed in other GWAS traits

This chapter has characterised for the first time the whole-genome human pancreatic islet methylome at single base-pair resolution. The human islet WGBS data described in this chapter showed that genome-wide DNA methylation information is critical to fully characterise the role of DNA methylation in islet regulation.

A comparison between methylation chip array technology (450k and 850k array) and the WGBS data clearly highlighted the differences in the ability of these methods to capture the DNA methylation status of human islets globally and near T2D GWAS regions. While the 450k array accurately determines DNA methylation at 480,000 interrogated CpG-sites, the design bias of the 450k array probes (found on both arrays) towards promoter and CpG-island regions limits the use of the 450k array to accurately characterise the methylation status across the genome, in particular, near distal regulatory regions. Also, the methylation arrays provide only selective coverage of GWAS regions, and they are not well-suited to study the functional role of methylation at GWAS loci. These observations are consistent with previous studies that have shown that most distal enhancers are enriched for GWAS signal while proximal regulatory elements such as promoters are not strongly associated with GWAS regions[76,82,84–86].

Subsequently, the whole-genome islet DNA methylation data described here was used to determine non-coding regulation annotations defined as LMRs and UMRs. These regulatory LMRs and UMRs showed consistent overlap with enhancers and promoters, respectively, and were found to be enriched in several other relevant genomic islet annotations, thus, suggesting that hypomethylated LMRs and UMRs represent important regulatory regions in the genome. Global enrichment analysis of these annotations in GWAS association data from various diseases found that enhancer-like islet LMRs showed strong enrichment specifically for T2D and FG GWAS signal. In addition to islet LMRs, pancreatic LMRs were also found to be enriched for T2D and FG GWAS regions suggesting shared methylation marks across these developmentally similar tissues. In contrast, promoter-like islet UMRs did not show enrichment in T2D-related traits which is consistent with the lack of coverage in GWAS region observed for the methylation array probes that focus on promoter regions. Overall these results are consistent with previous data that have found that tissue-specific methylation regulatory domains are critical for tissue function and enriched in GWAS signal[112].

While these data clearly highlighted that the WGBS provides a more comprehensive analysis of the islet DNA methylome, in particular, around GWAS regions, this study provided only limited information on the extent of variation in DNA methylation. The generation of large human pancreatic islet WGBS datasets is generally limited by the availability of these samples and the high costs associated with whole-genome sequencing. Thus, associations between DNA methylation changes, genetic variation and disease status represented by mQTLs and disease DMRs, could not be elucidated in this dataset. With sequencing costs decreasing over time, future studies should focus on accumulating larger numbers of high depth WGBS islet methylomes of both healthy and diabetic individuals in order to shed further light on the regulatory role of DNA methylation in human pancreatic islets and T2D development. Our research group is currently generating additional WGBS data for human islets to further improve understanding of the epigenetic DNA methylation mechanisms governing islet regulation and T2D development.

In conclusion, the data presented in this chapter provides a comprehensive view of the human pancreatic islet DNA methylome, identifies and characterises hypomethylated regulatory regions and links these regions to T2D-related GWAS traits, thus, highlighting the usefulness of WGBS

data in understanding islet regulation. The integration of islet epigenetic DNA methylation data and other islet epigenomic data to further characterise the biological mechanisms acting at T2D GWAS regions will be described in Chapter 5.

CHAPTER 4

Determining the role of open chromatin in islet regulation and at T2D GWAS loci

4.1 Introduction to open chromatin

As described in Chapter 1, DNA is associated with histone protein complexes which are collectively referred to as chromatin and determine DNA accessibility for DNA-binding proteins. Highly accessible DNA is usually found in active and transcribed regions in the genome while inaccessible chromatin is associated with inactivity. Changes in histones and chromatin structure play a significant role in cell regulation and are involved in development, DNA replication, transcription, cell-type specificity, and disease.

4.1.1 Approaches to assay the open chromatin landscape of cells

Recent advances in next-generation sequencing have facilitated large-scale approaches to investigate histone PTMs and other DNA-binding proteins and provided valuable insights into genome function. Chromatin Immunoprecipitation followed by next generation sequencing (ChIP-seq[68]) is one the most common methods used. It involves cross-linking of DNA and associated proteins followed by immunoprecipitation of proteins of interest using antibodies. The DNA is then purified from these DNA-protein complexes, PCR-amplified and sequenced. Accumulation of reads when analysing read density across the genome is indicative of histone marks or bound protein[118,211].

Until recently Formaldehyde-Assisted Isolation of Regulatory Elements followed by sequencing (FAIRE-seq,[212]) or DNase hypersensitivity followed by sequencing (DNase-seq[69]) were the most commonly used methods to directly study chromatin accessibility. While FAIRE-seq makes use of formaldehyde to cross-link DNA to proteins followed by sonication and phenol-chloroform extraction to releases nucleosome-free DNA, DNase-seq applies DNase which preferentially digests nucleosome-free DNA leading to an enrichment of nucleosome-depleted fragments. After DNA isolation, both FAIRE-seq and DNase-seq involve PCR amplification of the isolated DNA and sequencing of the amplified fragments to identify open chromatin regions in the genome enriched for sequenced reads. While all these technologies provide highly useful information on the location of histone marks, DNA-bound proteins or open chromatin regions, the large numbers of cells (100,000 to 10 million depending on the method) required for the experiments hinder the use of these methods in primary cell types. Investigating variability across individuals is particularly challenging using these technologies since primary tissue samples are often hard to collect in large numbers[118,211].

A novel method termed Assay for Transposase-Accessible Chromatin using sequencing (ATAC-seq) has partially overcome these limitations and requires only between 500-50,000 cells, thus, allowing researchers to study primary tissue samples from single individuals. ATAC-seq uses a Tn5 transposase charged with primers that are preferentially integrated into nucleosome-free open chromatin regulatory regions. After the transposition reaction, the open chromatin regions are PCR-amplified using the inserted primers and sequenced. ATAC-seq provides similar sensitivity and specificity in terms of signal compared to DNase-seq. However, it is much more time efficient and

requires much lower cellular input. Thus, ATAC-seq is highly useful for assaying primary tissue and determining variability in open chromatin across both healthy and diseased tissue[123].

4.1.2 The role of chromatin regulation in islet development, islet function and T2D

The application of these next-generation sequencing approaches, to fully differentiated human pancreatic islets and developmental pancreatic precursor cells, has shown that epigenetic factors contribute significantly to the development and mature function of pancreatic endocrine cells. Pancreatic tissue is derived from the endodermal germ layer that also gives rise to the gastrointestinal tract and the respiratory system. The progression from early endoderm into fully differentiated organs is stepwise through the induction of specific lineage TFs that regulate stage-specific gene expression programmes. While the TFs involved in the differentiation process are quite well characterised, the associated chromatin changes were not well understood until recently[213].

Early work in mouse Embryonic Stem Cells (ESCs) revealed that promoter regions can be marked by both active histone marks (H3K4me3) and repressive marks (H3K27me3). These promoters are described as being in a "bivalent state" and predominantly found near developmental regulatory genes which are inactive in ESCs but in a "poised" state prepared for later activation[214]. During developmental progression, H3K27me3 is removed from these bivalent domains leading to nearby gene activation (Fig4.1A). At the same time, the addition of H3K27me3 to active regions leads to inactivation, thereby, providing a flexible system of chromatin regulation that controls lineage-specific gene expression programmes[215].

Recent advances in *in vitro* pancreatic endocrine differentiation protocols have provided the opportunity to map the epigenome throughout the whole development of induced Pluripotent Stem Cells (iPSCs) into pancreatic endoderm. Through this approach, epigenetic chromatin changes can be tracked across multiple stages of development. Using these novel protocols, Xie et al 2013 confirmed that H3K27me3 regulates both activation and silencing of key endodermal regulators such as *SOX17* during pancreas development[216].

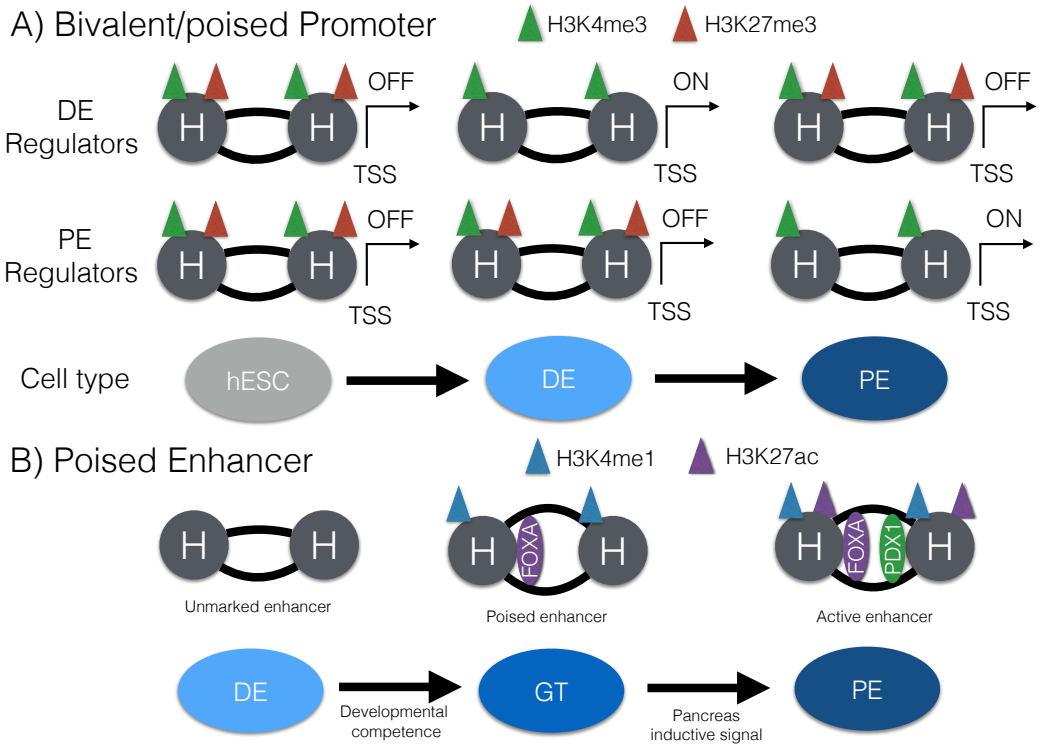


Figure 4.1: Regulation of bivalent promoters and enhancers (A) Bivalent promoters are marked by both active (H3K4me3, green) and inactive (H3K27me3, red) that leave these sites in a bivalent state poised for later activation. In hESCs (grey), developmental regulators for Definitive Endoderm (DE) and Pancreatic Endoderm (PE) are in a bivalent state marked by both H3K4me3 and H3K27me3. In the DE (light blue) stage, DE regulators become active by removal of H3K27me3 while PE regulators remain in the bivalent state. Upon transition to the PE stage (dark blue), PE regulators become active through loss of H3K27me3 while DE regulators are turned off again by addition of H3K27me3 (for simplicity not all developmental intermediates are shown)[216]. (B) Upon transition from DE (light blue) into Gut tube (GT, medium blue) inactive unmarked enhancers acquire developmental competence through accumulation of the mark H3K4me1 (blue) which allows pioneering transcription factors (FOXA, purple) to bind creating poised enhancer states. Inductive pancreatic signalling then leads to binding of lineage specific transcription factors (e.g. PDX1, green) at these poised sites which then create active enhancer states marked by H3K27ac (purple triangle). These active enhancer region then induce gene expression programmes that initiate the formation of PE (dark blue)[217]

As for poised promoters, recent studies also showed that, during both pancreatic and hepatic differentiation, poised chromatin marks accumulate at enhancers to provide developmental competence for cells. That is, lineage-specific enhancers become poised prior to lineage induction to provide cells with the ability to respond appropriately to environmental signals that initiate developmental progression of cells. In a sequential fashion, these enhancer regions first transition from unmarked chromatin to poised histone marks such as H3K4me1 and these poised enhancers are then bound by pioneer TFs including the FOXA family of TFs. Finally, lineage-specific endodermal TFs including PDX1 accumulate at these poised sites to mediate activation of enhancers (characterised by H3K27ac marks) and target gene expression[217] (Fig4.1B). Apart from early pancreatic development, chromatin state changes are also critical for the differentiation into the different pancreatic cell types. For instance, modifications of histone acetylation levels through inhibition of histone deacetylases (HDAC) influence commitment to differentiate into endocrine tissue as well as into endocrine islet cell subtypes (differentiated based on their hormone production) such as alpha (glucagon-producing), beta (insulin-producing), delta (somatostatin-producing), polypeptide (pancreatic polypeptide-producing) and epsilon (ghrelin-producing) cells[213,218].

In addition to enhancing the understanding of pancreatic development, these studies may also improve iPSC-to-pancreatic-endoderm differentiation protocols that could be used to generate physiologically- and disease-relevant islet cell models. As described in the introduction of the thesis, GWAS has placed pancreatic islets at the centre stage of T2D pathophysiology; however, current rodent and human cellular (EndoC-ßH1) beta-cell models are still functionally limited compared to human primary islet cells. While current *in vitro* protocols have advanced to the extent that they yield reasonable levels of islet-like cell types[219,220], late stage cell populations are often relatively heterogeneous and have immature function compared to human beta-cells. Epigenomic modifications may be responsible for these differences between mature human islets and *in vitro* generated islet models derived from iPSCs. For instance, Xie et al 2013 found that cells differentiated *in vitro* retained abnormal H3K27me3 repression signatures at important endocrine genes, such as *INS* and *GLP1*, compared to *in vivo* matured cells. Consistent with these abnormal chromatin marks, *in vitro* generated cells are functionally less mature compared to *in vivo* differentiated cells[216]. The source of these differences and of the variability in these iPSCs is still debated. Some studies show an individual's genotype to be the primary driver of transcriptional and epigenetic pro-

files (and differentiation capacity) when comparing iPSCs derived from multiple tissues of the same donor[221–223]. However, a body of work also supports an 'epigenetic memory' in mouse and human iPSCs[224–226]. That is, pluripotent iPSCs maintain, in particular at low passage number, epigenomic (and transcriptomic) signatures of the original donor cell including gene expression, methylation profiles, histone modifications, and areas of open chromatin. For instance, previously collaborators generated iPSCs (defined as BiPSCs) from human beta-cells and found that these cells were reprogrammed based on pluripotent marker expression and the ability to differentiate into all 3 germ layers[224]. However, a comparison of these BiPSCs derived from human beta-cells with iPSC derived from fibroblasts (FiPSCs) and human ESCs suggested that BiPSCs retained an epigenetic memory: histone H3 tail acetylation-ChIP-qPCR showed that the promoters of islet TFs *PDX1* and *INS* were significantly more open in BiPSCs, which was also reflected by hypomethylation of promoters for islet-relevant genes including *PDX1*. This translated into an enhanced capacity of these cells to differentiate down the endocrine pancreas lineage (versus FiPSCs and hESCs)[224].

Characterising these epigenetic differences during differentiation and development could be used to improve from 'iPSC to endocrine' pancreas differentiation protocols and enhance understanding of T2D development. For instance, using RNA-seq on cells from several stages of an iPSC-derived pancreatic endodermal developmental model, a recent study has shown stage-specific gene expression at more than half of the known T2D GWAS loci. This observation could indicate that these developmentally active genes may play a role in the genetic susceptibility of T2D[227] and that beta-cell development is linked to T2D pathogenesis. Further support is provided by the fact that T2D-loci were observed to be enriched in cell cycle genes[18]. Furthermore, the efficient *in vitro* generation of islet cells could also lead to the development of novel therapeutic approaches either directly through transplantation of functional cells or indirectly by using these cells as drug development platform.

Assays to investigate open chromatin have also been successfully used in adult islets to improve the understanding of islet regulation and T2D development. Gaulton et al published in 2010 the first high-resolution FAIRE-seq open chromatin map of adult human pancreatic islets to identify islet-specific regulatory regions[84]. These regions were found to be enriched for islet-specific TFs and to form clusters of distal open regulatory elements that are critical for islet gene expression

and biology. Additionally, Gaulton et al found allelic imbalance at the known T2D GWAS *TCF7L2* variant rs7903146 that modifies enhancer activity in cellular models, thus, highlighting potential molecular mechanisms governing T2D risk at this locus[84]. Parallel studies, which characterised the role of distal enhancer clusters in pancreatic islets and T2D development in more detail through the integration of ChIP-seq and open chromatin marks, are described in Chapter 5 Section 5.1.2. The development of novel technologies such as ATAC-seq has provided new opportunities to study rare primary tissue cell types that could previously not be investigated. A recent study took advantage of the low cellular input requirement of ATAC-seq and combined ATAC-seq open chromatin with RNA-seq expression data of human alpha and beta-cells to identify chromatin and gene signatures specific to either of these endocrine cell types and linked these regions to T2D GWAS variants[228].

In conclusion, determining the chromatin landscape of pancreatic developmental and adult cell types provides an efficient way of characterising both developmental and mature molecular processes involved in pancreatic endocrine and islet biology. However, most islet chromatin maps only represent a very static view due to the high cellular input required for both ChIP and FAIRE-seq.

The work in this chapter is divided into two main parts with the following aims:

- to generate and analyse human islet ATAC-seq open chromatin data from 29 individual human islet samples and link open chromatin regions to T2D development
- to use ATAC-seq to characterise the epigenomic landscape of human BiPSCs and FiPSCs samples to identify differential open chromatin sites (DOCS) that may explain the enhanced ability of BiPSCs compared to FiPSC to differentiate into insulin producing cells, and ultimately may be used to develop better endocrine differentiation protocols.

4.2 Methods

4.2.1 Samples processed for ATAC-seq

4.2.1.1 Human pancreatic islets

Human pancreatic islets (n=29) were retrieved from deceased donors from the Oxford DRWF Human Islet Isolation Facility and stored for 1-3 days in CMRL or UW. The latter were reactivated in CMRL for 1h before processing them further. Approximately 50 good quality islets (representing about 50,000 islet cells) per sample were hand-picked into 1.5ml Eppendorf tubes containing CMRL and immediately processed for ATAC-seq as described in Chapter 2 Section 2.3. These islet samples were used to characterise the open chromatin landscape in human pancreatic islets and near T2D GWAS loci.

4.2.1.2 iPSC sample generation and characterisation

As part of a collaborative effort, 5 BiPSCs from 3 donors were generated from human pancreatic islets by Liraz Shenhav at Tel Aviv University, Israel. Human pancreatic islets were received 2-4 days following isolation, dissociated into single cells, labelled for beta-cell lineage tracing and cultured as described[229]. Labelled Beta-Cell-Derived (BCD) cells were sorted using a FACS Aria cell sorter (Becton Dickinson, San Jose, CA). BiPSCs were generated as previously described[224] with the following modification: sorted BCD cells were expanded to passage 8-16 and transduced with a polycistronic lentiviral vector containing the 4 reprogramming TF *OCT4, SOX2, KLF4* and *c-MYC* (a gift of Gustavo Mostoslavsky). ES-like colonies emerged 3-7 weeks following transduction. Five of these colonies, originating from three different human donors, were isolated and expanded for further analysis. The beta-cell origin of iPSC clones was validated by DNA PCR analysis of the recombined reporter cassette as previously described[224].

Liraz Shenhav also obtained 5 FiPSC from 2 donors through the IMI/EU sponsored StemBANCC consortium via the Human Biomaterials Resource Centre, University of Birmingham and were generated as described in[227].

Characterisation and quality control of all these iPSCs in terms of pluripotency and karyotyping was performed by Liraz Shenhav at Tel Aviv University. The generated and quality controlled iPSCs were then processed using ATAC-seq as described in Chapter 2 Section 2.3 . The process up until the transposition reaction was performed by Liraz Shenhav at Tel Aviv University. The tagmented and purified DNA was then sent to OCDEM, University of Oxford, Oxford, UK on dry ice for further processing including PCR amplification, purification and QC of ATAC-seq DNA sequencing libraries.

4.2.2 Sequencing of ATAC-seq samples

Following the ATAC-seq processing in Israel (DNA isolation and tagmentation performed by Liraz Shenhav) and Oxford (PCR amplification, purification and QC performed by myself), libraries were sequenced at the High-Throughput Genomics group which is part of the Wellcome Trust Centre for Human Genetics, University of Oxford, Oxford, UK. Samples were sequenced as 4-6plex libraries across 1-3 Hiseq2500 lanes with 50bp paired-end read length.

4.2.3 Processing of reads

Raw FASTQ reads were processed with an in-house pipeline (first described in[230] and on the following website[231]). Specifically, library and sequencing quality was checked with FASTQC[200] and reads were mapped to the human genome hg19 via bowtie (version 1.1.0,[232]) with default settings but -m 2, and maxins 2000 which allows mapping of reads with a maximum number of 2 alignments and a maximum insert size of 2000bp. For reads that could not be aligned the first time, adapters were removed with Trim Galore[233] at the 3 prime end (settings -length 10, -qualFilter 20) to enhance the chance of mapping. The resulting trimmed reads were then mapped again with bowtie. Any remaining unmapped and trimmed reads were processed with FLASH (version 1.2.8 settings -m 9 -x 0.125,[234]). Since overlapping paired-end reads can often not be mapped by alignment software, FLASH combines the read pair and reconstructs a read pair without overlap. These are then realigned a third time using bowtie. PCR duplicates are then removed from the mapped

bam files using samtools rmdup function[203]. Additionally, all reads overlapping any of the "unmappable" UCSC Duke blacklisted hg19 regions are also removed from the final bam file.

4.2.4 Peak calling and read normalisation

Peaks were called from filtered bam files using MACS2 (settings FDR <0.01, –nomodel, -g hs)[235] to identify regions of open chromatin. Then peaks were merged for each cell type separately (human pancreatic islets, FiPSCs and BiPSC) using bedtools mergeBed function and used to normalise read depth across peaks by applying the bamnormalise and bamsummary function of the software package deeptools (using standard settings). PCs and Spearman's rho were calculated from normalised reads using R to compare samples across all peaks independent of cell types.

4.2.5 GWAS enrichment analysis of islet open chromatin peaks

As described in the main methods section ATAC-seq open chromatin regions were tested for enrichment in T2D, FG, BMI and TC GWAS regions. Open chromatin peaks were then split into likely islet enhancers and promoters based on distance to the TSS of expressed genes (obtained from Nica et al[236]). Proximal promoter regions were defined as within 2500bp upstream and 500bp downstream of the TSS while the remaining regions were classified as distal enhancers.

4.2.6 Identification of differentially open chromatin regions (DOCS) in BiPSCs and FiPSCs

Identification of DOCS was performed using the perl software diffReps[237] (standard settings were used with a defined fragment size of 50bp,) which normalised the distribution of reads for each sample. The software then applied a 1000bp sliding window with a 100bp step size throughout the genome to test each of these windows for differences in normalised read number between BiPSCs and FiPSCs using a negative binomial test. After analysing the whole genome, overlapping windows with a significant difference in normalised read depth (predefined as P<0.0001) were

merged together and the merged region were retested. Multiple testing correcting using an FDR approach was performed based on the number of merged and retested regions.

4.2.7 Regulatory chromatin state and TFBS information used for DOCS enrichment analysis

Expanded 18-state chromatin states from the Epigenome Roadmap data repository[82] were obtained for 98 cell types. Predefined endodermal weak enhancer and FOXA2 binding sites were downloaded from publicly available resources[217]. Islet TFBS (FOXA2, NKX2.2, NKX6.6, PDX1 and MAFB) were processed by Dr Kyle Gaulton as described in Chapter 2 Section 2.6.2 and non-islet TF (erythrocyte POLR2A (GEO:GSM935462), hESC BCL11A (GEO:GSM803396) and hESC SP1(GEO:GSM803377) were obtained from Encode[65].

4.2.8 Enrichment calculations for DOCS

Enrichment of DOCS in chromatin states was calculated by estimating the sum of log2 Fold Chance (log2FC) for each chromatin state separately based on true DOCS. This value was then compared to the mean of the sum of log2FC for each chromatin state derived from 1000 permutations of shuffled sites (true sites were shuffled using the shuffleBed function implemented in bedtools). For each state and each of the 98 cell types, P-values were calculated based on how often the shuffled sum of log2FC was higher compared to the true sum of log2FC (minimum P-value 0.001). For each of the 6 germ layer/pluripotent cell types (Ectoderm, Endoderm, Mesoderm, ESC, iPSC and Others) and for each chromatin state separately, the average P-value was calculated across all cells of a given type and chromatin state. These averaged P-values were FDR corrected and used for the analysis.

For the other datasets (including endodermal weak enhancer and FOXA2 TFBS, islet and non-islet TFBS) enrichment was calculated as described in the Chapter 2 Section 2.6.2. In short, enrichment was estimated based on comparing the true number of overlapping sites with any given annotation compared to the mean number of overlapping sites derived from 1000 permutations (tested sites

were randomly shuffled using shuffleBed function). P-values were calculated depending on how often the random number of overlapping sites exceeded the true number of overlapping sites (min P-value 0.001). The resulting P-values were adjusted for multiple-testing.

4.2.9 Gene and Pathway analysis of DOCS

Up- and downregulated DOCS were analysed separately using the software "Genomic Regions Enrichment of Annotations Tool" (GREAT)[238] with standard settings. To identify genes associated with DOCS, the 'basal+extension' method was used, which determined for each gene a regulatory domain. That is, each gene was assigned a region near its TSS (basal domain 5kb up and 1kb downstream of TSS) which was independent of the location of any other genes. The gene regulatory domain was then created by extending the basal domain near the TSS up to the next basal regulatory domain of a different gene (with a maximum extension of 1000kb). Once DOCS were linked to genes based on the overlap with a gene regulatory domain, DOCS were split into Bi-DOCS (more open in BiPSCs) and Fi-DOCS (more open in BiPSC). For each (Bi-DOCS and Fi-DOCS) two types of enrichment tests were performed with the whole genome as background. A binomial test determined the probability of an annotation to be assigned a random genomic region. This test was based on the probability of a part of a genomic region overlapping a gene regulatory domain compared to the total size of the genome. In addition to the binomial test a hypergeometric test was also performed determining gene set enrichment. A test gene set (including all genes at which test regions overlap the gene regulatory domain) was tested for enrichment in any given annotation compared to the whole genome gene set. Pathway and GO terms with a minimum binomial region fold enrichment and hypergeometric gene enrichment >2 and binomial and hypergeometric FDR <5% were reported with the associated genes.

In addition to evaluating enrichment of all Bi-DOCS and Fi-DOCS, the enrichment of Bi-DOCS and Fi-DOCS overlapping specific hESC chromatin states (derived from hESC H1 cell line) was determined by defining the foreground set as chromatin states overlapping DOCS while as background set all hESC-H1 chromatin states (of a given type) were used. The performed enrichment analysis was similar to the gene hypergeometric tests: the foreground set corresponded to the test set and enrichment in any given annotation was calculated compared to the background set (gene

enrichment >2 and FDR<0.05 was considered significant).

4.2.10 *In silico* validation of DOCS

Sample labels were randomised, mixed (group 1: 3 BiPSCs and 2 FiPSCs vs group2: 3 FiPSCs and 2 BiPSCs) and re-analysed using diffReps as described above to evaluate tissue of origin and genotype effects by comparing the number of true vs random DOCS.

To determine whether differences in karyotype largely influence the results, samples with abnormal karyotypes were removed (BiPSC-D.1, BiPSC-D.2) and the diffReps analysis was repeated in the same way as before. Correlation between DOCS identified in all samples and DOCS identified in karyotypically normal samples was calculated using Spearman's rho.

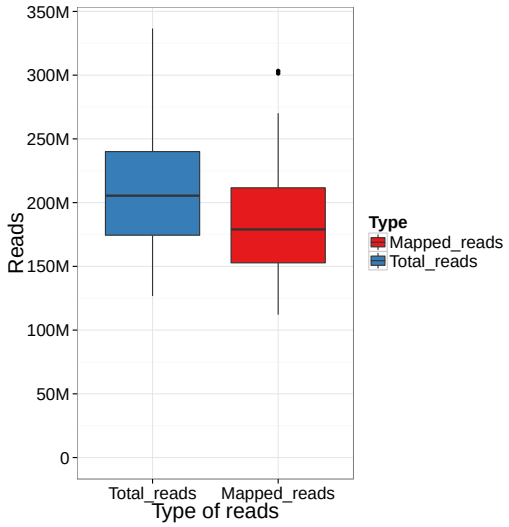
4.3 Results

4.3.1 Characterising the open chromatin landscape in human pancreatic islets using ATAC-seq

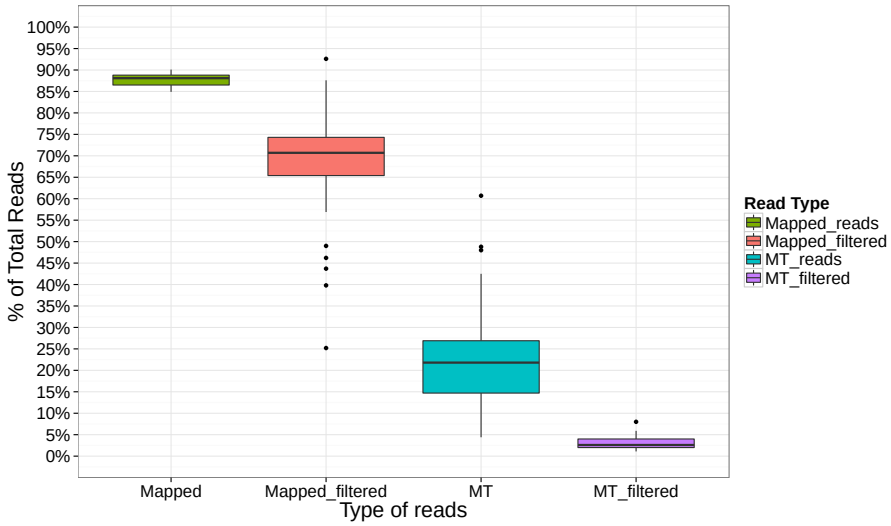
4.3.1.1 Identification of open chromatin regulatory regions

Twenty-nine, human pancreatic islet samples were processed using ATAC-seq and a median number of 205 million reads was obtained per sample (Fig4.2). Per sample a median number of 179 million reads (median mapping efficiency=88.1%) could be mapped to the genome (hg19). After filtering duplicated reads and reads within unmappable regions, the median number of reads per sample was reduced to 112 million. The large number of reads that were filtered out was related to the high percentage of mitochondrial PCR duplicates ranging from 4.4-60.7% (median=21.8%) which are generated during the ATAC-seq process: mitochondrial DNA is not associated with chromatin in the cell and is thus, easily accessible for the transposase.

Next, open chromatin regions were predicted using the software MACS2 which identifies regions



(A) Read number per sample

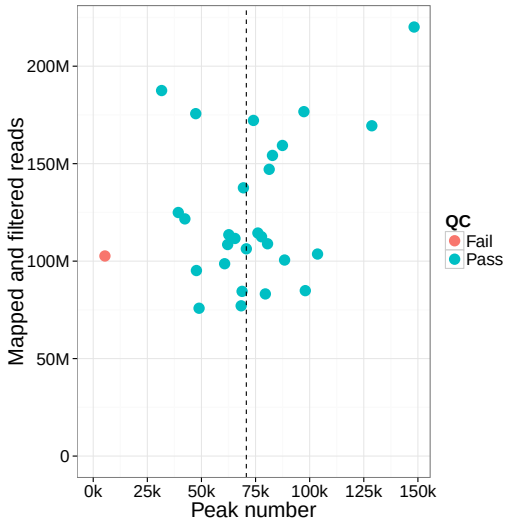


(B) Percentage of reads per sample

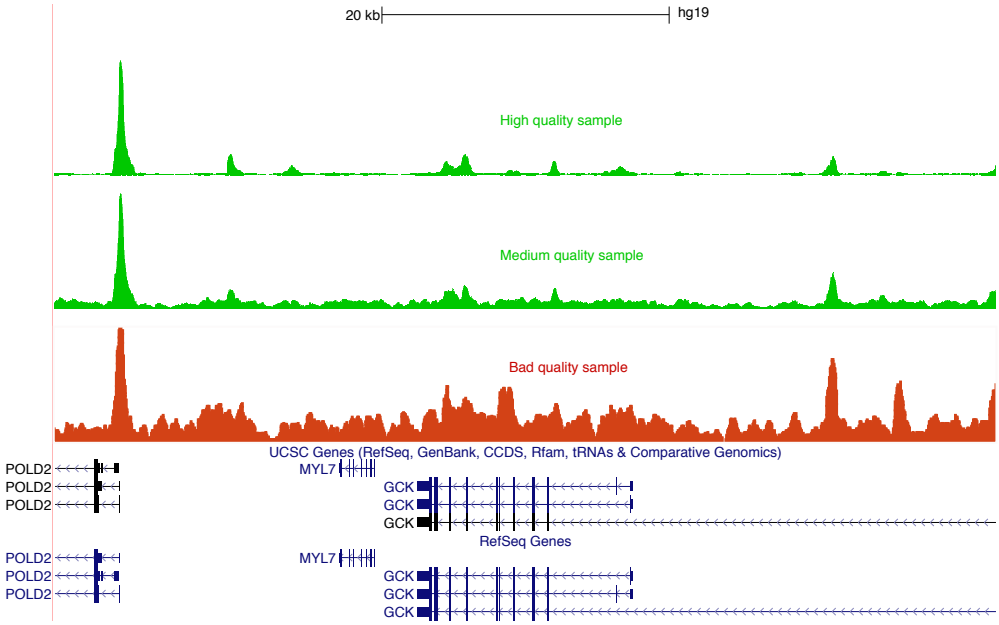
Figure 4.2: ATAC-seq read numbers per sample. (A) Distribution of the total (dark blue) and mapped (dark red) number of reads (y-axis) per sample in million. (B) Distribution of percentage (y-axis) of fully mapped reads before ("Mapped", green) and after ("Mapped_filtered", red) filtering by removing PCR duplicates compared to total number of reads per sample. In addition, the distribution of the percentage of reads belonging to the mitochondria genome before ("MT", blue) and after filtering by removing PCR duplicates ("MT_filtered", purple) reads compared to total number of all reads per sample.

enriched for ATAC-seq reads ("peaks") that are indicative of accessible DNA regulatory regions such as promoters and enhancers. Per sample between 5-148k peaks (median=71k) were identified at an FDR <1% (Fig4.3). Based on low peak number and low signal to noise ratio, one sample was removed from the analysis (Fig4.3). After filtering, overlapping sample peaks were merged to create a set of union islet peaks for the remaining analysis. This approach identified a total of 247.2k merged islet peaks present in at least one of the samples.

This merged dataset was then used to determine overlap with known promoter and enhancer regions from previously published islet regulatory studies. The large majority of ATAC-seq peaks overlapped either enhancer (59% of peaks) or promoter (15% of peaks) chromatin states obtained from[86] or[85]. Furthermore, the peaks were tested for enrichment in additional publicly available islet regulatory regions including islet eQTLs[159], genome-wide *CTCF* binding sites[160] and islet-relevant TFs (*FOXA2, MAFB, NKX2.2, NKX6.1* and *PDX1*)[85] of which the latter play an important role in both islet development and mature function. Consistent with the involvement of open chromatin regions in islet regulation, the merged peaks were found to be significantly enriched (Bonferroni corrected P <0.05) in islet eQTLs, CTCF binding sites and all tested islet TFBS (log2FE ranging from 3.3-3.9, Fig4.4) compared to randomised regions (see Chapter2 Section 2.6.2). Overall, these results show that ATAC-seq is able to accurately identify regulatory regions associated with high DNA accessibility. This is consistent with previous studies that found that the majority of DHS sites were located distal of the TSS and were strongly enriched for TF binding events[239,240]. Considering the high overlap with several other islet regulatory elements, Chapter 4 will describe the integration of the ATAC-seq data with various epigenomic annotations (including DNA methylation and ChIP-seq marks) to refine existing chromatin states.



(A) Peak number vs read number



(B) ATAC-seq read density distribution

Figure 4.3: Peak number and quality. (A) Scatter-plot showing read number (y-axis) and peak number (x-axis) for each sample. Samples with a very low number of peaks compared to all samples are removed from the analysis (red dot). The median peak number is indicated (black dashed line). (B) Genome-browser view of the ATAC-seq read density distribution (green and red) around the *GCK* locus (65kb region) in islets. One sample each of high quality (top, green), medium quality (middle, green) and bad quality (bottom, red) based on signal to noise ratio are shown. The bad quality sample was removed from the analysis which also corresponds to the failed sample with the low number of peaks.

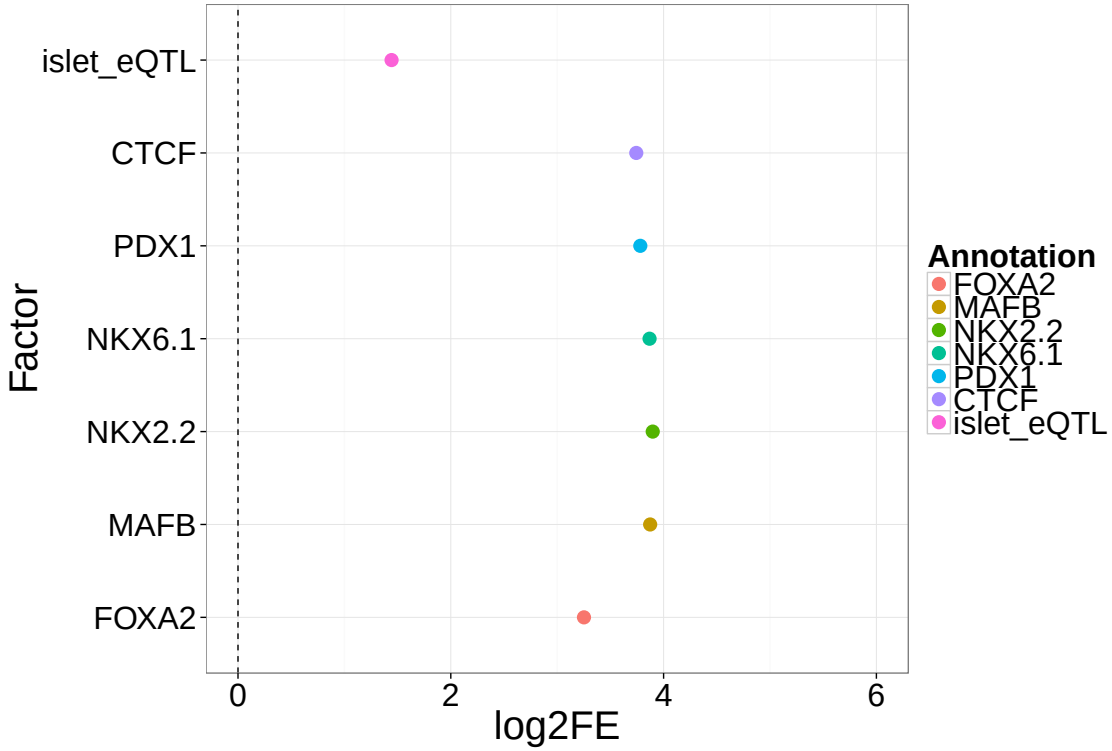


Figure 4.4: Enrichment of merged ATAC-seq open chromatin peaks in islet regulatory regions. Log2 Fold Enrichment (FE, x-axis) of merged ATAC-seq peaks in different genome annotations are shown (y-axis). From top to bottom these include islet eQTLs, genome-wide CTCF TFBS and islet relevant TFBS such as *PDX1, NKX6.1, NKX2.2, MAFB* and *FOXA2*. Enrichment is calculated compared to randomised regions created through shuffling the ATAC-seq peaks.

.

4.3.1.2 Determining the relationship between ATAC-seq islet open chromatin and T2D-associated GWAS regions

Since previous studies have found that distant highly accessible islet regulatory regions are enriched in GWAS signal for T2D-related traits[83,85,86], I next determined whether ATAC-seq open chromatin peaks show similar enrichment. The software FGWAS, which uses a linear hierarchical model (details in Chapter 2 section 2.6.1), was applied to estimate the enrichment of islet ATAC-seq open chromatin regions in T2D and FG GWAS signals. In agreement with results from former studies, all merged islet open chromatin regions were highly and significantly enriched in T2D- (log2FE=3.1, CI=2.4-3.8) and FG-associated (log2FE=3.4, CI=2.2-4.7) GWAS regions (Fig4.5, left panel).

To determine whether the enrichment was specific to T2D-related traits, FGWAS was applied to GWAS data from likely non-islet mediated traits such as Height and Total Cholesterol (TC). Islet open chromatin regions also showed no enrichment in Height (log2FE=2.7, CI=-0.96-0.82) and borderline significant enrichment in TC (log2FE=1.4, CI=0.1-2.2, Fig4.5, left panel) GWAS signal. To further dissect the contribution of these open chromatin peaks to the enrichment in GWAS traits, the regions were split into proximal (within 2000bp upstream and 500bp downstream of actively transcribed TSS in publicly available islet RNA-seq data[236]) or distal regions (non-proximal regions) and the enrichment analysis was repeated. Consistent with previous results, proximal promoter regulatory regions were not enriched in islet-mediated traits (T2D log2FE=2.7 CI=-20.0-4.2, FG log2FE=-6, CI=-50-3.3, Fig4.5, right panel) while distal islet regulatory elements, likely corresponding to enhancers, were strongly enriched (log2FE T2D=3.0 CI=2.2-3.6, FG=3.5 CI=2.3-4.7, Fig4.5, middle panel). In contrast to the T2D-related traits, distal islet elements did not show any significant enrichment in Height (log2FE=0.05 CI=-1.20-0.77) nor TC (log2FE=0.8 CI=-2.43-1.85) GWAS signal. Proximal islet regulatory regions were strongly enriched in TC (log2FE=2.5 CI=1.1-3.5) GWAS data while no significant enrichment was observed for Height (log2FE=1.05, CI=-20-2.19). The increase in enrichment for regions near the TSS in TC GWAS signal suggests that the enrichment is driven by promoter or gene regions, which is consistent with a former study that found TC enriched for coding regions[158]. ATAC-seq reads were also found to capture nucleosome structure and due to the nucleosomal patterns in the genome, fragment sizes of approximately

180bp, 360bp, 540bp are indicative of mono-, di- and tri-nucleosomes, respectively [123]. Since these nucleosomal regions may influence enrichment in GWAS signal, reads were filtered based on fragment size (reads with a fragment size >140bp were removed) for all samples and macs2 peak calling and the FGWAS enrichment analysis was repeated as before; however, no relevant differences in enrichment were detected for all samples combined (log2FE "All_reads"=3.11, CI=2.37-3.85, log2FE "Short_reads"=3.12, CI=2.36-3.79, Fig4.6). In contrast, when testing the peaks of 2 individual samples for enrichment in T2D GWAS signals, separately, without merging, shorter reads showed higher levels of enrichment (log2FE Sample_1=3.57, CI=2.71-4.26, log2FE Sample_2=3.65, CI=2.19-3.79) compared to all reads (log2FE Sample_1=3.57, log2FE Sample_2=3.65 Fig4.6). This indicates that nucleosomal reads in individual samples decrease GWAS enrichment because these do likely not correspond to distal regulatory elements which showed the strongest enrichment. However, after merging open chromatin peaks, no difference in GWAS enrichment between peaks derived from short- or all-fragment-sized reads could be detected which suggests that merging overlapping peaks across samples increases to some extent the noise in the signal.

Fig4.7 provides an example genome browser view of a peak overlapping the T2D-associated *CDC123* lead variant rs11257655 which is described in more detail in Chapter 5 Section 5.3.4 in the context of allelic imbalance in open chromatin.

Overall these data clearly show the importance of open chromatin regulatory regions in helping to understand GWAS data, in particular, indicating the role of islet regulatory regions in T2D-related traits. Additional analysis aiming to integrate different types of epigenomic data and to identify potential molecular mechanism at specific T2D-related GWAS hits is described in Chapter 5.

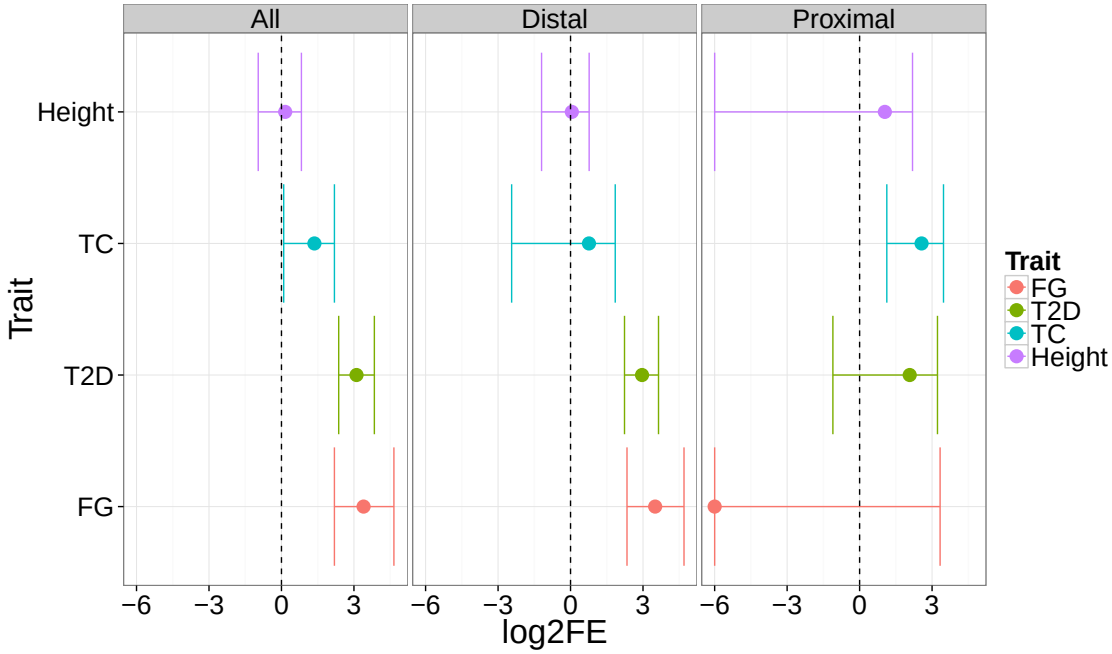


Figure 4.5: Enrichment of merged ATAC-seq open chromatin peaks in GWAS regions. Log2 Fold Enrichment (log2FE, x-axis) of merged ATAC-seq peaks in different GWAS traits (y-axis). Peaks are split into different categories dependent on location to active islet promoters based on publicly available RNA transcription levels. "All": all 247.1k ATAC-seq peaks independent of location, "Proximal": 15.4k of peaks within 2000bp upstream and 500 downstream of the Transcription Start Site (TSS); "Distal": 231.7k non-proximal peaks more distant from the TSS than 2000bp up and 500bp downstream. Abbreviations: Type 2 Diabetes (T2D), Body Mass Index (BMI), Fasting Glucose (FG). To improve visualisation the y-axis does not extend beyond -6.

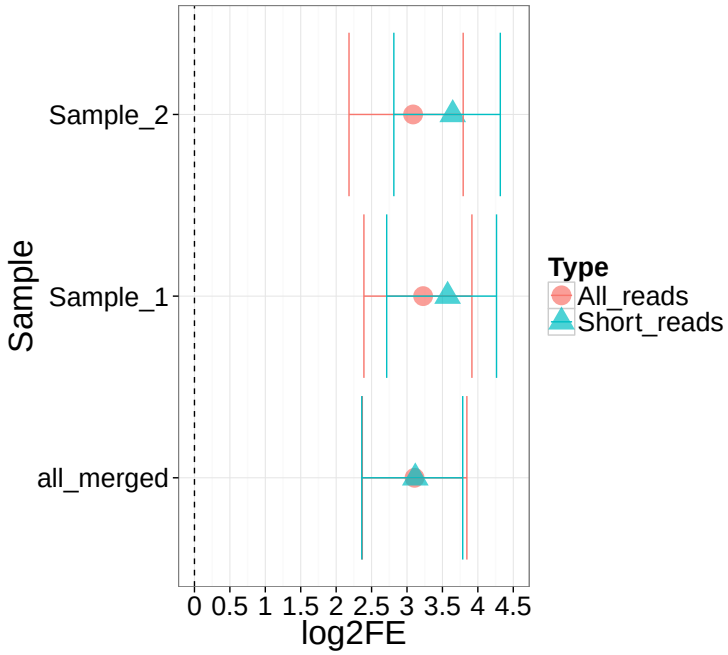


Figure 4.6: Enrichment of ATAC-seq peaks derived from all reads or nucleosome free short reads. Log2 Fold Enrichment (log2FE, x-axis) of ATAC-seq peaks across either all samples ("all_merged", peaks were merged across all samples) or 2 individual samples (y-axis) in T2D GWAS signal. Peaks are split into 2 different categories dependent on the filtering of reads before peak calling: "All_reads" (red circle): contains all reads including nucleosomal-associated reads with longer fragment size. "Short_reads" (blue triangle): excludes reads with a fragment size above 140bp to remove any nucleosomal-associated reads.

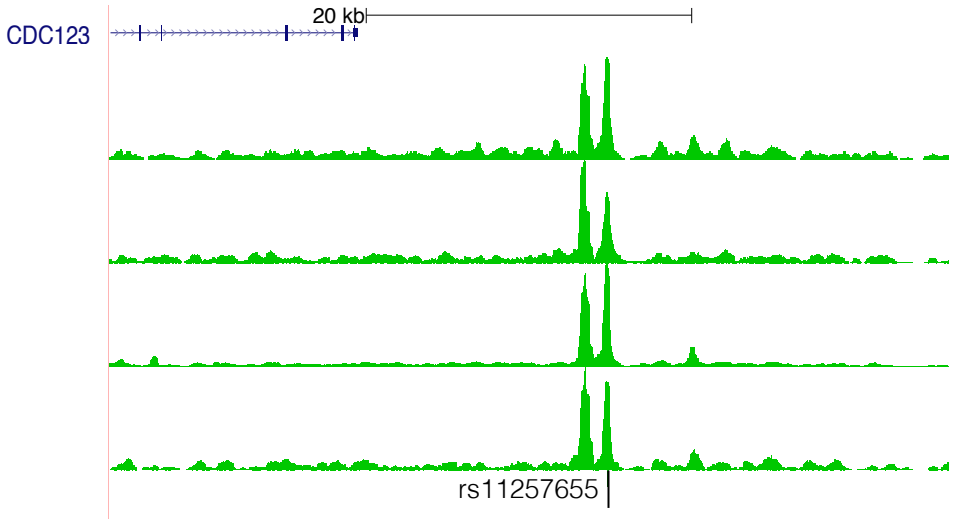


Figure 4.7: Peak overlapping T2D-associated *CDC123* GWAS lead variant rs11257655. GWAS lead variant rs11257655 at the *CDC123* locus (RefSeq gene shown in blue) overlaps an islet open chromatin peak in representative samples (green, shown is the averaged ATC-seq read density over 300bp windows).

4.3.2 Comparing the epigenomic landscape of beta-cell and fibroblast derived iPSCs

As described in the introduction of this Chapter, previous work from a collaborator (Prof Shimon Efrat) has shown that BiPSCs retained more open chromatin at islet relevant genes compared to FiPSCs which was associated with an enhanced ability to differentiate into insulin producing cells[224]. The retention of epigenetic marks related to the original cell type ("epigenetic memory") is one potential explanation for the enhanced differentiation ability of BiPSC.

This part of the chapter used ATAC-seq to characterise the epigenomic differences of these two iPSC cell lines and contrast their open chromatin profile with islet ATAC-seq data. Since several factors such as donor genotype[221–223] or differences in passage number[225] between BiPSCs (mean passage number=11) and FiPSCs (mean passage number= 23, Table4.1A) may influence the epigenomic landscape of these iPSCs, it remains unclear whether epigenetic memory is the primary driver of the observed epigenomic differences. Nevertheless, characterising these epigenomic differences between iPSC subtypes, which differ in their differentiation potential, can still provide relevant biological clues about improving "iPSC to endoderm" *in vitro* differentiation protocols.

4.3.2.1 Generation and QC of BiPSC and FiPCS

Through a collaboration, five FiPSC lines from two independent non-diabetic donors and five BiPSC lines from three independent non-diabetic donors were generated, assessed for pluripotency and quality controlled by Liraz Shenhav at Tel-Aviv University (Table4.1A). Specifically, the work of Liraz Shenhav showed the following:

- all lines were fully pluripotent based on the expression of key pluripotent genes (*NANOG, OCT4, TRA-1-60, REX1, DNMT3B* and *SOX2*) and the capacity to spontaneously differentiate into all three germ layers;
- BiPSCs had an enhanced ability to differentiate into the endodermal lineage *in vitro* compared to FiPSCs while no signs of lineage-specific commitment were observed for FiPSCs which is consistent with previous data[224];

- based on G-banding or an Illumina array all lines were karyotypically normal with the exception of lines BiPSC-D1 and D2 which both had 48 chromosomes (Table4.1A, the subsequent analysis will show that the effect of the abnormal karyotype was minimal).

Tissue of origin	Donor ID	Sample ID	Passage Number	Karyotype
Fibroblast derived	Individual A	FiPSC_A.1	25	46XX
Fibroblast derived	Individual A	FiPSC_A.2	24	46XX
Fibroblast derived	Individual A	FiPSC_A.3	26	46XX
Fibroblast derived	Individual B	FiPSC_B.1	21	46XX
Fibroblast derived	Individual B	FiPSC_B.2	18	46XX
Beta-cell derived	Individual C	BiPSC_C.1	12	46XX
Beta-cell derived	Individual C	BiPSC_C.2	10	46XX
Beta-cell derived	Individual D	BiPSC_D.1	10	48XX+20
Beta-cell derived	Individual D	BiPSC_D.2	10	48XX+12+20+add22
Beta-cell derived	Individual E	BiPSC_E.1	11	46XX

(A) iPSC sample characteristics

Sample	Purity
ISL_1	80%
ISL_2	70%
ISL_3	50%
ISL_4	50%
ISL_5	60%

(B) Islet characteristics

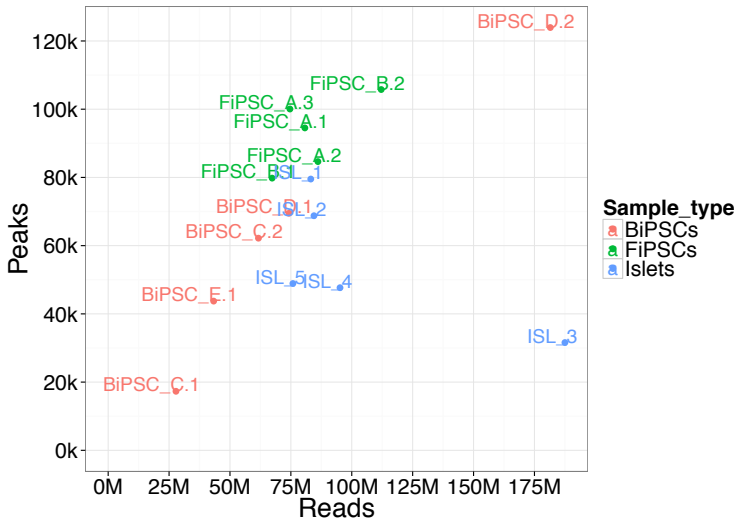
Table 4.1: iPSC and islet sample characteristics (A) Characteristics of iPSC samples including tissue of origin, donor individual, sample ID (which indicates donor), Passage Number and Karyotype. (B) Islet samples used for comparison with iPSCs with purity estimates.

4.3.2.2 Mapping the open chromatin landscape of BiPSCs and FiPSCs

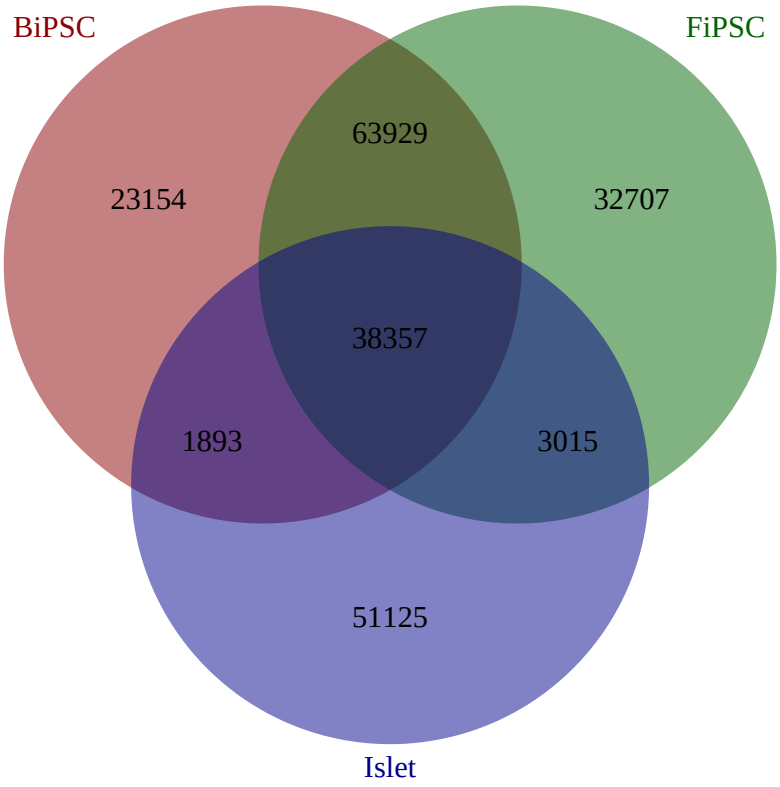
ATAC-seq was performed on these iPSCs lines to map the epigenomic landscape in both BiPSCs and FiPSCs and to identify DOCS between these iPSC lines that may explain the differences in their lineage commitment. In addition, the iPSC ATAC-seq data was compared to open chromatin ATAC-seq information from 5 human pancreatic islet samples, of which a large fraction represents beta-cells, to identify shared open chromatin characteristics between BiPSC and islets. For each iPSC sample, 60-334 million reads (median=151 million) were obtained of which 28-180 million mapped and filtered reads (median mapped and filtered read number=74 million) remained after processing. Using the software MACS2, between 17.3k-123.9k open chromatin peaks per sample at an FDR of 1% were identified (Fig4.8A). For the human islet samples, which represent a mix of 5 islets with varying read depth and peak number, 76-186 million mapped and filtered reads were obtained per sample and 32-80k peaks (FDR <1%) were identified by MACS2 (Fig4.8A). Subsequently, peaks were merged according to cell type which identified 127.3k BiPSC, 138.8k FiPSC and 94.4k islet merged peaks. A large fraction of merged peaks (80% of BiPSC and 75% of FiPSC peaks) was found in both iPSC cell types while a lower number was shared with human islets (30% of BiPSC and 29% of FiPSC peaks, (Fig4.8B).

Similarly, a significant and high correlation (median Spearman's rho=0.85) of normalised ATAC-seq read depth across all merged open chromatin peak regions (irrespective of donor cell type) was observed for all iPSC lines. In contrast, the correlation between iPSCs subtypes and human islets was considerably lower (median Spearman's rho=0.11 compared to within islet correlation with a median Spearman's rho=0.59, Fig4.9A). Principal component analysis (PCA) showed that the iPSC subtypes did not cluster according to donor genotype which was suggested previously to largely influence the transcriptomic and/or epigenomic profile of iPSCs[221–223]. Additionally, both FiPSC and BiPSC samples clustered separately from primary human pancreatic islet and based on the variance explained by PC1 and PC2 were more similar to each other compared to primary tissue (Fig4.9B).

These results confirm that both FiPSCs and BiPSCs largely share epigenomic features at the chromatin level that clearly distinguish them from primary human islets.

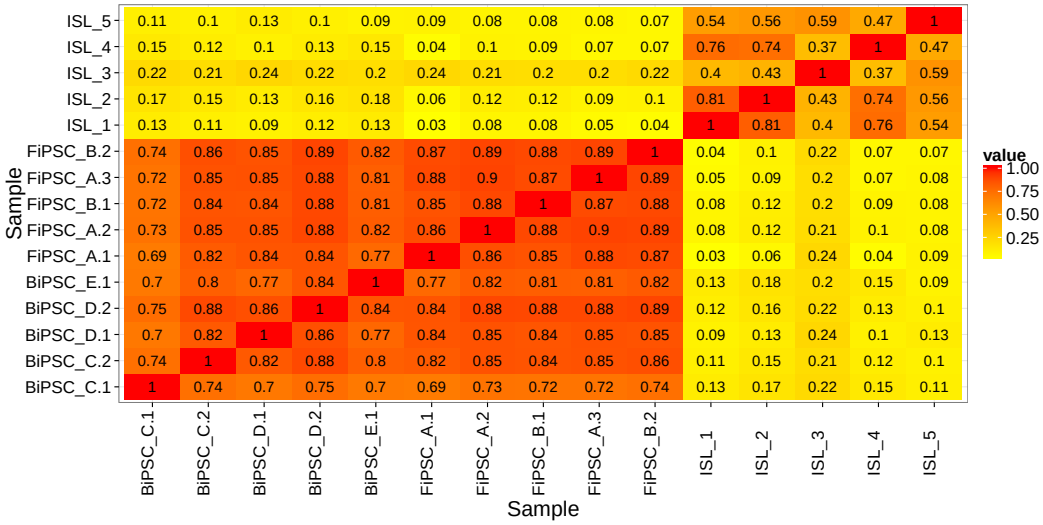


(A) Read and peak number per sample

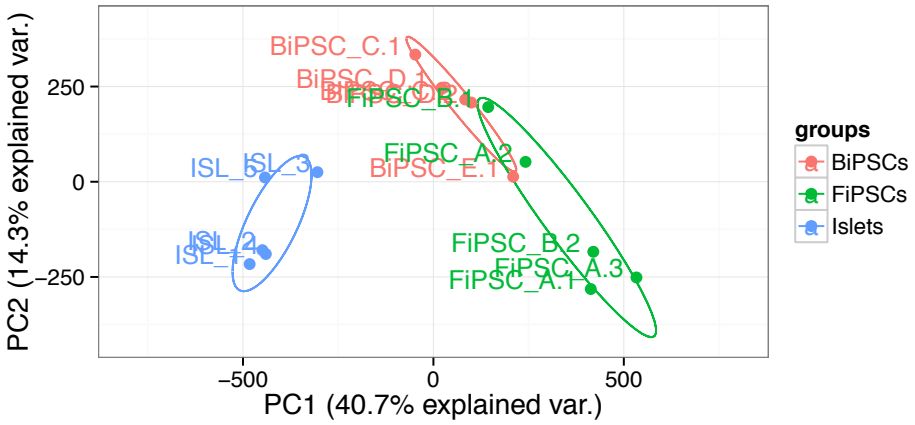


(B) Merged peak overlap

Figure 4.8: iPSC and selected islet samples read and peak number. (A) Peak number (y-axis) and read number (x-axis) for BiPSCs (red) , FiPSC (green) and selected islet samples (blue). (B) Venn Diagram indicating overlap of peaks merged according to cell subtype (BiPSCs: red, FiPSC: green and Islets: blue).



(A) Correlation read depth across peaks

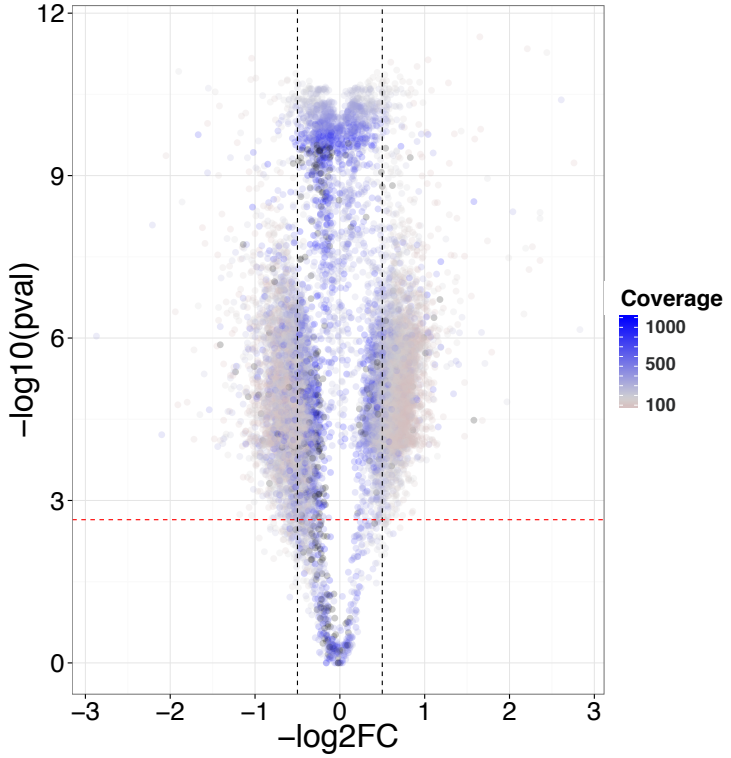


(B) PCA of read depth across peaks

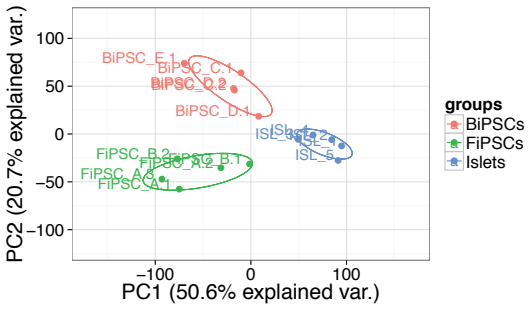
Figure 4.9: Correlation and PCA analysis across all merged peak regions. (A) Spearman's rho correlation heatmap of read depth across all merged open chromatin peaks for all iPSC and islet samples with the colour indicating the correlation level. (B). Principal component analysis (PCA) was performed on read depth across all peak regions for each sample and samples are plotted along PC1 (x-axis) and PC2 (y-axis). Colours indicate cell type: (BiPSCs: red, FiPSC: green and Islets: blue).

4.3.2.3 Identification of differential open chromatin regions (DOCS) between BiPSCs and FiPSC

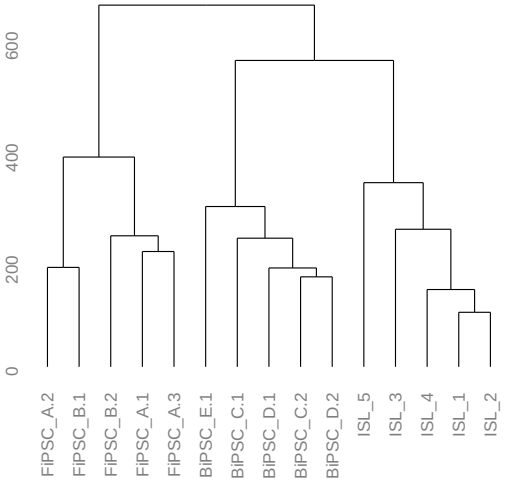
Despite the similarities of the iPSC subtypes, DOCS between BiPSCs and FiPSC could be detected using a genome-wide sliding window approach and negative binomial test as implemented in diffReps. In total, 8.3k significant DOCS (FDR< 5%) with a minimum absolute log2 fold change (log2FC) of 0.5 were identified. These include 4.8k (58%) sites that were characterised by increased open chromatin in BiPSCs compared to FiPSC (defined as Bi-DOCS) while the remaining 3.5k sites were more open in FiPSCs (defined as Fi-DOCS, Fig4.10A). As expected PCA and hierarchical clustering of normalised ATAC-seq read depth showed that DOCS clearly separate FiPSCs from BiPSCs. Consistent with the increased endodermal differentiation potential of BiPSC, across DOCS, BiPSCs were more similar to human islets than to FiPSCs (Fig4.10B and 4.10C). In addition, samples were not found to cluster according to donor genotype suggesting that genetic variation across samples is likely not the main factor driving the differences in open chromatin in this analysis.



(A) Volcano Plot of differential open chromatin analysis



(B) PCA of read depth across DOCS



(C) Dendrogram of read depth across DOCS

Figure 4.10: Identification of DOCS between BiPSC and FiPSC. (A) Volcano plot of differential open chromatin analysis between BiPSC and FiPSC. Log2FC are on the x-axis (positive indicates more open in BiPSCs compared to FiPSCs) and -log10 P-value on the y-axis. Horizontal red line indicates FDR of 0.05 while vertical black line indicate log2FC threshold of +/- 0.50 (B). Principal component analysis (PCA) was performed on read depth across all significant 8.3k DOCS for each sample and samples are plotted along PC1 (x-axis) and PC2 (y-axis). Colours indicate cell type: (BiPSCs: red, FiPSC: green and Islets: blue). (C) Dendrogram showing hierarchical clustering of samples based on read depth across DOCS.

4.3.2.4 DOCS are enriched for chromatin states involved in the regulation of development and stem cell function

To further understand the role of DOCS within the genome of the two iPSC subtypes, 18 predefined chromatin states from 98 Epigenome Roadmap cell types were obtained (website `http://egg2.wustl.edu/roadmap/web_portal/chr_state_learning.html#exp_18state`) provides more information about the relevant states). These states were derived from Chromatin Immunoprecipitation (ChIP) sequencing marks by the Epigenome Roadmap consortium in a systematic manner and are indicative of regulatory elements in the genome[76,82]. After grouping Epigenome Roadmap samples according to type of germ layer (for mature cells and tissues) and type of pluripotent cell (for iPSCs and ESCs), the chromatin states from each Epigenome Roadmap sample were overlapped with DOCS and the total log2FC for each chromatin state was estimated separately. Comparing these estimates to the total log2FC of randomised regions via permutation showed that, across all germ layers and stem cell types, DOCS were significantly (mean FDR adjusted P-value across all cell types of a given chromatin and germ layer/stem cell state <5%) enriched in regulatory chromatin states (highlighted in bold in Fig4.11A). These enriched states include active enhancers, active and bivalent promoters, weak and bivalent enhancers, flanking TSS sites, polycomb-repressed regions and one type of a gene enhancer (state "gene enhancer 2" which is characterised by H3K27ac, H3K4me3 and H3K36me3 chromatin marks). Analysing the magnitude of the significantly enriched states showed that the following ESC and iPSC states had the highest level of enrichment compared to the other cell types/germ layers (states are bold and underlined in Fig4.11A):

- active (ESC/iPSC log2FE=2.7-3.5) and gene enhancer (type 2 ESC/iPSC log2FE=2.3-2.7)
- bivalent (ESC/iPSC log2FE=2.9-3) and weak (ESC/iPSC log2FE=2.2-2.5) enhancer
- Flanking Transcription Start Sites (TSS, ESC/iPSC log2FE=2.1-2.8)
- active promoter (ESC/iPSC log2FE=2.0-2.1) and
- weak polycomb-repressed region (ESC/iPSC log2FE~1.2)

These results suggest that the identified DOCS may have an important regulatory function and thus, may help understand the differences in differentiation capability between FiPSCs and BiPSC.

To determine whether Bi-DOCS and Fi-DOCS have different regulatory roles, the analysis was specifically focused on significantly enriched (highlighted in bold in the previous Fig4.11A) chromatin states that showed differential enrichment between Bi-DOCS and Fi-DOCS. Across all tissues it was observed that Bi-DOCS (compared to Fi-DOCS) were most enriched (log2 Ratio of FE BI-DOCS VS Fi-DOCS >0) for the following states (highlighted in bold in Fig4.11B):

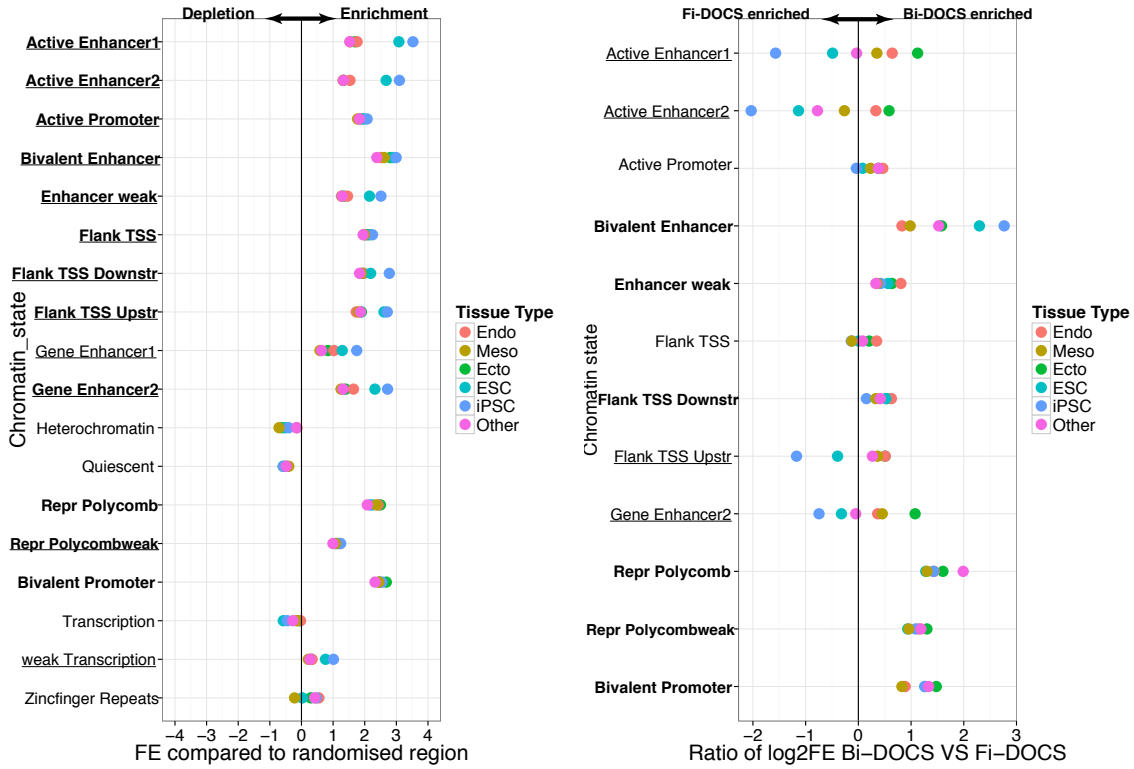
- bivalent enhancers (log2 Ratio FE=0.82-2.8)
- weak enhancers (log2 Ratio FE=0.3-0.8)
- Flanking TSS downstream (log2 Ratio FE=0.2-0.62)
- polycomb-repressed regions (log2 Ratio log2FE=1.0-2.0)
- bivalent promoters (log2 Ratio FE=0.8-1.5)

iPSC and hESC bivalent enhancers (ESC/iPSC log2 Ratio FE=2.3-2.8) showed the strongest differential enrichment (Bi-DOCS vs Fi-DOCS) compared to bivalent enhancers in any of the 3 germ layers (none ESC/iPSC log2 Ratio FE =0.8-1.6) indicating that these iPSC and hESC bivalent enhancers were highly accessible in BiPSCs.

None of the significantly enriched all DOCS states (highlighted in bold in the previous Fig4.11A) showed enrichment for Fi-DOCS across all tissue types. However, Fi-DOCS showed high enrichment (indicated by a log2 Ratio of Bi-DOCS VS Fi-DOCS FE < 0) in hESC and iPSC active and gene enhancer chromatin states (ESC/iPSC Ratio FE=-0.5 to -2.0) compared to Bi-DOCS (all Fi-DOCS hESC/iPSC enriched states are underlined in Fig4.11B).

Bivalent promoters, bivalent enhancers, polycomb-repressed regions and weak enhancers were previously shown to regulate developmental competence and differentiation, in particular for endoderm development and the endocrine pancreas lineage[216,217,241]. Thus, the enrichment of these regions in Bi-DOCS suggests that BiPSCs may be more primed for developmental differentiation

compared to FiPSCs. Similarly, the higher enrichment of Fi-DOCS in hESC and iPSC active enhancer regions indicates more stem-cell-like behaviour for FiPSC which may be explained by the extended passaging of FiPCS (compared to BiPSCs, see Table4.1A). In agreement with the observed highly stem-cell-like state of FiPSCs, higher passage number was previously found to be associated with increased stem-cell-like features and reduced levels of epigenetic memory[225].



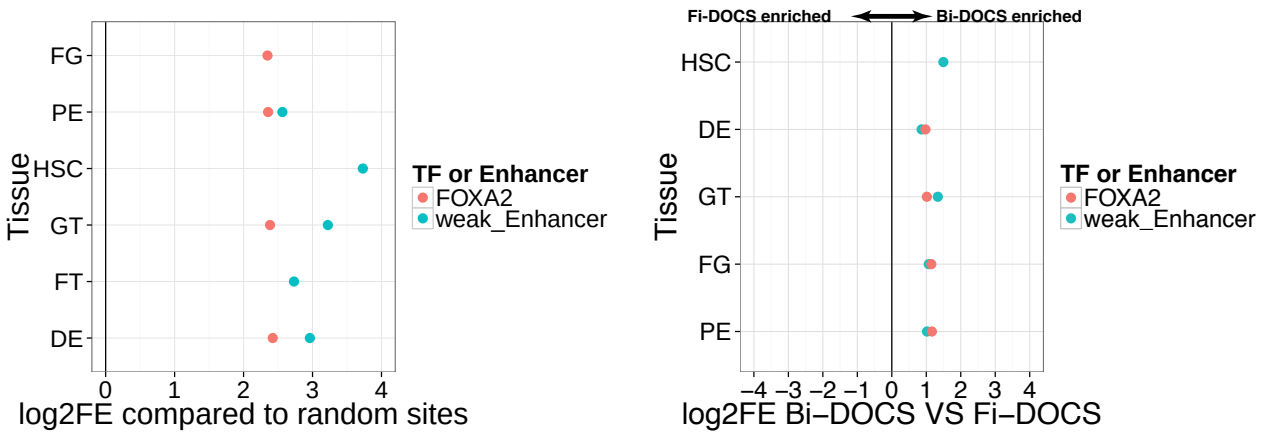
(A) All DOCS enriched in chromatin states

(B) Bi-DOCS VS Fi-DOCS enrichment

Figure 4.11: Enrichment of DOCS in Epigenome Roadmap chromatin states. (A) Log2FE of all DOCS (x-axis) in chromatin state regions (y-axis) derived from Epigenome Roadmap. States highlighted in bold indicate states that were enriched across all germ layer or pluripotent stem cell states (indicated by the colour). Underlined states indicate that ESC/iPSCs showed highest enrichment compared to other germ layers. (B). log2 Ratio of FE (x-axis) of Bi-DOCS VS Fi-DOCS across chromatin states (y-axis) enriched in plot (A). log2 Ratio of FE >0 indicates Bi-DOCS enrichment while log2 Ratio of FE <0 indicates Fi-DOCS enrichment. Bold states are enriched for Bi-DOCS while states underlined are enriched for Fi-DOCS in hESC and iPSC lines compared to any other germ layer/cell type.

4.3.2.5 Enrichment of Bi-DOCS in regions involved in endodermal development

To further define the regulatory role of Bi-DOCS, the overlap with regions known to be involved in endodermal development was determined. Previously generated and publicly available ChIP-seq weak enhancer states (based on H3K4me1 histone mark) and FOXA2 TFBS[217] were obtained for several stages of an experiment differentiating hESC into pancreatic endoderm. Specifically, these stages included HSC (Human Stem Cell, weak enhancers only), Definitive Endoderm (DE), Gut Tube (GT), Fore Gut (FG) and Pancreatic Endoderm (PE). FOXA family TFs are important for endodermal development including the pancreas and represent a group of pioneer factors that prime regulatory regions and facilitate other TFs to bind. DOCS were overlapped with these regulatory states and found to be significantly enriched (Bonferroni-corrected P <0.05) in both weak enhancer (log2FE for different stages ranged from 2.6-3.7) and FOXA2 TFBS (log2FE for different stages ranged from 2.3-2.4) across all developmental stages compared to randomised sites (Fig4.12A). Consistent with the hESC and iPSC Epigenome Roadmap results, it was observed that Bi-DOCS were significantly more enriched in endodermal weak enhancer states (log2 Ratio Bi-DOCS VS Fi-DOCS (log2 Ratio FE)=0.8-1.5) and FOXA2 TFBS (log2 Ratio FE=0.9-1.2) compared to Fi-DOCS (Fig4.12B).



