Transcriptional mechanisms influencing glycaemic traits and risk of type 2 diabetes

A thesis submitted to the University of Oxford



in candidature for the degree of Doctor of Philosophy
by

Anthony Joseph Payne

St John's College, University of Oxford
and
Wellcome Trust Centre for Human Genetics
Nuffield Department of Medicine

Trinity 2018

Declaration

This thesis is the original work of the author, and all analyses were performed using servers and infrastructure both from the McCarthy group and the Research Computing group at the Wellcome Trust Centre for Human Genetics in Oxford. DPhil funding was provided by the Rhodes Trust, the Natural Sciences and Engineering Research Council of Canada, and the Canadian Centennial Scholarship Fund.

The Genotype-Tissue Expression Project (GTEx) was supported by the Common Fund of the Office of the Director of the National Institutes of Health, and by National Cancer Institute, National Human Genome Research Institute, National Heart Lung and Blood Institute, National Institute on Drug Abuse, National Institute of Mental Health, and National Institute of Neurological Disorders and Stroke. GTEx datasets used throughout this thesis were accessed through dbGaP with accession phs000424.v7.p2.

Genome-wide association study meta-analysis summary data in Chapter 6 were published by the DIAbetes Genetics Replication And Meta-analysis (DIAGRAM) Consortium (`www.diagram-consortium.org`) and the Meta-Analyses of Glucose and Insulin-related traits Consortium (MAGIC; `https://www.magicinvestigators.org/`).

This thesis contains approximately 44,000 words (excluding figures, tables, appendices, and bibliography).

Transcriptional mechanisms influencing glycaemic traits and risk of type 2 diabetes

Anthony Joseph Payne, St John's College

Submitted in candidature for the degree of Doctor of Philosophy

Trinity 2018

Abstract

Genome-wide association studies (GWASs) have identified many genetic variants associated with type 2 diabetes (T2D) risk and glycaemic traits, most of which are noncoding. Although existing islet RNA sequencing (RNA-seq) studies have identified genes regulated by T2D-associated variants, sample sizes are small, and most known regulatory targets are messenger RNAs (mRNAs). The work presented in this thesis aims to characterise long noncoding RNA (lncRNA) expression in islets, identify candidate lncRNA regulatory mechanisms, and through incorporation of genetic data, identify genes with potential causal mediation of T2D-associated genetic variation.

Using data from 43 GTEx tissues, I first showed that lncRNAs are more tissue specific than mRNAs. Using 38 human islet samples with paired ribosomal rRNA depletion (RD) and poly-A-tailed RNA selection (PA) RNA-seq, I then showed that RD more comprehensively isolates lncRNAs, but the relative effective sequencing depth is approximately half that of PA. Using RD RNA-seq from 54 islet samples, I identified previously unannotated expressed regions, but comparison to lncRNAs from recently published islet-specific annotations showed high discordance between studies. Thus, although RD better captures lncRNAs, its use in transcriptome-wide islet analyses is not practical until sample sizes increase.

By analysing differential expression (DE) by T2D status in islets ($n = 207$) and 48 tissues from GTEx, I found 89 unique lncRNAs with DE in at least one GTEx tissue, and 75 with DE in islets. To give these functional context, co-expression modules were constructed in each tissue with lncRNAs and mRNAs. One common gene module between pancreas and islet was enriched for lncRNAs with DE, and contained mRNAs associated with insulin secretion regulation, ultimately pointing to lncRNA *CTB-12O2.1* as a candidate regulator of T2D-relevant gene expression.

I next incorporated genetic data with islet RNA-seq ($n = 426$) to demonstrate a modelling approach for identifying sparse, robust genetic models of gene expression. My approach is replicable, robust, and provides models that are sparser and explain more expression variation than LASSO or elastic net models.

Using these models along with expression quantitative trait locus models, I identified 33 candidate expression mediators of genetic associations with T2D and glycaemic traits. Amongst these were two lncRNAs that were transcriptionally associated with biologically relevant mRNAs, making them candidate regulatory targets of corresponding T2D-associated noncoding genetic variants.

The data presented throughout has therefore implicated previously unstudied lncRNAs in T2D risk and the regulation of insulin secretion, and has demonstrated how integration of disease-focussed islet RNA-seq and large-scale public datasets such as GTEx and GWAS meta-analyses can effectively identify novel regulatory mechanisms underlying genetic associations with T2D.

Acknowledgements

First and foremost, I would like to thank my supervisors Mark McCarthy and Cecilia Lindgren for providing me not only with the materials and tools to perform my research, but for mentorship, scientific guidance, flexibility, and patience. Coming from a non-biological background, both have constantly worked with me to support my growth, ultimately leading to the completion of this thesis.

I would next like to thank all those in Oxford that have interacted with me in various ways. From the people in my first week that I spoke with in awkward "What are you studying and where are you from?" conversations, to those that have become some of my closest friends. I sincerely thank my friends from far and near that have visited me or accommodated me. Further thanks to my friends from home that have always taken time to meet me for a quick pint at the Duke during my brief layovers at YYT, and a special shout-out to Mikayla and Chris for making my last year's living relaxed and enjoyable, and for never shying away from a mid-week tipple and card game when I have needed a break.

Members of the McCarthy group require a separate paragraph, as everyone (past and present) has made my entire degree possible through both scientific and social support. Particular thanks to those that have shared the 'dungeon' with me (Loukas, Juan, Jim, Patrick, Matthias, Anne, Martijn, Agata, Marta, Jason, Yunlong, and Moustafa), and to those from upstairs (Anubha, Neil, Will, Natalie, Wei, Hidetoshi). The scientific guidance and discussions, along with many not-so-scientific discussions, debates, and socials have kept me afloat in these four years. Thanks especially to Anne and Martijn who have made GTEx meetings particularly enjoyable and fruitful, and further thanks again to Martijn, who from day one has guided me into the world of RNA, islet biology, writing, etc. Without his input, patience, and guidance from my first days here I would have been lost in the sea of -omics.

Thanks especially to my family, all of whom have supported me in every way imaginable throughout the past four years. Without regular phone calls, visits, updates, and advice, I would have been lost many times.

Finally, I thank Amanda for putting up with my whining and groaning, for supporting me in my all that I do, and for always making me feel confident and proud. Without her love of travel, beer, food, and cribbage, and her persistent encouragement, the process of writing this dissertation and concluding my work would have been miles worse.

Thanks b'ys — long may your big jib draw.

Common abbreviations

2SMR	Two-sample Mendelian randomisation
bp	Base pairs
DE	Differential expression
eGene	Gene with at least one significantly associated eQTL
ENCODE	Encyclopedia of DNA Elements
eQTL	Expression quantitative trait locus
FDR	False discovery rate
FG	Fasting glucose
FI	Fasting insulin
GBA	Guilt by association
GO	Gene ontology
GTEx	Genotype-Tissue Expression Project
GWAS	Genome-wide association study
HbA1c	Glycated haemoglobin
HEK	Human embryonic kidney
InsPIRE	Integrated network for systematic analysis of Pancreatic Islet RNA Expression
IV	Instrument variable
IVW	Inverse-variance weighted
LASSO	Least absolute shrinkage and selection operator
LD	Linkage disequilibrium
lincRNA	Long intergenic noncoding RNA
lncRNA	Long noncoding RNA
MA	Meta-analysis
Mb	Megabase
MR	Mendelian randomisation
mRNA	Messenger RNA
NIER	Novel intergenic expressed region
NIH	National Institutes of Health
OGTT	Oral glucose tolerance test
PA	Poly-A-tailed RNA selection
PC	Principal component
RD	Ribosomal RNA depletion
RefSeq	The Reference Sequence
RNA-seq	RNA sequencing
rRNA	Ribosomal RNA
SMR	Summary Mendelian randomisation
SNP	Single nucleotide polymorphism
SV	Surrogate variable
T2D	Type 2 diabetes
TSS	Transcription start site
TWAS	Transcriptome-wide association study
WGCNA	Weighted correlation network analysis

Contents

1 Introduction **1**

1.1 The global impact of diabetes 1

1.2 Pathophysiology of T2D 3

1.2.1 Clinical diagnosis 3

1.2.2 Aetiology 4

1.3 Genetic influences on T2D 7

1.3.1 T2D GWASs and meta-analyses 7

1.3.2 T2D heritability 8

1.3.3 Challenges in functional understanding of variant associations . . 9

1.4 Transcriptome regulation in T2D 13

1.4.1 Tissue-specific transcriptomics 13

1.4.2 Long noncoding RNA regulatory mechanisms 15

1.4.3 Library preparation and sequencing 20

1.4.4 Gene expression associations with T2D 23

1.4.5 Genetic regulation of gene expression and T2D 27

1.5 Thesis aims 33

2 Data sources **34**

2.1 Genotype-Tissue Expression Project 34

2.2 InsPIRE islets 35

2.2.1 Genotype imputation 36

2.2.2 RNA sequencing 36
2.3 Oxford islets using ribosomal RNA depletion 37

3 The expression profile of lncRNAs and limitations in their detection 39

3.1 Introduction 39
3.1.1 Expression landscape of lncRNAs 39
3.1.2 Detection and quantification of lncRNAs 42
3.1.3 LncRNAs in human islets 44
3.2 Samples and methods 46
3.2.1 GTEx tissues 46
3.2.2 Pancreatic islets 47
3.2.3 Quantification of mapping rates 48
3.2.4 Novel gene annotation 48
3.3 Results 49
3.3.1 LncRNA tissue specificity 49
3.3.2 Comparison between rRNA depletion and poly-A-tailed RNA selection 56
3.3.3 Novel transcribed noncoding regions in islets 62
3.4 Discussion 69

4 The profile of T2D associated gene expression and potential lncRNA regulation 76

4.1 Introduction 76
4.1.1 RNA and phenotype associations 76
4.1.2 Gene co-expression 80
4.1.3 Functional relevance of co-expression clusters 81
4.2 Samples and methods 84
4.2.1 GTEx samples 84

4.2.2 Islet samples . . . 84
4.2.3 Differential expression . . . 85
4.2.4 Co-expression network construction . . . 85
4.2.5 Co-expression module analysis . . . 86
4.2.6 Gene ontology enrichment analysis . . . 87
4.3 Results . . . 88
4.3.1 Gene expression association with T2D in non-islet tissues . . . 89
4.3.2 Gene expression associations with T2D in islets . . . 97
4.3.3 *CTB-12O2.1* and *GLRA1* . . . 102
4.4 Discussion . . . 107

5 Genetic regulation of gene expression 112

5.1 Introduction . . . 112
5.2 Samples and methods . . . 115
5.2.1 RNA-seq filtering and normalisation . . . 115
5.2.2 Genotype filtering . . . 116
5.2.3 eQTL calculation . . . 116
5.2.4 R^2 calculation . . . 117
5.2.5 LASSO regression . . . 117
5.2.6 Linkage disequilibrium calculation . . . 118
5.2.7 LASSO model SNP filtering and final model calculation . . . 119
5.3 Results . . . 120
5.3.1 eQTL analysis . . . 120
5.3.2 Multi-variant gene expression models . . . 122
5.4 Discussion . . . 135

6 Gene expression mediation of T2D and glycaemic trait associated genetic variation **140**

6.1 Introduction . . . 140

6.1.1 Concordance between eQTLs and GWAS association variants . . 141

6.1.2 Multi-variant methods to identify effector genes . . . 145

6.1.3 Mendelian randomisation and its potential in gene expression studies . . . 147

6.2 Methods . . . 150

6.2.1 GWAS meta-analyses . . . 150

6.2.2 eQTL and GWAS concordance . . . 150

6.2.3 Mendelian randomisation . . . 151

6.2.4 S-PrediXcan . . . 153

6.3 Results . . . 154

6.3.1 eQTL and meta-analysis concordance . . . 154

6.3.2 Two-sample Mendelian randomisation and S-PrediXcan . . . 161

6.3.3 Integration with differential expression data . . . 170

6.4 Discussion . . . 181

7 General discussion **187**

Appendices **196**

Bibliography **217**

Chapter 1

Introduction

1.1 The global impact of diabetes

Diabetes is a condition that is characterised by higher than normal levels of blood glucose. The estimated prevalence of diabetes worldwide has increased from 4.7% in 1980 to 8.5% in 2014, corresponding to an increase from 108 million to 422 million affected individuals over the same period.[1] This rapid increase in diabetes cases is expected to continue, with a predicted 629 million diabetic individuals globally in 2045[2] (Figure 1.1). It is also estimated that in 2012, 1.5 million deaths globally were a direct result of diabetes, with a further 2.2 million deaths ultimately attributable to high blood glucose levels.[3]

Type 2 diabetes (T2D), the most common form of diabetes,[4] is often present but undiagnosed,[1] and the risk of cro-vascular disease (such as heart attack or stroke) also begins to increase in hyperglycaemic individuals even below levels of plasma glucose required for T2D diagnosis,[1] creating further challenges in tackling this global epidemic. Although traditionally known as a disease that affects primarily adults, an increased proportion of children globally are being diagnosed with T2D.

Alongside the growing health burden of diabetes, there is an associated

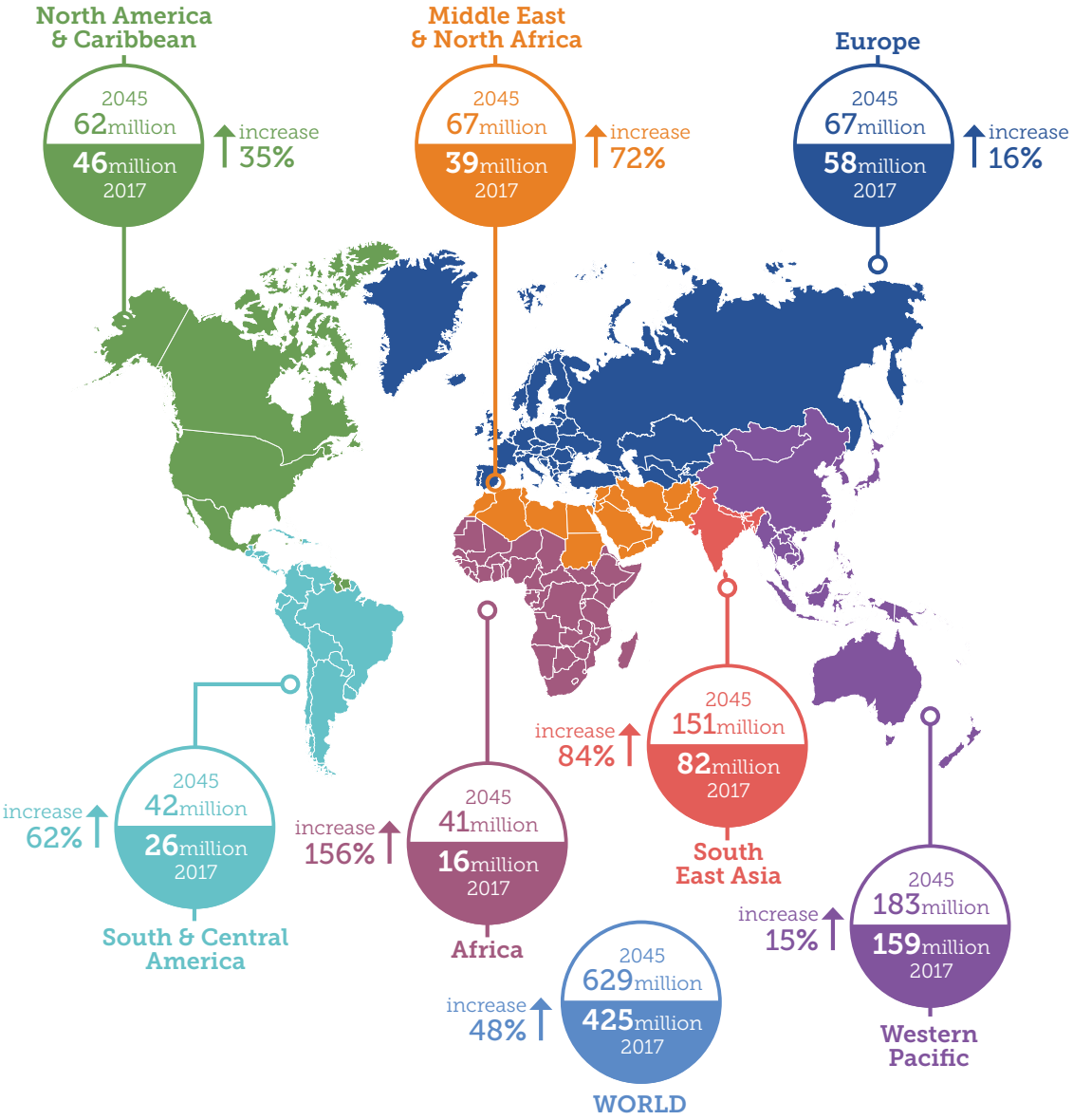


Figure 1.1: Number of people with diabetes worldwide and per region in 2017 and 2045 (20-79 years). The copyright of this figure belongs to the International Diabetes Federation, and it was originally printed in *IDF diabetes atlas, eighth edition* in 2017.[2] The International Diabetes Federation has granted permission to reproduce it in this thesis.

economic cost: in 2015, there was an estimated global spend of $850 billion as a direct result of diabetes (individuals aged 18-99), and this is expected to grow to $958 billion by 2045.[2] Given the immense, ever-growing burden of diabetes on global health and the global economy, expanding the fundamental understanding of the disease and its development is vital to combatting its long-term global consequences.

1.2 Pathophysiology of T2D

1.2.1 Clinical diagnosis

Clinical testing for diabetes is generally prompted by reported symptoms: increased thirst, increased urine, recurrent infection, unexplained weight loss, and occasionally drowsiness or coma.[4] The current World Health Organization diagnosis criteria for diabetes was established in 1998 and remained unchanged in the most recent review.[4,5] Based on the original 1998 criteria, diabetes can be diagnosed through measures of either fasting plasma glucose or 2-hour plasma glucose. In addition to direct diagnosis criteria, any individual with impaired fasting glucose should be tested with an oral glucose tolerance test (OGTT) to determine glucose tolerance status.[5] In 2011, the World Health Organization conducted a systematic review and consultation and deemed that glycated haemoglobin $> 6.5\%$ is also suitable for diagnosing diabetes if stringent quality assurance is in place, measurement is standardised, and there are no conditions that may affect its accurate measurement.[3]

T2D, which accounts for approximately 90% of all diabetes cases in the United Kingdom (and similar proportions worldwide),[6] is a form of diabetes that develops throughout life, with an average age at diagnosis of approximately 58.[7] T2D often goes undiagnosed until major complications arise,[1] since common early symptoms are sometimes mild and unreported despite possibly resulting from hyperglycaemic

progression.

1.2.2 Aetiology

T2D is characterised by two primary changes: a reduction in insulin sensitivity (primarily in skeletal muscle), and a reduction in insulin secretion.[8] Insulin is a hormone secreted by pancreatic beta-cells (contained in pancreatic islets of Langerhans), and along with the hormone glucagon, which is released by pancreatic alpha-cells, is vital in the regulation of plasma glucose levels. In short, during digestion, carbohydrates are processed into glucose, which is released into the blood, increasing glucose levels. As a reaction to increased glucose, the pancreas secretes insulin, which reduces plasma glucose levels through supporting the uptake of glucose in peripheral tissues and through glycogenesis in cells of the liver and muscle. In periods between meals when plasma glucose levels are lower than normal, the pancreas secretes glucagon, which supports glycogenolysis, ultimately maintaining plasma glucose levels. This constant feedback loop results in glucose homeostasis in healthy individuals.

In general T2D progression, early insulin resistance develops in muscle, and is compensated for by increases in plasma insulin levels. However, throughout development of T2D, continuous changes to insulin sensitivity coupled with a reduction in beta-cell function eventually lead to diabetic plasma glucose levels.[9] Though it is clear that there is a combination of effects of reduced insulin sensitivity and beta-cell dysfunction in the development of T2D, the proportion each contributes likely varies between cases and is not entirely understood.[10]

Insulin sensitivity

Insulin sensitivity is a broad term that describes the efficiency with which insulin reduces plasma glucose levels through cell uptake and glycogenesis. A reduction in insulin sensitivity, primarily reflected in skeletal muscle, tends to be

one of the earliest predictors of T2D.[8] Although insulin resistance in muscle generally begins many years before hyperglycaemia, this alone does not result in a hyperglycaemic state. The resulting increased levels of insulin in response to elevated blood glucose promote *de novo* hepatic lipogenesis, which in turn supports hepatic insulin resistance, ultimately disrupting normal glucose homeostasis.[8,9,11] This series of dysfunction is supported by the fact that individuals with type 1 diabetes (resulting from autoimmune death of beta-cells and reduced insulin secretion) have lower observed hepatic fat content compared to healthy individuals.[12]

Reductions in insulin sensitivity are known to be affected by various genetic factors. Early studies identified significant heritability in insulin sensitivity, and recent studies continue to estimate that heritability of insulin sensitivity may exceed 40%.[13–16] Many genetic mutations, both severe and common, have been identified as affecting insulin sensitivity.

Early studies found rare mutations in the insulin receptor gene *INSR* that result in severely impaired insulin sensitivity,[17,18] and since these studies, more than 120 *INSR* mutations,[19–21] along with common mutations in at least 12 further loci,[15,22,23] have been associated with measures of insulin sensitivity in humans. In addition to these, a genetic risk score consisting of 53 variants associated with higher fasting insulin, higher triglycerides, and lower HDL cholesterol was significantly associated with gold-standard measures of insulin sensitivity,[24] suggesting that many of these variants also likely contribute to changes in insulin sensitivity.

Pancreatic beta-cells and insulin secretion

Although insulin sensitivity plays a key role in the development of T2D, it is only when beta-cells fail to compensate for increased insulin resistance that glucose homeostasis becomes disrupted enough to cause a diagnosis of T2D.[25] Beta-cell

dysfunction is a broad term describing changes in beta-cell mass, phenotype, and function. As is the case with insulin sensitivity, there are multiple genetic factors influencing levels of insulin secretion.

Although not directly relevant to T2D development, monogenic forms of beta-cell dysfunction provide useful insight into molecular mechanisms that alter insulin secretion when disrupted. For example, various mutations in the *KCNJ11* (potassium voltage-gated channel subfamily J member 11) and *ABCC8* (ATP binding cassette subfamily C member 8) loci have direct and often severe effects on insulin secretion from birth, resulting in neonatal diabetes.[26–31] Both *KCNJ11* and *ABCC8* encode components of the ATP-sensitive potassium channel ($K_{ir}6.2$ and SUR1 respectively), and mutations in these genes can disrupt these channels in either direction, leading to hyperglycaemia or hyperinsulinaemic hypoglycaemia.[31,32]

Mutations in other genes such as *HNF1A* (Hepatocyte nuclear factor 1 alpha), *GCK* (glucokinase), *HNF4A* (hepatocyte nuclear factor 4 alpha), and *HNF1B* (hepatocyte nuclear factor 1 beta) are also known to impair beta-cell function resulting in monogenic forms of diabetes, with phenotypes varying significantly depending on the mutation.[33,34] For example, heterozygous inactivating mutations of *GCK* are a common cause of maturity-onset diabetes of the young,[35,36] while rare homozygous inactivating mutations can result in permanent neonatal diabetes.[37] Interestingly, in the case of *GCK*, activating mutations have the opposite effect, and can result in hyperinsulinism.[38]

In addition to these and other mutations that cause monogenic forms of diabetes, genome-wide association studies (GWASs) have identified common small-effect genetic variants that are associated with insulin secretion in healthy individuals. Prokopenko *et al.* meta-analysed GWAS summary data of glucose-stimulated insulin response during OGTTs from 6 studies (10,831 individuals), and in addition to seven previously published insulin-associated

variants discovered in previous GWASs, found a significant association at the *GRB10* locus.[39]

1.3 Genetic influences on T2D

1.3.1 T2D GWASs and meta-analyses

In 2007, one of the earliest waves of T2D GWASs emerged, each using T2D case-control data from different studies.[40–43] Although these studies together identified variant associations with T2D in at least ten regions (*TCF7L2*, *SLC30A8*, *HHEX/IDE*, *EXT2*, *IGF2BP2*, *CDKAL1*, *CDKN2A/CDKN2B*, *FTO*, *PPARG*, *KCNJ11*), they all suffered from small sample sizes ($n < 5{,}000$), and small numbers of genotyped single-nucleotide polymorphisms (SNPs) ($< 500{,}000$).

As the cost of genotyping and sequencing dropped, deep-sequenced genome-wide reference panels such as the 1000 Genomes Project[44] emerged. Using relatively inexpensive genotyping arrays designed to target tag variants for haplotype blocks from reference panels, modern GWASs based on genotype arrays can be imputed to cover substantially more genetic variants than have been directly genotyped, leading to high resolution GWASs. With the ever-decreasing cost of DNA sequencing, disease-focussed GWASs using whole genome sequencing are also emerging, such as a study by Fuchsberger *et al.*, which performed whole genome sequencing on a subset of T2D cases and controls to assess the impact on T2D heritability of low-frequency and previously unstudied variants.[45]

Given the number of independent T2D GWASs that have been performed across various populations and studies, there have been many recent efforts to meta-analyse results of independent GWASs to comprehensively identify genetic variants associated with T2D. The most recent T2D meta-analysis (MA),[46] which included data from 32 studies and a total of 74,124 T2D cases and 824,006 controls, was more than 3-fold larger than the next-largest MA published by Scott

et al. in 2017.[47] When combined with previous T2D GWAS and MA efforts,[45,47–79] the recent MA by Mahajan *et al.* brought the total number of known T2D associated loci to 283. At many of these loci, there are two or more independent association signals, together amounting to a total of more than 400 independent association signals across the genome.[46] The majority of these variant associations have been identified in Europeans, although nearly a quarter of these 283 loci harbour variants associated with T2D in non-European populations.[62–79]

1.3.2 T2D heritability

Early estimates of T2D heritability have ranged from 20%-80% depending on populations and study designs,[80] and it was suggested in a parental transmission study in 2000 that T2D risk is additive depending on the T2D status of each parent.[81] A more recent large-scale T2D heritability study using data from 34,166 pairs of same-sex twins from across eight cohorts estimated T2D heritability at 72% (95% confidence interval of 61-78%).[82] Despite high estimates of T2D heritability and a large number of T2D risk variants identified in large-scale GWASs and MAs, known significant T2D-associated genetic variants account for only a fraction of T2D heritability.[45]

In early days of GWASs, there was a common hypothesis that much of the 'missing heritability' in common diseases such as T2D was accounted for by unidentified low-frequency genetic variants with high penetrance.[83] However, this hypothesis has been unproven in the realm of T2D. A relatively recent study by Fuchsberger *et al.* performed whole-genome sequencing (5x coverage) and deep exome sequencing (82x) on 1,326 T2D cases and 1,331 controls and ultimately found that low-frequency variants contribute less to T2D heritability than do common variants.[45] Even after imputing a further 11,645 cases and 44,414 controls in addition to the directly genotyped individuals, results still suggested that low-frequency variation contributed minimally to T2D heritability. Further

simulations by Fuchsberger *et al.* also supported the hypothesis that common variation accounts for a greater proportion of T2D heritability than does low-frequency variation.

Though the rapidly increasing proportion of people with T2D worldwide is largely attributed to increasing obesity, given its high heritability, understanding genetic factors influencing T2D — even those with small effect sizes — is vital for a complete understanding of the disease and for development of T2D treatments.[84]

1.3.3 Challenges in functional understanding of variant associations

Despite the abundance of robust genetic variant associations with T2D, only a small subset has known clear function. One relatively early study by Dimas *et al.* attempted to classify T2D-associated variants based on physiological function. The authors first meta-analysed results of 9 OGTT studies (23,443 individuals) and 7 studies of intravenous-derived insulin sensitivity measures (4,180 individuals),[85] then by combining these MA results with those from the Meta-Analyses of Glucose and Insulin-Related Traits Consortium, they were able to cluster 16 known T2D-associated genetic loci based on their physiological mechanisms. Through hierarchical clustering, Dimas *et al.* placed four T2D association signals into an 'insulin resistance' class, two into a 'hyperglycaemic' class, one in a 'fasting proinsulin' class, and nine in a 'beta-cell' class.[85] However, the same study by Dimas *et al.* failed to classify the mechanism of action for 20 other T2D-associated loci that were assessed.

More recently, Wood *et al.* meta-analysed 10 GWASs (up to 5,567 total individuals) that measured glucose-stimulated insulin secretion based on intravenous-derived measures.[86] They showed that 21/76 known T2D risk alleles at the time were also associated with reduced first phase insulin secretion, including eight associations that were not previously shown to alter T2D risk

through effects on insulin secretion (*HNF1B*, *VPS13C/C2CD4A*, *FAF1*, *PTPRD*, *AP3S2*, *KCNK16*, *MAEA*, and *LPP*).

Although these two studies have uncovered the physiological relevance of some T2D-associated genetic variants, sample sizes were small, and even since publication of the analysis by Wood *et al.*, the number of known genetic variant associations with T2D has increased nearly 6-fold[46,54,86] rendering the results outdated. Physiological classification of variants also provides little insight into the actual molecular mechanisms of action of T2D risk variants.

The majority of T2D-associated variants are also in noncoding regions of the genome, which tend to be poorly annotated with little functional knowledge.[45,47] A recent study by Mahajan *et al.* showed that even at 13/37 T2D-associated regions harbouring coding variants, once the full credible variant sets were fine-mapped, evidence pointed to noncoding variants as the likely causal SNPs. This suggests that even at the approximately 10% of T2D-associated regions that harbour coding variants, the coding variants are instead tagging noncoding regulatory variants in some cases.[53] Thus, despite GWASs and MAs outlining many robust genetic associations with T2D, it is necessary to examine epigenetic and functional data to identify actual functional mechanisms through which noncoding variants act.

Until recently, noncoding regions of the genome had been poorly studied and annotated, making functional understanding of noncoding disease-associated variants challenging. In recent years, large-scale projects have been contributing to the functional characterisation of the genome in varying ways. Three major projects that contribute broadly to these efforts are as follows:

- **The National Institutes of Health (NIH) Roadmap Epigenomics Mapping Consortium**:[87] A project that has mapped DNA methylation, histone modifications, chromatin accessibility, and RNA expression across a range of human cell lines and tissues, including liver, adipose, muscle, and pancreas.

- **The Encyclopedia of DNA Elements (ENCODE) Project Consortium**:[88] Similar to the NIH Roadmap Epigenomics Mapping Consortium, ENCODE has mapped transcription, transcription factor binding sites, chromatin structure, and histone modifications. Selected assays have been executed on nearly 400 types of biosamples, categorised as immortalised cell lines, tissues, primary cells, *in vitro* differentiated cells, stem cells, or induced pluripotent stem cell lines. ENCODE tissues include adipose, liver, muscle, and pancreas.

- **The Genotype-Tissue Expression Project (GTEx)**:[89] This project has performed RNA sequencing (RNA-seq) in up to 38 tissue types per individual for more than 700 post-mortem donors, covering a total spectrum of 53 tissue types. The majority of these individuals also have accompanying whole genome sequencing, which allows for a thorough understanding of genetic regulation of gene expression in a wide spectrum of human tissues. As with the projects above, GTEx includes adipose, muscle, liver, and pancreas samples.

The main advantage of these large-scale public consortium efforts is that they provide functional information in a wide range of human tissues and cell types in an easy-to-access format, allowing researchers globally to perform functional analysis on variants of interest. However, even with the broad spectrum of tissues and cell-types covered, these projects do not contain data from beta-cell samples, and have few samples of pancreatic islets of Langerhans (the tissue type containing beta-cells). Given that disease-associated genetic mechanisms often act in a tissue- or cell-type-specific manner and that islets account for $<2\%$ of whole pancreas volume,[90] researchers have been complementing these wide-spectrum data releases with functional characterisation of islets and beta-cells to better understand T2D-associated genetic variation.

In an early genome-wide functional islet study, Gaulton *et al.* profiled

chromatin and identified open chromatin sites in human islets, and by comparing these to data derived from non-islet cells, were able to identify islet-selective open chromatin sites.[91] This islet-specific open chromatin data showed that the noncoding T2D-associated SNP rs7903146 at the *TCF7L2* locus altered islet enhancer activity and showed allelic imbalance in open chromatin data, suggesting a regulatory effect specific to islets. Pasquali *et al.* used two assays to identify open chromatin sites in islets, and further separated open chromatin sites into classes that displayed specific enrichment patterns of histone modification and CTCF binding.[92] It was then shown that T2D-associated GWAS variants were enriched in islet active enhancers, further narrowing their potential functional role and tissue-of-action in T2D pathogenesis. Gaulton *et al.* later combined epigenetic information derived from both ENCODE and Pasquali *et al.*[92] in an effort to identify the set of variants that are most likely causal (credible sets) at each T2D-associated genomic region.[93] This study ultimately showed that the T2D risk allele at GWAS-significant SNP rs10830963 led to higher *MTNR1B* (melatonin receptor 1B) expression in human islets, which has been suggested as the mechanism through which the SNP influences T2D pathogenesis.[93] Thurner *et al.* recently expanded islet chromatin annotations by combining assay for transposase-accessible chromatin with sequencing ($n = 17$) and whole-genome bisulphite sequencing ($n = 10$) with existing libraries of chromatin immunoprecipitation assays with sequencing to define 15 different islet chromatin states.[94] Thurner *et al.* further found significant enrichment of T2D GWAS signals in islet enhancers. This sequence of publications demonstrates the utility of tissue-specific functional data, but also makes it apparent that there is still much to be discovered in islets, evidenced by the fact that each of these studies has cumulatively expanded the resolution of islet annotations with no clear plateau.