(A) DOCS enrichment compared to random sites

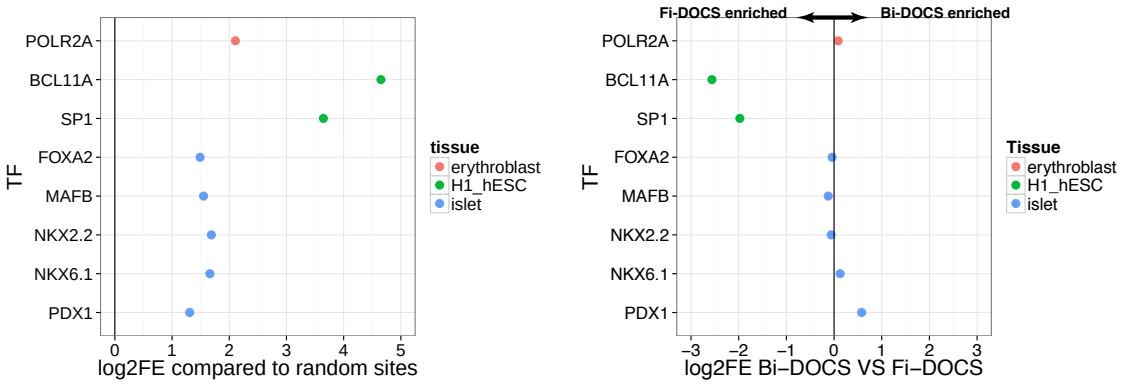
(B) Bi-DOCS VS Fi-DOCS enrichment

Figure 4.12: Enrichment of DOCS in weak enhancer and FOXA2 binding sites identified during endodermal development. (A) Log2FE of DOCS (x-axis) in weak enhancers (blue) and FOXA2 TFBS (red) across stages of endodermal development (y-axis). (B). log2 Ratio of FE (x-axis) of Bi-DOCS VS Fi-DOCS across weak enhancers and FOXA2 TFBS identified during different stages (y-axis) of endodermal development. log2 Ratio of FE >0 indicates Bi-DOCS enrichment while log2 Ratio of FE <0 indicates Fi-DOCS enrichment. HSC (Human Stem cell for weak enhancer only), Definitive Endoderm (DE), Gut Tube (GT), Fore Gut (FG) and Pancreatic Endoderm (PE).

4.3.2.6 Bi-DOCS enrichment is specific to endodermal developmental regions

To determine whether the enrichment of Bi-DOCS is specific to endodermal regulatory annotations, erythroblast (*POLR2A*) and hESC H1 TFBS (*BCL11A* and *SP1*) from the ENCODE project were obtained. In addition, publicly available TFBS that are active in mature human islets (*FOXA2, MAFB, NKX2.2, NKX6.1* and *PDX1*) were also used for the analysis to determine whether the effect is restricted to endodermal developmental regions. DOCS were overlapped with these TFBS from three different tissues to evaluate the enrichment of DOCS in these sites (Fig4.13). Significant enrichment (Bonferroni-corrected P<0.05) of DOCS across TFBS of all tissues (log2FE=1.3-4.65) was observed consistent with the fact that the DNA at these TFBS is, in general, more accessible compared to random regions (Fig4.13A). However, only TFBS in hESC (*BCL11A* and *SP1*) were differentially enriched between Bi-DOCS and Fi-DOCS with Bi-DOCS showing depletion in hESC TFBS (*BCL11A* log2FE Ratio= -2.6 and *SP1* log2FE Ratio= -2.0). The enrichment of Fi-DOCS in hESC TF is consistent with the enrichment of Fi-DOCS in hESC and iPSC Epigenome Roadmap active enhancer chromatin states described in the previous section. In contrast, no strong difference in enrichment between Bi-DOCS and Fi-DOCS could be found in human islet and erythrocyte TFBS (log2FE=-0.04 to 0.5, Fig4.13B).

These data indicate that Bi-DOCS were specifically enriched in regulatory regions associated with endodermal lineage commitment, such as endodermal weak enhancers and FOXA2 binding sites, whereas no Bi-DOCS specific enrichment was found in mature endodermal islets. Furthermore, Fi-DOCS showed depletion in stem cell active enhancer regions. These observations are in concordance with the aforementioned results that showed BiPSCs preferentially differentiate into endoderm during spontaneous differentiation *in-vitro* while Fi-DOCS show more stem cell-like characteristics.



(A) DOCS enrichment compared to random sites

(B) Bi-DOCS VS Fi-DOCS enrichment

Figure 4.13: Enrichment of DOCS in islet and non-islet TFs. (A) Log2FE of DOCS (x-axis) in erythroblast (red), hESC-H1 (green) and islet (blue) TFBS (y-axis) compared to random sites. (B). log2 Ratio of FE (x-axis) of Bi-DOCS VS Fi-DOCS across erythroblast (red), hESC-H1 (green) and islet (blue) TFBS (y-axis). log2 Ratio of FE >0 indicates Bi-DOCS enrichment while log2 Ratio of FE <0 indicates Fi-DOCS enrichment.

4.3.2.7 Gene and Pathway analysis for DOCS

To understand which genes and pathways may be influenced by DOCS, a pathway enrichment test was performed using GREAT[238]. Bi-DOCS (4.8k sites) with more accessible chromatin in BiPSCs were significantly enriched in 27 Gene Ontology (GO) Biological Process terms and 2 Molecular Signatures DataBase (MSigDB) Pathway terms (GREAT Binomial and Hypergeometric FDR <0.05 and binomial region fold enrichment and minimum gene enrichment >2, Table4.2 and Appendix Table1 for more information on associated genes). These include diabetes-relevant terms like 'Maturity-Onset Diabetes of the Young (MODY)', capturing forms of monogenic diabetes caused primarily by genes encoding TFs involved in endocrine pancreas development. Signalling terms relating to WNT and NOTCH pathways were also enriched. These pathways are critical for the induction of posterior endoderm and pancreatic progenitors, and pancreatic endocrine versus exocrine fate choices[242,243]. There was also enrichment of multiple terms relating to neuronal development which is consistent with shared characteristics between pancreatic endodermal and neuronal ectodermal tissue. Although these tissues are derived from different germ layers, previous studies have shown that several developmentally active TFs are expressed in both tissues and that the transcriptional profile as well as active chromatin marks between these cell types are similar[244,245]. Additionally, beta and neuronal cells also share features on the physiological level with both cell types being electrically excitable and able to respond to external stimuli with exocytosis[246].

GREAT enrichment analysis following stratification of Bi-DOCS according to chromatin state (using Bi-DOCS overlapping ESC bivalent promoters and enhancers, separately as foreground and all ESC chromatin states of the same type as background set) revealed similar significantly enriched terms (Hypergeometric FDR <0.05 and enrichment >2 compared to background) to those identified in the 'all DOCS' analysis (full list of associated terms shown in Appendix 2 while in the main text only terms related to pancreas/islet function are shown). Notably, bivalent enhancer (Table4.3) and promoter Bi-DOCS (Table4.4) were enriched for terms relating to glucose-sensing and insulin metabolism and secretion, as well as those relating to endocrine pancreas fate decisions and beta-cell development. These Bi-DOCS were linked to the aforementioned developmental regulators *FOXA2* as well as *NKX6.1, NKX2.2, PXD1* and *PAX6*. Full loss of *PDX1* is associated with pancreatic agenesis while partial loss leads to beta-cell dysfunction, beta-cell death and dia-

betes[247]. Similarly, both *NKX6.1* and *NKX2.2* were previously shown to be critical for beta-cell development as well as mature beta cell function[248,249]. *PAX6* deficiency is associated with pancreas and islet dysfunction during development and in the mature state. *Pax6* null mice suffer from reduced pancreatic endocrine cells while conditional knockout of *Pax6* is associated with loss of endocrine hormone expression and diabetes[250]. These findings suggest that Bi-DOCS are poised for appropriate developmental signalling cues facilitating a faster induction of genes involved in endocrine pancreas differentiation and mature function compared with FiPSCs.

Overall these data show that DOCS, which are more open in BiPSCs versus FiPSCs, are enriched for pathways with known roles in pancreatic endoderm induction, as well as those regulating more refined aspects of islet development and function.

Term	Description	Dir	RE	GE	BBH-P	HBH-P	Type
Go	neural tube patterning	Bi	3.3	2.7	8.7e-17	2.4e-05	Neurodevelopmental
Go	positive regulation of embryonic development	Bi	4.8	3.2	1.0e-16	9.3e-03	Developmental
Go	non-canonical Wnt receptor signaling pathway	Bi	3.5	2.1	2.4e-15	3.9e-02	WNT/NOTCH signalling
Go	hindbrain morphogenesis	Bi	2.8	2.3	4.1e-14	4.1e-04	Neurodevelopmental
Go	branching involved in mammary gland duct morphogenesis	Bi	3.2	2.5	5.2e-14	3.0e-03	Developmental
Go	cerebellum morphogenesis	Bi	2.9	2.4	4.9e-13	6.8e-04	Neurodevelopmental
Go	mammary gland duct morphogenesis	Bi	2.4	2.5	6.2e-12	1.8e-04	Developmental
Go	mammary gland epithelium development	Bi	2.1	2.1	9.1e-12	1.4e-04	Developmental
Go	dorsal spinal cord development	Bi	3.4	2.6	1.4e-11	3.8e-03	Neurodevelopmental

Go	cerebellum development	Bi	2.1	2.0	1.4e-10	1.1e-04	Neurodevelopmental
Go	cerebellar cortex morphogenesis	Bi	2.9	2.3	1.9e-10	6.6e-03	Neurodevelopmental
Go	spinal cord association neuron differentiation	Bi	3.8	2.7	5.6e-10	1.6e-02	Neurodevelopmental
Go	dorsal/ventral neural tube patterning	Bi	3.0	2.8	2.0e-09	9.5e-04	Neurodevelopmental
Go	negative regulation of Notch signaling pathway	Bi	3.7	2.5	1.1e-08	1.0e-02	WNT/NOTCH signalling
Go	cell differentiation in spinal cord	Bi	2.1	2.8	1.1e-08	2.6e-09	Neurodevelopmental
Go	cell differentiation in hindbrain	Bi	2.6	2.6	2.4e-08	1.6e-03	Neurodevelopmental
Go	cerebellar cortex development	Bi	2.4	2.0	2.8e-08	8.8e-03	Neurodevelopmental
Go	somitogenesis	Bi	2.0	2.0	4.6e-08	4.1e-04	Developmental
Go	white fat cell differentiation	Bi	3.4	2.7	1.3e-06	4.5e-02	Developmental
Go	embryonic axis specification	Bi	2.1	2.0	2.0e-06	1.9e-02	Developmental
Go	ventral spinal cord interneuron fate commitment	Bi	2.9	3.7	2.8e-06	3.1e-04	Neurodevelopmental
Go	regulation of insulin receptor signaling pathway	Bi	2.4	2.0	5.0e-06	3.3e-02	Diabetes/Insulin
Go	positive regulation of focal adhesion assembly	Bi	3.5	2.8	5.7e-06	1.9e-02	Other
Go	cerebellar cortex formation	Bi	2.5	2.4	6.4e-06	1.8e-02	Neurodevelopmental
Go	regulation of Notch signaling pathway	Bi	2.1	2.2	1.4e-05	1.0e-03	WNT/NOTCH signalling
Go	spinal cord dorsal/ventral patterning	Bi	2.3	2.5	4.3e-05	1.0e-02	Neurodevelopmental
Go	ventral spinal cord interneuron differentiation	Bi	2.5	3.4	5.0e-05	4.1e-04	Neurodevelopmental

MSigDB	Wnt signaling network	Bi	2.7	2.3	9.4e-08	2.8e-02	WNT/NOTCH signalling
MSigDB	Maturity onset diabetes of the young	Bi	2.1	2.4	3.0e-04	2.5e-02	Diabetes/Insulin
Go	dendrite morphogenesis	Fi	2.2	2.2	6.4e-04	4.0e-02	Neurodevelopmental
Go	long-term memory	Fi	2.2	2.6	4.6e-03	1.4e-02	Other
Go	positive regulation of acute inflammatory response	Fi	2.6	3.0	9.7e-03	6.1e-03	Other

Table 4.2: Enrichment terms associated with DOCS. Each row represents an enriched term associated with DOCS. Columns indicate Term (Go Biological Process or MSigDb term), Description, Direction (Dir with Bi indicating BiDOCs are associated while Fi indicates FiDOCs are associated with this term), Regional enrichment (RE), Gene Enrichment (GE), Binomial BH/FDR-adjusted P (BBH-P, for regional enrichment), Hypergeometric BH/FDR-adjusted P (HBH-P, for gene enrichment) and Type categorising the enriched term into broad biological classes.

Term	Description	State	FE	HBH-P	Gene
Go	regulation of transcription from RNA polymerase II promoter involved in detection of glucose	BivEnh	6.2	0.002	FOXA2
Go	negative regulation of transcription by glucose	BivEnh	6.2	0.002	FOXA2
Go	negative regulation of detection of glucose	BivEnh	6.2	0.002	FOXA2
Go	positive regulation of transcription from RNA polymerase II promoter by glucose	BivEnh	6.2	0.002	FOXA2
Go	negative regulation of transcription from RNA polymerase II promoter by glucose	BivEnh	6.2	0.002	FOXA2
Go	regulation of transcription by glucose	BivEnh	6.1	0.002	FOXA2
Go	detection of glucose	BivEnh	4.3	0.002	FOXA2, NKX6-1, PDX1
Go	regulation of insulin secretion involved in cellular response to glucose stimulus	BivEnh	2.9	0.03	ADRA2A, FOXA2, PIM3, TCF7L2
Go	insulin catabolic process	BivEnh	30.3	0.04	IDE
Go	CD8-positive, alpha-beta T cell differentiation	BivEnh	3.7	0.04	PAX1
MSigDB	Genes involved in Regulation of gene expression in beta cells	BivEnh	3.8	0.001	FOXA2, NKX2-2, NKX6-1, PAX6, PDX1, SLC2A2

MSigDB	Maturity onset diabetes of the young	BivEnh	2.6	0.03	FOXA2, HES1, NKX2-2, NKX6-1, PAX6, PDX1, SLC2A2
MSigDB	Genes involved in Regulation of beta-cell development	BivEnh	2.3	0.05	FOXA2, HES1, NKX2-2, NKX6-1, ONECUT3, PAX6, PDX1, SLC2A2

Table 4.3: Pancreas-related enrichment terms associated with bivalent enhancer BIDOCS. Each row represents an enriched term associated with BiDOCS overlapping ESC bivalent enhancer compared to all ESC bivalent enhancers. Columns indicate Term (Go Biological Process), Description, Chromatin State (State), Regional enrichment (RE), Hypergeometric BH/FDR-adjusted P (HBH-P, for gene enrichment) and Genes associated with the term.

Term	Description	State	FE	HBH-P	Gene
Go	detection of glucose	BivPro	6.7	0.002	FOXA2, NKX6-1
Go	pancreas development	BivPro	2.5	0.003	ALDH1A2, FOXA2, HES1, IGF1R, NKX2-2, NKX3-2, NKX6-1, NKX6-2, PAX2, PAX6, SHH, SLC2A2, SMAD2, SOX9, ZIC3
Go	negative regulation of detection of glucose	BivPro	10.2	0.003	FOXA2
Go	regulation of transcription from RNA polymerase II promoter involved in detection of glucose	BivPro	10.2	0.003	FOXA2
Go	negative regulation of transcription by glucose	BivPro	10.2	0.003	FOXA2
Go	negative regulation of transcription from RNA polymerase II promoter by glucose	BivPro	10.2	0.003	FOXA2
Go	positive regulation of transcription from RNA polymerase II promoter by glucose	BivPro	10.2	0.003	FOXA2
Go	regulation of transcription by glucose	BivPro	9.7	0.004	FOXA2

Go	endocrine pancreas development	BivPro	2.5	0.04	FOXA2, HES1, NKX2-2, NKX6-1, NKX6-2, PAX6, SLC2A2, SOX9
Go	response to glucose stimulus	BivPro	2.3	0.04	FOXA2, NKX6-1, PAX2, RAB11FIP2, SMAD2, SREBF1, ZNF236
MSigDB	Genes involved in Regulation of gene expression in beta cells	BivPro	4.3	0.01	FOXA2, NKX2-2, NKX6-1, PAX6, SLC2A2
MSigDB	Maturity onset diabetes of the young	BivPro	3.5	0.03	FOXA2, HES1, NKX2-2, NKX6-1, PAX6, SLC2A2

Table 4.4: Pancreas-related enrichment terms associated with bivalent promoter BIDOCS. Each row represents an enriched term associated with BiDOCS overlapping ESC bivalent promoter compared to all ESC bivalent promoter. Columns indicate Term (Go Biological Processes), Description, Chromatin State (State), Regional enrichment (RE), Hypergeometric BH/FDR-adjusted P (HBH-P, for gene enrichment) and Genes associated with the term.

4.3.2.8 *In silico* validation of identified DOCS

As the iPSC lines from each of the two subtypes were not derived from the same individuals, sample labels were randomly permuted between the FiPSCs and BiPSCs (to create two separate groups including either 2 Fi-DOCS and 3 Bi-DOCS or 2 Bi-DOCS and 3 Fi-DOCS) and the differential analysis was repeated to control for effects driven by genotype. The number of DOCS identified between the permuted samples was much lower than that observed in the original analysis (mean random sites=2.4k compared to 8.3k in the original analysis, Fig4.14). This suggests that genetic donor background is not the primary factor influencing the differential open chromatin analysis. To control for any differences driven by the two lines harbouring 48 chromosomes, samples with abnormal karyotypes (lines BiPSC-D 1-2) were removed, and the differential analysis was repeated using diffReps (Fig4.15). These data showed strong correlation with the data obtained from all iPSC lines (4.8k/8.3k DOCS overlapped with previous analysis and Spearman's rho=0.91) and (significant) DOCS were still associated with key pancreatic endodermal markers such as *FOXA2, NKX2-2,* and *PAX6*.

4.3.2.9 Cellular validation of selected DOCS using ChIP-qPCR

A number of DOCS (n=5) were selected for *in vitro* follow-up and validated by Liraz Shenhav by performing H3K4me3 ChIP-qPCR (a hallmark of open chromatin structure), using primers corresponding to Bi-DOCS that were more open in BiPSCs compared to FiPSCs. All five BiPSC lines showed higher chromatin enrichment levels in the promoter regions of *PDX1, NKX2-2, FOXA2* and *INS*, in comparison to the FiPSC lines (Fig4.16B). To highlight the open chromatin differences in the ATAC-seq data, Fig4.16B visualises six distinct DOCS between FiPSCs and BiPSCs around the *FOXA2* promoter. Overall the validation results confirm the presence of DOCS at key endodermal marker genes.

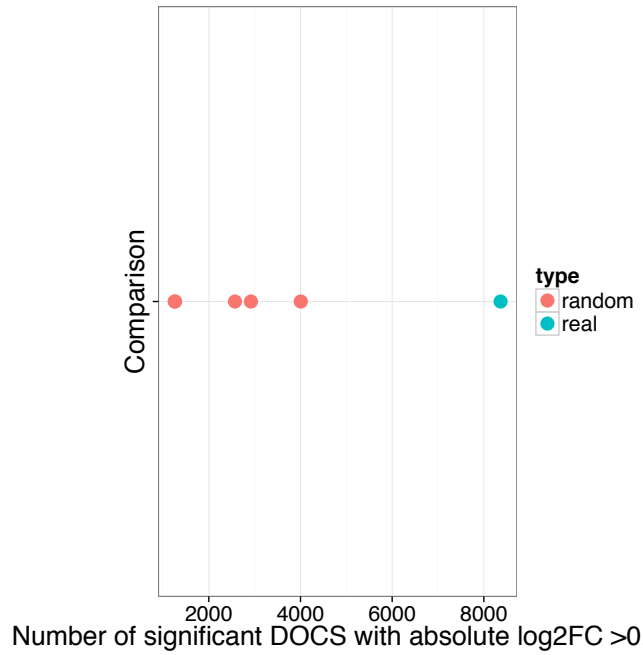
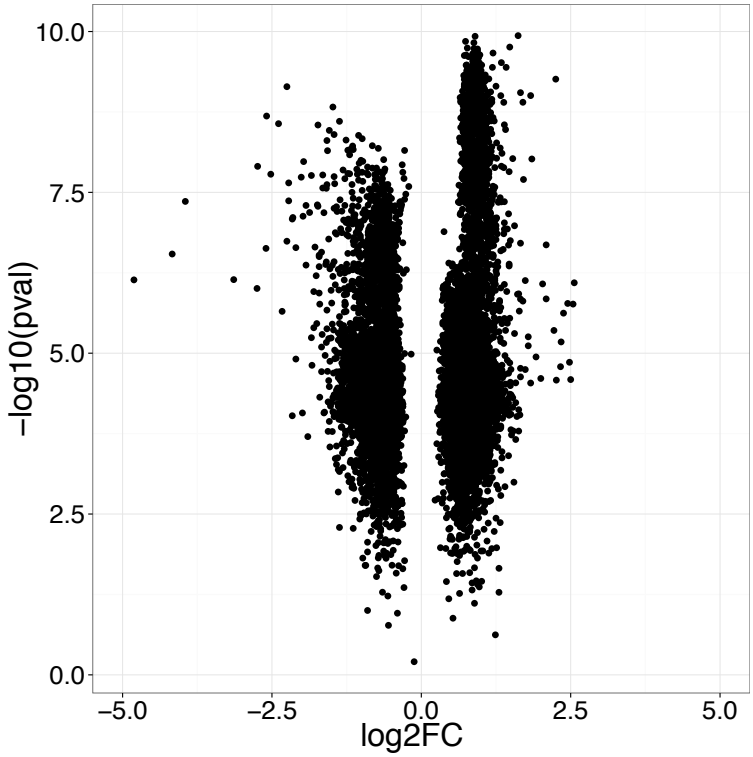
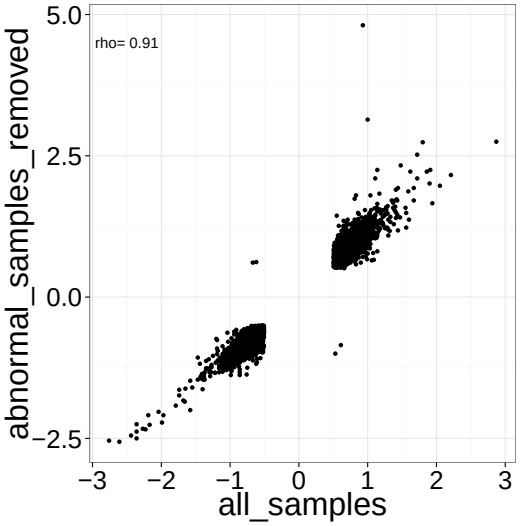


Figure 4.14: Number of significant real iPSC DOCS vs randomly permuted DOCS. Number of significant DOCS (FDR <5% and absolute log2FC >0.5 ,x-axis) from the real dataset (blue) compared to the randomly permuted dataset (red) to investigate potential effects of donor genotypes on DOCS analysis.

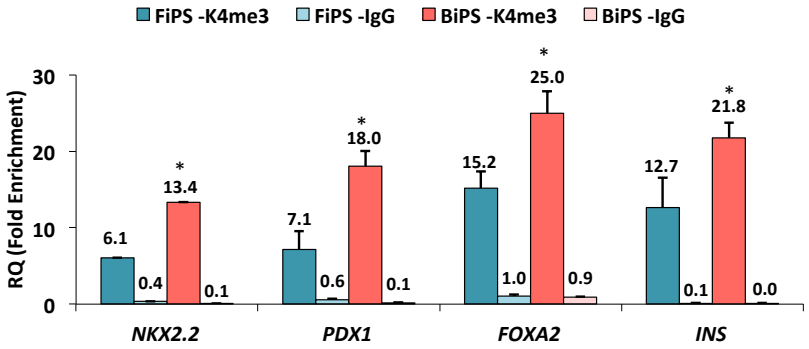


(A) Volcano plot of karyotypically normal samples

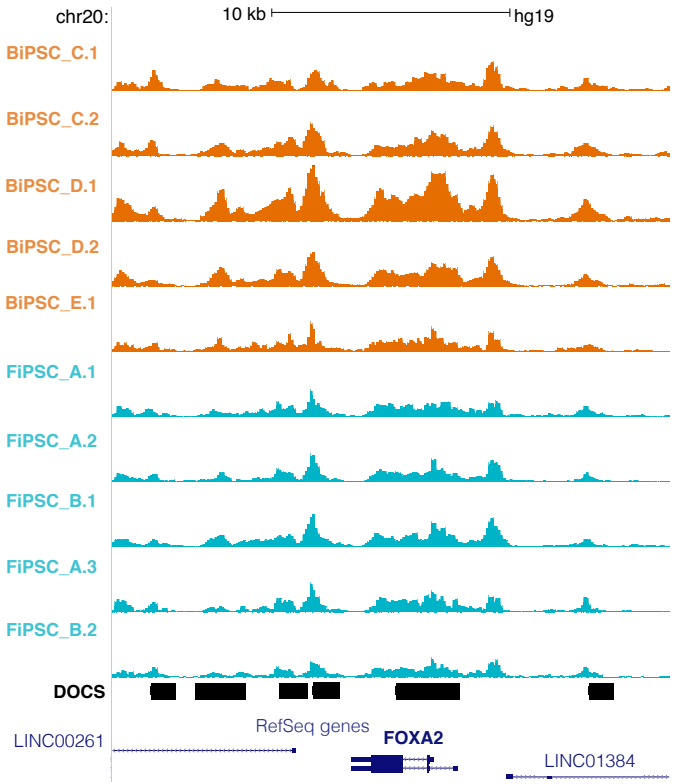


(B) Correlation all samples vs normal samples

Figure 4.15: DiffReps analysis of karyotypically normal samples. (A) Volcano plot showing log2FC (x-axis) and -log10 P-value (y-axis) of the diffReps analysis including only karyotypically normal samples (B). Correlation between log2FC of overlapping DOCS identified from the analysis including all samples (x-axis) and from the analysis including only karyotypically normal samples (y-axis).



(A) In vitro validation of DOCS



(B) FOXA2 region

Figure 4.16: Validation of selected DOCS. (A) ChIP analysis of histoneH3 tri-methyl Lys4 (H3K4me3) in BiPSCs (n=5), FiPSCs (n=4). RT-PCR was performed on bound and input DNA with primers recognizing *FOXA2, NKX2.2, INS* and *PDX1* promoter loci and are normalized to closed loci *CRYSTALLIN* and *TEX15* (closed state was validated using ATAC-seq) . Enrichment of open chromatin was validated using primer recognizing *APRT* promoter. Values are mean +/- SE and $*$ indicates P <0.05. *In vitro* validation performed by Liraz Shenhav at Tel-Aviv University. (B). *FOXA2* region comparing RPKM normalised read depth of BiPSC (orange) and FiPSC (blue) samples. Black bars near the bottom show 6 significant DOCS identified between the 2 iPSC sample types. The bottom track indicates the genes in the region with *FOXA2* being highlighted in bold.

4.4 Discussion

4.4.1 Open chromatin landscape in human pancreatic islets

The research in the first part of the chapter has dealt with the use of ATAC-seq to expand existing islet open chromatin information and to provide variable open chromatin maps for almost 30 individuals. These data will be used in Chapter 5 Section 5.3.4 for testing regions for allelic imbalance in open chromatin based on genotype. The analysis in the second part of the chapter used ATAC-seq to investigate open chromatin regions across iPSCs derived from both pancreatic beta-cells and fibroblasts, which differed in their differentiation potential, to identify open chromatin regions associated with enhanced endoderm differentiation. Specifically, the main results of this Chapter were that:

- ATAC-seq islet open chromatin regions showed substantial overlap with known islet promoter and enhancer regions and were enriched for islet eQTLs and TFBS.
- islet open chromatin regions, in particular, distal from the TSS, were substantially enriched in T2D and FG GWAS regions (on a global level)
- ATAC-seq in BiPSCs and FiPSC identified DOCS that were strongly enriched in regulatory regions involved in development
- Bi-DOCS were found to be associated with genes and TF related to early pancreatic development, thus, providing potential information on how to improve "iPSC to endoderm" differentiation

The observations for the human islet ATAC-seq data are in accordance with previous studies in islets that have shown that islet open chromatin regions correspond to islet regulatory elements such as enhancers and promoters and are strongly associated with islet-specific TF binding[84–86]. However, compared to previous studies which only generated FAIRE-seq data for up to 2 samples[84,85], the data presented here expanded the sample numbers considerably to up to almost 30 providing a more comprehensive view of islet open chromatin maps. Similar to islet studies, other

studies performed in non-islet tissue also confirmed the importance of accessible DNA regions in cell regulation[82].

Consistent with the regulatory role of open chromatin regions in human islets, these highly accessible peak regions were also globally enriched for T2D and FG GWAS signal which was strongest for distal elements (in relationship to the TSS) corresponding to enhancers. This is in accordance with previous data that showed islet enhancer chromatin states with accessible DNA are highly enriched for GWAS variants associated with T2D and other related traits[83–86]. To evaluate the enrichment of open chromatin in T2D GWAS loci in more detail, the analysis in Chapter 5 integrated different types of epigenomic marks to dissect the contribution of open chromatin, histone marks and DNA methylation to the T2D GWAS enrichment.

While the enrichment of distal enhancer regulatory elements was specific to T2D-related traits, islet open chromatin regions were also enriched in GWAS signal for TC which is a likely non-islet mediated trait. When dividing the regions into distal (likely enhancer) and proximal (likely promoter) regions it was observed that proximal islet open chromatin regions near promoters were strongly enriched for GWAS variants associated with TC. The enrichment for proximal open chromatin regions near the TSS is consistent with the almost significant UMR (unmethylated promoter regions identified from WGBS methylation data) enrichment in TC islet (and liver) GWAS data in the Chapter 3 Section 3.3.3.2. Furthermore, these results are in accordance with a recent study that found coding regions to be enriched in TC-associated GWAS variants[158].

Overall the human pancreatic islet ATAC-seq data has highlighted that islet open chromatin regions likely play a major role in islet regulation and the genetic susceptibility of T2D. While in this chapter the global enrichment of islet open chromatin peaks in T2D-related traits was evaluated, the analysis in Chapter 5 also explores the role of open chromatin at individual T2D GWAS loci. Specifically, the individual islet open chromatin data is used to identify allelic imbalance at GWAS variants to prioritise these variants for functional follow-up and provide potential molecular mechanisms governing T2D risk at these loci.

The second part of this Chapter has systematically investigated sites of open chromatin across the genome of iPSCs derived from both pancreatic beta-cells and skin fibroblasts. By overlaying this

information with analogous data from primary human islets, it was shown that despite derivation from different donor tissues, the open chromatin peaks in both iPSC subtypes cluster more closely together than their original cell type of origin, confirming their overall stem cell behaviour. Despite the global similarity in shared open chromatin peaks across both iPSC subtypes, DOCS between BiPSCs and FiPSCs could be identified which were strongly enriched for certain regulatory genomic annotations. Bi-DOCS (regions more open in BiPSCs compared to FiPCS) were enriched for endodermal weak enhancers, bivalent enhancers and bivalent promoters, confirming that they overlap regulatory motifs previously shown to facilitate 'developmental competence' and to influence endodermal lineage differentiation efficiency.

Through gene and pathway analysis Bi-DOCS were found to be enriched for GO terms relating to early pancreatic endoderm signalling pathways and TFs, including the WNT and NOTCH pathways and FOXA2 signalling. Both WNT and NOTCH signalling are two of the major pathways targeted by current protocols leading *in vitro* differentiation protocol[219,220] while FOXA2 is one of the master regulators of endodermal development. Along with FOXA1, FOXA2 has been shown to regulate expression of both NEUROG3 and PDX1, with *Foxa2-/-* mice displaying abnormal islet development and dying shortly after birth from severe hypoglycaemia[251]. *FOXA2* binding sites are enriched at weak endodermal enhancers, priming cells to receive extracellular cues signalling to differentiate into the various organs of the endodermal lineage, including pancreas, liver, and intestine[217]. Enrichment of *FOXA2* motifs in BiPSC upregulated DOCS suggests that this iPSC subtype retains 'memory' of its original endodermal cell type, whereby lineage-relevant poised enhancers are maintained. This is exemplified by the enhanced spontaneous differentiation potential of BiPSCs from embryoid bodies into both endoderm and pancreatic progenitors, compared with FiPSCs. Bi-DOCS may provide clues for optimisation of pluripotent stem cell directed differentiation into beta-like cells, for allowing robust differentiation of multiple cell lines from varied origins.

Maintenance of genomic 'bookmarks' in reprogrammed iPSCs and the true extent of epigenetic memory remains a subject of discussion within the field. It has been suggested that cell type of origin, at least for low passage numbers, influences the genomic landscape and lineage-specific differentiation efficiency of iPSCs, an effect also replicated in cloned animals where cell type of origin was shown to influence transcriptional and epigenetic profiles as well as *in vitro* differenti-

ation[224–226]. However, several studies suggest that donor genotype is the major influence on the transcriptomic and epigenetic profiles and differentiation capacity of iPSCs. Culture conditions, passage number, cell cycle phase, and reprogramming technology can also induce transcriptional and epigenetic differences across iPSC types[221–223]. Additional work will be required to determine whether the effect is truly dependent on the donor cell type, in particular, since the iPSC subtypes analysed in this chapter differed substantially in passage number and prolonged passaging was suggested to eradicate epigenetic memory effects[225]. Accordingly, the high passage number of FiPSCs may explain the observed enrichment of Fi-DOCS in ESC and iPSC active enhancer regions and the lack of enrichment in chromatin states related to fibroblast/mesoderm function. However, independent of the source of variation, the discovery of differences in open chromatin between BiPSC and FiPSC that differ in their endodermal differentiation potential, was still shown to provide biologically-relevant insights into endodermal development.

Overall the work presented in this chapter provided a valuable resource for researchers that could improve understanding of endodermal development and be used to generate enhanced islet cell models as well as to develop novel therapeutic approaches for diabetes.

CHAPTER 5

Integration of epigenomic annotation

5.1 Introduction to integration of epigenomic annotation

5.1.1 Interactions between epigenomic factors

As described in Chapter 1, DNA methylation, histone modifications, chromatin accessibility and transcription are highly interconnected and correlated in the human genome. DNA methylation found at CpG sites in the genome is generally associated with heterochromatin, transcriptional silencing and imprinting which is often mediated via changes in chromatin structure and histone PTMs [118].

While the exact mechanisms and the directionality of these interactions remain unclear, it was recently highlighted that integration of chromatin marks provides useful information for understanding genome regulation and disease susceptibility. Several recent large-scale studies collected epige-

nomic data to predict chromatin states from a large number of chromatin marks across multiple human tissues and cell lines[65,76,82,252]. These studies highlighted distinct associations between certain histone modifications and various other genomic annotations including TSS, TFBS and open chromatin regions. For instance, Ernst et al 2011[76] found that active enhancers and active promoters are associated with RNA Polymerase II activity and nucleosome depletion. Active Enhancers were also found to be enriched for tissue-specific TF binding events. Linking these enhancer and promoter states to TSS of nearby genes identified cell-type-specific clusters highlighting relevant biological terms and functions. An even larger study investigated chromatin states across up to 111 human tissue and cell types in combination with DNA accessibility and DNA methylation[82]. As described in other Chapters and consistent with other recent studies[131,239,240], the authors found that DNA accessibility was usually associated with activity, binding of regulatory TF and active histone marks while high levels of DNA methylation were associated with repression, compact chromatin and inactive marks. In contrast, intermediate and low methylation strongly correlated with enhancer and promoter histone marks, respectively, and were found in more open chromatin region. Consistent with the increased DNA accessibility and intermediate DNA methylation states, enhancers were enriched for TFBS motifs which specifically overlapped DNase footprints indicative of true binding events[82].

The co-occurrence of certain epigenomic factors suggests that DNA methylation and histone modifications are linked through shared biochemical pathways[253]. Proteins that bind to CpG regions depending on DNA methylation levels play a critical role in governing these interactions. For instance, MeCP2, a member of the MBD family, binds to methylated CpG sites and also associates with histone deacetylases and methyltransferases to influence histone modifications such as H3K9me[254]. Similar to DNA methylation, H3K9me3 is in general associated with repressed heterochromatin states. In addition to interactions with H3K9, DNA methylation is associated with H3K36 methylation, which is found predominantly in exonic regions and linked to increased levels of DNA methylation[255]. In contrast to H3K36 and H3K9 methylation, other histone variants such as H3K4me3 show strong negative correlation with DNA methylation potentially by blocking access to DNA methyltransferases which likely reduces DNA methylation levels at promoter and CpG-island regions[253,256]. These observations highlight the complexity of epigenomic patterns and annotations that regulate the genome.

Comparing genomic annotations across cell types and tissues can be highly useful to identify biologically relevant connections across developmental lineages and tissues. In accordance with the previous study, enhancer clusters were associated with cell-type specific functional modules: for instance, one study found associations between stem cell enhancers and developmentally active genes as well as immune cell enhancers and genes involved in the regulation of the immune system[82]. In addition to providing improved understanding of tissue specificity and cell function, enhancers also showed enrichment for GWAS signals specific to disease-relevant cell types (for more details see Chapter 1). Overall, these studies highlighted the complex relationship between transcription, DNA methylation, open chromatin and histone marks and showed that integrating these states is useful for understanding tissue and cell function as well as disease susceptibility.

5.1.2 Integration of genomic annotation to understand islet function and T2D development

Since pancreatic islets are central for T2D pathophysiology, several recent studies have focused on integrating islet genomic data such as transcription, chromatin and TFBS information to improve understanding of islet function and T2D genetic susceptibility. Parker et al 2013[86] combined histone marks from 10 cell types including islets to identify chromatin states associated with promoter, enhancer, insulator, transcriptional and repressive function. Across multiple cell types this approach identified both short and long enhancer states ("stretch enhancer") of which the latter were highly cell-type-specific. Integrating these stretch enhancers with RNA-seq data showed that they were also strongly correlated with the expression of nearby cell-type-specific genes, thus, suggesting that stretch enhancers are critical for cell identity and function. Across cell types these stretch enhancers were also enriched for tissue-relevant GWAS traits and, in particular, islet stretch enhancers showed strong enrichment in T2D and FG GWAS regions.

A parallel study[85] combined chromatin states defined from histone marks with TFBS and transcriptional data to improve understanding of the pancreatic islet regulome and T2D development. It was found that 5 islet relevant TFs including FOXA2, MAFB, NKX2.2, NKX6.1 and PDX1 formed a highly interconnected, cross-regulatory network with highly correlated binding sites throughout

the genome. Integration with chromatin accessibility information (derived from FAIRE-seq and H2AZ ChIP-seq data), CTCF binding sites and chromatin states derived from several histone marks (H3K4me1, H3K4me3, H3K27ac) showed that TFs bound to highly accessible chromatin regions in human islets. To further understand the function of these accessible states, the study investigated the relationship between these chromatin accessible sites and the expression of nearby genes. While islet promoter states were equally enriched for both islet-specific and non-specific genes, it could be shown that active enhancers, which were enriched for multiple TFs binding events, were strongly associated with islet-specific gene expression.

Consistent with the aforementioned studies[86], these results highlighted the importance of these active enhancer clusters (defined as "stretch enhancer" by Parker et al 2013) in regulating gene expression linked specifically to islet function. In addition, the study also investigated the role of these active TF bound enhancer states in the genetic susceptibility of T2D and found that the active islet enhancer clusters were specifically enriched in T2D and FG GWAS signals. This enrichment of islet enhancers in T2D GWAS signal remained even after removing known significantly associated T2D and FG GWAS loci, thus, suggesting that islet genomic annotations can also be used to identify novel GWAS signals that have not reached genome-wide significance. Functional follow-up of selected enhancer regions showed that the GWAS variant rs58692659 at the *ZFAND3* locus modified NEUROD1 binding by altering the TF motif, thus, providing a potentially functional mechanisms explaining T2D GWAS risk[85].

A follow-up study also combined genetic fine-mapping data with the epigenomic TFBS and chromatin state date to pinpoint likely causal T2D and FG GWAS variants and mechanisms governing T2D susceptibility[83]. Gaulton et al 2015 found strong enrichment of fine-mapped credible set variants in islet *FOXA2* ChIP-seq binding sites and identified a high PP variant in an FOXA2 bound enhancer at the *MTNR1B* locus. Functional follow-up indicated allele-specific differences in NEUROD1 binding activity associated with increased islet *MTNR1B* expression, thus, highlighting a potential molecular mechanism at this locus that may explain the T2D GWAS signal.

Parallel to equivalent efforts for other tissues and diseases[76,81,82,87,131,257], these islet studies have clearly highlighted the power of combining different genomic islet datasets to enhance the identification of islet regulatory states, improve understanding of the islet regulatory network and help

elucidate the role of GWAS variants in T2D development.

The work presented in this chapter aims to achieve the following

- to extend existing islet chromatin states by integrating different epigenomic datasets
- to combine refined islet chromatin states with genetic T2D and FG GWAS data and determine a maximum likelihood enrichment model combining multiple annotations
- to characterise the contribution of individual epigenomic factors to the islet chromatin state enrichment in T2D and FG GWAS signal
- to use the combined model to prioritise highly likely causal variants for testing allelic imbalance in open chromatin

5.2 Methods

5.2.1 ChIP-seq data processing for ChromHMM

Human islet ChIP-seq histone mark and TFBS raw data were obtained from various sources: H3K4me1, *CTCF* and H3K27ac (from Pasquali et al 2014[85]), H3K36me3 and H3K4me3 (from Moran et al 2012[258]) and H3K27me3 (from Encode/Epigenome Roadmap consortium `http://egg2.wustl.edu/roadmap/webportal/index.html`[82]). The fastq files were aligned to hg19 using bowtie1 (version 1.1.1[232]) with modified default settings (-m 1 which removes reads with more than 1 valid alignment and -n 1 which defines the maximum number of mismatches in seed) and PCR duplicates were removed from the aligned bam files with Picard tools (v1.119,[259]). The resulting reads were converted into bed format using the bedtools bamToBed function (version v2.21.0)[161] and extended by 200bp towards the 3' end to account for fragment size.

5.2.2 ChromHMM state predictions

Binarised 200bp density maps from the bed files of the 6 ChIP-seq marks were created using a Poisson distribution implemented in the BinaryBed function of the ChromHMM software as described in Ernst et al 2011 and 2012[76,260]. From these epigenomic density maps, 11 ChIP-only chromatin states were derived using a multivariate Hidden Markov Model implemented in the Learnmodel function (standard settings, h19 genome) of the software ChromHMM[260].

To generate additional sets of chromatin states based on ChIP-seq, ATAC-seq peaks and DNA methylation status the following approach was used. From a subset of the islet ATAC-seq data (n=17 available at the time of analysis) described in Chapter 4, open chromatin peaks were called by applying a departmental pipeline[230] that used sample-specific read depth and width parameters, which were chosen based on the signal to noise ratio of a given sample, to identify peaks (see website for details `http://userweb.molbiol.ox.ac.uk/public/telenius/PipeSite.html`). The presence of ATAC-seq peaks was binarised over 200bp windows in the genome. Similarly, whole-genome CpG methylation status (obtained from human islets in Chapter 3) was binarised into hypermethylation and hypomethylation based on a threshold of 60% methylation across 200bp windows of the genome.

These binarised 200bp ChIP-seq, ATAC-seq and DNA methylation maps were combined and used to generate 3 sets of chromatin states derived from ChIP and DNA methylation data (ChIP-Meth), ChIP and ATAC-seq data (ChIP-ATAC) or ChIP, ATAC-seq and DNA methylation data (ChIP-ATAC-Meth) using the Learnmodel ChromHMM function. As suggested by Ernst et al 2011[76], after evaluating models with up to 20 chromatin states, a 15 state model was chosen based on the resolution provided by identified states (see Fig5.1A for details). Enrichment of these states in the following genomic annotations was determined using the chromHMM OverlapEnrichment and NeighborhoodEnrichment function in chromHMM: CpG-islands, intergenic regions (Genome), Refseq gene and TSS regions as well as publicly available islet TFBS.

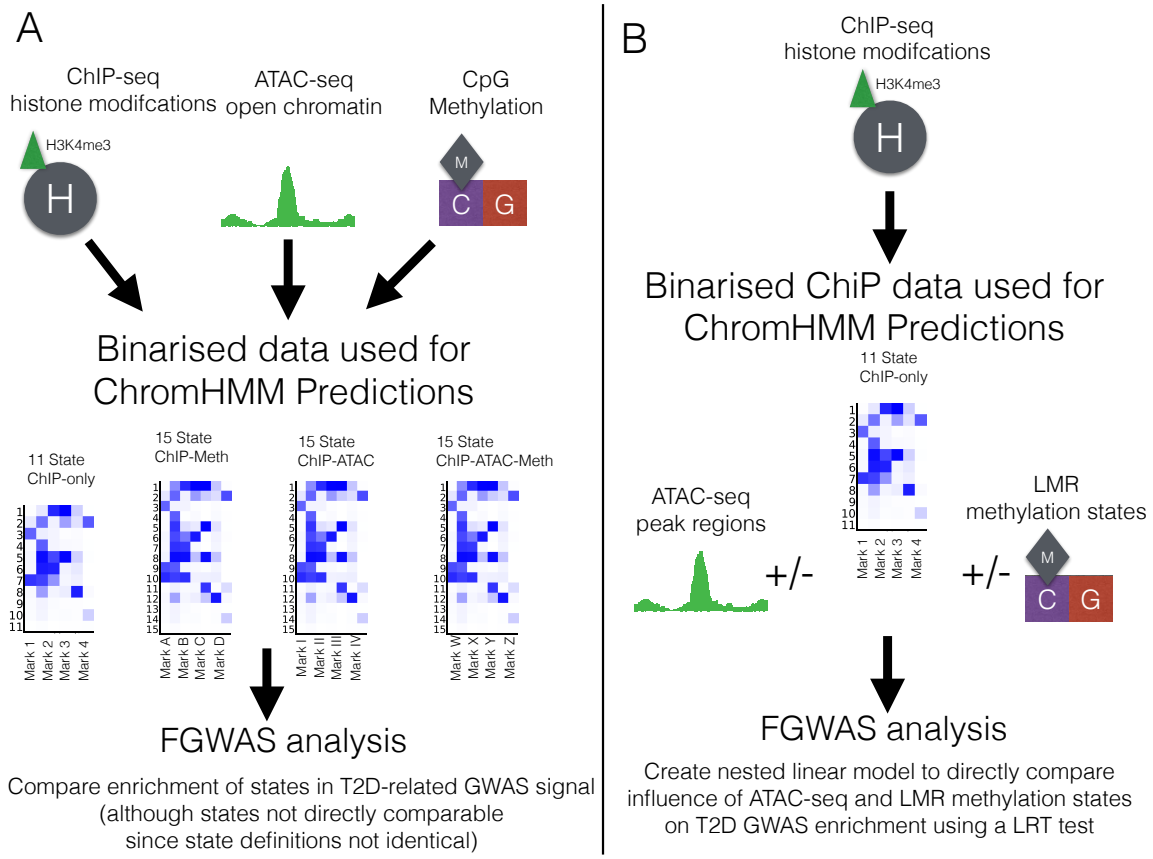


Figure 5.1: Outline of ChromHMM and FGWAS analysis of ChIP, ATAC-seq and DNA methylation.**(A)** Different combinations of epigenomic data (top) were combined to generate different sets of refined chromatin states (11 ChIP-only and 15 ChIP-Meth, ChIP-ATAC and ChIP-ATAC-Meth states, middle) using chromHMM. These sets of chromatin states were then tested for enrichment in T2D-related GWAS traits using FGWAS to both compare individual states and the best multi-state enrichment model (derived using cross-validation). However, chromatin states defined from a different set of epigenomic marks (ChIP-only, ChIP-Meth, ChIP-ATAC and ChIP-ATAC-Meth) are not equivalent and the enrichment can not be easily compared across models. Thus, in **(B)** a nested model approach was applied. That is, ChIP-only chromatin states were generated and FGWAS maximum likelihood models were defined using ChIP-only, LMR and/or ATAC-seq peak regulatory regions. The combination of all these annotations represented a nested linear model and the changes in maximum likelihood by adding/removing LMR and ATAC-seq states could be statistically evaluated using a Loglikelihood Ratio Test (LRT).

5.2.3 FGWAS enrichment analysis

As described in Chapter 2 section 2.6.1, FGWAS[158] uses a hierarchical modelling approach to determine the enrichment of annotations in GWAS signal. To determine the maximum likelihood model the following approach suggested by Pickrell et al 2014[158] was used for each set of chromatin states (ChIP-only, ChIP-ATAC, ChIP-Meth and ChIP-ATAC-Meth), separately. In accordance with the author's suggestion to test all significant annotations, hypomethylated LMRs (see Chapter 3 section 3.3.3.1 for more details), were also evaluated in this approach (together with the chromatin states) to determine the best FGWAS model.

1. A model was fitted for each annotation individually to identify all annotations that were significantly enriched with the trait.
2. The annotation with the highest increase (and enrichment) in the maximum loglikelihood was added to the model and the analysis is repeated with the additional annotation.
3. Annotations were added as long as they increase the maximum loglikelihood of the newly tested model.
4. A 10-fold cross-validation approach is used after determining a penalty parameter based on the maximum likelihood of a penalised log-likelihood function to avoid overfitting
5. Each annotation is dropped from the model and the maximum cross-validation likelihood is evaluated. If a reduced model has a higher cross-validation maximum likelihood, additional annotations are dropped until the model can not be further improved based on the maximum cross-validation likelihood. This model is described as the best fitted model and used for the remaining analysis.

5.2.4 Comparing the influence of open chromatin and DNA methylation on GWAS enrichment

Initially, similar enhancer chromatin states derived from the 4 different ChromHMM analyses (ChIP-only, ChIP-ATAC, ChIP-Meth, ChIP-ATAC-Meth) were compared. Similarity was deter-

mined based on shared histone chromatin marks according to the chromHMM emission parameters (approach visualised in Fig5.1A)

However considering that the chromatin states were derived from different sets of annotations across different runs of the ChromHMM a direct comparison is not fully possible. Hence, a nested model approach was used to further dissect the contribution of open chromatin and DNA methylation to the enrichment (approach visualised in Fig5.1B) . Specifically, an FGWAS analysis was performed that combined the ChIP-only chromHMM states with raw LMRs (representing DNA methylation) and ATAC-seq peaks (representing open chromatin). After determining the best maximum-likelihood cross-validation model (as described above) a log-likelihood ratio test was used on the maximum log-likelihood of the nested models in order to allow a formal statistical analysis. The following nested models were compared using FGWAS:

- a ChIP-only annotation model,
- a ChIP-annotation with hypomethylated LMRs model,
- a ChIP-annotation with ATAC-seq open chromatin peaks model
- a ChIP-annotation with both LMRs and ATAC-seq peaks model

5.2.5 Prioritising high posterior probability variants at significant loci for testing of allelic imbalance in open chromatin

After determining the prior probability of association as described in Chapter 2 section 2.6.1, Pickrell et al 2014 [158] showed that this information can be used to derive a posterior probability (PP) of association for both the region and the SNPs in a given region. Variants at significant GWAS regions with a high FGWAS PP (PP >0.10) were selected to determine allelic imbalance in ATAC-seq open chromatin data from heterozygous individuals. In addition to standard genome-wide significance ($P < 5x10^{-8}$), significant GWAS regions were determined based on a regional PP of 0.9 (representing the sum of the PPs of all SNPs in a given region). Pickrell observed that a regional PP of 0.9 or above can be used to identify sub-threshold GWAS loci.

5.2.6 Allelic imbalance in open chromatin

For high PP variants at significant GWAS regions allele counts were obtained from 28 individual ATAC-seq islet sample bam files using samtools mpileup function[203] and after filtering (only variants present in at least 2 individuals, with a per sample read depth >9 and at least 5 reads per allele were taken forward) the allelic imbalance was determined using a binomial test. Regions with significant (FDR-adjusted P <0.05) allelic imbalance in the binomial test were identified and followed-up using the software WASP (Version 0.2[261]) to deal with read-mapping and reference allele bias, which can often lead to spurious results. To remove reads associated with mapping bias, the WASP pipeline for mappability filtering (described at `https://github.com/bmvdgeijn/WASP/tree/master/mapping`) was used. In short, reads of the unfiltered bam file (within 1Mb of the high PP variant of interest) were identified that overlapped an SNP identified in 1000G data (phase1 v3 November 2010)[262]. For each read overlapping an SNP, the genotype of that SNP was changed to the alternative allele and the read was remapped. Any read that failed to realign in the same position in the genome was discarded. Ultimately, PCR duplicates were filtered using the WASP "rmdup_pe.py" script which removed reads independent of their mapping score to avoid any bias.

5.2.7 Determining overlap with TF motifs

The tool "Fimo" implemented in the "meme" software package[263] was applied to identify TF motifs that significantly (FDR <0.05) matched the sequence overlapping the SNP variant (20bp up and downstream). Based on position-specific scoring matrices of a given motif, fimo calculated a log-likelihood position score for each motif based on sequencing positions and converted these scores into P-values using dynamic programming followed by multiple testing correction using an FDR approach.

5.3 Results

5.3.1 Integration of epigenomic factors to refine chromatin states

Recent studies have shown that DNA regulatory elements can be derived from DNA methylation, chromatin accessibility or histone mark information and that there is substantial overlap between these different epigenomic marks. Also, integration of all of these factors has been suggested to significantly improve the enrichment of GWAS signals and help understand genetic disease associations.

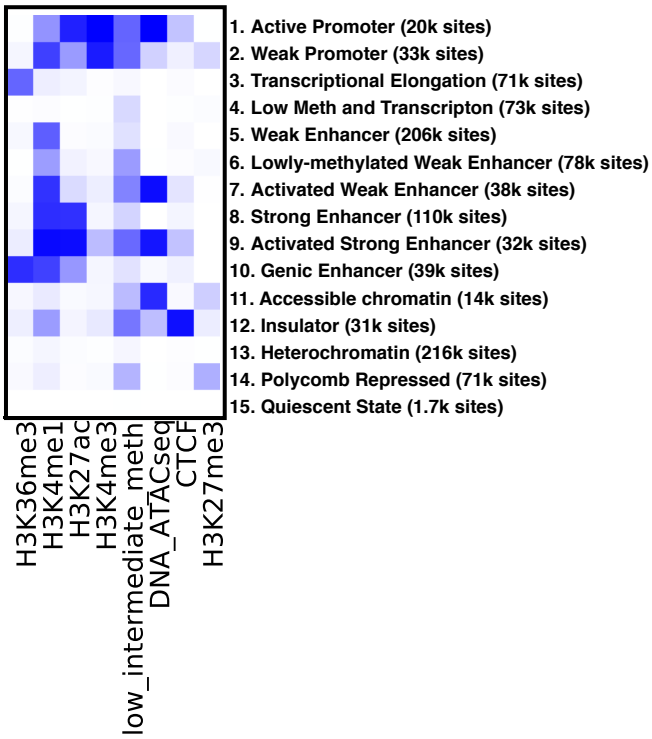
To refine islet chromatin states, I aimed to integrate different types of islet regulatory annotations, such as DNA hypomethylation and ATAC-seq open chromatin regions. As described in Chapter 3, WGBS DNA methylation data was used to determine hypomethylated regulatory regions termed LMRs and UMRs that showed substantial overlap with promoters and enhancers, respectively. In addition, in Chapter 4, ATAC-seq DNA accessibility information was used to determine highly accessible open chromatin regions which also correspond to both promoters and enhancers. To understand the link between islet DNA methylation and open chromatin, the overlap between ATAC-seq open chromatin sites (n=141k derived from 17 human islet samples as described in Section 5.2.2) and hypomethylated regulatory elements was determined. There was significant enrichment of both LMRs (FE=13.9) and UMRs (FE=10.7) in ATAC-seq peaks (compared to randomised region) and in total, 53% of LMRs (n=19.8k) and 98% of UMRs (n=13.3k) overlapped 16% (n=22.5) and 13% (n=18.5k), respectively, of all ATAC-seq peaks.

Due to the high interconnectivity, these different epigenomic regulatory data types were combined using chromHMM. Specifically, previously generated pancreatic islet chromatin ChIP-seq histone marks as well as ATAC-seq open chromatin peaks and WGBS hypomethylated regions (defined as having <60% methylation levels) were integrated in the chromHMM model to extend existing chromatin states (Fig5.2A). This analysis identified 15 chromatin states that could be broadly classified into promoter (H3K4me3 marks), transcribed (H3K36me3 mark), enhancer (H3K4me1 and H3K27ac marks), insulator (CTCF mark), repressed (H3K27me) and inactive states (lack of histone marks) based on the enrichment of histone marks. These broad classes could be further divided into

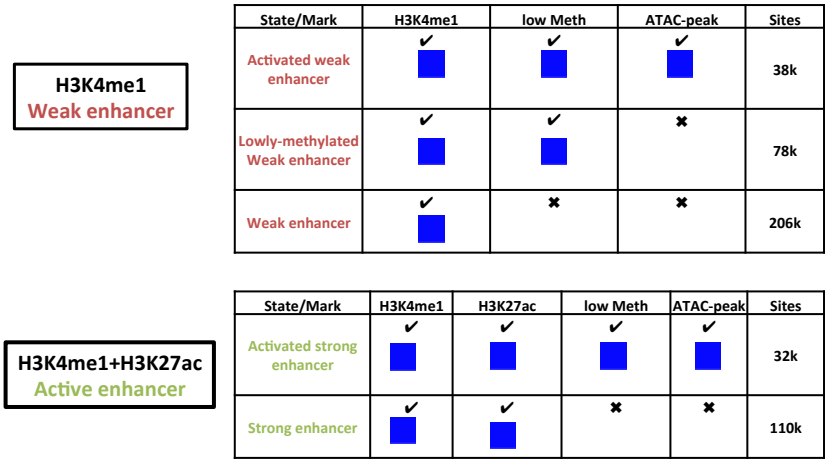
subtypes depending on their activity level. For instance, promoters (marked by H3K4me3) could be classified into active and weak (bivalent) promoters of which the former were associated with open chromatin while the latter were associated with H3K27me3.

In addition, the analysis identified also several enhancer classes including 2 subtypes of strong enhancers (usually marked by H3K4me1 and H3K27ac) and 3 subtypes of weak enhancer (usually marked by H3K4me1 only) that can be distinguished based on ATAC-seq open chromatin and DNA methylation levels. Specifically, the chromHMM model identified activated weak enhancers with low methylation and open chromatin (state 7, n=38k), lowly-methylated weak enhancers (state 6, n=78k) and weak enhancers (state 5, n=206k) with closed chromatin and hypermethylation as well as activated strong enhancers (state 9, n=32k) with open chromatin and hypomethylation and strong enhancers with closed chromatin and hypermethylation (state 8, n=110k, Fig5.2).

Consistent with previous studies the chromatin states were also associated with specific locations in the genome (Fig5.3). For instance, promoter states were enriched for CpG-island, TSS, exonic and genic regions while lowly-methylated transcribed, inactive heterochromatin and quiescent states are enriched for intergenic ("Genome %") regions. Also, this analysis highlighted the regulatory role of activated weak and strong enhancers, which are characterised by increased DNA accessibility and low DNA methylation, in the islet genome. These activated enhancers were strongly enriched for publicly available active islet TFBS (*FOXA2, MAFB, NKX2.2, NKX6.1* and *PDX1*) while no such enrichment was found for the remaining weak and strong enhancer regions.



(A) ChIP-ATAC-seq-Meth states



(B) Weak and Strong enhancers

Figure 5.2: ChromHMM ChIP-ATAC-Meth states. (A) 15 chromatin states (y-axis) were derived from ChIP histone marks, DNA methylation and ATAC-seq open chromatin annotations (x-axis) using chromHMM LearnModel function and for each state the relevant marks characterising the state are shown. The colour is based on the chromHMM emission parameters and a darker colour indicates a higher frequency of a mark at a given state. (B) Weak enhancers (marked by H3K4me1, top red) and strong enhancers (marked by H3K27ac and H3K4me1, bottom green) were subdivided by the chromHMM analysis according to methylation and ATAC-seq status. In total, 3 weak enhancer and 2 strong enhancer subtypes were identified.

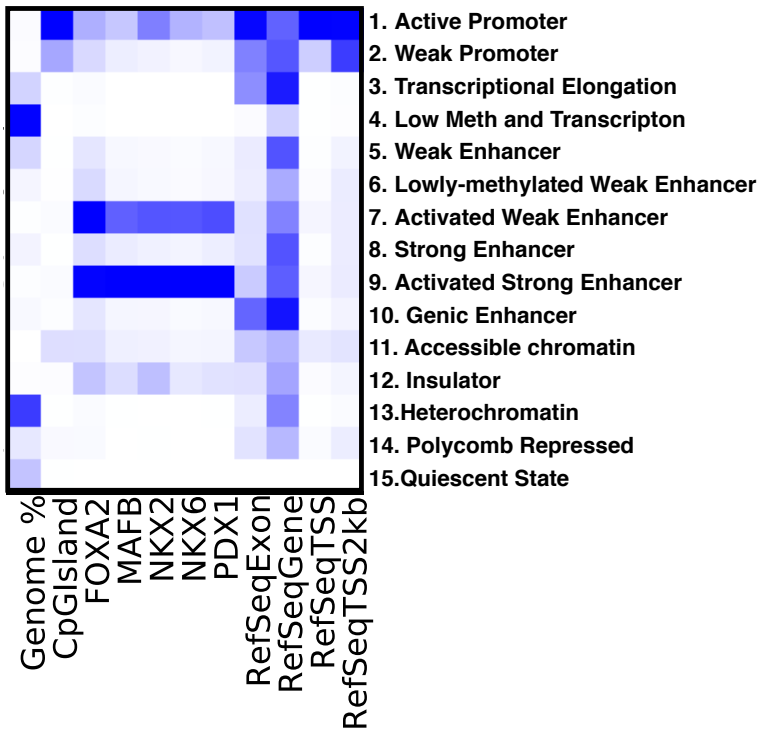


Figure 5.3: Enrichments of chromatin states in genomic annotations. For each state the enrichment in a given hg19 genome annotation is shown (darker colour indicates higher enrichment). The annotations include the following (all regions but TFBS were derived from UCSC genome browser): Intergenic (Percentage of Genome), CpG-island regions, publicly available islet TFBS[85] (*FOXA2, MAFB, NKX2.2 (NKX2), NKX6-1 (NKX6) and PDX1*), RefSeq-exons, RefSeq-genes, RefSeq Transcription Start site (TSS) and 2kb within a RefSeq TSS (RefSeqTSS2kb)).

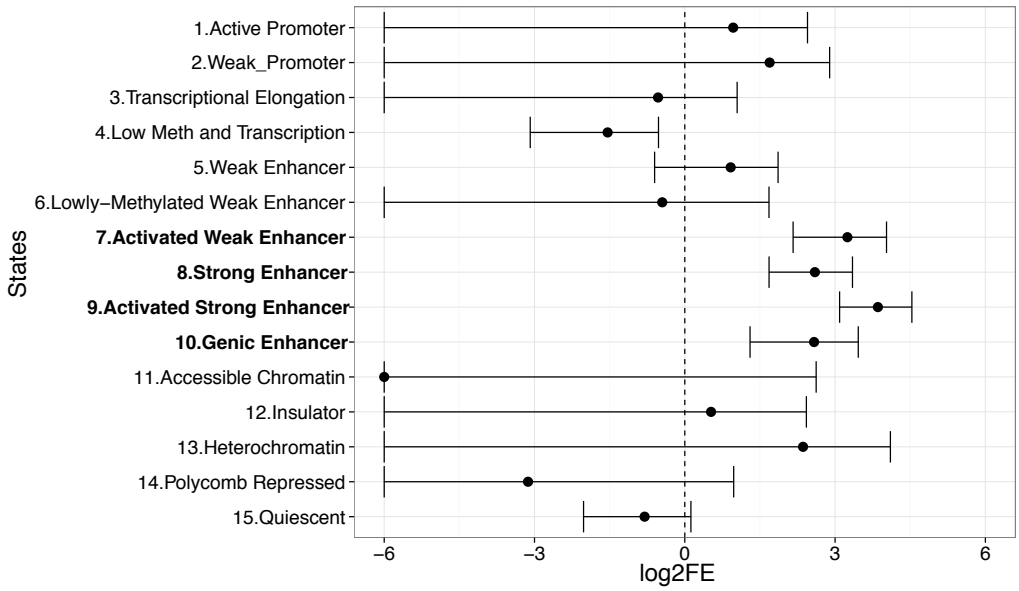
5.3.2 Enrichment of chromatin states in T2D and FG GWAS regions

Subsequently, the statistical modelling approach from FGWAS was used to determine the enrichment of these refined chromatin states in T2D and FG GWAS regions. In short, all chromHMM states were tested individually for enrichment (Fig5.4A) and the four significantly enriched chromatin states (activated strong enhancer, activated weak enhancer, strong enhancer and genic enhancer, shown in bold in Fig5.4A) were combined to find the model with the maximum likelihood estimated using a cross-validation approach to avoid overfitting (see Section 5.2.3). Although the chromatin states already included binarised methylation levels (above or below 60% methylation), LMRs, which showed significant enrichment in T2D and FG GWAS regions in Chapter 3, were also included in the cross-validation approach to determine the best enrichment model. This approach was chosen, since only 54% of all LMRs overlapped with the four significant enhancer chromatin states and the inclusion of LMRs in the cross-validation analysis is in accordance with the suggestion from Pickrell et al 2014[158]. The author recommended to include all enriched annotations when determining the maximum likelihood model.

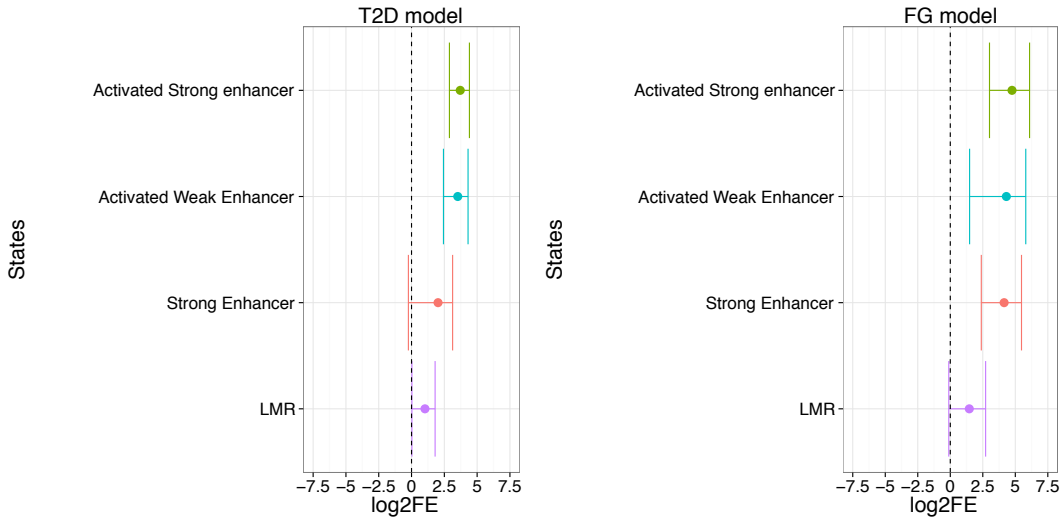
For T2D the maximum likelihood model combines activated strong enhancer (log2FE=3.7, CI=2.9 to 4.4, n=32k), activated weak enhancer (log2FE=3.5, CI=2.4 to 4.4, n=38k), strong enhancer (log2FE=2.0, CI= -0.2 to 3.1, n= 110k) and LMRs (log2FE=1, CI=0.02 to 1.8, n=38k) of which all but strong enhancers were significantly enriched (Fig5.4B , panel left). A very similar maximum likelihood model was predicted for FG which combined the same regulatory annotations (Fig5.4B, panel right): Activated strong enhancer (log2FE=4.8, CI=3.0-6.1, n=32k), activated weak enhancer (log2FE=4.3, CI=1.5 to 5.8, n=38k), strong enhancer (log2FE=4.1, CI=2.4 to 5.5, n=110k) and LMRs (log2FE=1.5, CI=-0.1 to 2.7, n=38k). Consistent with the single state enrichment analysis (Fig5.4A), it was observed that both activated weak and strong enhancer states with hypomethylation and open chromatin showed the strongest enrichment in the combined model (Fig5.4B). The increased enrichment is consistent with previous studies in both islet[85] and non-islet tissues[131] which found that additional epigenomic annotations can improve enhancer enrichment in GWAS signal.

Notably, LMRs increased the (cross-validation) maximum likelihood of both the T2D and FG enrichment models, despite the fact that the enrichment models already included chromatin states

with DNA hypomethylation status. This suggests that LMRs, which only partially overlapped the enhancer states, potentially include additional T2D-relevant regulatory regions that may not be fully captured by the chromHMM enhancer states in the models.



(A) FGWAS enrichment all states



(B) T2D and FG FGWAS best model

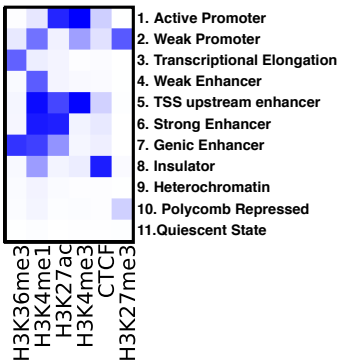
Figure 5.4: Enrichment of chromatin states in T2D and FG GWAS regions. (A) FGWAS Log2 Fold Enrichment including 95% CI (log2FE, x-axis) of all chromatin states (y-axis) in T2D GWAS regions. Significantly enriched annotations are shown in bold. (B) T2D (left) and FG (right) FGWAS maximum likelihood model determined through cross-validation. log2FE and 95% CI (x-axis) of annotations included in the maximum likelihood model (y-axis) are shown.

5.3.3 Characterising the role of DNA methylation and ATAC-seq chromatin accessibility information in T2D GWAS regions

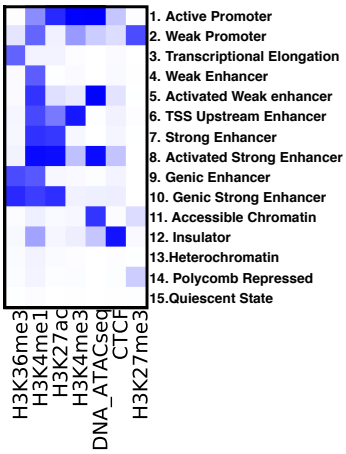
To further dissect the contribution of the newly integrated epigenetic marks (ATAC-seq DNA accessibility and WGBS methylation levels) to the enrichment in GWAS signals, the chromHMM analysis was repeated with different combinations of epigenetic marks. Specifically, only ChIP-seq marks (ChIP-only, excluding any methylation and ATAC-seq information) or ChIP-seq marks in combination with either ATAC-seq (ChIP-ATAC) or hypomethylation information (<60% methylation levels, ChIP-Meth) were used to predict chromatin states (Fig5.5A-5.5C). FGWAS was then applied to compare the enrichment of individual enhancer states across all of these different chromatin state sets Fig5.6. ATAC-seq information was found to have the most substantial effect on enrichment and increased T2D enrichment of gene enhancer (max log2FE for ATAC state (max ATAC log2FE)=3.1 vs max log2FE for non-ATAC state (max non-ATAC log2FE)=2.7), weak enhancer (max ATAC log2FE=3.2 vs max non-ATAC log2FE=1.1) and activated enhancer states (max ATAC log2FE=3.9 vs max non-ATAC log2FE=2.8) when comparing ChIP-ATAC and ChIP-ATAC-Meth to ChIP-only and ChIP-Meth states. In contrast, methylation status only improved the enrichment substantially for "Other Enhancers" (marked by H3K4me1, H3K27ac and H3K4me3), marginally enhanced the enrichment of strong and weak enhancers and for gene enhancers even decreased the enrichment (compared to the chromatin state sets).

In addition, the best (maximum likelihood) FGWAS enrichment model was determined for each new set of chromatin states (ChIP-only, ChIP-ATAC, ChIP-Meth) by combining significant chromatin states and LMRs and are shown together with the ChIP-ATAC-Meth maximum likelihood FGWAS enrichment model in Fig5.7. Similar to the comparison between chromatin state sets (ChIP-only, ChIP-ATAC, ChIP-Meth, ChIP-ATAC-Meth) in Fig5.6, a comparison within each chromatin state set (within each subplot) showed that ATAC-seq states often show the strongest enrichment (e.g. activated strong and weak enhancer within ChIP-ATAC-Meth and ChIP-ATAC). As for the single state analysis, methylation status contributed to the enrichment within chromatin sets (LMRs or hypomethylated enhancers were significantly enriched across all chromatin sets), however, is smaller compared to the ATAC-seq data.

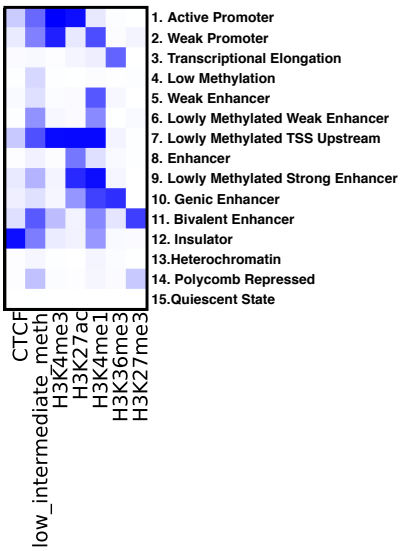
To formally evaluate the contribution of ATAC-seq and methylation information in T2D GWAS signal enrichment, another maximum likelihood model using FGWAS was determined. Specifically, nested FGWAS T2D enrichment models which combine ChIP-only chromatin states with raw ATAC-seq peaks and LMRs were created. The best model is shown in Figure 5.8 that included gene enhancer (log2FE=2.4, CI=1.14-3.3), strong enhancer (FE=1.3, CI=0.5-2.0), ATAC-seq peaks (FE=2.7, CI=2.0-3.4) and LMRs (FE=1.3, CI=0.4-2.0). Using a likelihood ratio test (LRT), the contribution of ATAC-seq and methylation information was determined by comparing the maximum likelihood estimate of the full model to models without LMRs or ATAC-seq peaks. Consistent with the single state analysis, it was found that both factors significantly improved the model based on the maximum likelihood (LRT ATAC-peaks included $P=6.6x10^{-7}$, LRT LMRs include P=0.02 compared to ChIP-only model), however, the ATAC-seq peak annotation had a much larger effect on the maximum likelihood than the LMRs (Fig5.8B). These results highlight the usefulness of integrating histone marks with both DNA accessibility and methylation regulatory states to refine chromatin states and facilitate understanding of T2D GWAS loci.



(A) ChIP-only states



(B) ChIP-ATAC states



(C) ChIP-Meth states

Figure 5.5: Comparing the effect of DNA methylation and ATAC-seq on chromatin state definitions. (A) 11 ChIP only chromatin states defined from histone marks and CTCF by chromHMM. (B) 15 ChIP-ATAC chromatin states defined from histone marks, CTCF and ATAC-seq data by chromHMM. (C) ChIP-Meth chromatin states defined from ChIP histone marks, CTCF and Methylation data by chromHMM. The colour is based on the chromHMM emission parameters and a darker colour indicates a higher frequency of a mark at a given state.

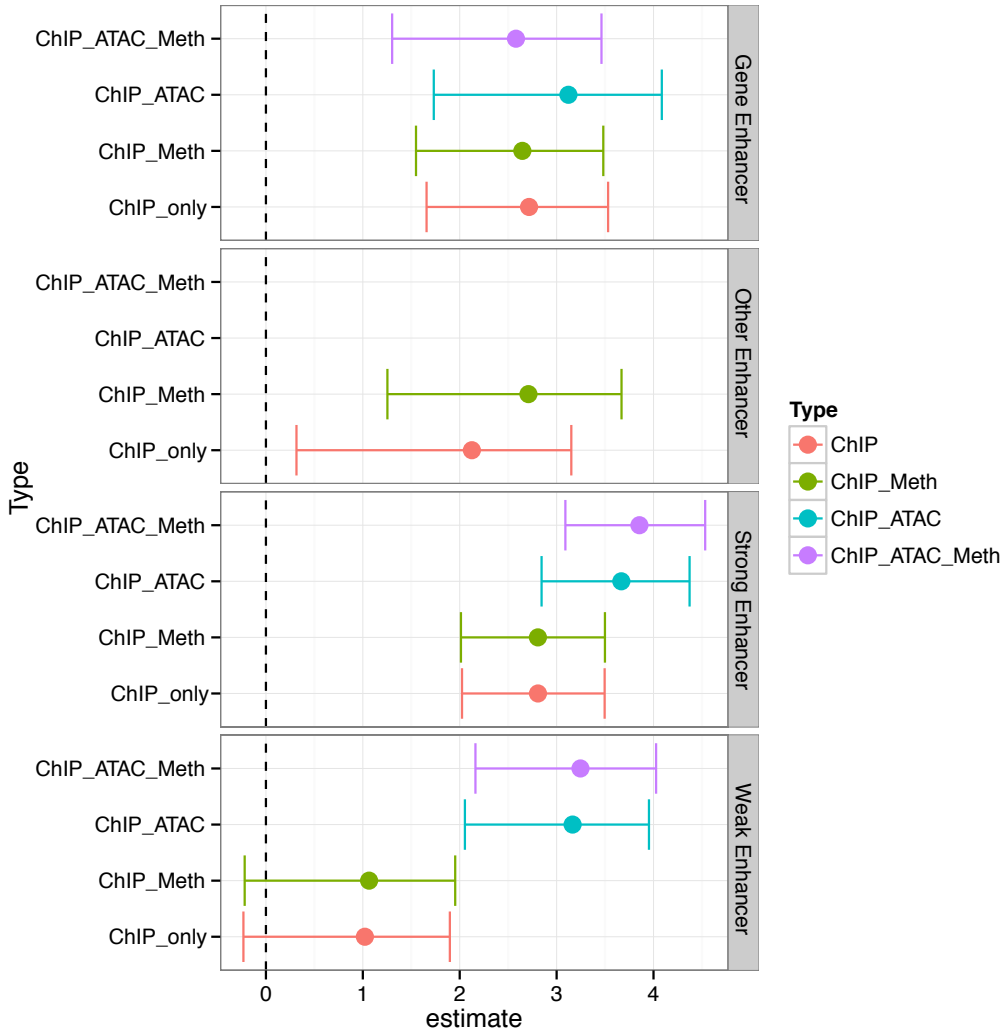


Figure 5.6: Comparing the enrichment in T2D GWAS regions of individual enhancer states defined from different epigenomic marks . log2FE (x-axis) for different enhancer states (grey panels) defined from different combinations of epigenetic marks (y-axis) including ChIP-ATAC-Meth, ChIP-ATAC, ChIP-Meth and ChIP-only. Enhancers are defined as follows: Strong enhancers are marked by both H3K4me1 and H3K27ac, weak Enhancers are defined by H3K4me1 only, gene enhancers are marked by H3K4me1 and H3K36me3, other enhancers are marked by H3K4me1, H3K4me3 and H3K27ac and are often referred to as TSS upstream regions (only included in the best FGWAS T2D model for ChIP-only and ChIP-Meth chromatin states).