In a manner similar to chromatin profiling, integration of RNA-seq data in genetics studies also provides an important layer of evidence to implicate

disease-associated SNPs in gene regulation. Expression quantitative trait locus (eQTL) analysis and allele-specific expression analysis have become widely used in studies to identify genetic regulators of gene expression across relevant tissues. Datasets with extensive RNA quantifications are emerging for both whole pancreatic islets and beta-cells, and these datasets, along with ever-growing GWASs and datasets of other epigenetic markers have helped to identify various tissue-specific genetic regulatory processes associated with T2D.[95–103] The general success and current limitations of existing islet RNA-seq studies for discovering T2D-relevant processes and regulatory effects of T2D-associated variants in humans are described in depth in the remainder of this chapter.

1.4 Transcriptome regulation in T2D

1.4.1 Tissue-specific transcriptomics

It is well established that gene expression and regulatory mechanisms often act in a cell-type- or tissue-specific way.[88,89,104,105] This means that when studying a complex disease, it may be necessary to assess various relevant tissues and cell-types to identify regulatory effects. As previously described, when assessing regulatory mechanisms directly relevant to T2D, it is particularly important to focus on skeletal muscle, adipose, liver, and pancreatic islets; however, to confidently determine the tissue or tissues of action of a regulatory mechanism, it is also important to perform analyses in other tissue panels as a comparative measure.

For example, RNA-seq data from skeletal muscle was used in one recent study to show that *ANK1* (ankyrin 1) expression was regulated in muscle by a variant associated with T2D.[106] The study showed that TR4 binding at rs508419 is linked to expression repression of *ANK1*, and that the T2D risk allele disrupts this process, resulting in increased *ANK1* expression in skeletal muscle. Although it

was suggested by Scott *et al.* that rs508419 resided in an active promoter that was specific to skeletal muscle and human skeletal muscle myoblasts, comparisons were only made to eight other tissue and cell types.[106] In fact, according to eQTL results released from GTEx version 7, rs508419 is significantly associated with expression of *ANK1* in heart, muscle, adipose, skin, breast, and tibial nerve (`www.gtexportal.org`), suggesting that the regulatory effects identified in skeletal muscle may be wider spread. It has further recently been suggested that variants influencing *ANK1* expression have minimal impact on T2D risk, and that the influence on T2D risk at this region is more likely driven by variant associations with human islet expression of *NKX6-3* (NK6 homeobox 3), a transcription factor that is important in islet development and function.[46]

In another T2D-relevant case, genetic variant associations with T2D were identified near *KLF14* (Kruppel-like factor 14), and it has further been shown that the maternally inherited T2D risk allele of rs4731702 was associated with reduced *KLF14* expression (based on microarray data) in adipose but not in other tissues.[51,60,107] Members of the Multiple Tissue Human Expression Resource Consortium analysed an independent dataset of microarray-based gene expression from adipose tissue and found that the same set of variants with T2D association was associated with expression of *KLF14* and ten other distal genes, and they found that *KLF14* acts as a master transcriptional regulator of these genes in adipose tissue.[108,109] Unlike rs508419 and *ANK1*, regulatory associations between rs4731702 and *KLF14* have not been widely observed in other tissue types.

As previously described, beta-cell dysfunction is one of the fundamental dysfunctions leading to T2D. As a result, several studies have attempted to profile the transcriptome of beta-cells in humans. Two primary methods have been used to identify beta-cell-specific genes and characterise the beta-cell transcriptome: single-cell sequencing in human islet samples,[110–112] and fluorescence-activated cell sorting to separate islet cell populations prior to RNA-seq.[96,113,114] Despite the

cell-type-specific transcriptional information gained through beta-cell transcriptome profiling, these beta-cell studies all suffer from small sample sizes, and are thus not suitable for transcriptome-wide methods such as eQTL and co-expression analyses. To compromise, many modern T2D transcriptomic studies have instead focussed RNA-seq efforts on whole pancreatic islets.[97–103,115,116] Although genes and mechanisms detected in whole islets cannot be assumed to act through beta-cells, the increased sample sizes available in these studies allow for identification of genes and mechanisms relevant to T2D pathogenesis and insulin secretion regulation that may have otherwise been missed in studies based on small-scale beta-cell datasets, or datasets of other tissue types.

1.4.2 Long noncoding RNA regulatory mechanisms

For many years, the noncoding portion of the human genome was largely regarded as 'junk'. In fact, the noncoding genome has often been referred to as 'junk DNA', likely originating from a paper titled 'So much "junk" DNA in our genome' by Susumu Ohno in 1972.[117] The view that noncoding genomic regions were not functional was historically a popular theory, although there remained scientists that believed there must still be some role for these regions.[118] With the introduction of next-generation sequencing in 2005, deeper sequencing of noncoding regions became possible, which allowed for more in-depth characterisation of noncoding genetic variants.[44,119]

As high-throughput sequencing became more standard, RNA-seq was also established as a common technique. Unlike microarray experiments, RNA-seq depends on no prior knowledge of genes or expressed regions, and since identification of expressed noncoding genes was previously difficult (since they do not contain open reading frames),[120] RNA-seq enabled identification of many noncoding genes that were previously unknown.[121] High-throughput sequencing therefore allowed for thorough classification of the noncoding genome both through

annotation of variants and noncoding genes that were not previously known.[121]

With the classification of noncoding genes, it soon became clear that much of the noncoding genome was actually transcribed, and a large fraction of these noncoding transcripts were classified as long noncoding RNAs (lncRNAs; noncoding genes greater than 200 nucleotides in length).[122] The existence and potential functional capabilities of lncRNAs were known previously, with a prime example being the *XIST* lncRNA (X-inactive specific transcript), which since 1996 has been proven to be the primary regulator of X chromosome inactivation in females.[123] However, with the vast increase in the number of known lncRNAs introduced by high-throughput sequencing, many researchers have shifted focus to this type of RNA, which is a trend clearly evidenced by the fact that the total number of indexed entries in PubMed related to lncRNAs increased from 1 in 2006 to more than 1,200 in 2016 (the term 'lncRNA' was not widely used prior to this).[124]

In general, lncRNAs are often spliced and polyadenylated in a manner similar to mRNAs.[125] GENCODE, one of the gold standard gene annotation resources, has classified more than 25% of all human genes as lncRNAs.[126] The majority of GENCODE lncRNAs fall into one of four categories (Figure 1.2):[126,127]

- long intergenic noncoding RNA (lincRNA) - no overlap with any other annotated gene
- sense intronic lncRNA - contained completely within the intron of a coding gene
- sense overlapping lncRNA - transcribed from a coding strand, and containing some genomic overlap with a coding gene (anywhere from partial coverage of a single exon to complete overlap)
- antisense lncRNA - transcribed from a coding gene's antisense strand and overlapping the coding gene's genomic coordinates

Any lncRNA that does not fit within one of these categories is defined by GENCODE simply as a processed transcript. Although these lncRNA classifications are based simply on location and strand of transcription, they still provide a useful means for stratifying this otherwise diverse class of genes.

Despite the vast abundance in humans and the increasing rate at which they are studied, many lncRNAs still have no identified biological function,[125] and due to the discrepancy between general abundance and lack of functional knowledge of lncRNAs,[125] the portion of lncRNAs that have true function has been widely debated.[128] Nevertheless, lncRNAs have been implicated in various broad cellular functions, and disruption of lncRNA expression has been shown to contribute to disease.[123, 125, 129–133] It has also been widely observed that lncRNAs tend to be more cell-type specific than mRNAs, which amplifies the importance of obtaining RNA-seq data from relevant tissues in order to infer their function.[131, 134, 135]

There is currently limited knowledge of lncRNAs in the pathogenesis of T2D, especially when compared to existing knowledge of microRNAs, a class of small noncoding RNAs.[136] The first broad in-depth look at islet-specific lncRNAs by Morán *et al.* identified more than 1,100 islet-expressed lncRNAs, the majority of which were antisense lncRNAs ($n = 761$) and lincRNAs ($n = 335$).[137] They further showed that nearly half of the lncRNAs identified in islets were islet-specific when compared in a panel of 18 tissue types, and that many of these islet-specific lncRNAs were directly resulting from beta-cell transcription.[137] The authors further identified lncRNAs that were associated with beta-cell differentiation, and highlighted the lncRNA *HI-LNC25* (human islet lncRNA 25) which was shown through knockdown experiments to regulate expression of *GLIS3* (GLIS family zinc finger 3); however, knockdown did not affect glucose-stimulated insulin secretion. Mutations in *GLIS3*, a gene that encodes an islet transcription factor, have been shown to cause a rare form of neonatal diabetes,[138] and *GLIS3* also harbours intronic variants associated with type 1 diabetes and T2D.[62, 139] Morán *et*

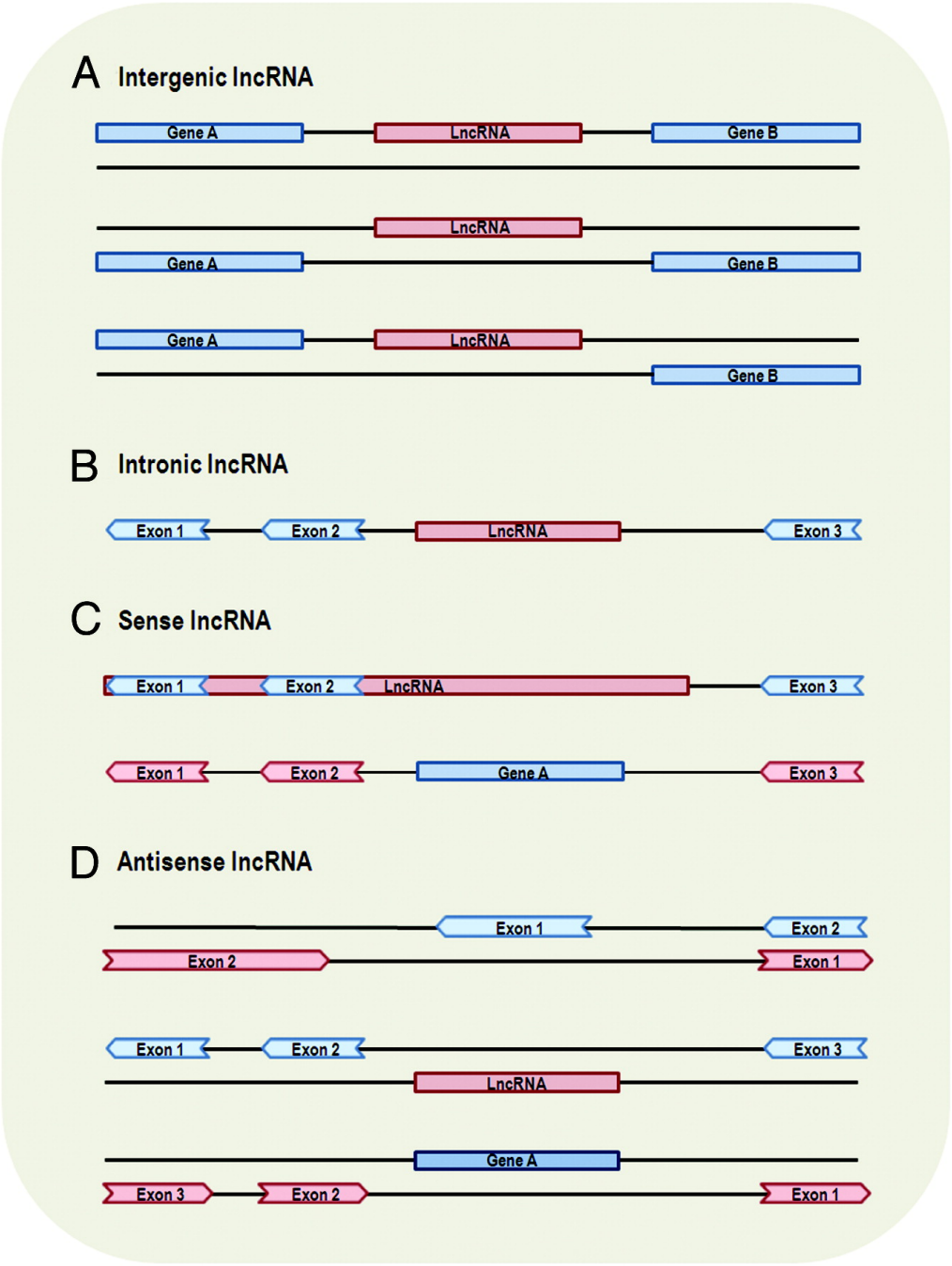


Figure 1.2: Primary lncRNA types. Blue represents coding genes, and pink represents lncRNA genes. A) intergenic lncRNAs contain no overlap with any other gene; B) sense intronic lncRNAs are completely contained within a coding gene's intron; C) sense overlapping lncRNAs overlap a coding gene (partial exonic overlap or complete containment); and D) antisense lncRNAs overlap any part of a coding gene (partial overlap, completely contained within a coding gene's boundaries, or completely covering a coding gene). This figure is part of an open-access article written by Ma *et al.* and published by Landes Bioscience under a Creative Commons Attribution-NonCommercial 3.0 Unported License.[127] The license and original article are available at `www.tandfonline.com/action/showCopyRight?scroll=top&doi=10.4161%2Frna.24604`.

al. further found that the expression of lncRNAs *KCNQ1OT1* (KCNQ1 opposite strand/antisense transcript 1) and *HI-LNC45* (human islet lncRNA 45) was dysregulated in islets from donors with T2D. Acute glucose stimulation experiments were also performed to test glucose-stimulated changes in expression of islet lncRNAs, and two lncRNAs — *HI-LNC78* and *HI-LNC80* — were upregulated in response to glucose stimulation in islets of five donors.

As a follow-up to the study by Morán *et al.*, Akerman *et al.* recently published a thorough broad extension of islet- and beta-cell-specific lncRNAs.[103] Using a similar detection pipeline as Morán *et al.*, except with total RNA-seq of islets from 41 individuals, Akerman *et al.* first defined a set of lncRNAs using islet RNA-seq, and filtered to identify a subset of 2,373 that were also expressed at a threshold of 0.05 fragments per million mapped reads in beta-cells (including previous definitions by Morán *et al.*). They identified a subset of 12 lncRNAs that were near protein-coding genes with important beta-cell function, and found that knockdown of 9/12 of these lncRNAs in human EndoC-βH1 cells caused transcriptional changes, and some of these lncRNAs had effects on expression levels of neighbouring genes. It was further shown that glucose-stimulated insulin response differed significantly after knocking down three of these lncRNAs in human EndoC-βH3 cells, suggesting that there are lncRNAs that not only regulate gene expression in beta-cells, but consequently alter glucose-stimulated insulin response in a human beta-cell line.

To explore the potential of lncRNAs in the regulation of insulin secretion in more detail, Akerman *et al.* performed follow-up experiments on lncRNA *HI-LNC71*, one of the lncRNAs of which knockdown impaired insulin secretion.[103] They showed that knockdown of *HI-LNC71* (which they refer to as *PLUTO* - *PDX1* locus upstream transcript) in EndoC-βH1 cells, human islet cells, and mouse MIN6 cells all resulted in decreased levels of PDX1 (pancreatic and duodenal homeobox 1), a transcription factor that is important in pancreas

development and beta-cell function. This study successfully built on the foundational analysis by Morán *et al.*, and not only defined a larger set of islet and beta-cell lncRNAs, but further demonstrated lncRNAs as functionally relevant to the regulation of gene expression and insulin secretion in islets.

Both studies described above are limited by small sample sizes. Even in the extended analysis by Akerman *et al.* (which used 5-fold more samples for lncRNA detection than the analysis by Morán *et al.*), islet lncRNA definitions were based on data from only 41 individuals. While both studies performed tissue specificity analysis of the identified lncRNAs, comparisons were made to a panel of only single samples from different tissue types. Given the small sample sizes of islets and the comparison to only single-sample RNA-seq libraries of other tissues, it is likely that many of the islet- and beta-cell-specific lncRNAs identified are not truly specific, and may be shared with other tissues or cell types. Since these studies were purely focussed on transcriptomics, there was, by design, no use of genetic data in either study. Given the abundance of noncoding genetic variation in humans associated with T2D risk,[53] incorporation of genetic data in such epigenetic studies is likely to uncover not only the identity of T2D-relevant regulatory lncRNAs, but the underlying genetic variants that are ultimately responsible for their regulation.

1.4.3 Library preparation and sequencing

In addition to studying relevant tissues in transcriptomic studies, it is important to understand the impact of the RNA-seq protocol used, particularly for studies focussed on lncRNAs. In modern RNA-seq studies, libraries are generally prepared using one of two methods: poly-A-tailed RNA selection (PA), and ribosomal RNA (rRNA) depletion (RD)[140,141] (Figure 1.3). For both protocols, total RNA is first extracted from the sample. In PA, poly-T oligonucleotides that are attached to magnetic beads are added to total RNA and selectively bind to polyadenylated transcripts. During a washing process, any unbound (thus

non-polyadenylated) transcript is removed.[142] In RD, a range of commercial kits such as RiboMinus™(Thermo Fisher Scientific) and Ribo-Zero® (Illumina, Inc.) are available that selectively target and deplete only rRNAs. Several studies have compared the utility of these preparations across various metrics and downstream analyses, typically using a small number of paired replicates for comparison.[140,141,143–148]

These studies all found high correlation between quantifications of coding genes using both library preparations along with significantly higher exonic read proportions using PA. Another common finding amongst these comparative studies is that RD libraries tend to result in quantification of a greater number of unique genes, with increases primarily in the number of lncRNAs detected. Given that RD aims to remove only rRNA while PA aims to remove everything that is not polyadenylated, the remaining transcripts after PA should be a subset of those transcripts remaining after rRNA depletion, which agrees with the pattern identified. Large-scale RNA-seq studies such as GTEx, The Genetic European Variation in Disease Consortium, and the Integrated network for systematic analysis of Pancreatic Islet RNA Expression (InsPIRE) collaboration have used PA for their sequencing, while ENCODE has libraries of both preparations.[88,89,149] Given that a large portion of the genome encodes RNA that may not be polyadenylated, RD datasets are necessary to accurately annotate the whole transcriptome, including noncoding, non-polyadenylated lncRNAs.[143]

Given that existing comparative studies have used only a small number of samples from a wide range of tissue and cell types, there have been differing conclusions in terms of mapping rates, abundance of unannotated intergenic reads, and general lncRNA quantifications. To fully understand the impact of the sequencing preparation method on downstream RNA-seq analyses, it is therefore necessary to perform comparisons between the sequencing methods using the specific tissue type of interest. Comparative studies do not currently exist for

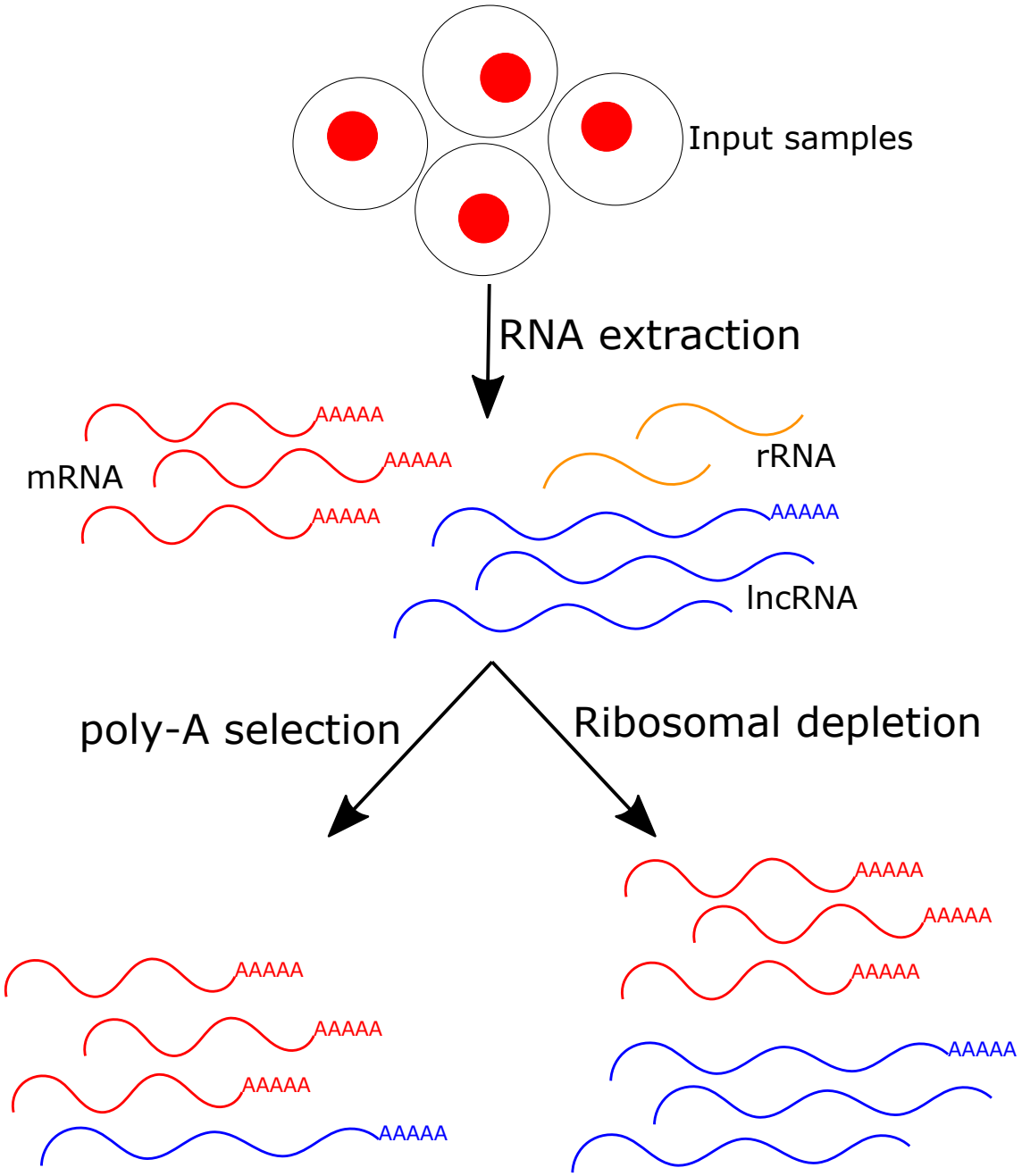


Figure 1.3: RNA-seq library preparations. Total RNA extracted from input tissues or cells consists of all transcript types, including messenger RNAs (mRNAs), lncRNAs, and rRNAs. For PA, polyadenylated transcripts are bound and retained for downstream sequencing while remaining RNA is washed out. For RD, rRNA is removed from total RNA, and all remaining product is sequenced.

pancreatic islets, making it difficult to estimate the true effect of sequencing preparation on downstream islet RNA-seq analysis. Since the majority of existing islet RNA-seq datasets (including data from the InsPIRE collaboration) have used PA, a thorough assessment of the strengths and limitations of PA compared to RD in islets is presented in Chapter 3.

1.4.4 Gene expression associations with T2D

To identify specific genes that are relevant to T2D, studies such as those by Morán *et al.* and Akerman *et al.* described above often assess purely transcriptomic mechanisms. Identifying genes that have expression in relevant tissues that is associated with T2D can prioritise genes for further study and can provide insight into the actual mechanisms and pathways through which they act. Several methods exist for identifying disease-associated genes and clusters of functionally related genes, the most common of which are described below along with published findings in the realm of T2D.

Differential expression

One of the most common methods for identifying disease-relevant genes is differential expression (DE) analysis, which simply tests the difference in gene expression between levels of a given variable (typically case-control).

DE analysis has been previously used to identify lncRNAs associated with T2D in several studies covering different tissue types. One relatively early analysis tested genes for DE in adipose between 12 pairs of monozygotic twins discordant for T2D and found 197 individual genes with different expression levels, including one lncRNA.[150] A more recent lncRNA study of whole blood RNA from 6 T2D cases and 6 healthy controls identified 55 lncRNAs with DE, of which 39 were upregulated and 16 were downregulated in blood from donors with T2D.[151]

One of the most compelling lncRNA DE findings in non-islet tissue surrounds

the maternally expressed lncRNA gene *H19*, which was found to have 5-fold reduced expression in muscle of humans with T2D compared to healthy controls.[152] It has been shown that *let-7* (lethal-7) microRNA binding to *H19* RNA counteracts expression-inhibiting interactions between *let-7* and both *INSR* (insulin receptor) and *LPL* (lipoprotein lipase).[153,154] Downregulation of *H19* in muscle of individuals with T2D results in increased endogenous *let-7*, ultimately leading to greater inhibition of *INSR* and *LPL* expression in muscle and dysregulation of glucose homeostasis.[153,154]

The previously described islet lncRNA studies by Morán *et al.* and Akerman *et al.* also each looked at lncRNA DE in islets between T2D cases and controls.[137] Morán *et al.* tested for DE of 14 lncRNAs between islet samples from 19 non-diabetic donors and 16 donors with T2D and found that expression of *KCNQ1OT1* and *HI-LNC45* were upregulated and downregulated respectively in islets of T2D donors. Akerman *et al.* re-quantified RNA-seq of 89 donors published by Fadista *et al.*[98] using their set of 2,373 *de novo* lncRNA definitions, and found that 15 lncRNAs had significant DE between samples from normoglycaemic and T2D donors.[103] Since the lncRNA definitions by Akerman *et al.* were derived using data from RD islet libraries, and the RNA-seq they used for DE was based on PA preparation, it is possible that a portion of their 2,373 lncRNA annotations describe non-polyadenylated lncRNAs that were not detected in data used for DE analysis.

In a recent thorough DE study in islets, Ottosson-Laakso *et al.* tested for expression changes based on acute glucose stimulation and long-term high glucose (hyperglycaemic and T2D patients), and further performed association tests between gene expression and insulin stimulatory index (ratio between secreted insulin in response to high and low glucose stimulation) to try and dissect primary and secondary effects on insulin secretion.[101] They identified 16 lncRNAs (amongst other genes) that had significant DE between islets of donors with and without

T2D that did not significantly change after acute glucose stimulation, and suggested that these are more likely than other genes to causally influence T2D risk. However, without formally assessing causality of these genes on T2D-related traits, and without addressing whether expression dysregulation in T2D donors is specific to islets or also occurs in other tissues, there is no way to confidently distinguish between primary and secondary associations with insulin secretion and T2D risk.

Co-expression network analysis

In addition to identifying independent genes that are associated with disease, it is also insightful to identify groups of genes exhibiting similar patterns of expression. Genes with similar expression profiles may be functionally related and involved in the same cellular processes, thus determining these clusters and identifying those that are relevant to disease can further pinpoint relevant genes and the specific mechanisms through which they act. Gene co-expression analysis is a commonly implemented means of identifying clusters of co-expressed genes. Assuming co-expression networks have been constructed using a combination of genes both with and without known functional annotations, function can be inferred for the unknown-function genes using a 'guilt by association' approach. In guilt by association analyses, it is simply assumed that genes in common co-expression clusters (high expression correlation) are associated with the same functional processes and pathways.

Wang *et al.* recently conducted pairwise co-expression analysis between lncRNAs and mRNAs with DE based on T2D status and identified ten lncRNA-mRNA pairs with near perfect expression correlation in peripheral blood, but other than correlation analysis with other measured phenotypes, no formal functional assessment was performed on the lncRNAs.[151] Fadista *et al.* found that expression of lncRNA *LOC283177* in RNA-seq from islets of 89 donors was

associated with glycated haemoglobin levels, and *LOC283177* also had strong co-expression with *MADD* (MAP kinase activating death domain), *SYT11* (synaptotagmin 11), and *PAX6* (paired box 6), all of which have roles in islet function.[98] Fadista *et al.* also showed that *LOC283177* had a significantly associated eQTL, and further showed that *LOC283177* had expression association with islet insulin exocytosis, suggesting that it is relevant to islet function.[98]

The most thorough islet lncRNA co-expression analysis was published by Akerman *et al.*[103] The authors constructed gene co-expression networks using RNA-seq from 64 human islet samples and identified 25 distinct gene co-expression modules each containing at least 100 genes. Using their set of 2,373 previously described beta-cell lncRNAs, they found seven co-expression modules in islets that were enriched for beta-cell lncRNAs. The largest of these seven beta-cell lncRNA enriched modules had an over-representation of genes associated with prototypical islet cell functions, and also contained islet transcription factors. They also identified a number of lncRNAs and cell-specific transcription factors that had high expression correlation amongst the islet samples examined.

Although these studies offer interesting insight into the co-expression patterns of lncRNAs and the pathways with which their expression is associated, they fail to validate whether these lncRNAs are actual regulators of expression. For example, Akerman *et al.* observed 292 lncRNAs that had highly significant expression correlation with adjacent protein coding genes, and classified these as "candidate *cis*-regulatory lncRNAs". Although there is evidence that some islet lncRNAs do have expression regulatory effects (such as *PLUTO* described above), there may be cases where the nearby lncRNAs are transcriptional noise associated with expression of nearby coding genes.[155] It is also possible that even if a lncRNA regulates the expression of a nearby coding gene, changes in expression of the paired lncRNAs and coding genes may be a consequence of elevated blood glucose in T2D donors, and may not have a causal impact on insulin secretion or T2D

risk.[101] The lncRNA co-expression studies by Fadista *et al.* and by Akerman *et al.* provide interesting insights into lncRNA co-expression patterns in islets and identify lncRNAs that are co-expressed with genes relevant to islet function, but without further validation either through incorporation of functional and genetic data or experimental follow-up, it is not clear whether the co-expression relationships identified are of actual direct relevance to healthy islet function.

1.4.5 Genetic regulation of gene expression and T2D

Identification of purely epigenetic regulatory transcriptomic mechanisms, such as those previously described, provides insight into the interactions between expressed genes, and the potential downstream phenotypes and diseases with which they are associated. It is also important to understand how these gene expression changes and mechanisms are actually controlled and differ between individuals. It has been shown in many studies that variation in gene expression is often associated with genetic variation, and this genetic regulatory architecture, much like gene expression itself, often exhibits tissue specificity.[89] A common hypothesis is that some of the GWAS-identified noncoding variant associations with disease exhibit this association through regulation of gene expression; therefore, identifying disease-associated genetic variants that are also associated with expression provides insight into their potential function. There are a variety of methods for assessing genetic regulation of gene expression and further identifying genetic regulatory effects with influence on downstream phenotype or disease, the most common of which are described below.

eQTL analysis and colocalisation

eQTLs are simply genetic variants that are associated with expression levels of a gene. The majority of eQTL studies look only at local effects (*cis*-eQTLs), as a full assessment of distal effects (*trans*-eQTLs) is essentially equivalent to

performing a full GWAS for the expression of each gene, and is therefore hampered by a massive multiple testing burden. A typical *cis*-eQTL study will test all variant-gene pairs where the variant is within 1 megabase of the gene's defined boundaries.

An early example of using eQTL analysis to identify a functional mechanism relevant to T2D is at the *KLF14* locus. Variants at this region were identified as being associated with T2D at a genome-wide significant level and were further suggested to primarily act through insulin action.[60] It was then shown in another study that the T2D risk allele of one of these SNPs, rs4731702, was associated with lower expression of *KLF14* in adipose tissue[51] and that this SNP (and others in high linkage disequilibrium) was enriched for *trans*-eQTL associations with many other genes in adipose tissue, suggesting that *KLF14* regulates expression of many adipose genes.[109] This successful application of eQTL analysis in adipose tissue demonstrated the importance of performing such analyses in relevant tissues, and in recent years, similar studies have been performed using expression data from human islets due to their importance in T2D pathogenesis.

One of the earliest RNA-seq-based islet eQTL studies by Fadista *et al.* tested for eQTLs in RNA-seq data from human islets of 89 donors and identified 640 eQTL genes, five of which had eQTL variants that corresponded to T2D associated SNPs.[98] A later study that used RNA-seq data from human islets of 118 different individuals at nearly twice the sequencing depth as the study by Fadista *et al.* identified 2,341 genes with at least one significant exon-level eQTL effect in islets, including eQTL variants at 14 loci that likely colocalised with known T2D or glycaemic trait associated SNPs.[100]

More recently, Varshney *et al.* used deep RNA-seq from islets of 112 individuals (31 donors in addition to 81 from the study by Fadista *et al.*[98]) to expand these analyses, identifying a total 28 eQTL variants (15 of which were previously unpublished), that were also significantly associated with T2D or

glycaemic traits.[102] Ottosson-Laakso *et al.* were able to identify 9 additional unpublished colocalised variants using RNA-seq from islets of 191 donors (including islets used by Fadista *et al.*[98]).