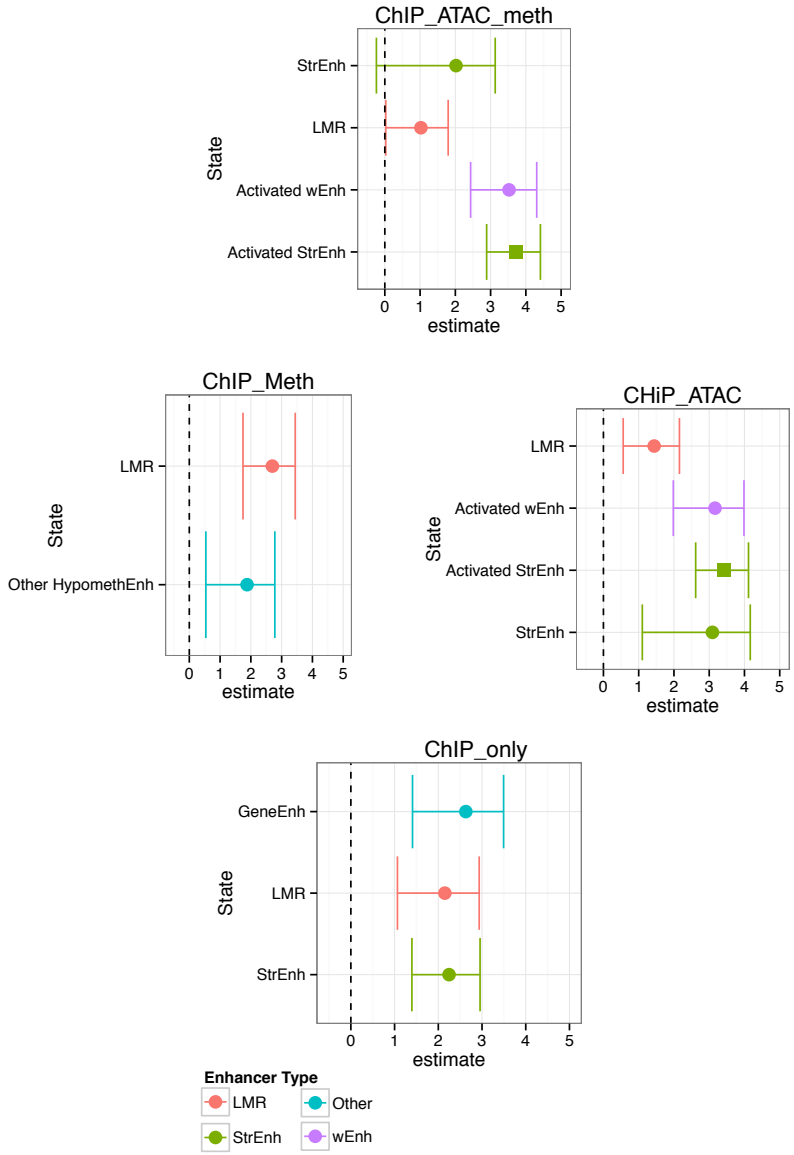
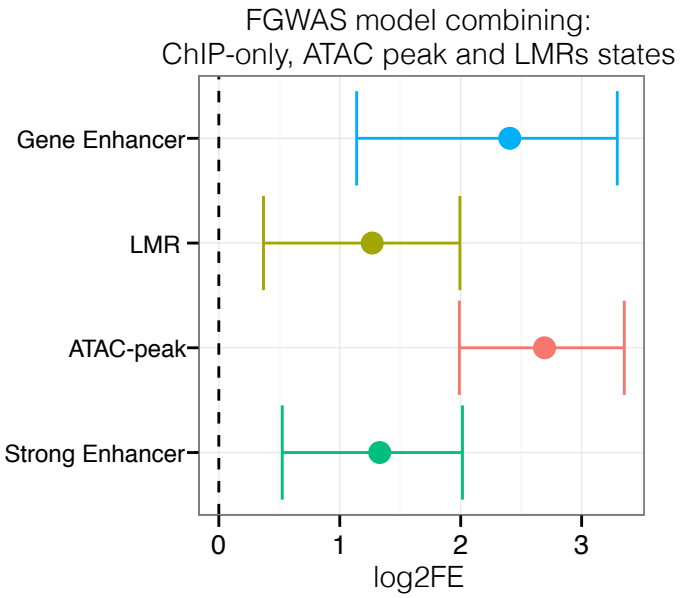
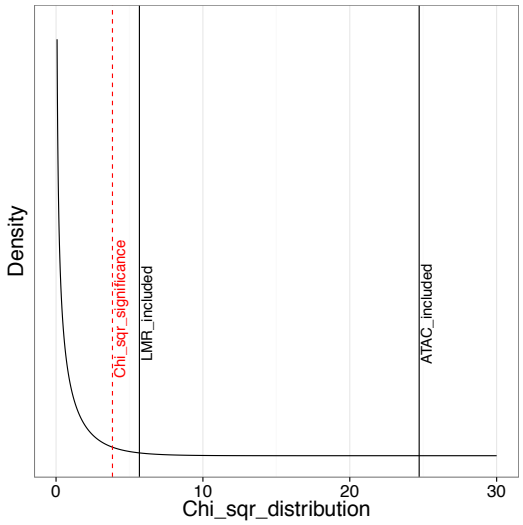


Figure 5.7: Comparing the enrichment in T2D GWAS regions of combined enhancer state FGWAS maximum likelihood models defined from different epigenomic marks. log2FE (x-axis) of enhancer state (y-axis) for different maximum likelihood FGWAS enrichment models including chromatin states from ChIP-ATAC-Meth, ChIP-ATAC, ChIP-Meth and ChIP-only. Enhancers are defined as follows: Strong enhancers are marked by both H3K4me1 and H3K27ac, weak enhancers are defined by H3K4me1 only, gene enhancers are marked by H3K4me1 and H3K36me3 and other enhancers are marked by H3K4me1, H3K4me3 and H3K27ac and are often referred to as TSS upstream regions. Activated strong and weak enhancers are additionally marked by open chromatin (and hypomethylation if available).



(A) Comparison combined models



(B) Comparison combined models

Figure 5.8: Effect of ATAC-seq and DNA methylation on T2D GWAS enrichment using a nested FGWAS model. (A) Maximum likelihood FGWAS nested model combining ChIP-only, ATAC-peaks and LMR states (y-axis) showing log2FE enrichment (x-axis).(B) Chi-square distribution with the indicated results of a maximum likelihood ratio test based on the maximum likelihood difference between a model including LMRs or ATAC-seq peaks compared to the ChIP-only model. The dashed line indicates significance (P-value <0.05).

5.3.4 Integrated epigenomic data facilitates identification of likely causal variants at T2D GWAS loci through allelic imbalance

Detection of global enrichment signals provides a basis for evaluating both the significance and functional impact of associated variants in GWAS regions, allowing variants that map into annotations that show global enrichment to be afforded extra weight. Hence, FGWAS was used to estimate PP, which were based on both association statistic and overlap of functional regulatory elements, to identify, at individual loci, potentially causal variants. The approach of performing global enrichment analysis of genomic annotations followed by identifying regulatory mechanisms at individual loci parallels the efforts by Gaulton et al 2015[83] that successfully analysed the role of islet ChIP-seq TFBS in T2D susceptibility.

In total, 49 significantly associated T2D regions were identified based on standard genome-wide significance ($P < 5x10^{-8}$) or having a regional reweighted FGWAS PP>0.9 (the latter threshold is based on the data from Pickrell et al 2014[158] who showed that this approach can be used to identify GWAS loci that did not pass standard genome-wide significance). Variants at these significant regions were prioritised for testing them for allelic imbalance in ATAC-seq open chromatin based on the following criteria. Variants were selected for testing of allelic imbalance if they had a high variant PP (reweighted PP >0.10) and overlapped with any of the 4 enriched regulatory states (activated strong and weak enhancer, strong enhancer and LMRs) that were included in the FGWAS T2D GWAS model (Fig5.4B).

In total, 46 high PP (PP>0.10) variants at 28 significant T2D GWAS regions (GWAS significance $P < 5x10^{-8}$ or regional reweighted FGWAS PP>0.9) were identified for testing of allelic imbalance (Fig5.9).

After further filtering based on the number of heterozygous samples (>2) and ATAC-seq read depth (depth>9 and at least 5 reads for each allele), 11 variants were tested for allelic imbalance in open chromatin as well as overlap for TF motifs. In total, 6 variants with significant allelic imbalance were identified (FDR <0.05) of which 3 remained significant after correcting for mapping bias using WASP (Table 5.1) suggesting that the others represented spurious results due to technical artefacts. For instance, at the *ADCY5* locus this approach identified an activated weak enhancer with

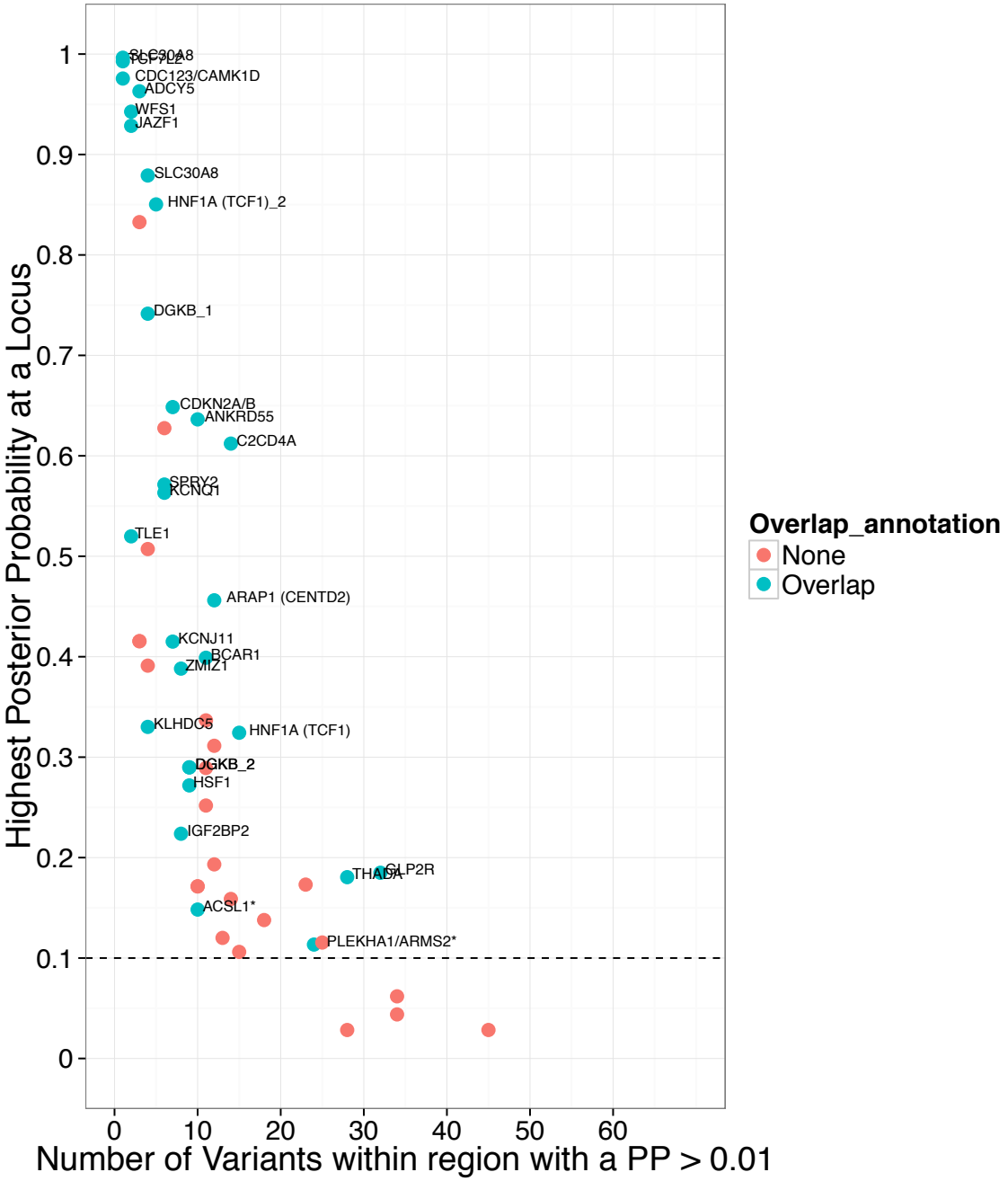


Figure 5.9: Identification of T2D GWAS loci and variants based on enhancer enrichment for testing of allelic imbalance in open chromatin. Identification of T2D GWAS loci and variants enriched for enhancer chromatin states using FGWAS PP. Per locus the highest PP variant is shown (y-axis) and the number of variants with PP >0.01 (x-axis). Loci with high PP variants (min PP >0.1, dashed horizontal line) that overlap one of the enhancer states or LMRs (green) are highlighted and the high PP variants were tested for allelic imbalance in open chromatin.

decreased chromatin accessibility (unadjusted binomial P=8.24x10^{-4}, Fig5.10A) for the rs11708067 risk A allele that is predicted to negatively affect HNF4A transcription factor binding by altering an overlapping HNF4A binding motif (Fig5.10B). The reduction in open chromatin and in HNF4A binding by the risk A allele is consistent with its previously identified effect on increased *ADCY* gene body methylation[99] (Fig5.10C) and decreased *ADCY5* transcript expression[227]. These findings combine previously and newly discovered regulatory effects at the *ADCY5* GWAS lead variant rs11708067 and provide a potential molecular mechanism of how the risk allele modifies a TFBS motif within an enhancer which in turn affects DNA accessibility, DNA methylation and gene expression to influence T2D susceptibility. Allelic imbalance in open chromatin was also found at the *KLHDC5* rs10842991 (unadjusted binomial=2.50x10^{-4}) and *CDC123* rs11257655 (unadjusted binomial P=5.52x10^{-8}) variants that overlapped integrated enhancer regions (Table 5.1). Consistent with higher chromatin accessibility the risk allele at the *CDC123* variant was previously associated with both enhanced binding of the important islet TF FOXA and increased expression of *CAMKD1*[264].

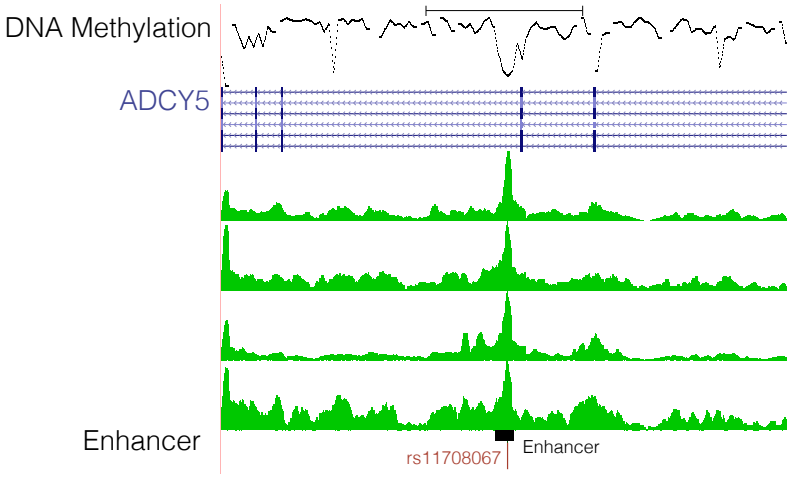
To highlight the usefulness of annotating GWAS regions with integrated hypomethylated enhancer, the remaining analysis focused on the *KLHDC5* GWAS locus which only passed the PP significance threshold (PP >0.9 but P <5x10^{-8}). At this locus, no distinct lead variant can be identified using fine mapping due to multiple variants (n=10) in high LD (r2=1) with the lead variant (based on 1000G data[262]). It was observed that the top credible variant rs10771372 (DIAGRAM PP based only on genetic data=0.05) and a neighbouring variant rs10842991 (PP=0.03) in full LD (r2=1) overlap a highly-accessible strong enhancer regions (Fig5.11A) which increased the reweighted PP for both rs10771372 (PP=0.32) and rs10842991 (PP=0.20). Thus, these variants were tested for allelic imbalance in open chromatin and it was found that only the risk C allele of the rs10842991 variant showed increased chromatin accessibility (binomial P=2.50x10^{-4}) at the chromatin state enhancer region. The rs10842991 variant overlapped a FIMO predicted PAX6 TFBS motif and the risk C allele was predicted to enhance PAX6 transcription factor binding (Fig5.11B). The absence of enhancers at this rs10842991 variant in any other Epigenome Roadmap tissue (according to haploreg v3[265]) strongly suggested that this regulatory effect may be islet-specific.

These data prioritised the *ADCY5* variant rs11708067 and *KLHDC5* variant rs10842991 for functional downstream analysis and suggested a potential molecular mechanism at these loci involving

modified transcription factor binding that may affect islet function and T2D susceptibility.

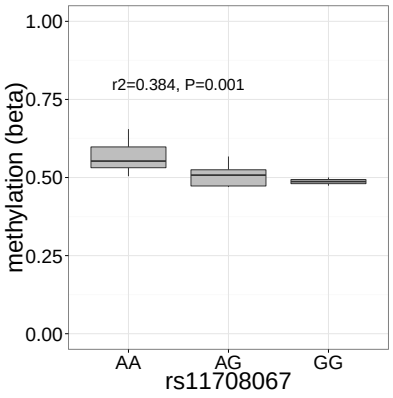
SNP	Locus	GWAS P	FGWAS T2D PP	AI effect size	AI P (WASP)	Direction of Effect
rs11708067	*ADCY5*	8.8x10-13	0.96	0.24 (11 A VS 34 G alleles)	8.24x10-4	risk allele A closed
rs11257655	*CDC123*	4.0x10-8	0.98	0.38 (204 C VS 330 T alleles)	5.52x10-8	risk allele T open
rs10842991	*KLHDC5*	7.3x10-7	0.20	0.60 (225 C VS 153 T alleles)	2.50x10-4	risk allele C open

Table 5.1: Variants with allelic imbalance in ATAC-seq open chromatin. For each variant, the ID, the Locus, the T2D GWAS P-value, the FGWAS T2D Posterior Probability (PP), the Allelic imbalance effect (AI effect size), the AI P-value after removing mapping bias with WASP and the direction of effect in terms of T2D risk is shown.



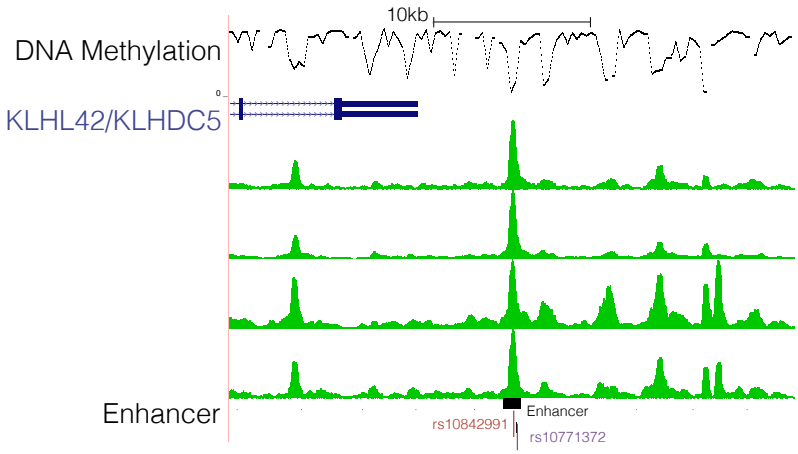
(A) ADCY5 Genome Browser view

(B) HNF4A motif

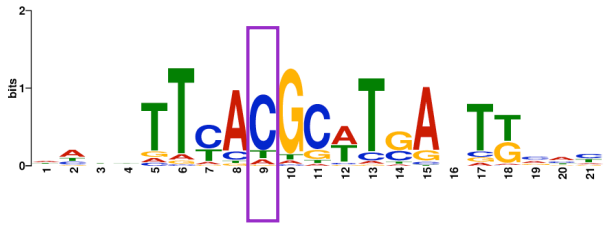


(C) ADCY5 mQTL

Figure 5.10: ADCY5 rs11708067 regulatory effects (A)-(C) ADCY5 rs11708067 lead variant overlapped a highly-accessible weak enhancer region with low methylation and open chromatin as shown in the genome browser view (A) and the risk A/T allele was associated with reduced open chromatin, a disrupted HNF4A binding site (the altered site is highlighted by a purple box on the opposite strand) (B) and was associated with increased methylation levels (y-axis, while genotypes are shown on the x-axis) (C).



(A) KLHDC5 Genome Browser view



(B) KLDHC5 PAX6 motif

Figure 5.11: KLDHC5 rs10842991 regulatory effect (A)-(B) *KLHDC5* rs10842991 overlaps a highly accessible weak enhancer region with low methylation and open chromatin (A) and shows allelic imbalance in islet ATAC-seq data which disrupts a PAX6 binding site (B). In addition another high PP variant (rs10771372) is shown that also overlaps the enhancer region but did not show allelic imbalance in open chromatin)

5.4 Discussion

The main results of this chapter are:

- islet ATAC-seq open chromatin and DNA methylation data help to refine islet chromatin states leading to the identification of strong and weak enhancer subsets
- activated enhancer defined by open chromatin and hypomethylation showed stronger enrichment in T2D and FG GWAS loci
- increased enrichment was driven by both ATAC-seq open chromatin and methylation status, however, open chromatin data had a substantially larger effect on GWAS enrichment compared to methylation
- allelic imbalance in open chromatin was found at 3 high PP variants at the T2D GWAS loci *ADCY5, CDC123* and *KLHDC5* potentially highlighting novel molecular mechanisms governing T2D risk at these loci.

The key challenge of understanding GWAS signals is to identify causal variants in regulatory DNA elements to prioritise variants for downstream analysis and determine functional mechanisms governing the effects on the phenotype. This understanding will be critical to use human genetics to characterise the biological networks responsible for disease pathogenesis and identify novel therapeutic approaches. The results from this chapter demonstrate that the integration of epigenomic islet annotations provides an effective way of refining islet regulatory states and highlighting potential molecular mechanisms at T2D and FG GWAS loci. The analysis showed that integrated chromatin enhancer states showed significantly higher enrichment in T2D and FG GWAS signals compared to chromatin states defined from histone marks only. Investigating the contribution to the enrichment in GWAS signal for DNA methylation and ATAC-seq DNA accessibility data separately showed that DNA accessibility adds significantly more to the overall enrichment compared to DNA methylation. However, the strongest enrichment was observed for chromatin enhancer states combining all genomic annotations including histone marks, DNA methylation and open chromatin, thus, indicating the advantage of an integrated analysis approach. Also, short and well-defined LMRs were found to further improve the enrichment model despite hypomethylation being

included in the chromHMM analysis. Thus, this data suggests that LMR may add additional information to the model that is not captured by any of the chromatin states used in the model.

The T2D enrichment model derived from refined chromatin states was highly useful for increasing the understanding of islet regulation and its role in T2D development. This is consistent with previous studies that also observed increased enrichment in GWAS signals for enhancer chromatin states that were combined with additional genomic annotations such as DNA accessibility or DNA methylation. For instance, Wang et al 2016 observed that enhancer chromatin states overlapping open chromatin and hypomethylated regions showed increased enrichment in GWAS loci associated with cardiac intervals and could be used to identify potential sub-threshold GWAS hits[131]. Similarly, increased enhancer enrichment in T2D GWAS loci was observed by Pasquali et al 2014 when combining open chromatin information and islet histone marks[85].

In addition, the enrichment model provided an opportunity to prioritise likely causal variants for testing allelic imbalance. After correcting for technical artefacts related to read bias using WASP[261], this approach identified 3 enhancer variants with significant allelic imbalance including *ADCY5*, *CDC123* and *KLHDC5*. At the *ADCY5* locus, the results suggest a potential functional mechanism mediating the previously described effect of the lead variant rs11708067 on T2D disease pathology[55,266]. The intronic rs11708067 risk A allele found in a highly-accessible weak enhancer region reduces binding of the HNF4A transcription factor which then may mediate the previously reported changes in gene expression[159] and methylation[99] of the *ADCY5* gene that is critical for coupling glucose to insulin secretion[267]. Similarly, at the *KLHDC5* locus, a credible set variant rs10842991 was determined that overlaps a differentially methylated and likely islet-specific highly-accessible strong enhancer region and shows allelic imbalance with the risk C allele increasing PAX6 transcription factor binding. Hence, this data suggests potential molecular mechanisms at these loci that may help understand T2D disease susceptibility.

While the data in this chapter has provided a complex integrated set of regulatory domains that is useful for understanding T2D-related GWAS signals, these regulatory annotations were derived from a rather modest number of islet samples from healthy donors due to the limited availability of primary human islets. Thus, the extent of variation in the epigenomic factors interrogated in this study is limited and it is not clear how this variation could directly affect islet regulation, partic-

ularly, during disease state. Also, the directional causality between these integrated genetic and epigenomic factors remains unclear. This is because previous studies have shown that the causal relationship between methylation changes and other molecular phenotypes such as gene expression is complicated with both active and passive relationships [101]. Future studies with a larger number of human islet sample including ones from healthy and diabetic individuals could enhance understanding of the causal interactions between DNA methylation, chromatin state, expression and T2D susceptibility. Specifically, an increase in sample number could be used to perform whole-genome islet chromatin QTL studies to highlight relevant functional regulatory mechanisms as has been done for eQTLs [130,159,268], mQTLs [96,97,99] and chromatin QTLs [124,269] in islets and/or other tissues. In addition, further insight into islet regulation could be provided by studies investigating disease-associated chromatin changes and the effect of environmental factors such as metabolites or pharmaceutical agents on the chromatin landscape. Our lab is currently increasing the number of available islet samples to facilitate such studies in the future.

In conclusion, the data in this chapter provided a comprehensive analysis of a larger number of islet epigenomic factors to accurately define islet regulatory elements and function. These regulatory annotations showed enrichment in T2D and FG GWAS regions and were highly useful to identify variants at T2D and FG GWAS loci that may govern T2D susceptibility.

CHAPTER 6

Identification of coding CNVs and their role in T2D development

6.1 Introduction to CNVs

6.1.1 The role of coding CNVs and PTV variants in T2D development

Structural variants (SVs) including CNVs represent large deletions, duplications or inversions in the genome which depending on the location may have severe effects on the phenotype. In non-coding regions, CNVs may alter the activity of regulatory elements, thereby influencing the expression of nearby genes while CNVs in coding regions may affect expression levels of dosage-sensitive genes, disrupt gene function, induce gene fusion events or unmask the effect of recessive alleles[270].

In accordance with their potential molecular effects, CNVs have been associated with several monogenic forms of disease. In addition, it was shown that CNVs also influence genetic susceptibility

of common diseases, may help explain missing heritability and provide potential causal mechanisms at known GWAS loci[150,270]. In 2010 an extensive study with 16,000 cases and 3000 controls was performed, investigating the role of common CNVs across eight complex diseases (including T2D) using a customised array[152]. While the authors investigated about 3.5k common CNVs (MAF >5%) representing about 50% of all common CNVs (above a size of 500bp) and estimated to have a power of 80% to detect any association with an OR above 1.4, only a handful of significant associations was found. For instance, for T2D a single CNV overlapping the gene *TSPAN8* was significantly associated with T2D in a region that was observed to be associated with T2D in SNP-based GWAS[271], thus, potentially providing a potential causal mechanism at this locus. However, the most associated SNP was only weakly correlated with the CNV (LD r^2=0.17 under the assumption of a bi-allelic CNV effect) which may indicate that the CNV is only in weak correlation with the true causal variant.

The study also found that most common variants are sufficiently well-tagged by common SNPs used in GWAS which suggests that effects of common CNVs on complex disease risk would have already been detected indirectly through SNP-based GWAS. Based on these results the authors concluded that common bi-allelic CNVs do not contribute significantly to complex disease risk[152]. However, the study was not able to determine the role of rare CNVs, in particular ones affecting the coding region and acting as potential PTVs or PDVs, in complex disease risk. As described in Chapter 1 Section 1.4.4, PTVs and PDVs refer to variants that alter protein function and are often referred to as LoF or GoF variants. However, this terminology does not necessarily accurately describe the effect of CNVs and other variants on the gene and protein level, hence, for coding CNVs the term PTVs and PDVs is used instead in this thesis.

Rare variants with strong effects including PTVs have also been suggested by some to explain missing heritability of complex diseases[150,272]. For instance, recent studies provide evidence that rare mutations or CNVs[272,273] cause a considerable number of autism cases. In contrast to autism, the results from a recent large-scale study investigating more than 100k samples did not find evidence that rare variants have a major role in T2D genetic susceptibility.While these data indicate that rare variants do not contribute substantially to the missing heritability, rare strong effect size variants may still be relevant and provide interesting insight into disease biology[64]. In particular, the discov-

ery of protective PTVs variants, which disrupt gene function, is of interest because these variants may facilitate drug development by highlighting *in vivo* validated drug targets. There are several examples of SNVs that represent rare PTVs of value in the field of T2D:

- ***SLC30A8***: in a cohort of approximately 150,000 individuals a highly T2D protective and rare PTV, which reduced the risk of T2D by 65%, was found in the gene *SLC30A8* encoding the Zinc transporter ZnT8 that is crucial for insulin secretion in pancreatic beta-cells[274]. While previous functional studies (in mice) provided conflicting evidence on the role of *SLC30A8* in T2D risk[275–281], this study showed how the use of large-scale genetic studies could improve understanding of T2D pathophysiology and provide potential therapeutic drug targets.

- ***TBC1D4***: in population isolates, which have extended LD patterns and higher chances of accumulating deleterious variants due to genetic drift, a Greenlandic PTV variant in the gene *TBC1D4* was observed that drastically increased the risk of T2D development (OR=10.3 which is much larger than any other T2D GWAS variant identified) and was extremely rare in other populations. Functional analysis indicated that the variant led to decreased GSIS, reduced glucose tolerance and T2D development via a reduction in TBC1D4 and GLUT4 levels in skeletal muscle[282].

- ***CDKN2A***: based on oral glucose tolerance tests in a small patient cohort and functional studies in a human beta-cell line model, rare PTVs of the gene *CDKN2A*, which is present in a region previously associated with T2D in GWAS, was found to affect insulin secretion and led to insulin resistance; however, the exact mechanisms causing insulin resistance remain unclear since the variant was too rare to recruit enough individuals for additional clinical tests[283].

Overall the results from these studies clearly show the benefit of determining the role PTVs in T2D biology, however, they also indicate the need for novel approaches to study these variants since characterising these rare variants in the general population requires either extremely large datasets, the aggregation of rare variants across genes or population isolates.

6.1.2 Detection methods for CNVs

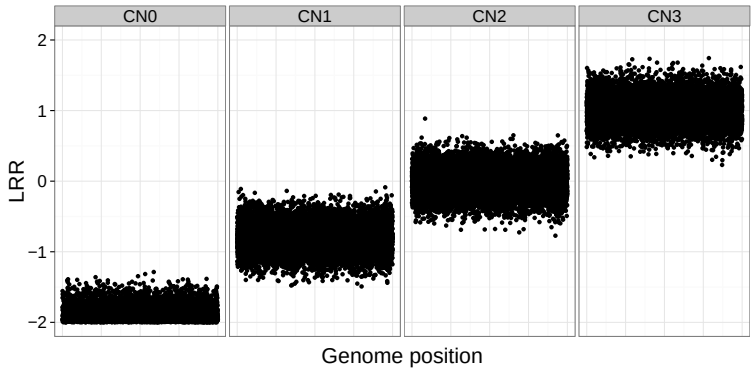
In addition to sample size restrictions, the accurate detection of rare PTVs and CNVs acting as PTVs or PDVs still remains challenging. Strong selective pressure is expected to act on most PTVs leading to decreased variation at such sites compared to the rest of the genome. In contrast, the expected level of sequencing errors should remain the same throughout, thus leading to a potential enrichment of false-positives at these sites of reduced variation[284].Detection of CNV is particularly difficult since no gold standards exist to validate the identified CNVs accurately. Several methods have been developed to determine copy number state throughout the genome. Cytogenetics using microscopy was among the earliest approaches to detect large deletions or duplications affecting chromosomal structure including trisomy 21.

6.1.2.1 CNV detection using microarray technology

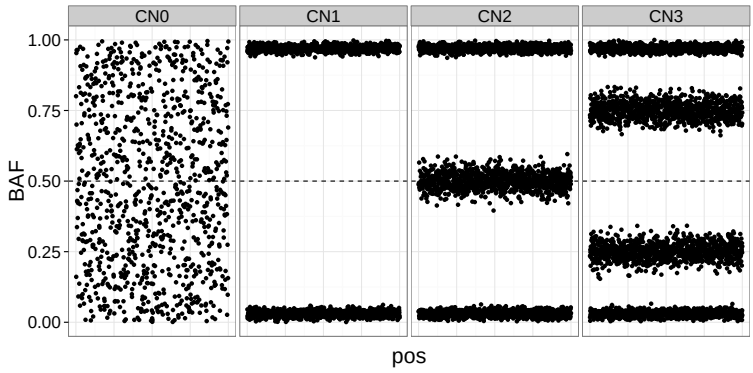
Since cytogenetics is limited in resolution, microarray technologies were developed that were based on comparative genomic hybridisation (CGH)[285]. The basic principle of CHG arrays is based on DNA probes which, depending on the type of array vary in size from less than 100bp up to several 100kb, cover large parts of the genome and are fixed upon glass slides. Then test and references DNA labelled with different fluorescence markers are hybridised to the probes and the fluorescence intensity is measured throughout the genome. Increases or decreases in the fluorescence signal of the test DNA compared to the reference DNA (assumed to be of normal copy number state) represent duplications or deletions in the genome, respectively. While CGH arrays have improved the discovery of CNVs, the resolution is still limited to deletions and duplications ranging from 40kb up to Megabases and higher resolution CGH arrays generally suffer from a poor signal to noise ratio leading to an increased number of false-positive results[286].

Advances in SNP genotyping arrays that nowadays provide relatively high-density coverage throughout the genome have led to the development of analysis methods to detect CNVs from SNP arrays. These methods rely on the normalised intensity measurement of the two alleles (defined as A and B) of a SNP which can be used to determine the Log R Ratio (LRR) representing the overall probe signal intensity as well as the B Allele Frequency (BAF) which denotes the intensity

ratio of the two alleles (Fig 6.1). Similar to CGH intensity measurements, the LRR corresponds to overall copy number state while the BAF provides an additional measurement indicating unusual patterns in the genotype distribution. Assuming a diploid copy number state, the BAF is expected to fall into one of three genotype categories (AA, AB, BB) while a deletion reduces the genotype classes (A or B) and a duplication increases the genotype classes (AAA, AAB, ABB, BBB), thus providing further information on the copy number state in a given region. Various analytical tools taking into account the measurements of LRR, BAF or both have been developed and successfully used to investigate genome-wide copy number states. Although SNP arrays provide a cost effective solution for CNV genotyping, the resolution of these arrays is still limited (10-40kb) since they typically rely on the distribution of SNP probes which were selected to tag haplotypes throughout the genome for GWAS but not CNV detection.[286,287].



(A) CNV-LRR-detection



(B) CNV-BAF-detection

Figure 6.1: Simulated Log R Ratio (LRR) and B Allele Frequency (BAF) plots for different Copy Number (CN0-CN3) states. Each data point refers to a single SNP. (A) The LRR representing total normalised signal intensity at a given SNP which in a normal diploid state clusters around 0. An LRR below 0 is associated with a deletion (CN1 or CN0) and an LRR above 0 is indicative of a duplication (CN3). (B) The BAF ranges from 0 to 1 and the number of BAF clusters, which represents the relative signal intensity of one SNP allele relative to the other allele, correlates with the number of possible genotypes and can be used to predict CN. Three clusters for each genotype class are expected in a diploid individual while less than 3 clusters are associated with deletions (CN0 and 1) and more than 3 clusters suggest a duplication (CN3).

6.1.2.2 CNV detection using Next-generation sequencing data

Due to reduced sequencing cost and increasing number of NGS datasets, several methods have been developed to identify CNVs from short reads (Fig6.2). The read pair (RP) method relies on significant deviations from the average insert size of the paired reads distribution which may be indicative of changes in copy number; a smaller insert size is indicative of deletions while a larger insert size suggests a duplication (Fig6.2A). However, this RP method is limited to the detection of medium-sized or large deletions considering that the insert size distribution of paired reads will be variable to some extent. In addition to using the insert size, paired reads can also provide potential evidence for changes in copy number through split reads (SR); that is, paired reads of which only one read maps properly while the second read fails to map in the genome (Fig6.2B). For smaller insertions and deletions, SP can provide valuable information about the breakpoint location; however, larger CNVs often can not be detected using this approach. While both RP and SR methods can determine changes in copy number state in the genome, they often can not accurately quantify the actual copy number. A Read Depth (RD) method has been developed that makes use of the correlation between regional depth and copy number state which allows for both detection and quantification of CNVs. In a set of samples, the normalised average read depth (similar to signal intensity measurements derived from microarray data) across windows in the genome is determined and a segmentation algorithm is applied to characterise the copy number state of each sample by comparing the sample read depth to the average population read depth[288,289].

Whole-genome sequencing provides the most comprehensive approach to analyse CNVs throughout the genome; however, the high-depth required for CNV detection makes this very costly. In contrast, whole-exome sequencing datasets, which have been used to identify exonic CNVs, provide relatively high-depth coverage of coding regions. Although exome capture methods may lead to variable RD across exonic regions, most tools for the discovery of CNVs from Whole-Exome Sequencing (WES) apply the RD approach, since it is independent of the location of the CNV breakpoints that may be outside of exonic regions.[290]

Despite the advances in detecting CNVs, it still remains very challenging. Considering these difficulties in accurately identifying CNVs, determining associations between copy number status and disease is even more difficult[288,290]. Over the years several studies have reported CNVs as-

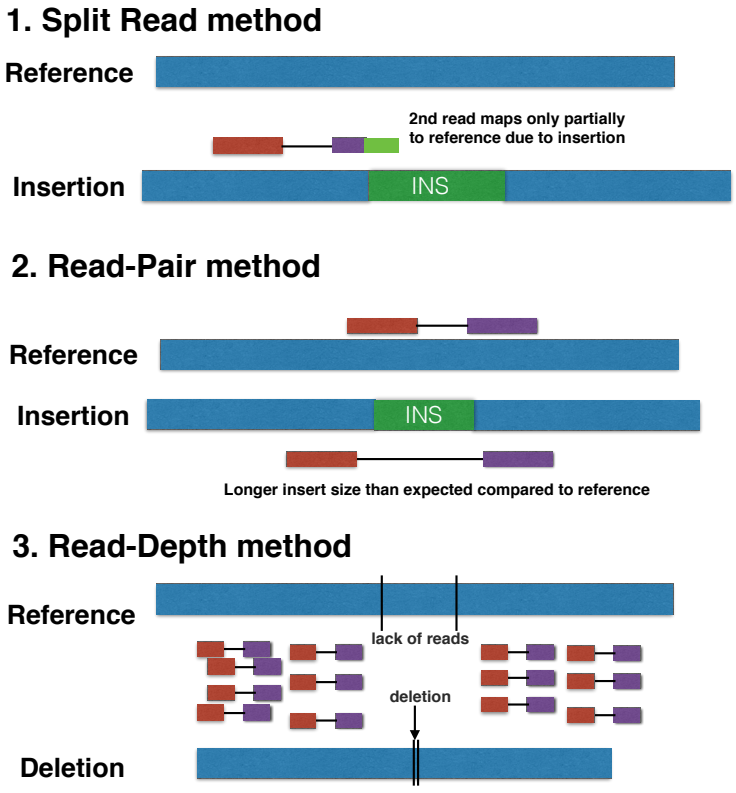


Figure 6.2: Approaches to identify deletions from WGS or WES. (1) The Split Read (SR) method identifies CNVs based on paired-end read-pairs in which one read does not or only partially align to the reference (an insertion example is shown). (2) The Read-Pair (RP) approach determines CNV regions based on changes in the insert size distribution. If an insertion/duplication is present the expected insert sizes of paired reads will increase while the presence of a deletion will decrease the insert size. (3) The RD approach uses differences in read depth to determine deletions (lower than expected read number, example shown) or duplications (higher than expected read number).

sociated with certain phenotypes; however, often these results could not be validated indicating a high false-positive rate. For instance, a recent study aimed to validate the association of 18 previously reported structural variants with obesity and despite estimated 80% of power to detect a significant association for most CNVs, the association of only one variant with obesity could be replicated[291]. Similarly, results at the *AMY1* locus which was suggested to be extremely strongly correlated with obesity[292], could not be replicated and likely represented technical errors related to CNV genotyping[293].

The research in this Chapter aims to evaluate the role of rare coding CNVs in the genetic susceptibility of T2D. For this purpose, coding CNVs were derived from large-scale exome sequencing and exome array datasets that were initially generated to study the role of rare and low-frequency SNVs in the genetic architecture of T2D. Specifically, the chapter is divided into two mains parts with the following aims:

- to characterise the role of rare coding CNVs, generated (by members of the ExAC consortium) from WES data, in T2D and BMI genetic susceptibility through association testing
- to develop a novel pipeline for the detection of rare coding deletions from large-scale exome chip array data and test the identified deletions for association with T2D.

6.2 Methods

6.2.1 CNV calls derived from ExAC exome-sequencing data

As described in Chapter 2 Section 2.5.2.1, ExAC is a consortium that aims to collect large-scale exome sequencing data from different studies and process it in a standardised way to generate summary level data for scientists to use.

6.2.1.1 ExAC samples used for CNV detection

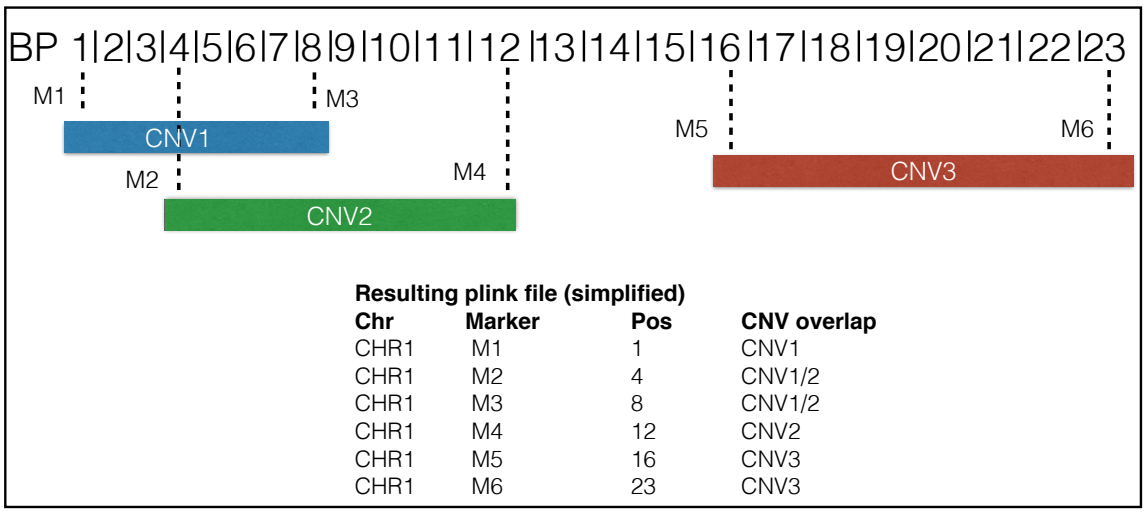
The samples used in this Chapter represent a subset of 14.5k of the 60k ExAC consortium samples. These include the following T2D-case-control studies: T2D-Genes (8980 individuals), GoT2D (2502 individuals) and SIGMA (3845 individuals) which combine genetic data from 5 different ancestry groups including African-American, East Asian, South Asian, Hispanics and European (Table 6.1). These WES samples were initially collected to characterise the role of genetic variation in T2D susceptibility and at a later point processed by ExAC for CNV calling. More information about the individual studies and samples is provided in Chapter 2 Section 2.5.2.1.

6.2.1.2 ExAC CNV detection and filtering

For this thesis, members of the ExAC consortium (Dr Douglas Ruderfer) have provided CNV calls from the aforementioned 14.5k T2D-case-control samples. CNVs were jointly called by the ExAC consortium using XHMM[294] as part of the 60k ExAC consortium samples which is described in the ExAC paper in detail[157]. In short, the GATK[295] DepthofCoverage function was used for calculating mean coverage per base across 220k exonic targets defined as Illumina ICE v1 targets plus additionally all coding regions in GENCODE v19[296] and after scaling the resulting matrix, normalisation was performed by correcting for 388 PCS to avoid batch effects and other biases. The CNVs were then called using the XHMM Viterbi algorithm. CNV calls were received in plink[297] .cnv format and .fam format from the ExAC consortium via sftp download from servers of the Broad Institute, Cambridge, USA. Once the CNV calls were provided by the ExAC consortium, the remaining work described in this chapter was performed by myself.

6.2.1.3 Filtering and processing of ExAC CNV calls

As suggested by the ExAC and XHMM authors[157,294], CNVs were filtered based on a CNV score of 60, which represents a PHRED-like quality score, and separated deletions and duplication calls using plink (v1.07)[297]. Plink "cnv-make-map" function was used to determine CNV surrogate markers based on start and stop positions of CNVs (Fig 6.3) to determine AC estimates.



Resulting plink file (simplified)

Chr	Marker	Pos	CNV overlap
CHR1	M1	1	CNV1
CHR1	M2	4	CNV1/2
CHR1	M3	8	CNV1/2
CHR1	M4	12	CNV2
CHR1	M5	16	CNV3
CHR1	M6	23	CNV3

Figure 6.3: Plink approach to identify surrogate markers corresponding to start and stop sites of CNVs. Three CNVs (blue, green and red) are shown across 23bp. Plink uses surrogate markers to define the overlap between CNVs based on start and stop positions of CNVs. 3 markers are identified for CNV1 (M1 at bp position 1, M2 at bp position 4 and M3 at bp position 8) and CNV2 (M2 at bp position 4, M3 at bp position 8 and M4 at bp position 12) while only 2 markers are identified for CNV3 (M6 at bp position 16 and M6 at bp position 23).

6.2.1.4 Association testing

To test for T2D and BMI association, the Firth Bias Corrected Test and the linear Wald Test, respectively, were used as implemented in the software "Efficient and Parallelizable Association Container Toolbox" (EPACTS v3.2.4,[298]). Deletions and duplications were (independently) collapsed based on gene overlap and the analysis of these gene-centred CNVs was performed for each study (and CNV type) separately. The Firth Bias-Correct test has been suggested to be highly suitable for rare variant tests[299] while the Wald test is frequently used in genetic association studies. Covariates included in the association analysis were sex, age, BMI (only for T2D) and PC1-4 to correct for population structure (PC1 for African American, East and South Asian samples while PC1-4 was used for European and T2D-Genes Hispanic samples. PC1-2 was used for Hispanic samples from SIGMA). PCs were derived from SNP-based exome sequencing data generated and provided by members of the relevant studies (Dr Anubha Mahajan provided PCs for T2D-Genes and GoT2D samples while Dr Josep Maria Mercader provided SIGMA PCs). The EPACTS association results were then meta-analysed using metal (version released on 2011-03-25[300]). QQplots and Manhattan plots were generated using the R package QQman[301].

6.2.1.5 Identifying gene-centred CNVs near publicly available GWAS signals

To provide further support for T2D- and BMI-associated CNVs, the overlap between associated CNVs and known T2D and BMI GWAS signals was determined, since a true CNV association signal would probably also be detected in previously studied variants that are in LD with the CNV.Similarly, Usher et al 2015 also tried to validate the AMY1 copy number association with BMI using SNP-based data (among several other approaches)[293].

To identify associated CNVs near GWAS signals T2D and BMI association summary level GWAS data was obtained from the DIAGRAM consortium (unpublished) and the ENGAGE GIANT consortium website (`http://diagram-consortium.org/2015_ENGAGE_1KG/`, more details about both the DIAGRAM and ENGAGE consortium are provided in Chapter 2 section 2.5.1). SNPs with suggestive genome-wide significance ($5x10^{-6}$) were identified. These SNPs were then overlapped with the CNV association signals using bedtools and any associated CNV within 100kb

of the SNP was reported. In addition, T2D-portal missense SNV T2D and BMI gene-burden test results were obtained, since a significant association between gene aggregated PTVs or missense variants, which likely affect gene function, also provides support for CNV-phenotype associations (particularly for deletions).

6.2.2 Detection of coding deletions from Exome chip array genotyping data

6.2.2.1 Exome chip samples

The Illumina Exome chip array[302] is a custom designed array covering more than 240,000 predominantly exonic SNPs identified in exome sequencing data of approximately 12,000 individuals (more information at `http://genome.sph.umich.edu/wiki/Exome_Chip_Design`). The exome chip array provides dense SNP coverage across exonic regions at low cost and high throughput and was initially developed as an intermediate solution between genotyping arrays, which investigate mainly common variants, and whole-exome sequencing studies that focus on coding variation across the whole-frequency spectrum. A successful CNV discovery pipeline could take advantage of the large-scale sample sets available for the exome chip and provide a cost-effective way of identifying rare coding CNVs.

I used exome chip genotyping data from 11,686 UK samples from 5 different studies (UKT2D, YDX, GoDARTs, Oxford Biobank, TwinsUK, further sample information is provided in Chapter 2 section 2.5.3) for the development of an exome chip deletion detection pipeline to identify coding deletions and test these variants for association with T2D (11,414 unrelated T2D-case-control samples were used for association testing). Pre-processing of the exome chip genotyping data was performed by members of each study individually.

6.2.2.2 In silico validation of deletion pipeline using GoT2D/UKT2D samples

To validate exome chip deletion calls, the exome chip CNV calls were compared to deletions identified from WGS and (at a later point) WES data. Specifically, a subset of approximately 650 exome chip samples (UKT2D study) were also whole-genome sequenced and whole-exome sequenced as part of the GoT2D project. This project applied a range of sequencing and genotyping technologies to efficiently query genetic variaiton across the whole genome with the aim to investigate lower-frequency variants and their role in T2D genetic susceptibility .

The GoT2D exome chip samples were used to validate the exome chip deletion discovery pipeline by initially comparing the exome chip deletions to deletion calls derived from WGS using GenomeSTRiP[303] (these WGS deletions were called by Ashish Kumar a member of the GoT2D consortium). After filtering WGS deletions based on overlap with at least 3 exome chip SNPs, in total 448 WGS deletions could be used for the initial validation. At a later point when the pipeline was already developed, ExAC consortium members (Dr Douglas Ruderfer) provided WES CNV calls from these same GoT2D samples (as part of the aforementioned ExAC study that investigated WES CNV calls in a total of 60k samples). After filtering, 427 high-quality (score >60) and rare (AF <1%) ExAC GoT2D WES deletion calls could be used for comparison to the exome chip deletion calls.

Specifically, validation for exome chip CNV calls in these GoT2D samples was performed in the following way:

1. the number of exome chip deletion regions was determined that were fully overlapping (nested within) a deletion call from the WES and WGS data

2. the genotyping specificity and sensitivity of the overlapping sites were calculated by determining the number of samples in which the genotypes from the exome chip CNV discovery pipeline were concordant with the GenomeSTRiP WGS and/or ExAC GoT2D WES data. As mentioned above, initially, when the pipeline was in development only WGS deletion calls were available and only at a later point could the ExAC GoT2D WES CNV calls be used for comparison.

6.2.2.3 Deletion discovery pipeline

Since it has been suggested to combine multiple copy number detection algorithms to increase CNV calling performance, in particular, to reduce false positive rates[304], three CNV calling algorithms (PennCNV[305], QuantiSNP[306] and Circular Binary Segmentation (CBS) algorithm[307]) were used for deletion discovery from the Illumina human exome chip array. Both PennCNV and QuantiSNP apply a Hidden Markov Model (HMM) for CNV detection using both SNP LRR and BAF while the CBS algorithm determines CNVs by segmenting the genome into copy number states using signal intensity values from the LRR ratio (Fig6.1 indicates how LRR and BAF can be used for CNV detection). Based on previous suggestions[304], this pipeline combines different approaches for CNV detection to take advantage of the different algorithms. The HMM-based algorithms are expected to be more specific considering the use of both the LRR and BAF for detection while the CBS approach is likely more sensitive.

The pipeline focussed initially on the discovery of deletions since only WGS deletion calls were available for validation when the pipeline was developed. In addition, deletions can be more accurately detected because lower copy number states are expected to have a larger influence on the LRR and BAF distribution (Fig6.1). Figure6.4 provides an overview of the pipeline.

6.2.2.3.1 Detection of deletions sites

Initially, the Hidden Markov Model-based algorithm PennCNV and QuantiSNP were used independently with standard settings, to predict copy number states using the SNP LRR and BAF. These measures were correlated to the total signal intensity and number of possible genotypes of a given SNP locus. As recommended by the authors[305,306], the data was corrected for GC-content to reduced periodicity in signal intensity that can influence CNV calls.

Samples with high variability (LRR Standard Deviation >0.24) and genomic periodicity (abs(Wavefactor) >0.04) in signal intensity were removed from the analysis. Deletion calls were then filtered based on size and the Log Bayes Factor (QuantiSNP) or the confidence scores (PennCNV). All CNV calls made by QuantiSNP with a Log Bayes Factor less than 30, which corresponds to a false-positive rate of <1%, were removed. Furthermore, small CNVs (<6,000bp) with a Log Bayes factor below

75 or a confidence score below 90 were also removed. The remaining filtered deletion calls were then merged together using bedtools to identify a unique set of calls. These thresholds were either chosen based on suggestions of the authors of PennCNV and QuantiSNP[305,306] or determined as most accurate based on comparison with the WGS deletion calls.

The merged calls were then intersected with call sets from each program individually using bedtools to identify deletion regions detected by both algorithms (Fig 6.4).

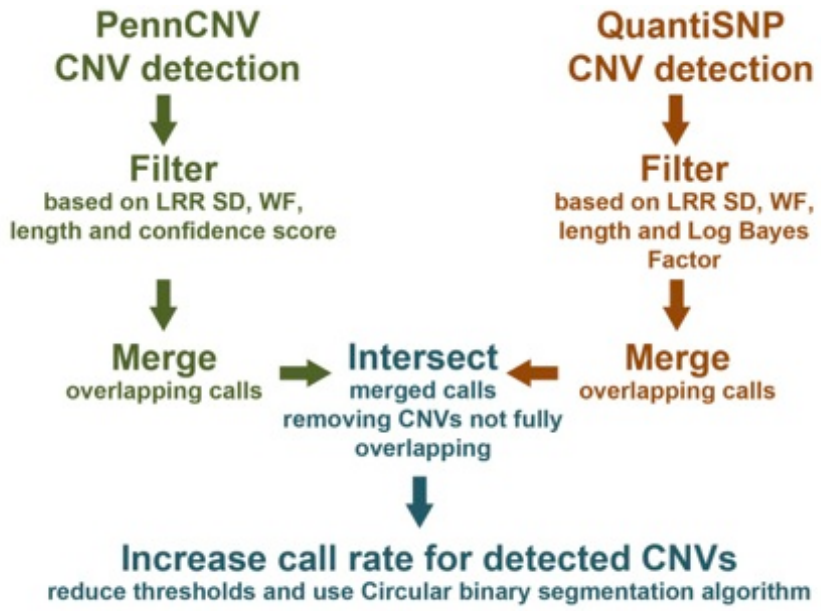


Figure 6.4: Pipeline for deletion discovery from human exome bead chip array. Initially, PennCVN and QuantiSNP were used independently to detect CNVs. Subsequently, samples with high signal variability and periodicity were removed from the analysis. Additionally, deletions were filtered based on length and confidence score (PennCNV) or Log Bayes Factor (QuantiSNP). The remaining deletions were merged together to identify a unique set of deletion regions detected by PennCVN and QuantiSNP, separately. Deletion regions identified by both PennCNV and QuantiSNP were then genotyped by the Circular Binary Segmentation (CBS) algorithm to increase sensitivity. Abbreviations: Log R Ratio (LRR), Standard Deviation (SD), Wave Factor (WF).

6.2.2.3.2 Improving genotyping sensitivity of the deletion pipeline

A first comparison between exome chip deletion calls and WGS deletions indicated good overlap between deletion sites and high genotyping specificity, which showed that deletion sites can be accurately identified; however, genotyping sensitivity was relatively low.

To improve genotyping accuracy in this set of high-quality deletion sites detected by PennCNV and QuantiSNP, the following approach was chosen:

1. CNV genotypes from QuantiSNP overlapping the high confident deletion set were re-included if they had a Log Bayes Factor ≥ 10;
2. The CBS algorithm, which defines segments of equal signal intensity, was applied to exome chip data to detect deletion genotypes using GC-corrected LRR values;
3. Deletion genotypes identified by the CBS algorithm (but not by PennCNV or QuantiSNP) were added to the calls if they fully overlapped any of the high-confident deletion regions.

6.2.2.4 Association testing

Since deletion breakpoints from exome chip SNP data can not be determined, 3 different approaches to collapse deletions were used to test identified deletions for association with T2D. Specifically, deletions were collapsed based on gene, SNP or deletion overlap as explained in Fig 6.5 .

After collapsing deletions, variants with Allele count >2 were tested for association with T2D using the Firth Bias corrected implemented in EPACTS[298]. As described above the test has been shown to be highly suitable for rare variant tests compared to other commonly used genetic association tests[299]. The following sample covariates were included in the model: two PCs derived from the SNP genotyping data (provided by Dr Anubha Mahajan a member of the McCarthy lab) to correct for population structure, sex, age and genomic wave factor (representing periodicity in exome chip SNP signal intensity) derived from QuantiSNP and PennCNV.

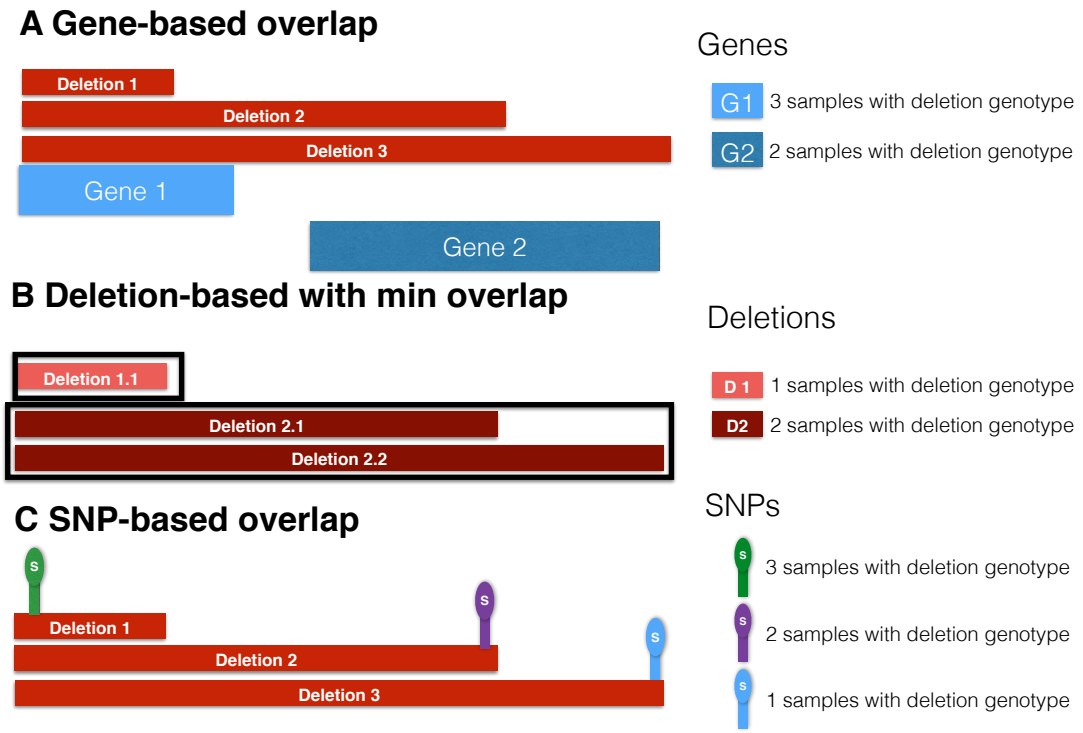


Figure 6.5: Approaches to collapse deletions for association testing. (A) The gene-based approach collapsed deletions (red) based on gene-overlap (dark and light blue). (B): The deletion-based method merged sub-deletions (dark red) that have a reciprocal overlap of at least 50% for the association analysis (indicated by a black box) while testing remaining sub-deletions independently (light red). (C) The SNP-based approach used SNP markers (green, purple, light blue) provided by the exome chip array as a proxy for deletion (red) genotypes.

6.2.2.5 *In silico* validation of genotyping sites through overlap with DGV

To determine the number of correctly identified deletion regions in the genome, the number of deletions was calculated that fully overlapped any CNV from the DGV database using bedtools intersectBed[161]. To exclude random overlap the 1,150 merged deletion sites were permuted 1,000 times by shifting the start and end position of deletions randomly between 1bp and 1Mb (shifting was performed to preserve structure of sub-deletions) and the P-value was estimated based on how often the number of randomly overlapping sites exceeded the number of true sites (minimum P-value 0.001).

6.2.2.6 *In vitro* deletion genotyping using Taqman assays

As described in more detail in Chapter 2 section 2.4, deletion genotyping was performed using predesigned CNV Taqman assays specific for the gene deletion region in combination with RNaseP CNV reference assay. Copy number was calculated using a ΔΔCt approach with the median Cycle Threshold (CT) value of diploid samples being used as reference.

Density estimations on the copy number states were derived using the densityMclust function, which is implemented in the R package mclust (version 4.4,[308,309]) and applied a Gaussian finite mixture model to the copy number states. Standard settings were used, and the optimal model was determined based on the Bayesian information criterion (BIC) criterion.

6.3 Results

6.3.1 Characterising the role of ExAC coding CNVs in T2D and obesity risk

6.3.1.1 Identification of exon sequencing CNVs

As part of a collaborative effort, the ExAC consortium provided coding CNV calls of 14.5k T2D-case-control individuals (including about 7.5k control and 7.1k case samples) from 5 different ethnic groups (see Table 6.1 and Chapter 2 Section 2.5.2.1 for further information) that were used for the analysis in this Chapter. These CNVs were identified by members of the ExAC consortium from a combined analysis of a total of 60k ExAC WES samples using XHMM (see Methods for details). In total, 611k exonic CNVs including 297k deletions and 315k duplications were detected in these 14.5k individuals by ExAC.

After filtering CNVs (according to instructions provided by ExAC) using a Phred-like XHMM CNV quality score threshold of 60 (which corresponds to a 99.9999% probability of a correct CNV call), 207k high-quality CNVs corresponding to 98k deletions and 109k duplications (Fig 6.6A) remained in the dataset. The median length of the filtered CNVs was 12.5kb with deletions being shorter (median 9.83kb) and duplications (median 14.8kb) longer.

The "cnv-make-map" function of the software plink was applied to use the locations of CNV start and end points as surrogate markers (Fig 6.3) onto which CNVs were collapsed to derive MAF estimates for overlapping CNVs. Through this approach 42.5k distinct surrogate markers (based on CNV start/end sites) were identified of which the vast majority (99.5%, Fig 6.6B) represents low frequency and rare variants (MAF <5%). This is consistent with the expected disruptive nature of exonic CNVs which would lead to strong negative selective pressure removing these variants from the population[284].

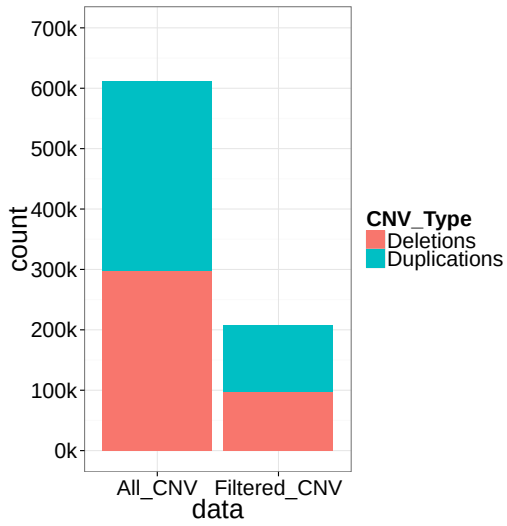
On an individual basis, it was found that each sample carried a median number of 13 CNVS (median of 6 deletions and 7 duplications per sample). Taking into account the size of the CNVs every

individual has a median of 447kb that are in a non-diploid state corresponding to 159kb deleted and 231kb duplicated sequences (Fig 6.7).

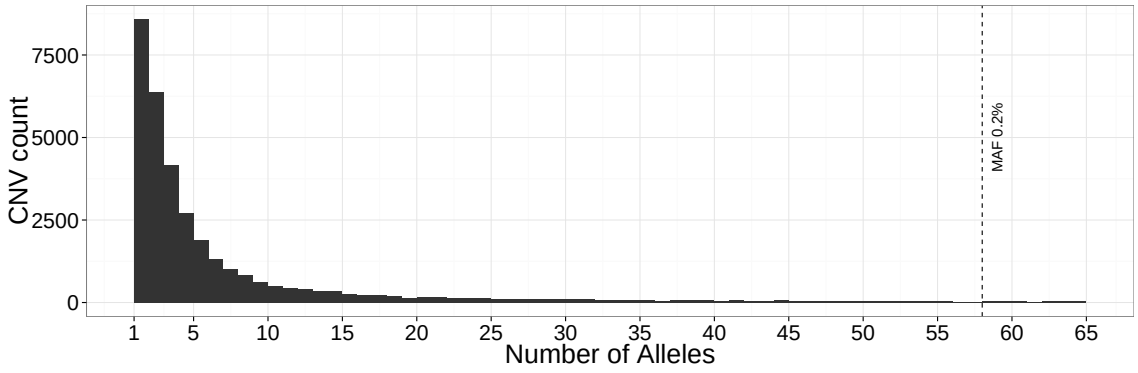
Next, the data was split according to ancestry to determine whether there are differences in CNV number per individual across cohorts with different ethnicity (Fig 6.8). Samples of African (median=23) and East Asian (median=17) descent showed the highest number of CNVs per individual compared to the other cohorts (median of 11-13 CNVs per individual depending on the cohort). This data is consistent with previous results that found individuals of African origin had also the highest number of SNVs potentially causing LoF[284] and the highest number of exonic CNVs in the whole ExAC dataset (60k samples)[157] ; although it can not be ruled out that these differences reflect batch effects associated with variation in read depth and processing of samples[157].

Ethnic-group	Individuals	Study
African-American	1787	T2D-genes
East-Asian	2146	T2D-genes
European	3279	T2D-genes
Hispanic-1	1420	T2D-genes
Hispanic-2	3763	SIGMA
South-Asian	2139	T2D-genes
Total	14534	All

Table 6.1: Number of individuals per ethnicity and study

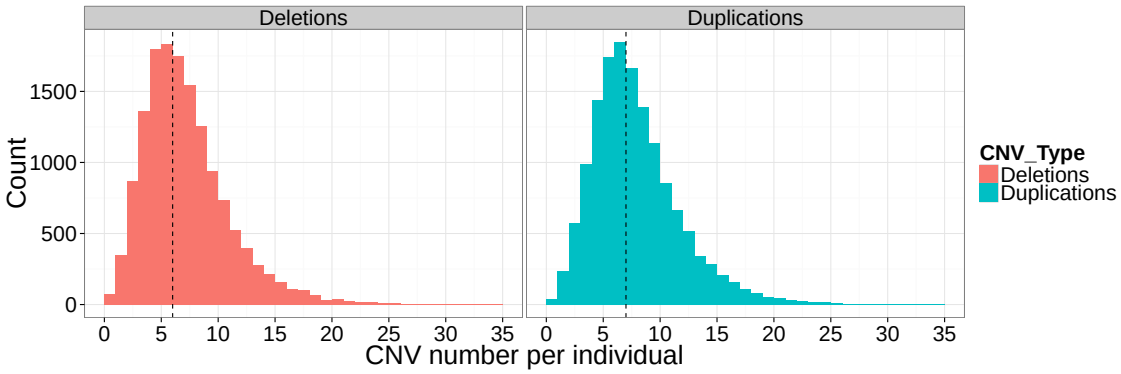


(A) CNV-number

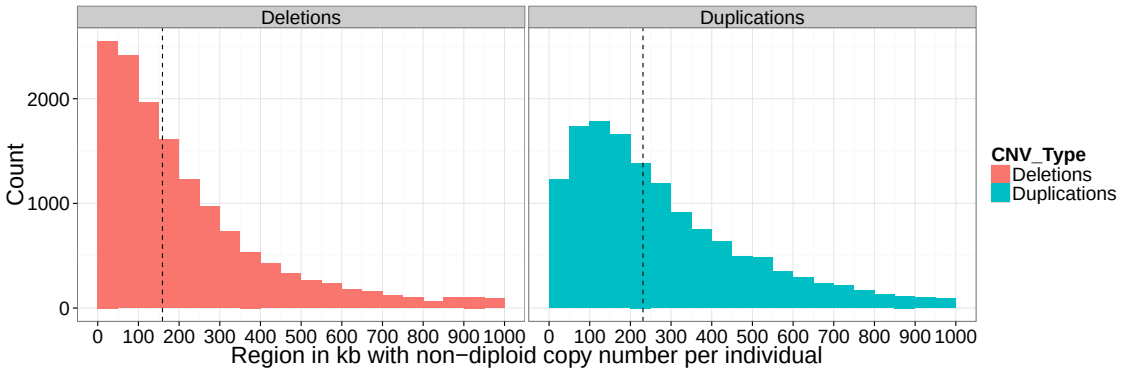


(B) CNV-AF

Figure 6.6: ExAC CNV number and allele frequency. (A) CNV number (y-axis) in the unfiltered and filtered dataset (x-axis) split into deletions (red) and duplications (blue). (B) Histogram of filtered CNVs showing allele number on the x-axis. To improve visualisation, the histogram threshold was set at 65 as the maximum allele count shown.



(A) CNV-number per individual



(B) Genomic region per individual with non-diploid copy number

Figure 6.7: ExAC CNV number per individual. (A)-(B) Histogram of (A) CNVs per individual (x-axis) and (B) the amount of the genome (in kb) affected per individual with abnormal copy number state. CNVs are split into deletions (red) and duplications (blue). To improve visual representation, the limits for the x-axis were set at 35 CNVs and at 1000kb. Dashed line indicates median.

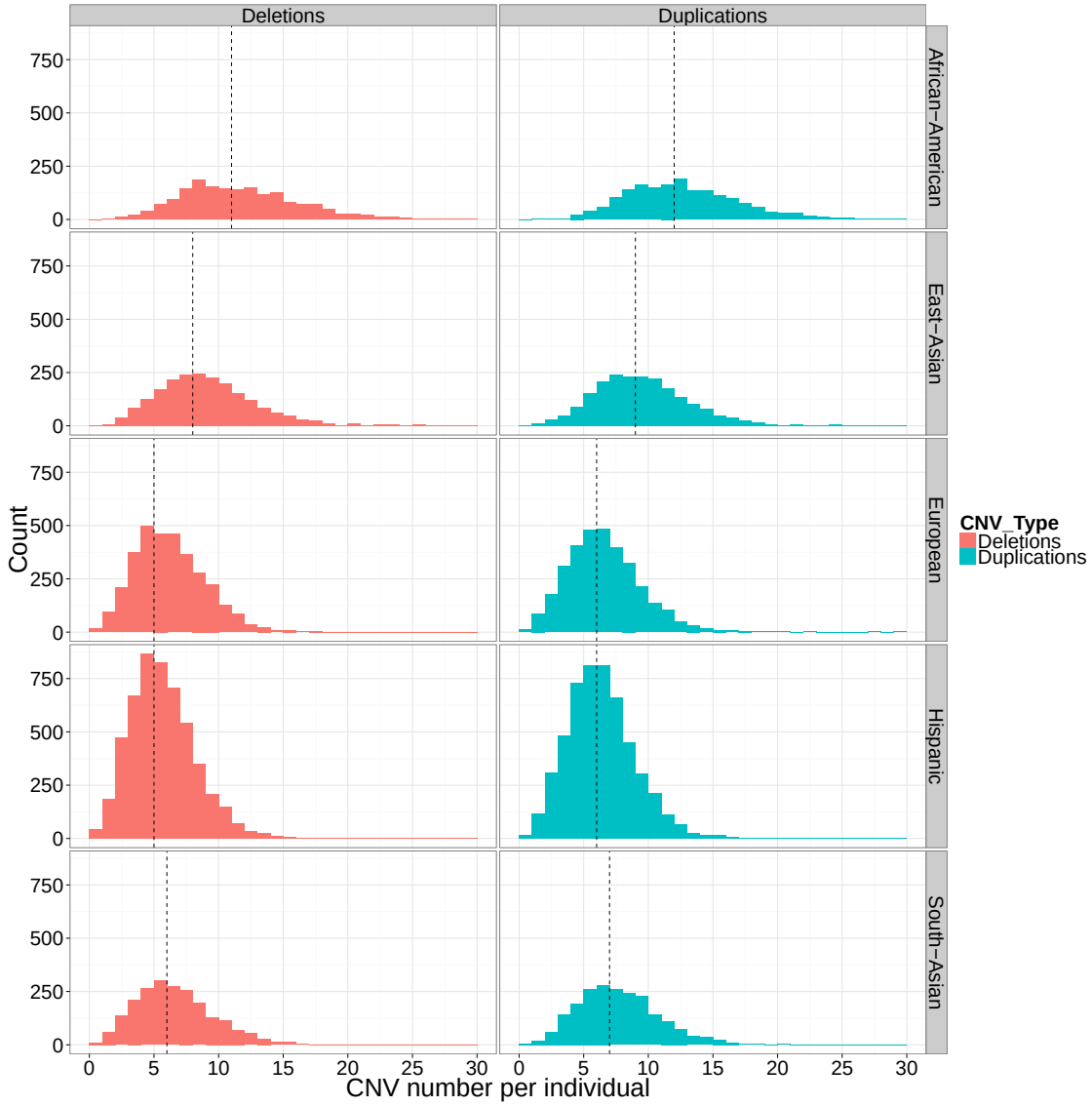
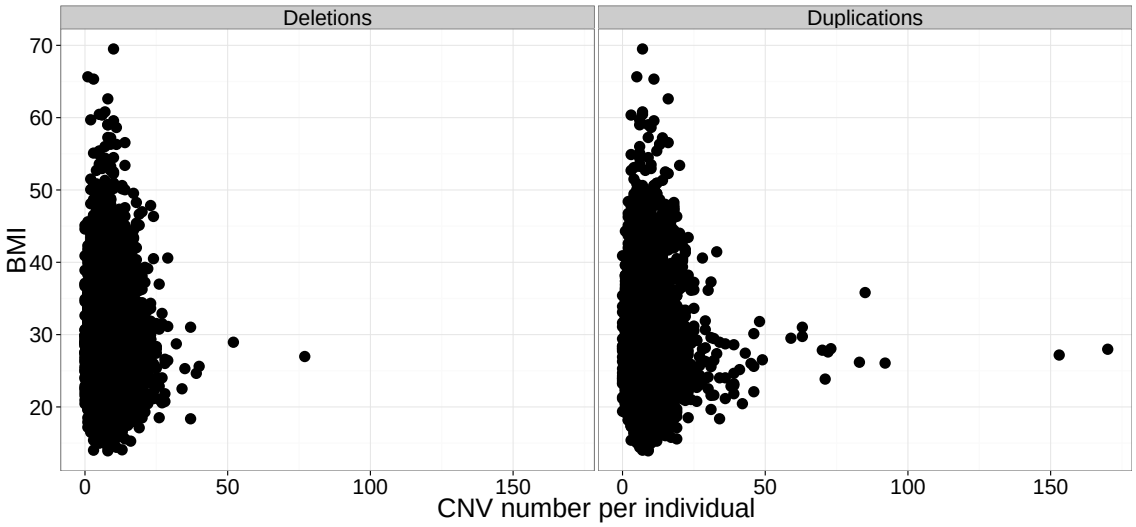


Figure 6.8: ExAC CNV number per individual and cohort. Histogram of CNV count per individual (x-axis) split per cohort and CNV type (deletions are red while duplications are shown in blue). The median number of CNVs per individual is indicated as dashed line for each cohort separately.

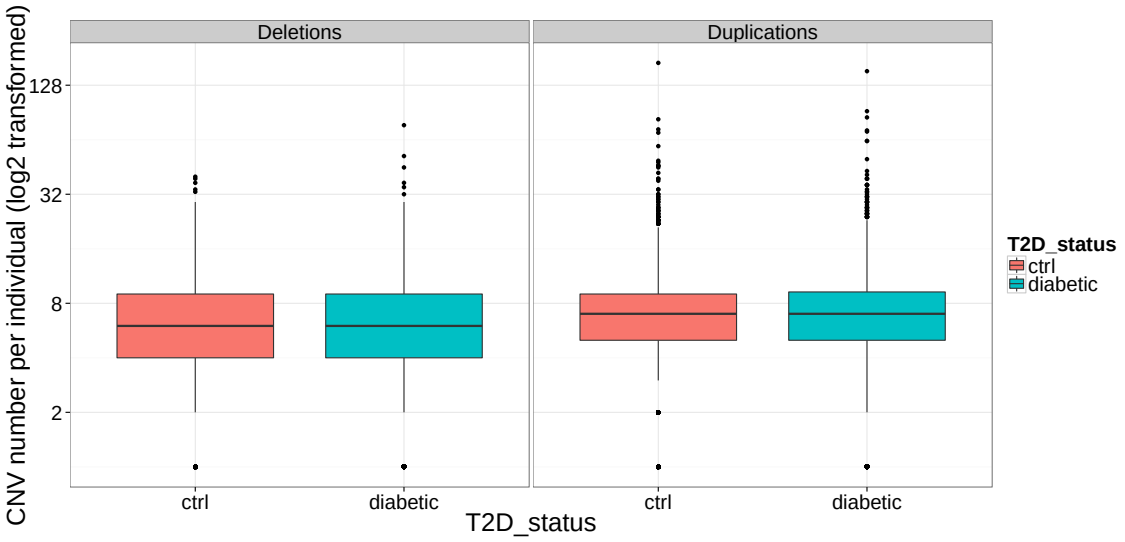
.

6.3.1.2 Determining WES CNV burden effects for T2D status and BMI

Next, the genetic CNV calls were integrated with phenotypic data (T2D status and BMI; for the BMI analysis, samples belonging to the SIGMA study had to be removed since permission was not obtained to test CNV-associations for BMI in these samples) to determine whether there is a CNV burden effect dependent on the phenotype. That is, differences in the CNV number per individual that are correlated with T2D status or BMI (Fig 6.9). Although the correlation analysis revealed a significant association between CNV number and both T2D status (P=0.024) and BMI (P=8.7×10^{-7}), the effect size (T2D rho=0.019, BMI rho=-0.048) was small suggesting that there is only a weak association between CNV number and these phenotypes. The analysis was then repeated for both T2D and BMI across the different ancestry groups to determine whether a CNV burden effect could be determined according to ethnic background. However, across ancestry groups, no significant association between CNV burden and either T2D status or BMI could be observed. Overall these results suggest that for both T2D status and BMI CNV burden per individual does not have a major impact on the phenotype.

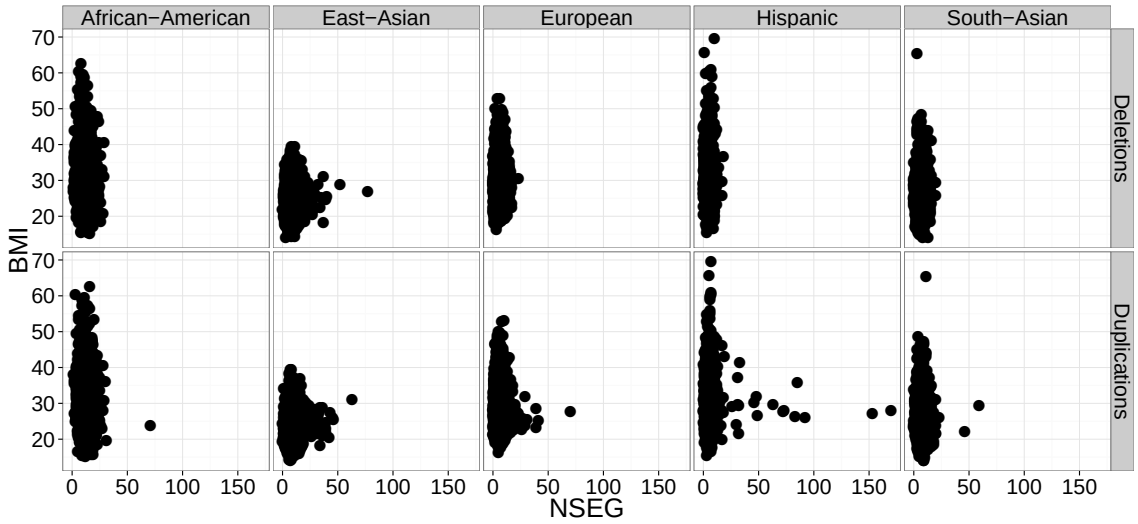


(A) CNV-burden BMI

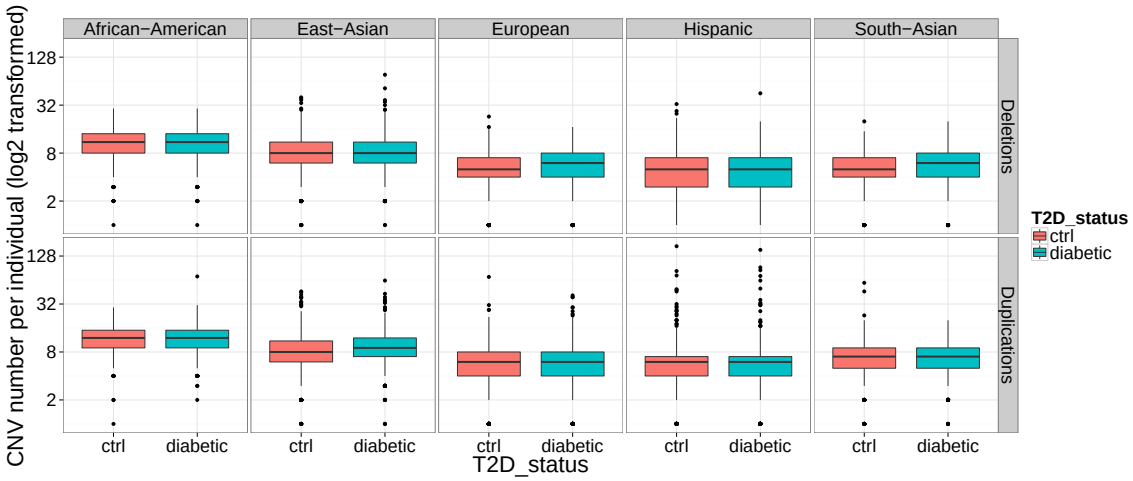


(B) CNV-burden T2D

Figure 6.9: BMI and T2D CNV burden per individual. (A) Correlation between BMI (y-axis) and CNV number per individual (x-axis). (B) Differences in CNV burden (CNVs per individuals, y-axis, log2 transformed for better visualisation) between T2D controls (ctrl, red) and cases (diabetic, blue).



(A) CNV-burden BMI per cohort



(B) CNV-burden T2D per cohort

Figure 6.10: BMI and T2D CNV burden per individual per cohort. (A) Correlation between BMI (y-axis) and deletion (top row, x-axis) and duplication number (bottom row, x-axis). (B) Differences in number of deletions (top, y-axis) and duplications (bottom, y-axis, log2 transformed for improved visualisation) between T2D controls (ctrl, red) and cases (diabetic, blue).

6.3.2 Testing rare coding WES CNVs for association with T2D and obesity

To identify genes in which a disruptive event may be correlated with T2D and obesity predisposition, deletions and duplications were collapsed separately based on overlap with hg19 gene annotations provided by plink (generated from the UCSC Table Browser RefSeq track data in May 2014). This approach is similar to an aggregate gene-burden test performed for SNVs in order to test functionally relevant units (genes) for association with disease, which decreases the number of tests performed and aggregates the AC of multiple variants for the tested functional unit. In total, this reduced the number of markers (representing genes) to analyse for associations drastically with deletions and duplications overlapping 5507 and 9369 protein-coding genes, respectively. The vast majority (99.9%) of these gene-centred markers was still low frequency (MAF <5%, Fig 6.11). After filtering based on AC (minimum AC >2), 1908 gene-centred deletions and 3603 gene-centred duplications were tested for associations with T2D and BMI using the software EPACTS[298]. Each cohort was analysed separately to reduce the influence of population structure and batch effects. The following covariates were included in the association model sex, age, and PCs 1-4 (the number of PCs was dependent on the study). For the T2D association testing, the Firth Bias-Corrected Logistic Likelihood Ratio Test was used and BMI was included as an additional covariate while the Linear Wald Test was used for the BMI association analysis. The EPACTS individual cohort results were then meta-analysed with the software tool METAL[300] to identify deletions and duplications associated with T2D and BMI across all cohorts. In total, a single gene-based deletion event was found that was associated with BMI at exome-wide significance ($P < 2.5 \times 10^{-6}$, Fig 6.12 and Fig 6.13) and that overlapped the gene *RP11-1396O13.13*. In addition, 149 T2D and 171 BMI nominally significant associations (P <0.05) were identified that will be described for each phenotype separately below.

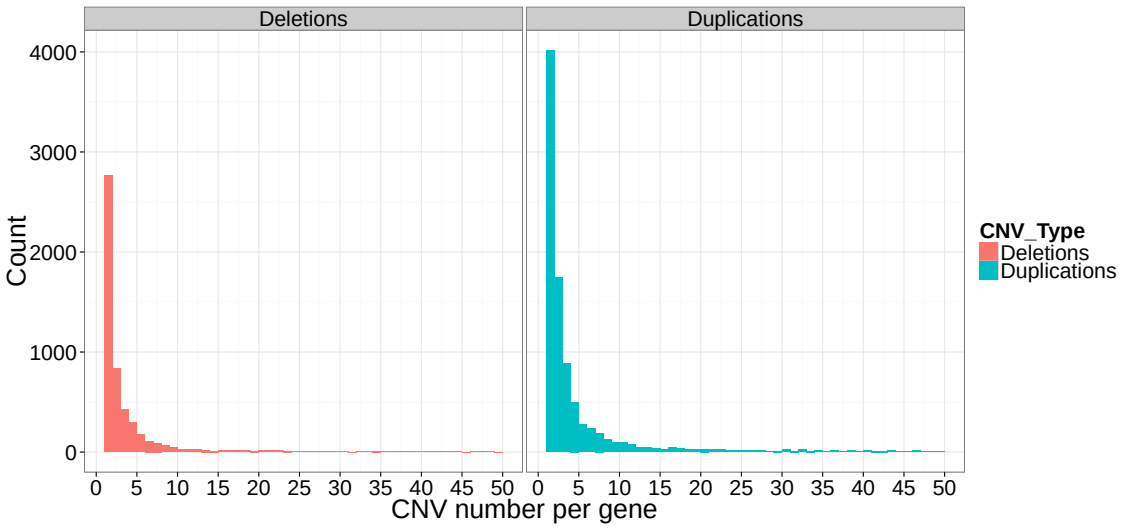
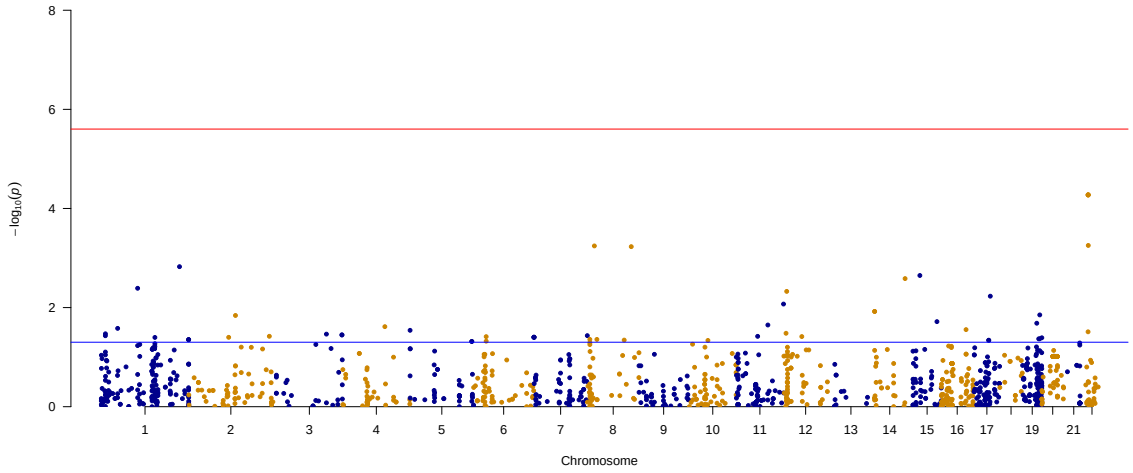
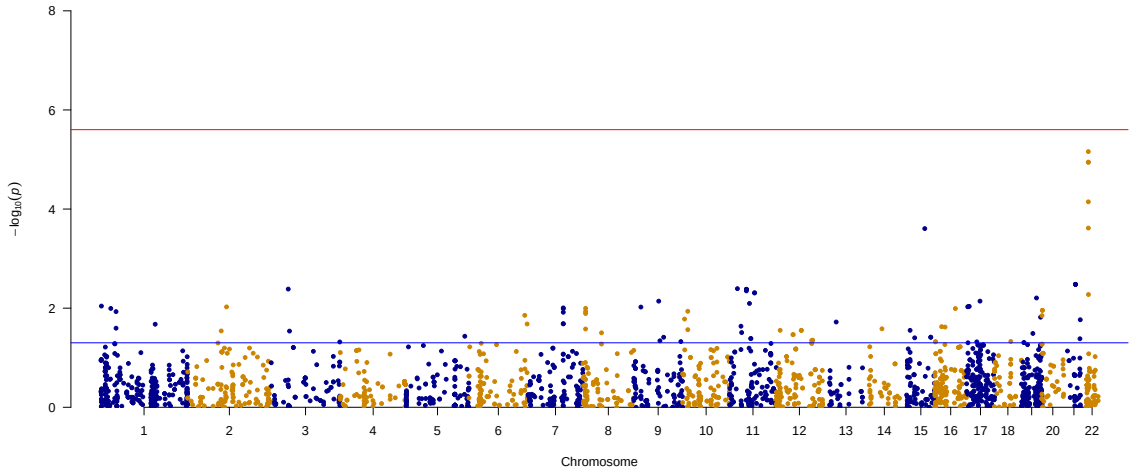


Figure 6.11: Gene deletion and duplication AC used for association testing. Histogram of CNV AC per gene split into deletions (red) and duplications (blue)

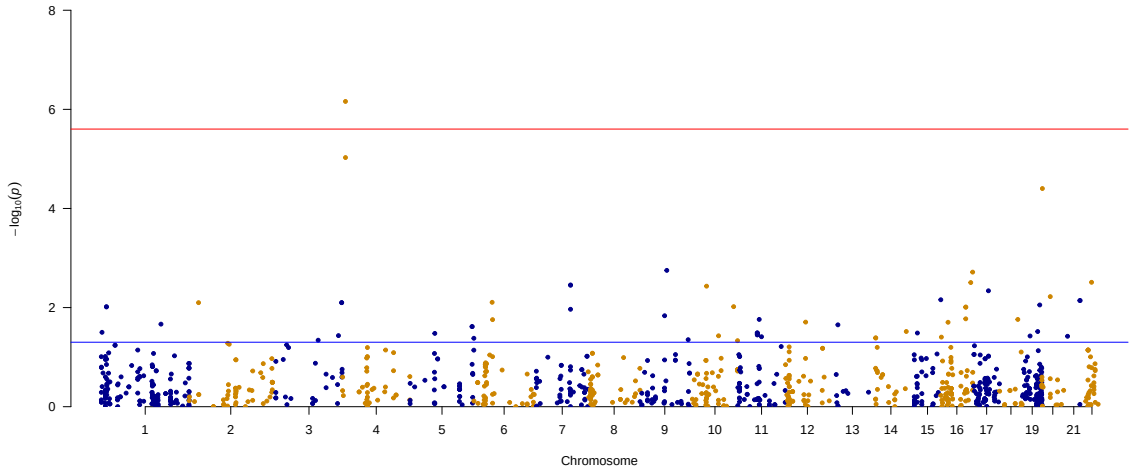


(A) T2D GWAS deletion manhattan plot

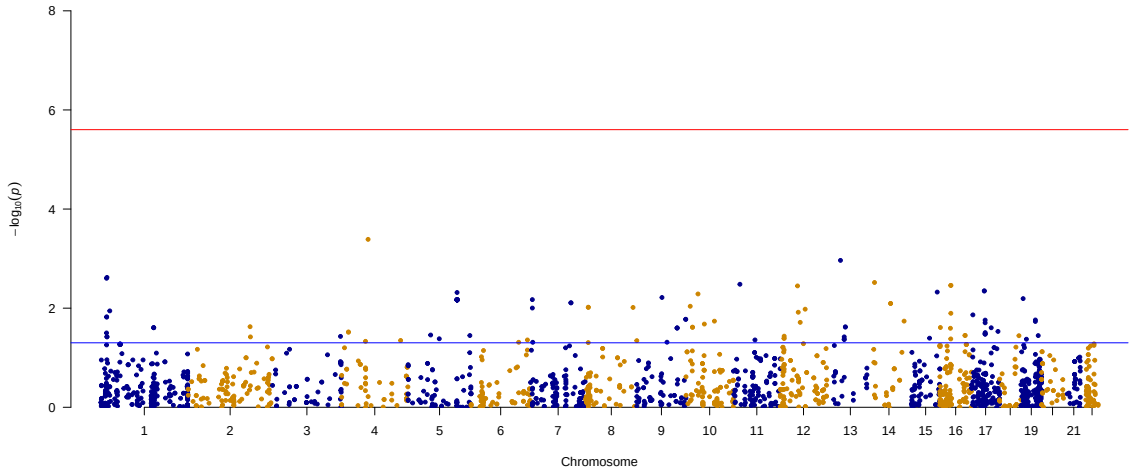


(B) T2D GWAS duplication manhattan plot

Figure 6.12: Manhattan plot of deletion and duplication T2D association results. (A)-(B) Manhattan plot with association results for T2D tested deletions (A) and duplications (B). Chromosomal coordinates are shown on the x-axis and the -logP value on the y-axis with a blue line indicating nominal significance (P<0.05) and a red line exome-wide significance ($2.5x10^{-6}$).



(A) BMI GWAS deletion manhattan plot



(B) BMI GWAS duplication manhattan plot

Figure 6.13: Manhattan plot of deletion and duplication BMI association results. (A)-(B) Manhattan plot with association results for BMI tested deletions (A) and duplications (B). Chromosomal coordinates are shown on the x-axis and the -logP value on the y-axis with a blue line indicating nominal significance (P<0.05) and a red line exome-wide significance ($2.5x10^{-6}$).

6.3.2.1 T2D gene based association results

6.3.2.1.1 T2D-GWAS meta-analysis results

For T2D, 58 deletions and 91 duplications achieved nominal significance (P <0.05, see Appendix Table 3) and the top 10 T2D-associated variants of each copy number type are shown in Table 6.2. To determine whether any of these nominally significant regions has been previously associated with T2D, CNVs were overlapped with DIAGRAM GWAS SNV summary level data. Overall, only an *ABCC3* gene deletion (P=0.006) and *ANKFY1* gene duplication (part of the *ZZEF1* locus, P=0.009) were identified within 100kb of regions that showed suggestive evidence for T2D GWAS association ($P < 5x10^{-6}$). To determine whether the direction of effect is consistent with previous GWAS association data, publicly available gene burden results of coding missense or protein truncating SNVs (category "possibly deleterious missense variants") were obtained from the T2D genetic online portal (website: `http://www.type2diabetesgenetics.org`). The data on the portal also includes SNV information on the WES ExAC dataset and directionally consistent gene-burden results may provide additional validation for the identified CNV associations; however, no significant T2D SNV gene-burden results for *ABCC3* and *ANFY1* were found in the portal data. While both genes are expressed in human islets (among the top 20% and 25% expressed genes based on publicly available islet expression data[236]), the biological role of the highlighted genes is not well characterised: *ABCC3* encodes a protein involved in transporting proteins across membranes[310] while *ANKFY1* encodes a double zinc-finger membrane protein that has been suggested to be involved in protein or vesicle transport[311]. Although further validation is required, the obtained results prioritise these variants for follow-up at these T2D GWAS loci.

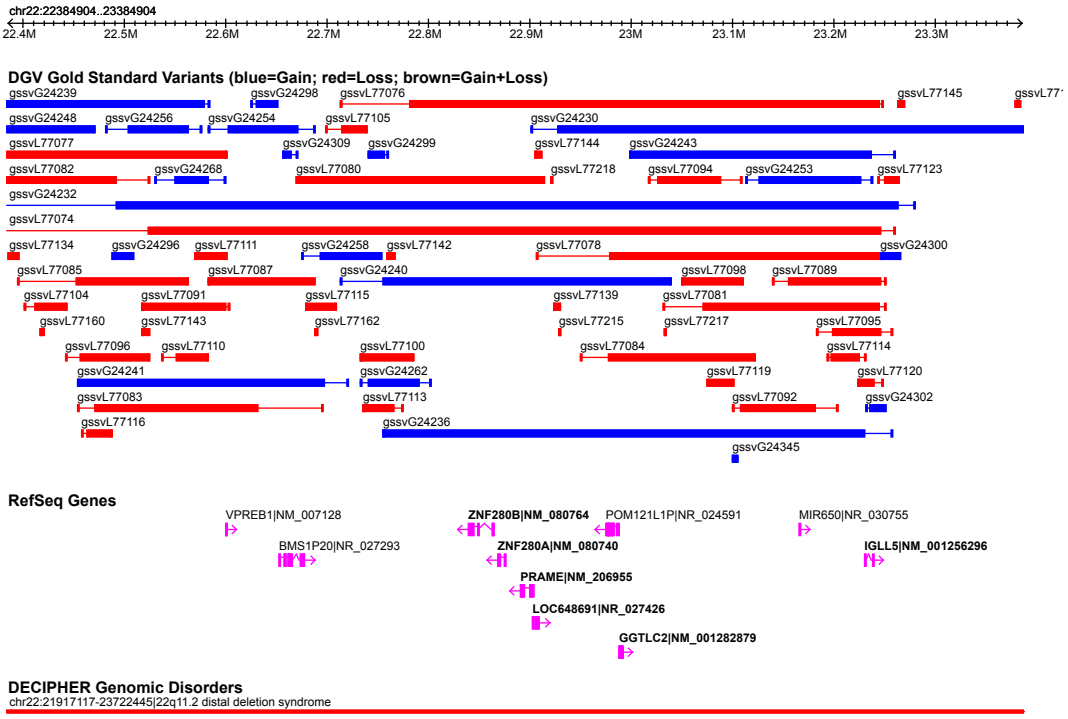
In addition to linking T2D variants to known T2D GWAS results, the role and CNV structure of the other highlighted genes was also investigated (Table 6.2). Consistent with the similarities in association signal, AC and chromosomal location, the top hits overlapping *LL22NC03-63E9.3, ZNF280A, ZNF280B, PRAME, GGTLC2* and *IGLL5* were all found in one large CNV region on chromosome 22 which is present in the Database of Genomic Variants (DGV)[312]. The region, which overlaps the T2D-associated CNVs, is involved in a rare heterogeneous disease named 22q11.2 distal deletion syndrome (OMIM Phenotype Number: 611867) and located directly next

to another CNV region associated with DiGeorge Syndrome (OMIM Phenotype Number: 188400, Fig 6.14A). Both disorders are characterised by intellectual disability and several other phenotypes including obesity (for DiGeorge Syndrome)[313]; however, no direct link to T2D was found. Most of these genes are biologically not well-characterised making it difficult to determine the role of these variants in T2D development: *LL22NC03-63E9.3* encodes a non-coding RNA of unknown function, *ZNF280A* and *ZNF280B* act as potential transcription factors (NCBI gene database[314]), *PRAME* is mainly expressed in tumour tissue[315], *GGTLC2* is involved in peptide processing[316] and *IGLL5* is an immunoglobulin gene (NCBI gene database[314]).

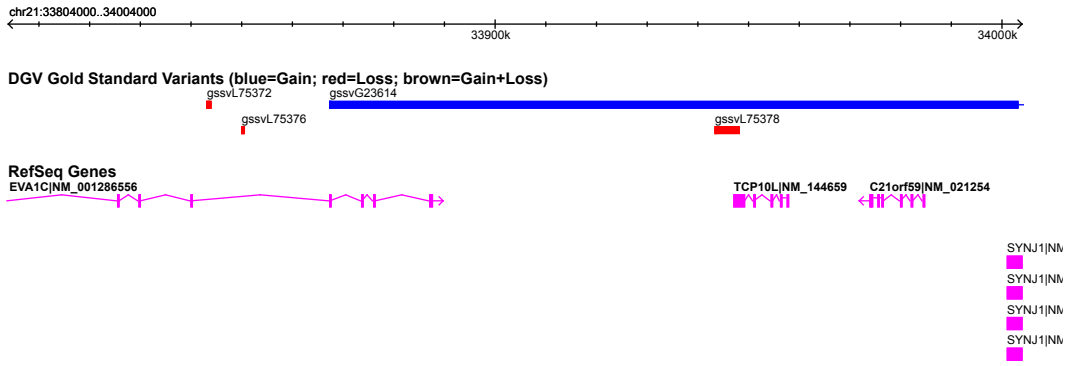
Similarly, the duplications on chromosome 21 (Table 6.2) are also in DGV and overlap a known large duplicated region spanning multiple genes potentially involved in Down syndrome[317]. While SNV gene-burden results found an association between missense variants in the gene *EVA1C* and T2D (P=0.02, OR=1.30), the function of the duplication overlapping genes is not well-characterised, thus making it unclear what the phenotypic effect of the duplication acting as PDV could be. *EVA1C* was suggested to play a role in the development of the murine nervous system[318], *TCP10L* potentially acts as transcriptional regulator[319] , *C21orf59* is involved in cilia function and associated with a ciliary dyskinesia affecting the respiratory and reproductive system, and *AP000275.65* is a paralogue of *C21orf59*[320] (Fig 6.14B). Although the nominally significant *EVA1C* gene-burden T2D-association results in SNV data support the CNV-association and prioritise this gene duplication for potential involvement in T2D development, the lack of understanding of the gene function makes the design of follow-up experiments difficult. Also, complex CNV structure is found at the *GOLGA6B* locus (Table 6.2) which appears to be more common in T2D control than case individuals; however, it is part of a low copy number repeat sequence on chromosome 15[321], thus, making accurate CNV genotyping difficult. For the other top genes the biological functions did not necessarily provide clear support for their role in T2D development (Table 6.2): *SLC18A1* is a monoamine transporter and was previously associated with neurological disorders[322], *WDYHV1* encodes a glutamine amidohydrolase[323], *HHIPL2* is an antagonist of hedgehog signalling and expressed in bone[324] and *TYRO3* has multiple roles related to cell adhesion, spermatogenesis, immunoregulation and phagocytosis[325]. Overall these results suggest that additional validation will be required in the future to determine whether the association can be replicated and to what extent these genes are involved in T2D development.

Gene	CNV Type	P.value	Zscore	Directionality	CHR	SBP	AC	GB-P	GB-OR
LL22NC03-63E9.3	Duplications	6.93E-06	-4.50	+ ? ? + ? ?	22	22,901,749	31	NAN	NAN
ZNF280A	Duplications	1.13E-05	-4.39	+ ? ? + ? ?	22	22,868,059	30	0.807	1.01
ZNF280B	Duplications	1.13E-05	-4.39	+ ? ? + ? ?	22	22,838,766	30	0.98	1
LL22NC03-63E9.3	Deletions	5.32E-05	-4.04	+ ? ? - ? ?	22	22,901,749	40	NAN	NAN
PRAME	Deletions	5.32E-05	-4.04	+ ? ? - ? ?	22	22,890,122	40	0.582	0.967
ZNF280A	Deletions	5.32E-05	-4.04	+ ? ? - ? ?	22	22,868,059	40	0.807	1.01
ZNF280B	Deletions	5.32E-05	-4.04	+ ? ? - ? ?	22	22,838,766	40	0.98	1
PRAME	Duplications	7.14E-05	-3.97	+ ? ? + ? ?	22	22,890,122	32	0.582	0.967
GGTLC2	Duplications	2.42E-04	-3.67	+ ? ? - ? ?	22	22,988,779	32	0.601	0.976
GOLGA6B	Duplications	2.48E-04	-3.66	+ + + + + +	15	72,947,078	528	0.766	0.803
IGLL5	Deletions	5.57E-04	-3.45	+ ? ? ? ? ?	22	23,229,959	20	0.0728	1.78
SLC18A1	Deletions	5.70E-04	3.45	? ? - ? ? ?	8	2e+07	8	0.16	1.07
WDYHV1	Deletions	5.89E-04	-3.44	+ ? ? - ? ?	8	124,428,964	68	0.666	1.01
HHIPL2	Deletions	1.49E-03	-3.18	? ? ? ? ? +	1	222,695,601	20	0.154	1.05
TYRO3	Deletions	2.25E-03	-3.06	+ ? ? ? ? ?	15	41,849,872	8	0.859	0.988
KIAA0125	Deletions	2.60E-03	-3.01	+ + ? - ? +	14	106,383,837	48	NAN	NAN
AP000275.65	Duplications	3.31E-03	2.94	? ? ? ? - ?	21	33,951,131	17	NAN	NAN
C21orf59	Duplications	3.31E-03	2.94	? ? ? ? - ?	21	33,964,388	17	0.206	0.798
EVA1C	Duplications	3.31E-03	2.94	? ? ? ? - ?	21	33,784,313	17	0.0195	1.3
TCP10L	Duplications	3.31E-03	2.94	? ? ? ? - ?	21	33,948,861	17	0.118	1.31

Table 6.2: Top 10 T2D-associated deletion and duplication CNVs. For each variant, the overlapping gene, the CNV type (deletion or duplication), the T2D-meta-analysis P-value, Z-score and Direction of effect per study are shown. Directionality for the studies is given in the following order: European, African America, East-Asian, T2D-genes Hispanics1 , South-Asian and SIGMA Hispanics2. In addition, chromosomal position (chromosome (CHR) and Start Base Position (SBP)), allele count (AC) and Gene-burden P-value (GB-P) and OR (GB-OR) for all possibly deleterious missense variants from the T2D genetics portal are shown (NA indicates not available).



(A) Complex CNV structure at 22q.11.2 DGV genome view



(B) Duplication at chromosome21 DGV genome view

Figure 6.14: Complex Genomic rearrangements at T2D associated CNVs. (A) DGV genome browser view of a region on chromosome22 containing the top T2D associated regions overlapping genes *ZNF280A, ZNF280B, PRAME, GGTLC2* and *IGLL5* which are indicated (bold, RefSeq track middle pink). While the gene-centred exome CNV calls indicate separate deletions per gene DGV Gold standard CNV calls (top) suggest a large shared deletion and duplication event spanning multiple genes. Also, the bottom track indicates overlap with the 22.q11.2 deletion syndrome region. (B) DGV genome browser view of multiple associated T2D gene-centred duplications including *EVA1C, TCP10L* and *C21orf59* (highlighted in bold, RefSeq track bottom pink) which likely represent one large duplication as supported by DGV CNV calls (top track, blue).

6.3.2.1.2 Very rare CNVs with imbalance in T2D-case-control samples

Since strong effect size variants are often very rare due to selective pressure, they are often hard to identify using traditional GWAS because of lack of power. Hence, a comparison of T2D case-control AC was performed in order to identify potentially interesting rare strong effect size variants that influence T2D risk. That is, it was determined whether rare collapsed gene-centred deletions and duplications showed strong imbalance in case VS control allele counts and whether these genes fall into regions that were previously associated with T2D based on DIAGRAM GWAS SNV summary level data. In total, this analysis identified 8 CNVs (3 deletions and 5 duplications) near suggestive T2D associated GWAS hits at which the difference in allele count between T2D affected and unaffected individuals was at least fourfold (Table 6.3). These CNVs include 1 T2D-protective gene deletion (*CBLC*), 4 T2D-protective gene duplications (*HEYL, YTHDC2, VPS13C, GAK*) as well as 2 T2D-causal gene deletions (*VPSC13C, ABCC3*) and 1 T2D-causal duplication (*ANKFY1*).

For *VPS13C*, which is part of the T2D *C2CD4A* locus, both deletions and duplications were identified with opposing effects. That is, *VPS13C* deletions showed evidence of being causal for T2D while duplications appeared to be protective against T2D. The location of the deletions and duplication in the *VPS13C* gene (Fig6.15) suggests that the opposing effects are likely not due to different parts of the gene being affected but indeed due to the contrasting effect on copy number. Consistent with the deletion PTV effect being causal for T2D, publicly available gene burden results for SNVs showed a nominally significant association with T2D (P=0.02, OR=1.67). *VPS13C* is also highly expressed in human pancreatic islets[326,327] and the T2D risk C allele of the likely causal SNP rs7163757 was found to decrease expression of *VPS13C*[328].

The case-control allelic imbalance analysis also highlighted several other potentially relevant T2D-genes (Table 6.3) as assessed by their expression levels (minimum among top 25% of all islet expressed genes) in pancreas and islets. However, further work will be required to establish their biological function in more detail. Apart from *ANFY1* and *ABCC3* which were identified in the association results, the following novel genes were highlighted: The gene *HEYL* encodes a TF involved in Notch signalling[329], shows high expression in human islets (one of the top 20% expressed genes according to publicly available islet expression data[236]) and combined knockout of *Heyl* and *Hey1* in mice leads to developmental cardiovascular defects[330]. *YTHDC2* which encodes an

RNA helicase associated with RNA processing[331] and *CBLC* which is involved in cell signalling[332] and the organisation of the Golgi network[333]. Both are expressed in pancreas and islets and CNVs in *YTHDC2* have been associated with pancreatic cancer[334]. *GAK* is also a highly expressed islet gene (top 12%), encodes a kinase that plays a role in vesicle transport[335] and is associated with Parkinson's Disease[336].

Gene	CNV Type	AFF	UNAFF	log2FE	GWAS Locus or SNP
HEYL	Duplication	0	6	-Inf	MACF1
YTHDC2	Duplication	1	6	-2.58	rs366219
VPS13C	Duplication	1	5	-2.32	C2CD4A
CBLC	Deletions	1	4	-2.00	APOE
GAK	Duplication	1	4	-2.00	MAEA

(A) Rare Protective Alleles

Gene	CNV Type	AFF	UNAFF	log2FE	GWAS Locus or SNP
VPS13C	Deletions	6	0	Inf	C2CD4A
ANKFY1	Duplications	7	1	2.81	ZZEF1
ABCC3	Deletions	9	2	2.17	rs898453

(B) Rare causal alleles

Table 6.3: Rare T2D-causal and protective CNVs with imbalance in allelic case and control counts. The table shows very rare T2D-causal and protective CNV variants near known GWAS loci or lead variants (column GWAS Locus or SNP) providing the number of individuals affected (AFF, diabetic individuals) and unaffected (UNAFF, control individuals) by T2D as well as the log2FE between affected and unaffected individuals.

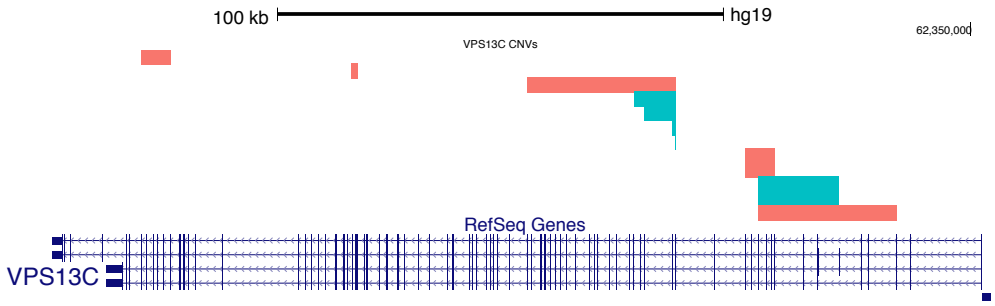


Figure 6.15: VPS13C deletion and duplication genome browser view. Location of deletions (red) and duplications (blue) are shown in relation to the *VPS13C* gene (bottom track, blue RefSeq transcripts are indicated). At the top, a genomic scale is provided to give an overview of the size of the deletions.

6.3.2.2 BMI association results

For BMI, the top 10 deletions and duplications, which include one exome-wide significant (P $<2.5x10^{-6}$) gene deletion overlapping *RP11-1396O13.13*, are presented in Table 6.4. In addition, 60 deletions and 101 duplications achieved nominal significant (P <0.05, see Appendix Table 4) associations with BMI. As for the T2D-associated CNVs, the nominal significant deletions and duplications were overlapped with ENGAGE BMI GWAS SNV summary level data to determine whether any of these regions were previously found to be associated with BMI. In total, 1 gene deletion in *PKHD1* and 2 gene duplications in *GDPD3* and *ERBB2IP* were found near BMI associated GWAS variants. The rare duplication with a strong protective effect against obesity overlapped the gene *GDPD3* and was found in the 16.p11.2 chromosomal region. Consistent with the association data, 16.p11.2 microduplications are associated with underweight[148] while 16.p11.2 microdeletions are associated with obesity[149,337], thus, providing validation of the association results. *PKHD1* mutations cause autosomal recessive polycystic kidney disease[338] while *ERBB2IP (ERBIN)* is involved in cell cycle regulation[339] and knockdown of *Erbin* in mice is under certain conditions associated with cardiac defects[340]. However, no obvious link could be established to obesity for these genes.

In addition to overlap with genome-wide significant GWAS regions, T2D portal BMI SNV (category of "possibly deleterious missense variants") aggregate gene-burden results were obtained for the top-associated hits (Table 6.4) and used to evaluate these CNV associations in more detail. For instance, BMI SNV gene-burden meta-analysis results showed a nominal association between BMI and missense variants in the genes *ITGAX* and *HNRNPC* which were also among the top 10-associated duplications (Table 6.4). While the direction of effect for the SNV gene-burden test was in the same direction of effect as the gene-centred CNV test, it remains unclear how the duplication may affect gene function and T2D development. In particular, since the biology of the affected genes does not necessarily indicate a role in obesity. *ITGAX* encodes a gene involved in the immune response including phagocytosis and neutrophil adherence while *HNRNPC* encodes heterogeneous nuclear ribonucleoproteins that are involved in RNA-processing (NCBI gene database[314]).

Apart from these 2 regions, another interesting candidate is the gene *SLC5A4* which is a glucose transporter that was found to play a role in insulin-independent, exercise-stimulated muscle glucose uptake and may decrease insulin resistance after exercise[341]. While no association was found in

the BMI SNV gene burden meta-analysis results across all cohorts (Table 6.4) including "possibly deleterious missense variants" , a nominally significant (P<0.05) association was observed between likely PTVs (category "possibly deleterious missense variants" P=0.0497, beta=0.58) and BMI in individuals with African-American ancestry. These results are consistent with the observation that *SLC5A4* deletions were associated with increased risk for obesity in the CNV analysis in African-American samples (P=0.003, the deletion was not present in any of the other cohorts). Further research will be required to replicate the associations and to determine whether these genes may be involved in the genetic susceptibility of obesity.

The remaining top results (Table 6.4) also indicated several regions that may represent spurious associations due to complex genomic regions including the shared duplication overlapping *DEFB131* and *RP11-1396O13.13*). *DEFB131* is a beta-defensin gene found in multiple gene clusters and involved in immune response[342]. Similarly, *SIRPD*, which is part of the SIRP protein family, is involved in immunoregulation and part of a gene cluster[343] and the RNA polymerase subunit *POLR2J2* (near *UKP3BL*) is also found in regions of gene clusters with multiple copies[344], thus, increasing the chance of incorrectly determining their copy number state.

Overall, these data highlight potentially relevant CNVs at a few loci such as *SLC5A4* that may play a role in obesity risk. Future work will be required to validate these results in different studies and to perform potential functional follow-up of genes of interest. In addition, the results presented here also highlighted that some associated regions are structurally complex and thus, will need extensive validation to determine whether these are true disease associations.

Gene	CNV Type	P.value	Zscore	Directionality	CHR	SBP	AC	GB-P	GB-B
RP11-1396O13.13	Deletions	6.92E-07	4.96	- - - - -	4	9,385,742	137	NAN	NAN
DEFB131	Deletions	9.39E-06	4.43	- - - - -	4	9,446,259	123	0.21	0.86
SIRPD	Deletions	3.96E-05	-4.11	+ ? ? ? ?	20	1,514,896	5	0.27	0.42
ART3	Duplications	4.09E-04	3.53	? ? - ? ?	4	76,932,336	10	0.55	0.03
FREM2	Duplications	1.09E-03	3.27	- ? - ? ?	13	39,261,265	48	0.74	-0.02
ALDH1A1	Deletions	1.77E-03	-3.13	? ? + ? ?	9	75,515,577	14	0.64	-0.27
ZNF276	Deletions	1.92E-03	-3.10	+ ? ? ? ?	16	89,786,807	4	1.00	0.0008
PADI4	Duplications	2.40E-03	-3.04	? ? + + ?	1	17,634,689	11	0.87	-0.01
FAM131C	Duplications	2.49E-03	-3.02	+ ? + ? +	1	16,384,263	22	0.98	-0.002
HNRNPC	Duplications	3.03E-03	-2.96	+ + + ? +	14	21,677,294	37	0.03	1.52
SLC5A4	Deletions	3.09E-03	-2.96	? + ? ? ?	22	32,614,464	19	0.60	0.06
ATP2C2	Deletions	3.13E-03	-2.96	? ? + ? +	16	84,402,132	25	0.26	-0.11
MRGPRX1	Duplications	3.29E-03	-2.94	+ + + + +	11	18,955,359	490	0.58	0.04
ITGAM	Duplications	3.47E-03	2.92	? - ? ? ?	16	31,271,310	5	0.50	-0.16
ITGAX	Duplications	3.47E-03	2.92	? - ? ? ?	16	31,366,454	6	0.02	-0.14
POLR2J2	Deletions	3.52E-03	2.92	- - - ? -	7	102,277,473	230	NAN	NAN
UPK3BL	Deletions	3.52E-03	2.92	- - - ? -	7	102,277,471	115	NAN	NAN
PRPH	Duplications	3.55E-03	-2.92	+ + ? + ?	12	49,687,034	25	0.78	-0.05
ASAH2C	Deletions	3.70E-03	2.90	? - ? ? -	10	4.8e+07	33	NAN	NAN
CWC25	Duplications	4.49E-03	-2.84	? ? ? ? +	17	36,956,686	4	0.35	-0.11

Table 6.4: Top 10 BMI-associated deletion and duplication CNVs. For each variant, the overlapping gene, the CNV type (deletion or duplication), the BMI-meta-analysis P-value, Z-score and Direction of effect per study are shown. Directionality is given in the following order: European, African America, East-Asian, T2D-genes Hispanics1 , South-Asian. In addition, chromosomal position (chromosome and start position), allele count (AC) and BMI Gene-burden P-value (GB-P) and beta (GB-B) for all possibly deleterious missense variants from the T2D genetics portal are shown (NA indicates results were not available).

6.3.3 Identification of deletions using the exome chip array

6.3.3.1 A novel pipeline developed for the detection of coding deletions from exome chip array data

While exome sequencing provides an effective way of predicting coding CNVs, the associated sequencing costs are still relatively high, and the analysis of the data is time-consuming, in particular, when trying to analysis rare CNV events that require large sample numbers for detection, association testing and replication. To overcome these limitations, the second part of this chapter describes a novel approach to identify rare coding deletions from the Illumina Human Exome Bead ChIP array which represents a cost-efficient high throughput array with dense exonic coverage. Apart from low cost and high throughput, the availability of very large datasets for multiple diseases is another advantage of the array compared to exome or whole-genome sequencing.

To detect rare coding deletions from the exome chip, the deletion pipeline, which combined multiple CNV detection algorithms (Fig 6.4), was applied to human exome chip data. Initially, the pipeline was validated by applying it to exome chip data from 650 GoT2D samples for which WGS (and at a later point) WES deletion calls were available for validation. Before filtering PennCNV[305] and QuantiSNP[306] detected 999 and 6580 deletion calls, respectively, in these 650 samples. After filtering and merging deletions, a total of 20 unique deletion sites (equivalent to SNP positions in the context of SNVs) with an average size of 162kb were identified in 650 samples, 18 (90%) of which were rare (MAF <1%) or low-frequency variants (MAF <0.05, Table 6.5).

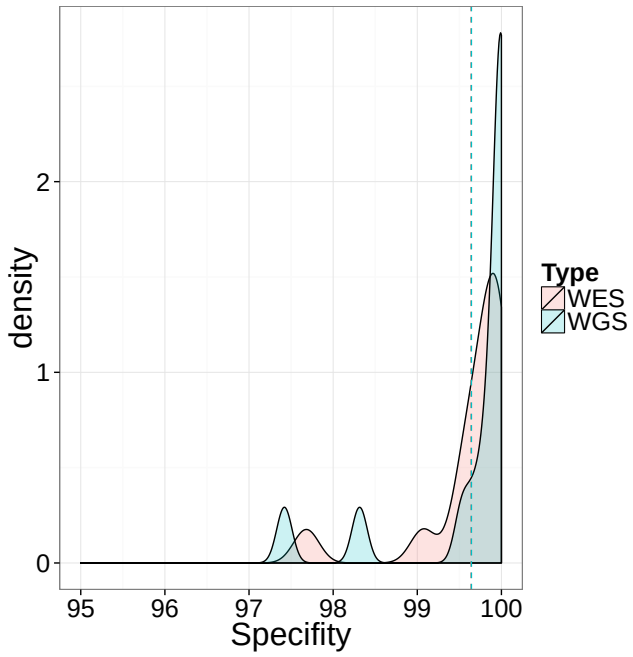
For validation, these deletion calls were overlapped with 448 WGS deletions, which overlapped at least 3 SNPs present on the exome chip, and 427 WES high-quality rare deletion calls generated from the same 650 GoT2D/UKT2D samples. In total, 15/20 (75%) deletion sites overlapped with 15/427 WES and 16/448 WGS CNV calls and there was good concordance between genotypes (based on the overlapping genotype between exome chip and WGS deletions, 1/16 WGS deletion was overlapping but did not correspond to any of the identified exome chip deletions suggesting a separate deletion event detected only by WGS). When comparing the deletion genotypes between overlapping exome chip and WGS deletion calls, the average genotyping sensitivity was 83.1% (although variable) and specificity was 99.6% (under the assumption that the WGS calls represent

the "true standard" for these sites). Similarly, high mean genotyping sensitivity (70.0%) and specificity (99.6%) was observed for the ExAC GoT2D CNV calls (Fig6.16). An example of the overlap between a small WGS, WES and exome chip deletion together with the LRR raw data of a deletion sample is shown in Figure6.17.

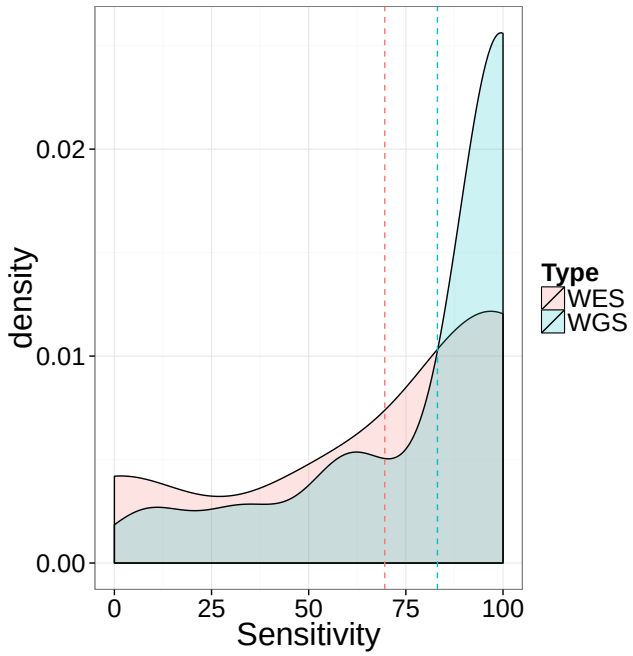
Overall, the concordance between the exome chip array deletions and the WGS and WES deletion datasets provided support for the detection accuracy of the newly developed pipeline especially considering the fact that neither the ExAC exome sequencing nor GoT2D WGS CNV calls can be regarded as gold standard.

Dataset	Deletions	Mean length (bp)	Rare	Low Freq	Common
650 samples	20	161,826	15	3	2
11.6k samples	208	124,865	185	20	3

Table 6.5: Exome CNV calls for validation set of 650 samples and 11.6k assocation data. For both sample sets (650 and 11.6k) the number of Deletions, the average deletion length, and the number of deletions across the AF spectrum is shown. Rare: MAF <0.01, Low Freq: MAF <0.05 and Common MAF >0.05



(A) Genotyping Specificity



(B) Genotyping Sensitivity

Figure 6.16: Genotyping specificity and sensitivity of 15 exome chip deletions overlapping WGS and WES deletion calls. Density (y-axis) plots showing specificity and sensitivity (x-axis) for the 15 exome chip deletions. Specificity and sensitivity distributions were calculated by considering the WES (red) or WGS (green) as "true" state.

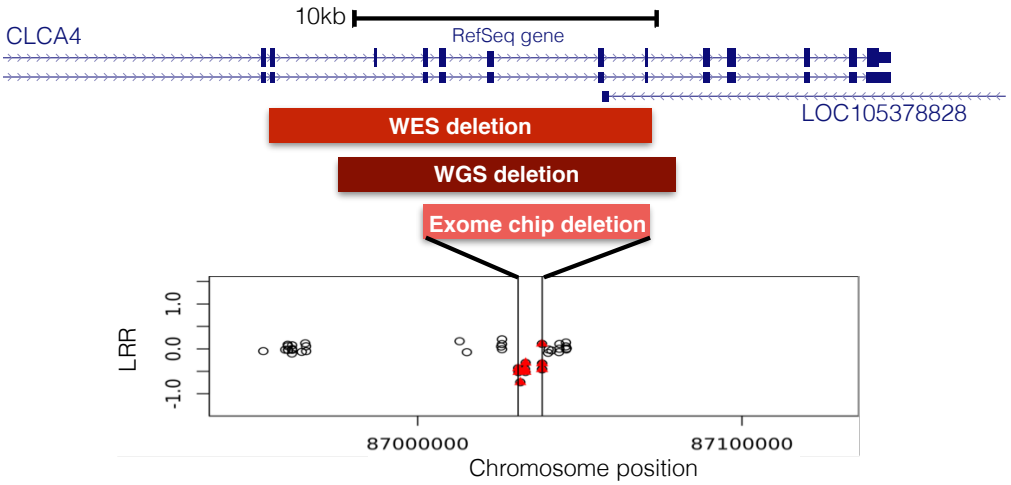


Figure 6.17: Overlap between small WGS, WES and exome chip deletion at the *CLCA4* gene Modified Genome Browser view of a WES deletion (top, medium red bar), WGS deletion (middle, dark red bar) and exome chip deletion (bottom, light red bar, size about 7.4kb). At the very bottom the SNP (dots) LRR values (y-axis) are shown along the chromosome (x-axis). Red coloured SNPs are found at the location of the exome chip deletion while the remaining SNPs are found in the region surrounding the deletion. (Different scales are provided for the genome-browser view and LRR plot).

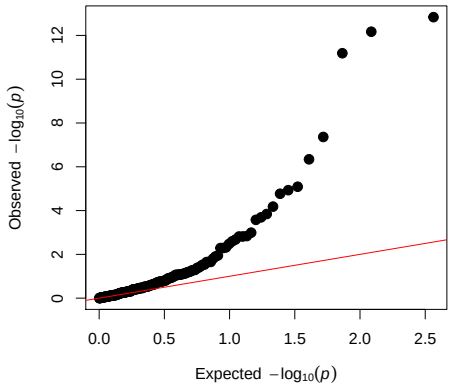
6.3.3.2 Testing exome chip deletions for association with T2D

To identify novel T2D associations and potentially validate some of the ExAC association results described in the first part of this Chapter (section 6.3.2.1), the newly developed pipeline was applied to 11.6k T2D-case-control samples of European ancestry (UK) including the 650 GoT2D samples that were shared between the exome chip and ExAC exome sequencing data. In total, PennCNV and QuantiSNP identified 288k and 404k deletions in these 11.6k samples. After filtering, and merging, 208 unique deletion sites with an average length of 125kb were identified. The vast majority of these deletions (205/208 or 99%) were low (MAF <5%) or rare (<1%) frequency events which suggests a similar allele frequency distribution as for the 650 GoT2D samples (Table 6.5).

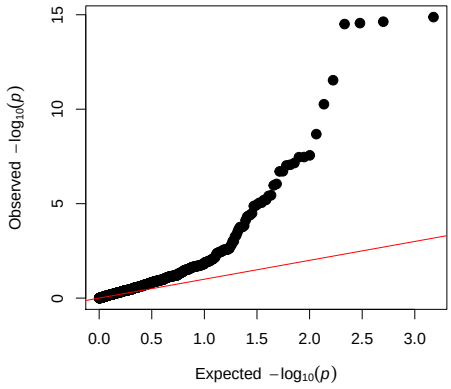
To test these deletions for association with T2D (in 11,414 unrelated samples with full phenotypic information including 3.9k cases and 7.5k controls), a gene-centred approach was applied. Similar to the ExAC gene-centred test, deletions were collapsed based on overlap with genes. After removing genes with an AC of less than three, 309 genes were tested for association combining all studies. The Firth bias corrected test was used for the association analysis with covariates correcting for sex, age, BMI and PCs derived from SNPs to account for population structure. In total, 5 significant associations (exome-wide significance P <2.5x10^{-6}) were identified. However, the association results were heavily inflated (Fig 6.18A).

Gene	PVALUE	BETA	AC	MAF
KIF7	1.46e-13	-0.657	1089	0.048
BAAT	6.77e-13	-0.988	552	0.024
C4orf40	6.51e-12	-0.604	1048	0.046
DDX28	4.35e-08	-1.477	156	0.007
RXFP3	4.56e-07	-1.854	95	0.004

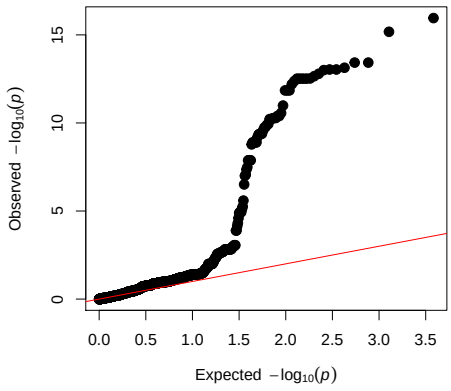
Table 6.6: Top 5 exome-wide significant T2D-associated deletions. For each variant, the overlapping gene, the T2D-case-control P-value and effect size (beta) are presented. In addition, allele count (AC) and Minor Allele Frequency (MAF) are shown .



(A) QQPlot gene-based testing



(B) QQPlot deletion-based testing



(C) QQPlot SNP-based testing

Figure 6.18: QQ-plot showing T2D-association results of deletions derived from 11.6k exome chip samples. QQplot showing expected (x-axis) vs observed -log10 P-values (y-axis) for gene-based (A), deletion-based (B) and SNP-based (C) T2D-association testing of exome chip deletion variants.

Similar association results were obtained for the two other approaches of collapsing deletions (Fig 6.5):

- a deletion-overlap-based approach identified 1150 merged deletions after merging deletions with at least 50% reciprocal bp overlap
- a SNP-based breakpoint approach used 4987 SNP probes on the array as surrogate markers for the deletion genotypes
- a total of 21 merged deletions and 72 proxy SNPs overlapping deletions were associated at exome wide significance (P value threshold $<2.5x10^{-6}$). However, these association results were also heavily inflated (Fig6.18B-6.18C).

Notably, the vast majority (all but one) of associated deletions had a T2D-protective effect. Considering the imbalance in case (3.9k) and control (7.5k) samples this may suggest spurious association results, in particular, since the majority of studies consisted of either only case (2.3/3.9k) or only control (5.4/7.5k) samples that were processed together; thus, amplifying potential batch effects. Additionally, the use of T2D-case-only or -control-only studies did prevent the application of a meta-analysis approach (as used for the ExAC CNV data). Hence, samples had to be tested for association as one cohort combined which could have further increased the influence of study-specific batch effects on the association results.

Next, the raw PennCNV and QuantiSNP exome chip deletion calls were used to try to replicate any of the T2D case-control allelic imbalance deletion results (*ABCC3, CBLC* and *VPSC13C*, see Table 6.3) of the ExAC sequencing deletion calls described in Section 6.3.2.1.2. Only *VPS13C* deletion calls were identified in the raw exome chip data; however, these 9 exome chip deletion calls (iden-ditfied by PennCNV) did not show any differences in cases (AC=5) compared to controls (AC=4). Overall these results suggest that the exome chip pipeline could neither be used to identify novel T2D associations nor to validate the ExAC CNV association results. Hence, additional validation will be required to determine the accuracy of the exome chip deletion pipeline.

6.3.3.3 Validation of identified exome chip deletions

6.3.3.3.1 *In silico* validation of deletions

To validate the identified exome chip CNVs, the deletion set, derived from merging deletions with reciprocal overlap, was used for *in silico* validation. These deletion were used since the deletion merging approach (Fig 6.5B) probably takes into account the structure of deletions most accurately. Specifically, these deletions were overlapped with CNVs present in DGV and the overlap was compared to random sites (derived from shifting the merged deletion sites 1000 times as described in section 6.2.2.5). Based on the permutation results, significant (P=0.02) overlap between the merged deletions (927/1150) and CNVs in DGV was found, indicating that the exome chip deletion pipeline accurately detects deletion sites in the genome.

6.3.3.3.2 Prioritisation of deletion variants for *in vitro* genotyping

Considering the significant and substantial overlap of deletion sites with DGV, the genotyping accuracy of the deletion pipeline was validated *in vitro*. Since accurate CNV genotyping *in vitro* is challenging, and a number of CNV genotyping methods exist, with varying success, depending on the deletion type, variants were prioritised to allow for efficient and cost-effective validation. Deletions were prioritised based on overlap with known deletion variants, preferentially identified from WGS, to obtain more accurate break point estimates. Additionally, deletions were selected for genotyping based on biological function of overlapping or nearby genes and low allele frequency threshold ($1\% < \text{MAF} < 5\%$). This allele frequency threshold was chosen to avoid testing rare or common CNVs of which the former may be present only in a few samples and not detected in any other dataset while the latter have likely been already tested for T2D-association by previous studies. Using these criteria, 4 deletions overlapping *KIF7, ANKRD35, DOCK8* and *TRIB3* were selected for *in vitro* genotyping (Table 6.19). In addition, exome chip deletions overlapping the gene *PTCHD3* were used as positive genotyping control considering that there was a clear overlap with a single WGS deletion identified in 1000 Genomes. Also, a previous study has already successfully genotyped *PTCHD3* deletions *in vitro*[345].

Deletion (gene)	Overlap known deletions	Biology (gene)	AF	Genotyping samples	T2D-signal
KIF7	Singleton sequencing deletion	Pancreas development + insulin secretion	Low	29 Deletion 7 Diploid	Exome-wide significant (P<2.5x10
ANKRD35	Multiple WGS deletions	*PIAS3* nearby is a sumoylation protein	Low	52 Deletion 44 Diploid	Exome-wide significant (P<2.5x10
DOCK8	Multiple WGS deletions	High expression in pancreas	Low	21 Deletion 75 Diploid	Bonferroni significant (P<1.16x10
TRIB3	Multiple WGS deletions	Disrupts insulin signalling	Low	37 Deletion 69 Diploid	NA
***PTCHD3* (control)**	Single WGS deletion	previously successfully genotyped	Low	22 Deletion 74 Diploid	NA

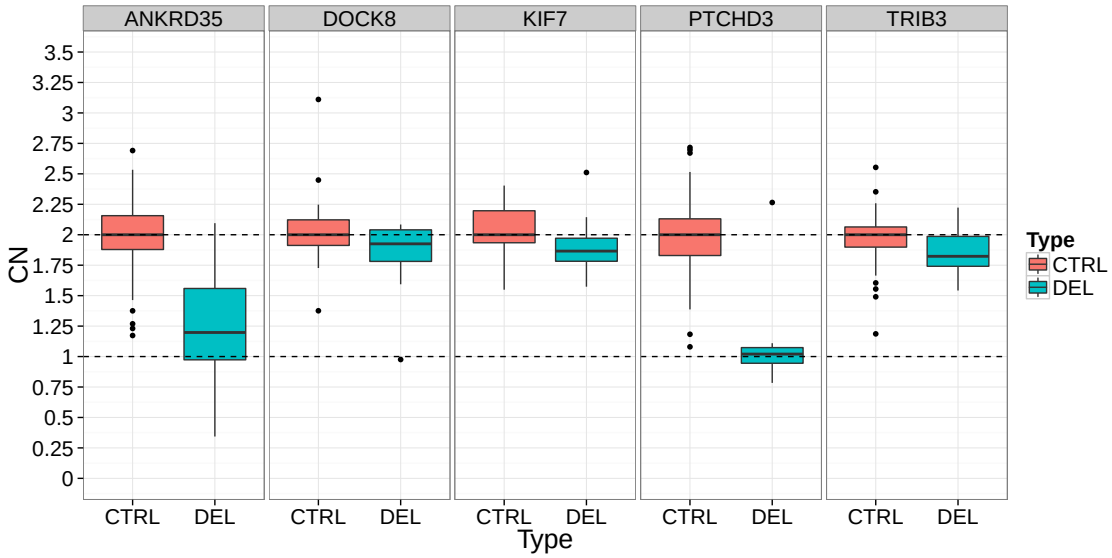
Figure 6.19: Prioritised deletions for in vitro genotyping. Each column describes a feature of prioritisation with the colour indicating whether a variant has good (green) or medium (yellow) support for functional follow-up. CNV variants were prioritised based on overlap with known DGV or GoT2D variants (preferentially individual sequencing variants), biology of the overlapping gene and the allele frequency. Additional columns indicate the number of samples that were *in vitro* genotyped and whether the variant was significantly associated with T2D (exome-wide significance and gene-based bonferroni adjusted multiple testing significance is indicated). References for *KIF7*[346–348] *PIAS3*[349], DOCK8[130], TRIB3[350] and PTCHD3[345].

6.3.3.3.3 *In vitro* genotyping of selected variants

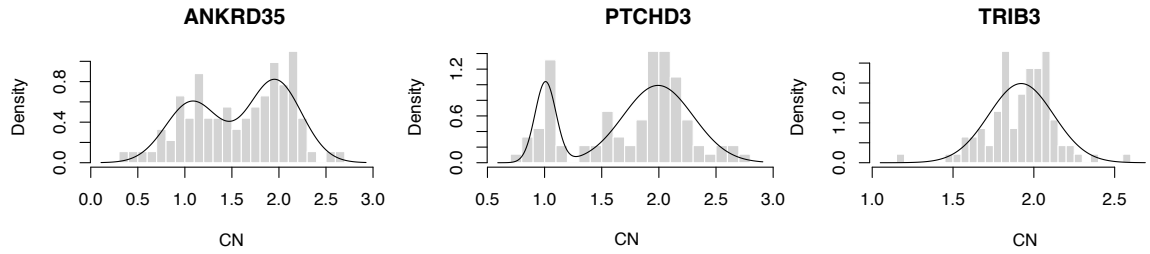
Taqman genotyping was performed for 4 prioritised gene deletions overlapping *ANKRD35, DOCK8, KIF7*, *TRIB3* and the deletion *PTCHD3*, which was used as positive control for genotyping, to determine copy number state per sample using the ΔΔCT method (see Chapter 2 Section 2.4 for more details). As shown in Fig 6.20A, for all gene deletion regions, predicted deletion samples had a lower copy number state compared to samples predicted to have a diploid genotype (predictions are based on the output from the exome chip deletion pipeline); however, only for 2 deletion regions (*ANKRD35* and *PTCHD3*) was the separation distinct enough that a Gaussian finite mixture modelling approach, as implemented in the R package mclust, determined two separate density distributions as best model (Fig 6.20B). For the remaining 3 deletions, mclust predicted only a single distribution as the best model (Fig 6.20B, only *TRIB3* shown).

In addition, to identifying separate distributions for *PTCHD3* and *ANKRD35*, mclust also provided probabilities for each sample to belong to either of the 2 distributions which could be used to estimate sample genotype probabilities based on the *in vitro* results. Specifically, samples that were predicted to be part of the lower copy number distribution were defined as samples with a deletion genotype while samples that were predicted to be part of the higher copy number distribution were defined as samples with a diploid genotype. These *in vitro*-based sample genotype predictions were then correlated with the *in silico* exome chip predicted genotypes and high correlation was observed for both *PTCHD3* (Pearson r=0.91) and *ANKRD35* (Pearson r=0.58). The correlation increased for both *PTDCHD3* (Pearson r=0.94) and *ANKRD35* (Pearson r=0.69) after removing samples that could not be clearly classified by mclust (probability <90% to belong to 1 of the 2 distributions).

In summary, these data clearly show that the exome chip pipeline can accurately predict deletion regions in the genome. Also, in accordance with the *in silico* exome chip results, *in vitro* genotyping copy number estimates of deletion samples were consistently lower compared to normal diploid samples; however, accurate deletion genotyping remains difficult for most sites using the *in silico* exome chip predictions. This is indicated by the inflated T2D-association results and the fact that only 2/5 deletion regions showed clear separation in *in vitro* copy number assays; thus, highlighting in general that it is difficult to accurately detect copy number states, which makes it difficult to test CNVs for association with binary case-control disease phenotypes.



(A) Taqman CN-estimates



(B) Distribution of CN estimates

Figure 6.20: In vitro genotyping to determine copy number state. (A) Copy Number (CN) estimates (y-axis) grouped according to exome chip deletion pipeline predictions (Control (CTRL) with normal CN or Deletion (DEL) sample, x-axis) for 5 gene deletions (*ANKRD35,DOCK8, KIF7, PTCHD3* and *TRIB3*). The dashed lines at CN 2 and 1 indicate diploid and heterozygous deletion state, respectively. (B) Clustering of CN states (x-axis) based on density (y-axis) estimation using Gaussian finite mixture modelling as implemented in mclust to determine whether CN states fall into distinct groups. Best estimate is shown as black line. Both *ANKRD35* and *PTCHD3* can be separated into 2 distributions while *TRIB3* is best characterised by a single distribution.

6.4 Discussion

The research in this Chapter has characterised the role of rare coding CNVs in the genetic risk of T2D and obesity and specifically, the analysis showed that

- the majority of coding CNVs is rare in line with their expected effect on gene function
- the CNV burden per individual is not associated with genetic T2D or obesity risk which is in contrast to results from previous studies that found CNV burden associated with obesity
- association testing identified a number of potentially relevant T2D- and obesity-associated coding CNVs among which deletions and duplications overlapping *VPS13C* were the most plausible biological candidates
- the deletion detection pipeline for the exome chip array likely can identify accurately deletion regions in the genome, however, accurate genotyping remains difficult, in particular, since no gold standard is available

The research in the first part of the chapter described WES coding CNVs derived from large-scale high depth exon sequencing T2D-case and control data. Consistent with results obtained from the full ExAC dataset of 60k individuals (of which the here presented data only represents a subset of 14.5k individuals) most of these variants were low-frequency which is expected considering the potential impact of some of these coding CNVs on gene function[157]. Similarly, large-scale coding SNV studies found that variants with large effects on gene function (missense, nonsense variants) are also very rare[284].

In addition to determining the general allele frequency distribution of these variants, the role of individual CNV burden in T2D and obesity risk was evaluated. Several neurodevelopmental disorders including schizophrenia and autism have been found to show significant CNV burden effects with affected individuals having a higher number of duplicated or deleted DNA sequences in the genome[351]. While the ExAC consortium found evidence for increased CNV burden in schizophrenia samples based on genic CNV intolerance scores[157], T2D and BMI were not associated with a CNV burden effect (based on CNV number). This suggests that deviations in overall exonic copy

number or disrupted gene sequence in the germline do not predispose to T2D and obesity. These results are inconsistent with a previous study that found large-scale rare deletions enriched in cases of extreme obesity[352]. However, differences in the CNV detection method (SNP array vs exome sequencing) may explain this difference considering that the SNP array may more accurately detect large deletions spanning substantial parts of the non-coding genome while exome sequencing is focused on the detection of variation in the coding sequence of the genome.

Apart from overall CNV burden, the large dataset provided an opportunity to study the effect of rare and low-frequency copy number variants in T2D and obesity across the whole exome. While several previous studies have linked rare copy number disorders to obesity[148,149,337,352,353], only a few studies have identified copy number variants potentially associated with T2D development[152,354,355]. The analysis in this chapter highlighted 157 rare gene-centred CNVs that either showed nominally significant associations with T2D (149 CNVs) or strong imbalances in T2D-case-control samples (8 CNVs); however, none reached exome-wide significance.

Using previously generated GWAS data (both on the single variant level and using PTV-based gene-burden results) to support the CNV association results, the analysis highlighted a number of genic CNVs that may be involved in the genetic risk of T2D.

For instance, at the known T2D GWAS locus *C2CD4A/VPS13C* , allelic imbalance in T2D cases and controls allele counts in opposite directions for VPS13C deletions (causal) and duplications (protective) provided further support for the functional involvement of *VPS13C* in T2D genetic susceptibility. Additionally, *VPS13C* is also associated with several other glycaemic traits including fasting proinsulin, 2h-glucose, 2h-insulin, fasting glucose[266,326,327,356–360]. The VPS13 family of proteins was shown to be involved in molecular processes related to membrane protein trafficking, Golgi structure, and the metabolism of phosphatidylinositol[361–363] and the gene *VPS13C* is highly expressed in human islets[326,327]. *VPS13C* has been suggested to be involved in insulin processing and beta-cell function, and a likely causal GWAS variant is associated with *VPS13C* expression in human islets[328]. Consistent with the deletion CNV being overrepresented in cases, a reduction in *VPS13C* gene expression is associated with higher T2D risk.

While the CNVs overlapping *VPS13C* represent the most plausible T2D-association result based on previous GWAS and functional data, the analysis also found several other genes with relevant biological function near T2D GWAS hits including *HEYL, YTHDC2, CBLC* and *GAK*. All of these genes showed relatively high expression in human pancreatic islets based on publicly available islet expression data[236] and had potentially biologically relevant functions related to development, TF binding activity and cell signalling[329,331–336]; thus, prioritising these genes at the GWAS loci for functional follow-up.

For the obesity association results, a confirmatory BMI-association signal was found for a deletion overlapping the gene *GDPD3* in the region 16p11.2 that was nominally associated with underweight. This region is linked to the 16p11.2 deletion and duplication syndrome[148,149,337] which are consistently associated with obesity and underweight, respectively, depending on the copy number state. In addition, several other interesting candidate genes were identified: For instance, deletions at *SLC5A4* were found to potentially increase obesity risk in African Americans and accordingly, PTV SNV in *SLC5A4* were associated with an increase in obesity in the same ancestry group; thus, supporting the observed associations. Previous studies suggested that *SLC5A4* encodes a glucose sensor found in cholinergic neurons of the intestine and skeletal muscle and regulates muscle activity[364]. In accordance with the association results, SLC5A4 was shown to increase glucose transport into skeletal muscle after exercise which may lead to a decrease in insulin resistance[341].These data support the observation that a reduction in *SLC5A4* may increase the risk for obesity which often is a direct cause of insulin resistance. In addition, recently SLC5A4 was also suggested to sense glucose in the portal vein and activate brain regions involved in food intake[365], thus, providing another potential link between *SLC5A4* deletions and obesity.

Similarly to *SLC5A4*, BMI SNV gene-burden results also provided some support for other CNVs, such as duplications overlapping *ITGAX* and *HNRNPC* in the involvement of obesity risk; however, based on the current understanding of the biological function of these genes, it remains unclear how the overlapping CNVs could be influencing genetic obesity risk.

The association analyses for both BMI and T2D also highlighted the difficulty in accurately studying the relationship between copy number and disease considering that the detection still remains challenging, in particular, in regions with complex copy number states. Several of the highlighted

variants were found in genomic regions with low copy number repeats or complex clusters of replicated genes including the beta-defensin *DEFB131* found in gene clusters[342]. Similarly, the T2D associated *GOLGA6B* region is also found in a region with low copy number repeats[321]. Due to the repeats and complex copy number state, the short reads generated by NGS can often not be mapped properly which may lead to spurious and conflicting results between studies. For instance, it was recently claimed that CNVs at the *AMY1* gene locus, which has a complex structure with multiple gene copies, may explain a large portion of BMI heritability[292]. However, a follow-up study did not find any significant association with obesity indicating that the previous results were observed due to genotyping errors[293].

The difficulty in accurately detecting CNVs and testing them for association with a phenotype was also highlighted by the T2D-association results for deletions identified from a newly developed exome chip discovery pipeline. As suggested by previous studies the pipeline combined multiple algorithms, harnessing their different properties, to detect deletions from the human exome bead chip which provides dense coverage of exonic regions at low cost. Although good concordance was observed between deletion genotypes derived from exome chip, exome sequencing and whole-genome sequencing data in a subset of 650 individuals, deletion detection in 11.6k exome chip samples followed by association testing led to highly inflated T2D-association results. *In silico* validation based on overlap with previously identified CNVs coupled with *in vitro* qPCR genotyping suggested that deletion sites in the genome can be accurately detected using the pipeline; however, precise genotyping was only possible for a subset of the deletions which provides a potential explanation for the inflated association results. In contrast to the results observed in this Chapter, parallel efforts from a different group have used the exome chip successfully for both the detection of large CNVs and the replication of schizophrenia-associated CNVs called from a different whole-genome array. Differences in the detection method, CNV size and cohort structure, in particular, the imbalance in T2D-case-control samples in this study, could explain the opposing results.

The analysis in this chapter provided an insight into the role of rare coding CNVs and their role in diabetes and obesity. The study also highlighted for the *C2CD4A/VPS13C* locus how rare coding CNVs can be used to prioritise genes in large GWAS regions with unknown effector transcript. However, the lack of existing gold standards for CNV calling and the low-frequency of the identi-

fied CNVs makes the replication and validation of these association results difficult. In addition, the underlying structure and size of the identified CNVs remains at least partially unclear since exome sequencing does not provide accurate information about breakpoints. Hence, it may be difficult to design cost-effective and accurate assays to follow up CNVs of interest highlighted by the analysis. Novel approaches such as a deletion discovery pipeline from the exome chip array may provide one way of cost-effective high-throughput CNV calling; however, extensive optimisation will be required to improve the deletion genotyping performance of the pipeline. Future large-scale CNV association studies will be required to validate the results paired with functional assays to determine the molecular mechanisms governing the CNV-associated T2D and obesity risk.

CHAPTER 7

General Discussion

The analyses performed in this thesis addressed the following primary objectives:

- to combine newly generated whole-genome human pancreatic islet DNA methylation and ATAC-seq open chromatin data with previously generated ChIP-seq histone marks to refine islet regulatory states
- to integrate these islet regulatory states with T2D GWAS data in order to define potential causal mechanisms at T2D GWAS loci governing genetic risk of T2D
- to characterise epigenetic changes involved in endodermal development to improve understanding of the molecular processes regulating endoderm and pancreas formation
- to detect exonic CNVs, which likely affect gene function, and identify variants associated with T2D susceptibility

The work in Chapter 3 extended the catalogue of available human islet epigenetic data and investigated for the first time the whole-genome DNA methylome of human islets providing the opportunity to study the link between DNA methylation, islet regulation and T2D development. A comparison between islet WGBS and Illumina 450k array methylation data clearly showed that WGBS provides a much more comprehensive analysis of the islet methylome, in particular, around GWAS regions. The inability to accurately characterise GWAS regions is predominantly based on the design of the 450k array that focusses on CpG-island and promoter regions which are often only weakly associated with disease compared to enhancer elements. Further analysis of the whole-genome islet methylation data identified two types of hypomethylated regulatory domains defined as UMRs and LMRs that corresponded to promoters and enhancers, respectively. These regions were strongly enriched for known islet regulatory annotations such as eQTLs, mQTLs and islet TFBS. Enhancer-like LMRs also showed enrichment in T2D GWAS signals.

In agreement with these results, it was previously demonstrated that methylation plays a major role in cellular regulation and transcriptional activity (reviewed in [366]). Specifically, LMRs were shown to be associated with TF binding [172,206,367], cell type specificity [112,172,206,368,369] and the expression levels of genes [370]. While only limited DNA methylation data exists for human pancreatic islets, previous studies have also confirmed a link between islet DNA methylation identified using the 450k array and islet regulation [99], T2D status [198] and ageing [199]. Correlations between genetic variation, gene expression and DNA methylation suggest that these factors form a complex regulatory network in human islets [99] and other tissues [101]. As mentioned above only limited DNA methylation data is currently available for human pancreatic islets , and the data in Chapter 3 substantially expanded the knowledge about the role of DNA methylation in islet regulation and T2D development.

Similar to Chapter 3, the analysis presented in Chapter 4 also extended the set of regulatory annotations in human islets. In addition, the interrogation of open chromatin in iPSC subtypes identified developmentally active open chromatin sites that likely govern endodermal formation and, thus, is also relevant for endodermal islet function. The first part of the chapter characterised the chromatin landscape of human islets using ATAC-seq and identified islet open chromatin regulatory elements enriched in T2D GWAS signals. These results are highly consistent with several previous studies that have determined the regulatory role of human islet open chromatin and active enhancers in the

genetic susceptibility of T2D[84–86]. Adding to these studies, the islet ATAC-seq data generated in this thesis extended the static views of the DNA accessibility landscape in human islets and provided open chromatin maps on an individual sample basis. These variable chromatin maps were exploited in Chapter 5 to highlight potential functional mechanisms at T2D GWAS loci through allelic imbalance in chromatin accessibility.

The second part of Chapter 4 investigated the open chromatin landscape of human iPSC subtypes that were generated from beta- and fibroblast-cells and differed in their propensity to differentiate into endoderm. To determine epigenetic markers associated with the enhanced differentiation potential into endodermal cells, the study identified differentially open chromatin regions (DOCS) between the iPSC subtypes and observed that Bi-DOCS (sites with more open chromatin in BiPSCs compared to FiPSCs) showed strong enrichment in genomic annotations related to pancreatic endodermal development. Consistently, gene and pathway enrichment analysis of these Bi-DOCS highlighted genes associated with both pancreas development and mature function as well as neurodevelopmental pathways. Specifically, NOTCH and WNT signalling and TFs such as FOXA2 and PDX1 were associated with Bi-DOCS and enhanced endodermal differentiation ability. In accordance with these results, the leading 'iPSC to endoderm' differentiation protocols, which generate functional human beta-cells *in vitro*, target the NOTCH and WNT pathway[219,220]. Additionally, FOXA2 was shown to provide developmental competence during endodermal pancreas formation by priming weak enhancer sites which are then bound by PDX1 to initiate stage-specific gene expression programmes[217].

Following the generation of novel epigenomic annotation in human islets, the work in Chapter 5 combines the whole-genome methylation and open chromatin datasets with human islet ChIP-seq annotations to refine existing chromatin states. These states were then used to evaluate the contribution of DNA methylation and open chromatin to T2D GWAS signal enrichment. This approach showed that DNA methylation and ATAC-seq data can be used to identify strong and weak enhancer subtypes that differ in chromatin accessibility, methylation status and the overall enrichment in T2D GWAS signals. Consistent with a more active state, it was found that hypomethylated open chromatin islet enhancers showed the highest levels of enrichment in T2D-GWAS loci. Previous studies have confirmed that the integration of histone ChIP-seq marks with chromatin

accessibility and/or hypomethylation information can improve GWAS enrichment of regulatory annotations[85,131]. Subsequently, posterior probability estimates were calculated taking into account both the GWAS association and chromatin state enrichment, to prioritise variants at GWAS loci for testing them in open chromatin imbalance. In total, at 3 GWAS loci including *ADCY5, CDC123* and *KLHDC5*, high posterior probability variants overlapping enriched enhancer states showed allelic imbalance in open chromatin, thus indicating potential molecular mechanisms governing T2D risk at these loci. For instance, at the *ADCY5* locus the T2D risk allele of the GWAS lead variant rs11708067 is predicted to reduce HNF4A binding and also associated with reduced *ADCY5* gene expression[159] and lower *ADCY5* promoter methylation[99]. Overall, these results indicate the power of an integrative epigenomic analysis to identify islet regulatory elements that enhance understanding of islet function and indicate potential mechanisms at T2D GWAS loci.

Chapter 6 aimed to characterise the role of coding CNVs in the genetic susceptibility of T2D and obesity. The first analysis of this chapter focussed on whole-exome deletions and duplications, which were generated externally by the ExAC consortium from a T2D-case-control exome sequencing study comprised of 14.5k individuals. In contrast to other complex diseases such as schizophrenia[371] and autism[372], no correlation between CNV number per individual and T2D or obesity could be identified. The lack of association between CNV burden and obesity is inconsistent with previous studies that have found that large rare deletion events are enriched in cases of extreme obesity[352]; however, differences in CNV detection methods (array vs sequencing) and target regions (coding vs non-coding CNVs) may explain this discrepancy between the results.

In addition to the CNV burden analysis, a gene-centred approach was used to identify genic CNVs that correlate with T2D status or BMI. A number of coding deletions and duplications were found to be associated with T2D or obesity including several variants overlapping genes with highly relevant biological function. For instance, imbalance in T2D-case-control CNV allele counts were observed in the gene *VPS13C* which is found at a GWAS locus associated with T2D-related traits[266,326,327,356–360] and encodes a protein suggested to influence beta-cell function and insulin processing[328]. Rare *VPS13C* deletions were only found in T2D-cases while all but one duplication were observed in non-diabetic individuals. These results are directionally consistent with *VPS13C* gene burden tests (of single nucleotide protein truncating variants) and an eQTL GWAS variant

which is associated with decreased *VPS13C* gene expression and increased risk of T2D. Similarly, deletions at the 16p11.2 locus were found to be associated with increased BMI which is strongly concordant with the fact that rare deletions in these regions were previously found to be linked to severe obesity[337].

The second part of Chapter 6 focussed on the development of a novel deletion detection pipeline from exome chip array data with the aim of validating CNV disease associations identified from the exome sequencing data. This approach paralleled the work performed by Szatkiewicz et al 2014[371] that have also used the exome array for validation and replication of CNV calls and CNVs disease-associations derived from a different whole-genome array. While the exome chip deletion calls generated in this thesis suggest that accurate detection of deletion sites is possible, the highly inflated T2D-association results and *in vitro* validation indicate that only some sites were accurately genotyped. This is inconsistent with the results from Szatkiewicz et al 2013 and 2014[371,373] which have successfully called large CNVs from the array and validated the association results. This may be explained by differences in the detection approach, the size of investigated CNVs and cohort structure (imbalances in number of cases vs controls).

Immediate steps following the work completed in this thesis would be to validate the results observed at disease associated GWAS loci as well as functional characterisation of regions involved in endodermal development. Through the integrative analysis, which combines multiple layers of genetic and epigenetic data, six strongly T2D-associated enhancer variants with allelic imbalance were identified at the T2D GWAS loci *ADCY5, CDC123* and *KLHDC5*, thus, suggesting potential molecular mechanisms governing T2D genetic susceptibility at these loci; however, these results require functional validation to confirm the observed associations. For instance, at the *ADCY5* variant rs11708067 at which genotype-dependent differences in HNF4A binding were predicted, Electrophoretic Mobility Shift Assays or ChIP-qPCR in a relevant human islet cell model such as the EndoC-ßH1 line could be used to confirm the suggested allele-specific HNF4A binding event (partially dependent on the genotype). This approach has been successfully used previously to validate differential TF binding events in human islets[83,85]. While the previously identified eQTL indicates that the variant influences *ADCY5* gene expression, further support for this variant regulating *ADCY5* could be provided by chromatin conformation capture. For instance, this experimental

technique helped to uncover the likely causal variant and mechanism at the obesity-associated locus *FTO*[374].

The epigenomic data created in Chapters 3-5 was generated from a rather small number of human pancreatic islets despite being one of the first (for the whole-genome DNA methylation dataset) and relatively large (for the ATAC-seq open chromatin) studies to investigate both DNA methylation and open chromatin in human islets. While these datasets provided highly useful information about the DNA methylation landscape and the variability in open chromatin, these datasets are still relatively small compared to other epigenetic or gene expression studies and do not include samples from diabetic individuals.