The 42 unique genes identified in these studies (in addition to three identified using microarray data by Taneera *et al.*[115]), are described in Table 6.1. Six of the forty-two unique genes identified in the studies above are lncRNAs: antisense lncRNAs *RP11-797D24.4*[101] and *SIX3-AS1*;[102] lincRNAs *LINC00261*,[101] *RP11-89K21.1*,[102] and *AC012354.6*;[102] and processed transcript *ABO*.[98] Since each of these lncRNAs has been identified in only a single study (and therefore has failed to 'replicate'), it is possible that their significant eQTL effects may be false positive findings. It nevertheless seems apparent from these studies that discovery of T2D- and glycaemic-trait-associated islet eQTLs has not yet plateaued, and analyses using increased sample sizes are thus likely to validate robust previously published discoveries, and identify new genes with potential mediating expression effects. The sample set of islets in the InsPIRE collaboration (described in Chapter 2), which includes islet RNA-seq from each of these islet eQTL studies in addition to samples from the Cell Isolation and Transplant Center at the Geneva University Hospital, amounts to the largest dataset of islet RNA-seq to date, therefore allowing for more powerful detection of eQTLs compared to any previous islet eQTL study.

Aggregate gene-level tests

While eQTL studies consider single-variant tests independently, gene-level aggregate tests that combine the effects of multiple variants into a single model have also been developed to assess transcriptomic regulation as a medium of action for disease-associated variants. Compared to single-variant tests and approaches, combining variant effects into gene-level tests has the potential to reduce multiple-testing burdens, exploit more comprehensive genetic regulatory data, and

in some cases may lead to better detection of expression-trait associations with small effects.[156,157] In a manner similar to eQTL and GWAS colocalisation analyses, variant associations with gene expression and phenotype are often derived from independent studies.

One such approach for gene-level analysis is transcriptome-wide association study (TWAS). TWAS uses a Bayesian linear mixed model to calculate the variant effects on expression for all variants within a *cis*-window of each gene, and uses a linear combination of these effect sizes with corresponding GWAS effects as an aggregate test statistic for each gene of interest.[157] However, it has been shown in an eQTL study of 4,896 peripheral blood samples that only 34.9% of genes with a significant eQTL had a secondary, independent eQTL, and further, only 11.7% of genes with an eQTL had 3 or more independent signals.[158] This implies that aggregate effects of hundreds or thousands of variants per gene likely provide little added benefit, and may even introduce unnecessary noise.

Another method, summary data-based Mendelian randomisation (SMR), approaches this issue from another extreme, considering only the most significant eQTL variant for each gene from an eQTL study and its corresponding effect from a GWAS of interest.[159] Though SMR uses only eQTL variants with evidence of association, it intrinsically ignores further variants that may have smaller independent effects on gene expression.

A different approach, PrediXcan, falls between these methods, using penalised regression to incorporate data from multiple SNPs that jointly explain each gene's expression while still reducing the variant sets compared to TWAS.[156] PrediXcan models the expression of each gene against all *cis*-SNPs using elastic net regression,[160] which essentially incorporates a coefficient penalty that combines ridge regression[161] and least absolute shrinkage and selection operator regression[162] penalties, allowing for sparser models than Bayesian linear mixed model as used in TWAS. However, the PrediXcan method depends on individual-level paired

genotype and phenotype data, which makes it inapplicable for many large-scale GWASs, where only summary data is published.

The S-PrediXcan method, an extension of PrediXcan, has attempted to mitigate this issue by requiring only summary-level data from both gene expression studies and GWASs.[163] As a result, the authors of S-PrediXcan have produced a database assessing gene expression effects that mediate variant effects on phenotype for more than 100 traits across multiple human tissues.

Although not widely used for gene expression mediation studies to date, one method that fully bridges the gaps between those previously described is inverse-variance weighted (IVW) two-sample Mendelian randomisation (2SMR).[164] 2SMR is an approach that tests for causal association of a risk factor against a phenotype of interest (each of which has been measured in an independent cohort) by applying instrument variable analysis, using genetic variants as instruments.[165] The most common framework for applying 2SMR analysis is the IVW framework, which relies solely on summary-level data and can therefore combine results from independent RNA-seq studies and GWASs in a manner similar to S-PrediXcan.[164,166] Although analytically similar to the gene-level methods described above, a valid formal Mendelian Randomisation analysis is strictly dependent on sufficiently meeting a series of assumptions (described in more detail in Chapter 6), which are not discussed or assessed in most analyses using the previously described methods.[167] It is therefore the case that analyses using TWAS, SMR, and S-PrediXcan may be considered 2SMR approaches if the instrument variable assumptions are adequately assessed and met.

IVW 2SMR has been widely used to assess potential causality of many risk factors on many outcomes. For example, IVW 2SMR has been used to show that increased body mass index increases the risk of developing T2D, with an estimated odds-ratio increase to T2D risk between 1.19 and 1.38 per unit of body mass index.[168] The IVW approach was also employed to suggest that lower birth weight

was causally associated with an increase to T2D risk.[169] One of the main benefits from applying IVW 2SMR to assess gene expression effects on phenotype compared to S-PrediXcan is the extent to which the IVW method has been critically assessed, improved, and studied.[170–173] For example, MR-Egger regression has been described by Bowden *et al.* as a method for detecting some violations of genetic instrument assumptions,[174] and Burgess *et al.* have used extensive simulations to estimate bias introduced in IVW 2SMR when the genetic effects on a risk factor and the genetic effects on the outcome of interest have been derived from overlapping samples.[175] Additionally, the importance of selecting valid genetic instruments for use in IVW 2SMR has been widely emphasised and discussed, whereas validity as instruments of variants selected in TWAS, SMR, PrediXcan, and S-PrediXcan have not been assessed in great detail.

Given the wide use of IVW 2SMR in the scientific community, it is seemingly surprising that it has rarely been used to assess gene expression as a risk factor on diseases of interest. In general, the IVW method requires a sparse set of independent genetic variants that are associated with the risk factor.[171] Unlike GWASs, where vast efforts have contributed to compiling lists of independent, significantly associated genetic variants, eQTL studies often find only 3 or fewer independent eQTL variants per gene,[158] while elastic net regression typically identifies many correlated variants with negligible effects on gene expression, making variant sets derived using both methods unsuitable for use in 2SMR. However, assuming sparse, yet comprehensive sets of independent variants associated with gene expression can be derived, well-powered and unbiased IVW 2SMR analysis can be applied to test for genetic variant associations with disease that are likely mediated through regulation of gene expression. The reduced multiple-testing burden and increased power from gene-level tests in this context, particularly with RNA-seq from pancreatic islets, may therefore allow for discovery of previously unknown biological mechanisms through which noncoding

T2D-associated genetic variants act.

1.5 Thesis aims

The work presented in this thesis aims to identify gene expression variation associated with the pathogenesis of T2D and glycaemic traits, and to outline RNA regulatory mechanisms associated with T2D development, with a particular focus on lncRNAs. More specifically, the aims are as follows:

- Characterise general differences in expression profiles of mRNAs and lncRNAs, assess the downstream differences between common RNA-seq preparations in islets, and assess the abundance of unannotated expressed regions in human islets using whole transcriptome RNA-seq.
- Describe the landscape of gene expression associations with T2D and identify sets of genes, including lncRNAs, that are putatively involved in the regulation of insulin secretion, and ultimately the pathogenesis of T2D.
- Demonstrate a modelling pipeline that sparsifies and improves genetic models of gene expression as compared to other penalised regression methods.
- Identify genes — particularly lncRNAs — with expression that mediates T2D-associated genetic variant effects by leveraging genetic models of islet gene expression and high-powered published MAs.

Chapter 2

Data sources

This chapter describes the primary data sets used for analyses throughout this thesis. A general overview of each data set is presented here, with further relevant details described in each chapter.

2.1 Genotype-Tissue Expression Project

The Genotype-Tissue Expression Project (GTEx) is a large-scale project that aims to provide a public source of human RNA sequencing (RNA-seq) data from a wide spectrum of human tissues along with genotype and phenotype data. A full description of GTEx including sample procurement, tissue harvesting, sequencing protocols, and data availability is available at `https://gtexportal.org/home` and in the most recent core GTEx publication.[89]

Data from version 7 of GTEx (the current public release) has been used in this thesis. Quantified gene count data and sample phenotype data were obtained directly from dbGaP with accession phs000424.v7.p2. Raw data processing, RNA-seq mapping, and quantification were performed by GTEx analysts.[89] As a member of the GTEx Analysis Working Group, I participated in discussions surrounding methods, RNA-seq mapping, and quality control, however I did not directly participate in preparation of the publicly available GTEx data.

General characteristics, filtering, and normalisation of subsets of GTEx samples used for analyses are described in detail in the relevant chapters.

2.2 InsPIRE islets

The Integrated network for systematic analysis of Pancreatic Islet RNA Expression (InsPIRE) collaboration is a consortium effort combining published islet RNA-seq and genotype data from four published sources[96,98,100,102] along with data from further samples from the same laboratories. In total, 442 samples have available RNA-seq data, and 426 of these samples have corresponding genotype data. Methods of procurement differed for each of the four contributors and have been previously published:

- Finland-United States Investigation of NIDDM Genetics (FUSION):[102] 39 samples, all with genotype
- Geneva, Switzerland:[96] 25 samples, 22 with genotype
- Lund, Sweden:[98,115] 204 samples, 191 with genotype
- Oxford, United Kingdom:[100] 174 samples, all with genotype

To create consistent data that could be jointly analysed, contributors used similar genotyping and RNA-seq protocols, and data were processed, mapped, and quality controlled jointly as described below. As a member of the InsPIRE collaboration through my work in Oxford, I have performed some analyses and taken part in regular discussions surrounding methods and results, but have not directly contributed to the procurement, preparation, mapping, quantification of RNA-seq, genotype imputation, or first phase genotype filtering.

The subsets of samples and data used for various analyses are described in detail in the relevant chapters.

2.2.1 Genotype imputation

Genotype imputation was led by Dr Martijn van de Bunt and Dr Ana Viñuela, and I did not directly contribute to it. All samples had been genotyped on either OmniExpress or Omni2.5 genotype arrays as described in contributor publications mentioned above. Genotypes were separated into two panels based on the array used, and each panel was quality-controlled by removing variants based on the following exclusion criteria:

- Minor allele frequency (MAF) < 0.05
- Genotype success rate < 0.95
- Palindromic with MAF > 0.4
- Hardy-Weinberg equilibrium $p < 10^{-6}$
- Not in 1000 Genomes reference panel[44]
- Inconsistent alleles compared to 1000 Genomes reference panel
- Inconsistent rsID mapping compared to 1000 Genomes reference panel
- MAF differing by $> 20\%$ compared to 1000 Genomes reference panel

The panels were each pre-phased with SHAPEIT v2[176] using IMPUTE2-supplied genetic maps.[177] Each panel was then imputed with IMPUTE2 v2.3.1 using the 1000 Genomes Phase 1 integrated variant set reference panel. Variants with INFO score ≤ 0.4 or with Hardy-Weinberg equilibrium $p \leq 10^{-6}$ were excluded. A combined variant call format file was created using the intersect of quality-passed variants in each panel. Genotype dosages were finally calculated using either the imputation probabilities or genotype calls (for genotyped variants).

2.2.2 RNA sequencing

Each contributor used poly-A-tailed RNA selection (PA) to prepare RNA, and all RNA was sequenced using either 100 base pair (bp) or 101bp paired-end reads. RNA-seq mapping and quantification were performed by other members of the

InsPIRE collaboration, and I did not directly participate in this work. Reads for all samples were mapped to the GRCh37 reference genome using the GEM mapper,[178] and exon-level quantifications were calculated based on GENCODE 19 annotations.[126] For overlapping exon definitions within a gene, the exons were merged to a single meta-exon for quantification.

After the initial quantifications were completed by members of the InsPIRE collaboration, since RNA-seq was mapped using a non-strand-specific protocol, I further removed any pair of exons from different genes that had overlapping genomic coordinates to eliminate the mapping ambiguity that may result from such exons.

2.3 Oxford islets using ribosomal RNA depletion

Human islet samples of 55 non-diabetic individuals were sequenced by the Oxford Genomics Centre in Oxford after ribosomal RNA depletion (RD) and have been used for various analyses in Chapter 3. Twenty-two of these samples came from individuals that were also part of the InsPIRE panel. Procurement of samples has been previously described.[100] In short, 38 samples were procured from the Oxford Centre for Islet Transplantation, and 17 were from Edmonton, Canada (7 extracted and thawed from the long-term cryo-preserved biobank[179] and 10 freshly isolated from donor pancreas[180]).

Since many islet samples had already been sequenced for our research group after PA (as described previously[100]), a subset of 38 of these (26 from Oxford, 3 Edmonton cryo-preserved, and 9 Edmonton freshly isolated) that had already been sequenced using PA (with measured strand orientation) and had enough remaining high-quality RNA for further RNA-seq were sequenced again after RD to create paired libraries for comparisons. In addition to these 38 samples with paired RNA-seq, 17 additional islet samples underwent RD and RNA-seq to build up the

dataset of available total RNA libraries from human islets.

RD, RNA preparation, and RNA-seq for all 55 samples were performed by the Oxford Genomics Centre in two separate batches as follows.

The first 20 samples were normalised at 500ng total RNA input for library preparation using Illumina TruSeq Stranded Total RNA Library Preparation Kit following RD using Ribo-Zero Gold rRNA Removal Kit (Human/Mouse/Rat). Samples were then pooled into two groups of 10 samples, each of which was sequenced on 3 Illumina HiSeq4000 lanes with 75bp paired-end reads. At a later date, the 35 remaining samples were normalised at 300ng total RNA input and prepared using the same protocol as the first 20. These 35 libraries were pooled together and sequenced on 8 Illumina HiSeq4000 lanes.

RNA-seq reads resulting from both RD and PA protocols were initially mapped and quantified by Dr Martijn van de Bunt, a post-doctoral fellow in the McCarthy Group at the time. Dr van de Bunt aligned reads to the GRCh37 reference genome using STAR 2.5.1b,[181] and further quantified gene-level counts based on GENCODE 19 annotations[126] using featureCounts 1.5.0-p2.[182] Further quantifications that I performed based on alternative transcriptome references are described in detail in Chapter 3.

Chapter 3

The expression profile of lncRNAs and limitations in their detection

3.1 Introduction

Before exploring potential function and regulatory mechanisms of long noncoding RNAs (lncRNAs) related to type 2 diabetes (T2D) and glycaemic traits, it is first important to understand their general expression landscape in humans, how this differs from messenger RNAs (mRNAs), and what limitations these differences impose for lncRNA studies in existing datasets.

3.1.1 Expression landscape of lncRNAs

Identification of noncoding genes was generally difficult until the widespread introduction of next-generation sequencing in the mid-2000s.[121] In contrast to microarray RNA quantification, RNA sequencing (RNA-seq) requires no prior knowledge of transcripts. This enabled identification of expressed noncoding genomic regions that were not previously known. Two primary large-scale annotation efforts have identified many noncoding RNAs in humans: The Reference Sequence (RefSeq) Database,[183] and the GENCODE collaboration.[126]

While the RefSeq database contains reference sequence data for more than 55,000 organisms,[183] the GENCODE collaboration, which is an extension of the Encyclopedia of DNA Elements Project, focusses specifically on reference transcriptomes for humans and mice.[126]

While both RefSeq and GENCODE provide vast databases of human exons, transcripts, and gene definitions, the GENCODE reference transcriptome provides annotations for a significantly greater number of unique exons and has greater genomic coverage for humans than does the RefSeq database.[184] The greater coverage and detail provided by GENCODE has resulted in its annotations being used for RNA quantification in many modern large-scale RNA-seq studies, including The Genotype-Tissue Expression Project (GTEx).[89] GTEx version 7 uses GENCODE version 19, which is the most recent GENCODE version based on Genome Reference Consortium human build 37.

GENCODE version 19 (`http://www.gencodegenes.org/releases/19.html`) provides a wide range of gene annotations including 20,345 coding genes and 13,870 lncRNA genes. Although the number of annotated coding genes has decreased from the earliest available version of GENCODE (3c, July 2009), the number of annotated lncRNAs has more than doubled from version 3c to 19. In another recent advance to lncRNA annotation, Hon *et al.* used RNA-seq from diverse human tissues and cell-lines ($n = 70$) to identify 27,919 potential lncRNA genes in humans.[185] These vast databases of lncRNA annotations coupled with the increasing affordability of high-throughput RNA-seq have led to wide study of lncRNAs in humans in recent years.

Using RNA-seq from different tissue and cell types, several studies have observed two primary expression patterns that are widespread for lncRNAs: lncRNAs tend to be expressed at lower levels than mRNAs, and lncRNAs tend to be expressed in a more tissue-specific manner than mRNAs.[186] One relatively early study across multiple tissue types observed lower expression of lncRNA genes than

of coding genes using RNA-seq data from human samples of 16 tissues types (3 samples per tissue type),[187] and a more recent study, which used RNA-seq data from 94 total samples across 20 tissue types, also found general low expression levels of lncRNAs when compared to mRNAs.[186] Analyses by Melé *et al.* using GTEx pilot data (9 tissue types) found that a substantial proportion of lncRNAs were expressed between 0.1 and 1 read per kilobase per million mapped reads, suggesting that low expression thresholds should be used when analysing lncRNAs.[188] Although the most recent core GTEx paper extensively assessed tissue specificity and tissue sharing of expression quantitative trait loci across tissues in GTEx version 6p,[89] a general transcriptomic overview has not been performed using data from recent GTEx releases. It has also been shown using *de novo* lncRNA assembly that existing reference transcriptomes (GENCODE, RefSeq, and Cabili *et al.*[189]) tend to capture only lncRNAs with high expression, and that there are potentially many more low-expressed lncRNAs that have not yet been annotated.[131,190]

A study by Derrien *et al.*[187] found lncRNAs to be expressed less ubiquitously than mRNAs, noting that 65% of coding genes were transcribed in all 16 analysed tissues, while only 11% of lncRNA genes were expressed in all tissues. Derrien *et al.* also observed higher expression variability between tissues for lncRNAs.[187] A study of lncRNAs by Jiang *et al.* similarly found 12% of lncRNAs to be expressed across all 20 tissues studied, and identified 14% of lncRNAs as tissue-specific.[186] Melé *et al.* also found greater tissue specificity for lncRNAs than for coding genes using GTEx pilot data.[188] Jiang *et al.* argue that an observed correlation between expression magnitude and tissue ubiquity suggests that widely expressed lncRNAs may be the most important part of the lncRNA transcriptome. However, Derrien *et al.* have raised the important consideration that this may be due to the fact that the general low expression of many lncRNAs may lead to false negative detection in some tissues simply due to sequencing depth and strict expression cut-off values.

Given the uncertainty in the true level of tissue specificity of lncRNAs, and that low expression and strict cut-offs may bias these estimates, it is surprising that the wide spectrum of RNA-seq data from recent versions of GTEx has not been used to analyse these patterns in detail. GTEx version 7 (the current public release) contains RNA-seq from a total of 53 tissue types collected from 714 donors, which has been sequenced at a median depth of 80 million total reads (`www.gtexportal.org`[89]). Despite GTEx tissue samples being sequenced using a poly-A-tailed RNA selection (PA) protocol, which may hamper detection of lncRNAs (discussed in greater detail below), the deep sequencing and breadth of tissue types make it an ideal data source for determining lncRNA specificity and general expression profiles. Further, all sample procurement, preparation, and sequencing protocols for all tissues in GTEx follow a common operating procedure, creating a homogeneous platform for assessing varying lncRNA effects across tissues with minimal technical bias. I will therefore use data from GTEx version 7 to further assess and expand on patterns of lncRNAs previously described.

3.1.2 Detection and quantification of lncRNAs

As described in Chapter 1, most public large-scale sequencing efforts perform RNA preparation using PA, which eliminates all non-polyadenylated RNAs before sequencing. It has been shown that lncRNAs do not always harbour poly-A tails,[191] and it has been suggested that the total number of non-polyadenylated lncRNAs has been underestimated due to the use of PA for RNA preparation in large-scale RNA-seq projects.[192]

The primary alternative method of RNA preparation used in modern studies is ribosomal RNA depletion (RD), which specifically targets and depletes only ribosomal RNAs (rRNAs). A previous direct comparison of lncRNA quantification between samples using PA and RD in human embryonic kidney (HEK) 293 T cells identified 447 lncRNAs that were detected using the RD protocol, but not the PA

protocol.[143] The same study also identified 324 lncRNAs using the PA protocol that were not detected using RD. Given that RNAs sequenced after PA should theoretically be a subset of those sequenced using RD (assuming appropriate adjustment for equivalent read depth), it is likely that the lncRNAs detected only with the PA protocol were low expressed genes that did not meet expression cut-off thresholds amongst total RNA, since reads in the study by Sultan *et al.* were quantified as reads per kilobase per million mapped reads, which divides counts by the total number of sequenced reads (a much larger relative denominator for RD than for PA).

Eight primary studies have compared multiple general aspects of PA and RD protocols using paired samples, seven of which had eight or fewer unique samples for comparison,[140,141,143,145–148] and the largest of which performed library preparation for 17 samples using both methods.[144] Studies are in general agreement that expression quantifications after PA and RD are correlated, although there were some discrepancies between findings. While some of these studies found that RD led to higher genomic mapping rates of reads,[143,145,147] others found PA to result in slightly higher mapping rates.[144,148] The studies consistently reported significantly higher exonic mapping rates using the PA protocol and a significantly higher proportion of reads after RD mapping to introns and intergenic regions. While intronic reads in many cases are likely to result from unprocessed pre-mRNAs, the higher intergenic mapping rates resulting from RD may be a result of uncharacterised non-polyadenylated noncoding RNAs.[143]

Given these findings and the fact that samples from both GTEx and the Integrated network for systematic analysis of Pancreatic Islet RNA Expression (InsPIRE) collaboration used in Chapters 4-6 have all been sequenced using a PA protocol, it is important to understand and discuss the extent to which the library preparation method may limit the power to detect lncRNAs. The sample sizes used for the studies described above were relatively small, thus I have performed a

similar comprehensive analysis of the strengths and weaknesses of each protocol using 38 human islet samples with paired PA and RD preparation. This amounts to more than double the samples from the next largest direct comparison of the preparations by Zhao *et al.*,[144] and results will be directly relevant and applicable to downstream analysis since human islets are the primary tissue of focus in later chapters.

3.1.3 LncRNAs in human islets

The potential limitations of lncRNA detection using PA can have a compounding effect on downstream lncRNA studies. Given the general tissue specificity of lncRNAs along with knowledge that PA eliminates the possibility of sequencing lncRNA transcripts that are non-polyadenylated, any non-polyadenylated lncRNA that is specific to one tissue or cell type will not have been annotated unless transcriptomic assemblies have been completed using total RNA of that specific tissue or cell type. Given that total RNA assemblies have not been created comprehensively for human tissues, it is likely that there are many unannotated lncRNAs in humans which will not be detected in studies that base quantifications on existing reference transcriptomes such as GENCODE.[193]

Several methods exist for detecting unannotated expressed regions and transcripts using RNA-seq, including Trinity,[194] Oases,[195] Cufflinks,[196] and StringTie.[197] StringTie is the most widely applicable method of *de novo* transcriptome assembly for modern RNA-seq studies, since it has been shown to more accurately reconstruct known transcripts than other methods,[197] and unlike its predecessor Cufflinks, is fast and computationally scalable to modern large RNA-seq datasets.[193] Some studies, rather than using existing methods for *de novo* gene and transcript assembly, have developed in-house methods that rely on RNA-seq read pile-ups and additional annotations, such as the methodology followed by Morán *et al.*, which tested for RNA pile-ups and assessed coding

potential, proximity to annotated genes, and H3K4me3 enrichment to identify unannotated lncRNAs expressed in human islets.[137]

The analysis by Morán *et al.* found that more than 19% of RNA-seq reads that mapped to the genome were outside of protein-coding regions. They further identified a total of 1,128 unannotated noncoding islet transcripts, 335 of which were intergenic. For their analysis, they used a pooled sample consisting of RNA-seq from three human islet PA libraries, three human islet RD libraries, two sorted beta-cell libraries, and two mouse islet libraries.[137] Given the small sample size, a recent study by Akerman *et al.* built on these annotations by putting approximately 5 billion pooled RNA-seq reads from RD libraries of 41 human islet samples through the Cufflinks method.[103] Using a similar methodology to the one described by Morán *et al.*, and combining their novel annotated lncRNAs with those from Morán *et al.*, they compiled a list of 2,226 islet lncRNAs.

The assembly by Akerman *et al.* is the most comprehensive *de novo* assembly of islet lncRNAs to date, and although it has not been replicated using independent data, it suggests an abundance of expressed lncRNAs in islets outside of previous annotations.[103] It is clear from all studies described throughout this section that many lncRNAs appear to be highly tissue-specific, and the detection of lncRNAs in general depends on a combination of the library preparation protocol used, the depth of sequencing, and the reference annotations used for RNA quantifications.

Given that lncRNAs are often low expressed, tissue-specific, non-polyadenylated, and incompletely annotated, this chapter has the following aims:

- Characterise the expression profiles of lncRNAs and how they differ from mRNAs across a spectrum of deep-sequenced human tissues from GTEx.
- Compare RNA-seq data from RD and PA libraries and assess the impact of PA on quantification of lncRNAs.

- Define a robust set of unannotated lncRNAs that are expressed in human islets to estimate the abundance of lncRNAs that may be missed by gene quantification based on GENCODE reference transcriptomes.

3.2 Samples and methods

3.2.1 GTEx tissues

GTEx is described in detail in section 2.1. RNA-seq of 43 tissue types (hereafter simply referred to as 'tissues') from GTEx version 7 were used in this chapter. Of the 53 tissues from GTEx with available RNA-seq, these 43 represented those that had a sample size $n \geq 50$ and were not specific to one sex. There were 48 tissues that met the sample size cut-off of $n \geq 50$ including the sex-specific tissues ovary, prostate, testis, uterus, and vagina. These five tissues were also removed for the following reasons:

- Analysing only tissues common to both sexes creates a more homogeneous environment for cross-tissue and joint-tissue analysis (identical covariate adjustments).
- The transcriptomic landscape differs significantly between sexes,[198,199] thus sex-specific tissues could bias the general transcriptomic overview presented here.
- The majority (43/53) of GTEx tissues had $n \geq 50$ and were not sex-specific, so a wide spectrum of tissues is still covered using these criteria.

Quantified read counts and phenotype information were downloaded from dbGaP as described in section 2.1, and read counts were filtered and normalised independently for each tissue. Read counts were first filtered to keep only genes with both > 6 gene reads and > 1 count per million reads in at least 10 samples

(thresholds based on those used by GTEx[89]). For remaining genes in each tissue, 1 raw count was added to each gene for each individual to eliminate taking the logarithm of 0 during normalisation. Trimmed mean of M values normalisation factors[200] were calculated using the `calcNormFactors` function from *edgeR* version 3.16.5[201] and applied to samples to normalise library sizes. Expression values were finally transformed using the `voom` function from *limma* version 3.30.9.[202]

To capture hidden systematic variation in RNA-seq expression values, surrogate variables (SVs) were calculated for each tissue using the `sva` function from *sva* 3.22.0 in R.[203] SV analysis is designed to calculate sources of hidden variation due to unknown, unmeasured, or complex factors in a manner similar to principal component analysis.[204] Unlike principal component analysis, SV analysis allows incorporation of explicitly known variables and ensures that the expression heterogeneity explained by the SVs is distinct from specified known covariates.[204] SVs were identified using a model explicitly defining sex, and residual expression values after regressing out all SVs and sex were used for all analyses in this chapter.

3.2.2 Pancreatic islets

RNA-seq from human islet samples of 55 individuals were used for analyses, as described in section 2.3. Of these 55 samples, 38 were sequenced using both RD and PA, and the remaining 17 were sequenced only after using RD. Primary alignment and gene quantification were performed by Dr Martijn van de Bunt as described in section 2.3 unless otherwise specified.

For each analysis, gene counts were filtered as described in section 3.2.1, except that there was no filtering based on raw read counts, and there was a requirement for a gene to be expressed at > 1 count per million reads in at least 5 individuals, rather than the cut-off of 10 individuals used for GTEx tissues. The requirement for expression in a smaller number of individuals for islets compared to GTEx tissues was due to the overall small sample size in islets compared to the majority

of GTEx tissues.

3.2.3 Quantification of mapping rates

Mapping rate metrics were based on GENCODE 19 feature definitions[126] and calculated using the `CollectRnaSeqMetrics` function from Picard tools version 2.10.9 (`http://broadinstitute.github.io/picard`). `CollectRnaSeqMetrics` uses a simple method of counting the number of bases that overlap each type of genomic annotation: coding sequence, untranslated exonic region, intronic region, rRNA region, or intergenic region.

3.2.4 Novel gene annotation

Detection of novel expressed regions was performed using StringTie.[197] StringTie is a method of *de novo* transcript assembly based on a network flow algorithm, and it has been shown to correctly assemble significantly more transcripts with both real and simulated RNA-seq data when compared to other methods.[197] StringTie also has the advantage of accepting a base reference transcriptome, which when supplied, forces the program to prioritise reads to these known annotated genes, and only uses reads that are not explained by the reference for novel transcript construction.

StringTie was run using the workflow suggested in its manual (`https://ccb.jhu.edu/software/stringtie/index.shtml?t=manual`):

i) StringTie version 1.3.3b was first run independently for each sample using the GENCODE 19 reference annotation (`-G` flag).

ii) StringTie was run using the `-merge` flag to combine references from i) into a non-redundant set of transcripts with a minimum transcript length of 200 base pairs (bp; `-m 200`).

iii) The gffcompare program[197] was used to merge the reference file from ii) with the GENCODE 19 reference file to assign a class code to each transcript.

Final transcripts were filtered to keep only previously annotated transcripts and unannotated intergenic transcripts (class codes '=' and 'u' respectively). Unannotated intergenic transcripts were further filtered to remove those with gene boundaries that fell within 10,000 bases of a known, previously annotated GENCODE gene. Reads were then re-quantified at the gene level based on this new reference transcriptome using featureCounts 1.5.0-p2,[182] and filtered and normalised as in Section 3.2.2.

3.3 Results

To emphasise the importance of performing lncRNA studies in relevant tissues, section 3.3.1 first highlights the tissue-specific nature of lncRNA expression. Given this demonstrated lncRNA tissue specificity, sections 3.3.2 and 3.3.3 then assess potential limitations in lncRNA detection and quantification specifically in islets, which is the primary tissue type of focus throughout the remaining chapters.

3.3.1 LncRNA tissue specificity

To understand the extent to which lncRNA analysis depends on the tissue type being studied, I used RNA-seq data from 43 GTEx tissues to assess lncRNA tissue specificity, and compared this to mRNA specificity. This builds on existing studies that have found varying levels of lncRNA tissue specificity,[186,187] and using only data from GTEx — which has been collected, sequenced, and processed following a standard procedure — allowed for assessment of tissue specificity using a homogeneous dataset with little technical variation. Since the human islet data used throughout this thesis has been processed independently using different procedures, islet RNA-seq data will not be used in comparisons, as high variation

in RNA-seq is expected when preparation and sequencing are performed by different centres using different methodologies.[205] Thus, this analysis provides insight into general profiles of lncRNAs and assesses their general tissue specificity to emphasise the importance of studying lncRNA biology in disease-relevant tissues, rather than attempting to identify lncRNAs unique to islets.

After filtering out low expressed genes in each of the 43 tissues analysed (section 3.2.1), a total of 27,069 genes were expressed in at least one tissue, representing 23 unique gene biotypes (Table 3.1). All GENCODE biotypes are described in full at `www.gencodegenes.org/gencode_biotypes.html`. There were 18,059 mRNAs expressed in at least one tissue, and 6,088 lncRNAs expressed in at least one tissue. Sixty-two percent of expressed mRNAs were expressed in all tissues, while only 11.4% of expressed lncRNAs were ubiquitously expressed (Figure 3.1). These proportions differed significantly ($p < 10^{-15}$), and were similar to a study of 16 tissue types by Derrien *et al.* that found 65% of coding genes and 11% of lncRNAs to be ubiquitously expressed,[187] and to a study by Jiang *et al.* that found 12% of lncRNAs to be expressed across 20 tissue types.[186] Across all biotypes, I found 12,208 genes expressed in all tissues. Given that thyroid had the maximum number of expressed genes (20,229), this means that a minimum of 60.3% of the genes expressed in any given tissue were ubiquitously expressed across all tissues. For Epstein-Barr virus transformed lymphocytes, the 'tissue' with the fewest expressed genes (15,750), 77.5% of its expressed genes were expressed in all tissues.