Several methylation studies in islets[99,196,198,199] or other tissues (for which samples are often more readily available) have shown that larger sample numbers allow for the identification of mQTLs[96–104] or disease differentially methylated regions (DMRs)[190,191,193,196,198,375–382] and can uncover highly relevant biology including T2D-relevant results[99,190,191,193,196,198]. However, most of these studies have used array-based or sequencing-based technologies that do not provide full whole-genome methylation levels at single base-pair resolution, thus, being limited in their ability to interrogate the full DNA methylome. In a similar way, large-scale RNA-seq studies in both islets[159,268,383] and other tissues[130,384] have proven useful to link genetic variants to genes through the identification of eQTLs for which enough statistical power is required to accurately detect the association. Based on these previous studies the collection of larger number of samples to investigate both DNA methylation or ATAC-seq open chromatin would provide several key advantages:

For DNA methylation, an increase in sample size would provide more information about the variability of whole-genome DNA methylation in human islets. This could provide the opportunity to perform Epigenome-Wide Association Studies (EWAS). These studies could detect interactions between genetic variation and islet DNA methylation as well as differentially methylated regions between diabetic and non-diabetic individuals on a true genome-wide scale compared to previous studies[99,196,198]. Similarly, the availability of large ATAC-seq open chromatin datasets would also be beneficial for the identification of genome-wide chromatin QTLs, disease-associated chromatin changes and differential open chromatin sites associated with exposure to different metabolites (glucose, fatty acids) or pharmaceutical agents. For instance, previously the effect of palmitate fatty

acid exposure on DNA methylation and transcription was investigated[385].

Currently our group is collecting additional human islet WGBS methylation and ATAC-seq data including samples exposed to high and low glucose conditions to enhance our understanding of epigenetic regulation in human islets and the effect of glucose exposure on the islet epigenome.

Human islets also consist of multiple different cell subtypes including alpha, beta, delta, epsilon and polypeptide cells and recent studies clearly have indicated that there are epigenetic and transcriptional differences between alpha and beta cells[228,236]. Since GWAS results have provided strong evidence for the functional role of beta-cells in T2D pathophysiology[18,19] and epigenetic studies have shown that genomic annotations are often tissue-specific[82], the characterisation of the epigenetic landscape of these cell subtypes could provide further insight into T2D development. For instance, Ackermann et al showed substantial overlap of alpha- and beta-cell-specific open chromatin signatures with T2D GWAS hits[228]. Advances in technology provide even the opportunity to investigate transcriptional and epigenetic profiles in single cells[386,387] and recent studies have applied RNA-seq to human islet cell subtypes to characterise differences across single islet endocrine cells[388,389].

Considering that the availability of primary human islets is often limited, and samples from human donors show high variability due to environmental factors, exocrine contamination, and differences in islet cell composition, future work would highly benefit from efficient iPSC to endocrine pancreas/islet differentiation protocols to generate human islet cell model. Until such protocols are available, the work provided in Chapter 4 is important to enhance the understanding of endodermal pancreas differentiation processes and characterise the key pathways associated with developmental progression. Future work targeting the identified developmental pathways, which were associated with enhanced differentiation ability in BiPSCs, could lead to improvements in these differentiation protocols. Immediate next steps related to the iPSC work in Chapter 4 should aim to replicate the results in a larger sample set, track epigenetic changes throughout development and integrate different types of epigenetic and transcriptional datasets. Our lab has previously published an RNA-seq study that investigated transcriptional changes in several stages of endodermal development[227] and we are currently collecting samples from a number of stages to interrogate both the DNA methylation and open chromatin landscape of these developmental cell types. In addition to providing

further insight into endodermal development, the RNA-seq study also highlighted that such an approach could help to determine whether T2D susceptibility variants rather act during development or directly influence mature islet function.

For most of the results in Chapter 3-5 no definitive statement can be made about causality considering that these results indicate only associations between epigenetic marks, which are variable over time, and T2D disease status. In contrast to GWAS which investigates associations between genetic variants that (usually) do not change over time, it remains unclear whether the epigenetic changes associated with T2D risk are a cause for or a consequence of the disease. Ultimately, efficient differentiation protocols could be used to obtain easily accessible cells from living donors to generate iPSCs harbouring genetically modified T2D-risk variants using CRISPR[390,391]. The effects of these genetic changes could then be studied throughout development and in mature islets, thus, allowing to dissect the direction of effect between phenotype and genetic and epigenetic variation. A recent publication also indicated that the CRISPR-Cas9 system could be modified to perform targeted methylation[392] or demethylation[393] of genomic regions, thus, potentially providing the opportunity to study the regulatory effects of changes in the DNA methylation pattern *in vitro* more readily.

Future work related to the analysis in Chapter 6 should involve extensive validation of CNV calls paired with *in vitro* functional studies of selected genes, such as *VPS13C*, with prior knowledge about biology and support from SNV genetic associations. Focussing on genes with known biology and GWAS support will facilitate the accurate choice of cellular models which is critical to determine the role of these CNVs in T2D risk. Depending on the predicted effect of CNVs, gene knockout experiments or overexpression studies in relevant models may help to confirm disease associations and understand the phenotypic effect of CNVs on the molecular level. Advances in qPCR technology, such as digital droplet PCR[394,395], have been shown to substantially improve CNV detection and genotyping assays, thus, making *in vitro* validation of prioritised CNVs more feasible. Furthermore, the association results should be replicated in different datasets which will require large sample sets for rare variants. Hence, further development and refinement of novel detection methods applicable to existing large-scale GWAS case-control datasets including exome chip array data could be highly useful. Alternatively, for rare strong effect size it may also be possible to replicate associations in population isolates in which rare variants may have risen to

increased allele frequency or in cohorts with extreme phenotypes that are often enriched for strong effect size variants.

In conclusion, the work performed in this thesis has targeted several fundamental challenges in understanding human islet function and genetic susceptibility of T2D. Firstly, the epigenomic landscape of human pancreatic islets and iPSCs was annotated to improve understanding of pancreatic islet developmental and mature function as described in Chapter 3 and 4. Secondly, the integration of epigenetic human islet datasets with T2D-related genetic association data was performed to prioritise T2D susceptibility variants at GWAS loci and highlight potential molecular mechanisms governing T2D risk (Chapter 5), thus, indicating the usefulness of combining multiple layers of epigenetic and genetic data to understand T2D genetic susceptibility. Thirdly, the knowledge regarding the genetic architecture of T2D was expanded through characterising the role of exonic CNVs in the genetic risk of T2D as described in Chapter 6.

In the future continued experimental and computational investigation of these results will be important to explain the genetic factors and mechanisms predisposing to T2D and critical for developing novel therapies for T2D to prevent and cure the global rise in T2D affected individuals.

Appendices

Additional Tables

Tables Chapter 4

Term	Description	Dir	Genes
Go	neural tube patterning	Bi	EN1, FGF8, FOXA2, GAS1, GBX2, GLI2, GLI3, GSC, HES1, IFT122, LRP6, MNX1, PAX6, PAX7, PSEN1, PTCH1, SHH, SSBP3, TCTN1, TMEM107, TULP3, WNT3A
Go	positive regulation of embryonic development	Bi	FOXA2, LHX1, NR2C2, OTX2, PLCB1, RBM19, WNT3A, WNT4
Go	non-canonical Wnt receptor signaling pathway	Bi	CELSR1, FZD1, FZD10, FZD2, FZD7, GRHL3, ROR2, SFRP1, SFRP5, WNT11, WNT4, WNT5A, WNT7A, ZNRF3
Go	hindbrain morphogenesis	Bi	ATP2B2, CACNA1A, DAB1, FAIM2, GAS1, GLI2, GNPAT, GSX2, HERC1, HES1, KNDC1, LHX1, LHX5, LMX1B, LRP6, NFIX, PCNT, RFX4, RORA, SKOR2, TRNP1, ULK1, WNT7A
Go	branching involved in mammary gland duct morphogenesis	Bi	BTRC, DDR1, EPHA2, ESR1, ETV4, LRP6, MSX2, NCOA3, PML, SRC, TBX3, TFAP2C, VDR, WNT4, WNT5A
Go	cerebellum morphogenesis	Bi	ATP2B2, CACNA1A, DAB1, FAIM2, GAS1, GLI2, GNPAT, HERC1, KNDC1, LHX1, LHX5, LMX1B, LRP6, NFIX, PCNT, RFX4, RORA, SKOR2, TRNP1, ULK1, WNT7A
Go	mammary gland duct morphogenesis	Bi	BTRC, CSF1R, DDR1, EPHA2, ESR1, ETV4, FGFR2, GLI2, LRP6, MSX2, NCOA3, NR3C1, NTN1, PML, PTCH1, PTHLH, SRC, TBX3, TFAP2C, VDR, WNT4, WNT5A
Go	mammary gland epithelium development	Bi	BTRC, CCND1, CEBPB, CSF1R, DDR1, EPHA2, ESR1, ETV4, FGF2, FGFR2, FOXB1, FOXF1, GLI2, ID2, LRP6, MGMT, MSX2, NCOA3, NR3C1, NTN1, PML, PTCH1, PTHLH, SRC, TBX3, TFAP2C, VDR, WNT3, WNT4, WNT5A, WNT7B, ZNF703
Go	dorsal spinal cord development	Bi	DRAXIN, GSX1, GSX2, LBX1, LHX1, LHX3, LHX5, MDGA1, PAX3, PAX7, PBX3, RFX4, UNCX, WNT3A
Go	cerebellum development	Bi	AGTR2, ATG7, ATP2B2, ATRN, CACNA1A, CDK5R1, CYP11A1, DAB1, FAIM2, GAS1, GBX2, GLI2, GNPAT, HERC1, KNDC1, LHX1, LHX5, LMX1A, LMX1B, LRP6, MDK, MYH10, NEUROD2, NFIX, NLGN4X, NR2C2, PCNT, PTPRS, RFX4, RORA, SEMA4C, SKOR2, SSTR2, TRNP1, ULK1, WNT7A, ZBTB18

Go	cerebellar cortex morphogenesis	Bi	ATP2B2, CACNA1A, FAIM2, GLI2, HERC1, KNDC1, LHX1, LHX5, NFIX, PCNT, RFX4, RORA, SKOR2, TRNP1, ULK1, WNT7A
Go	spinal cord association neuron differentiation	Bi	GSX1, GSX2, LBX1, LHX1, LHX3, LHX5, MDGA1, PAX3, PAX7, WNT3A
Go	dorsal/ventral neural tube patterning	Bi	FOXA2, GAS1, GLI2, GLI3, GSC, IFT122, MNX1, PAX7, PSEN1, PTCH1, SHH, TCTN1, TULP3, WNT3A
Go	negative regulation of Notch signaling pathway	Bi	BMP7, DLK1, EGFL7, FBXW7, GATA2, HEY1, HIF1AN, MAGEA1, NEURL, NFKBIA, NRARP, SLC35C1
Go	cell differentiation in spinal cord	Bi	ABT1, CACNA1A, CLN8, DBX1, DICER1, DRAXIN, GATA2, GLI2, GLI3, GSX1, GSX2, HOXC10, ISL1, LBX1, LHX1, LHX3, LHX4, LHX5, MDGA1, MNX1, NKX2-2, NKX6-1, NKX6-2, OLIG2, PAX3, PAX6, PAX7, PHOX2A, PTCH1, SHH, SOX1, SUFU, TBX20, TCTN1, WNT3A
Go	cell differentiation in hindbrain	Bi	ATP2B2, CACNA1A, FAIM2, FOXA2, GATA2, HERC1, KNDC1, LHX1, LHX5, NFIX, NOG, RORA, SKOR2, ULK1, WNT7A
Go	cerebellar cortex development	Bi	AGTR2, ATG7, ATP2B2, CACNA1A, FAIM2, GLI2, HERC1, KNDC1, LHX1, LHX5, MDK, MYH10, NEUROD2, NFIX, PCNT, RFX4, RORA, SKOR2, TRNP1, ULK1, WNT7A
Go	somitogenesis	Bi	AXIN2, BMPR1A, COBL, CREBBP, DLL1, EPB41L5, FOXA2, FOXB1, FOXC1, FOXF1, GDF3, KDM6A, LHX1, LRP6, MEOX1, NKX3-1, OTX2, PAX1, PCDH8, PLXNA2, PPP2R3A, PRKDC, PSEN1, ROR2, SFRP1, SMAD3, T, TBX18, TCF15, WNT3A, WNT5A
Go	white fat cell differentiation	Bi	AACS, ADIG, CEBPA, CTBP2, NCOR2, PRDM16, SIRT1, SNAI2
Go	embryonic axis specification	Bi	CITED1, COBL, EPB41L5, FOXA2, FRS2, GDF3, HOXD8, KDM6A, LHX1, OTX2, PCSK6, PLD6, PTCH1, SMAD2, T, TBX3, WNT5A, WNT7A
Go	ventral spinal cord interneuron fate commitment	Bi	DBX1, GLI2, GLI3, LHX3, NKX2-2, NKX6-1, NKX6-2, PAX6, SOX1, SUFU
Go	regulation of insulin receptor signaling pathway	Bi	GRB10, IGF2, IL1B, INS, KANK1, PRKCB, PRKCD, PRKCQ, PRKCZ, PTPN1, PTPN2, PTPRE, SIK2, SIRT1, SOCS1, SOCS3
Go	positive regulation of focal adhesion assembly	Bi	EPB41L5, ITGB1BP1, KDR, PTPRJ, SDC4, SFRP1, SMAD3, VEGFA, WNT4

Go	cerebellar cortex formation	Bi	ATP2B2, CACNA1A, FAIM2, HERC1, KNDC1, LHX1, LHX5, NFIX, RORA, SKOR2, ULK1, WNT7A
Go	regulation of Notch signaling pathway	Bi	BMP7, DLK1, DLL1, DTX1, EGFL7, EYA1, FBXW7, GATA2, HES1, HEY1, HIF1AN, ITGB1BP1, JAG1, LFNG, MAGEA1, MESP1, MFNG, NEURL, NFKBIA, NRARP, SLC35C1, SLC35C2, SOX2
Go	spinal cord dorsal/ventral patterning	Bi	DBX1, GLI2, GLI3, LHX3, NKX2-2, NKX6-1, NKX6-2, PAX6, SHH, SOX1, SUFU, TULP3
Go	ventral spinal cord interneuron differentiation	Bi	DBX1, GATA2, GLI2, GLI3, LHX3, NKX2-2, NKX6-1, NKX6-2, PAX6, SOX1, SUFU
MSigDB	Wnt signaling network	Bi	FZD1, FZD10, FZD2, FZD7, FZD8, FZD9, KREMEN1, LRP6, ROR2, RSPO1, RYK, WNT3, WNT3A, WNT5A, WNT7A, WNT7B
MSigDB	Maturity onset diabetes of the young	Bi	FOXA2, FOXA3, HES1, HNF1A, HNF4A, INS, MAFA, MNX1, NEUROG3, NKX2-2, NKX6-1, PAX6, PDX1, PKLR, SLC2A2
Go	dendrite morphogenesis	Fi	ATP7A, CACNA1A, CTNNA2, DACT1, DCX, EPHB1, FYN, KLF7, MAP2, MEF2A, PICALM, SDC2, SEMA3A, SHANK1, SLITRK5
Go	long-term memory	Fi	ADCY1, ADCY8, BTBD9, CRH, CTNS, GRIA1, GRIN1, NTRK2, PJA2, RELN, SGK1, SHANK1, TACR1
Go	positive regulation of acute inflammatory response	Fi	C3, CCR7, CNR1, IL1B, IL6, IL6ST, NPY5R, OSM, PIK3CG, PTGER3, PTGS2, TNF

Table 1: Enrichment terms associated with DOCS. Each row represents an enriched term associated with DOCS. Columns indicate Term (Go Biological Process or MSigDb term), Description, Direction (Dir with Bi indicating BiDOCs are associated while Fi indicates FiDOCs are associated with this term), Regional enrichment (RE), Gene Enrichment (GE), Binomial BH/FDR-adjusted P (BBH-P, for regional enrichment), Hypergeometric BH/FDR-adjusted P (HBH-P, for gene enrichment) and Type categorising the enriched term into broad biological classes.

Term	Description	State	FE	HBH-P	Gene
Go	olfactory behavior	BivEnh	71.2	1e-19	DRD4, GRIN1, SHANK1
Go	chemosensory behavior	BivEnh	54.4	8e-18	DRD4, GRIN1, SHANK1
Go	determination of affect	BivEnh	96.9	2e-17	SHANK1
Go	habituation	BivEnh	96.9	2e-17	SHANK1
Go	synapse maturation	BivEnh	90.9	3e-17	SHANK1
Go	sensory processing	BivEnh	85.5	7e-17	SHANK1
Go	positive regulation of excitatory postsynaptic membrane potential	BivEnh	39.1	1e-14	GRIN1, SHANK1
Go	long-term memory	BivEnh	20.6	1e-14	GRIN1, NTRK2, SGK1, SHANK1, TACR1
Go	vocalization behavior	BivEnh	58.2	2e-14	SHANK1
Go	positive regulation of membrane potential	BivEnh	30.7	3e-13	GRIN1, SHANK1
Go	righting reflex	BivEnh	44.1	6e-13	SHANK1
Go	nonassociative learning	BivEnh	33.0	2e-11	SHANK1
Go	cytoskeletal anchoring at plasma membrane	BivEnh	31.6	3e-11	SHANK1
Go	social behavior	BivEnh	15.0	7e-11	DRD4, GRIN1, SHANK1, VPS13A
Go	protein localization to synapse	BivEnh	26.9	2e-10	SHANK1
Go	negative regulation of actin filament bundle assembly	BivEnh	26.4	2e-10	SHANK1
Go	positive regulation of dendritic spine development	BivEnh	24.2	5e-10	SHANK1
Go	memory	BivEnh	8.0	6e-10	DRD4, GRIN1, HRH1, KLK8, NTRK2, SGK1, SHANK1, TACR1
Go	dendritic spine morphogenesis	BivEnh	18.4	1e-09	EPHB1, SHANK1
Go	dendritic spine organization	BivEnh	17.4	2e-09	EPHB1, SHANK1
Go	dendritic spine development	BivEnh	17.4	2e-09	EPHB1, SHANK1
Go	regulation of alpha-amino-3-hydroxy-5-methyl-4-isoxazole propionate selective glutamate receptor activity	BivEnh	19.1	6e-09	SHANK1
Go	reflex	BivEnh	18.6	8e-09	SHANK1
Go	multi-organism behavior	BivEnh	7.7	1e-08	DACH1, DRD4, GRIN1, SHANK1, TACR1, VPS13A
Go	regulation of glutamate receptor signaling pathway	BivEnh	16.3	3e-08	SHANK1
Go	associative learning	BivEnh	7.3	9e-08	DRD4, GRIN1, HRH1, SHANK1, TACR1
Go	regulation of postsynaptic membrane potential	BivEnh	8.6	5e-07	DRD4, GRIN1, SHANK1
Go	regulation of excitatory postsynaptic membrane potential	BivEnh	9.4	5e-07	GRIN1, SHANK1
Go	learning	BivEnh	5.1	3e-06	DRD4, GRIN1, HRH1, NTRK2, PTN, SHANK1, SLC6A1, TACR1
Go	regulation of actin filament bundle assembly	BivEnh	8.0	4e-06	SHANK1, TACR1
Go	regulation of dendritic spine development	BivEnh	10.3	4e-06	SHANK1
Go	dendrite morphogenesis	BivEnh	9.0	4e-06	EPHB1, SHANK1
Go	regulation of dendrite development	BivEnh	5.3	1e-05	DBN1, GRIN1, NTRK2, PTPRD, SHANK1
Go	maintenance of protein location in cell	BivEnh	7.1	5e-05	DBN1, SHANK1
Go	learning or memory	BivEnh	3.8	5e-05	DRD4, GRIN1, HRH1, KLK8, NTRK2, PTN, SGK1, SHANK1, SLC6A1, TACR1
Go	maintenance of location in cell	BivEnh	6.7	8e-05	DBN1, SHANK1
Go	maintenance of protein location	BivEnh	6.6	0.0001	DBN1, SHANK1
Go	regulation of transporter activity	BivEnh	5.3	0.0001	AKAP7, DRD4, SGK1, SHANK1
Go	cognition	BivEnh	3.5	0.0002	DRD4, GRIN1, HRH1, KLK8, NTRK2, PTN, SGK1, SHANK1, SLC6A1, TACR1
Go	regulation of ion transmembrane transporter activity	BivEnh	5.5	0.0002	AKAP7, DRD4, SHANK1
Go	maintenance of location	BivEnh	5.8	0.0003	DBN1, SHANK1

Go	synapse organization	BivEnh	4.0	0.0004	KLK8, KY, NTRK2, PTPRD, SHANK1, WNT7A
Go	membrane depolarization	BivEnh	5.2	0.0004	GRIN1, SHANK1
Go	regulation of transmembrane transporter activity	BivEnh	5.2	0.0004	AKAP7, DRD4, SHANK1
Go	regulation of receptor activity	BivEnh	6.2	0.0004	SHANK1
Go	neuromuscular process controlling balance	BivEnh	5.2	0.0009	POU4F2, SHANK1
Go	dendrite development	BivEnh	5.1	0.001	EPHB1, SHANK1
Go	negative regulation of cytoskeleton organization	BivEnh	5.2	0.002	SHANK1
Go	regulation of digestive system process	BivEnh	9.0	0.003	NEUROD1, SGK1, TAC4, TACR1
Go	amacrine cell differentiation	BivEnh	56.7	0.003	NEUROD1
Go	neuromuscular process	BivEnh	3.6	0.004	AGTPBP1, GRIN1, POU4F2, SHANK1
Go	regulation of intestinal epithelial structure maintenance	BivEnh	39.7	0.01	NEUROD1
Go	response to morphine	BivEnh	9.1	0.01	EDNRA, GHR, GRIN1, TACR1
Go	response to isoquinoline alkaloid	BivEnh	8.8	0.01	EDNRA, GHR, GRIN1, TACR1
Go	olfactory learning	BivEnh	36.1	0.01	DRD4, GRIN1
Go	negative regulation of type B pancreatic cell apoptotic process	BivEnh	16.5	0.02	NEUROD1, TCF7L2
Go	regulation of type B pancreatic cell apoptotic process	BivEnh	15.6	0.02	NEUROD1, TCF7L2
Go	intestinal epithelial cell development	BivEnh	15.6	0.02	NKX3-2
Go	response to ammonium ion	BivEnh	5.3	0.03	DRD4, EDNRA, GHR, GRIN1, HRH1, SLC6A1, TACR1
Go	regulation of neuron projection development	BivEnh	2.4	0.03	ANKRD1, DBN1, GRIN1, KLK8, NTRK2, PLXNA4, PTPRD, SHANK1, STK25, WNT7A
Go	response to heat	BivEnh	5.7	0.04	CALCA, MYOF, TACR1, TFEC
Go	regulation of retinal cell programmed cell death	BivEnh	9.4	1e-06	BDNF, BHLHE23
Go	post-embryonic eye morphogenesis	BivEnh	10.4	3e-06	BHLHE23
Go	negative regulation of retinal cell programmed cell death	BivEnh	10.4	3e-06	BHLHE23
Go	post-embryonic organ morphogenesis	BivEnh	8.5	1e-05	BHLHE23
Go	post-embryonic morphogenesis	BivEnh	5.7	4e-05	BHLHE23, GNAS
Go	non-canonical Wnt receptor signaling pathway via MAPK cascade	BivEnh	6.0	5e-05	FZD10, WNT4
Go	positive regulation of embryonic development	BivEnh	3.5	8e-05	FOXA2, LHX1, OTX2, RBM19, WNT3A, WNT4
Go	negative regulation of smooth muscle cell-matrix adhesion	BivEnh	25.3	0.0001	APOD
Go	negative regulation of cytokine production involved in inflammatory response	BivEnh	25.3	0.0001	APOD
Go	negative regulation of T cell migration	BivEnh	25.3	0.0001	APOD
Go	negative regulation of lipoprotein lipid oxidation	BivEnh	25.3	0.0001	APOD
Go	negative regulation of lipoprotein oxidation	BivEnh	25.3	0.0001	APOD
Go	apoptotic nuclear changes	BivEnh	3.3	0.0001	BIRC7, BLCAP, DEDD2, ETS1, SFRP1, TBX5
Go	branching involved in mammary gland duct morphogenesis	BivEnh	3.5	0.0002	BTRC, EPHA2, ETV4, MSX2, SRC, TBX3, TFAP2C, VDR, WNT4
Go	retinal rod cell development	BivEnh	6.2	0.0002	BHLHE23
Go	regulation of transcription from RNA polymerase III promoter	BivEnh	5.6	0.0002	CEBPA, FOXA2, RPTOR, ZNF76

Go	mammary gland duct morphogenesis	BivEnh	2.8	0.0002	BTRC, EPHA2, ETV4, FGFR2, GLI2, MSX2, NTN1, SRC, TBX3, TFAP2C, VDR, WNT4
Go	microtubule organizing center organization	BivEnh	4.8	0.0002	CLASP2, CROCC, HAUS7, NPM1
Go	negative regulation of monocyte chemotactic protein-1 production	BivEnh	21.7	0.0003	APOD
Go	negative regulation of Rho GTPase activity	BivEnh	6.4	0.0003	FZD10
Go	negative regulation of epithelial to mesenchymal transition	BivEnh	3.6	0.0003	FOXA2, SFRP1, TBX5
Go	non-canonical Wnt receptor signaling pathway via JNK cascade	BivEnh	6.3	0.0003	FZD10
Go	positive regulation of Rac GTPase activity	BivEnh	3.4	0.0004	ARHGAP27, ARHGEF16, EPHA2, FZD10, SFRP1
Go	regulation of embryonic development	BivEnh	2.1	0.0005	BMP7, FOXA2, FZD1, GATA3, GRHL3, HES1, LAMA3, LHX1, MESP1, NKD1, OTX2, RBM19, SFRP1, SHH, TENM4, WNT3A, WNT4
Go	non-canonical Wnt receptor signaling pathway	BivEnh	3.3	0.0005	FZD1, FZD10, GRHL3, SFRP1, WNT11, WNT4
Go	negative regulation of lymphocyte migration	BivEnh	13.0	0.0008	APOD, CCL2
Go	positive regulation of transcription from RNA polymerase III promoter	BivEnh	5.7	0.0008	CEBPA, FOXA2, RPTOR
Go	post-embryonic organ development	BivEnh	4.8	0.0008	BHLHE23, MYO1E
Go	dorsal/ventral neural tube patterning	BivEnh	2.8	0.001	FOXA2, GAS1, GLI2, GSC, PAX7, SHH, WNT3A
Go	ectoderm formation	BivEnh	5.0	0.001	FOXA2, LHX1
Go	negative regulation of neuron differentiation	BivEnh	2.1	0.001	BMP7, CNTN2, DLL1, FOXA2, HES1, ID2, IRX3, LBX1, LMX1A, MEIS1, NKX2-2, OLIG2, PAX6, SOX2, SOX3, SOX9, TLX3, TP73, ZNF536
Go	positive regulation of Ras GTPase activity	BivEnh	2.1	0.001	ARHGAP27, ARHGEF16, EPHA2, FZD10, GNAS, GRHL3, IQGAP2, NTF3, NTRK1, RALGAPA2, SFRP1, TBC1D22A, TBC1D7, THY1, WNT11, WNT4
Go	centrosome organization	BivEnh	4.5	0.001	CROCC, HAUS7, NPM1
Go	eye photoreceptor cell development	BivEnh	3.6	0.001	BHLHE23, CNGA3, PAX6, THY1
Go	cellular component disassembly involved in execution phase of apoptosis	BivEnh	2.5	0.002	BIRC7, BLCAP, DEDD2, ETS1, PKP1, PRKCQ, SFRP1, TBX5
Go	regulation of smooth muscle cell-matrix adhesion	BivEnh	15.2	0.002	APOD
Go	regulation of transcription from RNA polymerase II promoter involved in detection of glucose	BivEnh	6.2	0.002	FOXA2
Go	negative regulation of transcription by glucose	BivEnh	6.2	0.002	FOXA2
Go	negative regulation of detection of glucose	BivEnh	6.2	0.002	FOXA2
Go	positive regulation of transcription from RNA polymerase II promoter by glucose	BivEnh	6.2	0.002	FOXA2
Go	negative regulation of transcription from RNA polymerase II promoter by glucose	BivEnh	6.2	0.002	FOXA2
Go	neural tube patterning	BivEnh	2.2	0.002	EN1, FGF8, FOXA2, GAS1, GBX2, GLI2, GSC, HES1, PAX6, PAX7, SHH, SSBP3, WNT3A
Go	negative regulation of cell morphogenesis involved in differentiation	BivEnh	3.0	0.002	FOXA2, SFRP1, TBX5
Go	eye photoreceptor cell differentiation	BivEnh	2.9	0.002	BHLHE23, CNGA3, NKD1, OTX2, PAX6, SOX9, THY1
Go	regulation of transcription by glucose	BivEnh	6.1	0.002	FOXA2

Go	regulation of cell-cell adhesion mediated by integrin	BivEnh	3.7	0.002	ADA, FOXA2, PODXL, WNT3A
Go	ventral spinal cord interneuron fate determination	BivEnh	13.8	0.002	NKX2-2
Go	type B pancreatic cell fate commitment	BivEnh	13.8	0.002	NKX2-2
Go	photoreceptor cell development	BivEnh	3.3	0.002	BHLHE23, CNGA3, PAX6, THY1
Go	somite rostral/caudal axis specification	BivEnh	3.9	0.002	FOXA2, LHX1, OTX2
Go	carbon catabolite activation of transcription from RNA polymerase II promoter	BivEnh	5.9	0.002	FOXA2
Go	cellular response to nutrient	BivEnh	3.6	0.002	FOXA2, SFRP1, VDR
Go	regulation of focal adhesion assembly	BivEnh	3.4	0.002	ACVRL1, APOD, GREM1, SFRP1, SRC, WNT4
Go	stem cell maintenance	BivEnh	2.1	0.002	BCL9, CDX2, DICER1, DIS3L2, DLL1, HES1, HMGA2, KIT, MED24, MED27, NOTCH2, PRDM16, SFRP1, SOX2, SOX9, TBX3, TCF7L2, TCL1A, TFAP2C
Go	embryonic pattern specification	BivEnh	2.1	0.002	BMP7, COBL, DLL1, FGFR2, FOXA2, HOXD8, LHX1, MEOX1, OTX2, PCSK6, SHH, SMAD2, T, TBX3
Go	activation of Ras GTPase activity	BivEnh	3.8	0.002	ARHGEF16, EPHA2, NTF3, RALGAPA2, TBC1D7
Go	negative regulation of glucokinase activity	BivEnh	5.8	0.002	FOXA2
Go	carbon catabolite regulation of transcription	BivEnh	5.8	0.002	FOXA2
Go	mammary gland epithelium development	BivEnh	2.2	0.002	BTRC, EPHA2, ETV4, FGFR2, GLI2, ID2, MGMT, MSX2, NTN1, SRC, TBX3, TFAP2C, VDR, WNT4, WNT7B
Go	positive regulation of glial cell differentiation	BivEnh	2.8	0.002	DICER1, HES1, ID2, NKX2-2, NKX6-1, NKX6-2, NTF3, OLIG2, SHH, TENM4, TP73
Go	positive regulation of Rho GTPase activity	BivEnh	2.4	0.002	ARHGAP27, ARHGEF16, EPHA2, FZD10, GRHL3, SFRP1, TBC1D7, THY1, WNT4
Go	positive regulation of gastrulation	BivEnh	3.6	0.002	FOXA2, LHX1, OTX2, TENM4
Go	detection of glucose	BivEnh	4.3	0.002	FOXA2, NKX6-1, PDX1
Go	regulation of glucokinase activity	BivEnh	5.7	0.002	FOXA2
Go	positive regulation of cell-cell adhesion mediated by integrin	BivEnh	3.7	0.002	FOXA2, PODXL, WNT3A
Go	embryonic axis specification	BivEnh	2.4	0.003	COBL, FOXA2, HOXD8, LHX1, OTX2, PCSK6, SMAD2, T, TBX3
Go	negative regulation of lipoprotein metabolic process	BivEnh	12.6	0.003	APOD
Go	ncRNA catabolic process	BivEnh	12.6	0.003	DEDD2, DIS3L2
Go	photoreceptor cell differentiation	BivEnh	2.8	0.003	BHLHE23, CNGA3, NKD1, OTX2, PAX6, SOX9, THY1
Go	negative regulation of Ras GTPase activity	BivEnh	4.5	0.003	FZD10
Go	positive regulation of cell-cell adhesion mediated by cadherin	BivEnh	4.5	0.003	FOXA2, WNT3A
Go	parathyroid gland development	BivEnh	4.1	0.003	GATA3, PAX1, TBX1
Go	ectoderm development	BivEnh	3.4	0.003	FOXA2, GRHL3, LHX1, SHH
Go	mammary gland morphogenesis	BivEnh	2.2	0.003	BTRC, EPHA2, ETV4, FGFR2, GLI2, MSX2, NTN1, PAX3, SRC, TBX3, TFAP2C, VDR, WNT4
Go	peripheral nervous system axon regeneration	BivEnh	11.7	0.003	APOD
Go	execution phase of apoptosis	BivEnh	2.3	0.004	BIRC7, BLCAP, DEDD2, ETS1, PKP1, PRKCQ, SFRP1, TBX5
Go	regulation of Rho GTPase activity	BivEnh	2.2	0.004	ARHGAP27, ARHGEF16, EPHA2, EPHA4, EPHB3, FGD3, FZD10, GRHL3, SFRP1, TBC1D7, THY1, WNT4
Go	negative regulation of leukocyte migration	BivEnh	5.9	0.004	ADA, APOD, CCL2, GREM1

Go	regulation of action potential in neuron	BivEnh	2.2	0.004	ATRN, CHRNA1, CNTN2, DICER1, DPP6, EPB41L3, GNPAT, JAM3, MAL, MBP, MYRF, NAB1, NFASC, NKX6-2, NTF3, OLIG2, PMP22, POU3F1, POU3F2, SCN8A, TENM4
Go	regulation of cell-cell adhesion mediated by cadherin	BivEnh	4.2	0.004	FOXA2, WNT3A
Go	DNA catabolic process	BivEnh	2.6	0.004	BIRC7, SFRP1, SMUG1, UBE2V2, XPA, XRN2
Go	sclerotome development	BivEnh	5.8	0.004	PAX1
Go	sinoatrial node cell differentiation	BivEnh	5.8	0.004	MESP1, TBX3
Go	exit from mitosis	BivEnh	6.6	0.005	CTDP1, PPP2R2D
Go	positive regulation of cell-cell adhesion	BivEnh	2.5	0.006	BMP7, FOXA2, L1CAM, PODXL, SIRPG, WNT3A
Go	response to nutrient	BivEnh	2.0	0.006	ABCG5, ACSL1, ALDH1A2, BDNF, CCL2, CEBPA, COX4I1, CYP11A1, FOXA2, GNPAT, IGF1R, MGMT, PDX1, PITX2, SFRP1, SFTPC, SPARC, VDR
Go	regulation of cytokine production involved in inflammatory response	BivEnh	10.1	0.006	APOD
Go	Ral protein signal transduction	BivEnh	10.1	0.006	RALGAPA2
Go	primitive streak formation	BivEnh	3.1	0.006	FOXA2, LHX1, OTX2, T
Go	cell migration involved in coronary vasculogenesis	BivEnh	6.2	0.007	TBX5
Go	positive regulation of gliogenesis	BivEnh	2.5	0.007	DICER1, ETV5, HES1, ID2, NKX2-2, NKX6-1, NKX6-2, NTF3, OLIG2, SHH, TENM4, TP73
Go	pituitary gland development	BivEnh	2.0	0.008	ALDH1A2, ETS1, FGF8, GLI2, GSX1, HES1, HMGA2, PAX6, PITX2, POU3F2, SALL1, SOX2, SOX3, TCF7L2, WNT4
Go	DNA catabolic process, exonucleolytic	BivEnh	6.1	0.008	UBE2V2, XRN2
Go	activation of Ral GTPase activity	BivEnh	9.5	0.008	RALGAPA2
Go	rRNA catabolic process	BivEnh	13.5	0.008	DEDD2
Go	immunoglobulin secretion involved in immune response	BivEnh	13.5	0.008	POU2F2
Go	anterior/posterior axis specification	BivEnh	2.3	0.008	FOXA2, HOXD8, LHX1, OTX2, PCSK6, T, TBX3, WNT8A
Go	cellular response to prostaglandin E stimulus	BivEnh	5.9	0.009	GNAS, SFRP1
Go	positive regulation of oligodendrocyte differentiation	BivEnh	4.1	0.009	NKX2-2, OLIG2, SHH, TENM4, TP73
Go	regulation of branching involved in prostate gland morphogenesis	BivEnh	3.8	0.009	BMP7, FGFR2, SFRP1
Go	hepatocyte growth factor receptor signaling pathway	BivEnh	8.9	0.009	MUC20
Go	vasculogenesis involved in coronary vascular morphogenesis	BivEnh	5.7	0.01	TBX5
Go	sinoatrial node development	BivEnh	5.0	0.01	MESP1, TBX3
Go	cell migration involved in vasculogenesis	BivEnh	5.6	0.01	TBX5
Go	regulation of Ral protein signal transduction	BivEnh	8.4	0.01	RALGAPA2
Go	pancreatic A cell fate commitment	BivEnh	8.4	0.01	NKX2-2
Go	pancreatic PP cell fate commitment	BivEnh	8.4	0.01	NKX2-2
Go	positive regulation of histone methylation	BivEnh	4.3	0.01	PAX7, PAXIP1
Go	axis elongation	BivEnh	2.9	0.01	FGFR2, SFRP1, SHH, SOX9, TFAP2C, WNT3A
Go	sinoatrial node cell development	BivEnh	5.4	0.01	TBX3
Go	cellular response to prostaglandin stimulus	BivEnh	5.4	0.01	GNAS, SFRP1

Go	luteinizing hormone secretion	BivEnh	5.4	0.01	TBX3
Go	follicle-stimulating hormone secretion	BivEnh	5.4	0.01	TBX3
Go	atrioventricular bundle cell differentiation	BivEnh	5.4	0.01	TBX3
Go	immunoglobulin production involved in immunoglobulin mediated immune response	BivEnh	6.5	0.01	GCNT3, MSH2, POU2F2
Go	cellular response to ketone	BivEnh	4.2	0.01	GNAS, SFRP1, SRC
Go	branching involved in salivary gland morphogenesis	BivEnh	3.1	0.01	BMP7, FGF8, FGFR2, PLXNA1, SEMA3A, SHH, TGM2
Go	regulation of epithelial to mesenchymal transition	BivEnh	2.2	0.01	BAMBI, FOXA2, GREM1, SFRP1, SMAD2, TBX5, TCF7L2
Go	positive regulation of cell adhesion mediated by integrin	BivEnh	3.0	0.01	FOXA2, PODXL, WNT3A
Go	myelination	BivEnh	2.2	0.01	ATRN, CNTN2, DICER1, EPB41L3, GNPAT, JAM3, MAL, MBP, MYRF, NAB1, NFASC, NKX6-2, NTF3, OLIG2, PMP22, POU3F1, POU3F2, SCN8A, TENM4
Go	regulation of cell junction assembly	BivEnh	2.7	0.01	ACVRL1, APOD, GREM1, SFRP1, SRC, WNT4
Go	cellular response to BMP stimulus	BivEnh	2.7	0.01	ACVRL1, COL2A1, GATA3, GATA5, SFRP1
Go	regulation of monocyte chemotactic protein-1 production	BivEnh	8.0	0.01	APOD
Go	negative regulation of focal adhesion assembly	BivEnh	4.6	0.01	ACVRL1, APOD, SRC
Go	positive regulation of histone H3-K36 methylation	BivEnh	18.2	0.01	PAXIP1
Go	chaperone-mediated protein complex assembly	BivEnh	5.2	0.02	CCT6B, HSP90AA1
Go	bundle of His development	BivEnh	3.7	0.02	ID2, TBX3, TBX5
Go	axon ensheathment	BivEnh	2.2	0.02	ATRN, CNTN2, DICER1, EPB41L3, GNPAT, JAM3, MAL, MBP, MYRF, NAB1, NFASC, NKX6-2, NTF3, OLIG2, PMP22, POU3F1, POU3F2, SCN8A, TENM4
Go	regulation of JUN kinase activity	BivEnh	2.2	0.02	BIRC7, EPHA4, FZD10, MDFI, SFRP1, TAOK3, TP73
Go	negative regulation of chemokine production	BivEnh	7.6	0.02	APOD
Go	spinal cord association neuron differentiation	BivEnh	2.2	0.02	GSX1, GSX2, LBX1, LHX1, LHX5, MDGA1, PAX3, PAX7, WNT3A
Go	regulation of natural killer cell apoptotic process	BivEnh	7.2	0.02	BIRC7
Go	negative regulation of platelet-derived growth factor receptor signaling pathway	BivEnh	7.2	0.02	APOD
Go	positive regulation of extrinsic apoptotic signaling pathway	BivEnh	3.3	0.02	CYLD, DEDD2, SFRP1
Go	regulation of canonical Wnt receptor signaling pathway involved in controlling type B pancreatic cell proliferation	BivEnh	5.7	0.02	SFRP1, WNT3A
Go	regulation of cellular respiration	BivEnh	5.7	0.02	NOS2, PRDM16
Go	response to interleukin-6	BivEnh	3.5	0.02	FOXA2, SFTPC
Go	cardiac pacemaker cell differentiation	BivEnh	4.2	0.02	MESP1, TBX3
Go	cell differentiation in hindbrain	BivEnh	2.3	0.02	ATP2B2, FOXA2, KNDC1, LHX1, LHX5, NFIX
Go	axis elongation involved in somitogenesis	BivEnh	5.5	0.02	SFRP1, WNT3A
Go	cellular response to vitamin D	BivEnh	5.5	0.02	SFRP1, VDR
Go	regulation of gastrulation	BivEnh	2.4	0.03	FOXA2, LHX1, MESP1, OTX2, TENM4, WNT3A
Go	endocrine hormone secretion	BivEnh	3.1	0.03	CRHR1, GATA3, TBX3
Go	salivary gland development	BivEnh	2.5	0.03	BMP7, FGF8, FGFR2, PAX6, PLXNA1, SEMA3A, SHH, TFCP2L1, TGM2

Go	negative regulation of phosphatase activity	BivEnh	2.4	0.03	CAMSAP3, ELFN2, TMEM132D, URI1
Go	metanephric mesenchymal cell proliferation involved in metanephros development	BivEnh	4.5	0.03	BMP7, SHH
Go	response to prostaglandin E stimulus	BivEnh	4.5	0.03	GNAS, SFRP1
Go	sister chromatid segregation	BivEnh	3.6	0.03	DDX11, DIS3L2, KATNB1, KIF18A, NCAPD3, NCAPH
Go	mitotic sister chromatid segregation	BivEnh	3.6	0.03	DDX11, DIS3L2, KATNB1, KIF18A, NCAPD3, NCAPH
Go	immunoglobulin secretion	BivEnh	8.7	0.03	POU2F2
Go	cerebellar Purkinje cell-granule cell precursor cell signaling involved in regulation of granule cell precursor cell proliferation	BivEnh	2.8	0.03	GLI2, LHX1, LHX5, SHH
Go	cardiac conduction system development	BivEnh	3.1	0.03	ID2, MESP1, TBX3, TBX5
Go	regulation of exocyst localization	BivEnh	6.3	0.03	RALGAPA2
Go	positive regulation of cerebellar granule cell precursor proliferation	BivEnh	2.7	0.03	GLI2, LHX1, LHX5, SHH
Go	regulation of insulin secretion involved in cellular response to glucose stimulus	BivEnh	2.9	0.03	ADRA2A, FOXA2, PIM3, TCF7L2
Go	negative regulation of GTPase activity	BivEnh	2.7	0.03	FZD10, PTPRN2
Go	regulation of T cell migration	BivEnh	5.1	0.03	ADAM10, APOD
Go	regulation of cell adhesion mediated by integrin	BivEnh	2.3	0.03	ADA, EPHA2, FOXA2, PODXL, PTPN11, WNT3A
Go	regulation of lymphocyte migration	BivEnh	4.3	0.03	ADAM10, APOD, CCL2
Go	insulin catabolic process	BivEnh	30.3	0.04	IDE
Go	post-embryonic body morphogenesis	BivEnh	13.0	0.04	GNAS
Go	cardiac muscle cell fate commitment	BivEnh	3.2	0.04	TBX3, WNT3A, WNT8A
Go	cardiac cell fate commitment	BivEnh	2.8	0.04	MESP1, TBX3, TENM4, WNT3A, WNT8A
Go	base-excision repair	BivEnh	3.0	0.04	HMGA2, NEIL2, POLG, SMUG1
Go	activation of Cdc42 GTPase activity	BivEnh	8.1	0.04	ARHGEF16
Go	response to prostaglandin stimulus	BivEnh	4.2	0.04	GNAS, SFRP1
Go	nucleic acid phosphodiester bond hydrolysis	BivEnh	2.1	0.04	BIRC7, CPSF2, DICER1, DIS3L2, MGME1, SFRP1, TREX2, UBE2V2, XRN2
Go	hypothalamus development	BivEnh	2.5	0.04	CRHR1, ETS1, GSX1, PITX2, POU3F2, SOX3
Go	DNA-dependent transcription, termination	BivEnh	2.7	0.04	CPSF2, LZTS1, TAF1B, TTF1, XRN2
Go	CD8-positive, alpha-beta T cell differentiation	BivEnh	3.7	0.04	PAX1
Go	negative regulation of cardiac muscle cell proliferation	BivEnh	4.2	0.04	TBX5
Go	somatic stem cell maintenance	BivEnh	2.0	0.04	BCL9, CDX2, HES1, HMGA2, PRDM16, SFRP1, SOX2, SOX9, TCF7L2, TFAP2C
Go	sodium-independent organic anion transport	BivEnh	4.8	0.04	SLCO4A1
Go	neural crest cell fate commitment	BivEnh	5.8	0.04	SFRP1, WNT8A
Go	regulation of Cdc42 GTPase activity	BivEnh	3.3	0.05	ARHGAP27, ARHGEF16, EPHB3, FGD3
Go	negative regulation of NF-kappaB import into nucleus	BivEnh	3.6	0.05	BMP7, CYLD
Go	tissue regeneration	BivEnh	2.6	0.05	APOD, BCL9, PAX7
Go	negative regulation of protein import into nucleus	BivEnh	2.3	0.05	APOD, BMP7, CYLD, FZD1, MDFI, PKIG
Go	neural crest formation	BivEnh	5.6	0.05	SFRP1, WNT8A
Go	retinal bipolar neuron differentiation	BivEnh	4.0	0.05	BHLHE22, IRX5, VSX1, VSX2
Go	cardiac left ventricle formation	BivEnh	4.0	0.05	TBX5
Go	regulation of apoptotic process involved in morphogenesis	BivEnh	2.9	0.05	BMP7, PAX2, VDR
Go	positive regulation of JUN kinase activity	BivEnh	2.1	0.05	BIRC7, EPHA4, FZD10, MDFI, TAOK3
MSigDB	Genes involved in Regulation of gene expression in beta cells	BivEnh	3.8	0.001	FOXA2, NKX2-2, NKX6-1, PAX6, PDX1, SLC2A2

MSigDB	Maturity onset diabetes of the young	BivEnh	2.6	0.03	FOXA2, HES1, NKX2-2, NKX6-1, PAX6, PDX1, SLC2A2
MSigDB	Genes involved in RNA Polymerase I Promoter Opening	BivEnh	6.6	0.03	H3F3A, HIST1H2BJ, HIST1H2BL, HIST3H2BB
MSigDB	Genes involved in RNA Polymerase I Transcription	BivEnh	4.9	0.03	H3F3A, HIST1H2BJ, HIST1H2BL, HIST3H2BB, TAF1B
MSigDB	Genes involved in Association of TriC/CCT with target proteins during biosynthesis	BivEnh	5.5	0.02	FBXW4, KIFC3, XRN2
MSigDB	Genes involved in Protein folding	BivEnh	3.7	0.02	FBXW4, KIFC3, TBCA, TUBB2B, TUBB6, XRN2
MSigDB	Genes involved in Regulation of beta-cell development	BivEnh	2.3	0.05	FOXA2, HES1, NKX2-2, NKX6-1, ONECUT3, PAX6, PDX1, SLC2A2
Go	determination of affect	BivPro	71.5	2e-14	SHANK1
Go	habituation	BivPro	71.5	2e-14	SHANK1
Go	chemosensory behavior	BivPro	39.8	1e-13	DRD4, GRIN1, SHANK1
Go	sensory processing	BivPro	62.0	1e-13	SHANK1
Go	synapse maturation	BivPro	62.0	1e-13	SHANK1
Go	righting reflex	BivPro	44.3	8e-12	SHANK1
Go	vocalization behavior	BivPro	38.7	4e-11	SHANK1
Go	positive regulation of excitatory postsynaptic membrane potential	BivPro	25.6	3e-10	GRIN1, SHANK1
Go	positive regulation of membrane potential	BivPro	22.7	1e-09	GRIN1, SHANK1
Go	nonassociative learning	BivPro	27.3	2e-09	SHANK1
Go	negative regulation of actin filament bundle assembly	BivPro	26.6	2e-09	SHANK1
Go	cytoskeletal anchoring at plasma membrane	BivPro	25.1	4e-09	SHANK1
Go	long-term memory	BivPro	16.4	5e-09	GRIN1, SGK1, SHANK1
Go	reflex	BivPro	23.8	5e-09	SHANK1
Go	dendritic spine morphogenesis	BivPro	22.1	1e-08	SHANK1
Go	positive regulation of dendritic spine development	BivPro	21.1	2e-08	SHANK1
Go	dendritic spine development	BivPro	21.1	2e-08	SHANK1
Go	dendritic spine organization	BivPro	20.7	2e-08	SHANK1
Go	protein localization to synapse	BivPro	20.2	2e-08	SHANK1
Go	social behavior	BivPro	13.6	4e-08	DRD4, GRIN1, SHANK1
Go	regulation of alpha-amino-3-hydroxy-5-methyl-4-isoxazole propionate selective glutamate receptor activity	BivPro	16.6	2e-07	SHANK1
Go	memory	BivPro	7.6	6e-07	DRD4, GRIN1, HRH1, KLK8, SGK1, SHANK1
Go	regulation of glutamate receptor signaling pathway	BivPro	13.9	1e-06	SHANK1
Go	regulation of postsynaptic membrane potential	BivPro	9.3	3e-06	DRD4, GRIN1, SHANK1
Go	intestinal epithelial cell development	BivPro	37.2	3e-06	NKX3-2
Go	dendrite morphogenesis	BivPro	10.1	4e-06	SHANK1, SLITRK5
Go	associative learning	BivPro	7.6	6e-06	DRD4, GRIN1, HRH1, SHANK1
Go	regulation of excitatory postsynaptic membrane potential	BivPro	9.5	8e-06	GRIN1, SHANK1
Go	regulation of dendritic spine development	BivPro	10.4	1e-05	SHANK1
Go	regulation of actin filament bundle assembly	BivPro	8.9	6e-05	SHANK1
Go	dendrite development	BivPro	6.7	8e-05	FEZF2, SHANK1, SLITRK5
Go	multi-organism behavior	BivPro	6.4	0.0001	DRD4, GRIN1, SHANK1
Go	neuromuscular process controlling balance	BivPro	6.6	0.0003	POU4F2, SHANK1
Go	regulation of transporter activity	BivPro	5.2	0.0005	AKAP7, DRD4, SGK1, SHANK1
Go	regulation of ion transmembrane transporter activity	BivPro	5.4	0.0007	AKAP7, DRD4, SHANK1

Go	long-chain fatty acid catabolic process	BivPro	37.2	0.0007	CYP4F3, LIPE
Go	learning	BivPro	4.9	0.0008	DRD4, GRIN1, HRH1, SHANK1, SLC6A1
Go	regulation of transmembrane transporter activity	BivPro	5.2	0.001	AKAP7, DRD4, SHANK1
Go	maintenance of protein location in cell	BivPro	6.4	0.001	SHANK1
Go	regulation of receptor activity	BivPro	6.2	0.001	SHANK1
Go	membrane depolarization	BivPro	5.5	0.001	GRIN1, SHANK1
Go	maintenance of location in cell	BivPro	5.9	0.002	SHANK1
Go	negative regulation of cytoskeleton organization	BivPro	5.9	0.002	SHANK1
Go	intestinal epithelial cell differentiation	BivPro	12.4	0.002	NKX3-2
Go	maintenance of protein location	BivPro	5.8	0.002	SHANK1
Go	leukotriene B4 catabolic process	BivPro	46.5	0.005	CYP4F3
Go	neuromuscular process	BivPro	4.2	0.007	GRIN1, POU4F2, SHANK1
Go	maintenance of location	BivPro	5.0	0.007	SHANK1
Go	learning or memory	BivPro	3.4	0.008	DRD4, GRIN1, HRH1, KLK8, SGK1, SHANK1, SLC6A1
Go	regulation of dendrite development	BivPro	4.4	0.009	GRIN1, SHANK1
Go	synapse organization	BivPro	3.9	0.01	COL4A1, KLK8, SHANK1
Go	columnar/cuboidal epithelial cell development	BivPro	8.6	0.01	NKX3-2
Go	negative regulation of chondrocyte differentiation	BivPro	8.5	0.01	NKX3-2
Go	complement activation, alternative pathway	BivPro	31.0	0.02	C3, CFD
Go	cognition	BivPro	3.1	0.02	DRD4, GRIN1, HRH1, KLK8, SGK1, SHANK1, SLC6A1
Go	actin filament-based movement	BivPro	7.7	0.02	MYH14, TNNI3
MSigDB	Genes involved in Complement cascade	BivPro	24.8	0.02	C3, CFD, PROS1
MSigDB	Genes involved in Sema4D induced cell migration and growth-cone collapse	BivPro	14.1	0.02	MYH14
MSigDB	Genes involved in Sema4D in semaphorin signaling	BivPro	12.6	0.02	MYH14
MSigDB	Genes involved in Initial triggering of complement	BivPro	34.9	0.02	C3, CFD
MSigDB	Complement Pathway	BivPro	31.0	0.03	C3, CFD
MSigDB	Complement and coagulation cascades	BivPro	9.3	0.04	C3, CFD, PLG, PROS1
Go	regulation of retinal cell programmed cell death	BivPro	11.7	0.0009	BDNF, BHLHE23
Go	negative regulation of neuron differentiation	BivPro	3.0	0.001	BMP7, DLL1, FOXA2, HES1, ID2, IRX3, LBX1, MEIS1, NKX2-2, OLIG2, PAX6, SOX3, SOX9, TLX3, TTC3, ZNF536
Go	microtubule organizing center organization	BivPro	5.8	0.0008	CLASP2, CROCC, HAUS7, NPM1
Go	centrosome organization	BivPro	5.7	0.001	CROCC, HAUS7, NPM1
Go	embryonic axis specification	BivPro	3.8	0.001	COBL, FOXA2, HOXD8, LHX1, OTX2, SMAD2, T, TBX3
Go	embryonic pattern specification	BivPro	2.9	0.002	BMP7, COBL, DLL1, FGFR2, FOXA2, HOXD8, LHX1, OTX2, SHH, SMAD2, T, TBX3
Go	positive regulation of embryonic development	BivPro	5.3	0.002	FOXA2, LHX1, OTX2, RBM19, WNT3A, WNT4
Go	ectoderm formation	BivPro	8.5	0.002	FOXA2, LHX1
Go	stem cell maintenance	BivPro	2.9	0.002	BCL9, CDX2, DIS3L2, DLL1, HES1, HMGA2, KIT, NOTCH2, PRDM16, SFRP1, SOX9, TBX3, TFAP2C
Go	non-canonical Wnt receptor signaling pathway via MAPK cascade	BivPro	12.3	0.002	FZD10, WNT4
Go	detection of glucose	BivPro	6.7	0.002	FOXA2, NKX6-1

Go	anterior/posterior axis specification	BivPro	3.6	0.002	FOXA2, HOXD8, LHX1, OTX2, T, TBX3, WNT8A
Go	cellular response to nutrient	BivPro	5.9	0.002	BGLAP, CYP24A1, FOXA2, SFRP1, VDR
Go	positive regulation of transcription from RNA polymerase III promoter	BivPro	8.9	0.002	CEBPA, FOXA2
Go	pancreas development	BivPro	2.5	0.003	ALDH1A2, FOXA2, HES1, IGF1R, NKX2-2, NKX3-2, NKX6-1, NKX6-2, PAX2, PAX6, SHH, SLC2A2, SMAD2, SOX9, ZIC3
Go	positive regulation of glial cell differentiation	BivPro	4.1	0.003	HES1, ID2, NKX2-2, NKX6-1, NKX6-2, NTF3, OLIG2, SHH, TENM4
Go	segmentation	BivPro	2.6	0.003	COBL, DLL1, FOXA2, HOXD8, IRX1, IRX2, IRX3, LHX1, MAFB, OTX2, PAX1, PLXNA2, SFRP1, T, TBX3, WNT3A
Go	somite rostral/caudal axis specification	BivPro	6.0	0.003	FOXA2, LHX1, OTX2
Go	negative regulation of detection of glucose	BivPro	10.2	0.003	FOXA2
Go	regulation of transcription from RNA polymerase II promoter involved in detection of glucose	BivPro	10.2	0.003	FOXA2
Go	negative regulation of transcription by glucose	BivPro	10.2	0.003	FOXA2
Go	negative regulation of transcription from RNA polymerase II promoter by glucose	BivPro	10.2	0.003	FOXA2
Go	positive regulation of transcription from RNA polymerase II promoter by glucose	BivPro	10.2	0.003	FOXA2
Go	neuron fate commitment	BivPro	2.3	0.003	FOXA2, GSX1, GSX2, HES1, HOXC10, ID2, LBX1, LHX6, NKX2-2, NKX6-1, NKX6-2, OLIG2, OTX2, PAX3, PAX6, PAX7, SHH, TFAP2C, TLX3
Go	negative regulation of cell development	BivPro	2.4	0.003	BDNF, BMP7, FOXA2, HES1, HMGA2, ID2, MBP, NKX6-1, NKX6-2, NTN1, PAX6, SFRP1, TBX5, TTC3, WNT3A
Go	ureteric bud development	BivPro	2.3	0.003	BDNF, BMP7, CRLF1, FGF8, FGFR2, FOXD2, FOXD3, FOXD4L6, GATA3, HES1, LHX1, PAX2, SALL1, SDC4, SFRP1, SHH, SIM1, SMAD2, SOX9, TSHZ3, WNT4
Go	sinoatrial node cell differentiation	BivPro	9.7	0.004	MESP1, TBX3
Go	regulation of transcription by glucose	BivPro	9.7	0.004	FOXA2
Go	negative regulation of glucokinase activity	BivPro	9.7	0.004	FOXA2
Go	carbon catabolite activation of transcription from RNA polymerase II promoter	BivPro	9.7	0.004	FOXA2
Go	non-canonical Wnt receptor signaling pathway via JNK cascade	BivPro	12.8	0.004	FZD10
Go	signal transduction involved in regulation of gene expression	BivPro	3.9	0.004	FGF8, FOXA2, GSC, MESP1, MSX2, PAX6, PDGFRA, T
Go	positive regulation of gliogenesis	BivPro	3.7	0.004	ETV5, HES1, ID2, NKX2-2, NKX6-1, NKX6-2, NTF3, OLIG2, SHH, TENM4
Go	branching involved in mammary gland duct morphogenesis	BivPro	4.2	0.004	BTRC, MSX2, TBX3, TFAP2C, VDR, WNT4
Go	branching morphogenesis of an epithelial tube	BivPro	2.1	0.004	BMP7, BTRC, CLIC4, FGF8, FGFR2, FOXA2, FOXD2, FOXD3, FOXD4L6, LHX1, MSX2, NRARP, PAX2, PITX2, SALL1, SHH, SOX9, TBX3, TFAP2C, VDR, WNT4
Go	positive regulation of gastrulation	BivPro	4.9	0.004	FOXA2, LHX1, OTX2, TENM4
Go	carbon catabolite regulation of transcription	BivPro	9.2	0.004	FOXA2
Go	primitive streak formation	BivPro	4.9	0.004	FOXA2, LHX1, OTX2, T
Go	regulation of action potential in neuron	BivPro	2.8	0.005	ATRN, CHRNA1, DPP6, JAM3, MAL, MBP, NAB1, NKX6-2, NTF3, OLIG2, P2RX2, POU3F1, POU3F2, SCN8A, TENM4

Go	regulation of transcription from RNA polymerase III promoter	BivPro	7.2	0.005	CEBPA, FOXA2
Go	axis specification	BivPro	2.5	0.005	COBL, DLL1, FOXA2, HOXD8, LHX1, MDFI, OTX2, PAX6, PITX2, SFRP1, SMAD2, T, TBX3, WNT8A
Go	regulation of glucokinase activity	BivPro	8.8	0.005	FOXA2
Go	somatic stem cell maintenance	BivPro	3.5	0.005	BCL9, CDX2, HES1, HMGA2, PRDM16, SFRP1, SOX9, TFAP2C
Go	positive regulation of neuron differentiation	BivPro	2.4	0.006	BDNF, BMP7, ETV5, FEZ1, FOXA2, HOXD3, IRX3, NEUROG1, NEUROG2, NKX2-2, NKX6-1, SALL1
Go	regulation of epithelial cell differentiation	BivPro	2.5	0.006	ADAMTS9, ALOX15B, DLL1, GATA3, HES1, LHX1, MESP1, MSX2, NKX6-1, PAX2, PAX6, SOX9, TBX3, VDR
Go	post-embryonic eye morphogenesis	BivPro	10.9	0.006	BHLHE23
Go	negative regulation of Rho GTPase activity	BivPro	10.9	0.006	FZD10
Go	positive regulation of cell-cell adhesion mediated by cadherin	BivPro	6.7	0.007	FOXA2, WNT3A
Go	regulation of organ morphogenesis	BivPro	2.1	0.007	BDNF, BHLHE23, BMP7, CD34, ETV5, FGF3, FGF8, FGFR2, GATA3, HES1, HOXC11, LHX1, MESP1, PAX2, PDGFA, SFRP1, SHH, SOX9, TBX1, WNT3A, WNT4
Go	mammary gland duct morphogenesis	BivPro	3.2	0.007	BTRC, FGFR2, MSX2, NTN1, TBX3, TFAP2C, VDR, WNT4
Go	sinoatrial node development	BivPro	8.0	0.007	MESP1, TBX3
Go	regulation of hormone levels	BivPro	2.1	0.008	ALDH1A2, ALDH1A3, CACNA1H, CPLX1, DIO2, EDN3, ENPEP, FAM3B, FOXA2, GATA3, IDE, PDGFRA, PRLHR, PTPN11, RAB11FIP2, REN, SHH, SMAD2, TBX3, WNT4
Go	dorsal/ventral pattern formation	BivPro	2.1	0.008	FGF8, FOXA2, GAS1, GSC, GSX2, LHX1, LHX2, LMX1B, MDFI, NKX2-2, NKX6-1, NKX6-2, OTX2, PAX6, PAX7, SFRP1, SHH, SMAD2, WNT3A
Go	cell fate specification	BivPro	2.2	0.008	EYA2, FOXA2, GSX2, HOXC10, LBX1, NKX2-2, NKX6-1, NKX6-2, OLIG2, OTX2, PAX2, PAX6, SHH, SOX9, TBX1, TENM4, TLX3
Go	positive regulation of cell-cell adhesion mediated by integrin	BivPro	5.4	0.008	FOXA2, PODXL, WNT3A
Go	post-embryonic morphogenesis	BivPro	6.1	0.01	BHLHE23, GNAS
Go	negative regulation of retinal cell programmed cell death	BivPro	9.6	0.01	BHLHE23
Go	lysobisphosphatidic acid metabolic process	BivPro	23.0	0.01	ACP6
Go	regulation of embryonic development	BivPro	2.4	0.01	BMP7, FOXA2, GATA3, HES1, LAMA3, LHX1, MESP1, OTX2, RBM19, SFRP1, SHH, TENM4, WNT3A, WNT4
Go	gastrulation	BivPro	2.2	0.01	BMP7, CFC1, DUSP4, EYA2, FGF8, FOXA2, GSC, HMGA2, LHX1, MESP1, OTX2, PAX2, SMAD2, T, TENM4, TNRC6C, WNT3A
Go	neural tube patterning	BivPro	2.9	0.01	FGF8, FOXA2, GAS1, GSC, HES1, PAX6, PAX7, SHH, SSBP3, WNT3A
Go	cardiac cell fate commitment	BivPro	4.4	0.01	MESP1, TBX3, TENM4, WNT3A, WNT8A
Go	mammary gland epithelium development	BivPro	2.6	0.02	BTRC, FGFR2, ID2, MGMT, MSX2, NTN1, TBX3, TFAP2C, VDR, WNT4, WNT7B
Go	microtubule depolymerization	BivPro	12.3	0.02	KIF18A, KIF19

Go	nucleosome assembly	BivPro	3.2	0.02	H2AFY2, H3F3A, H3F3C, HIST1H2AH, HIST1H2BJ, HIST1H2BL, HIST3H2BB, NPM1, SOX9, TSPYL2
Go	dorsal/ventral neural tube patterning	BivPro	3.4	0.02	FOXA2, GAS1, GSC, PAX7, SHH, WNT3A
Go	regulation of coagulation	BivPro	3.0	0.02	ANXA4, CD34, F11, FOXA2, GP5, KRT1, PDGFA, PDGFRA, PRKCQ
Go	cell differentiation in spinal cord	BivPro	2.2	0.02	GSX1, GSX2, HOXC10, LBX1, LHX1, LHX5, MDGA1, NKX2-2, NKX6-1, NKX6-2, OLIG2, PAX3, PAX6, PAX7, SHH, WNT3A
Go	protein depolymerization	BivPro	8.5	0.02	KIF18A, KIF19, MICAL3
Go	sinoatrial node cell development	BivPro	8.5	0.02	TBX3
Go	post-embryonic organ morphogenesis	BivPro	8.5	0.02	BHLHE23
Go	atrioventricular bundle cell differentiation	BivPro	8.5	0.02	TBX3
Go	luteinizing hormone secretion	BivPro	8.5	0.02	TBX3
Go	follicle-stimulating hormone secretion	BivPro	8.5	0.02	TBX3
Go	regulation of cell-cell adhesion mediated by cadherin	BivPro	5.5	0.02	FOXA2, WNT3A
Go	response to nutrient	BivPro	2.4	0.02	ABCG5, ACSL1, ALDH1A2, BDNF, BGLAP, CD3E, CEBPA, CYP24A1, FOXA2, IGF1R, MGMT, PITX2, SFRP1, VDR
Go	regulation of gastrulation	BivPro	3.3	0.02	FOXA2, LHX1, MESP1, OTX2, TENM4, WNT3A
Go	nucleosome organization	BivPro	3.0	0.02	H2AFY2, H3F3A, H3F3C, HIST1H2AH, HIST1H2BJ, HIST1H2BL, HIST3H2BB, NFE2, NPM1, SOX9, TSPYL2
Go	sclerotome development	BivPro	8.1	0.02	PAX1
Go	negative regulation of epithelial to mesenchymal transition	BivPro	3.8	0.02	FOXA2, SFRP1, TBX5
Go	DNA packaging	BivPro	2.7	0.02	H2AFY2, H3F3A, H3F3C, HIST1H2AH, HIST1H2BJ, HIST1H2BL, HIST3H2BB, HMGA2, NCAPD3, NCAPH, NPM1, SOX9, TSPYL2
Go	cellular response to vitamin D	BivPro	11.1	0.02	BGLAP, CYP24A1, SFRP1, VDR
Go	ectoderm development	BivPro	4.1	0.02	FOXA2, LHX1, SHH
Go	columnar/cuboidal epithelial cell differentiation	BivPro	2.3	0.02	C10orf11, EDN3, FGF8, FGFR2, FOXD2, FOXD3, FOXD4L6, GATA5, HPS1, KIT, NKX2-2, NKX3-2, NKX6-1, PAX6, WNT4
Go	positive regulation of cell-cell adhesion	BivPro	3.3	0.02	BMP7, FOXA2, L1CAM, PODXL, WNT3A
Go	regulation of stem cell differentiation	BivPro	2.2	0.02	BAMBI, BMP7, FOXA2, HES1, HMGA2, PAX2, PDGFRA, SFRP1, SMAD2, SOX9, TBX5, WNT3A
Go	activation of Ral GTPase activity	BivPro	18.4	0.02	RALGAPA2
Go	response to anesthetic	BivPro	18.4	0.02	BDNF
Go	Ral protein signal transduction	BivPro	18.4	0.02	RALGAPA2
Go	response to fluoxetine	BivPro	18.4	0.02	BDNF
Go	spinal cord association neuron differentiation	BivPro	3.1	0.02	GSX1, GSX2, LBX1, LHX1, LHX5, MDGA1, PAX3, PAX7, WNT3A
Go	regulation of cell-cell adhesion mediated by integrin	BivPro	4.5	0.02	FOXA2, PODXL, WNT3A
Go	cardiac pacemaker cell differentiation	BivPro	6.1	0.02	MESP1, TBX3
Go	kidney morphogenesis	BivPro	2.6	0.02	GCNT3, HES1, IRX1, IRX2, IRX3, LHX1, PAX2, SALL1, SOX9, TSHZ3, WNT4, WNT7B
Go	parathyroid gland development	BivPro	5.1	0.02	GATA3, PAX1, TBX1
Go	formation of primary germ layer	BivPro	2.2	0.02	BMP7, DUSP4, EYA2, FGF8, FOXA2, HMGA2, LHX1, MESP1, PAX2, SMAD2, T, TNRC6C, WNT3A

Go	negative regulation of cell morphogenesis involved in differentiation	BivPro	3.4	0.02	FOXA2, SFRP1, TBX5, TTC3
Go	regulation of blood coagulation	BivPro	3.0	0.02	CD34, F11, FOXA2, GP5, KRT1, PDGFA, PDGFRA, PRKCQ
Go	cardiac muscle cell fate commitment	BivPro	4.9	0.03	TBX3, WNT3A, WNT8A
Go	regulation of glial cell differentiation	BivPro	2.7	0.03	EPHA4, HES1, HMGA2, ID2, NKX2-2, NKX6-1, NKX6-2, NTF3, OLIG2, SHH, TENM4
Go	metanephros development	BivPro	2.0	0.03	BMP7, CD34, FGF8, GATA3, HES1, HOXC11, ID2, IRX1, IRX2, IRX3, LHX1, PAX2, PDGFRA, SALL1, SHH, SOX9, TSHZ3, WNT4, WNT7B
Go	dopaminergic neuron differentiation	BivPro	3.8	0.03	FGF8, FOXA2, LMX1B
Go	chromatin assembly	BivPro	2.8	0.03	H2AFY2, H3F3A, H3F3C, HIST1H2AH, HIST1H2BJ, HIST1H2BL, HIST3H2BB, HMGA2, NPM1, SOX9, TSPYL2
Go	response to vitamin A	BivPro	7.0	0.03	ALDH1A2, BDNF, PITX2
Go	signal complex assembly	BivPro	9.4	0.03	CD3E, MAPK8IP2
Go	myotube differentiation involved in skeletal muscle regeneration	BivPro	15.3	0.03	BCL9
Go	chromatin assembly or disassembly	BivPro	2.6	0.03	H2AFY2, H3F3A, H3F3C, HIST1H2AH, HIST1H2BJ, HIST1H2BL, HIST3H2BB, HMGA2, NFE2, NPM1, SOX9, TSPYL2
Go	mesonephros development	BivPro	3.0	0.04	BMP7, FGF8, GATA3, LHX1, PAX2, REN, WNT4
Go	mammary gland morphogenesis	BivPro	2.5	0.04	BTRC, FGFR2, MSX2, NTN1, PAX3, TBX3, TFAP2C, VDR, WNT4
Go	regulation of action potential	BivPro	2.1	0.04	ATRN, CHRNA1, CNR2, DPP6, JAM3, MAL, MBP, NAB1, NKX6-2, NTF3, OLIG2, P2RX2, POU3F1, POU3F2, SCN4B, SCN8A, TENM4
Go	CDP-diacylglycerol biosynthetic process	BivPro	6.7	0.04	AGPAT2, AGPAT4, AGPAT9
Go	endocrine pancreas development	BivPro	2.5	0.04	FOXA2, HES1, NKX2-2, NKX6-1, NKX6-2, PAX6, SLC2A2, SOX9
Go	response to glucose stimulus	BivPro	2.3	0.04	FOXA2, NKX6-1, PAX2, RAB11FIP2, SMAD2, SREBF1, ZNF236
Go	detection of chemical stimulus	BivPro	2.4	0.04	ABCG1, FOXA2, NKX6-1, OR6S1, P2RX2, TAS1R2
Go	determination of bilateral symmetry	BivPro	2.2	0.04	ALDH1A2, CFC1, DLL1, DYNC2LI1, FGF8, LBX1, MESP1, NKX3-2, PITX2, SHH, T, TBX1, TBX3, WNT3A, ZIC3
Go	protein-DNA complex assembly	BivPro	2.6	0.04	H2AFY2, H3F3A, H3F3C, HIST1H2AH, HIST1H2BJ, HIST1H2BL, HIST3H2BB, NPM1, SOX9, TAF1B, TSPYL2
Go	regulation of apoptotic process involved in morphogenesis	BivPro	4.5	0.04	BMP7, PAX2, VDR
Go	negative regulation of GTPase activity	BivPro	4.5	0.04	FZD10, PTPRN2
Go	DNA conformation change	BivPro	2.4	0.04	H2AFY2, H3F3A, H3F3C, HIST1H2AH, HIST1H2BJ, HIST1H2BL, HIST3H2BB, HMGA2, NCAPD3, NCAPH, NPM1, PIF1, SOX9, TSPYL2
Go	CDP-diacylglycerol metabolic process	BivPro	6.4	0.04	AGPAT2, AGPAT4, AGPAT9
Go	protein-DNA complex subunit organization	BivPro	2.5	0.05	H2AFY2, H3F3A, H3F3C, HIST1H2AH, HIST1H2BJ, HIST1H2BL, HIST3H2BB, NFE2, NPM1, SOX9, TAF1B, TSPYL2

Go	development of primary male sexual characteristics	BivPro	2.0	0.05	FGF8, GATA3, HMGA2, KIT, MAMLD1, MGST1, MSH2, PDGFRA, PITX2, REN, SFRP1, SHH, SOX3, SOX9, TBX3, TFAP2C, WNT4
Go	regulation of gliogenesis	BivPro	2.4	0.05	EPHA4, ETV5, HES1, HMGA2, ID2, NKX2-2, NKX6-1, NKX6-2, NTF3, OLIG2, SHH, TENM4
Go	gastrulation with mouth forming second	BivPro	3.0	0.05	FOXA2, LHX1, OTX2, T, TENM4
Go	specification of symmetry	BivPro	2.2	0.05	ALDH1A2, CFC1, DLL1, DYNC2LI1, FGF8, LBX1, MESP1, NKX3-2, PITX2, SHH, T, TBX1, TBX3, WNT3A, ZIC3
Go	regulation of Ral protein signal transduction	BivPro	13.1	0.05	RALGAPA2
Go	response to hexose stimulus	BivPro	2.2	0.05	FOXA2, NKX6-1, PAX2, RAB11FIP2, SMAD2, SREBF1, ZNF236
MSigDB	Genes involved in Regulation of gene expression in beta cells	BivPro	4.3	0.01	FOXA2, NKX2-2, NKX6-1, PAX6, SLC2A2
MSigDB	Maturity onset diabetes of the young	BivPro	3.5	0.03	FOXA2, HES1, NKX2-2, NKX6-1, PAX6, SLC2A2

Table 2: Enrichment terms associated with ESC bivalent promoter and enhancer BIDOCS. Each row represents an enriched term associated with BiDOCS overlapping ESC bivalent enhancer or promoters compared to all ESC bivalent enhancers or promoter. Columns indicate Term (Go Biological Processes), Description, Chromatin State (State), Regional enrichment (RE), Hypergeometric BH/FDR-adjusted P (HBH-P, for gene enrichment) and Genes associated with the term.

Tables Chapter 6

Gene	CNV_Type	P.value	Zscore	Directionality
LL22NC03-63E9.3	Duplications	6.9e-06	-4.50	+ ? ? + ? ?
ZNF280A	Duplications	1.1e-05	-4.39	+ ? ? + ? ?
ZNF280B	Duplications	1.1e-05	-4.39	+ ? ? + ? ?
PRAME	Deletions	5.3e-05	-4.04	+ ? ? - ? ?
ZNF280B	Deletions	5.3e-05	-4.04	+ ? ? - ? ?
LL22NC03-63E9.3	Deletions	5.3e-05	-4.04	+ ? ? - ? ?
ZNF280A	Deletions	5.3e-05	-4.04	+ ? ? - ? ?
PRAME	Duplications	7.1e-05	-3.97	+ ? ? + ? ?
GGTLC2	Duplications	0.00024	-3.67	+ ? ? - ? ?
GOLGA6B	Duplications	0.00025	-3.66	+ + + + + +
IGLL5	Deletions	0.00056	-3.45	+ ? ? ? ? ?
SLC18A1	Deletions	0.00057	3.45	? ? - ? ? ?
WDYHV1	Deletions	0.00059	-3.44	+ ? ? - ? ?
HHIPL2	Deletions	0.0015	-3.18	? ? ? ? ? +
TYRO3	Deletions	0.0022	-3.06	+ ? ? ? ? ?
KIAA0125	Deletions	0.0026	-3.01	+ + ? - ? +
C21orf59	Duplications	0.0033	2.94	? ? ? ? - ?
EVA1C	Duplications	0.0033	2.94	? ? ? ? - ?
AP000275.65	Duplications	0.0033	2.94	? ? ? ? - ?
TCP10L	Duplications	0.0033	2.94	? ? ? ? - ?
SLC17A6	Duplications	0.004	2.88	? - ? ? ? ?
AMY2A	Deletions	0.0041	2.87	- - + - - -
QARS	Duplications	0.0041	-2.87	? ? ? ? ? +
OR4C3	Duplications	0.0042	2.87	? - ? ? ? ?
OR4X2	Duplications	0.0042	2.87	? - ? ? ? ?
OR4X1	Duplications	0.0042	2.87	? - ? ? ? ?
OR4S1	Duplications	0.0042	2.87	? - ? ? ? ?
OR4B1	Duplications	0.0045	2.84	? - ? ? ? ?
A2ML1	Deletions	0.0047	2.83	? - - ? ? -
LAMTOR1	Duplications	0.0049	-2.81	? ? ? ? + ?
LRTOMT	Duplications	0.0049	-2.81	? ? ? ? + ?
IGLL5	Duplications	0.0053	-2.79	+ ? ? + ? ?
ABCC3	Deletions	0.0059	-2.75	? ? ? ? ? +
PSG11	Duplications	0.0062	2.73	- - - - - -
KRTAP9-8	Duplications	0.0072	2.69	- ? ? - - -
APBA1	Duplications	0.0072	-2.69	+ + + + - +
SSRP1	Duplications	0.008	-2.65	+ ? ? ? ? ?
GLB1L3	Deletions	0.0085	-2.63	+ + + - + +
CDK11A	Duplications	0.0091	-2.61	+ + + + + +
RP11-1099M24.7	Duplications	0.0092	-2.60	+ ? ? ? ? ?

KCNAB3	Duplications	0.0092	-2.60	+ ? ? ? ? ?
ANKFY1	Duplications	0.0094	-2.60	+ ? ? ? ? ?
RGPD8	Duplications	0.0094	-2.60	? + ? ? ? ?
FOCAD	Duplications	0.0095	-2.59	? ? + ? ? ?
SPDYE2B	Duplications	0.0099	-2.58	+ + + - + +
DEFB4A	Duplications	0.01	2.57	- - - - + -
PHACTR4	Duplications	0.01	2.57	? ? - ? ? ?
CNOT1	Duplications	0.01	2.57	- ? ? - ? -
POLR2J2	Duplications	0.01	-2.57	+ + + + + +
SIRPB1	Duplications	0.011	2.54	- - - + - -
RP4-576H24.4	Duplications	0.011	2.54	- - - + - -
DCLRE1C	Duplications	0.012	-2.52	+ - + ? + +
SPAG11A	Duplications	0.012	2.52	- - - - + -
CDC20	Duplications	0.012	-2.52	? ? ? ? ? +
OR4N2	Deletions	0.012	2.51	? ? ? ? - ?
OR4Q3	Deletions	0.012	2.51	? ? ? ? - ?
UPK3BL	Duplications	0.012	-2.51	+ + + + + +
DEFB103A	Duplications	0.013	2.49	- - - - + -
LPA	Duplications	0.014	2.46	- + - - + -
CGB2	Deletions	0.014	-2.46	? + + ? + +
RBCK1	Duplications	0.014	2.45	? ? - ? ? ?
CCDC74A	Deletions	0.014	2.45	- - - + - -
KIR3DL2	Duplications	0.015	2.43	- - ? - + -
ASB13	Duplications	0.017	2.40	? - ? ? ? ?
C21orf58	Duplications	0.017	2.38	? ? - ? ? ?
FREM2	Duplications	0.019	-2.35	+ ? + ? ? -
PLIN1	Deletions	0.019	2.34	? ? - ? ? ?
RASA4	Duplications	0.021	-2.31	+ + + + + -
RP11-514P8.6	Duplications	0.021	-2.31	+ + + + + -
SHKBP1	Deletions	0.021	-2.31	+ + ? + + +
TCP10	Duplications	0.021	-2.31	- + + + + +
IQGAP3	Duplications	0.021	2.31	- ? ? ? ? ?
NAALAD2	Deletions	0.022	-2.28	+ + + + - +
CCDC73	Duplications	0.023	-2.27	? + ? ? ? ?
TMC7	Duplications	0.024	2.26	? ? - ? ? ?
ATXN2L	Duplications	0.024	2.26	? ? ? - ? ?
FABP2	Deletions	0.024	-2.25	+ ? + ? ? ?
ELOVL1	Duplications	0.025	-2.24	? ? ? ? ? +
KTN1	Duplications	0.026	-2.23	+ ? - ? + ?
CYP4A22	Deletions	0.026	-2.22	+ + + + + +
DEFB104A	Duplications	0.026	2.22	- - + - - -

MEIG1	Duplications	0.027	-2.21	+ - + ? + +
CLEC18C	Deletions	0.028	-2.20	+ + + ? ? ?
CHRFAM7A	Duplications	0.028	-2.20	+ + - + + +
TAS2R30	Duplications	0.028	2.20	- - ? ? ? ?
RAB21	Duplications	0.028	-2.20	? ? + ? ? ?
TBC1D15	Duplications	0.028	-2.20	? ? + ? ? ?
ZDHHC11B	Deletions	0.029	-2.19	- + - + + +
ANKRD36	Duplications	0.029	-2.19	- + + ? + ?
NEK4	Duplications	0.029	-2.18	? ? ? ? ? +
VPREB1	Deletions	0.031	-2.16	+ ? ? ? ? ?
EHF	Duplications	0.031	2.16	- ? ? ? ? ?
RB1CC1	Duplications	0.031	2.15	? ? - ? ? -
ANKRD27	Duplications	0.032	-2.14	+ ? ? ? + ?
ENO2	Deletions	0.033	2.13	? ? ? - ? ?
PRAMEF7	Deletions	0.034	-2.12	+ + + + + -
SENP1	Duplications	0.034	-2.12	+ ? ? ? ? ?
PFKM	Duplications	0.034	-2.12	+ ? ? ? ? ?
AADAC	Deletions	0.034	2.12	- ? ? - - -
MUC20	Deletions	0.036	2.10	+ - - - - -
MUC4	Deletions	0.036	2.10	+ - - - - -
KMT2C	Deletions	0.037	2.09	? - - ? ? -
SPDL1	Duplications	0.037	-2.09	? ? ? ? ? +
PRAMEF12	Deletions	0.037	-2.09	+ - - + + +
MFF	Deletions	0.038	2.08	? - ? ? ? ?
VPS37C	Deletions	0.038	-2.08	+ + ? + + +
CELA1	Deletions	0.038	-2.07	+ + ? ? ? ?
TICRR	Duplications	0.039	-2.07	? ? ? + ? ?
KIF7	Duplications	0.039	-2.07	? ? ? + ? ?
CLPSL1	Deletions	0.039	2.07	- - + - - -
IDNK	Duplications	0.039	-2.07	? ? ? ? ? +
STRC	Duplications	0.04	2.06	- - + - + -
COX19	Deletions	0.04	2.06	? ? ? ? ? -
SUN1	Deletions	0.04	2.06	? ? ? ? ? -
ADAP1	Deletions	0.04	2.06	? ? ? ? ? -
GET4	Deletions	0.04	2.06	? ? ? ? ? -
RGPD8	Deletions	0.04	-2.06	+ + ? ? + ?
SPRR2D	Deletions	0.04	-2.05	? + + - + ?
FCAR	Deletions	0.041	2.05	- ? ? ? - ?
PGA5	Duplications	0.041	2.04	- + - + - -
PGA4	Duplications	0.041	2.04	- + - + - -
COL18A1	Duplications	0.041	2.04	- ? ? - ? -

CGB	Deletions	0.043	-2.03	? + ? ? + ?
IGFL3	Deletions	0.043	-2.03	+ + - + + +
PTK2B	Deletions	0.044	-2.02	+ ? ? ? ? ?
ZNF705G	Deletions	0.044	2.02	- - - + - -
TXNRD1	Duplications	0.044	-2.01	? ? + ? ? ?
EID3	Duplications	0.044	-2.01	? ? + ? ? ?
OR2T1	Deletions	0.044	2.01	- - ? ? ? ?
OR2T4	Deletions	0.044	2.01	- - ? ? ? ?
OR2T6	Deletions	0.044	2.01	- - ? ? ? ?
GNPTAB	Duplications	0.045	2.01	? - ? ? ? ?
DCAF13	Deletions	0.045	-2.00	+ ? ? ? ? ?
ALDH1A1	Duplications	0.046	-2.00	? ? + ? ? +
ARL17B	Deletions	0.046	-2.00	? + ? ? + ?
LRRC37A	Deletions	0.046	-2.00	? + ? ? + ?
PCDH15	Deletions	0.046	-2.00	? ? ? ? ? +
ACAA2	Duplications	0.047	1.99	? ? ? ? ? -
TSC1	Duplications	0.047	-1.99	+ ? ? - ? ?
CLPS	Deletions	0.048	1.98	- - + - - -
CACNA1H	Duplications	0.048	-1.98	? ? ? ? ? +
TPSG1	Duplications	0.048	-1.98	? ? ? ? ? +
COPRS	Duplications	0.048	-1.98	? ? ? ? ? +
LRCH3	Duplications	0.048	1.98	- ? ? ? - ?
THOC3	Deletions	0.048	1.98	? - - ? ? ?
RP11-826N14.2	Deletions	0.048	1.98	? - - ? ? ?
CTD-2207O23.3	Duplications	0.05	1.96	? ? ? ? ? -
ARHGEF18	Duplications	0.05	1.96	? ? ? ? ? -
MINK1	Duplications	0.05	-1.96	+ + ? ? ? ?

Table 3: CNVs with nominal significant association for T2D For each variant, the overlapping gene, the CNV type (deletion or duplication), the T2D-meta-analysis P-value, Z-score and Direction of effect per study are shown. Directionality for the studies is given in the following order: European, African America, East-Asian, T2D-genes Hispanics1 , South-Asian and SIGMA Hispanics2

Gene	CNV_Type	P.value	Zscore	Directionality
RP11-1396O13.13	Deletions	6.9e-07	4.96	- - - - -
DEFB131	Deletions	9.4e-06	4.43	- - - - -
SIRPD	Deletions	4e-05	-4.11	+ ? ? ? ?
ART3	Duplications	0.00041	3.53	? ? - ? ?
FREM2	Duplications	0.0011	3.27	- ? - ? ?
ALDH1A1	Deletions	0.0018	-3.13	? ? + ? ?
ZNF276	Deletions	0.0019	-3.10	+ ? ? ? ?
PADI4	Duplications	0.0024	-3.04	? ? + + ?
FAM131C	Duplications	0.0025	-3.02	+ ? + ? +
HNRNPC	Duplications	0.003	-2.96	+ + + ? +
SLC5A4	Deletions	0.0031	-2.96	? + ? ? ?
ATP2C2	Deletions	0.0031	-2.96	? ? + ? +
MRGPRX1	Duplications	0.0033	-2.94	+ + + + +
ITGAM	Duplications	0.0035	2.92	? - ? ? ?
ITGAX	Duplications	0.0035	2.92	? - ? ? ?
UPK3BL	Deletions	0.0035	2.92	- - - ? -
POLR2J2	Deletions	0.0035	2.92	- - - ? -
PRPH	Duplications	0.0036	-2.92	+ + ? + ?
ASAH2C	Deletions	0.0037	2.90	? - ? ? -
CWC25	Duplications	0.0045	-2.84	? ? ? ? +
C17orf98	Duplications	0.0045	-2.84	? ? ? ? +
KANSL1	Deletions	0.0046	-2.83	+ - ? ? +
MCTP2	Duplications	0.0047	-2.83	+ ? ? ? ?
PCDHGA12	Duplications	0.0048	-2.82	+ ? ? ? ?
PARD3	Duplications	0.0052	-2.80	? ? + ? ?
CST1	Deletions	0.006	-2.75	? + ? ? ?
APBA1	Duplications	0.0061	-2.74	+ + - + +
PSPN	Duplications	0.0064	2.73	? ? ? ? -
RSPH10B	Duplications	0.0067	-2.71	+ + + + +
PCDHGB2	Duplications	0.0068	-2.71	+ ? ? ? ?
PCDHGA9	Duplications	0.0068	-2.71	+ ? ? ? ?
PCDHGA8	Duplications	0.0068	-2.71	+ ? ? ? ?
PCDHGB6	Duplications	0.0068	-2.71	+ ? ? ? ?
PCDHGA3	Duplications	0.0068	-2.71	+ ? ? ? ?
PCDHGA7	Duplications	0.0068	-2.71	+ ? ? ? ?
PCDHGB3	Duplications	0.0068	-2.71	+ ? ? ? ?
PCDHGA6	Duplications	0.0068	-2.71	+ ? ? ? ?
PCDHGA10	Duplications	0.0068	-2.71	+ ? ? ? ?
PCDHGB4	Duplications	0.0068	-2.71	+ ? ? ? ?
PCDHGA1	Duplications	0.0068	-2.71	+ ? ? ? ?

PCDHGA11	Duplications	0.0068	-2.71	+ ? ? ? ?
PCDHGB7	Duplications	0.0068	-2.71	+ ? ? ? ?
PCDHGA4	Duplications	0.0068	-2.71	+ ? ? ? ?
PCDHGB1	Duplications	0.0068	-2.71	+ ? ? ? ?
PCDHGA5	Duplications	0.0068	-2.71	+ ? ? ? ?
PCDHGA2	Duplications	0.0068	-2.71	+ ? ? ? ?
SPATA8	Deletions	0.007	-2.70	? ? + ? ?
C21orf90	Deletions	0.0072	-2.69	+ ? ? ? ?
TSPEAR	Deletions	0.0072	-2.69	+ ? ? ? ?
ST7	Duplications	0.0078	-2.66	? ? + ? ?
CAPZA2	Duplications	0.0078	-2.66	? ? + ? ?
PKHD1	Deletions	0.0078	-2.66	? ? + ? ?
MUC20	Deletions	0.008	-2.65	+ + + - +
MUC4	Deletions	0.008	-2.65	+ + + - +
DRC1	Deletions	0.008	-2.65	? ? + ? ?
ATP6V1D	Duplications	0.008	-2.65	? ? + ? ?
MPP5	Duplications	0.008	-2.65	? ? + ? ?
LHB	Deletions	0.0089	-2.62	? ? ? ? +
DHTKD1	Duplications	0.0092	2.61	? - ? - ?
DMBT1	Deletions	0.0096	2.59	- - - - -
DEFB104B	Duplications	0.0096	-2.59	? + + ? +
DEFB106B	Duplications	0.0096	-2.59	? + + ? +
CELA2A	Deletions	0.0097	2.59	- ? ? ? ?
CELA2B	Deletions	0.0097	2.59	- ? ? ? ?
ZFAT	Duplications	0.0097	-2.59	? ? + ? ?
DDX19A	Deletions	0.0098	-2.58	+ ? ? ? ?
RP11-529K1.3	Deletions	0.0098	-2.58	+ ? ? ? ?
PMS2	Duplications	0.0099	-2.58	+ + - + +
TSPAN8	Duplications	0.01	-2.56	? ? + ? ?
SPDYE2B	Deletions	0.011	2.55	- - - ? -
TMEM50A	Duplications	0.011	-2.53	? + + ? ?
CELA1	Duplications	0.012	2.51	- ? ? ? ?
ITGAD	Duplications	0.013	2.49	? - ? ? ?
SPNS3	Duplications	0.014	-2.47	? ? + ? ?
ANKRD20A4	Deletions	0.015	-2.44	+ + + - +
NECAP2	Duplications	0.015	-2.43	? ? + ? ?
SPATA21	Duplications	0.015	-2.43	? ? + ? ?
DPP7	Duplications	0.017	2.39	- ? ? ? ?
MAN1B1	Duplications	0.017	2.39	- ? ? ? ?
AARS	Deletions	0.017	2.39	? ? - ? ?
CYP2A7	Duplications	0.017	-2.38	+ + + + +

PGA3	Deletions	0.017	2.38	- - - - -
STARD6	Deletions	0.017	-2.38	? ? ? ? +
KRTAP9-6	Duplications	0.017	2.38	- ? ? - -
GSTA1	Deletions	0.018	2.38	? - - ? -
KIAA0125	Duplications	0.018	2.36	- ? - - -
SFTPA1	Duplications	0.018	-2.36	? + + ? ?
CTC-490E21.12	Duplications	0.019	-2.35	+ - + + +
RDH16	Duplications	0.019	-2.34	? + + ? -
STAC3	Deletions	0.02	2.33	? - ? ? ?
KRTAP9-8	Duplications	0.02	2.33	- ? ? - -
ACSM5	Deletions	0.02	-2.33	+ + + ? +
ASAH2B	Duplications	0.021	2.31	- - ? ? -
SELP	Deletions	0.022	2.30	? - ? ? ?
C1QTNF9B	Deletions	0.022	-2.29	+ - + ? +
PDE11A	Duplications	0.024	2.26	? ? ? ? -
SUGT1	Duplications	0.024	-2.26	? ? ? + ?
HNRNPA1L2	Duplications	0.024	-2.26	? ? ? + ?
THOC3	Deletions	0.024	-2.25	? + + ? ?
RP11-826N14.2	Deletions	0.024	-2.25	? + + ? ?
CACNB2	Duplications	0.024	-2.25	? ? + ? ?
NSUN6	Duplications	0.024	-2.25	? ? + ? ?
UBE2I	Duplications	0.025	-2.25	? ? ? ? +
SELENBP1	Duplications	0.025	-2.25	? ? + ? ?
POGZ	Duplications	0.025	-2.25	? ? + ? ?
PSMB4	Duplications	0.025	-2.25	? ? + ? ?
RAD51C	Duplications	0.025	-2.24	+ ? ? ? ?
KIAA1958	Duplications	0.025	-2.24	? + ? ? ?
INIP	Duplications	0.025	-2.24	? + ? ? ?
SNX30	Duplications	0.025	-2.24	? + ? ? ?
GDPD3	Duplications	0.025	2.23	? ? - ? ?
AFMID	Duplications	0.029	2.18	? ? - ? ?
KIAA0125	Deletions	0.03	2.17	- + ? - ?
PACRGL	Duplications	0.03	-2.17	+ ? ? ? ?
KCNIP4	Duplications	0.03	-2.17	+ ? ? ? ?
PSG9	Deletions	0.03	2.17	- + - - -
KRT34	Duplications	0.031	2.15	- - - - +
ARHGEF16	Deletions	0.032	2.15	? - ? ? ?
CLCNKA	Duplications	0.032	-2.15	+ + ? + +
TRPM1	Deletions	0.033	-2.14	+ ? ? ? ?
OR4C6	Deletions	0.033	-2.14	? + ? ? ?
OR5D14	Deletions	0.033	-2.14	? + ? ? ?

OR5D13	Deletions	0.033	-2.14	? + ? ? ?
KRTAP9-4	Duplications	0.033	2.13	- ? ? - -
KRTAP9-9	Duplications	0.033	2.13	- ? ? - -
GTF2H2	Deletions	0.033	-2.13	+ + - + +
KRTAP9-7	Duplications	0.034	2.12	- ? ? - -
ERBB2IP	Duplications	0.035	-2.11	? ? + ? ?
OR5L2	Deletions	0.035	-2.10	? + ? ? ?
OR5D18	Deletions	0.035	-2.10	? + ? ? ?
OR5L1	Deletions	0.035	-2.10	? + ? ? ?
TXNL4B	Duplications	0.036	2.10	? - - - ?
HPR	Duplications	0.036	2.10	? - - - ?
FAM153C	Duplications	0.036	2.10	- - - - -
CGB7	Duplications	0.036	2.10	- - - - -
NEDD4L	Duplications	0.036	-2.10	? + ? ? ?
PRB4	Duplications	0.037	2.09	- - ? ? ?
HRG	Deletions	0.037	-2.09	+ ? ? ? ?
TMEM254	Deletions	0.037	-2.08	? ? ? + ?
TNK2	Duplications	0.037	2.08	? - ? ? ?
ZNF257	Deletions	0.037	2.08	? - ? - ?
PADI1	Duplications	0.038	-2.08	? ? + ? ?
AC004824.2	Duplications	0.038	-2.08	? ? + ? ?
TPTE	Deletions	0.038	-2.08	+ - + + +
SETDB2	Duplications	0.038	-2.08	? ? + ? ?
TTN	Duplications	0.038	-2.08	+ - + + +
ALDH3B2	Deletions	0.039	-2.07	? ? + ? ?
IFT140	Deletions	0.039	-2.06	? + ? ? ?
GOLGA6B	Duplications	0.04	2.05	- - - - -
TAS2R31	Duplications	0.041	-2.05	+ ? ? ? ?
AC018630.1	Duplications	0.041	-2.05	+ ? ? ? ?
OR4N2	Deletions	0.041	-2.04	? ? ? ? +
OR4Q3	Deletions	0.041	-2.04	? ? ? ? +
GPR98	Duplications	0.041	-2.04	? ? + ? ?
BTNL3	Deletions	0.042	-2.04	? + + + +
KAT8	Duplications	0.042	-2.04	+ ? ? ? ?
CYP4F3	Duplications	0.042	2.03	? - ? ? ?
AL136218.1	Duplications	0.043	-2.03	? ? + ? ?
CAB39L	Duplications	0.043	-2.03	? ? + ? ?
ASRGL1	Duplications	0.044	-2.02	? + ? ? ?
AGPAT4	Duplications	0.044	-2.02	+ ? ? ? ?
TSC1	Deletions	0.044	2.01	- ? ? ? ?
SH3RF1	Duplications	0.045	-2.01	+ ? ? ? ?

C8orf33	Duplications	0.045	2.00	- - - ? -
IFT122	Deletions	0.046	2.00	? ? - ? ?
AL161645.2	Deletions	0.047	-1.99	? + + + +
UGT2B15	Duplications	0.047	-1.99	+ + ? + +
KIF27	Duplications	0.049	-1.97	? + ? ? ?
BCLAF1	Duplications	0.049	-1.97	+ + + + -
C1GALT1	Duplications	0.049	-1.97	+ ? ? ? ?
ZNF705G	Duplications	0.05	1.96	- - + - -

Table 4: CNVs with nominal significant association for BMI For each variant, the overlapping gene, the CNV type (deletion or duplication), the BMI-meta-analysis P-value, Z-score and Direction of effect per study are shown. Directionality for the studies is given in the following order: European, African America, East-Asian, T2D-genes Hispanics1 and South-Asian

Bibliography

[1] International Diabetes Foundation, IDF. IDF DIABETES ATLAS, 2015.

[2] American Diabetes Association. 2. Classification and Diagnosis of Diabetes. *Diabetes Care*, 39 Suppl 1(Supplement 1):S13–22, January 2016.

[3] American Diabetes Association. Diagnosis and classification of diabetes mellitus. American Diabetes Association, January 2014.

[4] Gerald M Reaven. Role of Insulin Resistance in Human Disease. *Diabetes*, 37(12):1595–1607, December 1988.

[5] Steven E Kahn, Mark E Cooper, and Stefano Del Prato. Pathophysiology and treatment of type 2 diabetes: perspectives on the past, present, and future. *Lancet*, 383(9922):1068–1083, March 2014.

[6] Stephen L Aronoff, Kathy Berkowitz, Barb Shreiner, and Laura Want. Glucose Metabolism and Regulation: Beyond Insulin and Glucagon. *Diabetes Spectrum*, 17(3):183–190, July 2004.

[7] Alan R Saltiel and C Ronald Kahn. Insulin signalling and the regulation of glucose and lipid metabolism. *Nature*, 414(6865):799–806, December 2001.

[8] B Ahrén. Autonomic regulation of islet hormone secretion–implications for health and disease. *Diabetologia*, 43(4):393–410, April 2000.

[9] T Kadowaki, C L Bevins, A Cama, K Ojamaa, B Marcus-Samuels, H Kadowaki, L Beitz, C McKeon, and S I Taylor. Two mutant alleles of the insulin receptor gene in a patient with extreme insulin resistance. *Science*, 240(4853):787–790, May 1988.

[10] Y Yoshimasa, S Seino, J Whittaker, T Kakehi, A Kosaki, H Kuzuya, H Imura, G I Bell, and D F Steiner. Insulin-resistant diabetes due to a point mutation that prevents insulin proreceptor processing. *Science*, 240(4853):784–787, May 1988.

[11] Stella George, Justin J Rochford, Christian Wolfrum, Sarah L Gray, Sven Schinner, Jenny C Wilson, Maria A Soos, Peter R Murgatroyd, Rachel M Williams, Carlo L Acerini, David B Dunger, David Barford, A Margot Umpleby, Nicholas J Wareham, Huw Alban Davies, Alan J Schafer, Markus Stoffel, Stephen O'Rahilly, and Ines Barroso. A family with severe insulin resistance and diabetes due to a mutation in AKT2. *Science*, 304(5675):1325–1328, May 2004.

[12] Oscar Rubio-Cabezas, Vishwajeet Puri, Incoronata Murano, Vladimir Saudek, Robert K Semple, Satya Dash, Caroline S S Hyden, William Bottomley, Corinne Vigouroux, Jocelyne Magré, Philippa Raymond-Barker, Peter R Murgatroyd, Anil Chawla, Jeremy N Skepper, V Krishna Chatterjee, Sara Suliman, Ann-Marie Patch, Anil K Agarwal, Abhimanyu Garg, Ines Barroso, Saverio Cinti, Michael P Czech, Jesús Argente, Stephen O'Rahilly, David B Savage, and LD Screening Consortium. Partial lipodystrophy and insulin resistant diabetes in a patient with a homozygous nonsense mutation in CIDEC. *EMBO Mol Med*, 1(5):280–287, August 2009.

[13] S N Jackson, T A Howlett, P G McNally, S O'Rahilly, and R C Trembath. Dunnigan-Kobberling syndrome: an autosomal dominant form of partial lipodystrophy. *QJM*, 90(1):27–36, January 1997.

[14] S Shackleton, D J Lloyd, S N Jackson, R Evans, M F Niermeijer, B M Singh, H Schmidt, G Brabant, S Kumar, P N Durrington, S Gregory, S O'Rahilly, and R C Trembath. LMNA,

encoding lamin A/C, is mutated in partial lipodystrophy. *Nat. Genet.*, 24(2):153–156, February 2000.

[15] Maura Agostini, Erik Schoenmakers, Catherine Mitchell, Istvan Szatmari, David Savage, Aaron Smith, Odelia Rajanayagam, Robert Semple, Jian'an Luan, Louise Bath, Anthony Zalin, Mourad Labib, Sudhesh Kumar, Helen Simpson, Dirk Blom, David Marais, John Schwabe, Ines Barroso, Richard Trembath, Nicholas Wareham, Laszlo Nagy, Mark Gurnell, Stephen O'Rahilly, and Krishna Chatterjee. Non-DNA binding, dominant-negative, human PPARgamma mutations cause lipodystrophic insulin resistance. *Cell Metab.*, 4(4):303–311, October 2006.

[16] Anil K Agarwal, Jean-Pierre Fryns, Richard J Auchus, and Abhimanyu Garg. Zinc metalloproteinase, ZMPSTE24, is mutated in mandibuloacral dysplasia. *Hum. Mol. Genet.*, 12(16):1995–2001, August 2003.

[17] Robert K Semple, David B Savage, Elaine K Cochran, Phillip Gorden, and Stephen O'Rahilly. Genetic Syndromes of Severe Insulin Resistance. *Endocrine Reviews*, 32(4):498–514, May 2011.

[18] Benjamin F Voight, Laura J Scott, Valgerdur Steinthorsdottir, Andrew P Morris, Christian Dina, Ryan P Welch, Eleftheria Zeggini, Cornelia Huth, Yurii S Aulchenko, Gudmar Thorleifsson, Laura J McCulloch, Teresa Ferreira, Harald Grallert, Najaf Amin, Guanming Wu, Cristen J Willer, Soumya Raychaudhuri, Steve A McCarroll, Claudia Langenberg, Oliver M Hofmann, Josée Dupuis, Lu Qi, Ayellet V Segrè, Mandy van Hoek, Pau Navarro, Kristin Ardlie, Beverley Balkau, Rafn Benediktsson, Amanda J Bennett, Roza Blagieva, Eric Boerwinkle, Lori L Bonnycastle, Kristina Bengtsson Boström, Bert Bravenboer, Suzannah Bumpstead, Noisël P Burtt, Guillaume Charpentier, Peter S Chines, Marilyn Cornelis, David J Couper, Gabe Crawford, Alex S F Doney, Katherine S Elliott, Amanda L Elliott, Michael R Erdos, Caroline S Fox, Christopher S Franklin, Martha Ganser, Christian Gieger, Niels Grarup, Todd Green, Simon Griffin, Christopher J Groves, Candace Guiducci, Samy Hadjadj, Neelam Hassanali, Christian Herder, Bo Isomaa, Anne U Jackson, Paul R V Johnson, Torben Jørgensen, Wen H L Kao, Norman Klopp, Augustine Kong, Peter Kraft, Johanna Kuusisto, Torsten Lauritzen, Man Li, Aloysius Lieverse, Cecilia M Lindgren, Valeriya

Lyssenko, Michel Marre, Thomas Meitinger, Kristian Midthjell, Mario A Morken, Narisu Narisu, Peter Nilsson, Katharine R Owen, Felicity Payne, John R B Perry, Ann-Kristin Petersen, Carl Platou, Christine Proença, Inga Prokopenko, Wolfgang Rathmann, N William Rayner, Neil R Robertson, Ghislain Rocheleau, Michael Roden, Michael J Sampson, Richa Saxena, Beverley M Shields, Peter Shrader, Gunnar Sigurdsson, Thomas Sparsø, Klaus Strassburger, Heather M Stringham, Qi Sun, Amy J Swift, Barbara Thorand, Jean Tichet, Tiinamaija Tuomi, Rob M van Dam, Timon W van Haeften, Thijs van Herpt, Jana V van Vliet-Ostaptchouk, G Bragi Walters, Michael N Weedon, Cisca Wijmenga, Jacqueline Witteman, Richard N Bergman, Stephane Cauchi, Francis S Collins, Anna L Gloyn, Ulf Gyllensten, Torben Hansen, Winston A Hide, Graham A Hitman, Albert Hofman, David J Hunter, Kristian Hveem, Markku Laakso, Karen L Mohlke, Andrew D Morris, Colin N A Palmer, Peter P Pramstaller, Igor Rudan, Eric Sijbrands, Lincoln D Stein, Jaakko Tuomilehto, Andre Uitterlinden, Mark Walker, Nicholas J Wareham, Richard M Watanabe, Goncalo R Abecasis, Bernhard O Boehm, Harry Campbell, Mark J Daly, Andrew T Hattersley, Frank B Hu, James B Meigs, James S Pankow, Oluf Pedersen, H-Erich Wichmann, Ines Barroso, Jose C Florez, Timothy M Frayling, Leif Groop, Rob Sladek, Unnur Thorsteinsdottir, James F Wilson, Thomas Illig, Philippe Froguel, Cornelia M van Duijn, Kari Stefansson, David Altshuler, Michael Boehnke, Mark I McCarthy, MAGIC investigators, and GIANT Consortium. Twelve type 2 diabetes susceptibility loci identified through large-scale association analysis. *Nat. Genet.*, 42(7):579–589, July 2010.

[19] Antigone S Dimas, Vasiliki Lagou, Adam Barker, Joshua W Knowles, Reedik Mägi, Marie-France Hivert, Andrea Benazzo, Denis Rybin, Anne U Jackson, Heather M Stringham, Ci Song, Antje Fischer-Rosinsky, Trine Welløv Boesgaard, Niels Grarup, Fahim A Abbasi, Themistocles L Assimes, Ke Hao, Xia Yang, Cécile Lecoeur, Ines Barroso, Lori L Bonnycastle, Yvonne Böttcher, Suzannah Bumpstead, Peter S Chines, Michael R Erdos, Jürgen Graessler, Peter Kovacs, Mario A Morken, Narisu Narisu, Felicity Payne, Alena Stančáková, Amy J Swift, Anke Tönjes, Stefan R Bornstein, Stephane Cauchi, Philippe Froguel, David Meyre, Peter E H Schwarz, Hans-Ulrich Häring, Ulf Smith, Michael Boehnke, Richard N Bergman, Francis S Collins, Karen L Mohlke, Jaakko Tuomilehto, Thomas Quertemous, Lars Lind, Torben Hansen, Oluf Pedersen, Mark Walker, Andreas F H Pfeiffer, Joachim

Spranger, Michael Stumvoll, James B Meigs, Nicholas J Wareham, Johanna Kuusisto, Markku Laakso, Claudia Langenberg, Josée Dupuis, Richard M Watanabe, Jose C Florez, Erik Ingelsson, Mark I McCarthy, Inga Prokopenko, and MAGIC investigators. Impact of type 2 diabetes susceptibility variants on quantitative glycemic traits reveals mechanistic heterogeneity. *Diabetes*, 63(6):2158–2171, June 2014.

[20] Andrew A Gumbs, Irvin M Modlin, and Garth H Ballantyne. Changes in Insulin Resistance Following Bariatric Surgery: Role of Caloric Restriction and Weight Loss. *OBES SURG*, 15(4):462–473, 2005.

[21] Steven E Nissen and Kathy Wolski. Effect of rosiglitazone on the risk of myocardial infarction and death from cardiovascular causes. *N. Engl. J. Med.*, 356(24):2457–2471, June 2007.

[22] Steven E Nissen and Kathy Wolski. Rosiglitazone revisited: an updated meta-analysis of risk for myocardial infarction and cardiovascular mortality. *Arch. Intern. Med.*, 170(14):1191–1201, July 2010.

[23] Philip D Home, Stuart J Pocock, Henning Beck-Nielsen, Paula S Curtis, Ramon Gomis, Markolf Hanefeld, Nigel P Jones, Michel Komajda, John J V McMurray, and RECORD Study Team. Rosiglitazone evaluated for cardiovascular outcomes in oral agent combination therapy for type 2 diabetes (RECORD): a multicentre, randomised, open-label trial. *Lancet*, 373(9681):2125–2135, June 2009.

[24] Kenneth W Mahaffey, Gail Hafley, Sheila Dickerson, Shana Burns, Sandra Tourt-Uhlig, Jennifer White, L Kristin Newby, Michel Komajda, John McMurray, Robert Bigelow, Philip D Home, and Renato D Lopes. Results of a reevaluation of cardiovascular outcomes in the RECORD trial. *Am. Heart J.*, 166(2):240–249.e1, August 2013.

[25] Raymond E Soccio, Eric R Chen, and Mitchell A Lazar. Thiazolidinediones and the Promise of Insulin Sensitization in Type 2 Diabetes. *Cell Metab.*, 20(4):573–591, October 2014.

[26] Hongying An and Ling He. Current understanding of metformin effect on the control of hyperglycemia in diabetes. *J. Endocrinol.*, 228(3):R97–106, March 2016.

[27] C Henzen. Monogenic diabetes mellitus due to defects in insulin secretion. *Swiss Med Wkly*, 2012.

[28] Mitsuhisa Komatsu, Masahiro Takei, Hiroaki Ishii, and Yoshihiko Sato. Glucose-stimulated insulin secretion: A newer perspective. *Journal of Diabetes Investigation*, 4(6):511–516, November 2013.

[29] Anna L Gloyn, Ewan R Pearson, Jennifer F Antcliff, Peter Proks, G Jan Bruining, Annabelle S Slingerland, Neville Howard, Shubha Srinivasan, José M C L Silva, Janne Molnes, Emma L Edghill, Timothy M Frayling, I Karen Temple, Deborah Mackay, Julian P H Shield, Zdenek Sumnik, Adrian van Rhijn, Jerry K H Wales, Penelope Clark, Shaun Gorman, Javier Aisenberg, Sian Ellard, Pål R Njølstad, Frances M Ashcroft, and Andrew T Hattersley. Activating Mutations in the Gene Encoding the ATP-Sensitive Potassium-Channel Subunit Kir6.2 and Permanent Neonatal Diabetes. *http://dx.doi.org/10.1056/NEJMoa032922*, 350(18):1838–1849, October 2009.

[30] S E Flanagan, E L Edghill, A L Gloyn, S Ellard, and A T Hattersley. Mutations in KCNJ11, which encodes Kir6.2, are a common cause of diabetes diagnosed in the first 6 months of life, with the phenotype determined by genotype. *Diabetologia*, 49(6):1190–1197, 2006.

[31] Anna L Gloyn, Catherine Diatloff-Zito, Emma L Edghill, Christine Bellanné-Chantelot, Sylvie Nivot, Régis Coutant, Sian Ellard, Andrew T Hattersley, and Jean Jacques Robert. KCNJ11 activating mutations are associated with developmental delay, epilepsy and neonatal diabetes syndrome and other neurological features. *Eur. J. Hum. Genet.*, 14(7):824–830, July 2006.

[32] Andrey P Babenko, Michel Polak, Hélène Cavé, Kanetee Busiah, Paul Czernichow, Raphael Scharfmann, Joseph Bryan, Lydia Aguilar-Bryan, Martine Vaxillaire, and Philippe Froguel. Activating Mutations in the ABCC8Gene in Neonatal Diabetes Mellitus. *N. Engl. J. Med.*, 355(5):456–466, August 2006.

[33] Martine Vaxillaire, Aurélie Dechaume, Kanetee Busiah, Hélène Cavé, Sabrina Pereira, Raphael Scharfmann, Guiomar Perez de Nanclares, Luis Castaño, Philippe Froguel, Michel Polak, and SUR1-Neonatal Diabetes Study Group. New ABCC8 mutations in relapsing neonatal diabetes and clinical features. *Diabetes*, 56(6):1737–1741, June 2007.

[34] Ewan R Pearson, Isabelle Flechtner, Pål R Njølstad, Maciej T Malecki, Sarah E Flanagan,

Brian Larkin, Frances M Ashcroft, Iwar Klimes, Ethel Codner, Violeta Iotova, Annabelle S Slingerland, Julian Shield, Jean Jacques Robert, Jens J Holst, Penny M Clark, Sian Ellard, Oddmund Søvik, Michel Polak, Andrew T Hattersley, and Neonatal Diabetes International Collaborative Group. Switching from insulin to oral sulfonylureas in patients with diabetes due to Kir6.2 mutations. *N. Engl. J. Med.*, 355(5):467–477, August 2006.

[35] P R Njølstad, O Søvik, A Cuesta-Muñoz, L Bjørkhaug, O Massa, F Barbetti, D E Undlien, C Shiota, M A Magnuson, A Molven, F M Matschinsky, and G I Bell. Neonatal diabetes mellitus due to complete glucokinase deficiency. *N. Engl. J. Med.*, 344(21):1588–1592, May 2001.

[36] B Glaser, P Kesavan, M Heyman, E Davis, A Cuesta, A Buchs, C A Stanley, P S Thornton, M A Permutt, F M Matschinsky, and K C Herold. Familial hyperinsulinism caused by an activating glucokinase mutation. *N. Engl. J. Med.*, 338(4):226–230, January 1998.

[37] P S Thornton, M S Satin-Smith, K Herold, B Glaser, K C Chiu, A Nestorowicz, M A Permutt, L Baker, and C A Stanley. Familial hyperinsulinism with apparent autosomal dominant inheritance: clinical and genetic differences from the autosomal recessive variant. *J. Pediatr.*, 132(1):9–14, January 1998.

[38] N Vionnet, M Stoffel, J Takeda, K Yasuda, G I Bell, H Zouali, S Lesage, G Velho, F Iris, and P Passa. Nonsense mutation in the glucokinase gene causes early-onset non-insulin-dependent diabetes mellitus. *Nature*, 356(6371):721–722, April 1992.

[39] Martine Vaxillaire, Amélie Bonnefond, and Philippe Froguel. The lessons of early-onset monogenic diabetes for the understanding of diabetes pathogenesis. *Best Pract. Res. Clin. Endocrinol. Metab.*, 26(2):171–187, April 2012.

[40] Divya Sri Priyanka Tallapragada, Seema Bhaskar, and Giriraj R Chandak. New insights from monogenic diabetes for "common" type 2 diabetes. *Front Genet*, 6:251, 2015.

[41] Valeriya Lyssenko, Anna Jonsson, Peter Almgren, Nicoló Pulizzi, Bo Isomaa, Tiinamaija Tuomi, Göran Berglund, David Altshuler, Peter Nilsson, and Leif Groop. Clinical risk factors, DNA variants, and the development of type 2 diabetes. *N. Engl. J. Med.*, 359(21):2220–2232, November 2008.

[42] Rashmi B Prasad and Leif Groop. Genetics of type 2 diabetes-pitfalls and possibilities. *Genes (Basel)*, 6(1):87–123, 2015.

[43] P Almgren, M Lehtovirta, B Isomaa, L Sarelin, M R Taskinen, V Lyssenko, T Tuomi, L Groop, and Botnia Study Group. Heritability and familiality of type 2 diabetes and related quantitative traits in the Botnia Study. *Diabetologia*, 54(11):2811–2819, November 2011.

[44] B Newman, J V Selby, M C King, C Slemenda, R Fabsitz, and G D Friedman. Concordance for type 2 (non-insulin-dependent) diabetes mellitus in male twins. *Diabetologia*, 30(10):763–768, October 1987.

[45] J Kaprio, J Tuomilehto, M Koskenvuo, K Romanov, A Reunanen, J Eriksson, J Stengård, and Y A Kesäniemi. Concordance for type 1 (insulin-dependent) and type 2 (non-insulin-dependent) diabetes mellitus in a population-based cohort of twins in Finland. *Diabetologia*, 35(11):1060–1067, November 1992.

[46] P Poulsen, K O Kyvik, A Vaag, and H Beck-Nielsen. Heritability of type II (non-insulin-dependent) diabetes mellitus and abnormal glucose tolerance–a population-based twin study. *Diabetologia*, 42(2):139–145, February 1999.

[47] F Medici, M Hawa, A Ianari, D A Pyke, and R D Leslie. Concordance rate for type II diabetes mellitus in monozygotic twins: actuarial analysis. *Diabetologia*, 42(2):146–150, February 1999.

[48] A Pal and M I McCarthy. The genetics of type 2 diabetes and its clinical relevance. *Clin. Genet.*, 83(4):297–306, April 2013.

[49] D Altshuler, J N Hirschhorn, M Klannemark, C M Lindgren, M C Vohl, J Nemesh, C R Lane, S F Schaffner, S Bolk, C Brewer, T Tuomi, D Gaudet, T J Hudson, M Daly, L Groop, and E S Lander. The common PPARgamma Pro12Ala polymorphism is associated with decreased risk of type 2 diabetes. *Nat. Genet.*, 26(1):76–80, September 2000.

[50] Anna L Gloyn, Michael N Weedon, Katharine R Owen, Martina J Turner, Bridget A Knight, Graham Hitman, Mark Walker, Jonathan C Levy, Mike Sampson, Stephanie Halford, Mark I

McCarthy, Andrew T Hattersley, and Timothy M Frayling. Large-scale association studies of variants in genes encoding the pancreatic beta-cell KATP channel subunits Kir6.2 (KCNJ11) and SUR1 (ABCC8) confirm that the KCNJ11 E23K variant is associated with type 2 diabetes. *Diabetes*, 52(2):568–572, February 2003.

[51] Robert Sladek, Ghislain Rocheleau, Johan Rung, Christian Dina, Lishuang Shen, David Serre, Philippe Boutin, Daniel Vincent, Alexandre Belisle, Samy Hadjadj, Beverley Balkau, Barbara Heude, Guillaume Charpentier, Thomas J Hudson, Alexandre Montpetit, Alexey V Pshezhetsky, Marc Prentki, Barry I Posner, David J Balding, David Meyre, Constantin Polychronakos, and Philippe Froguel. A genome-wide association study identifies novel risk loci for type 2 diabetes. *Nature*, 445(7130):881–885, February 2007.

[52] Diabetes Genetics Initiative of Broad Institute of Harvard and MIT, Lund University, and Novartis Institutes of BioMedical Research, Richa Saxena, Benjamin F Voight, Valeriya Lyssenko, Noël P Burtt, Paul I W de Bakker, Hong Chen, Jeffrey J Roix, Sekar Kathiresan, Joel N Hirschhorn, Mark J Daly, Thomas E Hughes, Leif Groop, David Altshuler, Peter Almgren, Jose C Florez, Joanne Meyer, Kristin Ardlie, Kristina Bengtsson Boström, Bo Isomaa, Guillaume Lettre, Ulf Lindblad, Helen N Lyon, Olle Melander, Christopher Newton-Cheh, Peter Nilsson, Marju Orho-Melander, Lennart Råstam, Elizabeth K Speliotes, Marja-Riitta Taskinen, Tiinamaija Tuomi, Candace Guiducci, Anna Berglund, Joyce Carlson, Lauren Gianniny, Rachel Hackett, Liselotte Hall, Johan Holmkvist, Esa Laurila, Marketa Sjögren, Maria Sterner, Aarti Surti, Margareta Svensson, Malin Svensson, Ryan Tewhey, Brendan Blumenstiel, Melissa Parkin, Matthew Defelice, Rachel Barry, Wendy Brodeur, Jody Camarata, Nancy Chia, Mary Fava, John Gibbons, Bob Handsaker, Claire Healy, Kieu Nguyen, Casey Gates, Carrie Sougnez, Diane Gage, Marcia Nizzari, Stacey B Gabriel, Gung-Wei Chirn, Qicheng Ma, Hemang Parikh, Delwood Richardson, Darrell Ricke, and Shaun Purcell. Genome-wide association analysis identifies loci for type 2 diabetes and triglyceride levels. *Science*, 316(5829):1331–1336, June 2007.

[53] Laura J Scott, Karen L Mohlke, Lori L Bonnycastle, Cristen J Willer, Yun Li, William L Duren, Michael R Erdos, Heather M Stringham, Peter S Chines, Anne U Jackson, Ludmila Prokunina-Olsson, Chia-Jen Ding, Amy J Swift, Narisu Narisu, Tianle Hu, Randall Pruim,

Rui Xiao, Xiao-Yi Li, Karen N Conneely, Nancy L Riebow, Andrew G Sprau, Maurine Tong, Peggy P White, Kurt N Hetrick, Michael W Barnhart, Craig W Bark, Janet L Goldstein, Lee Watkins, Fang Xiang, Jouko Saramies, Thomas A Buchanan, Richard M Watanabe, Timo T Valle, Leena Kinnunen, Goncalo R Abecasis, Elizabeth W Pugh, Kimberly F Doheny, Richard N Bergman, Jaakko Tuomilehto, Francis S Collins, and Michael Boehnke. A genome-wide association study of type 2 diabetes in Finns detects multiple susceptibility variants. *Science*, 316(5829):1341–1345, June 2007.

[54] Wellcome Trust Case Control Consortium. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature*, 447(7145):661–678, June 2007.

[55] Andrew P Morris, Benjamin F Voight, Tanya M Teslovich, Teresa Ferreira, Ayellet V Segrè, Valgerdur Steinthorsdottir, Rona J Strawbridge, Hassan Khan, Harald Grallert, Anubha Mahajan, Inga Prokopenko, Hyun Min Kang, Christian Dina, Tonu Esko, Ross M Fraser, Stavroula Kanoni, Ashish Kumar, Vasiliki Lagou, Claudia Langenberg, Jian'an Luan, Cecilia M Lindgren, Martina Müller-Nurasyid, Sonali Pechlivanis, N William Rayner, Laura J Scott, Steven Wiltshire, Loic Yengo, Leena Kinnunen, Elizabeth J Rossin, Soumya Raychaudhuri, Andrew D Johnson, Antigone S Dimas, Ruth J F Loos, Sailaja Vedantam, Han Chen, Jose C Florez, Caroline Fox, Ching-Ti Liu, Denis Rybin, David J Couper, Wen Hong L Kao, Man Li, Marilyn C Cornelis, Peter Kraft, Qi Sun, Rob M van Dam, Heather M Stringham, Peter S Chines, Krista Fischer, Pierre Fontanillas, Oddgeir L Holmen, Sarah E Hunt, Anne U Jackson, Augustine Kong, Robert Lawrence, Julia Meyer, John R B Perry, Carl G P Platou, Simon Potter, Emil Rehnberg, Neil Robertson, Suthesh Sivapalaratnam, Alena Stančáková, Kathleen Stirrups, Gudmar Thorleifsson, Emmi Tikkanen, Andrew R Wood, Peter Almgren, Mustafa Atalay, Rafn Benediktsson, Lori L Bonnycastle, Noël Burtt, Jason Carey, Guillaume Charpentier, Andrew T Crenshaw, Alex S F Doney, Mozhgan Dorkhan, Sarah Edkins, Valur Emilsson, Elodie Eury, Tom Forsen, Karl Gertow, Bruna Gigante, George B Grant, Christopher J Groves, Candace Guiducci, Christian Herder, Astradur B Hreidarsson, Jennie Hui, Alan James, Anna Jonsson, Wolfgang Rathmann, Norman Klopp, Jasmina Kravic, Kaarel Krjutškov, Cordelia Langford, Karin Leander, Eero Lindholm, Stéphane Lobbens, Satu Männistö, Ghazala Mirza, Thomas W Mühleisen, Bill Musk, Melissa Parkin, Loukianos Rallidis, Jouko Saramies, Bengt Sennblad, Sonia Shah,

Gunnar Sigurðsson, Angela Silveira, Gerald Steinbach, Barbara Thorand, Joseph Trakalo, Fabrizio Veglia, Roman Wennauer, Wendy Winckler, Delilah Zabaneh, Harry Campbell, Cornelia van Duijn, Andre G Uitterlinden, Albert Hofman, Eric Sijbrands, Goncalo R Abecasis, Katharine R Owen, Eleftheria Zeggini, Mieke D Trip, Nita G Forouhi, Ann-Christine Syvänen, Johan G Eriksson, Leena Peltonen, Markus M Nöthen, Beverley Balkau, Colin N A Palmer, Valeriya Lyssenko, Tiinamaija Tuomi, Bo Isomaa, David J Hunter, Lu Qi, Wellcome Trust Case Control Consortium, Meta-Analyses of Glucose and Insulin-related traits Consortium (MAGIC) Investigators, Genetic Investigation of ANthropometric Traits (GIANT) Consortium, Asian Genetic Epidemiology Network–Type 2 Diabetes (AGEN-T2D) Consortium, South Asian Type 2 Diabetes (SAT2D) Consortium, Alan R Shuldiner, Michael Roden, Ines Barroso, Tom Wilsgaard, John Beilby, Kees Hovingh, Jackie F Price, James F Wilson, Rainer Rauramaa, Timo A Lakka, Lars Lind, George Dedoussis, Inger Njølstad, Nancy L Pedersen, Kay-Tee Khaw, Nicholas J Wareham, Sirkka M Keinanen-Kiukaanniemi, Timo E Saaristo, Eeva Korpi-Hyövälti, Juha Saltevo, Markku Laakso, Johanna Kuusisto, Andres Metspalu, Francis S Collins, Karen L Mohlke, Richard N Bergman, Jaakko Tuomilehto, Bernhard O Boehm, Christian Gieger, Kristian Hveem, Stephane Cauchi, Philippe Froguel, Damiano Baldassarre, Elena Tremoli, Steve E Humphries, Danish Saleheen, John Danesh, Erik Ingelsson, Samuli Ripatti, Veikko Salomaa, Raimund Erbel, Karl-Heinz Jöckel, Susanne Moebus, Annette Peters, Thomas Illig, Ulf de Faire, Anders Hamsten, Andrew D Morris, Peter J Donnelly, Timothy M Frayling, Andrew T Hattersley, Eric Boerwinkle, Olle Melander, Sekar Kathiresan, Peter M Nilsson, Panos Deloukas, Unnur Thorsteinsdottir, Leif C Groop, Kari Stefansson, Frank Hu, James S Pankow, Josée Dupuis, James B Meigs, David Altshuler, Michael Boehnke, Mark I McCarthy, and DIAbetes Genetics Replication And Meta-analysis (DIAGRAM) Consortium. Large-scale association analysis provides insights into the genetic architecture and pathophysiology of type 2 diabetes. *Nat. Genet.*, 44(9):981–990, September 2012.

[56] Xiao Ou Shu, Jirong Long, Qiuyin Cai, Lu Qi, Yong-Bing Xiang, Yoon Shin Cho, E Shyong Tai, Xiangyang Li, Xu Lin, Wong-Ho Chow, Min Jin Go, Mark Seielstad, Wei Bao, Huaixing Li, Marilyn C Cornelis, Kai Yu, Wanqing Wen, Jiajun Shi, Bok-Ghee Han, Xue Ling Sim, Liegang Liu, Qibin Qi, Hyung-Lae Kim, Daniel P K Ng, Jong-Young Lee, Young-Jin

Kim, Chun Li, Yu-Tang Gao, Wei Zheng, and Frank B Hu. Identification of New Genetic Risk Variants for Type 2 Diabetes. *PLoS Genet.*, 6(9):e1001127, September 2010.

[57] E J Parra, J E Below, S Krithika, A Valladares, J L Barta, N J Cox, C L Hanis, N Wacher, J Garcia-Mena, P Hu, M D Shriver, Diabetes Genetics Replication, Meta-analysis DIAGRAM Consortium, J Kumate, P M McKeigue, J Escobedo, and M Cruz. Genome-wide association study of type 2 diabetes in a sample from Mexico City and a meta-analysis of a Mexican-American sample from Starr County, Texas. *Diabetologia*, 54(8):2038–2046, 2011.

[58] Minako Imamura, Shiro Maeda, Toshimasa Yamauchi, Kazuo Hara, Kazuki Yasuda, Takashi Morizono, Atsushi Takahashi, Momoko Horikoshi, Masahiro Nakamura, Hayato Fujita, Tatsuhiko Tsunoda, Michiaki Kubo, Hirotaka Watada, Hiroshi Maegawa, Miki Okada-Iwabu, Masato Iwabu, Nobuhiro Shojima, Toshihiko Ohshige, Shintaro Omori, Minoru Iwata, Hiroshi Hirose, Kohei Kaku, Chikako Ito, Yasushi Tanaka, Kazuyuki Tobe, Atsunori Kashiwagi, Ryuzo Kawamori, Masato Kasuga, Naoyuki Kamatani, DIAbetes Genetics Replication And Meta-analysis (DIAGRAM) Consortium, Yusuke Nakamura, and Takashi Kadowaki. A single-nucleotide polymorphism in ANK1 is associated with susceptibility to type 2 diabetes in Japanese populations. *Hum. Mol. Genet.*, 21(13):3042–3049, July 2012.

[59] Yoon Shin Cho, Chien-Hsiun Chen, Cheng Hu, Jirong Long, Rick Twee Hee Ong, Xueling Sim, Fumihiko Takeuchi, Ying Wu, Min Jin Go, Toshimasa Yamauchi, Yi-Cheng Chang, Soo-Heon Kwak, Ronald C W Ma, Ken Yamamoto, Linda S Adair, Tin Aung, Qiuyin Cai, Li-Ching Chang, Yuan-Tsong Chen, Yutang Gao, Frank B Hu, Hyung-Lae Kim, Sangsoo Kim, Young-Jin Kim, Jeannette Jen-Mai Lee, Nanette R Lee, Yun Li, Jian Jun Liu, Wei Lu, Jiro Nakamura, Eitaro Nakashima, Daniel Peng-Keat Ng, Wan Ting Tay, Fuu-Jen Tsai, Tien Yin Wong, Mitsuhiro Yokota, Wei Zheng, Rong Zhang, Congrong Wang, Wing Yee So, Keizo Ohnaka, Hiroshi Ikegami, Kazuo Hara, Young Min Cho, Nam H Cho, Tien-Jyun Chang, Yuqian Bao, Åsa K Hedman, Andrew P Morris, Mark I McCarthy, DIAGRAM Consortium, MuTHER Consortium, Ryoichi Takayanagi, Kyong Soo Park, Weiping Jia, Lee-Ming Chuang, Juliana C N Chan, Shiro Maeda, Takashi Kadowaki, Jong-Young Lee, Jer-Yuarn Wu, Yik Ying Teo, E Shyong Tai, Xiao Ou Shu, Karen L Mohlke, Norihiro Kato, Bok-Ghee Han, and Mark Seielstad. Meta-analysis of genome-wide association studies

identifies eight new loci for type 2 diabetes in east Asians. *Nat. Genet.*, 44(1):67–72, January 2012.

[60] Jaspal S Kooner, Danish Saleheen, Xueling Sim, Joban Sehmi, Weihua Zhang, Philippe Frossard, Latonya F Been, Kee Seng Chia, Antigone S Dimas, Neelam Hassanali, Tazeen Jafar, Jeremy B M Jowett, Xinzhong Li, Venkatesan Radha, Simon D Rees, Fumihiko Takeuchi, Robin Young, Tin Aung, Abdul Basit, Manickam Chidambaram, Debashish Das, Elin Grundberg, Åsa K Hedman, Zafar I Hydrie, Muhammed Islam, Chiea-Chuen Khor, Sudhir Kowlessur, Malene M Kristensen, Samuel Liju, Wei Yen Lim, David R Matthews, Jianjun Liu, Andrew P Morris, Alexandra C Nica, Janani M Pinidiyapathirage, Inga Prokopenko, Asif Rasheed, Maria Samuel, Nabi Shah, A Samad Shera, Kerrin S Small, Chen Suo, Ananda R Wickremasinghe, Tien Yin Wong, Mingyu Yang, Fan Zhang, DIAGRAM, MuTHER, Goncalo R Abecasis, Anthony H Barnett, Mark Caulfield, Panos Deloukas, Timothy M Frayling, Philippe Froguel, Norihiro Kato, Prasad Katulanda, M Ann Kelly, Junbin Liang, Viswanathan Mohan, Dharambir K Sanghera, James Scott, Mark Seielstad, Paul Z Zimmet, Paul Elliott, Yik Ying Teo, Mark I McCarthy, John Danesh, E Shyong Tai, and John C Chambers. Genome-wide association study in individuals of South Asian ancestry identifies six new type 2 diabetes susceptibility loci. *Nat. Genet.*, 43(10):984–989, October 2011.

[61] Nicholette D Palmer, Caitrin W McDonough, Pamela J Hicks, Bong H Roh, Maria R Wing, S Sandy An, Jessica M Hester, Jessica N Cooke, Meredith A Bostrom, Megan E Rudock, Matthew E Talbert, Joshua P Lewis, DIAGRAM Consortium, MAGIC Investigators, Assiamira Ferrara, Lingyi Lu, Julie T Ziegler, Michele M Sale, Jasmin Divers, Daniel Shriner, Adebowale Adeyemo, Charles N Rotimi, Maggie C Y Ng, Carl D Langefeld, Barry I Freedman, and Donald W Bowden. A Genome-Wide Association Search for Type 2 Diabetes Genes in African Americans. *PLoS ONE*, 7(1):e29202, January 2012.

[62] Huaixing Li, Wei Gan, Ling Lu, Xiao Dong, Xueyao Han, Cheng Hu, Zhen Yang, Liang Sun, Wei Bao, Pengtao Li, Meian He, Liangdan Sun, Yiqin Wang, Jingwen Zhu, Qianqian Ning, Yong Tang, Rong Zhang, Jie Wen, Di Wang, Xilin Zhu, Kunquan Guo, Xianbo Zuo, Xiaohui Guo, Handong Yang, Xianghai Zhou, DIAGRAM Consortium, AGEN-T2D Consortium,

Xuejun Zhang, Lu Qi, Ruth J F Loos, Frank B Hu, Tangchun Wu, Ying Liu, Liegang Liu, Ze Yang, Renming Hu, Weiping Jia, Linong Ji, Yixue Li, and Xu Lin. A Genome-Wide Association Study Identifies GRK5 and RASGRP1 as Type 2 Diabetes Loci in Chinese Hans. *Diabetes*, 62(1):291–298, January 2013.

[63] Anubha Mahajan, Min Jin Go, Weihua Zhang, Jennifer E Below, Kyle J Gaulton, Teresa Ferreira, Momoko Horikoshi, Andrew D Johnson, Maggie C Y Ng, Inga Prokopenko, Danish Saleheen, Xu Wang, Eleftheria Zeggini, Goncalo R Abecasis, Linda S Adair, Peter Almgren, Mustafa Atalay, Tin Aung, Damiano Baldassarre, Beverley Balkau, Yuqian Bao, Anthony H Barnett, Ines Barroso, Abdul Basit, Latonya F Been, John Beilby, Graeme I Bell, Rafn Benediktsson, Richard N Bergman, Bernhard O Boehm, Eric Boerwinkle, Lori L Bonnycastle, Noël Burtt, Qiuyin Cai, Harry Campbell, Jason Carey, Stephane Cauchi, Mark Caulfield, Juliana C N Chan, Li-Ching Chang, Tien-Jyun Chang, Yi-Cheng Chang, Guillaume Charpentier, Chien-Hsiun Chen, Han Chen, Yuan-Tsong Chen, Kee Seng Chia, Manickam Chidambaram, Peter S Chines, Nam H Cho, Young Min Cho, Lee-Ming Chuang, Francis S Collins, Marilyn C Cornelis, David J Couper, Andrew T Crenshaw, Rob M van Dam, John Danesh, Debashish Das, Ulf de Faire, George Dedoussis, Panos Deloukas, Antigone S Dimas, Christian Dina, Alex S F Doney, Peter J Donnelly, Mozhgan Dorkhan, Cornelia van Duijn, Josée Dupuis, Sarah Edkins, Paul Elliott, Valur Emilsson, Raimund Erbel, Johan G Eriksson, Jorge Escobedo, Tonu Esko, Elodie Eury, Jose C Florez, Pierre Fontanillas, Nita G Forouhi, Tom Forsen, Caroline Fox, Ross M Fraser, Timothy M Frayling, Philippe Froguel, Philippe Frossard, Yutang Gao, Karl Gertow, Christian Gieger, Bruna Gigante, Harald Grallert, George B Grant, Leif C Groop, Christopher J Groves, Elin Grundberg, Candace Guiducci, Anders Hamsten, Bok-Ghee Han, Kazuo Hara, Neelam Hassanali, Andrew T Hattersley, Caroline Hayward, Åsa K Hedman, Christian Herder, Albert Hofman, Oddgeir L Holmen, Kees Hovingh, Astradur B Hreidarsson, Cheng Hu, Frank B Hu, Jennie Hui, Steve E Humphries, Sarah E Hunt, David J Hunter, Kristian Hveem, Zafar I Hydrie, Hiroshi Ikegami, Thomas Illig, Erik Ingelsson, Muhammed Islam, Bo Isomaa, Anne U Jackson, Tazeen Jafar, Alan James, Weiping Jia, Karl-Heinz Jöckel, Anna Jonsson, Jeremy B M Jowett, Takashi Kadowaki, Hyun Min Kang, Stavroula Kanoni, Wen Hong L Kao, Sekar Kathiresan, Norihiro Kato, Prasad Katulanda, Sirkka M Keinanen-Kiukaanniemi, Ann M

Kelly, Hassan Khan, Kay-Tee Khaw, Chiea-Chuen Khor, Hyung-Lae Kim, Sangsoo Kim, Young-Jin Kim, Leena Kinnunen, Norman Klopp, Augustine Kong, Eeva Korpi-Hyövälti, Sudhir Kowlessur, Peter Kraft, Jasmina Kravic, Malene M Kristensen, S Krithika, Ashish Kumar, Jesus Kumate, Johanna Kuusisto, Soo-Heon Kwak, Markku Laakso, Vasiliki Lagou, Timo A Lakka, Claudia Langenberg, Cordelia Langford, Robert Lawrence, Karin Leander, Jen-Mai Lee, Nanette R Lee, Man Li, Xinzhong Li, Yun Li, Junbin Liang, Samuel Liju, Wei Yen Lim, Lars Lind, Cecilia M Lindgren, Eero Lindholm, Ching-Ti Liu, Jian Jun Liu, Stéphane Lobbens, Jirong Long, Ruth J F Loos, Wei Lu, Jian'an Luan, Valeriya Lyssenko, Ronald C W Ma, Shiro Maeda, Reedik Mägi, Satu Männistö, David R Matthews, James B Meigs, Olle Melander, Andres Metspalu, Julia Meyer, Ghazala Mirza, Evelin Mihailov, Susanne Moebus, Viswanathan Mohan, Karen L Mohlke, Andrew D Morris, Thomas W Mühleisen, Martina Müller-Nurasyid, Bill Musk, Jiro Nakamura, Eitaro Nakashima, Pau Navarro, Peng-Keat Ng, Alexandra C Nica, Peter M Nilsson, Inger Njølstad, Markus M Nöthen, Keizo Ohnaka, Twee Hee Ong, Katharine R Owen, Colin N A Palmer, James S Pankow, Kyong Soo Park, Melissa Parkin, Sonali Pechlivanis, Nancy L Pedersen, Leena Peltonen, John R B Perry, Annette Peters, Janani M Pinidiyapathirage, Carl G P Platou, Simon Potter, Jackie F Price, Lu Qi, Venkatesan Radha, Loukianos Rallidis, Asif Rasheed, Wolfgang Rathmann, Rainer Rauramaa, Soumya Raychaudhuri, N William Rayner, Simon D Rees, Emil Rehnberg, Samuli Ripatti, and ... Robertson. Genome-wide trans-ancestry meta-analysis provides insight into the genetic architecture of type 2 diabetes susceptibility. *Nat. Genet.*, 46(3):234–244, February 2014.

[64] Christian Fuchsberger, Jason Flannick, Tanya M Teslovich, Anubha Mahajan, Vineeta Agarwala, Kyle J Gaulton, Clement Ma, Pierre Fontanillas, Loukas Moutsianas, Davis J McCarthy, Manuel A Rivas, John R B Perry, Xueling Sim, Thomas W Blackwell, Neil R Robertson, N William Rayner, Pablo Cingolani, Adam E Locke, Juan Fernandez Tajes, Heather M Highland, Josée Dupuis, Peter S Chines, Cecilia M Lindgren, Christopher Hartl, Anne U Jackson, Han Chen, Jeroen R Huyghe, Martijn van de Bunt, Richard D Pearson, Ashish Kumar, Martina Müller-Nurasyid, Niels Grarup, Heather M Stringham, Eric R Gamazon, Jaehoon Lee, Yuhui Chen, Robert A Scott, Jennifer E Below, Peng Chen, Jinyan Huang, Min Jin Go, Michael L Stitzel, Dorota Pasko, Stephen C J Parker, Tibor V Varga, Todd

Green, Nicola L Beer, Aaron G Day-Williams, Teresa Ferreira, Tasha Fingerlin, Momoko Horikoshi, Cheng Hu, Iksoo Huh, Mohammad Kamran Ikram, Bong-Jo Kim, Yongkang Kim, Young-Jin Kim, Min-Seok Kwon, Juyoung Lee, Selyeong Lee, Keng-Han Lin, Taylor J Maxwell, Yoshihiko Nagai, Xu Wang, Ryan P Welch, Joon Yoon, Weihua Zhang, Nir Barzilai, Benjamin F Voight, Bok-Ghee Han, Christopher P Jenkinson, Teemu Kuulasmaa, Johanna Kuusisto, Alisa Manning, Maggie C Y Ng, Nicholette D Palmer, Beverley Balkau, Alena Stančáková, Hanna E Abboud, Heiner Boeing, Vilmantas Giedraitis, Dorairaj Prabhakaran, Omri Gottesman, James Scott, Jason Carey, Phoenix Kwan, George Grant, Joshua D Smith, Benjamin M Neale, Shaun Purcell, Adam S Butterworth, Joanna M M Howson, Heung Man Lee, Yingchang Lu, Soo-Heon Kwak, Wei Zhao, John Danesh, Vincent K L Lam, Kyong Soo Park, Danish Saleheen, Wing Yee So, Claudia H T Tam, Uzma Afzal, David Aguilar, Rector Arya, Tin Aung, Edmund Chan, Carmen Navarro, Ching-Yu Cheng, Domenico Palli, Adolfo Correa, Joanne E Curran, Denis Rybin, Vidya S Farook, Sharon P Fowler, Barry I Freedman, Michael Griswold, Daniel Esten Hale, Pamela J Hicks, Chiea-Chuen Khor, Satish Kumar, Benjamin Lehne, Dorothée Thuillier, Wei Yen Lim, Jianjun Liu, Yvonne T van der Schouw, Marie Loh, Solomon K Musani, Sobha Puppala, William R Scott, Loic Yengo, Sian-Tsung Tan, Herman A Taylor, Farook Thameem, Gregory Wilson, Tien Yin Wong, Pål Rasmus Njølstad, Jonathan C Levy, Massimo Mangino, Lori L Bonnycastle, Thomas Schwarzmayr, João Fadista, Gabriela L Surdulescu, Christian Herder, Christopher J Groves, Thomas Wieland, Jette Bork-Jensen, Ivan Brandslund, Cramer Christensen, Heikki A Koistinen, Alex S F Doney, Leena Kinnunen, Tonu Esko, Andrew J Farmer, Liisa Hakaste, Dylan Hodgkiss, Jasmina Kravic, Valeriya Lyssenko, Mette Hollensted, Marit E Jørgensen, Torben Jørgensen, Claes Ladenvall, Johanne Marie Justesen, Annemari Käräjämäki, Jennifer Kriebel, Wolfgang Rathmann, Lars Lannfelt, Torsten Lauritzen, Narisu Narisu, Allan Linneberg, Olle Melander, Lili Milani, Matt Neville, Marju Orho-Melander, Lu Qi, Qibin Qi, Michael Roden, Olov Rolandsson, Amy Swift, Anders H Rosengren, Kathleen Stirrups, Andrew R Wood, Evelin Mihailov, Christine Blancher, Mauricio O Carneiro, Jared Maguire, Ryan Poplin, Khalid Shakir, Timothy Fennell, Mark DePristo, Martin Hrabé de Angelis, Panos Deloukas, Anette P Gjesing, Goo Jun, Peter Nilsson, Jacquelyn Murphy, Robert Onofrio, Barbara Thorand, Torben Hansen, Christa Meisinger,

Frank B Hu, Bo Isomaa, Fredrik Karpe, Liming Liang, Annette Peters, Cornelia Huth, Stephen P O'Rahilly, Colin N A Palmer, Oluf Pedersen, Rainer Rauramaa, Jaakko Tuomilehto, Veikko Salomaa, Richard M Watanabe, Ann-Christine Syvänen, Richard N Bergman, Dwaipayan Bharadwaj, Erwin P Bottinger, Yoon Shin Cho, Giriraj R Chandak, Juliana C N Chan, Kee Seng Chia, Mark J Daly, Shah B Ebrahim, Claudia Langenberg, Paul Elliott, Kathleen A Jablonski, Donna M Lehman, Weiping Jia, Ronald C W Ma, and Polli... The genetic architecture of type 2 diabetes. *Nature*, 536(7614):41–47, August 2016.

[65] ENCODE Project Consortium. An integrated encyclopedia of DNA elements in the human genome. *Nature*, 489(7414):57–74, September 2012.

[66] Bradley E Bernstein, John A Stamatoyannopoulos, Joseph F Costello, Bing Ren, Aleksandar Milosavljevic, Alexander Meissner, Manolis Kellis, Marco A Marra, Arthur L Beaudet, Joseph R Ecker, Peggy J Farnham, Martin Hirst, Eric S Lander, Tarjei S Mikkelsen, and James A Thomson. The NIH Roadmap Epigenomics Mapping Consortium. *Nat. Biotechnol.*, 28(10):1045–1048, October 2010.

[67] David Adams, Lucia Altucci, Stylianos E Antonarakis, Juan Ballesteros, Stephan Beck, Adrian Bird, Christoph Bock, Bernhard Boehm, Elias Campo, Andrea Caricasole, Fredrik Dahl, Emmanouil T Dermitzakis, Tariq Enver, Manel Esteller, Xavier Estivill, Anne Ferguson-Smith, Jude Fitzgibbon, Paul Flicek, Claudia Giehl, Thomas Graf, Frank Grosveld, Roderic Guigó, Ivo Gut, Kristian Helin, Jonas Jarvius, Ralf Küppers, Hans Lehrach, Thomas Lengauer, Åke Lernmark, David Leslie, Markus Loeffler, Elizabeth Macintyre, Antonello Mai, Joost H A Martens, Saverio Minucci, Willem H Ouwehand, Pier Giuseppe Pelicci, Hèléne Pendeville, Bo Porse, Vardhman Rakyan, Wolf Reik, Martin Schrappe, Dirk Schübeler, Martin Seifert, Reiner Siebert, David Simmons, Nicole Soranzo, Salvatore Spicuglia, Michael Stratton, Hendrik G Stunnenberg, Amos Tanay, David Torrents, Alfonso Valencia, Edo Vellenga, Martin Vingron, Jörn Walter, and Spike Willcocks. BLUEPRINT to decode the epigenetic signature written in blood. *Nat. Biotechnol.*, 30(3):224–226, March 2012.

[68] Artem Barski, Suresh Cuddapah, Kairong Cui, Tae-Young Roh, Dustin E Schones, Zhibin Wang, Gang Wei, Iouri Chepelev, and Keji Zhao. High-resolution profiling of histone methylations in the human genome. *Cell*, 129(4):823–837, May 2007.

[69] Lingyun Song and Gregory E Crawford. DNase-seq: a high-resolution technique for mapping active gene regulatory elements across the genome from mammalian cells. *Cold Spring Harb Protoc*, 2010(2):pdb.prot5384–pdb.prot5384, February 2010.

[70] M Frommer, L E McDonald, D S Millar, C M Collis, F Watt, G W Grigg, P L Molloy, and C L Paul. A genomic sequencing protocol that yields a positive display of 5-methylcytosine residues in individual DNA strands. *PNAS*, 89(5):1827–1831, March 1992.

[71] Yuanyuan Li and Trygve O Tollefsbol. DNA Methylation Detection: Bisulfite Genomic Sequencing Analysis. In *Epigenetics Protocols*, pages 11–21. Humana Press, Totowa, NJ, August 2011.

[72] Zhong Wang, Mark Gerstein, and Michael Snyder. RNA-Seq: a revolutionary tool for transcriptomics. *Nat. Rev. Genet.*, 10(1):57–63, January 2009.

[73] Ugrappa Nagalakshmi, Zhong Wang, Karl Waern, Chong Shou, Debasish Raha, Mark Gerstein, and Michael Snyder. The Transcriptional Landscape of the Yeast Genome Defined by RNA Sequencing. *Science*, 320(5881):1344–1349, June 2008.

[74] Lucas D Ward and Manolis Kellis. Interpreting noncoding genetic variation in complex traits and human disease. *Nat. Biotechnol.*, 30(11):1095–1106, November 2012.

[75] Dirk S Paul, Nicole Soranzo, and Stephan Beck. Functional interpretation of non-coding sequence variation: Concepts and challenges. *Bioessays*, 36(2):191–199, February 2014.

[76] Jason Ernst, Pouya Kheradpour, Tarjei S Mikkelsen, Noam Shoresh, Lucas D Ward, Charles B Epstein, Xiaolan Zhang, Li Wang, Robbyn Issner, Michael Coyne, Manching Ku, Timothy Durham, Manolis Kellis, and Bradley E Bernstein. Mapping and analysis of chromatin state dynamics in nine human cell types. *Nature*, 473(7345):43–49, May 2011.

[77] Matthew T Maurano, Richard Humbert, Eric Rynes, Robert E Thurman, Eric Haugen, Hao Wang, Alex P Reynolds, Richard Sandstrom, Hongzhu Qu, Jennifer Brody, Anthony Shafer, Fidencio Neri, Kristen Lee, Tanya Kutyavin, Sandra Stehling-Sun, Audra K Johnson, Theresa K Canfield, Erika Giste, Morgan Diegel, Daniel Bates, R Scott Hansen, Shane Neph, Peter J Sabo, Shelly Heimfeld, Antony Raubitschek, Steven Ziegler, Chris Cotsapas, Nona

Sotoodehnia, Ian Glass, Shamil R Sunyaev, Rajinder Kaul, and John A Stamatoyannopoulos. Systematic Localization of Common Disease-Associated Variation in Regulatory DNA. *Science*, 337(6099):1190–1195, September 2012.

[78] Robin Andersson, Claudia Gebhard, Irene Miguel-Escalada, Ilka Hoof, Jette Bornholdt, Mette Boyd, Yun Chen, Xiaobei Zhao, Christian Schmidl, Takahiro Suzuki, Evgenia Ntini, Erik Arner, Eivind Valen, Kang Li, Lucia Schwarzfischer, Dagmar Glatz, Johanna Raithel, Berit Lilje, Nicolas Rapin, Frederik Otzen Bagger, Mette Jørgensen, Peter Refsing Andersen, Nicolas Bertin, Owen Rackham, A Maxwell Burroughs, J Kenneth Baillie, Yuri Ishizu, Yuri Shimizu, Erina Furuhata, Shiori Maeda, Yutaka Negishi, Christopher J Mungall, Terrence F Meehan, Timo Lassmann, Masayoshi Itoh, Hideya Kawaji, Naoto Kondo, Jun Kawai, Andreas Lennartsson, Carsten O Daub, Peter Heutink, David A Hume, Torben Heick Jensen, Harukazu Suzuki, Yoshihide Hayashizaki, Ferenc Müller, The FANTOM Consortium, Alistair R R Forrest, Piero Carninci, Michael Rehli, and Albin Sandelin. An atlas of active enhancers across human cell types and tissues. *Nature*, 507(7493):455–461, March 2014.

[79] Batool Akhtar-Zaidi, Richard Cowper-Sal lari, Olivia Corradin, Alina Saiakhova, Cynthia F Bartels, Dheepa Balasubramanian, Lois Myeroff, James Lutterbaugh, Awad Jarrar, Matthew F Kalady, Joseph Willis, Jason H Moore, Paul J Tesar, Thomas Laframboise, Sanford Markowitz, Mathieu Lupien, and Peter C Scacheri. Epigenomic Enhancer Profiling Defines a Signature of Colon Cancer. *Science*, 336(6082):736–739, May 2012.

[80] Olivia Corradin, Alina Saiakhova, Batool Akhtar-Zaidi, Lois Myeroff, Joseph Willis, Richard Cowper-Sal lari, Mathieu Lupien, Sanford Markowitz, and Peter C Scacheri. Combinatorial effects of multiple enhancer variants in linkage disequilibrium dictate levels of gene expression to confer susceptibility to common traits. *Genome Res.*, 24(1):1–13, January 2014.

[81] Dennis J Hazelett, Suhn Kyong Rhie, Malaina Gaddis, Chunli Yan, Daniel L Lakeland, Simon G Coetzee, Ellipse GAME-ON consortium, Practical consortium, Brian E Henderson, Houtan Noushmehr, Wendy Cozen, Zsofia Kote-Jarai, Rosalind A Eeles, Douglas F Easton, Christopher A Haiman, Wange Lu, Peggy J Farnham, and Gerhard A Coetzee. Comprehensive Functional Annotation of 77 Prostate Cancer Risk Loci. *PLoS Genet.*, 10(1):e1004102, January 2014.

[82] Roadmap Epigenomics Consortium, Anshul Kundaje, Wouter Meuleman, Jason Ernst, Misha Bilenky, Angela Yen, Alireza Heravi-Moussavi, Pouya Kheradpour, Zhizhuo Zhang, Jianrong Wang, Michael J Ziller, Viren Amin, John W Whitaker, Matthew D Schultz, Lucas D Ward, Abhishek Sarkar, Gerald Quon, Richard S Sandstrom, Matthew L Eaton, Yi-Chieh Wu, Andreas R Pfenning, Xinchen Wang, Melina Claussnitzer, Yaping Liu, Cristian Coarfa, R Alan Harris, Noam Shoresh, Charles B Epstein, Elizabeta Gjoneska, Danny Leung, Wei Xie, R David Hawkins, Ryan Lister, Chibo Hong, Philippe Gascard, Andrew J Mungall, Richard Moore, Eric Chuah, Angela Tam, Theresa K Canfield, R Scott Hansen, Rajinder Kaul, Peter J Sabo, Mukul S Bansal, Annaick Carles, Jesse R Dixon, Kai-How Farh, Soheil Feizi, Rosa Karlic, Ah-Ram Kim, Ashwinikumar Kulkarni, Daofeng Li, Rebecca Lowdon, GiNell Elliott, Tim R Mercer, Shane J Neph, Vitor Onuchic, Paz Polak, Nisha Rajagopal, Pradipta Ray, Richard C Sallari, Kyle T Siebenthall, Nicholas A Sinnott-Armstrong, Michael Stevens, Robert E Thurman, Jie Wu, Bo Zhang, Xin Zhou, Arthur E Beaudet, Laurie A Boyer, Philip L De Jager, Peggy J Farnham, Susan J Fisher, David Haussler, Steven J M Jones, Wei Li, Marco A Marra, Michael T McManus, Shamil Sunyaev, James A Thomson, Thea D Tlsty, Li-Huei Tsai, Wei Wang, Robert A Waterland, Michael Q Zhang, Lisa H Chadwick, Bradley E Bernstein, Joseph F Costello, Joseph R Ecker, Martin Hirst, Alexander Meissner, Aleksandar Milosavljevic, Bing Ren, John A Stamatoyannopoulos, Ting Wang, and Manolis Kellis. Integrative analysis of 111 reference human epigenomes. *Nature*, 518(7539):317–330, February 2015.