On the other end of this spectrum, there were 3,156 lncRNAs (51.8%) that were expressed in 10 or fewer tissues, with a smaller proportion of coding genes (9.5%) expressed in 10 or fewer tissues ($p < 10^{-15}$). It was also observed that 966 lncRNAs (15.9%) were expressed in a single tissue, while significantly fewer coding genes (303; 1.7%) were expressed in only one tissue ($p < 10^{-15}$; Figure 3.1). The U-shaped distributions of expression across tissues are similar to those previously

Category	Biotype	Count	Total count
protein coding		18,059	18,059
lncRNA	long intergenic noncoding RNA	2,916	6,088
	antisense	2,482	
	sense intronic	290	
	processed transcript	267	
	sense overlapping	126	
	3' overlapping	7	
other	pseudogene	2,468	2,922
	immunoglobulin V	131	
	T-cell receptor V	82	
	miscellaneous RNA	52	
	immunoglobulin V pseudogene	47	
	small nucleolar RNA	33	
	polymorphic pseudogene	27	
	microRNA	17	
	small nuclear RNA	13	
	immunoglobulin J	13	
	immunoglobulin C	12	
	T-cell receptor J	8	
	T-cell receptor C	5	
	immunoglobulin D	5	
	immunoglobulin C pseudogene	5	
	T-cell receptor V pseudogene	4	

Table 3.1: Biotypes of genes expressed in at least one tissue. The number of unique genes of each biotype and broad category that were expressed in at least 1/43 GTEx tissues analysed. Biotype definitions are from GENCODE version 19.

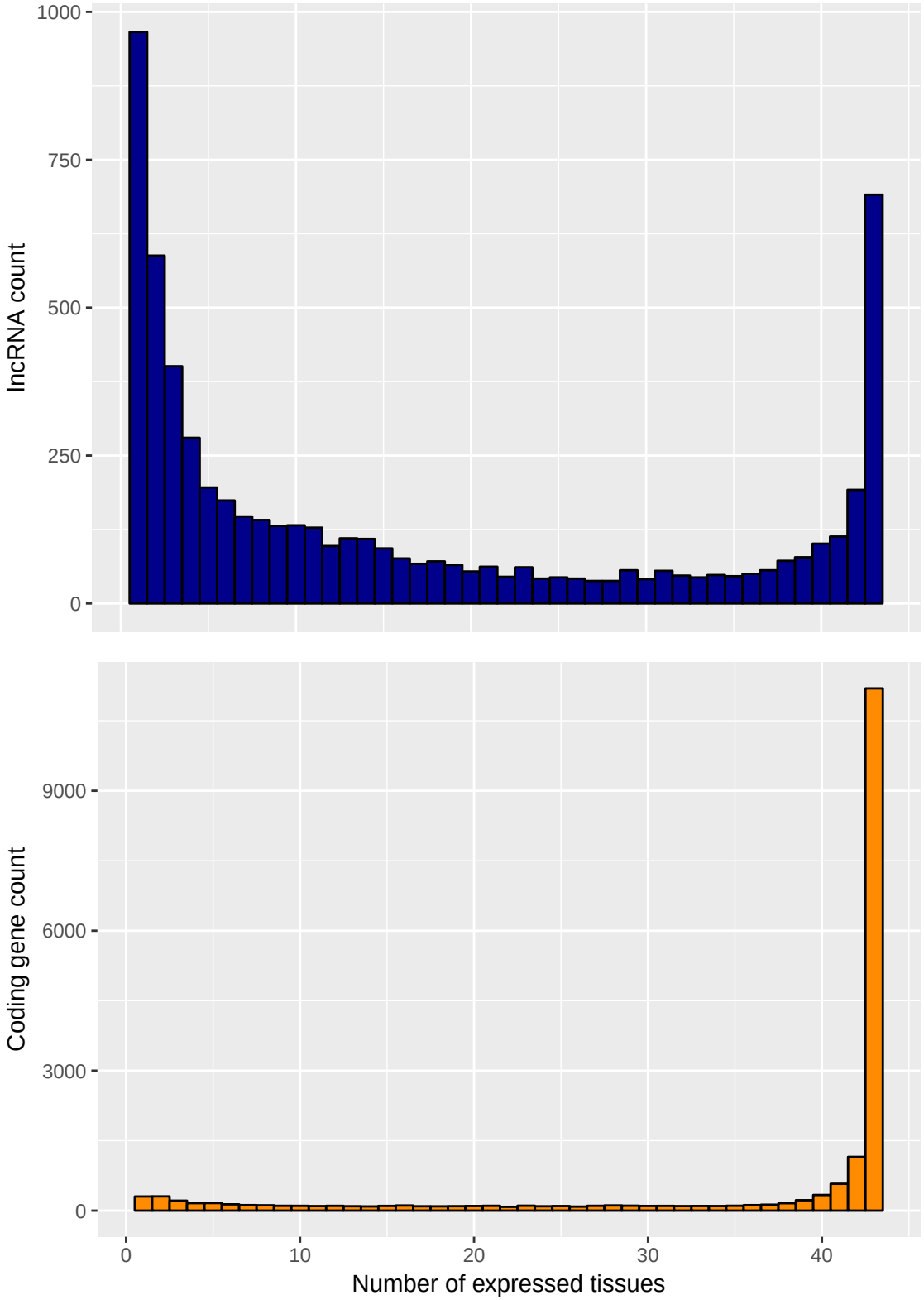


Figure 3.1: Number of tissues in which each unique lncRNA and coding gene is expressed. Histogram showing the number of tissues from GTEx (x-axis) in which each unique lncRNA gene (blue) and coding gene (orange) is expressed.

published by Melé *et al.* using GTEx pilot data. Liver had the highest number of lncRNAs that were expressed only in that tissue (103), and the proportion of unique lncRNAs expressed per tissue differed significantly across tissues (χ^2 test $p \approx 0$). Tissue sample size was not significantly correlated with the absolute number or proportion of unique lncRNA genes expressed ($p = 0.057$ and $p = 0.088$ respectively), indicating that sample size was likely not a limiting factor in detecting unique lncRNAs (Figure 3.2). Sample size, number of expressed genes, and number of unique lncRNAs and coding genes expressed in each tissue are detailed in Table A.1.

The specific profiles of the two most abundant lncRNA biotypes, antisense and long intergenic noncoding RNA (lincRNA), were examined to identify any potential differences in tissue specificity (Figure 3.3). The tissue specificity distributions of lincRNAs and antisense lncRNAs were not independent of the lncRNA type (Chi-squared p=0.0002). However, this was due largely to the extremes of the distributions, as the same test, excluding genes expressed in either 1, 2, 42, or 43 tissues, was not significant (p=0.43). Compared to lincRNAs, a significantly higher proportion of antisense lncRNAs was expressed in all tissues (8.8% and 11.8% respectively; $p = 0.0003$), while the proportion of tissue-specific lincRNAs was significantly higher than the proportion of tissue-specific antisense lncRNAs (18.4% and 14.4% respectively; $p = 8.7 \times 10^{-5}$). These differing distributional tails may be due to the fact that antisense lncRNAs, unlike lincRNAs, have been shown to be expressed in patterns correlated to coding genes, albeit at lower magnitude.[206] There may therefore be a subset of antisense lncRNAs correlated closely with tissue-ubiquitous coding gene expression.

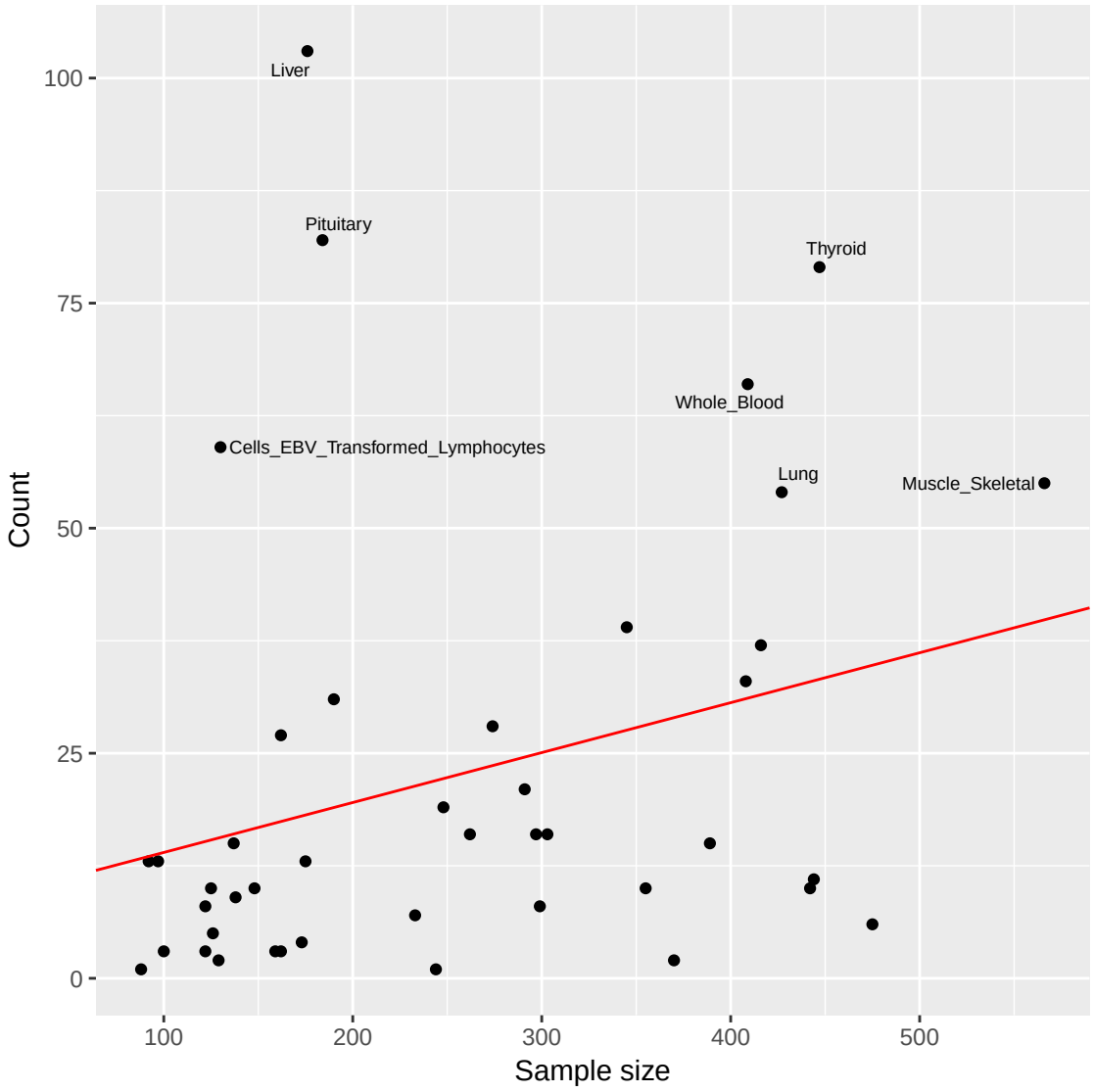


Figure 3.2: Correlation between sample size and detection of uniquely expressed lncRNAs. Scatterplot showing the correlation between tissue sample size (x-axis) and number of lncRNAs expressed only in that tissue (y-axis). Sample size and uniquely expressed lncRNA counts were not significantly correlated ($r = 0.29$, $p = 0.057$). The best-fit line is indicated in red.

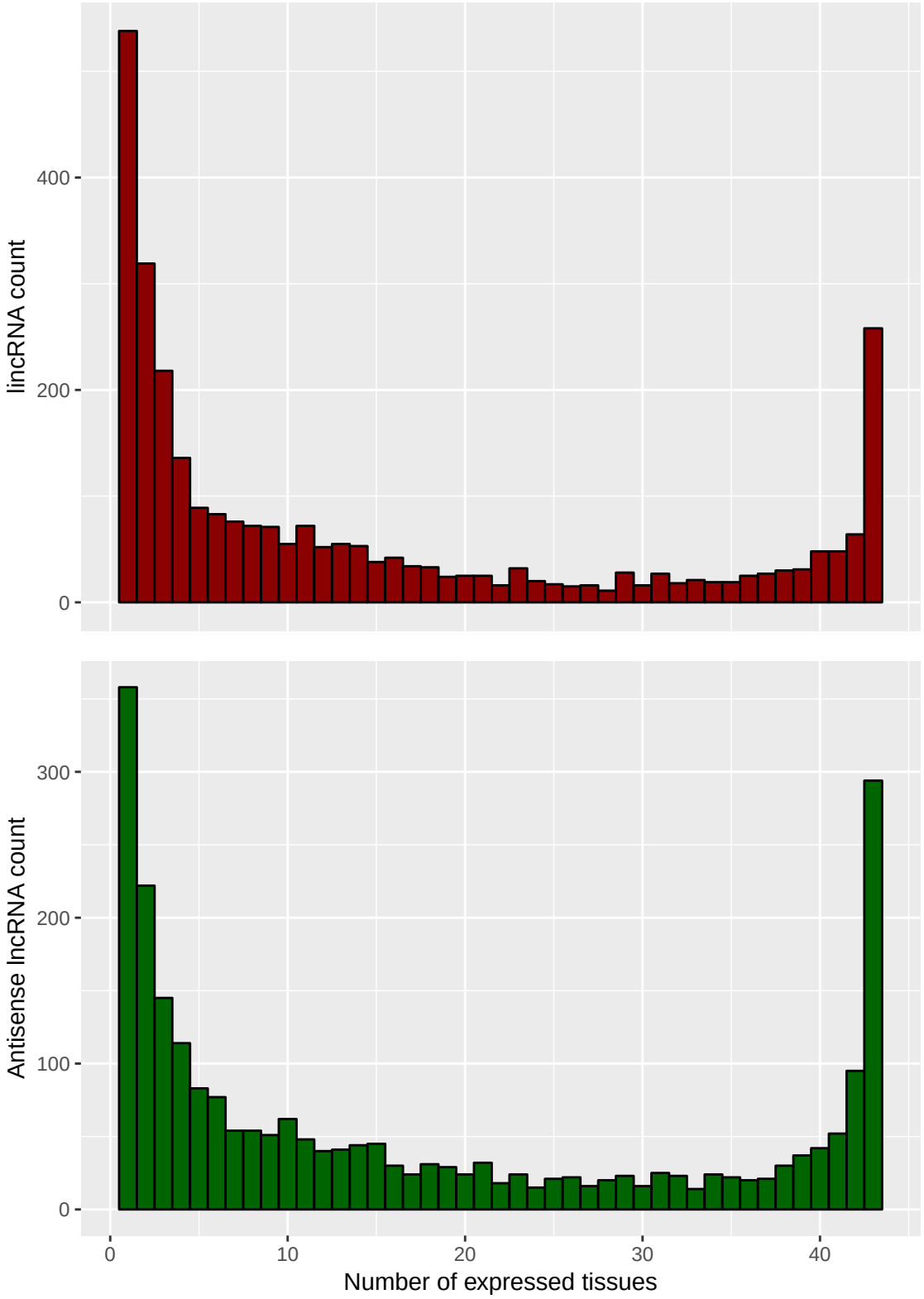


Figure 3.3: Number of tissues in which each unique lincRNA and antisense lncRNA is expressed. Histogram showing the number of tissues from GTEx (x-axis) in which each unique lincRNA (red) and antisense lncRNA (green) is expressed.

3.3.2 Comparison between rRNA depletion and poly-A-tailed RNA selection

Since both GTEx and InsPIRE used PA prior to RNA-seq and it has been shown that this preparation method limits detection of some lncRNAs,[143] I next directly compared PA to RD to assess strengths and limitations of each of these preparation methods. PA preparation kits selectively bind RNAs with poly-A tails and wash away remaining RNA, while RD preparation kits selectively target and deplete rRNAs (detailed description in Chapter 1). RNA from pancreatic islet samples of 38 humans was sequenced after using each of these two preparation methods (described in section 2.3), which is a larger dual-sequenced dataset than any of those described previously[140,141,143,144] . Having paired RNA-seq libraries for each sample allowed for individual- and aggregate-level comparisons of the preparation methods.

Read mapping

All 38 samples were sequenced using strand-specific protocols. The primary difference between the sets being compared was the read length: RNA-seq for PA samples used 100bp paired-end reads, while RNA-seq for RD samples used 75bp paired-end reads. The PA libraries had been completed prior to the design of this experiment, and at the time of sequencing for RD samples, 75bp was the read length used by the laboratory. It has been shown that shorter reads typically result in lower mapping rates,[207] so this consideration will be taken into account when interpreting results.

Basic read mapping statistics are described in full for all samples in Table A.2. The median read depth for PA samples was 62.7 million pairs of reads (range 26.6 to 77.6 million), while there were a median 81.7 million pairs of reads for RD samples (range 69.4 to 119.5 million). Given differences in read depth between PA samples and RD samples, mapping statistics will primarily be described as

proportions to enable direct comparison of performance between the preparations. Both methods successfully depleted rRNA, with a maximum of 216 reads mapping to rRNA in any sample-preparation pair. Samples using RD had a median read mapping rate of 95.7%, while samples using PA had a median mapping rate of 92.2%. Approximately 2.15/2.35 billion combined read pairs across all PA samples were successfully mapped (91.7%), while approximately 3.16/3.32 billion combined read pairs across RD samples were successfully mapped (95.5%). These significantly different mapping rates ($p < 10^{-15}$) are consistent with a previous study that also found higher mapping rates for samples using RD compared to samples using PA.[143] Given that lower RNA-seq read lengths typically lead to a lower read mapping rate within the same protocol,[207] the fact that RNA-seq of RD samples had significantly higher mapping rates despite shorter reads than the PA RNA-seq suggests that mapping rates for RD samples may have been even higher if sequenced with the same read length as PA samples.

The median proportions of mapped reads that were exonic were 65.2% and 89.4% respectively for RD and PA, with 65.3% and 87.2% of pooled mapped read pairs mapping to exonic regions for RD and PA respectively (Table 3.2). Respective median mapping rates to introns of annotated genes for RD and PA were 27.9% and 7.5%, and median intergenic mapping rates were 4.9% and 3.3% (Table 3.2). These proportions were all in line with other studies[140,141,143] and show that a higher proportion of reads after PA map to exonic regions compared to paired RD libraries. It is likely that the high rate of intronic reads after RD results from retention of unprocessed pre-mRNA transcripts.[143] This would partially explain the higher overall mapping rates of RD RNA-seq, since mature RNA reads that span exon boundaries (and therefore do not map continuously to a reference genome) have lower mapping accuracy than RNA-seq reads spanning a continuous DNA sequence (which would result from any unprocessed RNA).[181] The higher intergenic mapping rates for samples using RD may result from

	Total mapped	Exonic (%)	Intronic (%)	Intergenic (%)
PA	2.15	1.88 (87.2)	0.17 (7.8)	0.11 (5.0)
RD	3.16	2.07 (65.3)	0.87 (27.4)	0.23 (7.3)

Table 3.2: Mapping statistics for pooled reads. The number and percentage of reads mapping to: a) overall human reference, b) exonic regions, c) intronic regions, and d) intergenic regions. All read counts are in billions of pairs of reads, and percentages for exonic, intronic, and intergenic reads represent the proportion of successfully mapped reads.

non-polyadenylated transcripts that are yet uncharacterised.[143]

Read quantification

After assessing read mapping statistics above, I next analysed the read quantification rates for each protocol (the proportion of mapped reads that were successfully assigned to an annotated GENCODE exon). A median 80.9% of the exonic reads per RD sample were successfully quantified, compared to a median 86.0% quantification of exonic reads for PA samples. In total 1.63/2.07 billion mapped exonic read pairs across RD samples were successfully quantified (78.7%), while 1.61/1.88 billion mapped exonic read pairs across PA samples were successfully quantified (85.8%). There were four samples that had a higher exonic quantification rate using RD compared to PA.

The highest expressed genes were next compared between preparation methods. It would be expected to see *INS* (the insulin gene) and *GCG* (the glucagon gene) expressed at high levels, since human islets are made primarily of beta- and alpha-cells.[110] For PA samples, the gene responsible for the highest proportion of total reads was *INS* (2.4%), with *GCG* having the third highest proportion of reads (1.7%). For RD samples, *INS* and *GCG* were the fifth and seventh highest quantified genes (2.3% and 1.3% of reads respectively). For the RD samples, *RN7SL2*, *RN7SL1*, *RN7SK*, and *RPPH1* accounted for 8.4%, 8.0%, 4.8%, and 2.6% of total reads respectively (combined 23.8% of total reads). These four

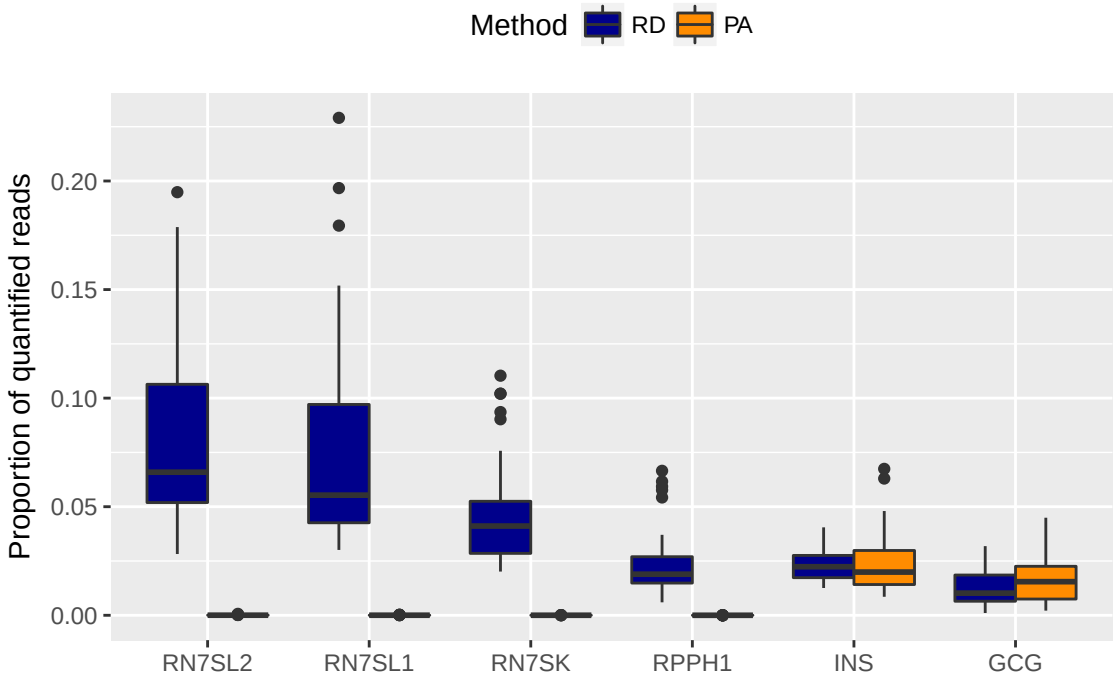


Figure 3.4: Distribution of read proportions mapping to top expressed genes after RD. The distribution of the proportion of reads from each sample (y-axis) mapping to *RN7SL2*, *RN7SL1*, *RN7SK*, and *RPPH1* (the four highest expressed genes in RD samples) as well as to *INS* and *GCG* (x-axis). Blue boxes represent RD samples, and orange represents PA samples.

abundant genes are each detected in the PA samples, albeit contributing a much smaller proportion of reads (0.007%, 0.003%, 0.001%, and 0.0003% respectively; Figure 3.4). This indicates that there may be both polyadenylated and non-polyadenylated versions of the transcripts, with the latter being more common.

Various subsets of these four genes have been identified as having extremely high expression using RD protocols in various tissues.[143,208,209] *RN7SL1* and *RN7SL2* (along with *RN7SL3*) encode the 7SL RNA component of the signal recognition particle,[210] *RPPH1* is an RNA component of the RNase P ribonucleoprotein,[143] and *RN7SK* encodes the 7SK RNA, which is abundant in humans and along with several proteins regulates RNA polymerase II activity.[211] Given the vast abundance of these genes in the RD samples, they were excluded for all remaining analyses in this section.

After filtering quantified reads as described in section 3.2.2, 19,604 genes remained for RD libraries, and 17,741 remained for PA libraries. Of the 17,741 genes passing filtering for the PA libraries, 17,265 (97.3%) also passed filtering for the RD libraries. Of these 17,265 genes that were expressed using both protocols, the expression correlation was high between the two preparation methods ($r = 0.949$), with higher correlation when comparing mRNAs only ($r = 0.950$), and lower correlation when comparing lncRNAs ($r = 0.861$; Figure 3.5). Many lncRNAs cluster toward the upper-left of Figure 3.5, indicating that these lncRNAs are detected at higher levels in RD samples than they are in PA samples. There were a total of 969 lncRNAs and 437 coding genes detected in only the RD samples, and 109 lncRNAs and 159 coding genes detected in only the PA samples (Table 3.3). A previous study by Sultan *et al.* that compared the protocols found 447 lncRNAs uniquely with RD and 324 uniquely with PA.[143] Interestingly only 10 of the lncRNAs detected uniquely with PA by Sultan *et al.* were detected uniquely by PA in my analysis, while 176 of the lncRNAs identified only using RD by Sultan *et al.* were also detected only in samples using RD in my analysis. The overall differences of identified lncRNAs between studies is likely due to the different tissue types used (HEK 239 cells by Sultan *et al.*), although the agreement of 176 lncRNAs being detecting using only RD in both studies provides robust evidence that these genes are transcribed and generally not polyadenylated.

Though many lncRNAs are detected using PA, the abundance of lncRNAs detected only in RD samples highlights the fact that lncRNA analyses using PA RNA-seq data are limited in their ability to assess lncRNA function. Results from this section do not imply that lncRNA function cannot be assessed using PA RNA-seq libraries. They instead show that analyses of lncRNAs using PA RNA-seq libraries are not exhaustive due to the inability to detect many lncRNAs after the removal of non-polyadenylated RNAs.

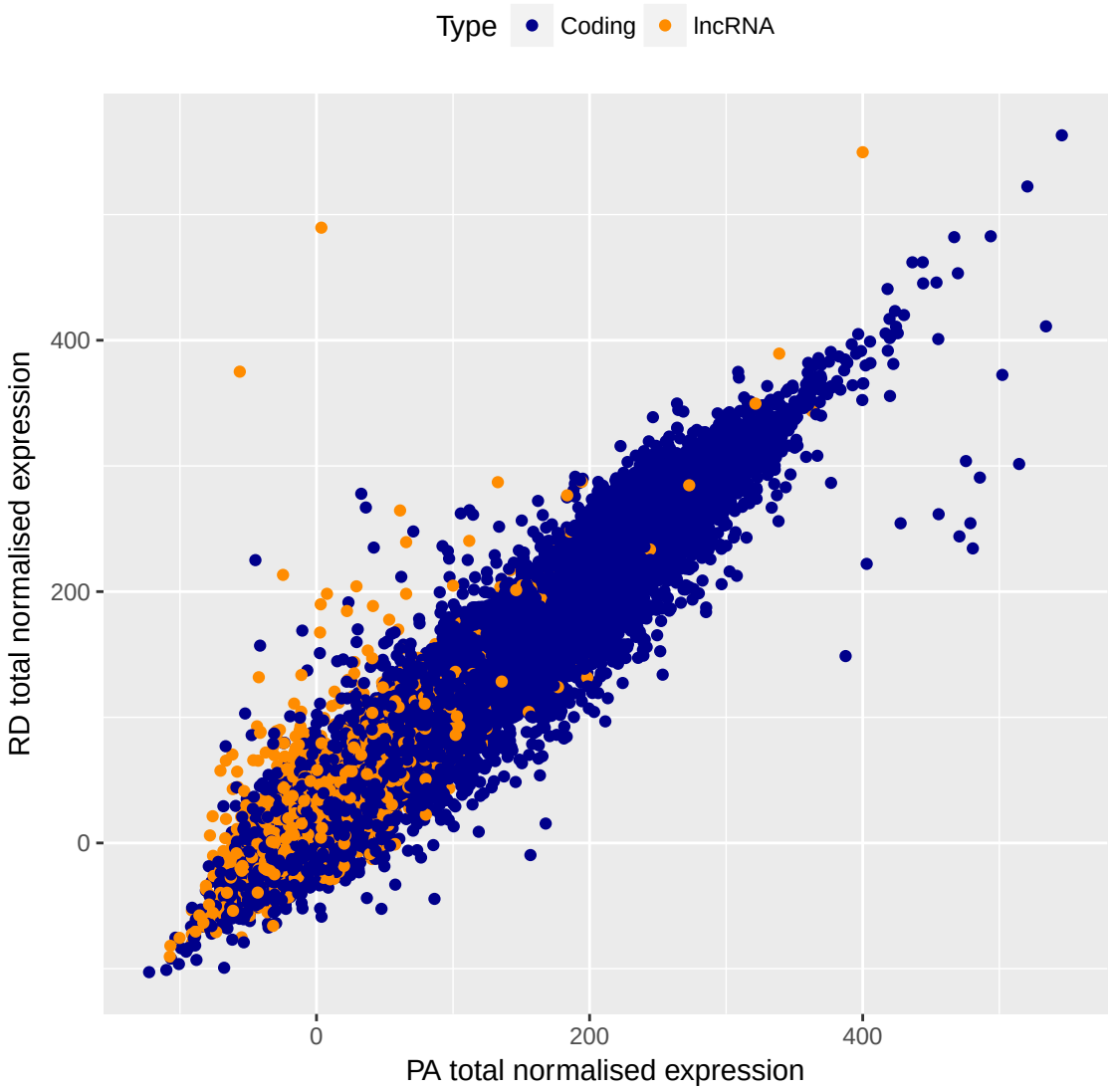


Figure 3.5: Correlation of RNA expression between RD and PA protocols. Each point represents one of the 17,265 genes that were quantified using both RD (y-axis) and PA (x-axis) protocols. Blue points represent mRNAs, and orange points represent lncRNAs. Both axes represent the sum of normalised counts combined across all samples.

Biotype	# only with RD	# only with PA
Coding	437	159
lincRNA	401	38
Antisense	293	64
Sense intronic	218	2
Processed transcript	32	4
Sense overlapping	22	1
3' overlapping	3	0

Table 3.3: Unique genes detected using each protocol. The number of coding genes and lncRNAs (separated by biotype) that were detected uniquely using each of the RD and PA samples.

3.3.3 Novel transcribed noncoding regions in islets

Samples and contamination

Given the above results that show RD leads to a higher percentage of reads mapping to intergenic regions, I next used RD samples to identify unannotated lncRNA genes in islets. This analysis provides an estimate of the abundance of robust lncRNAs expressed in islets that would otherwise be unobserved in quantifications based on GENCODE annotations. This section, along with the results of section 3.3.2, provide insight into the general sensitivity of lncRNA detection using GENCODE annotations and PA libraries, since all analyses in Chapters 4-6 use these protocols.

A total of 55 rRNA-depleted RNA-seq samples of human pancreatic islets were available for analyses, including the 38 used in Section 3.3.2. Sequencing and mapping of these 55 samples are described in section 2.3. One of the additional samples (not one of the 38 used in the above analysis) had an exceptionally low mapping rate (43.3% of bases aligned). The most abundant read sequences present in the RNA-seq for this sample were determined using FastQC (`http://www.bioinformatics.babraham.ac.uk/projects/fastqc`), and all unique read sequences constituting at least 0.3% of the total reads for the sample were tested against the nucleotide database in BLAST[212] to identify the source of

contamination (full list of tested reads in Table A.3). These sequences each mapped to at least one strain of *pseudomonas* bacteria. I then re-mapped the raw RNA-seq reads to the *pseudomonas aeruginosa* PAO1 reference genome[213] using STAR 2.5.1b,[181] and 9.17% of reads were successfully aligned. This suggests a high level of contamination, so the sample was excluded from all further analyses.

This contaminated sample raises an important concern when working with RNA-seq data. The same sample had been sequenced using an unstranded PA protocol, which showed no evidence of contamination (94.7% mapping rate to human). This likely represents a case where the raw tissue sample contained bacterial contamination, but due to the fact that most bacterial RNAs are not adenylated (and if they are harbour only short poly-A tails),[214] the contamination was not sequenced using the PA protocol. This has implications for both RNA preparation methods. It shows that RD is not robust to bacterial contamination, and that any contamination is likely to be sequenced and alter the effective sequencing quality and depth of samples. It also suggests that bacterial contamination may not be detectable when sequencing after PA. Given the high contamination rate, it is likely that bacterial contamination happened post cell death in this case. However, the finding suggests that with PA, in cases where live cells have bacterial infection or contamination that could cause significant changes in gene expression,[215] the samples are likely to be processed and analysed without ever being flagged.

Novel transcription

StringTie[197] was used to detect novel lncRNA genes (previously unannotated expressed intergenic regions) based on aligned RNA-seq data from the 54 remaining islet samples (description in section 3.2.4). StringTie was run using GENCODE version 19 as a reference transcriptome, which allowed the program to prioritise known transcripts and build on existing annotations. Since accurate

detection of novel isoforms from short-read data is difficult,[216] to be more conservative in detection, only novel genes detected in intergenic regions were carried forward, which eliminated the possibility of pre-mRNAs and known intronic regions being processed as novel genes.[155] In addition, any novel gene that fell within 10,000 bases of a known annotated gene was discarded, as these may be products of alternative cleavage, promoter- or terminator-associated transcription, or alternative exons of known genes.[155]

In total, 518 novel lncRNAs were identified (hereafter referred to as NIERs - novel intergenic expressed regions), containing 1,298 exons: 415 (80.1%) single-exon genes and 103 (19.9%) genes with multiple exons. In GENCODE 19, 37.9% of genes are single-exon genes, but the significantly higher portion of single-exon genes in the set of NIERs is consistent with a study that found 83% of a set of putative functional lncRNAs in humans and mice to contain only a single exon.[217]

Gene-level reads for all 54 RD samples were then re-quantified with featureCounts 1.5.0-p2[182] using a) the original GENCODE reference transcriptome annotations, and b) an expanded reference transcriptome annotation that included the original GENCODE annotations as well as the 518 NIER definitions. Read counts were then filtered based on expression and normalised as described in Section 3.2.2.