[83] Kyle J Gaulton, Teresa Ferreira, Yeji Lee, Anne Raimondo, Reedik Mägi, Michael E Reschen, Anubha Mahajan, Adam Locke, N William Rayner, Neil Robertson, Robert A Scott, Inga Prokopenko, Laura J Scott, Todd Green, Thomas Sparsø, Dorothée Thuillier, Loic Yengo, Harald Grallert, Simone Wahl, Mattias Frånberg, Rona J Strawbridge, Hans Kestler, Himanshu Chheda, Lewin Eisele, Stefan Gustafsson, Valgerdur Steinthorsdottir, Gudmar Thorleifsson, Lu Qi, Lennart C Karssen, Elisabeth M van Leeuwen, Sara M Willems, Man Li, Han Chen, Christian Fuchsberger, Phoenix Kwan, Clement Ma, Michael Linderman, Yingchang Lu, Soren K Thomsen, Jana K Rundle, Nicola L Beer, Martijn van de Bunt, Anil Chalisey, Hyun Min Kang, Benjamin F Voight, Goncalo R Abecasis, Peter Almgren, Damiano Baldassarre, Beverley Balkau, Rafn Benediktsson, Matthias Blüher, Heiner

Boeing, Lori L Bonnycastle, Erwin P Bottinger, Noël P Burtt, Jason Carey, Guillaume Charpentier, Peter S Chines, Marilyn C Cornelis, David J Couper, Andrew T Crenshaw, Rob M van Dam, Alex S F Doney, Mozhgan Dorkhan, Sarah Edkins, Johan G Eriksson, Tonu Esko, Elodie Eury, João Fadista, Jason Flannick, Pierre Fontanillas, Caroline Fox, Paul W Franks, Karl Gertow, Christian Gieger, Bruna Gigante, Omri Gottesman, George B Grant, Niels Grarup, Christopher J Groves, Maija Hassinen, Christian T Have, Christian Herder, Oddgeir L Holmen, Astradur B Hreidarsson, Steve E Humphries, David J Hunter, Anne U Jackson, Anna Jonsson, Marit E Jørgensen, Torben Jørgensen, Wen Hong L Kao, Nicola D Kerrison, Leena Kinnunen, Norman Klopp, Augustine Kong, Peter Kovacs, Peter Kraft, Jasmina Kravic, Cordelia Langford, Karin Leander, Liming Liang, Peter Lichtner, Cecilia M Lindgren, Eero Lindholm, Allan Linneberg, Ching-Ti Liu, Stéphane Lobbens, Jian'an Luan, Valeriya Lyssenko, Satu Männistö, Olga McLeod, Julia Meyer, Evelin Mihailov, Ghazala Mirza, Thomas W Mühleisen, Martina Müller-Nurasyid, Carmen Navarro, Markus M Nöthen, Nikolay N Oskolkov, Katharine R Owen, Domenico Palli, Sonali Pechlivanis, Leena Peltonen, John R B Perry, Carl G P Platou, Michael Roden, Douglas Ruderfer, Denis Rybin, Yvonne T van der Schouw, Bengt Sennblad, Gunnar Sigurðsson, Alena Stančáková, Gerald Steinbach, Petter Storm, Konstantin Strauch, Heather M Stringham, Qi Sun, Barbara Thorand, Emmi Tikkanen, Anke Tönjes, Joseph Trakalo, Elena Tremoli, Tiinamaija Tuomi, Roman Wennauer, Steven Wiltshire, Andrew R Wood, Eleftheria Zeggini, Ian Dunham, Ewan Birney, Lorenzo Pasquali, Jorge Ferrer, Ruth J F Loos, Josée Dupuis, Jose C Florez, Eric Boerwinkle, James S Pankow, Cornelia van Duijn, Eric Sijbrands, James B Meigs, Frank B Hu, Unnur Thorsteinsdottir, Kari Stefansson, Timo A Lakka, Rainer Rauramaa, Michael Stumvoll, Nancy L Pedersen, Lars Lind, Sirkka M Keinanen-Kiukaanniemi, Eeva Korpi-Hyövälti, Timo E Saaristo, Juha Saltevo, Johanna Kuusisto, Markku Laakso, Andres Metspalu, Raimund Erbel, Karl-Heinz Jöcke, Susanne Moebus, Samuli Ripatti, Veikko Salomaa, Erik Ingelsson, Bernhard O Boehm, Richard N Bergman, Francis S Collins, Karen L Mohlke, Heikki Koistinen, Jaakko Tuomilehto, Kristian Hveem, Inger Njølstad, Panagiotis Deloukas, Peter J Donnelly, Timothy M Frayling, Andrew T Hattersley, Ulf de Faire, Anders Hamsten, Thomas Illig, Annette Peters, Stephane Cauchi, Rob Sladek, Philippe Froguel, Torben Hansen, Oluf Pedersen, Andrew D Morris, Collin N A Palmer, Sekar Kathiresan, Olle

Melander, Peter M Nilsson, Leif C Groop, Ines Barroso, Claudia Langenberg, Nicholas J Wareham, Christopher A O'Callaghan, Anna L Gloyn, David Altshuler, Michael Boehnke, Tanya M Teslovich, Mark I McCarthy, Andrew P Morris, and DIAbetes Genetics Replication And Meta-analysis (DIAGRAM) Consortium. Genetic fine mapping and genomic annotation defines causal mechanisms at type 2 diabetes susceptibility loci. *Nat. Genet.*, 47(12):1415–1425, December 2015.

[84] Kyle J Gaulton, Takao Nammo, Lorenzo Pasquali, Jeremy M Simon, Paul G Giresi, Marie P Fogarty, Tami M Panhuis, Piotr Mieczkowski, Antonio Secchi, Domenico Bosco, Thierry Berney, Eduard Montanya, Karen L Mohlke, Jason D Lieb, and Jorge Ferrer. A map of open chromatin in human pancreatic islets. *Nat. Genet.*, 42(3):255–259, March 2010.

[85] Lorenzo Pasquali, Kyle J Gaulton, Santiago A Rodríguez-Seguí, Loris Mularoni, Irene Miguel-Escalada, Ildem Akerman, Juan J Tena, Ignasi Morán, Carlos Gómez-Marín, Martijn van de Bunt, Joan Ponsa-Cobas, Natalia Castro, Takao Nammo, Inês Cebola, Javier García-Hurtado, Miguel Angel Maestro, François Pattou, Lorenzo Piemonti, Thierry Berney, Anna L Gloyn, Philippe Ravassard, José Luis Gómez-Skarmeta, Ferenc Müller, Mark I McCarthy, and Jorge Ferrer. Pancreatic islet enhancer clusters enriched in type 2 diabetes risk-associated variants. *Nat. Genet.*, 46(2):136–143, February 2014.

[86] Stephen C J Parker, Michael L Stitzel, D Leland Taylor, Jose Miguel Orozco, Michael R Erdos, Jennifer A Akiyama, Kelly Lammerts van Bueren, Peter S Chines, Narisu Narisu, NISC Comparative Sequencing Program, Brian L Black, Axel Visel, Len A Pennacchio, Francis S Collins, National Institutes of Health Intramural Sequencing Center Comparative Sequencing Program Authors, and NISC Comparative Sequencing Program Authors. Chromatin stretch enhancer states drive cell-specific gene regulation and harbor human disease risk variants. *Proc. Natl. Acad. Sci. U.S.A.*, 110(44):17921–17926, October 2013.

[87] Adam E Locke, Bratati Kahali, Sonja I Berndt, Anne E Justice, Tune H Pers, Felix R Day, Corey Powell, Sailaja Vedantam, Martin L Buchkovich, Jian Yang, Damien C Croteau-Chonka, Tonu Esko, Tove Fall, Teresa Ferreira, Stefan Gustafsson, Zoltán Kutalik, Jian'an Luan, Reedik Mägi, Joshua C Randall, Thomas W Winkler, Andrew R Wood, Tsegaselassie Workalemahu, Jessica D Faul, Jennifer A Smith, Jing Hua Zhao, Wei Zhao, Jin Chen,

Rudolf Fehrmann, Åsa K Hedman, Juha Karjalainen, Ellen M Schmidt, Devin Absher, Najaf Amin, Denise Anderson, Marian Beekman, Jennifer L Bolton, Jennifer L Bragg-Gresham, Steven Buyske, Ayse Demirkan, Guohong Deng, Georg B Ehret, Bjarke Feenstra, Mary F Feitosa, Krista Fischer, Anuj Goel, Jian Gong, Anne U Jackson, Stavroula Kanoni, Marcus E Kleber, Kati Kristiansson, Unhee Lim, Vaneet Lotay, Massimo Mangino, Irene Mateo Leach, Carolina Medina-Gomez, Sarah E Medland, Michael A Nalls, Cameron D Palmer, Dorota Pasko, Sonali Pechlivanis, Marjolein J Peters, Inga Prokopenko, Dmitry Shungin, Alena Stančáková, Rona J Strawbridge, Yun Ju Sung, Toshiko Tanaka, Alexander Teumer, Stella Trompet, Sander W van der Laan, Jessica van Setten, Jana V van Vliet-Ostaptchouk, Zhaoming Wang, Loic Yengo, Weihua Zhang, Aaron Isaacs, Eva Albrecht, Johan Ärnlöv, Gillian M Arscott, Antony P Attwood, Stefania Bandinelli, Amy Barrett, Isabelita N Bas, Claire Bellis, Amanda J Bennett, Christian Berne, Roza Blagieva, Matthias Blüher, Stefan Böhringer, Lori L Bonnycastle, Yvonne Böttcher, Heather A Boyd, Marcel Bruinenberg, Ida H Caspersen, Yii-Der Ida Chen, Robert Clarke, E Warwick Daw, Anton J M de Craen, Graciela Delgado, Maria Dimitriou, Alex S F Doney, Niina Eklund, Karol Estrada, Elodie Eury, Lasse Folkersen, Ross M Fraser, Melissa E Garcia, Frank Geller, Vilmantas Giedraitis, Bruna Gigante, Alan S Go, Alain Golay, Alison H Goodall, Scott D Gordon, Mathias Gorski, Hans-Jörgen Grabe, Harald Grallert, Tanja B Grammer, Jürgen Gräßler, Henrik Grönberg, Christopher J Groves, Gaëlle Gusto, Jeffrey Haessler, Per Hall, Toomas Haller, Göran Hallmans, Catharina A Hartman, Maija Hassinen, Caroline Hayward, Nancy L Heard-Costa, Quinta Helmer, Christian Hengstenberg, Oddgeir Holmen, Jouke Jan Hottenga, Alan L James, Janina M Jeff, Asa Johansson, Jennifer Jolley, Thorhildur Juliusdottir, Leena Kinnunen, Wolfgang Koenig, Markku Koskenvuo, Wolfgang Kratzer, Jaana Laitinen, Claudia Lamina, Karin Leander, Nanette R Lee, Peter Lichtner, Lars Lind, Jaana Lindström, Ken Sin Lo, Stéphane Lobbens, Roberto Lorbeer, Yingchang Lu, François Mach, Patrik K E Magnusson, Anubha Mahajan, Wendy L McArdle, Stela McLachlan, Cristina Menni, Sigrun Merger, Evelin Mihailov, Lili Milani, Alireza Moayyeri, Keri L Monda, Mario A Morken, Antonella Mulas, Gabriele Müller, Martina Müller-Nurasyid, Arthur W Musk, Ramaiah Nagaraja, Markus M Nöthen, Ilja M Nolte, Stefan Pilz, Nigel W Rayner, Frida Renstrom, Rainer Rettig, Janina S Ried, Stephan Ripke, Neil R Robertson, Lynda M Rose, Serena

Sanna, Hubert Scharnagl, Salome Scholtens, Fredrick R Schumacher, William R Scott, Thomas Seufferlein, Jianxin Shi, Albert Vernon Smith, Joanna Smolonska, Alice V Stanton, Valgerdur Steinthorsdottir, Kathleen Stirrups, Heather M Stringham, Johan Sundström, Morris A Swertz, Amy J Swift, Ann-Christine Syvänen, Sian-Tsung Tan, Bamidele O Tayo, Barbara Thorand, Gudmar Thorleifsson, Jonathan P Tyrer, Hae-Won Uh, Liesbeth Vandenput, Frank C Verhulst, Sita H Vermeulen, Niek Verweij, Judith M Vonk, Lindsay L Waite, Helen R Warren, Dawn Waterworth, Michael N Weedon, Lynne R Wilkens, Christina Willenborg, Tom Wilsgaard, Mary K Wojczynski, Andrew Wong, Alan F Wright, Qunyuan Zhang, Eoin P Brennan, Murim Choi, Zari Dastani, Alexander W Drong, and Per an... Eriksson. Genetic studies of body mass index yield new insights for obesity biology. *Nature*, 518(7538):197–206, February 2015.

[88] Olivier Harismendy, Dimple Notani, Xiaoyuan Song, Nazli G Rahim, Bogdan Tanasa, Nathaniel Heintzman, Bing Ren, Xiang-Dong Fu, Eric J Topol, Michael G Rosenfeld, and Kelly A Frazer. 9p21 DNA variants associated with coronary artery disease impair interferon-γ signalling response. *Nature*, 470(7333):264–268, February 2011.

[89] Melina Claussnitzer, Simon N Dankel, Kyoung-Han Kim, Gerald Quon, Wouter Meuleman, Christine Haugen, Viktoria Glunk, Isabel S Sousa, Jacqueline L Beaudry, Vijitha Puviindran, Nezar A Abdennur, Jannel Liu, Per-Arne Svensson, Yi-Hsiang Hsu, Daniel J Drucker, Gunnar Mellgren, Chi-Chung Hui, Hans Hauner, and Manolis Kellis. FTOObesity Variant Circuitry and Adipocyte Browning in Humans. *N. Engl. J. Med.*, 373(10):895–907, September 2015.

[90] Scott Smemo, Juan J Tena, Kyoung-Han Kim, Eric R Gamazon, Noboru J Sakabe, Carlos Gómez-Marín, Ivy Aneas, Flavia L Credidio, Débora R Sobreira, Nora F Wasserman, Ju Hee Lee, Vijitha Puviindran, Davis Tam, Michael Shen, Joe Eun Son, Niki Alizadeh Vakili, Hoon-Ki Sung, Silvia Naranjo, Rafael D Acemel, Miguel Manzanares, Andras Nagy, Nancy J Cox, Chi-Chung Hui, José Luis Gómez-Skarmeta, and Marcelo A Nóbrega. Obesity-associated variants within FTO form long-range functional connections with IRX3. *Nature*, 507(7492):371–375, March 2014.

[91] Fumitaka Inoue and Nadav Ahituv. Decoding enhancers using massively parallel reporter assays. *Genomics*, 106(3):159–164, September 2015.

[92] A P Morris. *Fine Mapping Type 2 Diabetes Susceptibility Loci*, volume 23 of *Type 2 Diabetes and Related Traits*. Karger Publishers, Basel, September 2014.

[93] Benjamin F Voight, Hyun Min Kang, Jun Ding, Cameron D Palmer, Carlo Sidore, Peter S Chines, Noël P Burtt, Christian Fuchsberger, Yanming Li, Jeanette Erdmann, Timothy M Frayling, Iris M Heid, Anne U Jackson, Toby Johnson, Tuomas O Kilpeläinen, Cecilia M Lindgren, Andrew P Morris, Inga Prokopenko, Joshua C Randall, Richa Saxena, Nicole Soranzo, Elizabeth K Speliotes, Tanya M Teslovich, Eleanor Wheeler, Jared Maguire, Melissa Parkin, Simon Potter, N William Rayner, Neil Robertson, Kathleen Stirrups, Wendy Winckler, Serena Sanna, Antonella Mulas, Ramaiah Nagaraja, Francesco Cucca, Ines Barroso, Panos Deloukas, Ruth J F Loos, Sekar Kathiresan, Patricia B Munroe, Christopher Newton-Cheh, Arne Pfeufer, Nilesh J Samani, Heribert Schunkert, Joel N Hirschhorn, David Altshuler, Mark I McCarthy, Goncalo R Abecasis, and Michael Boehnke. The Metabochip, a Custom Genotyping Array for Genetic Studies of Metabolic, Cardiovascular, and Anthropometric Traits. *PLoS Genet.*, 8(8):e1002793, August 2012.

[94] Marie P Fogarty, Tami M Panhuis, Swarooparani Vadlamudi, Martin L Buchkovich, and Karen L Mohlke. Allele-Specific Transcriptional Activity at Type 2 Diabetes–Associated Single Nucleotide Polymorphisms in Regions of Pancreatic Islet Open Chromatin at the JAZF1 Locus. *Diabetes*, 62(5):1756–1762, May 2013.

[95] En Li and Yi Zhang. DNA methylation in mammals. *Cold Spring Harb Perspect Biol*, 6(5):a019133–a019133, May 2014.

[96] Eilis Hannon, Helen Spiers, Joana Viana, Ruth Pidsley, Joe Burrage, Therese M Murphy, Claire Troakes, Gustavo Turecki, Michael C O'Donovan, Leonard C Schalkwyk, Nicholas J Bray, and Jonathan Mill. Methylation QTLs in the developing brain and their enrichment in schizophrenia risk loci. *Nat. Neurosci.*, 19(1):48–54, January 2016.

[97] Joseph L McClay, Andrey A Shabalin, Mikhail G Dozmorov, Daniel E Adkins, Gaurav Kumar, Srilaxmi Nerella, Shaunna L Clark, Sarah E Bergen, Swedish Schizophrenia Consor-

tium, Christina M Hultman, Patrik K E Magnusson, Patrick F Sullivan, Karolina A Aberg, and Edwin J C G van den Oord. High density methylation QTL analysis in human blood via next-generation sequencing of the methylated genomic DNA fraction. *Genome Biol.*, 16(1):291, 2015.

[98] Jianxin Shi, Crystal N Marconett, Jubao Duan, Paula L Hyland, Peng Li, Zhaoming Wang, William Wheeler, Beiyun Zhou, Mihaela Campan, Diane S Lee, Jing Huang, Weiyin Zhou, Tim Triche, Laufey Amundadottir, Andrew Warner, Amy Hutchinson, Po-Han Chen, Brian S I Chung, Angela C Pesatori, Dario Consonni, Pier Alberto Bertazzi, Andrew W Bergen, Mathew Freedman, Kimberly D Siegmund, Benjamin P Berman, Zea Borok, Nilanjan Chatterjee, Margaret A Tucker, Neil E Caporaso, Stephen J Chanock, Ite A Laird-Offringa, and Maria Teresa Landi. Characterizing the genetic basis of methylome diversity in histologically normal human lung tissue. *Nat Commun*, 5:3365, 2014.

[99] Anders H Olsson, Petr Volkov, Karl Bacos, Tasnim Dayeh, Elin Hall, Emma A Nilsson, Claes Ladenvall, Tina Rönn, and Charlotte Ling. Genome-wide associations between genetic and epigenetic variation influence mRNA expression and insulin secretion in human pancreatic islets. *PLoS Genet.*, 10(11):e1004735, November 2014.

[100] Elin Grundberg, Eshwar Meduri, Johanna K Sandling, Åsa K Hedman, Sarah Keildson, Alfonso Buil, Stephan Busche, Wei Yuan, James Nisbet, Magdalena Sekowska, Alicja Wilk, Amy Barrett, Kerrin S Small, Bing Ge, Maxime Caron, So-Youn Shin, Multiple Tissue Human Expression Resource Consortium, Mark Lathrop, Emmanouil T Dermitzakis, Mark I McCarthy, Timothy D Spector, Jordana T Bell, and Panos Deloukas. Global analysis of DNA methylation variation in adipose tissue from twins reveals links to disease-associated variants in distal regulatory elements. *Am. J. Hum. Genet.*, 93(5):876–890, November 2013.

[101] Maria Gutierrez-Arcelus, Tuuli Lappalainen, Stephen B Montgomery, Alfonso Buil, Halit Ongen, Alisa Yurovsky, Julien Bryois, Thomas Giger, Luciana Romano, Alexandra Planchon, Emilie Falconnet, Deborah Bielser, Maryline Gagnebin, Ismael Padioleau, Christelle Borel, Audrey Letourneau, Periklis Makrythanasis, Michel Guipponi, Corinne Gehrig, Stylianos E Antonarakis, and Emmanouil T Dermitzakis. Passive and active DNA methylation and the interplay with genetic variation in gene regulation. *Elife*, 2(0):e00523–e00523, 2013.

[102] Alexander W Drong, George Nicholson, Åsa K Hedman, Eshwar Meduri, Elin Grundberg, Kerrin S Small, So-Youn Shin, Jordana T Bell, Fredrik Karpe, Nicole Soranzo, Tim D Spector, Mark I McCarthy, Panos Deloukas, Mattias Rantalainen, Cecilia M Lindgren, and MolPAGE Consortia. The presence of methylation quantitative trait loci indicates a direct genetic influence on the level of DNA methylation in adipose tissue. *PLoS ONE*, 8(2):e55923, 2013.

[103] Jordana T Bell, Athma A Pai, Joseph K Pickrell, Daniel J Gaffney, Roger Pique-Regi, Jacob F Degner, Yoav Gilad, and Jonathan K Pritchard. DNA methylation patterns associate with genetic and gene expression variation in HapMap cell lines. *Genome Biol.*, 12(1):R10, 2011.

[104] Dandan Zhang, Lijun Cheng, Judith A Badner, Chao Chen, Qi Chen, Wei Luo, David W Craig, Margot Redman, Elliot S Gershon, and Chunyu Liu. Genetic control of individual differences in gene-specific methylation in human brain. *Am. J. Hum. Genet.*, 86(3):411–419, March 2010.

[105] Holger Heyn. A symbiotic liaison between the genetic and epigenetic code. *Front Genet*, 5(R10):113, 2014.

[106] Huda Y Zoghbi and Arthur L Beaudet. Epigenetics and Human Disease. *Cold Spring Harb Perspect Biol*, 8(2):a019497, February 2016.

[107] D H Ledbetter, V M Riccardi, S D Airhart, R J Strobel, B S Keenan, and J D Crawford. Deletions of chromosome 15 as a cause of the Prader-Willi syndrome. *N. Engl. J. Med.*, 304(6):325–329, February 1981.

[108] R D Nicholls, J H Knoll, M G Butler, S Karam, and M Lalande. Genetic imprinting suggested by maternal heterodisomy in nondeletion Prader-Willi syndrome. *Nature*, 342(6247):281–285, November 1989.

[109] R E Magenis, M G Brown, D A Lacy, S Budden, and S LaFranchi. Is Angelman syndrome an alternate result of del(15)(q11q13)? *Am. J. Med. Genet.*, 28(4):829–838, December 1987.

[110] Kristel R van Eijk, Simone de Jong, Eric Strengman, Jacobine E Buizer-Voskamp, René S Kahn, Marco P Boks, Steve Horvath, and Roel A Ophoff. Identification of schizophrenia-

associated loci by combining DNA methylation and gene expression data from whole blood. *Eur. J. Hum. Genet.*, 23(8):1106–1110, August 2015.

[111] E R Gamazon, J A Badner, L Cheng, C Zhang, D Zhang, N J Cox, E S Gershon, J R Kelsoe, T A Greenwood, C M Nievergelt, C Chen, R McKinney, P D Shilling, N J Schork, E N Smith, C S Bloss, J I Nurnberger, H J Edenberg, T Foroud, D L Koller, W A Scheftner, W Coryell, J Rice, W B Lawson, E A Nwulia, M Hipolito, W Byerley, F J McMahon, T G Schulze, W H Berrettini, J B Potash, P P Zandi, P B Mahon, M G McInnis, S Zöllner, P Zhang, D W Craig, S Szelinger, T B Barrett, and C Liu. Enrichment of cis-regulatory gene expression SNPs and methylation quantitative trait loci among bipolar disorder susceptibility variants. *Mol. Psychiatry*, 18(3):340–346, March 2013.

[112] Michael J Ziller, Hongcang Gu, Fabian Müller, Julie Donaghey, Linus T-Y Tsai, Oliver Kohlbacher, Philip L De Jager, Evan D Rosen, David A Bennett, Bradley E Bernstein, Andreas Gnirke, and Alexander Meissner. Charting a dynamic DNA methylation landscape of the human genome. *Nature*, 500(7463):477–481, August 2013.

[113] Anna Portela and Manel Esteller. Epigenetic modifications and human disease. *Nat. Biotechnol.*, 28(10):1057–1068, October 2010.

[114] Minjia Tan, Hao Luo, Sangkyu Lee, Fulai Jin, Jeong Soo Yang, Emilie Montellier, Thierry Buchou, Zhongyi Cheng, Sophie Rousseaux, Nisha Rajagopal, Zhike Lu, Zhen Ye, Qin Zhu, Joanna Wysocka, Yang Ye, Saadi Khochbin, Bing Ren, and Yingming Zhao. Identification of 67 histone marks and histone lysine crotonylation as a new type of histone modification. *Cell*, 146(6):1016–1028, September 2011.

[115] Zhixin Tian, Nikola Tolić, Rui Zhao, Ronald J Moore, Shawna M Hengel, Errol W Robinson, David L Stenoien, Si Wu, Richard D Smith, and Ljiljana Paša-Tolić. Enhanced top-down characterization of histone post-translational modifications. *Genome Biol.*, 13(10):R86, 2012.

[116] Bing Li, Michael Carey, and Jerry L Workman. The role of chromatin during transcription. *Cell*, 128(4):707–719, February 2007.

[117] L Siggens and K Ekwall. Epigenetics, chromatin and genome organization: recent advances from the ENCODE project. *J. Intern. Med.*, 276(3):201–214, September 2014.

[118] Chloe M Rivera and Bing Ren. Mapping human epigenomes. *Cell*, 155(1):39–55, September 2013.

[119] Zhibin Wang, Chongzhi Zang, Kairong Cui, Dustin E Schones, Artem Barski, Weiqun Peng, and Keji Zhao. Genome-wide mapping of HATs and HDACs reveals distinct functions in active and inactive genes. *Cell*, 138(5):1019–1031, September 2009.

[120] Dustin E Schones, Kairong Cui, Suresh Cuddapah, Tae-Young Roh, Artem Barski, Zhibin Wang, Gang Wei, and Keji Zhao. Dynamic regulation of nucleosome positioning in the human genome. *Cell*, 132(5):887–898, March 2008.

[121] Geeta J Narlikar, Ramasubramanian Sundaramoorthy, and Tom Owen-Hughes. Mechanisms and functions of ATP-dependent chromatin-remodeling enzymes. *Cell*, 154(3):490–503, August 2013.

[122] F Petrij, R H Giles, H G Dauwerse, J J Saris, R C Hennekam, M Masuno, N Tommerup, G J van Ommen, R H Goodman, and D J Peters. Rubinstein-Taybi syndrome caused by mutations in the transcriptional co-activator CBP. *Nature*, 376(6538):348–351, July 1995.

[123] Jason D Buenrostro, Paul G Giresi, Lisa C Zaba, Howard Y Chang, and William J Greenleaf. Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position. *Nat. Methods*, 10(12):1213–1218, December 2013.

[124] Jacob F Degner, Athma A Pai, Roger Pique-Regi, Jean-Baptiste Veyrieras, Daniel J Gaffney, Joseph K Pickrell, Sherryl De Leon, Katelyn Michelini, Noah Lewellen, Gregory E Crawford, Matthew Stephens, Yoav Gilad, and Jonathan K Pritchard. DNaseâĂŁI sensitivity QTLs are a major determinant of human expression variation. *Nature*, 482(7385):390–394, February 2012.

[125] Harm-Jan Westra and Lude Franke. From genome to function by studying eQTLs. *Biochim. Biophys. Acta*, 1842(10):1896–1902, October 2014.

[126] Antigone S Dimas, Samuel Deutsch, Barbara E Stranger, Stephen B Montgomery, Christelle Borel, Homa Attar-Cohen, Catherine Ingle, Claude Beazley, Maria Gutierrez-Arcelus,

Magdalena Sekowska, Marilyne Gagnebin, James Nisbett, Panos Deloukas, Emmanouil T Dermitzakis, and Stylianos E Antonarakis. Common regulatory variation impacts gene expression in a cell type-dependent manner. *Science*, 325(5945):1246–1250, September 2009.

[127] Alexandra C Nica, Leopold Parts, Daniel Glass, James Nisbet, Amy Barrett, Magdalena Sekowska, Mary Travers, Simon Potter, Elin Grundberg, Kerrin Small, Åsa K Hedman, Veronique Bataille, Jordana Tzenova Bell, Gabriela Surdulescu, Antigone S Dimas, Catherine Ingle, Frank O Nestle, Paola di Meglio, Josine L Min, Alicja Wilk, Christopher J Hammond, Neelam Hassanali, Tsun-Po Yang, Stephen B Montgomery, Steve O'Rahilly, Cecilia M Lindgren, Krina T Zondervan, Nicole Soranzo, Ines Barroso, Richard Durbin, Kourosh Ahmadi, Panos Deloukas, Mark I McCarthy, Emmanouil T Dermitzakis, Timothy D Spector, and MuTHER Consortium. The architecture of gene regulatory variation across multiple human tissues: the MuTHER study. *PLoS Genet.*, 7(2):e1002003, 2011.

[128] Jingyuan Fu, Marcel G M Wolfs, Patrick Deelen, Harm-Jan Westra, Rudolf S N Fehrmann, Gerard J Te Meerman, Wim A Buurman, Sander S M Rensen, Harry J M Groen, Rinse K Weersma, Leonard H van den Berg, Jan Veldink, Roel A Ophoff, Harold Snieder, David van Heel, Ritsert C Jansen, Marten H Hofker, Cisca Wijmenga, and Lude Franke. Unraveling the regulatory mechanisms underlying tissue-dependent genetic variation of gene expression. *PLoS Genet.*, 8(1):e1002431, January 2012.

[129] Stacey L Edwards, Jonathan Beesley, Juliet D French, and Alison M Dunning. Beyond GWASs: illuminating the dark road from association to function. *Am. J. Hum. Genet.*, 93(5):779–797, November 2013.

[130] GTEx Consortium. The Genotype-Tissue Expression (GTEx) project. *Nat. Genet.*, 45(6):580–585, June 2013.

[131] Xinchen Wang, Nathan R Tucker, Gizem Rizki, Robert Mills, Peter Hl Krijger, Elzo de Wit, Vidya Subramanian, Eric Bartell, Xinh-Xinh Nguyen, Jiangchuan Ye, Jordan Leyton-Mange, Elena V Dolmatova, Pim van der Harst, Wouter de Laat, Patrick T Ellinor, Christopher Newton-Cheh, David J Milan, Manolis Kellis, and Laurie A Boyer. Discovery and validation of sub-threshold genome-wide association study loci using epigenomic signatures. *Elife*, 5:1061, 2016.

[132] D E Reich and E S Lander. On the allelic spectrum of human disease. *Trends Genet.*, 17(9):502–510, September 2001.

[133] Jonathan K Pritchard and Nancy J Cox. The allelic architecture of human disease genes: common disease-common variant...or not? *Hum. Mol. Genet.*, 11(20):2417–2423, October 2002.

[134] J K Pritchard. Are rare variants responsible for susceptibility to complex diseases? *Am. J. Hum. Genet.*, 69(1):124–137, July 2001.

[135] Walter Bodmer and Carolina Bonilla. Common and rare variants in multifactorial susceptibility to common diseases. *Nat. Genet.*, 40(6):695–701, June 2008.

[136] Niels Grarup, Camilla H Sandholt, Torben Hansen, and Oluf Pedersen. Genetic susceptibility to type 2 diabetes and obesity: from genome-wide association studies to rare variants and beyond. *Diabetologia*, 57(8):1528–1541, August 2014.

[137] Jian Yang, Andrew Bakshi, Zhihong Zhu, Gibran Hemani, Anna A E Vinkhuyzen, Sang Hong Lee, Matthew R Robinson, John R B Perry, Ilja M Nolte, Jana V van Vliet-Ostaptchouk, Harold Snieder, LifeLines Cohort Study, Tonu Esko, Lili Milani, Reedik Mägi, Andres Metspalu, Anders Hamsten, Patrik K E Magnusson, Nancy L Pedersen, Erik Ingelsson, Nicole Soranzo, Matthew C Keller, Naomi R Wray, Michael E Goddard, and Peter M Visscher. Genetic variance estimation with imputed variants finds negligible missing heritability for human height and body mass index. *Nat. Genet.*, 47(10):1114–1120, October 2015.

[138] Andrew R Wood, Tonu Esko, Jian Yang, Sailaja Vedantam, Tune H Pers, Stefan Gustafsson, Audrey Y Chu, Karol Estrada, Jian'an Luan, Zoltán Kutalik, Najaf Amin, Martin L Buchkovich, Damien C Croteau-Chonka, Felix R Day, Yanan Duan, Tove Fall, Rudolf Fehrmann, Teresa Ferreira, Anne U Jackson, Juha Karjalainen, Ken Sin Lo, Adam E Locke, Reedik Mägi, Evelin Mihailov, Eleonora Porcu, Joshua C Randall, André Scherag, Anna A E Vinkhuyzen, Harm-Jan Westra, Thomas W Winkler, Tsegaselassie Workalemahu, Jing Hua Zhao, Devin Absher, Eva Albrecht, Denise Anderson, Jeffrey Baron, Marian Beekman, Ayse Demirkan, Georg B Ehret, Bjarke Feenstra, Mary F Feitosa, Krista Fischer, Ross M Fraser,

Anuj Goel, Jian Gong, Anne E Justice, Stavroula Kanoni, Marcus E Kleber, Kati Kristiansson, Unhee Lim, Vaneet Lotay, Julian C Lui, Massimo Mangino, Irene Mateo Leach, Carolina Medina-Gomez, Michael A Nalls, Dale R Nyholt, Cameron D Palmer, Dorota Pasko, Sonali Pechlivanis, Inga Prokopenko, Janina S Ried, Stephan Ripke, Dmitry Shungin, Alena Stančáková, Rona J Strawbridge, Yun Ju Sung, Toshiko Tanaka, Alexander Teumer, Stella Trompet, Sander W van der Laan, Jessica van Setten, Jana V van Vliet-Ostaptchouk, Zhaoming Wang, Loic Yengo, Weihua Zhang, Uzma Afzal, Johan Ärnlöv, Gillian M Arscott, Stefania Bandinelli, Amy Barrett, Claire Bellis, Amanda J Bennett, Christian Berne, Matthias Blüher, Jennifer L Bolton, Yvonne Böttcher, Heather A Boyd, Marcel Bruinenberg, Brendan M Buckley, Steven Buyske, Ida H Caspersen, Peter S Chines, Robert Clarke, Simone Claudi-Boehm, Matthew Cooper, E Warwick Daw, Pim A De Jong, Joris Deelen, Graciela Delgado, Josh C Denny, Rosalie Dhonukshe-Rutten, Maria Dimitriou, Alex S F Doney, Marcus Dörr, Niina Eklund, Elodie Eury, Lasse Folkersen, Melissa E Garcia, Frank Geller, Vilmantas Giedraitis, Alan S Go, Harald Grallert, Tanja B Grammer, Jürgen Gräßler, Henrik Grönberg, Lisette C P G M de Groot, Christopher J Groves, Jeffrey Haessler, Per Hall, Toomas Haller, Göran Hallmans, Anke Hannemann, Catharina A Hartman, Maija Hassinen, Caroline Hayward, Nancy L Heard-Costa, Quinta Helmer, Gibran Hemani, Anjali K Henders, Hans L Hillege, Mark A Hlatky, Wolfgang Hoffmann, Per Hoffmann, Oddgeir Holmen, Jeanine J Houwing-Duistermaat, Thomas Illig, Aaron Isaacs, Alan L James, Janina Jeff, Berit Johansen, Asa Johansson, Jennifer Jolley, Thorhildur Juliusdottir, Juhani Junttila, Abel N Kho, Leena Kinnunen, Norman Klopp, Thomas Kocher, Wolfgang Kratzer, Peter Lichtner, Lars Lind, Jaana Lindström, Stéphane Lobbens, Mattias Lorentzon, Yingchang Lu, Valeriya Lyssenko, Patrik K E Magnusson, Anubha Mahajan, Marc Maillard, Wendy L McArdle, Colin A McKenzie, Stela McLachlan, Paul J McLaren, Cristina Menni, Sigrun Merger, Lili Milani, Alireza Moayyeri, Keri L Monda, Mario A Morken, Gabriele Müller, Martina Müller-Nurasyid, Arthur W Musk, Narisu Narisu, Matthias Nauck, Ilja M Nolte, Markus M Nöthen, Laticia Oozageer, Stefan Pilz, Nigel W Rayner, Frida Renstrom, Neil R Robertson, Lynda M Rose, Ronan Roussel, Serena Sanna, Hubert Scharnagl, Salome Scholtens, Fredrick R Schumacher, Heribert Schunkert, Robert A Scott, Joban Sehmi, Thomas Seufferlein, Jianxin Shi, Karri Silventoinen, Johannes H Smit, Albert Vernon

Smith, Joanna Smolonska, Alice V Stanton, Kathleen Stirrups, David J Stott, Heather M Stringham, Johan Sundström, Morris A Swertz, Ann-Christine Syvänen, Bamidele O Tayo, Gudmar Thorleifsson, Jonathan P Tyrer, Suzanne van Dijk, Natasja M van Schoor, Nathalie van der Velde, Diana van Heemst, Floor V A van Oort, Sita H Vermeulen, Niek Verweij, Judith M Vonk, Lindsay L Waite, Melanie Waldenberger, Roman Wennauer, Lynne R Wilkens, Christina Willenborg, Tom Wilsgaard, Mary K Wojczynski, Andrew Wong, Alan F Wright, and Qunyu... Zhang. Defining the role of common variation in the genomic and biological architecture of adult human height. *Nat. Genet.*, 46(11):1173–1186, November 2014.

[139] Fowzan S Alkuraya. Human knockout research: new horizons and opportunities. *Trends Genet.*, 31(2):108–115, February 2015.

[140] Gero Hütter, Daniel Nowak, Maximilian Mossner, Susanne Ganepola, Arne Müssig, Kristina Allers, Thomas Schneider, Jörg Hofmann, Claudia Kücherer, Olga Blau, Igor W Blau, Wolf K Hofmann, and Eckhard Thiel. Long-term control of HIV by CCR5 Delta32/Delta32 stem-cell transplantation. *N. Engl. J. Med.*, 360(7):692–698, February 2009.

[141] Yali Xue, Allan Daly, Bryndis Yngvadottir, Mengning Liu, Graham Coop, Yuseob Kim, Pardis Sabeti, Yuan Chen, Jim Stalker, Elizabeth Huckle, John Burton, Steven Leonard, Jane Rogers, and Chris Tyler-Smith. Spread of an inactive form of caspase-12 in humans is due to recent positive selection. *Am. J. Hum. Genet.*, 78(4):659–670, April 2006.

[142] Alexis Battle, Zia Khan, Sidney H Wang, Amy Mitrano, Michael J Ford, Jonathan K Pritchard, and Yoav Gilad. Genomic variation. Impact of regulatory variation from RNA to protein. *Science*, 347(6222):664–667, February 2015.

[143] Vagheesh M Narasimhan, Yali Xue, and Chris Tyler-Smith. Human Knockout Carriers: Dead, Diseased, Healthy, or Improved? *Trends in Molecular Medicine*, 22(4):341–351, April 2016.

[144] Catarina D Campbell and Evan E Eichler. Properties and rates of germline mutations in humans. *Trends Genet.*, 29(10):575–584, October 2013.

[145] Fady M Mikhail. Copy number variations and human genetic disease. *Current Opinion in Pediatrics*, 26(6):646–652, December 2014.

[146] Donna M McDonald-McGinn and Kathleen E Sullivan. Chromosome 22q11.2 Deletion Syndrome (DiGeorge Syndrome/Velocardiofacial Syndrome). *Medicine*, 90(1):1–18, January 2011.

[147] U Francke. Williams-Beuren syndrome: genes and mechanisms. *Hum. Mol. Genet.*, 8(10):1947–1954, 1999.

[148] Sebastien Jacquemont, Alexandre Reymond, Flore Zufferey, Louise Harewood, Robin G Walters, Zoltán Kutalik, Danielle Martinet, Yiping Shen, Armand Valsesia, Noam D Beckmann, Gudmar Thorleifsson, Marco Belfiore, Sonia Bouquillon, Dominique Campion, Nicole de Leeuw, Bert B A de Vries, Tonu Esko, Bridget A Fernandez, Fernando Fernández-Aranda, José Manuel Fernández-Real, Mònica Gratacòs, Audrey Guilmatre, Juliane Hoyer, Marjo-Riitta Järvelin, R Frank Kooy, Ants Kurg, Cédric Le Caignec, Katrin Männik, Orah S Platt, Damien Sanlaville, Mieke M Van Haelst, Sergi Villatoro Gomez, Faida Walha, Bai-Lin Wu, Yongguo Yu, Azzedine Aboura, Marie-Claude Addor, Yves Alembik, Stylianos E Antonarakis, Benoît Arveiler, Magalie Barth, Nathalie Bednarek, Frédérique Béna, Sven Bergmann, Mylène Beri, Laura Bernardini, Bettina Blaumeiser, Dominique Bonneau, Armand Bottani, Odile Boute, Han G Brunner, Dorothée Cailley, Patrick Callier, Jean Chiesa, Jacqueline Chrast, Lachlan Coin, Charles Coutton, Jean-Marie Cuisset, Jean-Christophe Cuvellier, Albert David, Bénédicte de Freminville, Bruno Delobel, Marie-Ange Delrue, Bénédicte Demeer, Dominique Descamps, Gérard Didelot, Klaus Dieterich, Vittoria Disciglio, Martine Doco-Fenzy, Séverine Drunat, Bénédicte Duban-Bedu, Christèle Dubourg, Julia S El-Sayed Moustafa, Paul Elliott, Brigitte H W Faas, Laurence Faivre, Anne Faudet, Florence Fellmann, Alessandra Ferrarini, Richard Fisher, Elisabeth Flori, Lukas Forer, Dominique Gaillard, Marion Gerard, Christian Gieger, Stefania Gimelli, Giorgio Gimelli, Hans J Grabe, Agnès Guichet, Olivier Guillin, Anna-Liisa Hartikainen, Délphine Heron, Loyse Hippolyte, Muriel Holder, Georg Homuth, Bertrand Isidor, Sylvie Jaillard, Zdenek Jaros, Susana Jiménez-Murcia, Géraldine Joly Helas, Philippe Jonveaux, Satu Kaksonen, Boris Keren, Anita Kloss-Brandstätter, Nine V A M Knoers, David A Koolen, Peter M Kroisel, Florian Kronenberg, Audrey Labalme, Emilie Landais, Elisabetta Lapi, Valérie Layet, Solenn Legallic, Bruno Leheup, Barbara Leube, Suzanne Lewis, Josette Lucas, Kay D MacDermot, Pall Magnusson, Christian Marshall, Michèle Mathieu-Dramard, Mark I McCarthy,

Thomas Meitinger, Maria Antonietta Mencarelli, Giuseppe Merla, Alexandre Moerman, Vincent Mooser, Fanny Morice-Picard, Mafalda Mucciolo, Matthias Nauck, Ndeye Coumba Ndiaye, Ann Nordgren, Laurent Pasquier, Florence Petit, Rolph Pfundt, Ghislaine Plessis, Evica Rajcan-Separovic, Gian Paolo Ramelli, Anita Rauch, Roberto Ravazzolo, Andre Reis, Alessandra Renieri, Cristobal Richart, Janina S Ried, Claudine Rieubland, Wendy Roberts, Katharina M Roetzer, Caroline Rooryck, Massimiliano Rossi, Evald Saemundsen, Véronique Satre, Claudia Schurmann, Engilbert Sigurdsson, Dimitri J Stavropoulos, Hreinn Stefansson, Carola Tengström, Unnur Thorsteinsdottir, Francisco J Tinahones, Renaud Touraine, Louis Vallée, Ellen van Binsbergen, Nathalie Van der Aa, Catherine Vincent-Delorme, Sophie Visvikis-Siest, Peter Vollenweider, Henry Völzke, Anneke T Vulto-van Silfhout, Gerard Waeber, Carina Wallgren-Pettersson, Robert M Witwicki, Simon Zwolinksi, Joris Andrieux, Xavier Estivill, James F Gusella, Omar Gustafsson, Andres Metspalu, Stephen W Scherer, Kari Stefansson, Alexandra I F Blakemore, Jacques S Beckmann, and Philippe Froguel. Mirror extreme BMI phenotypes associated with gene dosage at the chromosome 16p11.2 locus. *Nature*, 478(7367):97–102, October 2011.

[149] R G Walters, S Jacquemont, A Valsesia, A J de Smith, D Martinet, J Andersson, M Falchi, F Chen, J Andrieux, S Lobbens, B Delobel, F Stutzmann, J S El-Sayed Moustafa, J-C Chèvre, C Lecoeur, V Vatin, S Bouquillon, J L Buxton, O Boute, M Holder-Espinasse, J-M Cuisset, M-P Lemaitre, A-E Ambresin, A Brioschi, M Gaillard, V Giusti, F Fellmann, A Ferrarini, N Hadjikhani, D Campion, A Guilmatre, A Goldenberg, N Calmels, J-L Mandel, C Le Caignec, A David, B Isidor, M-P Cordier, S Dupuis-Girod, A Labalme, D Sanlaville, M Béri-Dexheimer, P Jonveaux, B Leheup, K Ounap, E G Bochukova, E Henning, J Keogh, R J Ellis, K D Macdermot, M M van Haelst, C Vincent-Delorme, G Plessis, R Touraine, A Philippe, V Malan, M Mathieu-Dramard, J Chiesa, B Blaumeiser, R F Kooy, R Caiazzo, M Pigeyre, B Balkau, R Sladek, S Bergmann, V Mooser, D Waterworth, A Reymond, P Vollenweider, G Waeber, A Kurg, P Palta, T Esko, A Metspalu, M Nelis, P Elliott, A-L Hartikainen, M I McCarthy, L Peltonen, L Carlsson, P Jacobson, L Sjöström, N Huang, M E Hurles, S O'Rahilly, I S Farooqi, K Männik, M-R Jarvelin, F Pattou, D Meyre, A J Walley, L J M Coin, A I F Blakemore, P Froguel, and J S Beckmann. A new highly penetrant form of obesity due to deletions on chromosome 16p11.2. *Nature*, 463(7281):671–675, February 2010.

[150] Teri A Manolio, Francis S Collins, Nancy J Cox, David B Goldstein, Lucia A Hindorff, David J Hunter, Mark I McCarthy, Erin M Ramos, Lon R Cardon, Aravinda Chakravarti, Judy H Cho, Alan E Guttmacher, Augustine Kong, Leonid Kruglyak, Elaine Mardis, Charles N Rotimi, Montgomery Slatkin, David Valle, Alice S Whittemore, Michael Boehnke, Andrew G Clark, Evan E Eichler, Greg Gibson, Jonathan L Haines, Trudy F C Mackay, Steven A McCarroll, and Peter M Visscher. Finding the missing heritability of complex diseases. *Nature*, 461(7265):747–753, October 2009.

[151] Hreinn Stefansson, Andreas Meyer-Lindenberg, Stacy Steinberg, Brynja Magnusdottir, Katrin Morgen, Sunna Arnarsdottir, Gyda Bjornsdottir, G Bragi Walters, Gudrun A Jonsdottir, Orla M Doyle, Heike Tost, Oliver Grimm, Solveig Kristjansdottir, Heimir Snorrason, Solveig R Davidsdottir, Larus J Gudmundsson, Gudbjorn F Jonsson, Berglind Stefansdottir, Isafold Helgadottir, Magnus Haraldsson, Birna Jonsdottir, Johan H Thygesen, Adam J Schwarz, Michael Didriksen, Tine B Stensbøl, Michael Brammer, Shitij Kapur, Jonas G Halldorsson, Stefan Hreidarsson, Evald Saemundsen, Engilbert Sigurdsson, and Kari Stefansson. CNVs conferring risk of autism or schizophrenia affect cognition in controls. *Nature*, 505(7483):361–366, January 2014.

[152] Wellcome Trust Case Control Consortium, Nick Craddock, Matthew E Hurles, Niall Cardin, Richard D Pearson, Vincent Plagnol, Samuel Robson, Damjan Vukcevic, Chris Barnes, Donald F Conrad, Eleni Giannoulatou, Chris Holmes, Jonathan L Marchini, Kathy Stirrups, Martin D Tobin, Louise V Wain, Chris Yau, Jan Aerts, Tariq Ahmad, T Daniel Andrews, Hazel Arbury, Anthony Attwood, Adam Auton, Stephen G Ball, Anthony J Balmforth, Jeffrey C Barrett, Ines Barroso, Anne Barton, Amanda J Bennett, Sanjeev Bhaskar, Katarzyna Blaszczyk, John Bowes, Oliver J Brand, Peter S Braund, Francesca Bredin, Gerome Breen, Morris J Brown, Ian N Bruce, Jaswinder Bull, Oliver S Burren, John Burton, Jake Byrnes, Sian Caesar, Chris M Clee, Alison J Coffey, John M C Connell, Jason D Cooper, Anna F Dominiczak, Kate Downes, Hazel E Drummond, Darshna Dudakia, Andrew Dunham, Bernadette Ebbs, Diana Eccles, Sarah Edkins, Cathryn Edwards, Anna Elliot, Paul Emery, David M Evans, Gareth Evans, Steve Eyre, Anne Farmer, I Nicol Ferrier, Lars Feuk, Tomas Fitzgerald, Edward Flynn, Alistair Forbes, Liz Forty, Jayne A Franklyn, Rachel M Freathy, Polly Gibbs, Paul Gilbert, Omer Gokumen, Katherine Gordon-Smith,

Emma Gray, Elaine Green, Chris J Groves, Detelina Grozeva, Rhian Gwilliam, Anita Hall, Naomi Hammond, Matt Hardy, Pile Harrison, Neelam Hassanali, Husam Hebaishi, Sarah Hines, Anne Hinks, Graham A Hitman, Lynne Hocking, Eleanor Howard, Philip Howard, Joanna M M Howson, Debbie Hughes, Sarah Hunt, John D Isaacs, Mahim Jain, Derek P Jewell, Toby Johnson, Jennifer D Jolley, Ian R Jones, Lisa A Jones, George Kirov, Cordelia F Langford, Hana Lango Allen, G Mark Lathrop, James Lee, Kate L Lee, Charlie Lees, Kevin Lewis, Cecilia M Lindgren, Meeta Maisuria-Armer, Julian Maller, John Mansfield, Paul Martin, Dunecan C O Massey, Wendy L McArdle, Peter McGuffin, Kirsten E McLay, Alex Mentzer, Michael L Mimmack, Ann E Morgan, Andrew P Morris, Craig Mowat, Simon Myers, William Newman, Elaine R Nimmo, Michael C O'Donovan, Abiodun Onipinla, Ifejinelo Onyiah, Nigel R Ovington, Michael J Owen, Kimmo Palin, Kirstie Parnell, David Pernet, John R B Perry, Anne Phillips, Dalila Pinto, Natalie J Prescott, Inga Prokopenko, Michael A Quail, Suzanne Rafelt, Nigel W Rayner, Richard Redon, David M Reid, Renwick, Susan M Ring, Neil Robertson, Ellie Russell, David St Clair, Jennifer G Sambrook, Jeremy D Sanderson, Helen Schuilenburg, Carol E Scott, Richard Scott, Sheila Seal, Sue Shaw-Hawkins, Beverley M Shields, Matthew J Simmonds, Debbie J Smyth, Elilan Somaskantharajah, Katarina Spanova, Sophia Steer, Jonathan Stephens, Helen E Stevens, Millicent A Stone, Zhan Su, Deborah P M Symmons, John R Thompson, Wendy Thomson, Mary E Travers, Clare Turnbull, Armand Valsesia, Mark Walker, Neil M Walker, Chris Wallace, Margaret Warren-Perry, Nicholas A Watkins, John Webster, Michael N Weedon, Anthony G Wilson, Matthew Woodburn, B Paul Wordsworth, Allan H Young, Eleftheria Zeggini, Nigel P Carter, Timothy M Frayling, Charles Lee, Gil McVean, Patricia B Munroe, Aarno Palotie, Stephen J Sawcer, Stephen W Scherer, David P Strachan, Chris Tyler-Smith, Matthew A Brown, Paul R Burton, Mark J Caulfield, Alastair Compston, Martin Farrall, Stephen C L Gough, Alistair S Hall, Andrew T Hattersley, Adrian V S Hill, Christopher G Mathew, Marcus Pembrey, Jack Satsangi, Michael R Stratton, Jane Worthington, Panos Deloukas, Audrey Duncanson, Dominic P Kwiatkowski, Mark I McCarthy, Willem Ouwehand, Miles Parkes, Nazneen Rahman, John A Todd, Nilesh J Samani, and Peter Donnelly. Genome-wide association study of CNVs in 16,000 cases of eight common diseases and 3,000 shared controls. *Nature*, 464(7289):713–720, April 2010.

[153] P Chomczynski and N Sacchi. Single-step method of RNA isolation by acid guanidinium thiocyanate-phenol-chloroform extraction. *Anal. Biochem.*, 162(1):156–159, April 1987.

[154] Jason D Buenrostro, Beijing Wu, Howard Y Chang, and William J Greenleaf. ATAC-seq: A Method for Assaying Chromatin Accessibility Genome-Wide. *Curr Protoc Mol Biol*, 109:21.29.1–9, 2015.

[155] Momoko Horikoshi, Reedik MÓŞgi, Martijn van de Bunt, Ida Surakka, Antti-Pekka Sarin, Anubha Mahajan, Letizia Marullo, Gudmar Thorleifsson, Sara HÓŞgg, Jouke Jan Hottenga, Claes Ladenvall, Janina S Ried, Thomas W Winkler, Sara M Willems, Natalia Pervjakova, Tonu Esko, Marian Beekman, Christopher P Nelson, Christina Willenborg, Steven Wiltshire, Teresa Ferreira, Juan Fernandez, Kyle J Gaulton, Valgerdur Steinthorsdottir, Anders Hamsten, Patrik K E Magnusson, Gonneke Willemsen, Yuri Milaneschi, Neil R Robertson, Christopher J Groves, Amanda J Bennett, Terho LehtimÓŞki, Jorma S Viikari, Johan Rung, Valeriya Lyssenko, Markus Perola, Iris M Heid, Christian Herder, Harald Grallert, Martina Müller-Nurasyid, Michael Roden, Elina Hyppönen, Aaron Isaacs, Elisabeth M van Leeuwen, Lennart C Karssen, Evelin Mihailov, Jeanine J Houwing-Duistermaat, Anton J M de Craen, Joris Deelen, Aki S Havulinna, Matthew Blades, Christian Hengstenberg, Jeanette Erdmann, Heribert Schunkert, Jaakko Kaprio, Martin D Tobin, Nilesh J Samani, Lars Lind, Veikko Salomaa, Cecilia M Lindgren, P Eline Slagboom, Andres Metspalu, Cornelia M van Duijn, Johan G Eriksson, Annette Peters, Christian Gieger, Antti Jula, Leif Groop, Olli T Raitakari, Chris Power, Brenda W J H Penninx, Eco de Geus, Johannes H Smit, Dorret I Boomsma, Nancy L Pedersen, Erik Ingelsson, Unnur Thorsteinsdottir, Kari Stefansson, Samuli Ripatti, Inga Prokopenko, Mark I McCarthy, Andrew P Morris, and ENGAGE Consortium. Discovery and Fine-Mapping of Glycaemic and Obesity-Related Trait Loci Using High-Density Imputation. *PLoS Genet.*, 11(7):e1005230, July 2015.

[156] Ida Surakka, Momoko Horikoshi, Reedik Mägi, Antti-Pekka Sarin, Anubha Mahajan, Vasiliki Lagou, Letizia Marullo, Teresa Ferreira, Benjamin Miraglio, Sanna Timonen, Johannes Kettunen, Matti Pirinen, Juha Karjalainen, Gudmar Thorleifsson, Sara Hägg, Jouke Jan Hottenga, Aaron Isaacs, Claes Ladenvall, Marian Beekman, Tonu Esko, Janina S Ried, Christopher P Nelson, Christina Willenborg, Stefan Gustafsson, Harm-Jan Westra, Matthew

Blades, Anton J M de Craen, Eco J de Geus, Joris Deelen, Harald Grallert, Anders Hamsten, Aki S Havulinna, Christian Hengstenberg, Jeanine J Houwing-Duistermaat, Elina Hyppönen, Lennart C Karssen, Terho Lehtimäki, Valeriya Lyssenko, Patrik K E Magnusson, Evelin Mihailov, Martina Müller-Nurasyid, John-Patrick Mpindi, Nancy L Pedersen, Brenda W J H Penninx, Markus Perola, Tune H Pers, Annette Peters, Johan Rung, Johannes H Smit, Valgerdur Steinthorsdottir, Martin D Tobin, Natalia Tsernikova, Elisabeth M van Leeuwen, Jorma S Viikari, Sara M Willems, Gonneke Willemsen, Heribert Schunkert, Jeanette Erdmann, Nilesh J Samani, Jaakko Kaprio, Lars Lind, Christian Gieger, Andres Metspalu, P Eline Slagboom, Leif Groop, Cornelia M van Duijn, Johan G Eriksson, Antti Jula, Veikko Salomaa, Dorret I Boomsma, Christine Power, Olli T Raitakari, Erik Ingelsson, Marjo-Riitta Järvelin, Unnur Thorsteinsdottir, Lude Franke, Elina Ikonen, Olli Kallioniemi, Vilja Pietiäinen, Cecilia M Lindgren, Kari Stefansson, Aarno Palotie, Mark I McCarthy, Andrew P Morris, Inga Prokopenko, Samuli Ripatti, and ENGAGE Consortium. The impact of low-frequency and rare variants on lipid levels. *Nat. Genet.*, 47(6):589–597, June 2015.

[157] Douglas M Ruderfer, Tymor Hamamsy, Monkol Lek, Konrad J Karczewski, David Kavanagh, Kaitlin E Samocha, Exome Aggregation Consortium, Mark J Daly, Daniel G MacArthur, Menachem Fromer, and Shaun M Purcell. Patterns of genic intolerance of rare copy number variation in 59,898 human exomes. *Nat. Genet.*, August 2016.

[158] Joseph K Pickrell. Joint analysis of functional genomic data and genome-wide association studies of 18 human traits. *Am. J. Hum. Genet.*, 94(4):559–573, April 2014.

[159] Martijn van de Bunt, Jocelyn E Manning Fox, Xiaoqing Dai, Amy Barrett, Caleb Grey, Lei Li, Amanda J Bennett, Paul R Johnson, Raymond V Rajotte, Kyle J Gaulton, Emmanouil T Dermitzakis, Patrick E MacDonald, Mark I McCarthy, and Anna L Gloyn. 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(12):e1005694, December 2015.

[160] Jesse D Ziebarth, Anindya Bhattacharya, and Yan Cui. CTCFBSDB 2.0: a database for CTCF-binding sites and genome organization. *Nucleic Acids Res.*, 41(Database issue):D188–94, January 2013.

[161] Aaron R Quinlan and Ira M Hall. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics*, 26(6):841–842, March 2010.

[162] W James Kent, Charles W Sugnet, Terrence S Furey, Krishna M Roskin, Tom H Pringle, Alan M Zahler, and David Haussler. The human genome browser at UCSC. *Genome Res.*, 12(6):996–1006, June 2002.

[163] R Core Team. R: A Language and Environment for Statistical Computing. *R Foundation for Statistical Computing*, 2013.

[164] M Dowle, T Short, S Lianoglou, R Saporta, A Srinivasan, and E Antonyan. data.table: Extension of data.frame. *(2014)*, 2014.

[165] H Wickham. ggplot2. *Springer Science Business Media*, 2009.

[166] Dirk Schübeler. Function and information content of DNA methylation. *Nature*, 517(7534):321–326, January 2015.

[167] L Venolia and S M Gartler. Comparison of transformation efficiency of human active and inactive X-chromosomal DNA. *Nature*, 302(5903):82–83, March 1983.

[168] E Li, C Beard, and R Jaenisch. Role for DNA methylation in genomic imprinting. *Nature*, 366(6453):362–365, November 1993.

[169] A C Bell and G Felsenfeld. Methylation of a CTCF-dependent boundary controls imprinted expression of the Igf2 gene. *Nature*, 405(6785):482–485, May 2000.

[170] B Hendrich and A Bird. Identification and characterization of a family of mammalian methyl-CpG binding proteins. *Mol. Cell. Biol.*, 18(11):6538–6547, November 1998.

[171] A P Bird and A P Wolffe. Methylation-induced repression–belts, braces, and chromatin. *Cell*, 99(5):451–454, November 1999.

[172] Michael B Stadler, Rabih Murr, Lukas Burger, Robert Ivanek, Florian Lienert, Anne Schöler, Erik van Nimwegen, Christiane Wirbelauer, Edward J Oakeley, Dimos Gaidatzis, Vijay K Tiwari, and Dirk Schübeler. DNA-binding factors shape the mouse methylome at distal regulatory regions. *Nature*, 480(7378):490–495, December 2011.

[173] Rafael A Irizarry, Christine Ladd-Acosta, Bo Wen, Zhijin Wu, Carolina Montano, Patrick Onyango, Hengmi Cui, Kevin Gabo, Michael Rongione, Maree Webster, Hong Ji, James B Potash, Sarven Sabunciyan, and Andrew P Feinberg. The human colon cancer methylome shows similar hypo- and hypermethylation at conserved tissue-specific CpG island shores. *Nat. Genet.*, 41(2):178–186, February 2009.

[174] Peter A Jones. Functions of DNA methylation: islands, start sites, gene bodies and beyond. *Nat. Rev. Genet.*, 13(7):484–492, July 2012.

[175] Marina Bibikova, Bret Barnes, Chan Tsan, Vincent Ho, Brandy Klotzle, Jennie M Le, David Delano, Lu Zhang, Gary P Schroth, Kevin L Gunderson, Jian-Bing Fan, and Richard Shen. High density DNA methylation array with single CpG site resolution. *Genomics*, 98(4):288–295, October 2011.

[176] Sebastian Moran, Carles Arribas, and Manel Esteller. Validation of a DNA methylation microarray for 850,000 CpG sites of the human genome enriched in enhancer sequences. *Epigenomics*, 8(3):389–399, March 2016.

[177] Helena Kilpinen and Emmanouil T Dermitzakis. Genetic and epigenetic contribution to complex traits. *Hum. Mol. Genet.*, 21(R1):R24–8, October 2012.

[178] Rosanna Weksberg, Adam C Smith, Jeremy Squire, and Paul Sadowski. Beckwith-Wiedemann syndrome demonstrates a role for epigenetic control of normal development. *Hum. Mol. Genet.*, 12 Spec No 1:R61–8, April 2003.

[179] I K Temple and J P H Shield. Transient neonatal diabetes, a disorder of imprinting. *J. Med. Genet.*, 39(12):872–875, December 2002.

[180] C N Hales, D J Barker, P M Clark, L J Cox, C Fall, C Osmond, and P D Winter. Fetal and infant growth and impaired glucose tolerance at age 64. *BMJ*, 303(6809):1019–1022, October 1991.

[181] Magnus Kaijser, Anna-Karin Edstedt Bonamy, Olof Akre, Sven Cnattingius, Fredrik Granath, Mikael Norman, and Anders Ekbom. Perinatal risk factors for diabetes in later life. *Diabetes*, 58(3):523–526, March 2009.

[182] K Pilgaard, K Færch, B Carstensen, P Poulsen, C Pisinger, O Pedersen, D R Witte, T Hansen, T Jørgensen, and A Vaag. Low birthweight and premature birth are both associated with type 2 diabetes in a random sample of middle-aged Danes. *Diabetologia*, 53(12):2526–2530, December 2010.

[183] A C Ravelli, J H van der Meulen, R P Michels, C Osmond, D J Barker, C N Hales, and O P Bleker. Glucose tolerance in adults after prenatal exposure to famine. *Lancet*, 351(9097):173–177, January 1998.

[184] D R McCance, D J Pettitt, R L Hanson, L T Jacobsson, W C Knowler, and P H Bennett. Birth weight and non-insulin dependent diabetes: thrifty genotype, thrifty phenotype, or surviving small baby genotype? *BMJ*, 308(6934):942–945, April 1994.

[185] Vanessa Derenji Ferreira de Mello, Leena Pulkkinen, Marianne Lalli, Marjukka Kolehmainen, Jussi Pihlajamäki, and Matti Uusitupa. DNA methylation in obesity and type 2 diabetes. *Ann. Med.*, 46(3):103–113, May 2014.

[186] Bastiaan T Heijmans, Elmar W Tobi, Aryeh D Stein, Hein Putter, Gerard J Blauw, Ezra S Susser, P Eline Slagboom, and L H Lumey. Persistent epigenetic differences associated with prenatal exposure to famine in humans. *Proc. Natl. Acad. Sci. U.S.A.*, 105(44):17046–17049, November 2008.

[187] Elmar W Tobi, L H Lumey, Rudolf P Talens, Dennis Kremer, Hein Putter, Aryeh D Stein, P Eline Slagboom, and Bastiaan T Heijmans. DNA methylation differences after exposure to prenatal famine are common and timing- and sex-specific. *Hum. Mol. Genet.*, 18(21):4046–4053, November 2009.

[188] Christopher G Bell, Sarah Finer, Cecilia M Lindgren, Gareth A Wilson, Vardhman K Rakyan, Andrew E Teschendorff, Pelin Akan, Elia Stupka, Thomas A Down, Inga Prokopenko, Ian M Morison, Jonathan Mill, Ruth Pidsley, International Type 2 Diabetes 1q Consortium, Panos Deloukas, Timothy M Frayling, Andrew T Hattersley, Mark I McCarthy, Stephan Beck, and Graham A Hitman. Integrated genetic and epigenetic analysis identifies haplotype-specific methylation in the FTO type 2 diabetes and obesity susceptibility locus. *PLoS ONE*, 5(11):e14040, 2010.

[189] Gidon Toperoff, Dvir Aran, Jeremy D Kark, Michael Rosenberg, Tatyana Dubnikov, Batel Nissan, Julio Wainstein, Yechiel Friedlander, Ephrat Levy-Lahad, Benjamin Glaser, and Asaf Hellman. Genome-wide survey reveals predisposing diabetes type 2-related DNA methylation variations in human peripheral blood. *Hum. Mol. Genet.*, 21(2):371–383, January 2012.

[190] Wei Yuan, Yudong Xia, Christopher G Bell, Idil Yet, Teresa Ferreira, Kirsten J Ward, Fei Gao, A Katrina Loomis, Craig L Hyde, Honglong Wu, Hanlin Lu, Yuan Liu, Kerrin S Small, Ana Viñuela, Andrew P Morris, María Berdasco, Manel Esteller, M Julia Brosnan, Panos Deloukas, Mark I McCarthy, Sally L John, Jordana T Bell, Jun Wang, and Tim D Spector. An integrated epigenomic analysis for type 2 diabetes susceptibility loci in monozygotic twins. *Nat Commun*, 5:5719, 2014.

[191] John C Chambers, Marie Loh, Benjamin Lehne, Alexander Drong, Jennifer Kriebel, Valeria Motta, Simone Wahl, Hannah R Elliott, Federica Rota, William R Scott, Weihua Zhang, Sian-Tsung Tan, Gianluca Campanella, Marc Chadeau-Hyam, Loic Yengo, Rebecca C Richmond, Martyna Adamowicz-Brice, Uzma Afzal, Kiymet Bozaoglu, Zuan Yu Mok, Hong Kiat Ng, François Pattou, Holger Prokisch, Michelle Ann Rozario, Letizia Tarantini, James Abbott, Mika Ala-Korpela, Benedetta Albetti, Ole Ammerpohl, Pier Alberto Bertazzi, Christine Blancher, Robert Caiazzo, John Danesh, Tom R Gaunt, Simon de Lusignan, Christian Gieger, Thomas Illig, Sujeet Jha, Simon Jones, Jeremy Jowett, Antti J Kangas, Anuradhani Kasturiratne, Norihiro Kato, Navaratnam Kotea, Sudhir Kowlessur, Janne Pitkäniemi, Prakash Punjabi, Danish Saleheen, Clemens Schafmayer, Pasi Soininen, E Shyong Tai, Barbara Thorand, Jaakko Tuomilehto, Ananda Rajitha Wickremasinghe, Soterios A Kyrtopoulos, Timothy J Aitman, Christian Herder, Jochen Hampe, Stephane Cauchi, Caroline L Relton, Philippe Froguel, Richie Soong, Paolo Vineis, Marjo-Riitta Järvelin, James Scott, Harald Grallert, Valentina Bollati, Paul Elliott, Mark I McCarthy, and Jaspal S Kooner. Epigenome-wide association of DNA methylation markers in peripheral blood from Indian Asians and Europeans with incident type 2 diabetes: a nested case-control study. *Lancet Diabetes Endocrinol*, 3(7):526–534, July 2015.

[192] Ines Florath, Katja Butterbach, Jonathan Heiss, Melanie Bewerunge-Hudler, Yan Zhang, Ben Schöttker, and Hermann Brenner. Type 2 diabetes and leucocyte DNA methylation: an

epigenome-wide association study in over 1,500 older adults. *Diabetologia*, 59(1):130–138, January 2016.

[193] Emma Nilsson, Per Anders Jansson, Alexander Perfilyev, Petr Volkov, Maria Pedersen, Maria K Svensson, Pernille Poulsen, Rasmus Ribel-Madsen, Nancy L Pedersen, Peter Almgren, João Fadista, Tina Rönn, Bente Klarlund Pedersen, Camilla Scheele, Allan Vaag, and Charlotte Ling. Altered DNA methylation and differential expression of genes influencing metabolism and inflammation in adipose tissue from subjects with type 2 diabetes. *Diabetes*, 63(9):2962–2976, September 2014.

[194] Marloes Dekker Nitert, Tasnim Dayeh, Peter Volkov, Targ Elgzyri, Elin Hall, Emma Nilsson, Beatrice T Yang, Stefan Lang, Hemang Parikh, Ylva Wessman, Holger Weishaupt, Joanne Attema, Mia Abels, Nils Wierup, Peter Almgren, Per Anders Jansson, Tina Rönn, Ola Hansson, Karl-Fredrik Eriksson, Leif Groop, and Charlotte Ling. Impact of an exercise intervention on DNA methylation in skeletal muscle from first-degree relatives of patients with type 2 diabetes. *Diabetes*, 61(12):3322–3332, December 2012.

[195] Rasmus Ribel-Madsen, Mario F Fraga, Stine Jacobsen, Jette Bork-Jensen, Ester Lara, Vincenzo Calvanese, Agustín F Fernández, Martin Friedrichsen, Birgitte F Vind, Kurt Højlund, Henning Beck-Nielsen, Manel Esteller, Allan Vaag, and Pernille Poulsen. Genome-wide analysis of DNA methylation differences in muscle and fat from monozygotic twins discordant for type 2 diabetes. *PLoS ONE*, 7(12):e51302, 2012.

[196] Michael Volkmar, Sarah Dedeurwaerder, Daniel A Cunha, Matladi N Ndlovu, Matthieu Defrance, Rachel Deplus, Emilie Calonne, Ute Volkmar, Mariana Igoillo-Esteve, Najib Naamane, Silvia Del Guerra, Matilde Masini, Marco Bugliani, Piero Marchetti, Miriam Cnop, Decio L Eizirik, and François Fuks. DNA methylation profiling identifies epigenetic dysregulation in pancreatic islets from type 2 diabetic patients. *EMBO J.*, 31(6):1405–1426, March 2012.

[197] Mary E Travers, Deborah J G Mackay, Marloes Dekker Nitert, Andrew P Morris, Cecilia M Lindgren, Andrew Berry, Paul R Johnson, Neil Hanley, Leif C Groop, Mark I McCarthy, and Anna L Gloyn. 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(3):987–992, March 2013.

[198] Tasnim Dayeh, Petr Volkov, Sofia Salö, Elin Hall, Emma Nilsson, Anders H Olsson, Clare L Kirkpatrick, Claes B Wollheim, Lena Eliasson, Tina Rönn, Karl Bacos, and Charlotte Ling. 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 Genet.*, 10(3):e1004160, March 2014.

[199] Karl Bacos, Linn Gillberg, Petr Volkov, Anders H Olsson, Torben Hansen, Oluf Pedersen, Anette Prior Gjesing, Hans Eiberg, Tiinamaija Tuomi, Peter Almgren, Leif Groop, Lena Eliasson, Allan Vaag, Tasnim Dayeh, and Charlotte Ling. Blood-based biomarkers of age-associated epigenetic changes in human islets associate with insulin secretion and diabetes. *Nat Commun*, 7:11089, 2016.

[200] Simon R Andrews. FASTQC: A quality control tool for high throughput sequence data. *http://www.bioinformatics.babraham.ac.uk/projects/fastqc/*, 2010.

[201] Felix Krueger and Simon R Andrews. Bismark: a flexible aligner and methylation caller for Bisulfite-Seq applications. *Bioinformatics*, 27(11):1571–1572, June 2011.

[202] Ben Langmead and Steven L Salzberg. Fast gapped-read alignment with Bowtie 2. *Nat. Methods*, 9(4):357–359, April 2012.

[203] Heng Li, Bob Handsaker, Alec Wysoker, Tim Fennell, Jue Ruan, Nils Homer, Gabor Marth, Goncalo Abecasis, Richard Durbin, and 1000 Genome Project Data Processing Subgroup. The Sequence Alignment/Map format and SAMtools. *Bioinformatics*, 25(16):2078–2079, August 2009.

[204] Robert Lowe, Carolina Gemma, Huriya Beyan, Mohammed I Hawa, Alexandra Bazeos, R David Leslie, Alexandre Montpetit, Vardhman K Rakyan, and Sreeram V Ramagopalan. Buccals are likely to be a more informative surrogate tissue than blood for epigenome-wide association studies. *Epigenetics*, 8(4):445–454, April 2013.

[205] Dimos Gaidatzis, Lukas Burger, Rabih Murr, Anita Lerch, Sophie Dessus-Babus, Dirk Schübeler, and Michael B Stadler. DNA sequence explains seemingly disordered methylation

levels in partially methylated domains of Mammalian genomes. *PLoS Genet.*, 10(2):e1004143, February 2014.

[206] Lukas Burger, Dimos Gaidatzis, Dirk Schübeler, and Michael B Stadler. Identification of active regulatory regions from DNA methylation data. *Nucleic Acids Res.*, 41(16):e155–e155, September 2013.

[207] Martin J Aryee, Andrew E Jaffe, Hector Corrada-Bravo, Christine Ladd-Acosta, Andrew P Feinberg, Kasper D Hansen, and Rafael A Irizarry. Minfi: a flexible and comprehensive Bioconductor package for the analysis of Infinium DNA methylation microarrays. *Bioinformatics*, 30(10):1363–1369, May 2014.

[208] Randall J Pruim, Ryan P Welch, Serena Sanna, Tanya M Teslovich, Peter S Chines, Terry P Gliedt, Michael Boehnke, Goncalo R Abecasis, and Cristen J Willer. LocusZoom: regional visualization of genome-wide association scan results. *Bioinformatics*, 26(18):2336–2337, September 2010.

[209] Cristen J Willer and Karen L Mohlke. Finding genes and variants for lipid levels after genome-wide association analysis. *Curr. Opin. Lipidol.*, 23(2):98–103, April 2012.

[210] Sarah Voisin, Markus Sällman Almén, Galina Y Zheleznyakova, Lina Lundberg, Sanaz Zarei, Sandra Castillo, Fia Ence Eriksson, Emil K Nilsson, Matthias Blüher, Yvonne Böttcher, Peter Kovacs, Janis Klovins, Mathias Rask-Andersen, and Helgi B Schiöth. Many obesity-associated SNPs strongly associate with DNA methylation changes at proximal promoters and enhancers. *Genome Med*, 7(1):103, 2015.

[211] Maria Tsompana and Michael J Buck. Chromatin accessibility: a window into the genome. *Epigenetics Chromatin*, 7(1):33, 2014.

[212] Paul G Giresi, Jonghwan Kim, Ryan M McDaniell, Vishwanath R Iyer, and Jason D Lieb. FAIRE (Formaldehyde-Assisted Isolation of Regulatory Elements) isolates active regulatory elements from human chromatin. *Genome Res.*, 17(6):877–885, June 2007.

[213] Ruiyu Xie, Andrea C Carrano, and Maike Sander. A systems view of epigenetic networks regulating pancreas development and β-cell function. *Wiley Interdiscip Rev Syst Biol Med*, 7(1):1–11, January 2015.

[214] Bradley E Bernstein, Tarjei S Mikkelsen, Xiaohui Xie, Michael Kamal, Dana J Huebert, James Cuff, Ben Fry, Alex Meissner, Marius Wernig, Kathrin Plath, Rudolf Jaenisch, Alexandre Wagschal, Robert Feil, Stuart L Schreiber, and Eric S Lander. A bivalent chromatin structure marks key developmental genes in embryonic stem cells. *Cell*, 125(2):315–326, April 2006.

[215] Fabio Mohn, Michael Weber, Michael Rebhan, Tim C Roloff, Jens Richter, Michael B Stadler, Miriam Bibel, and Dirk Schübeler. Lineage-specific polycomb targets and de novo DNA methylation define restriction and potential of neuronal progenitors. *Mol. Cell*, 30(6):755–766, June 2008.

[216] Ruiyu Xie, Logan J Everett, Hee-Woong Lim, Nisha A Patel, Jonathan Schug, Evert Kroon, Olivia G Kelly, Allen Wang, Kevin A D'Amour, Allan J Robins, Kyoung-Jae Won, Klaus H Kaestner, and Maike Sander. Dynamic chromatin remodeling mediated by polycomb proteins orchestrates pancreatic differentiation of human embryonic stem cells. *Cell Stem Cell*, 12(2):224–237, February 2013.

[217] Allen Wang, Feng Yue, Yan Li, Ruiyu Xie, Thomas Harper, Nisha A Patel, Kayla Muth, Jeffrey Palmer, Yunjiang Qiu, Jinzhao Wang, Dieter K Lam, Jeffrey C Raum, Doris A Stoffers, Bing Ren, and Maike Sander. Epigenetic priming of enhancers predicts developmental competence of hESC-derived endodermal lineage intermediates. *Cell Stem Cell*, 16(4):386–399, April 2015.

[218] Cécile Haumaitre, Olivia Lenoir, and Raphael Scharfmann. Histone deacetylase inhibitors modify pancreatic cell fate determination and amplify endocrine progenitors. *Mol. Cell. Biol.*, 28(20):6373–6383, October 2008.

[219] Alireza Rezania, Jennifer E Bruin, Payal Arora, Allison Rubin, Irina Batushansky, Ali Asadi, Shannon O'Dwyer, Nina Quiskamp, Majid Mojibian, Tobias Albrecht, Yu Hsuan Carol Yang, James D Johnson, and Timothy J Kieffer. Reversal of diabetes with insulin-producing cells derived in vitro from human pluripotent stem cells. *Nat. Biotechnol.*, 32(11):1121–1133, November 2014.

[220] Felicia W Pagliuca, Jeffrey R Millman, Mads Gürtler, Michael Segel, Alana Van Dervort,

Jennifer Hyoje Ryu, Quinn P Peterson, Dale Greiner, and Douglas A Melton. Generation of functional human pancreatic β cells in vitro. *Cell*, 159(2):428–439, October 2014.

[221] Courtney K Burrows, Nicholas E Banovich, Bryan J Pavlovic, Kristen Patterson, Irene Gallego Romero, Jonathan K Pritchard, and Yoav Gilad. Genetic Variation, Not Cell Type of Origin, Underlies the Majority of Identifiable Regulatory Differences in iPSCs. *PLoS Genet.*, 12(1):e1005793, January 2016.

[222] Foad Rouhani, Natsuhiko Kumasaka, Miguel Cardoso de Brito, Allan Bradley, Ludovic Vallier, and Daniel Gaffney. Genetic background drives transcriptional variation in human induced pluripotent stem cells. *PLoS Genet.*, 10(6):e1004432, June 2014.

[223] Aija Kyttälä, Roksana Moraghebi, Cristina Valensisi, Johannes Kettunen, Colin Andrus, Kalyan Kumar Pasumarthy, Mahito Nakanishi, Ken Nishimura, Manami Ohtaka, Jere Weltner, Ben Van Handel, Olavi Parkkonen, Juha Sinisalo, Anu Jalanko, R David Hawkins, Niels-Bjarne Woods, Timo Otonkoski, and Ras Trokovic. Genetic Variability Overrides the Impact of Parental Cell Type and Determines iPSC Differentiation Potential. *Stem Cell Reports*, 6(2):200–212, February 2016.

[224] Ori Bar-Nur, Holger A Russ, Shimon Efrat, and Nissim Benvenisty. 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, July 2011.

[225] Jose M Polo, Susanna Liu, Maria Eugenia Figueroa, Warakorn Kulalert, Sarah Eminli, Kah Yong Tan, Effie Apostolou, Matthias Stadtfeld, Yushan Li, Toshi Shioda, Sridaran Natesan, Amy J Wagers, Ari Melnick, Todd Evans, and Konrad Hochedlinger. Cell type of origin influences the molecular and functional properties of mouse induced pluripotent stem cells. *Nat. Biotechnol.*, 28(8):848–855, August 2010.

[226] K Kim, A Doi, B Wen, K Ng, R Zhao, P Cahan, J Kim, M J Aryee, H Ji, L I R Ehrlich, A Yabuuchi, A Takeuchi, K C Cunniff, H Hongguang, S McKinney-Freeman, O Naveiras, T J Yoon, R A Irizarry, N Jung, J Seita, J Hanna, P Murakami, R Jaenisch, R Weissleder, S H Orkin, I L Weissman, A P Feinberg, and G Q Daley. Epigenetic memory in induced pluripotent stem cells. *Nature*, 467(7313):285–290, September 2010.

[227] Martijn van de Bunt, Majlinda Lako, Amy Barrett, Anna L Gloyn, Mattias Hansson, Mark I McCarthy, Nicola L Beer, and Christian Honoré. Insights into islet development and biology through characterization of a human iPSC-derived endocrine pancreas model. *Islets*, 8(3):83–95, April 2016.

[228] Amanda M Ackermann, Zhiping Wang, Jonathan Schug, Ali Naji, and Klaus H Kaestner. Integration of ATAC-seq and RNA-seq identifies human alpha cell and beta cell signature genes. *Mol Metab*, 5(3):233–244, March 2016.

[229] Holger A Russ, Yael Bar, Philippe Ravassard, and Shimon Efrat. In vitro proliferation of cells derived from adult human beta-cells revealed by cell-lineage tracing. *Diabetes*, 57(6):1575–1583, June 2008.

[230] Deborah Hay, Jim R Hughes, Christian Babbs, James O J Davies, Bryony J Graham, Lars L P Hanssen, Mira T Kassouf, A Marieke Oudelaar, Jacqueline A Sharpe, Maria C Suciu, Jelena Telenius, Ruth Williams, Christina Rode, Pik-Shan Li, Len A Pennacchio, Jacqueline A Sloane-Stanley, Helena Ayyub, Sue Butler, Tatjana Sauka-Spengler, Richard J Gibbons, Andrew J H Smith, William G Wood, and Douglas R Higgs. Genetic dissection of the α-globin super-enhancer in vivo. *Nat. Genet.*, 48(8):895–903, August 2016.

[231] Jelena Telenius. Pipesite. DNase, ChIP and ATAC-seq pipeline. 2016.

[232] Ben Langmead, Cole Trapnell, Mihai Pop, and Steven L Salzberg. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol.*, 10(3):R25, 2009.

[233] Felix Krueger. Trim galore. 2012.

[234] Tanja Magoč and Steven L Salzberg. FLASH: fast length adjustment of short reads to improve genome assemblies. *Bioinformatics*, 27(21):2957–2963, November 2011.

[235] Yong Zhang, Tao Liu, Clifford A Meyer, Jérôme Eeckhoute, David S Johnson, Bradley E Bernstein, Chad Nusbaum, Richard M Myers, Myles Brown, Wei Li, and X Shirley Liu. Model-based analysis of ChIP-Seq (MACS). *Genome Biol.*, 9(9):R137, 2008.

[236] Alexandra C Nica, Halit Ongen, Jean-Claude Irminger, Domenico Bosco, Thierry Berney, Stylianos E Antonarakis, Philippe A Halban, and Emmanouil T Dermitzakis. Cell-type, allelic, and genetic signatures in the human pancreatic beta cell transcriptome. *Genome Res.*, 23(9):1554–1562, September 2013.

[237] Li Shen, Ning-Yi Shao, Xiaochuan Liu, Ian Maze, Jian Feng, and Eric J Nestler. diffReps: detecting differential chromatin modification sites from ChIP-seq data with biological replicates. *PLoS ONE*, 8(6):e65598, 2013.

[238] Cory Y McLean, Dave Bristor, Michael Hiller, Shoa L Clarke, Bruce T Schaar, Craig B Lowe, Aaron M Wenger, and Gill Bejerano. GREAT improves functional interpretation of cis-regulatory regions. *Nat. Biotechnol.*, 28(5):495–501, May 2010.

[239] Robert E Thurman, Eric Rynes, Richard Humbert, Jeff Vierstra, Matthew T Maurano, Eric Haugen, Nathan C Sheffield, Andrew B Stergachis, Hao Wang, Benjamin Vernot, Kavita Garg, Sam John, Richard Sandstrom, Daniel Bates, Lisa Boatman, Theresa K Canfield, Morgan Diegel, Douglas Dunn, Abigail K Ebersol, Tristan Frum, Erika Giste, Audra K Johnson, Ericka M Johnson, Tanya Kutyavin, Bryan Lajoie, Bum-Kyu Lee, Kristen Lee, Darin London, Dimitra Lotakis, Shane Neph, Fidencio Neri, Eric D Nguyen, Hongzhu Qu, Alex P Reynolds, Vaughn Roach, Alexias Safi, Minerva E Sanchez, Amartya Sanyal, Anthony Shafer, Jeremy M Simon, Lingyun Song, Shinny Vong, Molly Weaver, Yongqi Yan, Zhancheng Zhang, Zhuzhu Zhang, Boris Lenhard, Muneesh Tewari, Michael O Dorschner, R Scott Hansen, Patrick A Navas, George Stamatoyannopoulos, Vishwanath R Iyer, Jason D Lieb, Shamil R Sunyaev, Joshua M Akey, Peter J Sabo, Rajinder Kaul, Terrence S Furey, Job Dekker, Gregory E Crawford, and John A Stamatoyannopoulos. The accessible chromatin landscape of the human genome. *Nature*, 489(7414):75–82, September 2012.

[240] Alan P Boyle, Lingyun Song, Bum-Kyu Lee, Darin London, Damian Keefe, Ewan Birney, Vishwanath R Iyer, Gregory E Crawford, and Terrence S Furey. High-resolution genome-wide in vivo footprinting of diverse transcription factors in human cells. *Genome Res.*, 21(3):456–464, March 2011.

[241] Philipp Voigt, Wee-Wei Tee, and Danny Reinberg. A double take on bivalent promoters. *Genes Dev.*, 27(12):1318–1338, June 2013.

[242] M Cristina Nostro, Farida Sarangi, Shinichiro Ogawa, Audrey Holtzinger, Barbara Corneo, Xueling Li, Suzanne J Micallef, In-Hyun Park, Christina Basford, Michael B Wheeler, George Q Daley, Andrew G Elefanty, Edouard G Stanley, and Gordon Keller. Stage-specific signaling through TGFβ family members and WNT regulates patterning and pancreatic specification of human pluripotent stem cells. *Development*, 138(5):861–871, March 2011.

[243] Corentin Cras-Méneur, Lin Li, Raphael Kopan, and M Alan Permutt. Presenilins, Notch dose control the fate of pancreatic endocrine progenitors during a narrow developmental window. *Genes Dev.*, 23(17):2088–2101, September 2009.

[244] Geert A Martens, Lei Jiang, Karine H Hellemans, Geert Stangé, Harry Heimberg, Finn C Nielsen, Olivier Sand, Jacques Van Helden, Leentje Van Lommel, Frans Schuit, Frans K Gorus, and Daniel G Pipeleers. Clusters of conserved beta cell marker genes for assessment of beta cell phenotype. *PLoS ONE*, 6(9):e24134, 2011.

[245] Joris van Arensbergen, Javier García-Hurtado, Ignasi Morán, Miguel Angel Maestro, Xiaobo Xu, Mark Van de Casteele, Anouchka L Skoudy, Matteo Palassini, Harry Heimberg, and Jorge Ferrer. Derepression of Polycomb targets during pancreatic organogenesis allows insulin-producing beta-cells to adopt a neural gene activity program. *Genome Res.*, 20(6):722–732, June 2010.

[246] Daniel Eberhard. Neuron and beta-cell evolution: learning about neurons is learning about beta-cells. *Bioessays*, 35(7):584–584, July 2013.

[247] K Fujimoto and K S Polonsky. Pdx1 and other factors that regulate pancreatic beta-cell survival. *Diabetes Obes Metab*, 11 Suppl 4(s4):30–37, November 2009.

[248] Brandon L Taylor, Fen-Fen Liu, and Maike Sander. Nkx6.1 is essential for maintaining the functional state of pancreatic beta cells. *Cell Rep*, 4(6):1262–1275, September 2013.

[249] Michelle J Doyle and Lori Sussel. Nkx2.2 regulates beta-cell function in the mature islet. *Diabetes*, 56(8):1999–2007, August 2007.

[250] Alan W Hart, Sebastien Mella, Jacek Mendrychowski, Veronica van Heyningen, and Dirk A Kleinjan. The developmental regulator Pax6 is essential for maintenance of islet cell function in the adult mouse pancreas. *PLoS ONE*, 8(1):e54173, 2013.

[251] Maria E Wilson, David Scheel, and Michael S German. Gene expression cascades in pancreatic development. *Mech. Dev.*, 120(1):65–80, January 2003.

[252] Joost H A Martens and Hendrik G Stunnenberg. BLUEPRINT: mapping human blood cell epigenomes. *Haematologica*, 98(10):1487–1489, October 2013.

[253] Nathan R Rose and Robert J Klose. Understanding the relationship between DNA methylation and histone lysine methylation. *Biochim. Biophys. Acta*, 1839(12):1362–1372, December 2014.

[254] François Fuks, Paul J Hurd, Daniel Wolf, Xinsheng Nan, Adrian P Bird, and Tony Kouzarides. The methyl-CpG-binding protein MeCP2 links DNA methylation to histone methylation. *J. Biol. Chem.*, 278(6):4035–4040, February 2003.

[255] Marco Morselli, William A Pastor, Barbara Montanini, Kevin Nee, Roberto Ferrari, Kai Fu, Giancarlo Bonora, Liudmilla Rubbi, Amander T Clark, Simone Ottonello, Steven E Jacobsen, and Matteo Pellegrini. In vivo targeting of de novo DNA methylation by histone modifications in yeast and mouse. *Elife*, 4:e06205, 2015.

[256] Michael Weber, Ines Hellmann, Michael B Stadler, Liliana Ramos, Svante Pääbo, Michael Rebhan, and Dirk Schübeler. Distribution, silencing potential and evolutionary impact of promoter DNA methylation in the human genome. *Nat. Genet.*, 39(4):457–466, April 2007.

[257] Marc A Schaub, Alan P Boyle, Anshul Kundaje, Serafim Batzoglou, and Michael Snyder. Linking disease associations with regulatory information in the human genome. *Genome Res.*, 22(9):1748–1759, September 2012.

[258] Ignasi Morán, Ildem Akerman, Martijn van de Bunt, Ruiyu Xie, Marion Benazra, Takao Nammo, Luis Arnes, Nikolina Nakić, Javier García-Hurtado, Santiago Rodríguez-Seguí, Lorenzo Pasquali, Claire Sauty-Colace, Anthony Beucher, Raphael Scharfmann, Joris van Arensbergen, Paul R Johnson, Andrew Berry, Clarence Lee, Timothy Harkins, Valery Gmyr, François Pattou, Julie Kerr-Conte, Lorenzo Piemonti, Thierry Berney, Neil Hanley, Anna L Gloyn, Lori Sussel, Linda Langman, Kenneth L Brayman, Maike Sander, Mark I McCarthy, Philippe Ravassard, and Jorge Ferrer. Human β cell transcriptome analysis uncovers lncRNAs

that are tissue-specific, dynamically regulated, and abnormally expressed in type 2 diabetes. *Cell Metab.*, 16(4):435–448, October 2012.

[259] Broad Institute. Picard Tools. *http://broadinstitute.github.io/picard/*.

[260] Jason Ernst and Manolis Kellis. ChromHMM: automating chromatin-state discovery and characterization. *Nat. Methods*, 9(3):215–216, March 2012.

[261] Bryce van de Geijn, Graham McVicker, Yoav Gilad, and Jonathan K Pritchard. WASP: allele-specific software for robust molecular quantitative trait locus discovery. *Nat. Methods*, 12(11):1061–1063, November 2015.

[262] 1000 Genomes Project Consortium, Adam Auton, Lisa D Brooks, Richard M Durbin, Erik P Garrison, Hyun Min Kang, Jan O Korbel, Jonathan L Marchini, Shane McCarthy, Gil A McVean, and Goncalo R Abecasis. A global reference for human genetic variation. *Nature*, 526(7571):68–74, October 2015.

[263] Charles E Grant, Timothy L Bailey, and William Stafford Noble. FIMO: scanning for occurrences of a given motif. *Bioinformatics*, 27(7):1017–1018, April 2011.

[264] Marie P Fogarty, Maren E Cannon, Swarooparani Vadlamudi, Kyle J Gaulton, and Karen L Mohlke. Identification of a regulatory variant that binds FOXA1 and FOXA2 at the CDC123/CAMK1D type 2 diabetes GWAS locus. *PLoS Genet.*, 10(9):e1004633, September 2014.

[265] Lucas D Ward and Manolis Kellis. HaploReg: a resource for exploring chromatin states, conservation, and regulatory motif alterations within sets of genetically linked variants. *Nucleic Acids Res.*, 40(Database issue):D930–4, January 2012.

[266] Josée Dupuis, Claudia Langenberg, Inga Prokopenko, Richa Saxena, Nicole Soranzo, Anne U Jackson, Eleanor Wheeler, Nicole L Glazer, Nabila Bouatia-Naji, Anna L Gloyn, Cecilia M Lindgren, Reedik Mägi, Andrew P Morris, Joshua Randall, Toby Johnson, Paul Elliott, Denis Rybin, Gudmar Thorleifsson, Valgerdur Steinthorsdottir, Peter Henneman, Harald Grallert, Abbas Dehghan, Jouke Jan Hottenga, Christopher S Franklin, Pau Navarro, Kijoung Song, Anuj Goel, John R B Perry, Josephine M Egan, Taina Lajunen,

Niels Grarup, Thomas Sparsø, Alex Doney, Benjamin F Voight, Heather M Stringham, Man Li, Stavroula Kanoni, Peter Shrader, Christine Cavalcanti-Proença, Meena Kumari, Lu Qi, Nicholas J Timpson, Christian Gieger, Carina Zabena, Ghislain Rocheleau, Erik Ingelsson, Ping An, Jeffrey O'Connell, Jian'an Luan, Amanda Elliott, Steven A McCarroll, Felicity Payne, Rosa Maria Roccasecca, François Pattou, Praveen Sethupathy, Kristin Ardlie, Yavuz Ariyurek, Beverley Balkau, Philip Barter, John P Beilby, Yoav Ben-Shlomo, Rafn Benediktsson, Amanda J Bennett, Sven Bergmann, Murielle Bochud, Eric Boerwinkle, Amélie Bonnefond, Lori L Bonnycastle, Knut Borch-Johnsen, Yvonne Böttcher, Eric Brunner, Suzannah J Bumpstead, Guillaume Charpentier, Yii-Der Ida Chen, Peter Chines, Robert Clarke, Lachlan J M Coin, Matthew N Cooper, Marilyn Cornelis, Gabe Crawford, Laura Crisponi, Ian N M Day, Eco J C de Geus, Jerome Delplanque, Christian Dina, Michael R Erdos, Annette C Fedson, Antje Fischer-Rosinsky, Nita G Forouhi, Caroline S Fox, Rune Frants, Maria Grazia Franzosi, Pilar Galan, Mark O Goodarzi, Jürgen Graessler, Christopher J Groves, Scott Grundy, Rhian Gwilliam, Ulf Gyllensten, Samy Hadjadj, Göran Hallmans, Naomi Hammond, Xijing Han, Anna-Liisa Hartikainen, Neelam Hassanali, Caroline Hayward, Simon C Heath, Serge Hercberg, Christian Herder, Andrew A Hicks, David R Hillman, Aroon D Hingorani, Albert Hofman, Jennie Hui, Joe Hung, Bo Isomaa, Paul R V Johnson, Torben Jørgensen, Antti Jula, Marika Kaakinen, Jaakko Kaprio, Y Antero Kesaniemi, Mika Kivimaki, Beatrice Knight, Seppo Koskinen, Peter Kovacs, Kirsten Ohm Kyvik, G Mark Lathrop, Debbie A Lawlor, Olivier Le Bacquer, Cécile Lecoeur, Yun Li, Valeriya Lyssenko, Robert Mahley, Massimo Mangino, Alisa K Manning, María Teresa Martínez-Larrad, Jarred B McAteer, Laura J McCulloch, Ruth McPherson, Christa Meisinger, David Melzer, David Meyre, Braxton D Mitchell, Mario A Morken, Sutapa Mukherjee, Silvia Naitza, Narisu Narisu, Matthew J Neville, Ben A Oostra, Marco Orrù, Ruth Pakyz, Colin N A Palmer, Giuseppe Paolisso, Cristian Pattaro, Daniel Pearson, John F Peden, Nancy L Pedersen, Markus Perola, Andreas F H Pfeiffer, Irene Pichler, Ozren Polasek, Danielle Posthuma, Simon C Potter, Anneli Pouta, Michael A Province, Bruce M Psaty, Wolfgang Rathmann, Nigel W Rayner, Kenneth Rice, Samuli Ripatti, Fernando Rivadeneira, Michael Roden, Olov Rolandsson, Annelli Sandbaek, Manjinder Sandhu, Serena Sanna, Avan Aihie Sayer, Paul Scheet, Laura J Scott, Udo Seedorf, Stephen J Sharp, Beverley Shields, Gunnar Sigurethsson,

Eric J G Sijbrands, Angela Silveira, Laila Simpson, Andrew Singleton, Nicholas L Smith, Ulla Sovio, Amy Swift, Holly Syddall, Ann-Christine Syvänen, Toshiko Tanaka, Barbara Thorand, Jean Tichet, Anke Tönjes, Tiinamaija Tuomi, Andre G Uitterlinden, Ko Willems van Dijk, Mandy van Hoek, Dhiraj Varma, Sophie Visvikis-Siest, Veronique Vitart, Nicole Vogelzangs, Gerard Waeber, Peter J Wagner, Andrew Walley, G Bragi Walters, Kim L Ward, Hugh Watkins, Michael N Weedon, Sarah H Wild, Gonneke Willemsen, Jaqueline C M Witteman, John W G Yarnell, Eleftheria Zeggini, Diana Zelenika, Björn Zethelius, Guangju Zhai, Jing Hua Zhao, M Carola Zillikens, DIAGRAM Consortium, GIANT Consortium, and Global BPgen Consortium an... New genetic loci implicated in fasting glucose homeostasis and their impact on type 2 diabetes risk. *Nat. Genet.*, 42(2):105–116, February 2010.

[267] David J Hodson, Ryan K Mitchell, Lorella Marselli, Timothy J Pullen, Silvia Gimeno Brias, Francesca Semplici, Katy L Everett, Dermot M F Cooper, Marco Bugliani, Piero Marchetti, Vanessa Lavallard, Domenico Bosco, Lorenzo Piemonti, Paul R Johnson, Stephen J Hughes, Daliang Li, Wen-Hong Li, A M James Shapiro, and Guy A Rutter. ADCY5 couples glucose to insulin secretion in human islets. *Diabetes*, 63(9):3009–3021, September 2014.

[268] João Fadista, Petter Vikman, Emilia Ottosson Laakso, Inês Guerra Mollet, Jonathan Lou Esguerra, Jalal Taneera, Petter Storm, Peter Osmark, Claes Ladenvall, Rashmi B Prasad, Karin B Hansson, Francesca Finotello, Kristina Uvebrant, Jones K Ofori, Barbara Di Camillo, Ulrika Krus, Corrado M Cilio, Ola Hansson, Lena Eliasson, Anders H Rosengren, Erik Renström, Claes B Wollheim, and Leif Groop. Global genomic and transcriptomic analysis of human pancreatic islets reveals novel genes influencing glucose metabolism. *Proc. Natl. Acad. Sci. U.S.A.*, 111(38):13924–13929, September 2014.

[269] Natsuhiko Kumasaka, Andrew J Knights, and Daniel J Gaffney. Fine-mapping cellular QTLs with RASQUAL and ATAC-seq. *Nat. Genet.*, 48(2):206–213, February 2016.

[270] Feng Zhang, Wenli Gu, Matthew E Hurles, and James R Lupski. Copy number variation in human health, disease, and evolution. *Annu Rev Genomics Hum Genet*, 10(1):451–481, 2009.

[271] Eleftheria Zeggini, Laura J Scott, Richa Saxena, Benjamin F Voight, Jonathan L Marchini, Tianle Hu, Paul I W de Bakker, Goncalo R Abecasis, Peter Almgren, Gitte Andersen, Kristin

Ardlie, Kristina Bengtsson Boström, Richard N Bergman, Lori L Bonnycastle, Knut Borch-Johnsen, Noël P Burtt, Hong Chen, Peter S Chines, Mark J Daly, Parimal Deodhar, Chia-Jen Ding, Alex S F Doney, William L Duren, Katherine S Elliott, Michael R Erdos, Timothy M Frayling, Rachel M Freathy, Lauren Gianniny, Harald Grallert, Niels Grarup, Christopher J Groves, Candace Guiducci, Torben Hansen, Christian Herder, Graham A Hitman, Thomas E Hughes, Bo Isomaa, Anne U Jackson, Torben Jørgensen, Augustine Kong, Kari Kubalanza, Finny G Kuruvilla, Johanna Kuusisto, Claudia Langenberg, Hana Lango, Torsten Lauritzen, Yun Li, Cecilia M Lindgren, Valeriya Lyssenko, Amanda F Marvelle, Christa Meisinger, Kristian Midthjell, Karen L Mohlke, Mario A Morken, Andrew D Morris, Narisu Narisu, Peter Nilsson, Katharine R Owen, Colin N A Palmer, Felicity Payne, John R B Perry, Elin Pettersen, Carl Platou, Inga Prokopenko, Lu Qi, Li Qin, Nigel W Rayner, Matthew Rees, Jeffrey J Roix, Anelli Sandbaek, Beverley Shields, Marketa Sjögren, Valgerdur Steinthorsdottir, Heather M Stringham, Amy J Swift, Gudmar Thorleifsson, Unnur Thorsteinsdottir, Nicholas J Timpson, Tiinamaija Tuomi, Jaakko Tuomilehto, Mark Walker, Richard M Watanabe, Michael N Weedon, Cristen J Willer, Wellcome Trust Case Control Consortium, Thomas Illig, Kristian Hveem, Frank B Hu, Markku Laakso, Kari Stefansson, Oluf Pedersen, Nicholas J Wareham, Ines Barroso, Andrew T Hattersley, Francis S Collins, Leif Groop, Mark I McCarthy, Michael Boehnke, and David Altshuler. Meta-analysis of genome-wide association data and large-scale replication identifies additional susceptibility loci for type 2 diabetes. *Nat. Genet.*, 40(5):638–645, May 2008.

[272] Jon McClellan and Mary-Claire King. Genetic heterogeneity in human disease. *Cell*, 141(2):210–217, April 2010.

[273] Maja Bucan, Brett S Abrahams, Kai Wang, Joseph T Glessner, Edward I Herman, Lisa I Sonnenblick, Ana I Alvarez Retuerto, Marcin Imielinski, Dexter Hadley, Jonathan P Bradfield, Cecilia Kim, Nicole B Gidaya, Ingrid Lindquist, Ted Hutman, Marian Sigman, Vlad Kustanovich, Clara M Lajonchere, Andrew Singleton, Junhyong Kim, Thomas H Wassink, William M McMahon, Thomas Owley, John A Sweeney, Hilary Coon, John I Nurnberger, Mingyao Li, Rita M Cantor, Nancy J Minshew, James S Sutcliffe, Edwin H Cook, Geraldine Dawson, Joseph D Buxbaum, Struan F A Grant, Gerard D Schellenberg, Daniel H Geschwind, and Hakon Hakonarson. Genome-wide analyses of exonic copy number

variants in a family-based study point to novel autism susceptibility genes. *PLoS Genet.*, 5(6):e1000536, June 2009.

[274] Jason Flannick, Gudmar Thorleifsson, Nicola L Beer, Suzanne B R Jacobs, Niels Grarup, Noël P Burtt, Anubha Mahajan, Christian Fuchsberger, Gil Atzmon, Rafn Benediktsson, John Blangero, Don W Bowden, Ivan Brandslund, Julia Brosnan, Frank Burslem, John Chambers, Yoon Shin Cho, Cramer Christensen, Desirée A Douglas, Ravindranath Duggirala, Zachary Dymek, Yossi Farjoun, Timothy Fennell, Pierre Fontanillas, Tom Forsen, Stacey Gabriel, Benjamin Glaser, Daniel F Gudbjartsson, Craig Hanis, Torben Hansen, Astradur B Hreidarsson, Kristian Hveem, Erik Ingelsson, Bo Isomaa, Stefan Johansson, Torben Jørgensen, Marit Eika Jørgensen, Sekar Kathiresan, Augustine Kong, Jaspal Kooner, Jasmina Kravic, Markku Laakso, Jong-Young Lee, Lars Lind, Cecilia M Lindgren, Allan Linneberg, Gisli Masson, Thomas Meitinger, Karen L Mohlke, Anders Molven, Andrew P Morris, Shobha Potluri, Rainer Rauramaa, Rasmus Ribel-Madsen, Ann-Marie Richard, Tim Rolph, Veikko Salomaa, Ayellet V Segrè, Hanna Skärstrand, Valgerdur Steinthorsdottir, Heather M Stringham, Patrick Sulem, E Shyong Tai, Yik Ying Teo, Tanya Teslovich, Unnur Thorsteinsdottir, Jeff K Trimmer, Tiinamaija Tuomi, Jaakko Tuomilehto, Fariba Vaziri-Sani, Benjamin F Voight, James G Wilson, Michael Boehnke, Mark I McCarthy, Pål R Njølstad, Oluf Pedersen, Go-T2D Consortium, T2D-GENES Consortium, Leif Groop, David R Cox, Kari Stefansson, and David Altshuler. Loss-of-function mutations in SLC30A8 protect against type 2 diabetes. *Nat. Genet.*, 46(4):357–363, April 2014.

[275] Tamara J Nicolson, Elisa A Bellomo, Nadeeja Wijesekara, Merewyn K Loder, Jocelyn M Baldwin, Armen V Gyulkhandanyan, Vasilij Koshkin, Andrei I Tarasov, Raffaella Carzaniga, Katrin Kronenberger, Tarvinder K Taneja, Gabriela da Silva Xavier, Sarah Libert, Philippe Froguel, Raphael Scharfmann, Volodymir Stetsyuk, Philippe Ravassard, Helen Parker, Fiona M Gribble, Frank Reimann, Robert Sladek, Stephen J Hughes, Paul R V Johnson, Myriam Masseboeuf, Remy Burcelin, Stephen A Baldwin, Ming Liu, Roberto Lara-Lemus, Peter Arvan, Frans C Schuit, Michael B Wheeler, Fabrice Chimienti, and Guy A Rutter. Insulin storage and glucose homeostasis in mice null for the granule zinc transporter ZnT8 and studies of the type 2 diabetes-associated variants. *Diabetes*, 58(9):2070–2083, September 2009.

[276] K Lemaire, M A Ravier, A Schraenen, J W M Creemers, R Van de Plas, M Granvik, L Van Lommel, E Waelkens, F Chimienti, G A Rutter, P Gilon, P A in't Veld, and F C Schuit. Insulin crystallization depends on zinc transporter ZnT8 expression, but is not required for normal glucose homeostasis in mice. *Proc. Natl. Acad. Sci. U.S.A.*, 106(35):14872–14877, September 2009.

[277] Lynley D Pound, Suparna A Sarkar, Richard K P Benninger, Yingda Wang, Adisak Suwanichkul, Melanie K Shadoan, Richard L Printz, James K Oeser, Catherine E Lee, David W Piston, Owen P McGuinness, John C Hutton, David R Powell, and Richard M O'Brien. Deletion of the mouse Slc30a8 gene encoding zinc transporter-8 results in impaired insulin secretion. *Biochem. J.*, 421(3):371–376, August 2009.

[278] N Wijesekara, F F Dai, A B Hardy, P R Giglou, A Bhattacharjee, V Koshkin, F Chimienti, H Y Gaisano, G A Rutter, and M B Wheeler. Beta cell-specific Znt8 deletion in mice causes marked defects in insulin processing, crystallisation and secretion. *Diabetologia*, 53(8):1656–1668, August 2010.

[279] Lynley D Pound, Suparna A Sarkar, Alessandro Ustione, Prasanna K Dadi, Melanie K Shadoan, Catherine E Lee, Jay A Walters, Masakazu Shiota, Owen P McGuinness, David A Jacobson, David W Piston, John C Hutton, David R Powell, and Richard M O'Brien. The physiological effects of deleting the mouse SLC30A8 gene encoding zinc transporter-8 are influenced by gender and genetic background. *PLoS ONE*, 7(7):e40972, 2012.

[280] A B Hardy, N Wijesekara, I Genkin, K J Prentice, A Bhattacharjee, D Kong, F Chimienti, and M B Wheeler. Effects of high-fat diet feeding on Znt8-null mice: differences between β-cell and global knockout of Znt8. *Am. J. Physiol. Endocrinol. Metab.*, 302(9):E1084–96, May 2012.

[281] Motoyuki Tamaki, Yoshio Fujitani, Akemi Hara, Toyoyoshi Uchida, Yoshifumi Tamura, Kageumi Takeno, Minako Kawaguchi, Takahiro Watanabe, Takeshi Ogihara, Ayako Fukunaka, Tomoaki Shimizu, Tomoya Mita, Akio Kanazawa, Mica O Imaizumi, Takaya Abe, Hiroshi Kiyonari, Shintaro Hojyo, Toshiyuki Fukada, Takeshi Kawauchi, Shinya Nagamatsu, Toshio Hirano, Ryuzo Kawamori, and Hirotaka Watada. The diabetes-susceptible gene

SLC30A8/ZnT8 regulates hepatic insulin clearance. *J. Clin. Invest.*, 123(10):4513–4524, October 2013.

[282] Ida Moltke, Niels Grarup, Marit E Jørgensen, Peter Bjerregaard, Jonas T Treebak, Matteo Fumagalli, Thorfinn S Korneliussen, Marianne A Andersen, Thomas S Nielsen, Nikolaj T Krarup, Anette P Gjesing, Juleen R Zierath, Allan Linneberg, Xueli Wu, Guangqing Sun, Xin Jin, Jumana Al-Aama, Jun Wang, Knut Borch-Johnsen, Oluf Pedersen, Rasmus Nielsen, Anders Albrechtsen, and Torben Hansen. A common Greenlandic TBC1D4 variant confers muscle insulin resistance and type 2 diabetes. *Nature*, 512(7513):190–193, August 2014.

[283] Aparna Pal, Thomas P Potjer, Soren K Thomsen, Hui Jin Ng, Amy Barrett, Raphael Scharfmann, Tim J James, D Timothy Bishop, Fredrik Karpe, Ian F Godsland, Hans F A Vasen, Julia Newton-Bishop, Hanno Pijl, Mark I McCarthy, and Anna L Gloyn. Loss-of-Function Mutations in the Cell-Cycle Control Gene CDKN2A Impact on Glucose Homeostasis in Humans. *Diabetes*, 65(2):527–533, February 2016.

[284] Daniel G MacArthur, Suganthi Balasubramanian, Adam Frankish, Ni Huang, James Morris, Klaudia Walter, Luke Jostins, Lukas Habegger, Joseph K Pickrell, Stephen B Montgomery, Cornelis A Albers, Zhengdong D Zhang, Donald F Conrad, Gerton Lunter, Hancheng Zheng, Qasim Ayub, Mark A DePristo, Eric Banks, Min Hu, Robert E Handsaker, Jeffrey A Rosenfeld, Menachem Fromer, Mike Jin, Xinmeng Jasmine Mu, Ekta Khurana, Kai Ye, Mike Kay, Gary Ian Saunders, Marie-Marthe Suner, Toby Hunt, If H A Barnes, Clara Amid, Denise R Carvalho-Silva, Alexandra H Bignell, Catherine Snow, Bryndis Yngvadottir, Suzannah Bumpstead, David N Cooper, Yali Xue, Irene Gallego Romero, 1000 Genomes Project Consortium, Jun Wang, Yingrui Li, Richard A Gibbs, Steven A McCarroll, Emmanouil T Dermitzakis, Jonathan K Pritchard, Jeffrey C Barrett, Jennifer Harrow, Matthew E Hurles, Mark B Gerstein, and Chris Tyler-Smith. A systematic survey of loss-of-function variants in human protein-coding genes. *Science*, 335(6070):823–828, February 2012.

[285] O P Kallioniemi, A Kallioniemi, D Sudar, D Rutovitz, J W Gray, F Waldman, and D Pinkel. Comparative genomic hybridization: a rapid new method for detecting and mapping DNA amplification in tumors. *Semin. Cancer Biol.*, 4(1):41–46, February 1993.

[286] Nigel P Carter. Methods and strategies for analyzing copy number variation using DNA microarrays. *Nat. Genet.*, 39(7 Suppl):S16–21, July 2007.

[287] C Yau and C C Holmes. CNV discovery using SNP genotyping arrays. *Cytogenet. Genome Res.*, 123(1-4):307–312, 2008.

[288] Haley J Abel and Eric J Duncavage. Detection of structural DNA variation from next generation sequencing data: a review of informatic approaches. *Cancer Genet*, 206(12):432–440, December 2013.

[289] Mehdi Pirooznia, Fernando S Goes, and Peter P Zandi. Whole-genome CNV analysis: advances in computational approaches. *Front Genet*, 6:138, 2015.

[290] Latha Kadalayil, Sajjad Rafiq, Matthew J J Rose-Zerilli, Reuben J Pengelly, Helen Parker, David Oscier, Jonathan C Strefford, William J Tapper, Jane Gibson, Sarah Ennis, and Andrew Collins. Exome sequence read depth methods for identifying copy number changes. *Brief. Bioinformatics*, 16(3):380–392, May 2015.

[291] Robin G Walters, Lachlan J M Coin, Aimo Ruokonen, Adam J de Smith, Julia S El-Sayed Moustafa, Sebastien Jacquemont, Paul Elliott, Tonu Esko, Anna-Liisa Hartikainen, Jaana Laitinen, Katrin Männik, Danielle Martinet, David Meyre, Matthias Nauck, Claudia Schurmann, Rob Sladek, Gudmar Thorleifsson, Unnur Thorsteinsdottir, Armand Valsesia, Gerard Waeber, Flore Zufferey, Beverley Balkau, François Pattou, Andres Metspalu, Henry Völzke, Peter Vollenweider, Kari Stefansson, Marjo-Riitta Järvelin, Jacques S Beckmann, Philippe Froguel, and Alexandra I F Blakemore. Rare genomic structural variants in complex disease: lessons from the replication of associations with obesity. *PLoS ONE*, 8(3):e58048, 2013.

[292] Mario Falchi, Julia Sarah El-Sayed Moustafa, Petros Takousis, Francesco Pesce, Amélie Bonnefond, Johanna C Andersson-Assarsson, Peter H Sudmant, Rajkumar Dorajoo, Mashael Nedham Al-Shafai, Leonardo Bottolo, Erdal Ozdemir, Hon-Cheong So, Robert W Davies, Alexandre Patrice, Robert Dent, Massimo Mangino, Pirro G Hysi, Aurélie Dechaume, Marlène Huyvaert, Jane Skinner, Marie Pigeyre, Robert Caiazzo, Violeta

Raverdy, Emmanuel Vaillant, Sarah Field, Beverley Balkau, Michel Marre, Sophie Visvikis-Siest, Jacques Weill, Odile Poulain-Godefroy, Peter Jacobson, Lars Sjostrom, Christopher J Hammond, Panos Deloukas, Pak Chung Sham, Ruth McPherson, Jeannette Lee, E Shyong Tai, Robert Sladek, Lena M S Carlsson, Andrew Walley, Evan E Eichler, François Pattou, Timothy D Spector, and Philippe Froguel. Low copy number of the salivary amylase gene predisposes to obesity. *Nat. Genet.*, 46(5):492–497, May 2014.

[293] Christina L Usher, Robert E Handsaker, Tonu Esko, Marcus A Tuke, Michael N Weedon, Alex R Hastie, Han Cao, Jennifer E Moon, Seva Kashin, Christian Fuchsberger, Andres Metspalu, Carlos N Pato, Michele T Pato, Mark I McCarthy, Michael Boehnke, David M Altshuler, Timothy M Frayling, Joel N Hirschhorn, and Steven A McCarroll. Structural forms of the human amylase locus and their relationships to SNPs, haplotypes and obesity. *Nat. Genet.*, 47(8):921–925, August 2015.

[294] Menachem Fromer, Jennifer L Moran, Kimberly Chambert, Eric Banks, Sarah E Bergen, Douglas M Ruderfer, Robert E Handsaker, Steven A McCarroll, Michael C O'Donovan, Michael J Owen, George Kirov, Patrick F Sullivan, Christina M Hultman, Pamela Sklar, and Shaun M Purcell. Discovery and statistical genotyping of copy-number variation from whole-exome sequencing depth. *Am. J. Hum. Genet.*, 91(4):597–607, October 2012.

[295] Aaron McKenna, Matthew Hanna, Eric Banks, Andrey Sivachenko, Kristian Cibulskis, Andrew Kernytsky, Kiran Garimella, David Altshuler, Stacey Gabriel, Mark Daly, and Mark A DePristo. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res.*, 20(9):1297–1303, September 2010.

[296] Jennifer Harrow, Adam Frankish, Jose M Gonzalez, Electra Tapanari, Mark Diekhans, Felix Kokocinski, Bronwen L Aken, Daniel Barrell, Amonida Zadissa, Stephen Searle, If Barnes, Alexandra Bignell, Veronika Boychenko, Toby Hunt, Mike Kay, Gaurab Mukherjee, Jeena Rajan, Gloria Despacio-Reyes, Gary Saunders, Charles Steward, Rachel Harte, Michael Lin, Cédric Howald, Andrea Tanzer, Thomas Derrien, Jacqueline Chrast, Nathalie Walters, Suganthi Balasubramanian, Baikang Pei, Michael Tress, Jose Manuel Rodriguez, Iakes Ezkurdia, Jeltje van Baren, Michael Brent, David Haussler, Manolis Kellis, Alfonso Valencia, Alexandre Reymond, Mark Gerstein, Roderic Guigó, and Tim J Hubbard. GENCODE: the reference

human genome annotation for The ENCODE Project. *Genome Res.*, 22(9):1760–1774, September 2012.

[297] Shaun Purcell, Benjamin Neale, Kathe Todd-Brown, Lori Thomas, Manuel A R Ferreira, David Bender, Julian Maller, Pamela Sklar, Paul I W de Bakker, Mark J Daly, and Pak C Sham. PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am. J. Hum. Genet.*, 81(3):559–575, September 2007.

[298] Hyun Min Kang. EPACTS. Efficient and Parallelizable Association Container Toolbox. *http://genome.sph.umich.edu/wiki/EPACTS*, 2016.

[299] Xuefeng Wang. Firth logistic regression for rare variant association tests. *Front Genet*, 5:187, 2014.

[300] Goncalo Abecasis. Metal. *http://csg.sph.umich.edu//abecasis/Metal/*, 2016.

[301] Stephen D Turner. qqman: an R package for visualizing GWAS results using Q-Q and manhattan plots. *bioRxiv*, page 005165, May 2014.

[302] Jeroen R Huyghe, Anne U Jackson, Marie P Fogarty, Martin L Buchkovich, Alena Stančáková, Heather M Stringham, Xueling Sim, Lingyao Yang, Christian Fuchsberger, Henna Cederberg, Peter S Chines, Tanya M Teslovich, Jane M Romm, Hua Ling, Ivy McMullen, Roxann Ingersoll, Elizabeth W Pugh, Kimberly F Doheny, Benjamin M Neale, Mark J Daly, Johanna Kuusisto, Laura J Scott, Hyun Min Kang, Francis S Collins, Goncalo R Abecasis, Richard M Watanabe, Michael Boehnke, Markku Laakso, and Karen L Mohlke. Exome array analysis identifies new loci and low-frequency variants influencing insulin processing and secretion. *Nat. Genet.*, 45(2):197–201, February 2013.

[303] Robert E Handsaker, Joshua M Korn, James Nemesh, and Steven A McCarroll. Discovery and genotyping of genome structural polymorphism by sequencing on a population scale. *Nat. Genet.*, 43(3):269–276, March 2011.

[304] Laura Winchester, Christopher Yau, and Jiannis Ragoussis. Comparing CNV detection methods for SNP arrays. *Brief Funct Genomic Proteomic*, 8(5):353–366, September 2009.

[305] Kai Wang, Mingyao Li, Dexter Hadley, Rui Liu, Joseph Glessner, Struan F A Grant, Hakon Hakonarson, and Maja Bucan. PennCNV: an integrated hidden Markov model designed for high-resolution copy number variation detection in whole-genome SNP genotyping data. *Genome Res.*, 17(11):1665–1674, November 2007.

[306] Stefano Colella, Christopher Yau, Jennifer M Taylor, Ghazala Mirza, Helen Butler, Penny Clouston, Anne S Bassett, Anneke Seller, Christopher C Holmes, and Jiannis Ragoussis. QuantiSNP: an Objective Bayes Hidden-Markov Model to detect and accurately map copy number variation using SNP genotyping data. *Nucleic Acids Res.*, 35(6):2013–2025, 2007.

[307] Adam B Olshen, E S Venkatraman, Robert Lucito, and Michael Wigler. Circular binary segmentation for the analysis of array-based DNA copy number data. *Biostatistics*, 5(4):557–572, October 2004.

[308] Chris Fraley and Adrian E Raftery. Model-Based Clustering, Discriminant Analysis, and Density Estimation. *Journal of the American Statistical Association*, 97(458):611–631, December 2011.

[309] C Fraley and Raftery, A. mclust Version 4 for R: Normal Mixture Modeling for Model-Based Clustering, Classification, and Density Estimation . *Technical Report No 597, Department of Statistics, University of Washington*, 2012.

[310] M Dean and R Allikmets. Evolution of ATP-binding cassette transporter genes. *Curr. Opin. Genet. Dev.*, 5(6):779–785, December 1995.

[311] K Ito, N Ishii, A Miyashita, K Tominaga, H Kuriyama, H Maruyama, M Shirai, M Naito, M Arakawa, and R Kuwano. Molecular cloning of a novel 130-kDa cytoplasmic protein, Ankhzn, containing Ankyrin repeats hooked to a zinc finger motif. *Biochem. Biophys. Res. Commun.*, 257(1):206–213, April 1999.

[312] Jeffrey R MacDonald, Robert Ziman, Ryan K C Yuen, Lars Feuk, and Stephen W Scherer. The Database of Genomic Variants: a curated collection of structural variation in the human genome. *Nucleic Acids Res.*, 42(Database issue):D986–92, January 2014.

[313] Joanna S Amberger, Carol A Bocchini, François Schiettecatte, Alan F Scott, and Ada Hamosh. OMIM.org: Online Mendelian Inheritance in Man (OMIM®), an online catalog of human genes and genetic disorders. *Nucleic Acids Res.*, 43(Database issue):D789–98, January 2015.

[314] Garth R Brown, Vichet Hem, Kenneth S Katz, Michael Ovetsky, Craig Wallin, Olga Ermolaeva, Igor Tolstoy, Tatiana Tatusova, Kim D Pruitt, Donna R Maglott, and Terence D Murphy. Gene: a gene-centered information resource at NCBI. *Nucleic Acids Res.*, 43(Database issue):D36–42, January 2015.

[315] Kenjiro Mitsuhashi, Akihiro Masuda, Yan-Hua Wang, Masayuki Shiseki, and Toshiko Motoji. Prognostic significance of PRAME expression based on immunohistochemistry for diffuse large B-cell lymphoma patients treated with R-CHOP therapy. *Int. J. Hematol.*, 100(1):88–95, July 2014.

[316] Nora Heisterkamp, John Groffen, David Warburton, and Tam P Sneddon. The human gamma-glutamyltransferase gene family. *Hum. Genet.*, 123(4):321–332, May 2008.

[317] James D Weisfeld-Adams, Amanda K Tkachuk, Kenneth N Maclean, Naomi L Meeks, and Stuart A Scott. A de novo 2.78-Mb duplication on chromosome 21q22.11 implicates candidate genes in the partial trisomy 21 phenotype. *npj Genomic Med.*, 1:16003, March 2016.

[318] Gregory James, Simon R Foster, Brian Key, and Annemiek Beverdam. The expression pattern of EVA1C, a novel Slit receptor, is consistent with an axon guidance role in the mouse nervous system. *PLoS ONE*, 8(9):e74115, 2013.

[319] Zheng Chen, Long Yu, Hai Wu, Jianqiang Yu, Lisha Zhang, Daojun Jiang, Lijie Ma, Dan Li, and Shouyuan Zhao. Identification of a novel liver-specific expressed gene, TCP10L, encoding a human leucine zipper protein with transcription inhibition activity. *J. Hum. Genet.*, 48(11):556–563, 2003.

[320] Christina Austin-Tse, Jan Halbritter, Maimoona A Zariwala, Renée M Gilberti, Heon Yung Gee, Nathan Hellman, Narendra Pathak, Yan Liu, Jennifer R Panizzi, Ramila S Patel-King, Douglas Tritschler, Raqual Bower, Eileen O'Toole, Jonathan D Porath, Toby W Hurd, Moumita Chaki, Katrina A Diaz, Stefan Kohl, Svjetlana Lovric, Daw-Yang Hwang,

Daniela A Braun, Markus Schueler, Rannar Airik, Edgar A Otto, Margaret W Leigh, Peadar G Noone, Johnny L Carson, Stephanie D Davis, Jessica E Pittman, Thomas W Ferkol, Jeffry J Atkinson, Kenneth N Olivier, Scott D Sagel, Sharon D Dell, Margaret Rosenfeld, Carlos E Milla, Niki T Loges, Heymut Omran, Mary E Porter, Stephen M King, Michael R Knowles, Iain A Drummond, and Friedhelm Hildebrandt. Zebrafish Ciliopathy Screen Plus Human Mutational Analysis Identifies C21orf59 and CCDC65 Defects as Causing Primary Ciliary Dyskinesia. *Am. J. Hum. Genet.*, 93(4):672–686, October 2013.

[321] M A Pujana, M Nadal, M Gratacòs, B Peral, K Csiszar, R González-Sarmiento, L Sumoy, and X Estivill. Additional complexity on human chromosome 15q: identification of a set of newly recognized duplicons (LCR15) on 15q11-q13, 15q24, and 15q26. *Genome Res.*, 11(1):98–111, January 2001.

[322] D Peter, J P Finn, I Klisak, Y Liu, T Kojis, C Heinzmann, A Roghani, R S Sparkes, and R H Edwards. Chromosomal localization of the human vesicular amine transporter genes. *Genomics*, 18(3):720–723, December 1993.

[323] Mi Seul Park, Eduard Bitto, Kyung Rok Kim, Craig A Bingman, Mitchell D Miller, Hyun-Jung Kim, Byung Woo Han, and George N Phillips. Crystal structure of human protein N-terminal glutamine amidohydrolase, an initial component of the N-end rule pathway. *PLoS ONE*, 9(10):e111142, 2014.

[324] Yuriko Katoh and Masaru Katoh. Comparative genomics on HHIP family orthologs. *Int. J. Mol. Med.*, 17(2):391–395, February 2006.

[325] Christoph Heiring, Björn Dahlbäck, and Yves A Muller. Ligand recognition and homophilic interactions in Tyro3: structural insights into the Axl/Tyro3 receptor tyrosine kinase family. *J. Biol. Chem.*, 279(8):6952–6958, February 2004.

[326] Richa Saxena, Marie-France Hivert, Claudia Langenberg, Toshiko Tanaka, James S Pankow, Peter Vollenweider, Valeriya Lyssenko, Nabila Bouatia-Naji, Josée Dupuis, Anne U Jackson, W H Linda Kao, Man Li, Nicole L Glazer, Alisa K Manning, Jian'an Luan, Heather M Stringham, Inga Prokopenko, Toby Johnson, Niels Grarup, Trine W Boesgaard, Cécile Lecoeur, Peter Shrader, Jeffrey O'Connell, Erik Ingelsson, David J Couper, Kenneth Rice,

Kijoung Song, Camilla H Andreasen, Christian Dina, Anna Köttgen, Olivier Le Bacquer, François Pattou, Jalal Taneera, Valgerdur Steinthorsdottir, Denis Rybin, Kristin Ardlie, Michael Sampson, Lu Qi, Mandy van Hoek, Michael N Weedon, Yurii S Aulchenko, Benjamin F Voight, Harald Grallert, Beverley Balkau, Richard N Bergman, Suzette J Bielinski, Amélie Bonnefond, Lori L Bonnycastle, Knut Borch-Johnsen, Yvonne Böttcher, Eric Brunner, Thomas A Buchanan, Suzannah J Bumpstead, Christine Cavalcanti-Proença, Guillaume Charpentier, Yii-Der Ida Chen, Peter S Chines, Francis S Collins, Marilyn Cornelis, Gabriel J Crawford, Jerome Delplanque, Alex Doney, Josephine M Egan, Michael R Erdos, Mathieu Firmann, Nita G Forouhi, Caroline S Fox, Mark O Goodarzi, Jürgen Graessler, Aroon Hingorani, Bo Isomaa, Torben Jørgensen, Mika Kivimaki, Peter Kovacs, Knut Krohn, Meena Kumari, Torsten Lauritzen, Claire Lévy-Marchal, Vladimir Mayor, Jarred B McAteer, David Meyre, Braxton D Mitchell, Karen L Mohlke, Mario A Morken, Narisu Narisu, Colin N A Palmer, Ruth Pakyz, Laura Pascoe, Felicity Payne, Daniel Pearson, Wolfgang Rathmann, Annelli Sandbaek, Avan Aihie Sayer, Laura J Scott, Stephen J Sharp, Eric Sijbrands, Andrew Singleton, David S Siscovick, Nicholas L Smith, Thomas Sparsø, Amy J Swift, Holly Syddall, Gudmar Thorleifsson, Anke Tönjes, Tiinamaija Tuomi, Jaakko Tuomilehto, Timo T Valle, Gerard Waeber, Andrew Walley, Dawn M Waterworth, Eleftheria Zeggini, Jing Hua Zhao, GIANT Consortium, MAGIC investigators, Thomas Illig, H-Erich Wichmann, James F Wilson, Cornelia van Duijn, Frank B Hu, Andrew D Morris, Timothy M Frayling, Andrew T Hattersley, Unnur Thorsteinsdottir, Kari Stefansson, Peter Nilsson, Ann-Christine Syvänen, Alan R Shuldiner, Mark Walker, Stefan R Bornstein, Peter Schwarz, Gordon H Williams, David M Nathan, Johanna Kuusisto, Markku Laakso, Cyrus Cooper, Michael Marmot, Luigi Ferrucci, Vincent Mooser, Michael Stumvoll, Ruth J F Loos, David Altshuler, Bruce M Psaty, Jerome I Rotter, Eric Boerwinkle, Torben Hansen, Oluf Pedersen, Jose C Florez, Mark I McCarthy, Michael Boehnke, Ines Barroso, Robert Sladek, Philippe Froguel, James B Meigs, Leif Groop, Nicholas J Wareham, and Richard M Watanabe. Genetic variation in GIPR influences the glucose and insulin responses to an oral glucose challenge. *Nat. Genet.*, 42(2):142–148, February 2010.

[327] Rona J Strawbridge, Josée Dupuis, Inga Prokopenko, Adam Barker, Emma Ahlqvist, Denis Rybin, John R Petrie, Mary E Travers, Nabila Bouatia-Naji, Antigone S Dimas, Alexandra

Nica, Eleanor Wheeler, Han Chen, Benjamin F Voight, Jalal Taneera, Stavroula Kanoni, John F Peden, Fabiola Turrini, Stefan Gustafsson, Carina Zabena, Peter Almgren, David J P Barker, Daniel Barnes, Elaine M Dennison, Johan G Eriksson, Per Eriksson, Elodie Eury, Lasse Folkersen, Caroline S Fox, Timothy M Frayling, Anuj Goel, Harvest F Gu, Momoko Horikoshi, Bo Isomaa, Anne U Jackson, Karen A Jameson, Eero Kajantie, Julie Kerr-Conte, Teemu Kuulasmaa, Johanna Kuusisto, Ruth J F Loos, Jian'an Luan, Konstantinos Makrilakis, Alisa K Manning, María Teresa Martínez-Larrad, Narisu Narisu, Maria Nastase Mannila, John Ohrvik, Clive Osmond, Laura Pascoe, Felicity Payne, Avan A Sayer, Bengt Sennblad, Angela Silveira, Alena Stančáková, Kathy Stirrups, Amy J Swift, Ann-Christine Syvänen, Tiinamaija Tuomi, Ferdinand M van 't Hooft, Mark Walker, Michael N Weedon, Weijia Xie, Björn Zethelius, DIAGRAM Consortium, GIANT Consortium, MuTHER Consortium, CARDIoGRAM Consortium, C4D Consortium, Halit Ongen, Anders Mälarstig, Jemma C Hopewell, Danish Saleheen, John Chambers, Sarah Parish, John Danesh, Jaspal Kooner, Claes-Göran Ostenson, Lars Lind, Cyrus C Cooper, Manuel Serrano-Ríos, Ele Ferrannini, Tom J Forsen, Robert Clarke, Maria Grazia Franzosi, Udo Seedorf, Hugh Watkins, Philippe Froguel, Paul Johnson, Panos Deloukas, Francis S Collins, Markku Laakso, Emmanouil T Dermitzakis, Michael Boehnke, Mark I McCarthy, Nicholas J Wareham, Leif Groop, François Pattou, Anna L Gloyn, George V Dedoussis, Valeriya Lyssenko, James B Meigs, Ines Barroso, Richard M Watanabe, Erik Ingelsson, Claudia Langenberg, Anders Hamsten, and Jose C Florez. Genome-wide association identifies nine common variants associated with fasting proinsulin levels and provides new insights into the pathophysiology of type 2 diabetes. *Diabetes*, 60(10):2624–2634, October 2011.

[328] Zenobia B Mehta, Nicholas Fine, Timothy J Pullen, Matthew C Cane, Ming Hu, Pauline Chabosseau, Gargi Meur, Antonio Velayos-Baeza, Anthony P Monaco, Lorella Marselli, Piero Marchetti, and Guy A Rutter. Changes in the expression of the type 2 diabetes-associated gene VPS13C in the β-cell are associated with glucose intolerance in humans and mice. *Am. J. Physiol. Endocrinol. Metab.*, 311(2):E488–507, August 2016.

[329] Tatsuya Iso, Larry Kedes, and Yasuo Hamamori. HES and HERP families: multiple effectors of the Notch signaling pathway. *J. Cell. Physiol.*, 194(3):237–255, March 2003.

[330] Andreas Fischer, Christian Steidl, Toni U Wagner, Esra Lang, Peter M Jakob, Peter Friedl, Klaus-Peter Knobeloch, and Manfred Gessler. Combined loss of Hey1 and HeyL causes congenital heart defects because of impaired epithelial to mesenchymal transition. *Circ. Res.*, 100(6):856–863, March 2007.

[331] Atsushi Tanabe, Junpei Konno, Kenya Tanikawa, and Hiroeki Sahara. Transcriptional machinery of TNF-α-inducible YTH domain containing 2 (YTHDC2) gene. *Gene*, 535(1):24–32, February 2014.

[332] M M Keane, S A Ettenberg, M M Nau, P Banerjee, M Cuello, J Penninger, and S Lipkowitz. cbl-3: a new mammalian cbl family protein. *Oncogene*, 18(22):3365–3375, June 1999.

[333] Wan Yin Lee, Germaine Goh, Joanne Chia, Adrian Boey, Natalia V Gunko, and Frederic Bard. The Ubiquitin Ligase CBLC Maintains the Network Organization of the Golgi Apparatus. *PLoS ONE*, 10(9):e0138789, 2015.

[334] Daniele Fanale, Juan Lucio Iovanna, Ezequiel Luis Calvo, Patrice Berthezene, Pascal Belleau, Jean Charles Dagorn, Giuseppe Bronte, Giuseppe Cicero, Viviana Bazan, Christian Rolfo, Daniele Santini, and Antonio Russo. Germline copy number variation in the YTHDC2 gene: does it have a role in finding a novel potential molecular target involved in pancreatic adenocarcinoma susceptibility? *Expert Opin. Ther. Targets*, 18(8):841–850, August 2014.

[335] T Greener, X Zhao, H Nojima, E Eisenberg, and L E Greene. Role of cyclin G-associated kinase in uncoating clathrin-coated vesicles from non-neuronal cells. *J. Biol. Chem.*, 275(2):1365–1370, January 2000.

[336] Ze-Gang Ma, Feng He, and Jian Xu. Quantitative assessment of the association between GAK rs1564282 C/T polymorphism and the risk of Parkinson's disease. *J Clin Neurosci*, 22(7):1077–1080, July 2015.

[337] Elena G Bochukova, Ni Huang, Julia Keogh, Elana Henning, Carolin Purmann, Kasia Blaszczyk, Sadia Saeed, Julian Hamilton-Shield, Jill Clayton-Smith, Stephen O'Rahilly, Matthew E Hurles, and I Sadaf Farooqi. Large, rare chromosomal deletions associated with severe early-onset obesity. *Nature*, 463(7281):666–670, February 2010.

[338] Christopher J Ward, Marie C Hogan, Sandro Rossetti, Denise Walker, Tam Sneddon, Xiaofang Wang, Vicky Kubly, Julie M Cunningham, Robert Bacallao, Masahiko Ishibashi, Dawn S Milliner, Vicente E Torres, and Peter C Harris. The gene mutated in autosomal recessive polycystic kidney disease encodes a large, receptor-like protein. *Nat. Genet.*, 30(3):259–269, March 2002.

[339] Dan Liu, Ming Shi, Hao Zhang, Lu Qian, Ming Yu, Meiru Hu, Ruihong Zhang, Tianyou Wang, Caili Han, Huijun Duan, and Ning Guo. c-Myb regulates cell cycle-dependent expression of Erbin: an implication for a novel function of Erbin. *PLoS ONE*, 7(8):e42903, 2012.

[340] Inbal Rachmin, Sagi Tshori, Yoav Smith, Amit Oppenheim, Sylvie Marchetto, Gillian Kay, Roger S-Y Foo, Noa Dagan, Eliahu Golomb, Dan Gilon, Jean-Paul Borg, and Ehud Razin. Erbin is a negative modulator of cardiac hypertrophy. *Proc. Natl. Acad. Sci. U.S.A.*, 111(16):5902–5907, April 2014.

[341] Francisco Castaneda, Jennifer E Layne, and Carmen Castaneda. Skeletal muscle sodium glucose co-transporters in older adults with type 2 diabetes undergoing resistance training. *Int J Med Sci*, 3(3):84–91, 2006.

[342] Amar A Patil, Yibin Cai, Yongming Sang, Frank Blecha, and Guolong Zhang. Cross-species analysis of the mammalian beta-defensin gene family: presence of syntenic gene clusters and preferential expression in the male reproductive tract. *Physiol. Genomics*, 23(1):5–17, September 2005.

[343] Ellen M van Beek, Fiona Cochrane, A Neil Barclay, and Timo K van den Berg. Signal regulatory proteins in the immune system. *J. Immunol.*, 175(12):7781–7787, December 2005.

[344] S Grandemange, S Schaller, S Yamano, S Du Manoir, G V Shpakovski, M G Mattei, C Kedinger, and M Vigneron. A human RNA polymerase II subunit is encoded by a recently generated multigene family. *BMC Mol. Biol.*, 2(1):14, 2001.

[345] Mohammad M Ghahramani Seno, Benjamin Y M Kwan, Ka Ki M Lee-Ng, Rainald Moessner, Anath C Lionel, Christian R Marshall, and Stephen W Scherer. Human PTCHD3 nulls:

rare copy number and sequence variants suggest a non-essential gene. *BMC Med. Genet.*, 12(1):45, 2011.

[346] Mu He, Radhika Subramanian, Fiona Bangs, Tatiana Omelchenko, Karel F Liem, Tarun M Kapoor, and Kathryn V Anderson. The kinesin-4 protein Kif7 regulates mammalian Hedgehog signalling by organizing the cilium tip compartment. *Nat. Cell Biol.*, 16(7):663–672, July 2014.

[347] Sarah C Goetz and Kathryn V Anderson. The primary cilium: a signalling centre during vertebrate development. *Nat. Rev. Genet.*, 11(5):331–344, May 2010.

[348] Setsu Endoh-Yamagami, Marie Evangelista, Deanna Wilson, Xiaohui Wen, Jan-Willem Theunissen, Khanhky Phamluong, Matti Davis, Suzie J Scales, Mark J Solloway, Frederic J de Sauvage, and Andrew S Peterson. The mammalian Cos2 homolog Kif7 plays an essential role in modulating Hh signal transduction during development. *Curr. Biol.*, 19(15):1320–1326, August 2009.

[349] Ching-Shyi Wu and Lee Zou. The SUMO (Small Ubiquitin-like Modifier) Ligase PIAS3 Primes ATR for Checkpoint Activation. *J. Biol. Chem.*, 291(1):279–290, January 2016.

[350] Keyong Du, Stephan Herzig, Rohit N Kulkarni, and Marc Montminy. TRB3: A tribbles Homolog That Inhibits Akt/PKB Activation by Insulin in Liver. *Science*, 300(5625):1574–1577, June 2003.

[351] Suhani H Almal and Harish Padh. Implications of gene copy-number variation in health and diseases. *J. Hum. Genet.*, 57(1):6–13, January 2012.

[352] Kai Wang, Wei-Dong Li, Joseph T Glessner, Struan F A Grant, Hakon Hakonarson, and R Arlen Price. Large copy-number variations are enriched in cases with moderate to extreme obesity. *Diabetes*, 59(10):2690–2694, October 2010.

[353] Joan C Han, Qing-Rong Liu, MaryPat Jones, Rebecca L Levinn, Carolyn M Menzie, Kyra S Jefferson-George, Diane C Adler-Wailes, Ethan L Sanford, Felicitas L Lacbawan, George R Uhl, Owen M Rennert, and Jack A Yanovski. Brain-derived neurotrophic factor and obesity in the WAGR syndrome. *N. Engl. J. Med.*, 359(9):918–927, August 2008.

[354] Heun-Sik Lee, Sanghoon Moon, Jun Ho Yun, MeeHee Lee, Mi Yeong Hwang, Young-Jin Kim, Bok-Ghee Han, Jeong-Min Kim, and Bong-Jo Kim. Genome-wide copy number variation study reveals KCNIP1 as a modulator of insulin secretion. *Genomics*, 104(2):113–120, August 2014.

[355] Manasa Prabhanjan, Raviraj V Suresh, Megha N Murthy, and Nallur B Ramachandra. Type 2 diabetes mellitus disease risk genes identified by genome wide copy number variation scan in normal populations. *Diabetes Res. Clin. Pract.*, 113:160–170, March 2016.

[356] T W Boesgaard, N Grarup, T Jørgensen, K Borch-Johnsen, Meta-Analysis of Glucose and Insulin-Related Trait Consortium (MAGIC), T Hansen, and O Pedersen. Variants at DGKB/TMEM195, ADRA2A, GLIS3 and C2CD4B loci are associated with reduced glucose-stimulated beta cell function in middle-aged Danish people. *Diabetologia*, 53(8):1647–1655, August 2010.

[357] N Grarup, M Overvad, T Sparsø, D R Witte, C Pisinger, T Jørgensen, T Yamauchi, K Hara, S Maeda, T Kadowaki, T Hansen, and O Pedersen. The diabetogenic VPS13C/C2CD4A/C2CD4B rs7172432 variant impairs glucose-stimulated insulin response in 5,722 non-diabetic Danish individuals. *Diabetologia*, 54(4):789–794, April 2011.

[358] Erik Ingelsson, Claudia Langenberg, Marie-France Hivert, Inga Prokopenko, Valeriya Lyssenko, Josée Dupuis, Reedik Mägi, Stephen Sharp, Anne U Jackson, Themistocles L Assimes, Peter Shrader, Joshua W Knowles, Björn Zethelius, Fahim A Abbasi, Richard N Bergman, Antje Bergmann, Christian Berne, Michael Boehnke, Lori L Bonnycastle, Stefan R Bornstein, Thomas A Buchanan, Suzannah J Bumpstead, Yvonne Böttcher, Peter Chines, Francis S Collins, Cyrus C Cooper, Elaine M Dennison, Michael R Erdos, Ele Ferrannini, Caroline S Fox, Jürgen Graessler, Ke Hao, Bo Isomaa, Karen A Jameson, Peter Kovacs, Johanna Kuusisto, Markku Laakso, Claes Ladenvall, Karen L Mohlke, Mario A Morken, Narisu Narisu, David M Nathan, Laura Pascoe, Felicity Payne, John R Petrie, Avan A Sayer, Peter E H Schwarz, Laura J Scott, Heather M Stringham, Michael Stumvoll, Amy J Swift, Ann-Christine Syvänen, Tiinamaija Tuomi, Jaakko Tuomilehto, Anke Tönjes, Timo T Valle, Gordon H Williams, Lars Lind, Ines Barroso, Thomas Quertermous, Mark Walker, Nicholas J Wareham, James B Meigs, Mark I McCarthy, Leif Groop, Richard M

Watanabe, Jose C Florez, and MAGIC investigators. Detailed physiologic characterization reveals diverse mechanisms for novel genetic Loci regulating glucose and insulin metabolism in humans. *Diabetes*, 59(5):1266–1275, May 2010.

[359] Jan Windholz, Peter Kovacs, Anke Tönjes, Kathrin Dittrich, Susann Blüher, Wieland Kiess, Michael Stumvoll, and Antje Körner. Effects of genetic variants in ADCY5, GIPR, GCKR and VPS13C on early impairment of glucose and insulin metabolism in children. *PLoS ONE*, 6(7):e22101, 2011.

[360] Toshimasa Yamauchi, Kazuo Hara, Shiro Maeda, Kazuki Yasuda, Atsushi Takahashi, Momoko Horikoshi, Masahiro Nakamura, Hayato Fujita, Niels Grarup, Stephane Cauchi, Daniel P K Ng, Ronald C W Ma, Tatsuhiko Tsunoda, Michiaki Kubo, Hirotaka Watada, Hiroshi Maegawa, Miki Okada-Iwabu, Masato Iwabu, Nobuhiro Shojima, Hyoung Doo Shin, Gitte Andersen, Daniel R Witte, Torben Jørgensen, Torsten Lauritzen, Annelli Sandbaek, Torben Hansen, Toshihiko Ohshige, Shintaro Omori, Ikuo Saito, Kohei Kaku, Hiroshi Hirose, Wing Yee So, Delphine Beury, Juliana C N Chan, Kyong Soo Park, E Shyong Tai, Chikako Ito, Yasushi Tanaka, Atsunori Kashiwagi, Ryuzo Kawamori, Masato Kasuga, Philippe Froguel, Oluf Pedersen, Naoyuki Kamatani, Yusuke Nakamura, and Takashi Kadowaki. A genome-wide association study in the Japanese population identifies susceptibility loci for type 2 diabetes at UBE2E2 and C2CD4A-C2CD4B. *Nat. Genet.*, 42(10):864–868, October 2010.

[361] L Rampoldi, C Dobson-Stone, J P Rubio, A Danek, R M Chalmers, N W Wood, C Verellen, X Ferrer, A Malandrini, G M Fabrizi, R Brown, J Vance, M Pericak-Vance, G Rudolf, S Carrè, E Alonso, M Manfredi, A H Németh, and A P Monaco. A conserved sorting-associated protein is mutant in chorea-acanthocytosis. *Nat. Genet.*, 28(2):119–120, June 2001.

[362] Jae-Sook Park, Simon Halegoua, Shosei Kishida, and Aaron M Neiman. A conserved function in phosphatidylinositol metabolism for mammalian Vps13 family proteins. *PLoS ONE*, 10(4):e0124836, 2015.

[363] Floriane Limoge, Laurence Faivre, Thomas Gautier, Jean-Michel Petit, Elodie Gautier, David Masson, Gaëtan Jego, Salima El Chehadeh-Djebbar, Nathalie Marle, Virginie

Carmignac, Valérie Deckert, Marie-Claude Brindisi, Patrick Edery, Jamal Ghoumid, Edward Blair, Laurent Lagrost, Christel Thauvin-Robinet, and Laurence Duplomb. Insulin response dysregulation explains abnormal fat storage and increased risk of diabetes mellitus type 2 in Cohen Syndrome. *Hum. Mol. Genet.*, 24(23):6603–6613, December 2015.

[364] Ana Diez-Sampedro, Bruce A Hirayama, Christina Osswald, Valentin Gorboulev, Katharina Baumgarten, Christopher Volk, Ernest M Wright, and Hermann Koepsell. A glucose sensor hiding in a family of transporters. *Proc. Natl. Acad. Sci. U.S.A.*, 100(20):11753–11758, September 2003.

[365] G Mithieux. Metabolic effects of portal vein sensing. *Diabetes Obes Metab*, 16 Suppl 1(S1):56–60, September 2014.

[366] Adrian Bird. DNA methylation patterns and epigenetic memory. *Genes Dev.*, 16(1):6–21, January 2002.

[367] Angelika Feldmann, Robert Ivanek, Rabih Murr, Dimos Gaidatzis, Lukas Burger, and Dirk Schübeler. Transcription factor occupancy can mediate active turnover of DNA methylation at regulatory regions. *PLoS Genet.*, 9(12):e1003994, 2013.

[368] Emily Hodges, Antoine Molaro, Camila O Dos Santos, Pramod Thekkat, Qiang Song, Philip J Uren, Jin Park, Jason Butler, Shahin Rafii, W Richard McCombie, Andrew D Smith, and Gregory J Hannon. Directional DNA methylation changes and complex intermediate states accompany lineage specificity in the adult hematopoietic compartment. *Mol. Cell*, 44(1):17–28, October 2011.

[369] Gary C Hon, Nisha Rajagopal, Yin Shen, David F McCleary, Feng Yue, My D Dang, and Bing Ren. Epigenetic memory at embryonic enhancers identified in DNA methylation maps from adult mouse tissues. *Nat. Genet.*, 45(10):1198–1206, October 2013.

[370] Hong Hu, Jingting Xu, and Yang Dai. Regulatory elements in low-methylated regions predict directional change of gene expression. *IEEE J Biomed Health Inform*, 19(4):1293–1300, July 2015.

[371] J P Szatkiewicz, C O'Dushlaine, G Chen, K Chambert, J L Moran, B M Neale, M Fromer, D Ruderfer, S Akterin, S E Bergen, A Kähler, P K E Magnusson, Y Kim, J J Crowley,

E Rees, G Kirov, M C O'Donovan, M J Owen, J Walters, E Scolnick, P Sklar, S Purcell, C M Hultman, S A McCarroll, and P F Sullivan. Copy number variation in schizophrenia in Sweden. *Mol. Psychiatry*, 19(7):762–773, July 2014.

[372] Brian Hon-Yin Chung, Victoria Qinchen Tao, and Winnie Wan-Yee Tso. Copy number variation and autism: new insights and clinical implications. *J. Formos. Med. Assoc.*, 113(7):400–408, July 2014.

[373] J P Szatkiewicz, B M Neale, C O'Dushlaine, M Fromer, J I Goldstein, J L Moran, K Chambert, A Kähler, P K E Magnusson, C M Hultman, P Sklar, S Purcell, S A McCarroll, and P F Sullivan. Detecting large copy number variants using exome genotyping arrays in a large Swedish schizophrenia sample. *Mol. Psychiatry*, 18(11):1178–1184, November 2013.

[374] Melina Claussnitzer, Simon N Dankel, Kyoung-Han Kim, Gerald Quon, Wouter Meuleman, Christine Haugen, Viktoria Glunk, Isabel S Sousa, Jacqueline L Beaudry, Vijitha Puviindran, Nezar A Abdennur, Jannel Liu, Per-Arne Svensson, Yi-Hsiang Hsu, Daniel J Drucker, Gunnar Mellgren, Chi-Chung Hui, Hans Hauner, and Manolis Kellis. FTO Obesity Variant Circuitry and Adipocyte Browning in Humans. *N. Engl. J. Med.*, 373(10):895–907, September 2015.

[375] Tammy Sadler, Jeffrey M Bhasin, Yaomin Xu, Jill Barnholz-Sloan, Yanwen Chen, Angela H Ting, and Eleni Stylianou. Genome-wide analysis of DNA methylation and gene expression defines molecular characteristics of Crohn's disease-associated fibrosis. *Clin Epigenetics*, 8(1):30, 2016.

[376] Verity F Oliver, Andrew E Jaffe, Jin Song, Guohua Wang, Pingwu Zhang, Kari E Branham, Anand Swaroop, Charles G Eberhart, Donald J Zack, Jiang Qian, and Shannath L Merbs. Differential DNA methylation identified in the blood and retina of AMD patients. *Epigenetics*, 10(8):698–707, 2015.

[377] A Córdova-Palomera, M Fatjó-Vilas, C Gastó, V Navarro, M-O Krebs, and L Fañanás. Genome-wide methylation study on depression: differential methylation and variable methylation in monozygotic twins. *Transl Psychiatry*, 5(4):e557, 2015.

[378] María Luisa Mansego, Fermín I Milagro, María Ángeles Zulet, María J Moreno-Aliaga, and José Alfredo Martínez. Differential DNA Methylation in Relation to Age and Health Risks of Obesity. *Int J Mol Sci*, 16(8):16816–16832, 2015.

[379] Fiona Allum, Xiaojian Shao, Frédéric Guénard, Marie-Michelle Simon, Stephan Busche, Maxime Caron, John Lambourne, Julie Lessard, Karolina Tandre, Åsa K Hedman, Tony Kwan, Bing Ge, Multiple Tissue Human Expression Resource Consortium, Lars Rönnblom, Mark I McCarthy, Panos Deloukas, Todd Richmond, Daniel Burgess, Timothy D Spector, André Tchernof, Simon Marceau, Mark Lathrop, Marie-Claude Vohl, Tomi Pastinen, and Elin Grundberg. Characterization of functional methylomes by next-generation capture sequencing identifies novel disease-associated variants. *Nat Commun*, 6:7211, 2015.

[380] L F Wockner, E P Noble, B R Lawford, R McD Young, C P Morris, V L J Whitehall, and J Voisey. Genome-wide DNA methylation analysis of human brain tissue from schizophrenia patients. *Transl Psychiatry*, 4(1):e339, 2014.

[381] Xiaojing Xu, Shaoyong Su, Vernon A Barnes, Carmen De Miguel, Jennifer Pollock, Dennis Ownby, Hidong Shi, Haidong Zhu, Harold Snieder, and Xiaoling Wang. A genome-wide methylation study on obesity: differential variability and differential methylation. *Epigenetics*, 8(5):522–533, May 2013.

[382] Yun Liu, Martin J Aryee, Leonid Padyukov, M Daniele Fallin, Espen Hesselberg, Arni Runarsson, Lovisa Reinius, Nathalie Acevedo, Margaret Taub, Marcus Ronninger, Klementy Shchetynsky, Annika Scheynius, Juha Kere, Lars Alfredsson, Lars Klareskog, Tomas J Ekström, and Andrew P Feinberg. Epigenome-wide association data implicate DNA methylation as an intermediary of genetic risk in rheumatoid arthritis. *Nat. Biotechnol.*, 31(2):142–147, February 2013.

[383] Miriam Cnop, Baroj Abdulkarim, Guy Bottu, Daniel A Cunha, Mariana Igoillo-Esteve, Matilde Masini, Jean-Valery Turatsinze, Thasso Griebel, Olatz Villate, Izortze Santin, Marco Bugliani, Laurence Ladriere, Lorella Marselli, Mark I McCarthy, Piero Marchetti, Michael Sammeth, and Decio L Eizirik. RNA sequencing identifies dysregulation of the human pancreatic islet transcriptome by the saturated fatty acid palmitate. *Diabetes*, 63(6):1978–1993, June 2014.

[384] Alexander Gusev, Arthur Ko, Huwenbo Shi, Gaurav Bhatia, Wonil Chung, Brenda W J H Penninx, Rick Jansen, Eco J C de Geus, Dorret I Boomsma, Fred A Wright, Patrick F Sullivan, Elina Nikkola, Marcus Alvarez, Mete Civelek, Aldons J Lusis, Terho Lehtimäki, Emma Raitoharju, Mika Kähönen, Ilkka Seppälä, Olli T Raitakari, Johanna Kuusisto, Markku Laakso, Alkes L Price, Päivi Pajukanta, and Bogdan Pasaniuc. Integrative approaches for large-scale transcriptome-wide association studies. *Nat. Genet.*, 48(3):245–252, March 2016.

[385] Elin Hall, Petr Volkov, Tasnim Dayeh, Karl Bacos, Tina Rönn, Marloes Dekker Nitert, and Charlotte Ling. Effects of palmitate on genome-wide mRNA expression and DNA methylation patterns in human pancreatic islets. *BMC Med*, 12(1):103, 2014.

[386] Matthias Farlik, Nathan C Sheffield, Angelo Nuzzo, Paul Datlinger, Andreas Schönegger, Johanna Klughammer, and Christoph Bock. Single-cell DNA methylome sequencing and bioinformatic inference of epigenomic cell-state dynamics. *Cell Rep*, 10(8):1386–1397, March 2015.

[387] Jason D Buenrostro, Beijing Wu, Ulrike M Litzenburger, Dave Ruff, Michael L Gonzales, Michael P Snyder, Howard Y Chang, and William J Greenleaf. Single-cell chromatin accessibility reveals principles of regulatory variation. *Nature*, 523(7561):486–490, July 2015.

[388] Jin Li, Johanna Klughammer, Matthias Farlik, Thomas Penz, Andreas Spittler, Charlotte Barbieux, Ekaterine Berishvili, Christoph Bock, and Stefan Kubicek. Single-cell transcriptomes reveal characteristic features of human pancreatic islet cell types. *EMBO Rep.*, 17(2):178–187, February 2016.

[389] Yue J Wang, Jonathan Schug, Kyoung-Jae Won, Chengyang Liu, Ali Naji, Dana Avrahami, Maria L Golson, and Klaus H Kaestner. Single cell transcriptomics of the human endocrine pancreas. *Diabetes*, page db160405, June 2016.

[390] Le Cong, F Ann Ran, David Cox, Shuailiang Lin, Robert Barretto, Naomi Habib, Patrick D Hsu, Xuebing Wu, Wenyan Jiang, Luciano A Marraffini, and Feng Zhang. Multiplex genome engineering using CRISPR/Cas systems. *Science*, 339(6121):819–823, February 2013.

[391] Prashant Mali, Luhan Yang, Kevin M Esvelt, John Aach, Marc Guell, James E DiCarlo,

Julie E Norville, and George M Church. RNA-guided human genome engineering via Cas9. *Science*, 339(6121):823–826, February 2013.

[392] Aleksandar Vojta, Paula Dobrinić, Vanja Tadić, Luka Bočkor, Petra Korać, Boris Julg, Marija Klasić, and Vlatka Zoldoš. Repurposing the CRISPR-Cas9 system for targeted DNA methylation. *Nucleic Acids Res.*, 44(12):5615–5628, July 2016.

[393] Xingxing Xu, Yonghui Tao, Xiaobo Gao, Lei Zhang, Xufang Li, Weiguo Zou, Kangcheng Ruan, Feng Wang, Guo-Liang Xu, and Ronggui Hu. A CRISPR-based approach for targeted DNA demethylation. *Cell Discov*, 2:16009, 2016.

[394] Benjamin J Hindson, Kevin D Ness, Donald A Masquelier, Phillip Belgrader, Nicholas J Heredia, Anthony J Makarewicz, Isaac J Bright, Michael Y Lucero, Amy L Hiddessen, Tina C Legler, Tyler K Kitano, Michael R Hodel, Jonathan F Petersen, Paul W Wyatt, Erin R Steenblock, Pallavi H Shah, Luc J Bousse, Camille B Troup, Jeffrey C Mellen, Dean K Wittmann, Nicholas G Erndt, Thomas H Cauley, Ryan T Koehler, Austin P So, Simant Dube, Klint A Rose, Luz Montesclaros, Shenglong Wang, David P Stumbo, Shawn P Hodges, Steven Romine, Fred P Milanovich, Helen E White, John F Regan, George A Karlin-Neumann, Christopher M Hindson, Serge Saxonov, and Bill W Colston. High-throughput droplet digital PCR system for absolute quantitation of DNA copy number. *Anal. Chem.*, 83(22):8604–8610, November 2011.

[395] Erica Mazaika and Jason Homsy. Digital Droplet PCR: CNV Analysis and Other Applications. *Curr Protoc Hum Genet*, 82:7.24.1–13, 2014.