After filtering to remove low expressed genes, 99.9% of genes (19,512/19,533) detected using the unexpanded GENCODE reference were also detected with the expanded reference, suggesting that the additional NIERs did not capture reads that would have otherwise mapped to the original reference genes. Of the full set of 518 NIERs identified using StringTie, 218 passed final expression filtering thresholds, providing evidence that these regions are truly expressed in the data. The 300 remaining NIERs detected by StringTie did not meet my RNA-seq expression thresholds, suggesting that they are either expressed at extremely low

levels, or were false findings from StringTie.

The 218 NIERs that were expressed after filtering thresholds were applied were distributed across all chromosomes, ranging from 3 on chromosomes 3 and 21, to 31 on chromosome 2, and ranged in length from 203 to 219,899 base pairs. The ranks amongst all expressed genes of median expression values of NIERs ranged from 2 (second lowest expressed gene) to 19,293 (97.7th percentile), indicating that while some detected NIERs have low expression, others are highly expressed. A total of 197/218 expressed NIERs (90.4%) were amongst the 50% of detected expressed genes with the lowest expression, and the median expression rank of NIERs was 3,073.5 (∼16th percentile; Figure 3.6). While these numbers initially seem particularly low, they are nearly identical to the corresponding numbers for annotated lincRNAs: lincRNA ranks ranged from 12 to 19,747 (fifth highest expressed gene), 1,081/1,169 (92.5%) expressed lincRNAs were amongst the 50% lowest expressed genes, and the median rank of annotated lincRNAs was 3,963 (∼20th percentile; Figure 3.6). Although the overall rank distributions of NIERs and annotated lincRNAs differ significantly ($p = 8.04 \times 10^{-10}$), the properties listed above indicate some similarity between the expression profiles of NIERs and annotated lincRNAs.

Comparison to lncRNAs identified by Akerman *et al.*

A recent study by Akerman *et al.* performed *de novo* transcriptome assembly of lncRNAs using RNA-seq from rRNA-depleted human pancreatic islet samples of 41 individuals,[103] then combined their novel annotations with those previously published by Morán *et al.*[137] Given the similarity in sample size as well as the use of RD by Akerman *et al.*, I compared these published lncRNA annotations to NIERs from my assembly.

The publication identified a total of 1,297 novel lncRNA transcripts that were not detected as isoforms of known genes, and combined these with 727 islet

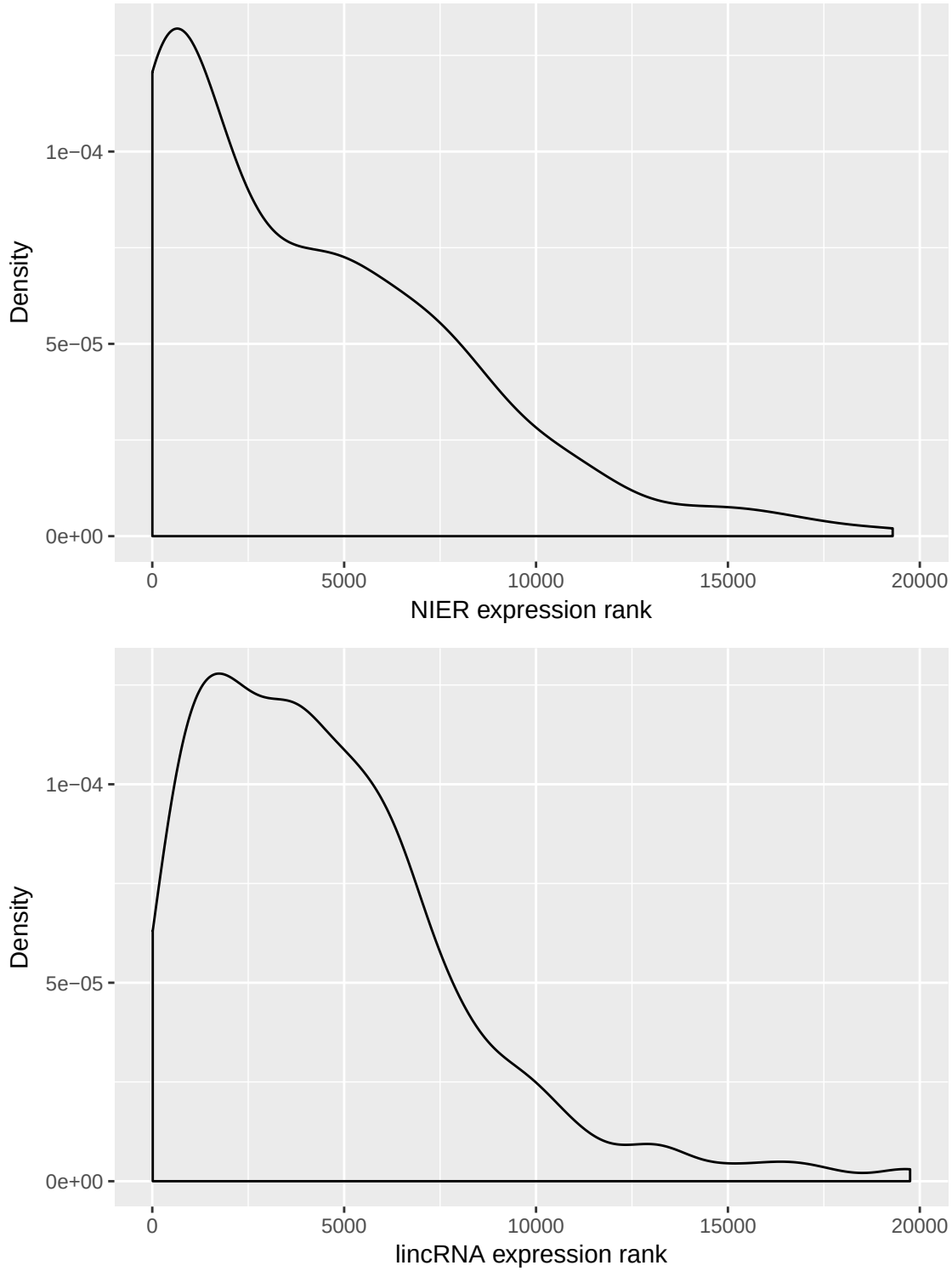


Figure 3.6: Density of NIER and lincRNA expression ranks. The density (y-axis) of the expression rank (x-axis) of all NIERs and lincRNAs passing expression filtering thresholds.

lncRNA definitions by Morán *et al.*,[137] resulting in a total of 2,024 transcripts. The goal of Akerman *et al.* was to identify an exhaustive list of potential lncRNA genes in islets, while the goal of my study was to assess the abundance of unannotated islet lncRNAs with minimal possibility of representing abundant pre-mRNAs, unknown exons of annotated genes, alternate isoforms, or transcription associated with promoters or terminators. Thus, to perform as direct a comparison as possible, of the 2,024 transcripts from Akerman *et al.*, only those that did not fall within 10,000 bases of genes annotated in GENCODE 19 were carried forward for comparison to NIERs from section 3.3.3.

A total of 47 lncRNAs from Akerman *et al.* were retained after distance filtering, including 14 originally described by Morán *et al.* A new transcriptome reference was created using GENCODE 19 genes as well as these 47 published islet lncRNAs, and all 54 samples from this chapter were re-quantified. Similar to the reference transcriptome augmented with NIERs in section 3.3.3, 99.9% of genes that were expressed using the original GENCODE transcriptome reference were also expressed when the 47 lncRNAs from Akerman *et al.* were included in the reference. Of the 47 lncRNAs added to the reference, 34 (72%) were expressed: all 14 from Morán *et al.*, and 20 from Akerman *et al.*

These 34 expressed genes were then tested for overlap with the 218 expressed NIERs described in section 3.3.3. Thirteen directly overlapped NIERs: 5 from Morán *et al.* and 8 from Akerman *et al.* One further gene from Morán *et al.* was within 10,000 bases of a NIER. The other 20/34 expressed genes from Akerman *et al.* are likely independent of NIERs identified in this chapter. The thirteen lncRNAs from Akerman *et al.* that directly overlapped NIERs were spread across 8 chromosomes, ranged from 1,105 to 23,721 nucleotides in length, and contained between 1 and 4 exons (summarised in Table 3.4).

It is important to note several differences in the study by Akerman *et al.* that likely led to detection of fewer intergenic lncRNA transcripts compared to my

Gene ID	Chr:Start-End	Source	Exon count
HI-LNC264	chr1:121484329-121485434	Moran *et al.*	1
HI-LNC329	chr2:43037341-43038974	Moran *et al.*	1
HI-LNC402	chr2:232469640-232474863	Moran *et al.*	1
HI-LNC85	chr4:43765001-43768392	Moran *et al.*	2
HI-LNC2668	chr4:129533106-129556827	Akerman *et al.*	4
HI-LNC2681	chr4:153510416-153513584	Akerman *et al.*	2
HI-LNC2966	chr7:24443379-24452592	Akerman *et al.*	2
HI-LNC3219	chr9:29214075-29225301	Akerman *et al.*	2
HI-LNC3590	chr12:25537044-25539285	Akerman *et al.*	2
HI-LNC968	chr13:95200353-95201818	Moran *et al.*	1
HI-LNC3791	chr13:108987693-108991843	Akerman *et al.*	2
HI-LNC4615	chr21:38932288-38937238	Akerman *et al.*	2
HI-LNC4616	chr21:38936330-38940985	Akerman *et al.*	3

Table 3.4: lncRNAs from Akerman *et al.* that overlap NIERs. Summaries of the 13 lncRNAs published by Akerman *et al.*[103] that directly overlapped genomic coordinates of at least one of the 218 NIERs defined in section 3.3.3.

analysis:

- Akerman *et al.* performed *de novo* assembly using Cufflinks,[196] while this chapter used StringTie, which was shown in its original publication to correctly assemble 53% more known transcripts than Cufflinks.[197]

- Akerman *et al.* first removed any fragment falling in the top 143 most expressed genes, while no pre-filtering was performed in this chapter. Although not explicitly stated, these fragments may have been removed due to a known issue in Cufflinks that causes the program to lag indefinitely if there are regions with exceptionally high numbers of reads (such as at *RN7SL2*). Given that some NIERs identified in my study were above the 95th percentile of expression magnitude amongst all expressed genes, this may have resulted in missed novel lncRNAs in the study by Akerman *et al.*

- Akerman *et al.* removed single-exon transcripts, while the majority (80.1%) of assembled NIERs in my study consisted of a single exon.

Despite these differences, the direct overlap of 13/47 lncRNAs published by

Akerman *et al.* with NIERs defined in my analysis provides robust evidence that these 13 islet lncRNA genes are likely real expressed regions. The identification of additional NIERs independent of those described by Akerman *et al.* (which itself was an expansion to definitions by Morán *et al.*) supports the notion that *de novo* assembly and replication testing using increased sample sizes and new datasets will allow for continued expansion and refinement of islet lncRNA annotations.

Although these results show that there are likely unannotated expressed lncRNAs in islets that are not defined in GENCODE, the comparison with the study by Akerman *et al.* implies that independent lncRNA detection efforts can lead to highly variable results depending on the sample set and methods used. The results presented here act as a reminder that in islets (and likely other cell types) there is likely true transcription happening outside of the known annotated transcriptome. Similar to limitations introduced when using PA compared to RD described in section 3.3.2, pre-existing general transcript references (such as GENCODE) may therefore create a limitation in lncRNA detection that needs to be considered in interpretation of downstream results. Thus, until robust tissue- and cell-type-specific assemblies are created from total RNA of large sample sets, lncRNA analyses are unlikely to be exhaustive.

3.4 Discussion

This chapter has characterised the overall expression differences between lncRNAs and mRNAs, has assessed downstream RNA-seq differences between two commonly used RNA preparation methods, and has followed this comparison with identification of unannotated lncRNAs in human islets. The main results of the analyses were as follows:

- Significantly more lncRNAs than mRNAs are tissue-specific, while more mRNAs are expressed across all tissues: thus lncRNA analysis in

disease-relevant tissue may be important even more so than for mRNA analysis.

- Compared to PA, RNA-seq after RD results in more intronic and intergenic reads, has reduced effective sequencing depth due to four highly expressed lncRNAs, and is more prone to sequencing bacterial contamination.
- mRNA quantification correlation is high between RNA-seq after RD and PA, but RD results in greater depth and quantification of lncRNAs.
- 218 unannotated NIERs in human islets were detected using RD and RNA-seq, and 13 replicated previously published *de novo* islet lncRNA annotations.

The first section of this chapter assessed the tissue uniqueness of lncRNA expression using 43 human tissue types from GTEx, and compared these properties with mRNAs in the same tissues. Although tissue specificity of eQTLs has been analysed by the GTEx Consortium,[89] a comparative analysis of the tissue specificity of lncRNAs and mRNAs as performed in my analysis has not been done. A minimum of 60% of the genes expressed in each GTEx tissue were ubiquitously expressed in all tissues, which is consistent with a past result finding the majority of mRNAs to be ubiquitously expressed across 16 tissue or cell types.[218] I further found that lncRNAs were generally expressed in fewer tissues than mRNAs. This trend has been identified in other studies,[135,143,219] but given that lncRNAs are generally expressed at low levels, it is difficult to differentiate between true tissue specificity and apparent tissue-specific effects caused by combined low expression and insufficient sequencing depth. One possible way to assess whether apparent tissue specificity of lncRNAs is due to sequencing depth would be to downsample RNA-seq libraries, and test whether the number of detected tissue-specific lncRNAs changes. Although pancreatic islet RNA-seq was not included in this analysis (to maintain high homogeneity between samples), the observed tissue

specificity of lncRNAs confirms the importance of studying lncRNA expression effects in disease-relevant tissues. This finding has led to the choice of performing downstream T2D- and glycaemic-trait-associated analyses using RNA-seq from human pancreatic islets, the tissue containing insulin-secreting beta-cells.

Since lncRNAs can be non-polyadenylated,[125] I next performed a thorough comparison of RNA-seq data resulting from two common RNA preparations: PA and RD. There were 38 islet samples available, each sequenced using both protocols, making it the largest direct pairwise comparison of the preparations. The RD protocol resulted in higher genomic mapping rates than the PA protocol, which is consistent with a previous study[143] and may result from an abundance of unprocessed RNAs that map at higher rates than spliced RNAs.[181] The increased mapping rate for samples using RD comes despite them being sequenced with shorter reads than for the PA samples, which would typically (all else equal) result in a lower mapping rate.[207] This suggests that if RD samples were also sequenced using the same read length as the PA samples (100bp), the mapping rate for RD samples may have been even higher. It was also found that RD resulted in higher intronic mapping rates and lower exonic mapping rates for nearly all samples, which may be a result of sequencing unprocessed pre-mRNAs.[207] A general small increase in intergenic reads was also observed with the RD protocol, which was in part explained by the presence of unannotated expressed intergenic regions described in section 3.3.3. This hypothesis was further supported by two observations: i) the two protocols resulted in a high correlation between expression levels of coding genes but lower correlation for lncRNA expression, and ii) RD resulted in detected expression of significantly more lncRNAs (particularly lincRNAs) than PA.

Once reads were quantified, it was noted that four lncRNAs were expressed at exceptionally high levels after RD (accounting for a combined 23.8% of quantified reads): *RN7SL2*, *RN7SL1*, *RN7SK*, and *RPPH1*. While each of these genes has

been identified as having exceptionally high expression in various studies using RD protocols, no study to my knowledge has found all four to be expressed at high rates. In addition to these lncRNAs, other highly expressed lncRNAs have been identified in other studies using RD in different tissue and cell types, such as *SNORD3A* in a study of HEK 293 T cells,[207] and *RMRP* in a study of 17 tumour samples from breast and prostate.[144] In one study, it was suggested that aberrantly expressed lncRNAs could be depleted by adding specific probes.[207] However, given the small number of RD assessments and the fact that different genes have high expression in different studies, it is currently difficult to determine which genes are highly expressed in all tissues and cell types and which are tissue- or cell-type specific.

After exclusion of reads that mapped to the four previously mentioned highly expressed lncRNAs, only 37.5% of remaining raw sequenced reads from libraries using RD were successfully quantified. In contrast, 68.5% of reads from PA libraries were quantified. This suggests that it is necessary to sequence approximately 83% more reads if using RD compared to PA to quantify the same number of reads (after accounting for genes with extreme expression). This estimate does not account for the fact that different studies, such as that by Akerman *et al.* described in this chapter, may wish to remove even more highly expressed genes, reducing the effective sequencing depth further.[103] In terms of mRNA depth, 34.8% of reads after RD were ultimately quantified as mRNAs, while 65.8% of reads after PA were quantified as mRNAs. This means that if a study requires a particular mRNA sequencing depth, approximately 89% more reads are required using RD compared to PA. Thus, if the goal of a study is to capture mRNA variation, PA is more efficient, but to capture an image of the full transcriptome with high coverage of lncRNAs and pre-mRNAs, RD is more appropriate.

One RD sample used in this study contained a large portion of RNA belonging

to *pseudomonas* bacteria. The same sample had also been sequenced using the PA protocol, which showed no sign of contamination. Since bacterial RNAs tend not to be polyadenylated, it is likely that the sample itself was contaminated, but the contamination was only detected in the RD sample.[214] There is a possibility that contamination occurred during sequencing preparation, and this could be determined by re-sequencing the sample using both protocols. The potential for skewed RNA-seq results due to undetected bacterial contamination could have implications in downstream analyses. However, in practice, correction for hidden sources of variation using principal components, probabilistic estimation of expression residuals,[220] or SVs,[203] should mitigate this issue.

In the final section of this chapter, RNA-seq data was used to identify 218 NIERs in human pancreatic islets. These NIERs represent regions of the genome with consistent evidence of transcription across samples, and due to strict filtering thresholds, are unlikely to be transcriptional byproducts of other genes. The largest similar study that defined novel lncRNA transcripts in human islets used data from 41 individuals.[103] Although the detection and filtering pipelines used by Akerman *et al.* were different from those used in my analysis, a subset of their defined lncRNAs met the same filtering criteria that I used.[103] Several of these published islet lncRNAs were expressed in the RNA-seq data from my analysis, and a further subset overlapped genomic boundaries of NIERs defined in section 3.3.3. Thirty-four of the genes identified by Akerman *et al.* were expressed at detectable levels in the samples used in my analysis, which provides robust confirmation of their expression in human islets. At the time of my analysis, the RNA-seq data used by Akerman *et al.* was not publicly available, so an independent replication analysis of the NIERs I identified has not been performed.

Although it was shown in section 3.3.3 that there are true lncRNAs expressed in human islets that are not part of the GENCODE reference transcriptome, NIER definitions from my analysis and lncRNAs from Akerman *et al.* will not be used

for quantification of islets in later chapters. The exclusion of these in following chapters is for the following reasons:

- The majority of lncRNAs detected in section 3.3.3 and by Akerman *et al.* do not overlap, indicating high detection variability between methods and samples.

- Although 14 NIERs were identified by both my analysis and that of Akerman *et al.*, and 34 lncRNAs defined by Akerman *et al.* were expressed in the islet samples from my analysis, including these definitions in downstream analysis would account for at most a 3% increase in expressed lncRNAs beyond those defined in GENCODE.

- Analyses in Chapters 4-6 use RNA-seq from GTEx and InsPIRE, which have both been quantified using GENCODE 19 prior to my analyses.

The detection of novel lncRNAs nevertheless emphasises the fact that pre-compiled, non-tissue-specific transcriptome references such as GENCODE are not exhaustive. By comparing my results to results of Akerman *et al.*, it was shown that with small sample sizes (both $n < 55$), detection of unannotated lncRNAs is highly variable depending on the dataset and method of detection. Thus, it is important to note that there are lncRNAs in islets beyond GENCODE definitions, but robust, replicable detection is currently limited.

The availability of large-scale RNA-seq datasets from many tissue and cell types is rapidly expanding, which means that researchers are often able to analyse RNA-seq data from tissues that are relevant to the phenotype or disease being studied. However, this chapter has outlined various factors that should be considered when analysing and interpreting RNA-seq data, such as the gene biotypes being studied, limitations of small sample sizes, and the RNA preparation method used. The utility of RD RNA-seq was demonstrated through the identification of previously undescribed expressed regions in human islets, but its

inefficiencies were also highlighted. The detection of NIERs in islets also highlighted the fact that there are lncRNAs in islets beyond those described by GENCODE, but robust detection is currently limited due to small sample sizes and discrepancies between detection methods. With these observations in mind, the following chapters of this thesis further exploit RNA-seq data from the GTEx project and from a large sample set of human pancreatic islets to identify gene expression patterns and regulatory effects of lncRNAs that are associated with, and in some cases contribute to, the pathogenesis of T2D.

Chapter 4

The profile of T2D associated gene expression and potential lncRNA regulation

4.1 Introduction

The previous chapter discussed factors that impact the sequencing and quantification of long noncoding RNAs (lncRNAs) in humans, and established a set of factors that should be considered in analyses involving RNA-sequencing (RNA-seq) of lncRNAs. With these observations in mind, the remainder of this thesis will assess in several ways the functional potential of lncRNAs in type 2 diabetes (T2D) development and regulation of glycaemic traits. The focus of this chapter is on analyses that do not require genetic information, while Chapters 5 and 6 will expand on these analyses with the incorporation of genetic data.

4.1.1 RNA and phenotype associations

The availability of individual-level RNA-seq data and phenotype data allows for direct evaluation of associations between expression of genes and phenotype.

The simplest type of association, which is commonly employed in RNA-seq studies, is generally referred to as differential expression (DE) analysis. DE analysis consists of association testing between levels of gene expression and a phenotypic variable of interest:[221]

- **Discrete trait DE** tests for differences in expression between individuals of each disease state. A simple example is performance of a t-test between the mean expression of individuals with and without T2D.
- **Continuous trait DE** tests for a significant regression coefficient for the association between the phenotype and gene expression. An example would be a test of significance in a linear regression model with gene expression as an explanatory variable for body mass index.

In addition to these simple examples, many further methods have been developed specifically for identifying DE effects. Some of the most widely used R packages specifically designed for DE analysis are *DESeq*,[222] *edgeR*,[201] *NOISeq*,[223] and *limma*.[202] The selection of method for DE analysis is important when RNA-seq studies consist of a small number of replicates,[221,224] but as the number of biological replicates increases in DE studies, results from methods tend to converge. It was shown by Schurch *et al.* that for RNA-seq studies with at least 12 biological replicates, 13 different DE methods the authors tested captured a similar number of true differentially expressed genes across varying levels of expression fold changes between states.[225]

DE analyses have been used to identify disease-associated lncRNAs in many settings. Some DE studies are focussed on biomarker detection, such as an analysis between gastric cancer specimens and healthy paired tissue that identified five lncRNAs (*CTD-2541M15*, *UCA1*, *AP000459*, *LINC01071*, and *MEG3*) as robust biomarkers in the early detection of gastric cancer,[226] and another study that found lncRNAs *C1orf132* and *TMPO-AS1* to have robust DE profiles between

lung adenocarcinoma and normal tissue.[227] Outside the cancer realm, recent studies have used DE analysis to classify two lncRNAs that are robust biomarkers for tuberculosis diagnosis,[228] 11 candidate lncRNA biomarkers for early diagnosis of acute myocardial infarction,[229] and several hundred lncRNAs with expression profiles associated with the three replication cycle processes of HIV,[230] to name a few.

Several studies have also used DE analysis to identify gene expression that is associated with T2D status and glycaemic traits, and rather than biomarker detection, focus on identifying genes with causal effects on T2D development. The most comprehensive DE study in human islets was published recently by Ottosson-Laakso *et al.*, and performed three separate DE analyses:[101]

- DE before and after glucose stimulation in islets of 31 normoglycaemic donors
- DE before and after glucose stimulation in islets of 14 hyperglycaemic donors
- DE between unstimulated islets of 81 normoglycaemic donors, 19 hyperglycaemic donors, and 16 donors with T2D

Based on overlaps between results of these three sets of tests, Ottosson-Laakso *et al.* attempted to dissect causality of gene expression effects by first suggesting that genes with significant DE after both acute and long-term glucose exposure (hyperglycaemic and T2D islets) were likely results of glucotoxicity. Similarly, they hypothesised that genes with DE between islets of varying glycaemic state but not after acute glucose stimulation were the strongest candidate causal genes in T2D-related processes. Of the 392 genes in the latter category, there were 16 lncRNAs: 7 antisense, 8 long intergenic noncoding RNAs, and 1 processed transcript. Although the authors did not specifically discuss lncRNAs, these results demonstrate the utility of DE analysis for their detection.

RNA-seq of islets from 89 donors published by Fadista *et al.*[98] was recently re-quantified by Akerman *et al.* using their expanded *de novo* lncRNA definitions

(described in Chapter 3). Akerman *et al.* identified 15 lncRNAs in islets with significant DE between samples from normoglycaemic donors and donors with T2D,[103] but made no distinction between causal and secondary effects. The studies by Ottosson-Laakso *et al.* and Akerman *et al.* are the largest published analyses of lncRNA differential expression in human islets; however, results are not directly comparable, since the analysis by Akerman *et al.* re-quantified genes based on their *de novo* lncRNA definitions described in Chapter 3. A study by Wang *et al.* used peripheral blood samples to identify 55 lncRNAs with DE between healthy individuals and individuals with newly diagnosed T2D,[151] but other studies focussing specifically on lncRNA expression associations with T2D in tissues other than islet are not common.

The studies described above demonstrate the utility of DE analysis for identifying lncRNAs that may be of functional relevance to T2D. However, the studies of islets suffer from relatively small sample sizes ($n < 90$), which may limit exhaustive detection of differentially expressed lncRNAs. There are also few studies assessing lncRNA expression associations with T2D in tissues other than islet, despite availability of RNA-seq from other T2D-relevant tissues such as skeletal muscle, adipose, liver, and pancreas. Thus, lncRNA associations with T2D beyond those currently published are likely to exist. Although the causal direction of association between gene expression dysregulation and T2D development cannot be confidently determined from DE analysis alone, the focus of studying DE in this chapter is to dissect the tissue specificity of dysregulation in T2D under the hypothesis that genes with DE across many tissues may result from overall elevated plasma glucose and other broad phenotypes associated with T2D. Causality and causal inference using genetic data will be described in detail throughout Chapter 6.

4.1.2 Gene co-expression

Construction and analysis of gene co-expression networks is common in many studies assessing regulatory elements of gene expression. Gene co-expression networks broadly describe the relatedness of gene expression patterns, and can be applied simply to identify clusters of genes with similar expression patterns, or to determine more complex relationships such as pairwise relatedness and geodesic distance between genes. Any gene co-expression analysis generally consists of three steps:[231] i) calculation of pairwise relatedness between genes, ii) definition of edges based on a minimum relatedness value, and iii) clustering of genes based on edge definitions.

Each of the three steps in co-expression network construction can be performed in a number of ways. For defining relatedness (co-expression) between pairs of genes, Pearson correlation is used most often.[232–234] However, Spearman correlation can also be used, and in the widely used weighted correlation network analysis (WGCNA) method, exponentiated correlation is used.[235] In the typical second step of co-expression network construction, selection of a suitable cut-off value for defining edges is performed before clustering. Vandepoele *et al.*, for example, used a permutation approach to determine correlation cut-off thresholds corresponding to percentiles.[236] For some methods such as WGCNA, hierarchical clustering is performed using all genes ('skipping' step 2) and then branches of the resulting dendrogram are cut to identify final gene clusters.[235,237] The clustering step in gene co-expression analysis can be performed using many possible methods, though k-means clustering and hierarchical clustering are most common.[231,238]

Hierarchical clustering typically uses an agglomerative approach, and begins by assigning each individual gene to its own cluster. Then, genes (and eventually clusters) with high similarity are iteratively merged until all genes are connected.[238,239] The k-means clustering algorithm is an iterative approach that first selects a random centroid value for each of k clusters, and places each gene

into the cluster with the nearest centroid[238] (based on a similarity measure). The centroid of each cluster is then re-defined as the mean of genes in that cluster, and the approach is iteratively repeated until the clusters do not change between iterations.[238,240]

Given its ease of implementation, WGCNA is the most commonly used gene co-expression construction pipeline.[231] The use of hierarchical clustering in WGCNA also assumes no prior knowledge of the number of distinct clusters, which is advantageous for exploratory co-expression analysis where there is no clear prior knowledge of gene clustering. Once gene co-expression networks and clusters are defined, it is possible to infer the functional roles of gene clusters, which can then be used to predict function of specific genes — particularly those with limited prior functional annotations such as lncRNAs.[231]

4.1.3 Functional relevance of co-expression clusters

There are two common approaches used to identify potential function of co-expression clusters: functional enrichment analysis[231,241] and eigengene analysis.[242]

Functional enrichment analysis uses existing databases that relate genes to functional activity, and tests for statistically significant over-representation of particular functional classes or pathways within a gene co-expression cluster. Two of the most common databases used for such analyses are the Gene Ontology Consortium's gene ontology (GO) terms,[243,244] and the Kyoto Encyclopedia of Genes and Genomes database project's pathways.[245] Once co-expression clusters have been defined, a variety of online tools are available to test for enrichment of annotations such as DAVID[246,247] and PANTHER;[248] enrichments can also be calculated in R with simple packages such as GOstats,[249] which implements a hypergeometric test to determine if GO terms are over-represented amongst genes in a cluster.

A gene cluster's eigengene simply refers to the first principal component of the co-expression cluster, and is essentially an individual-level aggregate measure of all genes in the cluster.[235] Thus, a simple eigengene analysis can be performed by associating the cluster eigengene to a phenotype of interest. A significant association between a module eigengene and a phenotype of interest implies that the overall expression of the co-expression cluster is associated with that phenotype.[250,251]

Regardless of the method used to determine module function, results from functional tests are then typically used in 'guilt by association' (GBA) approaches. In GBA analysis of co-expression clusters, it is simply assumed that if a cluster is associated with a particular function, all genes in that cluster are associated in some way with that same function.[231,252] Although a study by Gillis and Pavlidis showed that GBA is often an oversimplification of biology and does not always hold true,[253] GBA is still widely used for assigning function to co-expressed genes.[251,254–256]

A recent analysis by Axelsson *et al.* used WGCNA to define co-expression networks using microarray expression data from islets of 64 humans. They identified 56 distinct co-expression clusters, and the eigengene of one cluster was significantly correlated with T2D status and glycated haemoglobin[251] (HbA1c). Based on follow-up of genes in the co-expression cluster, they determined that the module is also enriched for genes in regions with islet-selective open chromatin, and were able to partially replicate the module in an independent sample of islets from 59 donors. They further identified *SOX5* (SRY-Box 5) as a regulator of the module's expression, and through experimental validation showed that *SOX5* over-expression improves insulin response to glucose stimulation, and that the improvement was 3-fold stronger in islets from donors with T2D.[251]

In the lncRNA study by Akerman *et al.* described previously, gene co-expression networks were constructed with WGCNA using RNA-seq from 64

human islet samples.[103] The authors identified 25 gene modules, seven of which were enriched for beta-cell-specific lncRNAs. Functional enrichment analysis was then used to show that five of these modules were enriched for genes associated with islet enhancer clusters, two were enriched for genes involved in messenger RNA translation and metabolic pathways, three were enriched for islet-specific transcription factor genes, and the largest of the seven modules was enriched for GO terms associated with prototypical islet cell functions.[103]

These studies together demonstrate the utility of co-expression analysis and GBA approaches to identify both gene clusters and individual genes that are relevant to T2D and glycaemic traits. Both studies used co-expression as a method of identifying candidate genes and gene sets for follow-up analysis. The study by Akerman *et al.* also demonstrates that lncRNAs are active and co-expressed with genes relevant to T2D and beta-cell function. However, much like other RNA-seq studies in islets, those described above have relatively small sample sizes, limiting their power to detect lncRNA associations.

Given the demonstrated utility of DE and co-expression analysis for identifying putative functional lncRNAs and gene clusters related to T2D, this chapter aims to extend and improve these analyses under the following aims:

- Identify gene expression associations with T2D using RNA-seq from a wide spectrum of human tissues including the largest such study of RNA-seq from human islets, and test for tissue distribution of effects.
- Construct gene co-expression networks using RNA-seq from a wide range of human tissues and identify clusters enriched for T2D-associated lncRNAs.
- Identify biologically relevant lncRNAs by assessing the functional relevance of enriched co-expression clusters to the regulation of insulin secretion.

4.2 Samples and methods

4.2.1 GTEx samples

Genotype-Tissue Expression Project (GTEx) samples and processing are described in section 2.1. For all analyses in this chapter, any sample with unavailable T2D status was excluded, and tissues with remaining $n < 50$ were excluded from analyses (bladder, ectocervix, endocervix, fallopian tube, and kidney). The 48 remaining tissues are listed with sample sizes and number of T2D cases in Table A.4.

After filtering samples, raw read counts were filtered and normalised as described in section 3.2.1.

Surrogate variables (SVs) were then calculated for each tissue as described in section 3.2.1. For all tissues, T2D was explicitly defined in the SV model, and for tissues that were not unique to a single sex, sex was also explicitly defined in the SV model.

4.2.2 Islet samples

Of the 443 RNA-sequenced islet samples from the Integrated network for systematic analysis of Pancreatic Islet RNA Expression (InsPIRE) collaboration (described in section 2.2), data from a subset of 207 samples with known T2D status was analysed in this chapter: 3 samples from Geneva, and all 204 Lund samples. Oxford samples were not included here since all Oxford samples included in the InsPIRE collaboration are from non-diabetic donors.

RNA-seq data from these 207 samples was normalised in the same way as GTEx RNA-seq described in section 4.2.1, and SVs were calculated using a model that explicitly defined sex and T2D status.

4.2.3 Differential expression

DE analysis was performed for each tissue using the `eBayes` function from the *limma* package version 3.30.9 in R.[202]

First, a design matrix was created with the `design` function from the *BioGenerics* package version 0.20.0, which is included as part of the *limma* package. Each tissue's design matrix consisted of an intercept term, all SVs, sex, and T2D status. Linear models were then produced for each gene using the `lmFit` function in *limma*. Finally, the `eBayes` function was used to assess DE for each gene based on T2D status, and final results were extracted with the `topTable` function from *limma*.

The `eBayes` method uses variation information from all genes to inform and improve estimation of per-gene variation, which reduces the false positive rate and provides greater power for detecting DE in low-variance genes.[202] Given that all samples were sequenced using poly-A-tailed RNA selection and that many lncRNAs are detected at low levels under this protocol (Chapter 3), the empirical bayes methodology of `eBayes` was used to maximise discovery of differentially expressed lncRNAs.[202]

4.2.4 Co-expression network construction

Before construction of co-expression networks, all SVs and sex were regressed out of normalised expression values for each gene independently using a simple linear model, and residual values were used for network construction. T2D status was not regressed out of normalised expression values. Effects of T2D on gene expression were maintained to allow for clustering of genes with similar patterns of T2D-based DE.

WGCNA version 1.51 was used in R for co-expression network construction in each tissue.[235] Networks were constructed and modules were selected based on guidelines in the WGCNA R tutorial available from `labs.genetics.ucla.edu/`

horvath/CoexpressionNetwork/Rpackages/WGCNA/Tutorials/.

All analysed tissues had $n > 40$, so a soft-thresholding power of 6 was used for all tissues as recommended in WGCNA documentation (up-to-date as of December 2017). A minimum cluster size of 20 genes was enforced, and a cut height of 0.2 was used when merging clusters (corresponding to a correlation of 0.8).

4.2.5 Co-expression module analysis

Enrichment of T2D-relevant lncRNAs in modules was calculated using the Fisher combined scores of lncRNA T2D DE p-values in each module. When all p-values are independent, Fisher's combined score assumes a χ^2_{2k} distribution for a combined score of k p-values.[257] However, since expression of multiple genes in the same tissue is unlikely to be completely independent, this null distribution may no longer be valid.[258] To address this concern, the following process was implemented to assess enrichment:

i) The Fisher combined score F_m was used to combine T2D DE p-values of all lncRNAs in each co-expression module m:

$$F_m = -2 \sum_{i \in l_m} \log p_i$$

$F_m =$ Fisher score of DE p-values for lncRNAs in module m

$l_m =$ list of all lncRNAs in module m

$p_i =$ DE p-value for lncRNA i in module m

ii) A null distribution was calculated for each module by calculating 10,000 permuted scores. For each permutation, a random set of expressed lncRNAs the same size as l_m was selected from the same tissue, and the Fisher score was calculated for their T2D DE p-values.

iii) A p-value was calculated for each F_m based on the null distribution calculated

in ii).

A significant module enrichment p-value after correction for multiple tests signifies that there is a higher-than-random enrichment of lncRNA T2D DE within the module.

4.2.6 Gene ontology enrichment analysis

GO[243] term enrichment analysis was performed on the genes in each module using a hypergeometric test implemented with the `hyperGTest` function from the *GOstats* package version 2.40.0 in R.[249] Gene IDs were matched to GO term annotations using the *org.Hs.eg.db* package version 3.4.0 in R.[259] Each enrichment test was limited to an environment of genes that were expressed in that tissue to alleviate potential sampling bias.[260]

Each GO annotation is associated with a list of gene IDs that are part of that annotation. For each module in each tissue, each GO annotation is tested independently. Under the null hypergeometric distribution, the probability $P(X = k)$ that k genes within a module of a tissue have a certain GO annotation is calculated as

$$P(X = k) = \frac{\binom{K}{k}\binom{N-K}{n-k}}{\binom{N}{n}}$$

where N is the total number of genes expressed in that tissue, n is the total number of genes in the module, K is the total number of genes expressed in the tissue with that GO annotation, and k is the number of genes in the module with that GO annotation. Significant multiple-test-corrected p-values indicate that more genes in a module have that particular GO annotation than would be expected at random, suggesting that genes in the module may be associated with the process depicted by the GO annotation.

4.3 Results

Since dysfunction of pancreatic islets is one of the primary changes leading to development of T2D, gene expression differences based on T2D status that are specific to islets are more likely than tissue-shared expression differences to be associated directly with regulation of insulin secretion. By using RNA-seq data from a spectrum of GTEx tissues, as well as RNA-seq from islets, it is possible to identify genes with T2D DE patterns and associated functional processes that are specific to islets. I reasoned that islet-specific T2D-associated expression differences and processes are more likely of direct relevance to T2D pathogenesis compared to expression changes shared across many tissues, which may suggest more widespread secondary expression effects triggered by elevated blood glucose or other broad phenotypes resulting from T2D. Although the analyses presented in this chapter are not sufficient evidence that expression effects are causally associated with T2D, islet-associated patterns identified here have been used in conjunction with genetic data in Chapter 6 to partially dissect causality and identify likely causal functional processes.

A recent publication by Ottosson-Laakso *et al.* has demonstrated the utility of DE analysis in islets to highlight genes with potential functional relevance to T2D, and my analysis not only uses a significantly larger sample size to assess these effects, but adds the analysis layer of tissue specificity, which was not assessed by Ottosson-Laakso *et al.*[101]

I therefore first performed tests of DE within tissues from GTEx, then performed co-expression analysis and enrichment tests to identify lncRNAs that are relevant to T2D development in non-islet tissues. By further performing similar analyses in data from islets, I assessed the tissue specificity of effects to identify those that are likely islet-specific and therefore likely to be associated with the regulation of insulin secretion.

4.3.1 Gene expression association with T2D in non-islet tissues

Differential expression

After filtering GTEx samples based on availability of T2D status and sample size ($n > 50$), 48 GTEx tissues remained with sample sizes ranging from 88 (substantia nigra) to 565 (skeletal muscle), with a median 187 samples (Figure 4.1 and Table A.4). There were a median 19.7% T2D cases across the tissues analysed (minimum 14.4% in minor salivary gland, maximum 26.5% in caudate; Figure 4.1 and Table A.4). The number of genes expressed per tissue ranged from 15,750 (Epstein-Barr virus transformed lymphocytes) to 24,945 (testis) with a median 18,537 expressed genes.

DE analysis was then performed independently per tissue to identify the genes with significant differences in expression level between donors with and without T2D. After correcting for multiple testing per tissue with a false discover rate (FDR) of 5%, 1,425 gene-tissue pairs were significant (Figure 4.2), representing 1,354 unique genes. Of the 1,354 unique genes with significant DE tests, 1,238 associations were coding genes, 89 were lncRNAs, and 27 were other biotypes (1 T-cell receptor V gene, 1 small nuclear RNA, and 25 pseudogenes). The proportion of lncRNAs with DE amongst all differentially expressed genes (6.3%) was significantly smaller than the overall proportion of lncRNAs tested (12.8%, $p = 2.74 \times 10^{-13}$).

Only 22/48 tissue types analysed had any differentially expressed genes, and 38.6% of significant DE effects were in skeletal muscle. Change in skeletal muscle insulin resistance (which is itself a complex phenotype) is one of the primary defects in the development of T2D,[261] and systematic dysregulation of genes in muscle from T2D donors has been shown before.[262–265] It is therefore not surprising to observe a high number of genes with DE in skeletal muscle based on

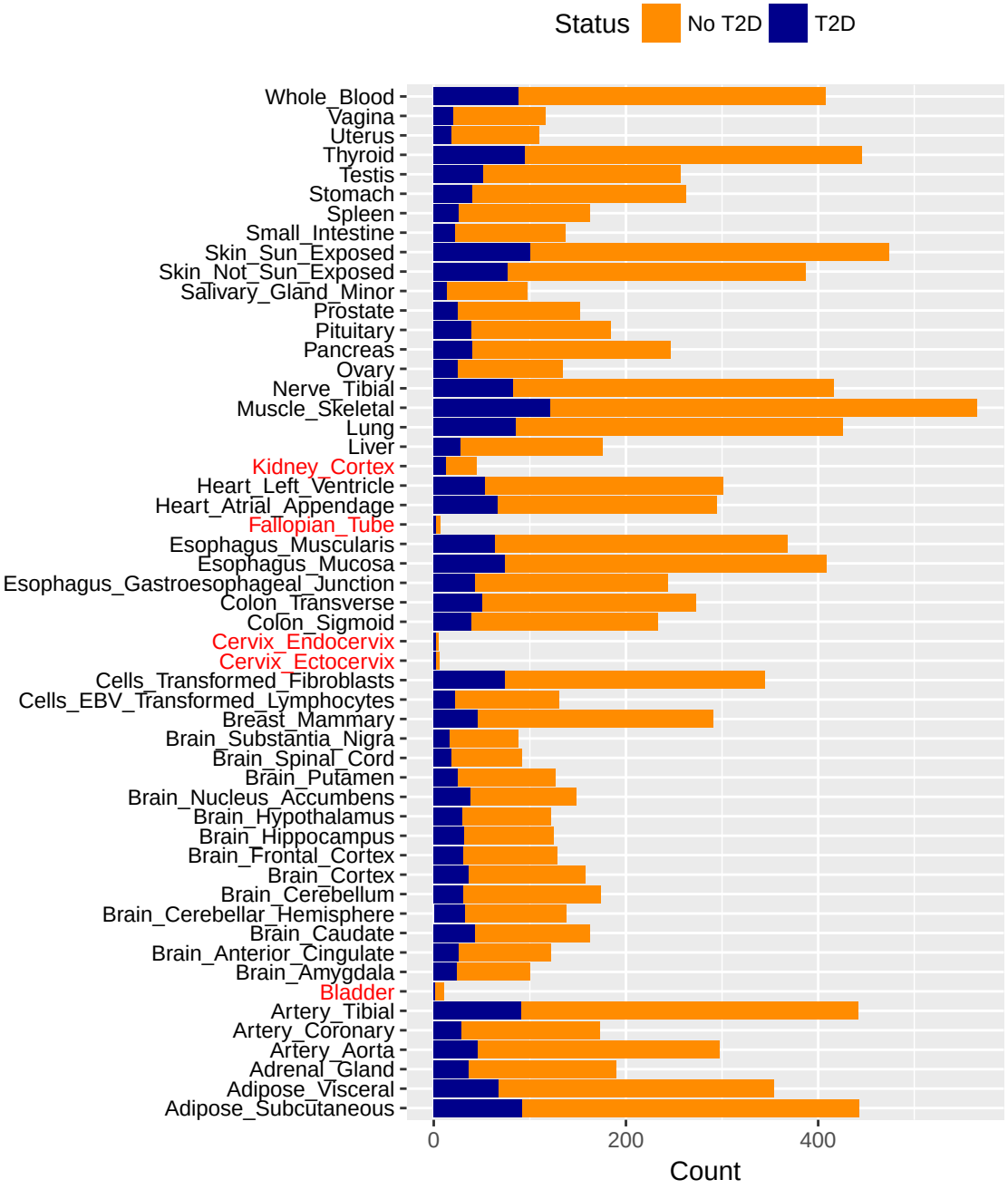


Figure 4.1: Summary of GTEx tissue samples. The number of samples per tissue type with known T2D status (x-axis), and the distribution of T2D cases and controls per tissue (bar colour). The names of tissues that were excluded in this chapter ($n < 50$) are shown in red.

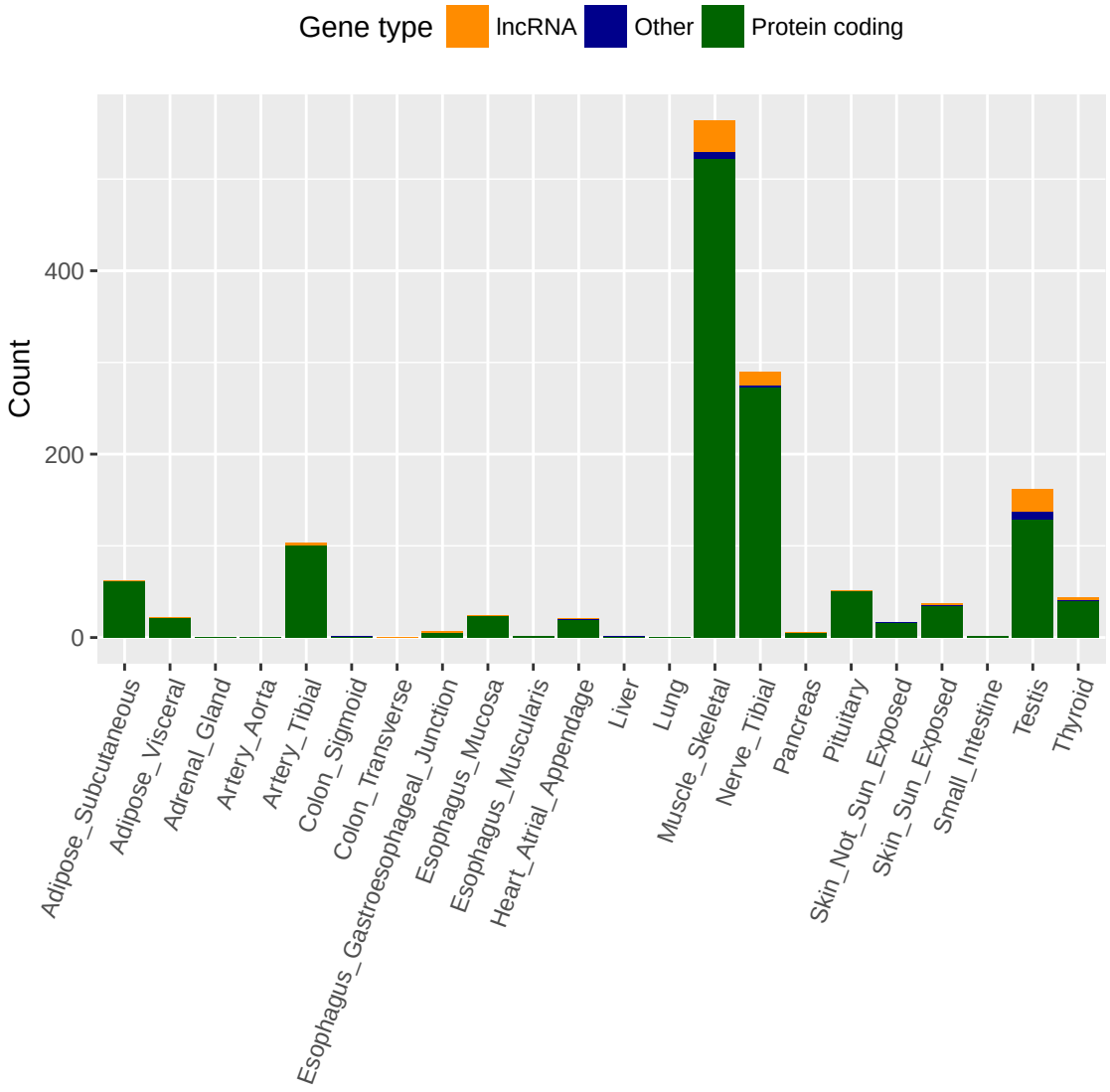


Figure 4.2: Distribution of T2D DE effects by tissue. The number of genes significantly differentially expressed between T2D cases and controls per tissue (y-axis). Bar colour indicates the gene biotype distributions (lncRNA, coding, or any other biotype).

T2D status. Although determining whether these expression changes are primary (changes in gene expression that causally influence insulin resistance or T2D development) or secondary (changes in expression that are a result of physiological change in donors with T2D) with expression data alone is difficult, identification of many of these genes only in skeletal muscle is suggestive that they may represent primary associations.

There were 1,238 unique coding genes with DE in at least one tissue, and 1,173 of these (94.7%) were differentially expressed in a single tissue. Of the 89 unique lncRNAs with DE in at least one tissue, 88 were significant in a single tissue (Figure 4.3). The proportions of tissue-specific lncRNAs and coding genes were not significantly different ($p = 0.13$). A recent study found a similar lack of tissue sharing of genes with DE based on sex across GTEx tissues;[199] however, existing published T2D DE studies have generally focussed on only a single tissue.[98,150,266] The number of genes detected with significant DE in each tissue was also correlated with sample size ($r = 0.54$; Figure 4.4), in line with previous evidence that increased sample sizes improve power to detect differential gene expression.[267,268]

DE analysis across GTEx tissues has shown that the majority (95.1%) of unique genes (all biotypes) with DE based on T2D status were dysregulated in only one tissue, and most of the DE associations occurred in skeletal muscle; however, the apparent tissue specificity of many of these genes may be due to lack of power to detect effects, evidenced by a modest correlation between tissue sample size and number of differentially expressed genes.

Gene co-expression and lncRNA enrichment

Since many lncRNAs do not have any annotated function, co-expression networks were next constructed to identify co-expressed clusters of genes containing T2D-associated lncRNAs. These clusters were then used to identify potential function of lncRNAs using a GBA approach.

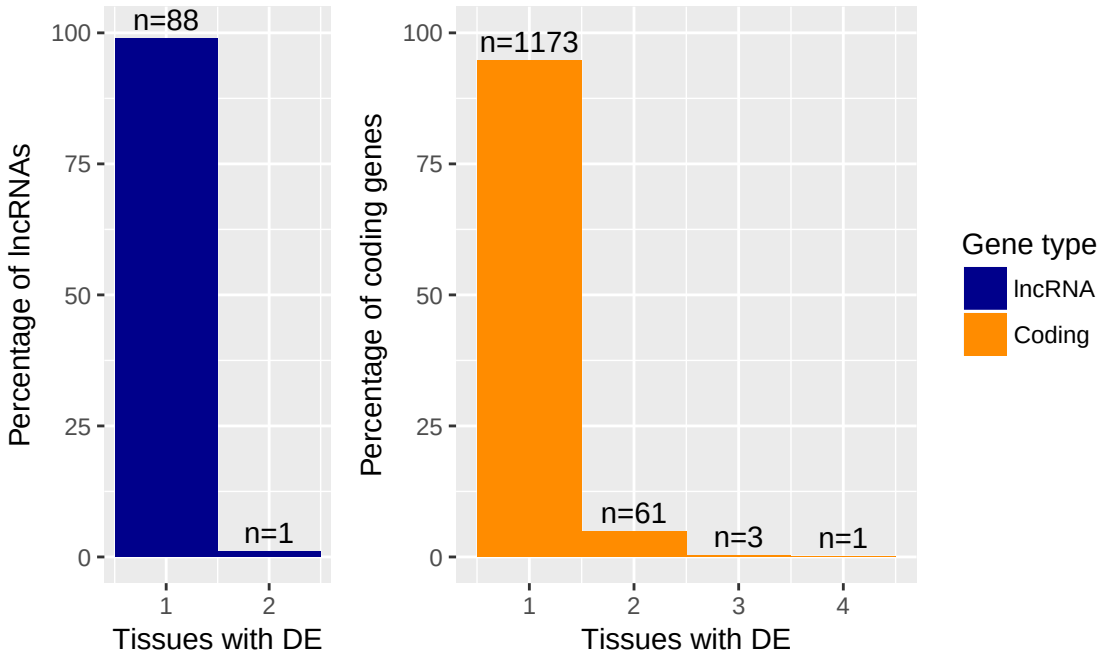


Figure 4.3: Tissue sharing of T2D differentially expressed genes. The number of tissues in which each unique lncRNA (left plot) and coding gene (right plot) has significant DE based on T2D status.

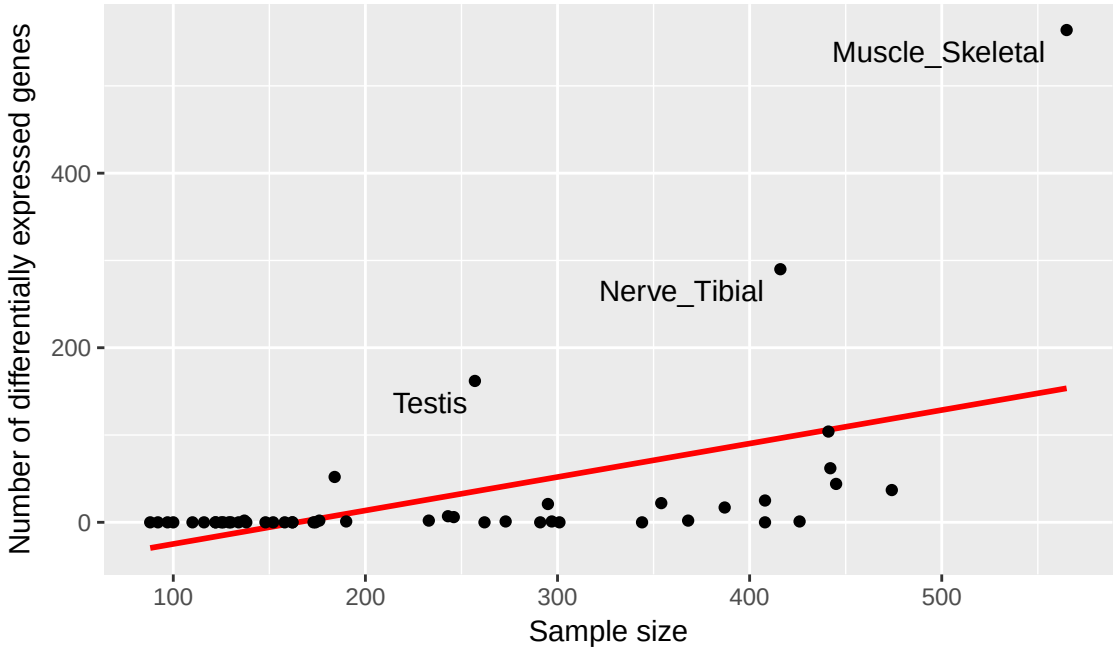


Figure 4.4: Correlation between DE detection and sample size. The number of genes with significant DE based on T2D status (y-axis) has correlation $r = 0.52$ (red line) with tissue sample size (x-axis).

Mutually exclusive gene co-expression modules were constructed using the WGCNA method.[235] Using normalised and residualised gene expression values, WGCNA first calculates pairwise correlations between genes and exponentiates these correlations to emphasise the strongest gene correlations (soft thresholding). Hierarchical clustering is then used to cluster genes based on the exponentiated similarity matrix, and the clustering dendrogram is cut to define final modules of co-expressed genes.

After performing WGCNA analysis on each of the 48 GTEx tissues, there were a median 101.5 gene co-expression modules per tissue (minimum 46 in skeletal muscle, maximum 251 in substantia nigra). There were a median 65 genes per module across tissues, and the largest module consisted of 2,762 genes in mammary tissue (Figure 4.5). There were a median 35.8% of genes per tissue that were not assigned to any module.

Next, to identify gene clusters involving lncRNAs with T2D DE, I tested for enrichment of lncRNA DE p-values in each module using the T2D DE results from section 4.3.1. The Fisher combined score of DE p-values of lncRNAs was first calculated for each module. Since expression of genes within a single module is inherently correlated, the null χ^2 distribution normally followed by the Fisher score is no longer valid.[258] To address this, for each module, I performed a permutation analysis to define a null distribution. For each of 10,000 permutations, I selected a random subset of lncRNA DE p-values from the full set of lncRNAs expressed in the relevant tissue and calculated the Fisher score. The p-value for each module was then calculated as the fraction of permutations that resulted in a higher score than the lncRNAs in the module. For example, if 18 permutations resulted in a higher score than the score of the observed lncRNAs, the module would be assigned a p-value of 0.0018.

Thirty-four modules in 23 tissues were significantly enriched for lncRNA T2D DE (FDR<0.05; Table A.5). Other than anterior cingulate, which had 3

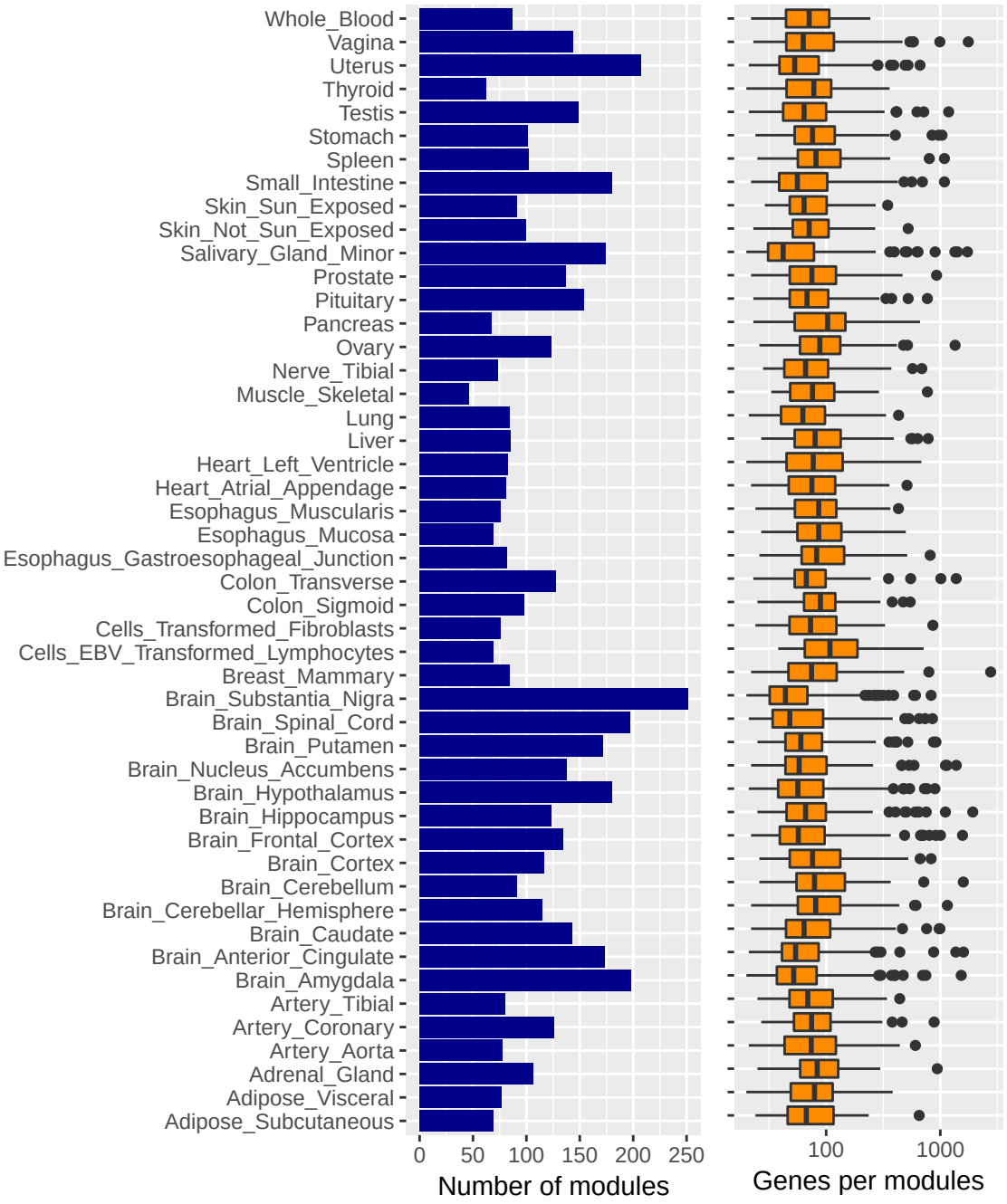


Figure 4.5: Co-expression module distribution by tissue. The number of co-expression gene modules identified in each tissue (x-axis, left plot) ranged from 46 to 251. Boxplots (right plot) indicate the distribution of the number of genes per module on a $\log_{10}$ scale (x-axis).

significantly enriched modules, no tissue had more than 2 enriched modules.

To identify potential biological roles of each of the modules that were enriched for lncRNA T2D DE, I performed enrichment tests for GO terms in each of the 34 enriched modules. A significant enrichment for a GO term in a module indicates that there are more genes with that GO annotation in the module than would be expected by random for a module that size. Twenty-nine unique modules in 20 tissues contained a total of 3,586 over-represented GO terms (2,284 unique). Modules directly relevant to glycaemic regulation were identified by performing text searches of 'insulin', 'glucose', and 'glucagon' against all significant GO terms in each of the enriched modules. Three modules had matching over-represented GO terms:

- Adrenal gland: 'negative regulation of insulin secretion involved in cellular response to glucose stimulus'
- Pancreas: 'insulin secretion' and 'regulation of insulin secretion'
- Mammary tissue: 'glucose transport' and 'regulation of glucose transport'

Despite all three of these modules being generally enriched for lncRNA T2D DE p-values, only one of these — the module in pancreas — contained genes that were independently significantly differentially expressed. In the pancreas module, there were also 41 over-represented GO terms in addition to the two associated with regulation of insulin secretion, many of which are also relevant to glucose homeostasis such as 'hormone secretion', 'regulation of hormone levels', and 'secretion by cell'. All over-represented GO terms in the pancreas module are listed in Table A.6.

The pancreas module with enrichment for insulin-related GO terms contained six genes that were significantly differentially expressed between pancreas tissue of T2D and non-diabetic donors: *G6PC2*, *HHATL*, *HADH*, *IAPP*, *PPP1R1A*, and lncRNA *CTB-12O2.1*. The pairwise expression correlations between all genes in

this module are shown in Figure 4.6. *CTB-12O2.1* had the minimum DE p-value of all lncRNAs in pancreas, and amongst the 89 unique lncRNAs with DE in at least one tissue, *CTB-12O2.1* had the fifth smallest DE p-value. Given that *CTB-12O2.1* is i) differentially expressed based on T2D status in pancreas but no other GTEx tissue and ii) co-expressed with genes enriched for regulation of insulin secretion GO annotations, under a GBA assumption, *CTB-12O2.1* may also be associated with regulation of insulin secretion. In an effort to determine whether or not the observed T2D-associated expression effects of *CTB-12O2.1* in whole pancreas were attributable to pancreatic islets (and thus more likely directly relevant to insulin secretion), I next extended analyses to an independent sample of human islets. Although it is possible that any association effects replicated in islets may stem from exocrine cells in islet samples,[98] replication in islets would still provide improved insight into potential mechanisms of action.

4.3.2 Gene expression associations with T2D in islets

To further examine gene expression associations with T2D and glycaemic regulation and determine those that are likely specific to islets, and to further assess the relevance of *CTB-12O2.1* in T2D pathogenesis, analyses were repeated using RNA-seq data from human pancreatic islets of 207 individuals, 32 of whom had T2D.

DE analysis identified 1,764 genes in islets with DE (FDR$<$0.05) between samples with and without T2D: 1,652 coding genes, 75 lncRNAs, and 37 genes of other biotypes. As was the case with DE in GTEx tissues, the proportion of differentially expressed genes that were lncRNAs was significantly smaller than the proportion of tested genes that were lncRNAs ($p = 1.29 \times 10^{-11}$), which may in part be due to their general low expression and expression variation. The number of T2D DE genes identified here is in line with a previous analysis that identified 1,619 genes with DE based on measures of HbA1c in a subset of 89 of the samples

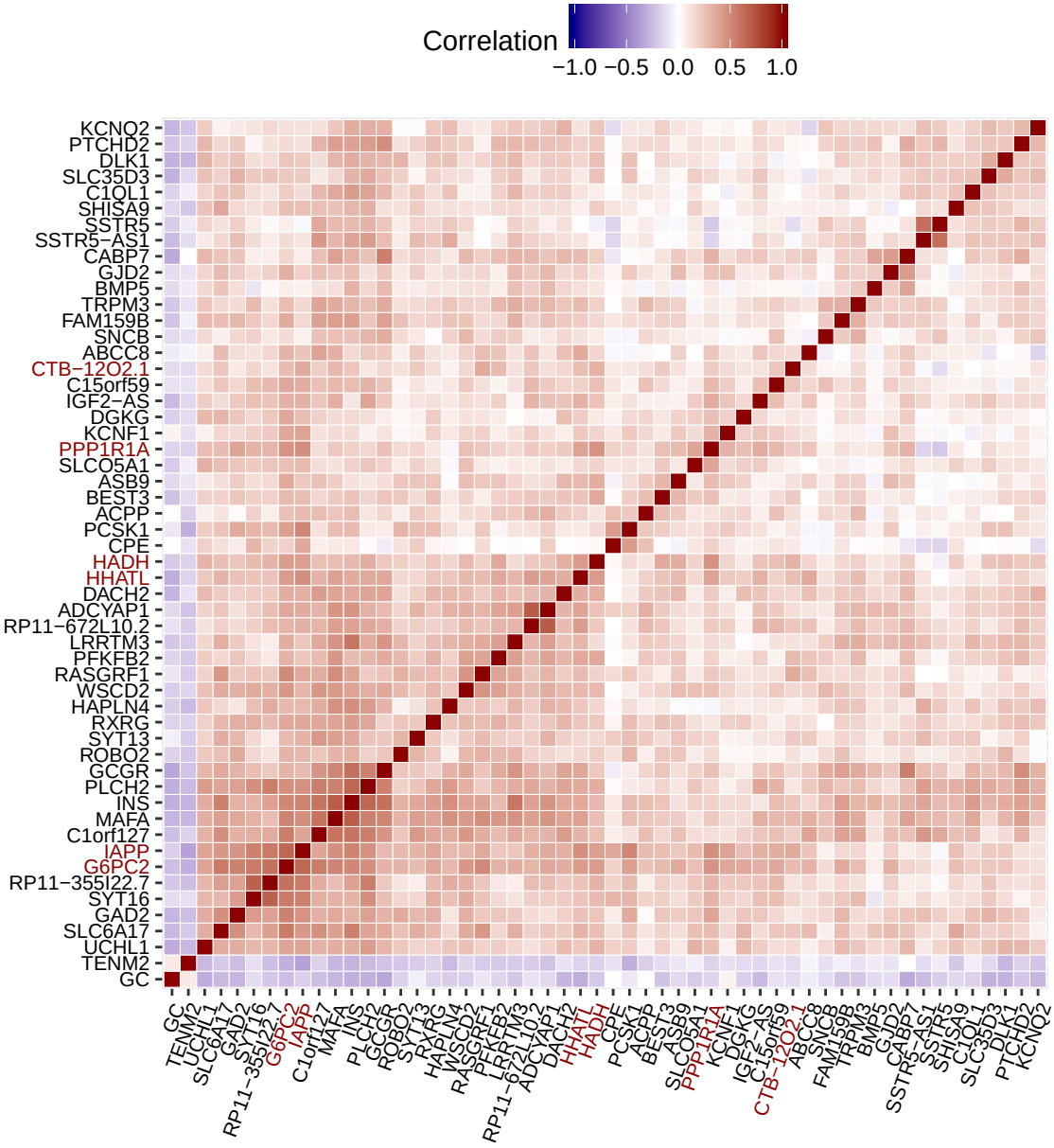


Figure 4.6: Correlation heatmap for genes in insulin regulatory pancreas module. A heatmap indicating the pairwise correlation values between each pair of genes in the pancreas co-expression module with enrichment for insulin secretion GO terms. Red squares indicate positive correlation, and blue squares indicate negative correlation. Genes with significant DE based on T2D status are indicated in red font.

used here.[98] Of the 1,764 genes with DE in islets, only 140 (7.9%) were differentially expressed in at least one GTEx tissue (using the same per-tissue FDR corrections for GTEx described above), indicating that the majority of these T2D gene associations are likely specific to islets.

Gene co-expression analysis was then performed in islets in the same manner as for GTEx tissues above, and identified 123 mutually exclusive co-expression modules, with 26.7% of genes not being assigned to any module. Using the same method for lncRNA T2D DE enrichment as was used for GTEx tissues, one module containing 180 genes was significantly enriched for lncRNA T2D associations.

To determine whether the genes in this cluster also exhibited co-expression in any of the GTEx tissues, I calculated the proportion of genes from each GTEx module in each tissue that was present in this 180-gene islet module (Figure 4.7). Of all modules identified across GTEx tissues, the one with the greatest proportion of genes overlapping this islet module was the pancreas module previously described, with 25/54 (46.3%) of its genes overlapping. This was a significantly higher proportion ($p = 0.012$) than the module with the next highest overlap, which was a module in amygdala with 13.0% of its genes also present in the islet cluster. This shows that the islet cluster is closely related to the insulin-associated pancreas module described above, and no other module across GTEx tissues showed the same level of gene sharing.

Among the 25 genes that overlapped between the pancreas and islet modules mentioned above were all six genes that had significant DE in the pancreas module (*G6PC2*, *HHATL*, *HADH*, *IAPP*, *PPP1R1A*, and *CTB-12O2.1*). Five of these six (all but *G6PC2*) were significantly differentially expressed between islets of T2D and non-diabetic donors. As further support of consistency between genes identified in both the islet and pancreas modules, 21/25 of the overlapping genes had consistent direction of effect in DE analysis of both tissues (lower expression in samples from donors with T2D; Figure 4.8).

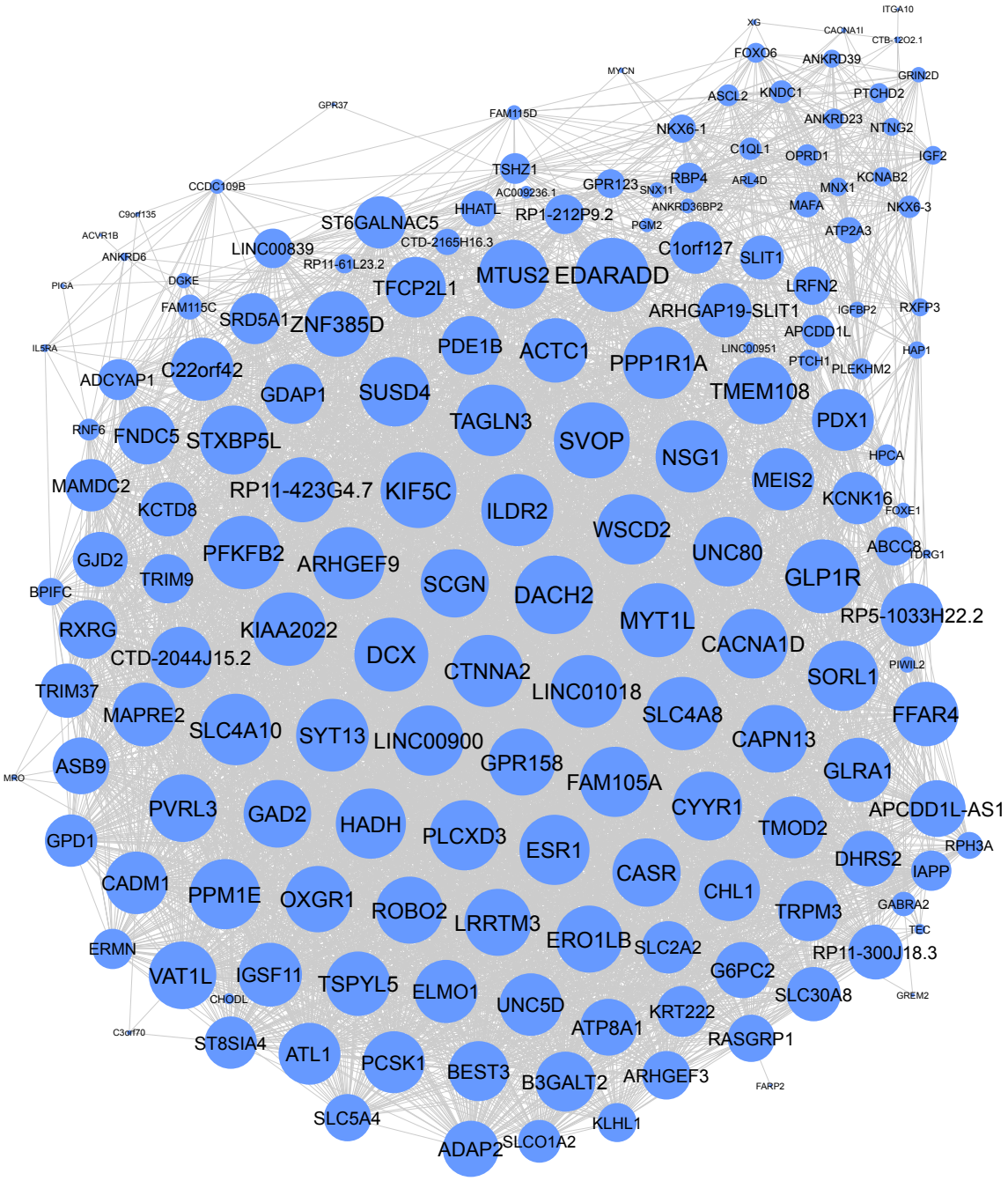


Figure 4.7: Network connectivity graph of genes in the T2D lncRNA enriched islet module. Edges are drawn between all pairs of genes with expression correlation of $r > 0.5$. The size of each gene node is representative of the number of edges.

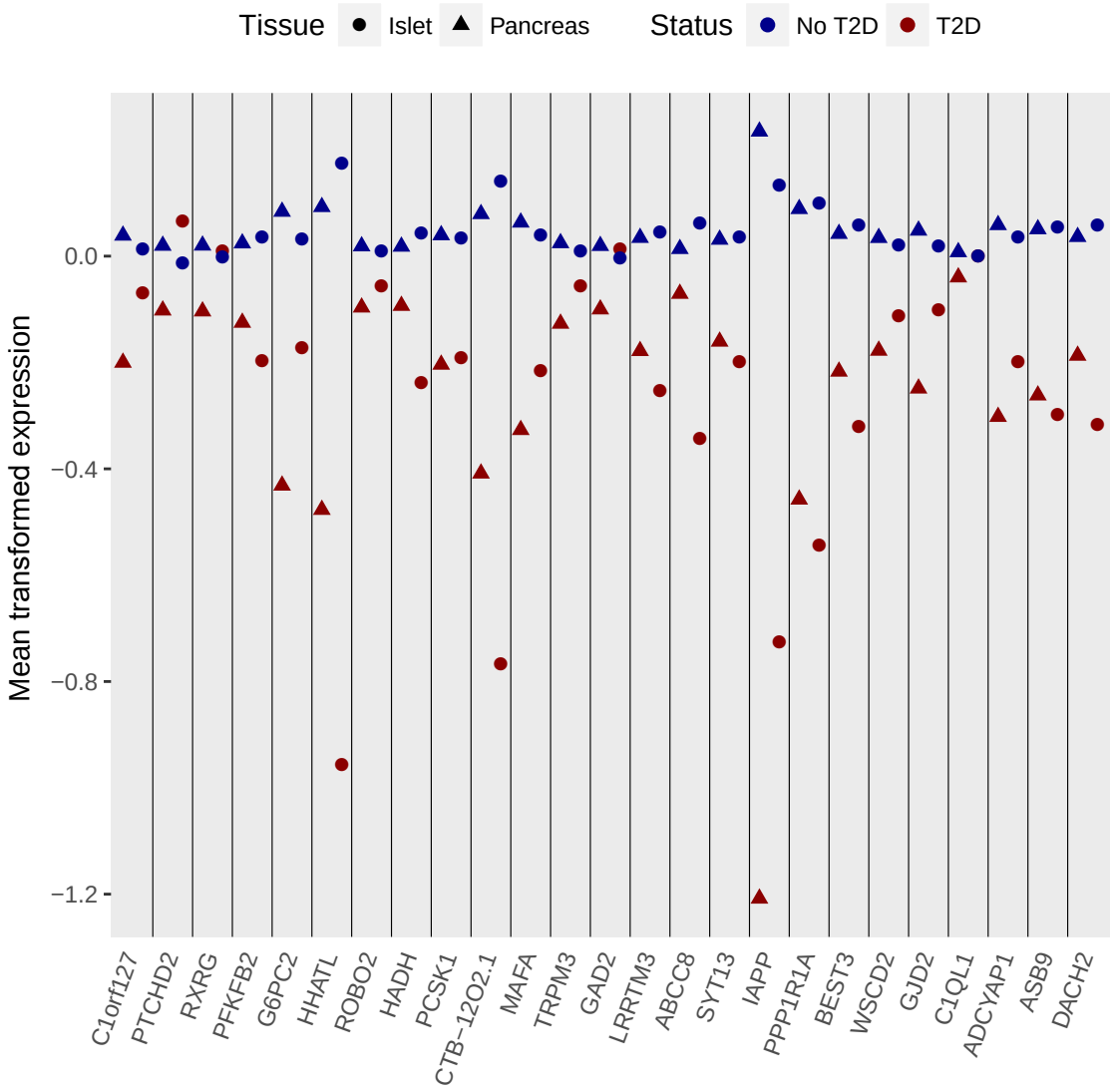


Figure 4.8: Expression of genes overlapping between islet and pancreas modules. The mean normalised expression values (y-axis; normalisation described in section 3.2.1) are shown for each gene that was present in both the insulin associated islet and pancreas modules. The tissue is indicated by the shape of the point, and within each tissue, expression is categorised by T2D status. All but 4/25 genes (*PTCHD2*, *RXRG*, *GAD2*, and *C1QL1*) exhibited consistent direction of effect between tissues, with higher expression in samples without T2D.

GO term	P-value
insulin secretion	2.05e-08
regulation of insulin secretion	2.26e-07
positive regulation of insulin secretion	0.000719
negative regulation of insulin secretion	0.00844
insulin processing	0.0347
detection of glucose	0.000803
glucose metabolic process	0.00492
response to glucose	0.00835
glucose transmembrane transport	0.0129
glucose homeostasis	0.0211
glucose 6-phosphate metabolic process	0.0342

Table 4.1: Enriched GO terms in islets. All insulin- or glucose-related GO terms with significant enrichment in the T2D-associated lncRNA-enriched islet module.

GO enrichment analysis of the T2D-associated lncRNA-enriched islet module identified 489 over-represented GO terms (FDR<0.05), including 11 terms matching text queries of 'insulin', 'glucose', and 'glucagon', echoing results from pancreas and confirming the module's relevance to glycaemic regulation (Table 4.1).

DE, co-expression, and GO enrichment analyses across a spectrum of tissues from GTEx have therefore together identified a pancreas-specific cluster of genes with relevance to insulin secretion. This cluster included one lncRNA, *CTB-12O2.1*, with tissue-specific T2D DE and expression association with other insulin-regulation-associated genes. Identification of a similar cluster of co-expressed genes (including *CTB-12O2.1*) in an independent sample of pancreatic islets verified a robust association between the expression of *CTB-12O2.1* and T2D, marking it as an ideal candidate for follow-up analyses.

4.3.3 *CTB-12O2.1* and *GLRA1*

Analyses in the above sections pointed to the lncRNA *CTB-12O2.1* as having potential involvement in regulation of insulin:

- *CTB-12O2.1* is significantly differentially expressed between T2D cases and controls in islets and pancreas, but none of the other 47 GTEx tissues analysed.

- *CTB-12O2.1* expression is associated with expression of genes involved in the regulation of insulin secretion in both whole pancreas and islet, suggesting *CTB-12O2.1* may be involved in similar processes.

CTB-12O2.1 was expressed at low levels across all GTEx tissues and in islets (Figure 4.9), and no tissue was an expression outlier. Despite its general low expression, *CTB-12O2.1* passed expression filtering thresholds for 36/48 GTEx tissues and for islets (more than 6 reads and CPM >1 in at least 10 samples). Although widely expressed across tissues, DE analysis by T2D status identified significant DE only in pancreas and islet (Figure 4.10).

CTB-12O2.1 is an intergenic lncRNA located on chromosome 5, with *GLRA1* (glycine receptor alpha 1) as its most proximal coding gene (Figure 4.11). Given that there has been growing evidence of lncRNAs that regulate expression of proximal coding genes,[269] the relationship between *CTB-12O2.1* and *GLRA1* was examined further. Four of the six tissues with expression of both *CTB-12O2.1* and *GLRA1* exhibited expression correlation between the genes ($r > 0.5$): cerebellum, cerebellar hemisphere, hypothalamus, and islets. The strongest expression correlation, $r = 0.67$, was observed in islets, and this effect was consistent when the data was subsetted to test only samples from donors with T2D ($r = 0.65$) and nondiabetic donors ($r = 0.57$; Figure 4.12). *GLRA1* was significantly down-regulated in islets of T2D donors compared to non-diabetic islets ($p = 3.44 \times 10^{-15}$), but there were no significant expression changes in other tissues. This is consistent with a past microarray study that identified *GLRA1* as being the most significant gene associated with insulin secretion and HbA1c in islets.[95] The same study also found down-regulation of *GLRA1* expression in islets exposed to short-term high glucose.[95] A more recent study by Ottosson-Laakso *et*

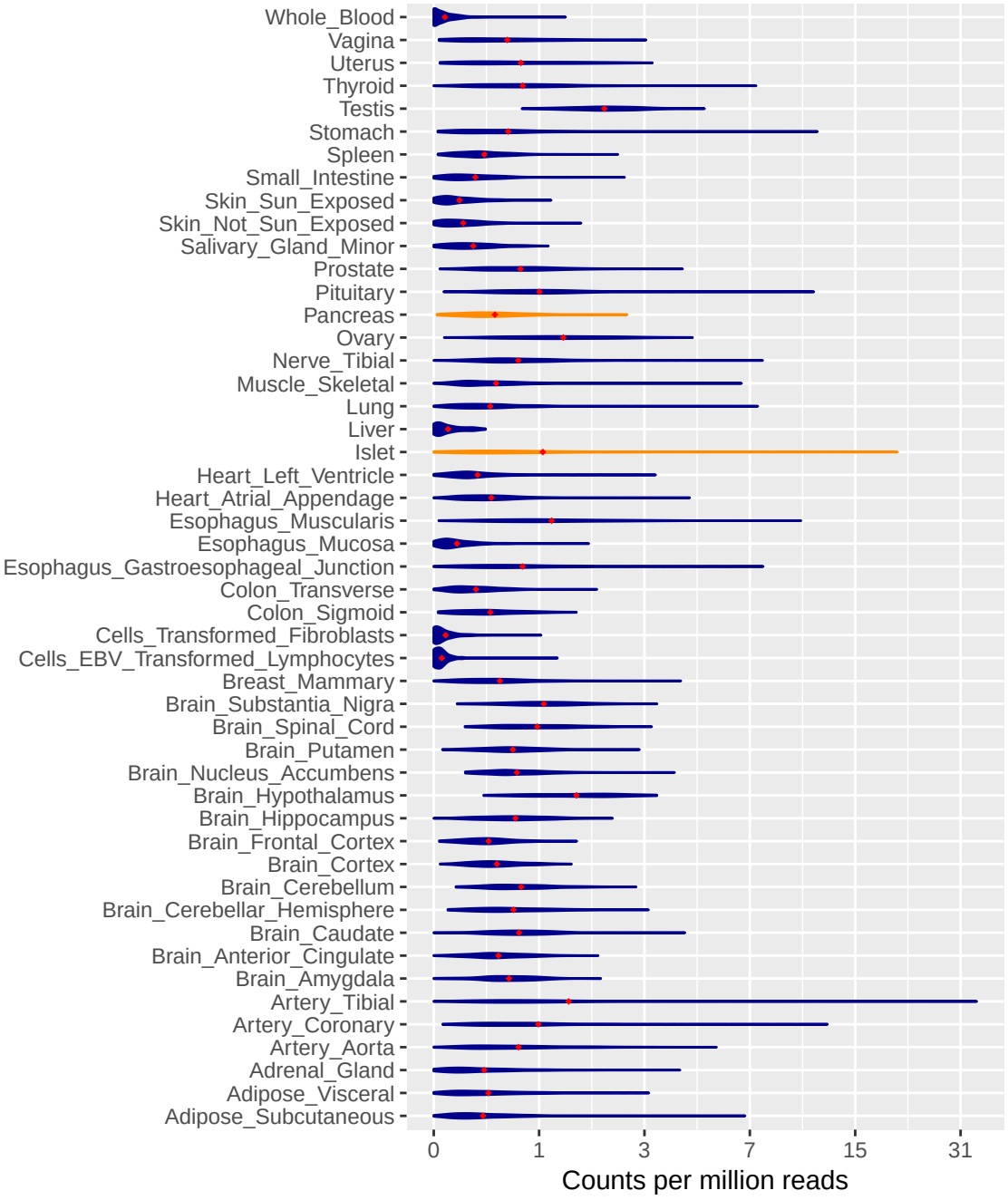


Figure 4.9: Distribution of *CTB-12O2.1* expression across tissues. For each tissue in which *CTB-12O2.1* was expressed, violin plots represent the distribution across samples of its expression in counts per million (x-axis, $\log_2$ scale). Pancreas and islet are shown in orange, and each red point indicates the mean expression in that tissue.

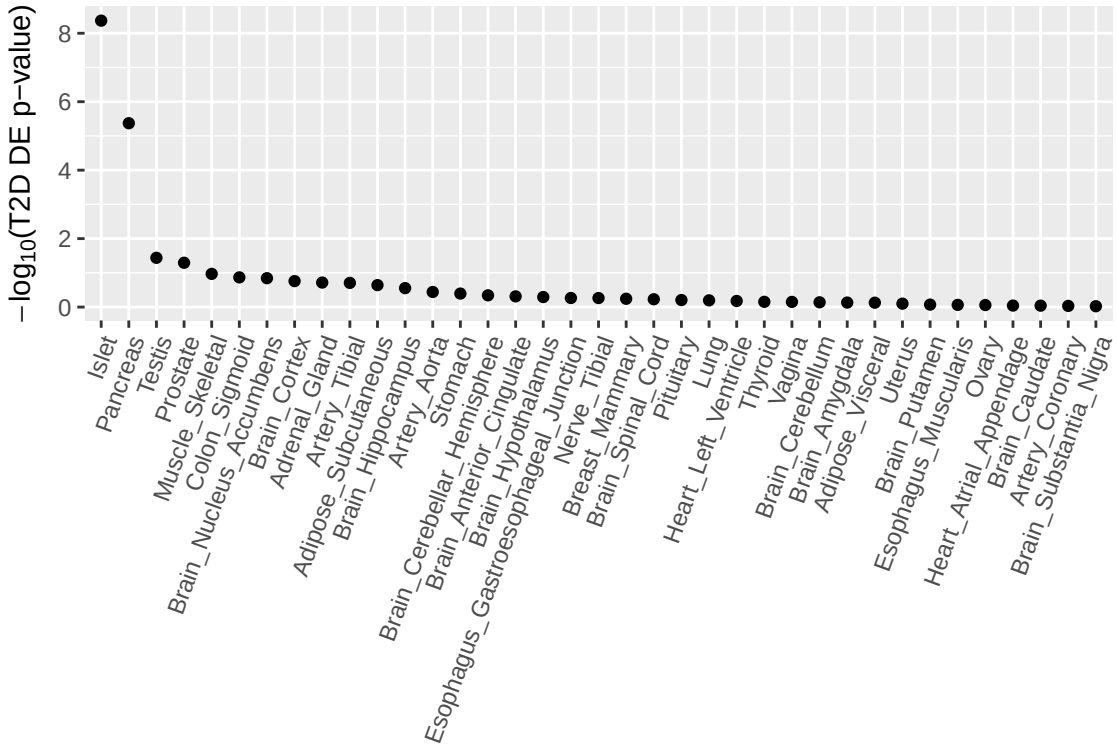


Figure 4.10: *CTB-12O2.1* T2D DE. $\log_{10}$ p-values for tests of *CTB-12O2.1* DE by T2D status in each tissue (y-axis). Only islet and pancreas had significant effects.

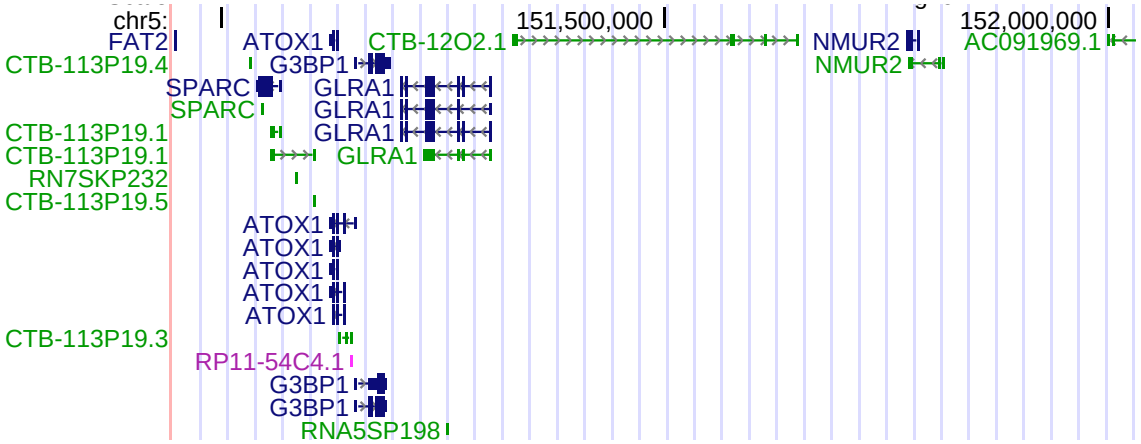


Figure 4.11: *CTB-12O2.1* and proximal genes. Genomic location of *CTB-12O2.1* and its proximal genes, including *GLRA1*. Generated by the Human Genome Browser.[270]

al. found that both *CTB-12O2.1* (named using the Ensembl ID ENSG00000254226 in their study) and *GLRA1* were down-regulated in islets in two separate comparison tests:[101]

- islets from donors with abnormal glucose tolerance compared to normoglycaemic islets
- normoglycaemic islets after incubation at 18.9 mM of glucose compared to 5.5 mM of glucose

As a result of these findings, Ottosson-Laakso *et al.* suggested that these genes (amongst others that had similar consistency of effects) were most likely to be a consequence of hyperglycaemia and less likely to be causally involved with T2D development.[101] It is important to note that a subset of the 116 islet samples used by Ottosson-Laakso *et al.* for DE analysis are part of the InsPIRE dataset, and were thus also included in my analysis.

Whether consequential or causal, consistent evidence between my analysis and that of the two studies mentioned above suggests that both *GLRA1* and the proximal long intergenic noncoding RNA *CTB-12O2.1* are correlated in islet expression and have expression that is strongly associated with elevated blood

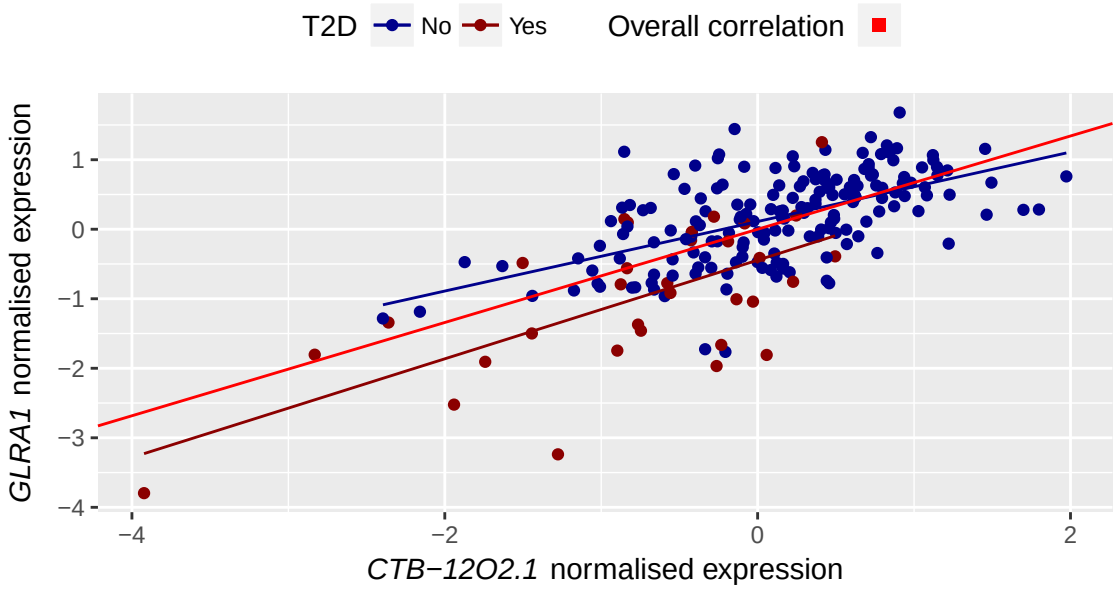


Figure 4.12: Expression correlation between *CTB-12O2.1* and *GLRA1* in islets. The correlation between normalised expression values in islets for *CTB-12O2.1* (x-axis) and *GLRA1* (y-axis). Points are separated into T2D cases (dark red) and controls (blue), with corresponding coloured best-fit lines for each subset. The bright red line indicates the correlation for all combined samples ($r = 0.67$).

glucose. The exact mechanisms behind these associations, as well as determination of whether *CTB-12O2.1* is a regulator of *GLRA1* or simply a transcriptional by-product, will require functional experimentation.

4.4 Discussion

This chapter has characterised the landscape of gene expression associations with T2D across a spectrum of human tissues, and has highlighted a lncRNA with T2D-relevant associations in both whole pancreas and islets. The main findings of the chapter were as follows:

- Nearly 3,000 genes were identified with T2D DE in at least one tissue, and more than half of the significant T2D DE associations were in islets.
- Co-expression clustering and enrichment analyses identified a cluster of genes

in both pancreas and islets that was associated with the regulation of insulin secretion.

- Integration of RNA-seq analyses pointed to one lncRNA, *CTB-12O2.1*, as being both associated with T2D status and co-expressed with genes involved in regulation of insulin secretion in only pancreas and islets.

- *CTB-12O2.1* and its most proximal coding gene *GLRA1* have correlated expression in islets and are both down-regulated in islets from donors with T2D, suggesting they may be functionally related.[95,101]

The first part of this chapter described differences in gene expression between T2D cases and controls across a spectrum of 48 human tissues in GTEx. This analysis provided a baseline of DE results with which results from islets could be compared to identify tissue specificity of effects. Of these 48 tissues, skeletal muscle had the greatest number of genes with expression differing between T2D and non-T2D samples, which is consistent with the fact that change to insulin resistance in skeletal muscle is one of the primary defects in T2D development.[261] This is consistent with previous studies that have observed systematic dysregulation of genes expressed in skeletal muscle of donors with T2D.[262–265] While there were 62 and 2 genes respectively with DE in subcutaneous adipose and liver (two other tissues that are central to T2D development), skeletal muscle contained 564 genes — nearly 10 times more — with DE. When DE analysis was extended to RNA-seq data from islets, there were more genes identified with DE in islets alone than there were combined across all GTEx tissues (using the same analysis pipeline and significance threshold). A previously published DE analysis using a subset of the islets analysed in this chapter identified a similar number of genes associated with HbA1c.[98] Given that pancreatic islets and changes to beta-cell function and mass are central to the pathophysiology of T2D,[271] it is not surprising to see significantly more differentially expressed genes than in the other

tissues studied. Whether or not these changes in RNA expression are a result of T2D-relevant changes in islets, or are causally associated with T2D development, is not clear without further analysis or experimentation.

Co-expression analyses were then used to identify clusters of similarly expressed genes that were enriched for T2D-associated lncRNAs. Functional enrichment analysis was then performed across co-expression modules in GTEx tissues, leading to identification of a gene cluster in pancreas — including lncRNA *CTB-12O2.1* — that contained genes involved in the regulation of insulin secretion. Performing the same series of analyses in RNA-seq of pancreatic islets identified a co-expression cluster containing almost half of the genes in the pancreas cluster (including *CTB-12O2.1*), supporting the hypothesis that *CTB-12O2.1* and the associated set of genes are relevant in insulin secretion regulation. A recent study by Akerman *et al.* employed an approach similar to the one in this chapter to infer lncRNA function in islets: they first conducted co-expression analysis with WGCNA, then tested for enrichment of beta-cell specific lncRNAs and islet-specific transcription factor genes. This led to the identification of *PLUTO*, a lncRNA that through functional follow-up was shown to regulate expression of *PDX1*, a known regulator of beta-cell function which has been implicated in T2D pathogenesis.[103] This recent publication provides a promising example of how the methods I employed in this chapter have been used as a first step in successfully identifying a true functional lncRNA.

CTB-12O2.1, a lncRNA that exhibited significant downregulation in islets and pancreas of T2D donors, was also found to be highly correlated in islets with its most proximal coding gene *GLRA1*. *GLRA1* was also significantly down-regulated in glucose-stimulated islets in two previous studies.[95,101] Considering *GLRA1*'s genomic proximity and correlation with *CTB-12O2.1* expression, it is likely that both genes are part of the same regulatory process. Although a previous knockdown experiment of *GLRA1* expression in islets did not causally affect

insulin secretion,[95] it cannot be assumed that *CTB-12O2.1* expression does not have a causal effect on insulin secretion without further study. The knockdown experiment of *GLRA1* was also not followed by RNA-seq, so it remains to be seen whether *CTB-12O2.1* expression is affected by changes in *GLRA1* expression.

There are several considerations when interpreting the results of this chapter. In Chapter 3, it was shown that 969 expressed lncRNAs in a different sample of islets were undetected using poly-A-tailed RNA selection, which is the protocol that has been used for all samples described in this chapter. The analyses above therefore do not exhaustively characterise T2D-associated lncRNAs, particularly those that are non-polyadenylated. There was also an observed correlation between the number of genes with significant DE and the tissue sample size in GTEx tissues, indicating that detection of DE effects may be limited with the current GTEx sample sizes. The null distributions used for the module lncRNA enrichment analysis may be biased since the permuted sampling of lncRNAs was random, while the co-expressed lncRNAs tested may be biologically associated (for example, by sharing a transcription factor binding site[272]). This bias could lead to false positive enrichment. However, this was not of concern since the lncRNA enrichment analysis was used only as a method of module prioritisation for follow-up GO enrichment tests, and many co-expressed genes in the single pancreas module that was followed up were also co-expressed in an independent sample of islets.

It is also important to note that although the results presented above (DE, co-expression, and GO enrichment) were consistent in whole pancreas and islet tissues, given that pancreatic islet samples contain varying proportions of exocrine cells[98] and islets account for $<2\%$ of total pancreas volume,[90] it is difficult to discern whether the patterns detected were driven by islet endocrine cells (such as insulin-secreting beta-cells), pancreatic exocrine cells, or a combination of cell types. Other analysts in the InsPIRE collaboration performed tissue deconvolution

in the InsPIRE islet RNA-seq and estimated that pancreas exocrine contamination ranged from 1.8% to 91.7% per sample (mean 33.5%). They further estimated that beta-cells constituted a median 58.8% of the endocrine portions of the samples. Beta-cells are therefore likely the primary cell type making up the InsPIRE islets, but given the modest portion of pancreas exocrine cells and non-beta endocrine cells, it is possible that transcriptomic patterns identified using RNA-seq from InsPIRE islets may stem from these exocrine and non-beta cells.

While the identification of *CTB-12O2.1* as a relevant lncRNA in insulin secretion regulation offers a promising outlook on the potential role of lncRNAs in T2D pathogenesis, expression data is not sufficient for inferring causality or function. The following chapters describe the integration of genetic data with islet RNA-seq data to further characterise the gene expression associations with T2D and regulation of insulin secretion that have been described above.

Chapter 5

Genetic regulation of gene expression

5.1 Introduction

As described in Chapter 1, genetic models of gene expression using both single variants and multiple joint single nucleotide polymorphisms (SNPs) have become widely used in recent years to identify genes with expression that mediates genetic variant effects on phenotype. Similar to the abundance of genome-wide association study (GWAS) summary results being continuously published across a wide spectrum of phenotypes and diseases,[273] large-scale expression quantitative trait locus (eQTL) datasets across a spectrum of human tissues and cell-types such as those in the Genotype-Tissue Expression Project (GTEx) are becoming widely available.[89] In addition to GTEx and other multi-tissue databases,[88] many smaller-scale tissue- and cell-type-specific eQTL studies have been designed to better understand relevant functional mechanisms. Gene expression data from human bronchial epithelial cells was recently used to identify cell-type-specific gene expression associations with asthma-associated SNPs,[274] while eQTL data from monocytes and T-cells has been used to identify variants associated with Alzheimer's disease that also regulated gene expression only in monocytes.[275] In the realm of type 2 diabetes (T2D) and glycaemic traits, several recent studies

have successfully leveraged eQTL data from both whole human islets and beta-cells in order to identify tissue- and cell-type-specific mechanisms through which T2D- and glycaemic-trait-associated variants and genes may act.[96,98,100–102,111,115]

As an example, one study by van de Bunt *et al.* used RNA sequencing (RNA-seq) from human islets of 118 individuals, the majority of which are included in the Integrated network for systematic analysis of Pancreatic Islet RNA Expression (InsPIRE) collaboration dataset, to identify 14 effector transcripts that mediated GWAS SNP associations with T2D and glycaemic traits.[100] This result was then used to inform an experiment showing that increased expression of *ZMIZ1* (zinc finger MIZ-Type Containing 1 gene) impaired glucose- and calcium chloride-induced insulin secretion. A more recent study by Varshney *et al.* used RNA-seq from 112 islet samples to identify 28 GWAS variants that likely also represented eQTLs in islets.[102] Similar studies are described more exhaustively in Chapter 6. Although gene-level disease association tests that incorporate multiple SNPs per gene can lower the multiple testing burden, exploit more comprehensive genetic data, and identify associations with small effects,[156,157] multi-variant gene expression tests for T2D association have only been published for non-islet data.

The PrediXcan method has been used with RNA-seq from lymphoblastoid cell lines to identify one significant gene association with T2D (although the gene was not listed),[276] while a different study found no significant T2D associations using multi-variant PrediXcan models based on whole blood RNA-seq.[277] The most comprehensive such analysis used the S-PrediXcan method with genotypes and RNA-seq of 44 tissues from GTEx to identify 90 significant associations with T2D, 37 of which were genes outside of known T2D-associated loci.[278] Given the tissue-specific nature of gene expression regulation and the importance of pancreatic islets in the pathogenesis of T2D, along with the success of co-localisation efforts using islet expression data[98,100–102,115] and multi-variant analysis using non-islet data,[276,278] it follows that multi-variant gene-level tests for

T2D association in pancreatic islets may reveal new regulatory pathways relevant to glycaemic traits and T2D.

Although methods such as two-sample Mendelian randomisation (2SMR)[164] and S-PrediXcan[163] are suitable for identifying cases in which gene expression may mediate SNP effects on phenotype, these methods typically rely on two assumptions: i) SNPs selected for each gene-based test should be significantly associated with gene expression, and ii) SNPs should be uncorrelated in order to make sound statistical inference.[174,279] While these assumptions are described in greater detail in Chapter 6, it is important to note that the multi-variant expression studies described above all used elastic net regression[160] to identify sets of SNPs associated with each gene's expression and determine final SNP association coefficients. Elastic net regression incorporates a penalty which is a weighted mixture of ridge regression[161] and least absolute shrinkage and selection operator[162] (LASSO) regression penalties. Ridge regression results in models with smoothed coefficients but performs no variable selection, while LASSO regression's penalty shrinks unassociated variable coefficients to 0, leading to sparse models (thus acting as variable selection). Elastic net regression, which is a mix of ridge and LASSO, thus inherently shrinks regression coefficients of some variables to 0 (corresponding to SNP selection in this application) but to a lesser extent than LASSO regression, and it has been shown before that elastic net regression still often maintains many variables that are highly correlated which introduces multicollinearity, making resulting coefficients unstable and difficult to interpret.[280]

It has been shown that genetic regulation of gene expression is sparse in nature, and that despite LASSO regression performing stricter and sparser variable selection than elastic net, genetic models of expression based on LASSO regression have similar predictive power to elastic net models.[162,281] Thus, given the downstream requirement in 2SMR analyses that SNPs be uncorrelated and associated with gene expression, LASSO regression may provide the most suitable

out-of-the-box platform for sparse SNP selection in gene expression models. Although using LASSO for SNP selection in genetic models of gene expression results in models that are more sparse than elastic net regression models, whether or not SNPs selected by LASSO regression are truly independent has not been explored in this context. Thus, in preparation for eQTL co-localisation and 2SMR analyses in Chapter 6, using genotype and RNA-seq data from the InsPIRE collaboration, this chapter has the following aims:

- Compare the predictive power of LASSO regression models to models of only the most significant eQTL per gene.
- Assess the robustness of penalised regression analysis as used in methods such as PrediXcan and S-PrediXcan and offer a solution to improve reproducibility of results.
- Demonstrate the inability of LASSO regression to select truly uncorrelated sets of expression-associated SNPs for each gene.
- Implement a SNP filtering and modelling pipeline to be used in addition to LASSO regression, which results in increased model sparsity, decreased multicollinearity, and improved predictive performance.

5.2 Samples and methods

5.2.1 RNA-seq filtering and normalisation

RNA-seq data from human islets of 426 individuals from the InsPIRE collaboration (section 2.2) along with corresponding genotype data was used for all analyses in this chapter. RNA-seq, mapping, and quantification are described in section 2.2.2.

Gene counts were filtered and normalised as described in section 3.2.1, and 25 surrogate variables were identified using a model explicitly defining sex. Residual

expression values after regressing out all SVs and sex were used for analyses in this chapter.

5.2.2 Genotype filtering

Primary genotype filtering and imputation as well as genotype dosage calculation were performed by other members of the InsPIRE collaboration as described in section 2.2.1. Using this pre-filtered dataset, I removed any SNP with a final sample minor allele frequency < 0.05, and further removed all SNPs with missing data. SNPs with missing entries were removed since analysis with the *glmnet* package in R used in section 5.3.2 requires all covariates (SNPs) to be complete. Only autosomal SNPs remained after this filtering, since only autosomal genotypes from the Geneva subset of islets were included in the InsPIRE dataset. A final set of 4,860,212 SNPs was used for analyses.

5.2.3 eQTL calculation

eQTL analysis was performed using residualised expression values described in section 5.2.1 and genetic dosage information for all 4,860,212 filtered SNPs described in section 5.2.2. All eQTL calculations were performed using *Matrix eQTL* version 2.1.1 in R.[282] For each gene-SNP pair specified, a simple linear additive model assuming independent homoskedastic errors was used to model the SNP dosage effect on the gene's expression (parameters `useModel=modelLINEAR` and `errorCovariance=numeric()`). Since residualised expression values were used as input in this analysis, no covariates were specified to the software.

Only *cis*-eQTLs were calculated, meaning that for each gene, *Matrix eQTL* modelled any SNP that was either a) within 1 megabase (Mb) of the start or end of the gene, or b) anywhere within the gene's defined boundaries.

5.2.4 R^2 calculation

All analyses describing model fit R^2 values were calculated in the following way:

- The coefficients for the relevant model of p SNPs and an intercept ($(p+1) \times 1$ vector) were directly extracted.
- The coefficient vector was multiplied by the $426 \times (p+1)$ matrix of covariates for all 426 samples, which consisted of the combined unit intercept vector (426×1) and genotype matrix for all modelled SNPs ($426 \times p$).
- The true expression values were modelled against the fitted expression values (calculated in the previous step) using a simple linear model, and R^2 was determined from this model.

5.2.5 LASSO regression

Given a matrix of observations $x_i \in \mathbb{R}^p$ of p covariates (SNP genotypes in this case) and responses $y_i \in \mathbb{R}, i = 1, ..., N$ for N individuals (gene expression in this case), LASSO regression aims to minimise the function

$$\min_{(\beta_0,\beta)\in\mathbb{R}^{p+1}} \frac{1}{2N} \sum_{i=1}^{N} \left(y_i - \beta_0 - x_i^T\beta\right)^2 + \lambda\|\beta\|_1$$

with $\lambda > 0$.[162, 283] This essentially represents a standard linear regression model with an added coefficient penalty $\lambda\|\beta\|_1$, which results in coefficients of unassociated covariates being shrunk to zero. This ultimately reflects an automated method of covariate selection when $p \ggg N$ as is the case in this analysis, where there are generally many more SNPs in the *cis* region than there are samples.

Cross-validated LASSO regression was run with the `cv.glmnet` function from *glmnet* version 2.0-5 in R, which uses coordinate descent to minimise the LASSO function.[283] This function partitions the data into a user-determined number of

random partitions (k), calculates a sequence of potential values of λ, and iteratively a) leaves out one partition at at time, b) calculates coefficients from the remaining data for all values of λ, and c) fits the excluded partition based on these coefficients for all values of λ. Finally, the average error across all k iterations is reported for each λ. Ten-fold cross-validation was used in this analysis (`nfolds=10`), and mean-squared error was used to measure quality of model fit (`type.measure='mse'`). After cross-validated mean-squared error values were used to determine the optimal value of λ for each gene, the final coefficients were calculated using the full dataset (no partitions) based on this value of λ.

For LASSO runs with 200 iterations that were used to find consistent and robust values of λ (described in section 5.3.2), each iteration was run as above, and the mean-squared error at each value of λ for each full cross-validated LASSO iteration was recorded. Finally the value of λ with the minimum average mean-squared error across all full cross-validated LASSO iterations was selected, and the final model coefficients were calculated using the full dataset (no partitions) based on this value of λ.

5.2.6 Linkage disequilibrium calculation

Pairwise linkage disequilibrium (LD) was calculated in the study sample for pairs of SNPs using the `LD` function from the *genetics* package version 1.3.8.1 in R.[284] This package can only calculate LD from genotypes, so all SNP dosages (continuous scale from 0 to 2) were rounded to the nearest genotype for each individual in order to estimate pairwise LD.

LD was estimated using the study sample rather than an external source such as the 1000 Genomes Project[44] to represent the LD structure in the data from which LASSO SNP selection was performed, and accurately benchmark its ability to deal with multicollinearity amongst potential covariates.

5.2.7 LASSO model SNP filtering and final model calculation

After iterative LASSO regression was used to select a set of associated SNPs for each gene's expression, these SNPs were further filtered to remove any below a defined threshold of contribution to gene expression (described in section 5.3.2), which also led to a reduction in the number of pairwise SNPs in LD in a single gene's model.

First, SNPs selected for each gene's model were determined as those with non-zero coefficients from the final models based on 200 iterations of LASSO regression. The genotypes of this full set of SNPs were modelled against the gene's expression in a simple linear model, using all 426 individuals (the 'full model'). From the full model, the SNP with the most significant contribution to gene expression was defined as the SNP with the most significant coefficient ('top SNP'). This top SNP was then modelled alone against the gene's expression, and the model R^2 was calculated. For any case where the model containing the top SNP only recovered at least 95% of the full model's R^2 for model fit, the top SNP alone was selected as optimal and no further SNPs were considered in the final model. The strict threshold of 95% was selected to ensure that the predictive power of the final models would be similar to that of the models selected by LASSO.

For cases where the model containing the top SNP alone did not recover at least 95% of the full model's R^2 for model fit, SNPs were iteratively added to the subset model, in an order determined by the rank of their corresponding coefficient p-values from the full model. As soon as the regression model based on the subset of SNPs from the full model recovered at least 95% of the full model's R^2, the SNPs in the subset model were used in the final model with no consideration for further SNPs.

Finally, for each gene, ridge regression was used to model the final filtered set of SNPs against gene expression. Ridge regression was performed using the *glmnet*

package version 2.0-5 in R,[283] and aims to minimise a function similar to that of LASSO regression described in section 5.2.5, except the LASSO penalty $\lambda\|\beta\|_1$ is replaced with the ridge penalty $\lambda\|\beta\|_2^2$. The `cv.glmnet` function was used with the same settings as described in section 5.2.5 for LASSO regression, except with `alpha=0` (which corresponds to the ridge penalty). The optimal value of λ was selected by averaging the resulting mean-squared error from 25 iterations of `cv.glmnet` for each gene. The reduction from 200 iterations (used for LASSO) to 25 iterations was implemented to improve computational efficiency, since all SNPs in the final models were already filtered to remove those with sub-threshold contributions, and SNP selection was not being performed with ridge regression (thus minimal variability between iterations was expected). The final SNP coefficients were then calculated using ridge regression with the full dataset (no partitions) based on this optimal value of λ.

5.3 Results

5.3.1 eQTL analysis

To understand genetic regulation of gene expression in the simplest sense, standard univariate *cis*-eQTL analysis was performed using RNA-seq from human pancreatic islet samples of 426 individuals. Although eQTL analysis has been performed previously on the InsPIRE collaboration data by Dr Martijn van de Bunt and Dr Ana Viñuela, the purpose of this section is to present results of a similar analysis using SNP filtering, gene filtering, and covariate adjustments that are consistent with other analyses in this thesis. This allows for direct comparison of eQTL results with results from other analyses.

Of the 19,605 genes that passed RNA-seq filtering (section 5.2.1), 19,503 had at least 1 SNP within the 1Mb *cis* range. Of these 19,503 genes, 2,503 were long noncoding RNAs (lncRNAs; 12.8%). There were approximately 67.6 million

SNP-gene pairs tested, with a median 3,488 SNPs tested per gene (range from 1 to 19,446). eQTLs were calculated using *Matrix eQTL* version 2.1.1 in R,[282] which identified 11,394 genes with at least one significantly associated SNP (eGenes) at a false discover rate (FDR) of 5%. The proportion of lncRNAs that were eGenes was significantly higher than the proportion of overall genes that were eGenes (p=0.008), although the proportions differed only slightly at 61.2% for lncRNAs and 58.4% overall.

At a more stringent FDR threshold of 1% (used in the core InsPIRE collaboration analysis), I identified 6,761 eGenes compared to the 4,312 eGenes identified in the InsPIRE analysis. This discrepancy is due to differences in the analysis pipelines, primarily pertaining to four factors:

- The analysis presented here considered all SNPs within 1Mb of any part of a gene definition, while the InsPIRE paper considered only SNPs within 1Mb of the transcription start site (TSS).
- The InsPIRE analysis corrected for 25 expression principal components (PCs; the number that maximised eQTL detection), sex, genotype PCs, and the laboratory of origin, while I corrected for sex and SVs, which are similar to PCs but the number of SVs is selected algorithmically and calculation takes known covariates into account (unlike PCs).
- RNA-seq filtering differed between the analyses, and the InsPIRE paper only considering coding and lncRNA genes (17,914 genes compared to the 19,605 tested here).
- The InsPIRE analysis used a permutation-based FDR calculation, while my analysis used a Benjamini-Hochberg correction across all tests performed.[285–287]

To maintain consistency of results, the 11,394 genes identified at an FDR of 5% will be used in remaining comparisons.

It has been shown in previous studies that the top eQTL SNP associations for genes are not uniformly distributed through the *cis* region of analysis and tend to cluster near the TSS.[89,288] For the top SNP association of each of the 11,394 eGenes identified, the median distance from the TSS was 80.9 kilobases (Figure 5.1), and the distribution of distances from the TSS was significantly different from a null uniform distribution (Kolmogorov-Smirnov test $p < 10^{-15}$). The median distance from the top SNP to the TSS is smaller for eGenes selected under stricter FDR thresholds, indicating that SNPs with the strongest gene expression associations tend to be closer to the TSS than other expression associated SNPs (Figure 5.2).

Genetic variants explain differing amounts of gene expression variation, with the most significant associated SNP for each eGene explaining between 2.7% and 74.6% of the associated eGene's expression variation (median 3.9%). The distributions of top SNP model R^2 per eGene differed only marginally between the full set of eGenes, only lncRNA eGenes, and only non-lncRNA eGenes (minimum paired Kolmogorov-Smirnov $p = 0.04$; Table 5.1). Considering that past studies have reported mean heritability of gene expression across all genes between 14% and 32% depending on the normalisation method used,[289,290] it is clear that the top associated SNP with each gene contributes only a portion to its overall heritability in many cases. Under the hypothesis that there are further *cis* variants that contribute to gene expression regulation of many genes, the following section uses a method of multi-variant gene expression modelling to identify a robust set of SNPs that captures a greater proportion of genetic heritability of each gene's expression than single eQTLs alone.

5.3.2 Multi-variant gene expression models

To identify an exhaustive set of SNPs that contribute to the expression variability of each gene, it would be necessary to perform a full GWAS for each

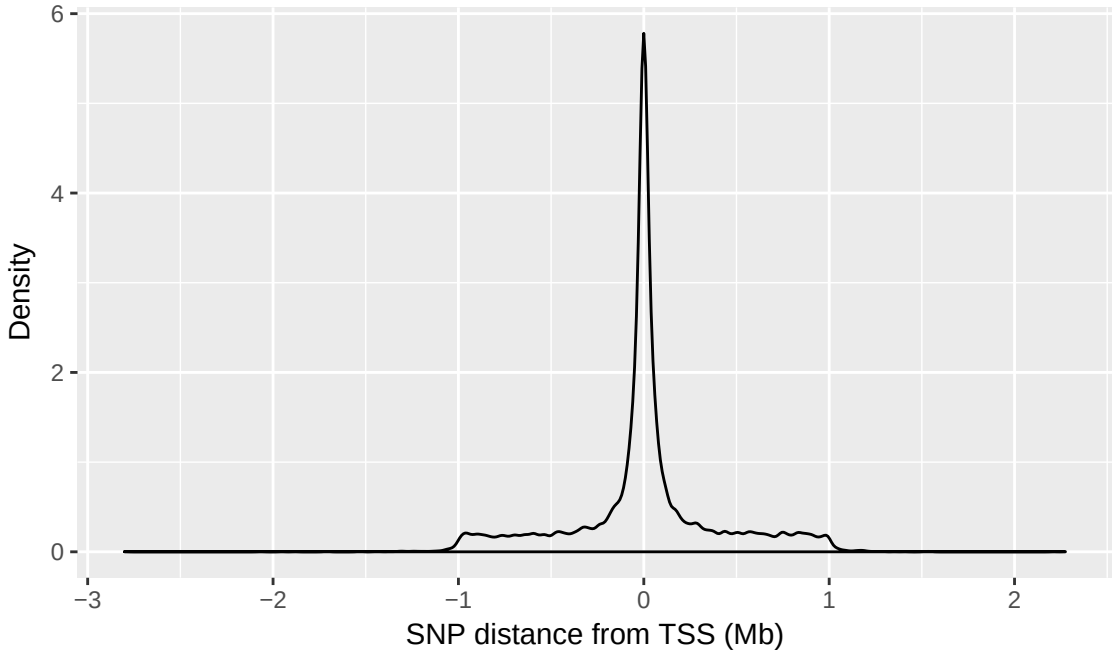


Figure 5.1: Distance from TSS for the top associated SNP of each eGene. Density plot outlining the distance from the TSS for the top eQTL SNP association in each eGene (x-axis) identified at an FDR threshold of 0.05.

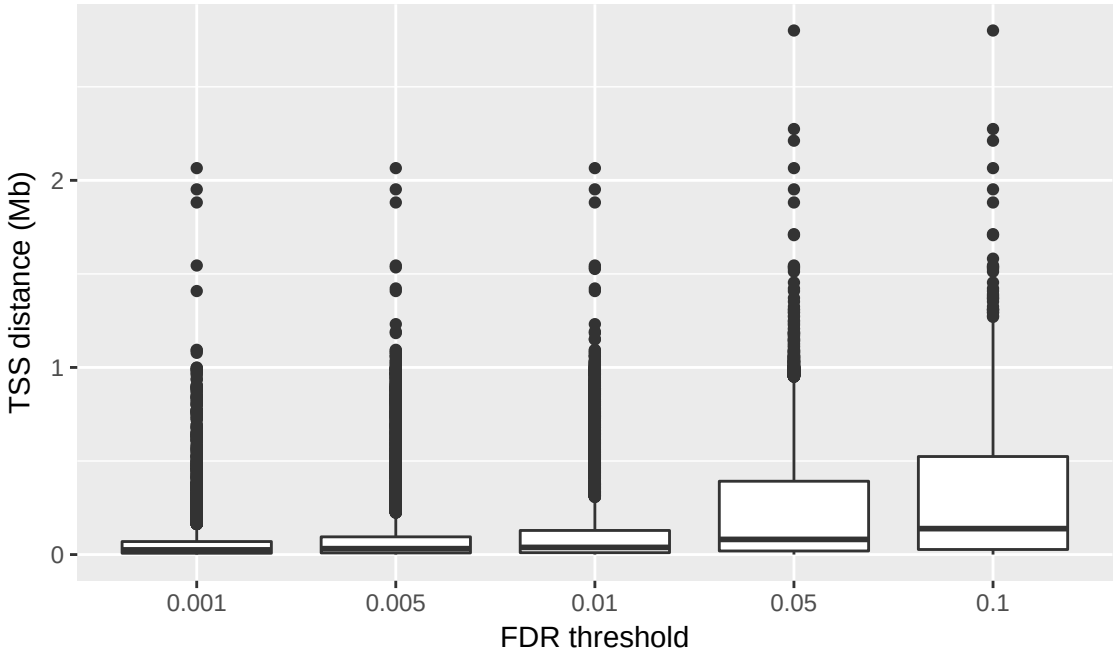


Figure 5.2: Distance from TSS for the top associated SNP of each eGene at varying thresholds. Each boxplot shows the distribution of distance from the TSS for the top associated eQTL SNP of each eGene (y-axis) across several eGene selection thresholds (x-axis). Higher FDR thresholds result in a higher median SNP distance from the TSS.

Subset	Minimum	Maximum	Median
All genes	2.7%	74.6%	3.9%
lncRNA	2.7%	72.4%	4.0%
non-lncRNA	2.7%	74.6%	3.9%

Table 5.1: Proportion of expression variation (R^2) explained by the top single eQTL per gene. The minimum, maximum, and median eGene expression variation explained (model R^2) by the single top associated eQTL variant per eGene, for different subsets of eGenes.

individual gene. In this study, such an analysis would consist of 95 billion individual tests, with whole genome sequencing studies potentially requiring more than one trillion tests. These enormous multiple testing burdens would make it impossible to confidently identify any but the strongest individual SNP-gene associations. It has been shown in previous studies that distal (*trans*) genetic variation contributes little to gene expression variability, epitomised by GTEx, where a complete *trans*-eQTL study across all tissues, genes, and SNPs identified only 673 significant associations (at a liberal FDR of 10%), compared to the 152,869 *cis*-eQTLs identified in the same publication (with a stricter FDR of 5%).[89] Given that most *trans*-eQTL effects are either not abundant or not detectable in existing studies, it is common practice to restrict even multi-variant models of gene expression to the same *cis* region used in standard eQTL studies. While *cis* variants will not exhaustively explain the heritability of gene expression variation, using all variants in a gene's *cis* region has been previously shown to explain approximately 3 times more of the heritability of gene expression than the top associated SNP alone in a microarray expression study of blood samples from 2,494 individual twins.[289]

Given that there are a median 3,488 *cis* SNPs per expressed gene in the data presented here, many of which form highly correlated LD clusters, it is desirable to identify a small subset of SNPs that explain the majority of the heritability explained by the entire *cis* region. LASSO regression is a common method of

penalised regression that is often used for variable selection.[162] LASSO regression can consider a large set of covariates (SNPs) that may or may not contribute to an outcome (gene expression), and linearly models these SNPs with an added coefficient penalty, allowing for many covariate coefficients to be optimised at 0 (section 5.2.5). This regression method is often used as an automated means of identifying a sparse joint model of SNPs that together capture the majority of expression heritability explained by the entire *cis* region.

LASSO regression

Cross-validated LASSO regression was implemented using *glmnet* version 2.0-5 in R[283] to select a sparse set of *cis* SNPs that were jointly associated with the expression of each gene (section 5.2.5). This implementation automatically generates a sequence of potential LASSO penalty weights (λ) to use based on the data, and then randomly partitions the data into a user-determined number of partitions to ultimately select the value of λ that results in the model with the smallest mean-squared error. A popular wrapper for applying *glmnet* to select gene expression models from pools of SNPs, PrediXcan,[156] selects a value of λ from a single cross-validation iteration (with a random partition of data), despite the authors of *glmnet* recommending that users run cross-validated models many times in order to reduce randomness of results. Given the discrepancy between advice from the authors of *glmnet* and the single iteration implementation used in the highly cited PrediXcan method, I first assessed the extent to which randomness in cross-validation partitions affects results between iterations of *glmnet*.

To test the concordance between independent single iterations of cross-validated LASSO regression with *glmnet*, using the same set of genes and SNPs as for eQTL analysis (section 5.3.1), two independent identical single iterations of cross-validated LASSO regression were performed. In many cases, LASSO models were optimised using the intercept-only model, indicating that no

SNPs contributed significantly to the expression variation of those genes. In total, 11,145 genes (57.1%) had at least one SNP selected in both of the two iterations. However, the initial LASSO iteration also selected at least one SNP for 1,305 genes that were given intercept-only models in the second iteration, while the second iteration resulted in models with SNPs for 1,283 genes that were optimised with intercept-only models in the first iteration. This shows that for each iteration, more than 10% of genes with models inclusive of at least one SNP were unique to that iteration only. Of the 11,145 genes for which LASSO selected at least one SNP in both iterations, only 3,822 (34.3%) had identical models. To compare the overall model performance of the iterations, the model R^2 was calculated for each of the 11,145 overlapping genes for each iteration (section 5.2.4). Although there was strong correlation between iterations for model R^2 ($r = 0.931$), Figure 5.3 shows high variation between iterations, particularly for genes with model $R^2 < 0.4$.

Given the discrepancy between independent iterations of *glmnet*, I performed two further independent runs of LASSO regression, except in each run, the optimal value of λ was selected by averaging 200 random cross-validated iterations of LASSO regression as recommended by the authors of *glmnet*.[283] The number of iterations was selected based on computation feasibility. There were 11,800 genes for which both runs selected at least one associated SNP, and there were only 89 and 101 genes respectively with at least one SNP selected by only one of the first or second runs. The model R^2 was almost perfectly correlated between runs for the 11,800 genes with SNPs selected in both runs ($r = 0.997$), indicating that although the results do not perfectly replicate, there is little variation between runs of *glmnet* when λ is selected by averaging over 200 iterations (Figure 5.4). Of the 11,800 genes modelled in both runs, 10,580 (89.7%) had identical models between runs — an improvement of 55.4 percentage points over using only a single iteration for each run.

As suggested by the authors of *glmnet*, averaging model error across many

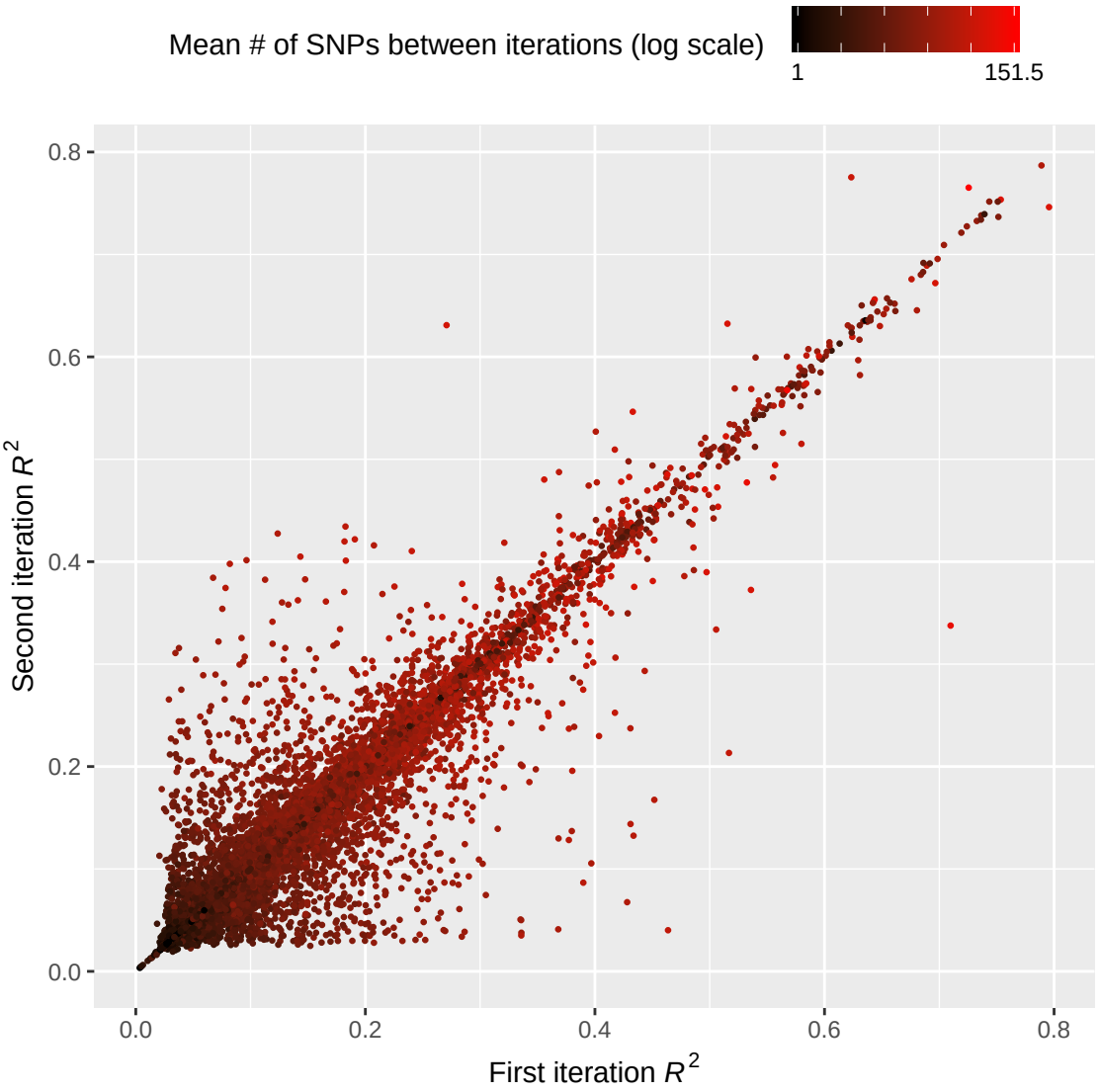


Figure 5.3: Model R^2 concordance between LASSO iterations. Scatterplot showing the model R^2 concordance between two independant identical LASSO iterations for 11,145 genes with at least one selected SNP in each iteration. The colour represents the mean number of SNPs in each gene's model between the two iterations (log scale; black indicating a single SNP, red indicating the maximum observed number). Although R^2 values between iterations were correlated ($r = 0.931$), high variation between iterations can be seen, particularly for genes with model $R^2 < 0.4$.

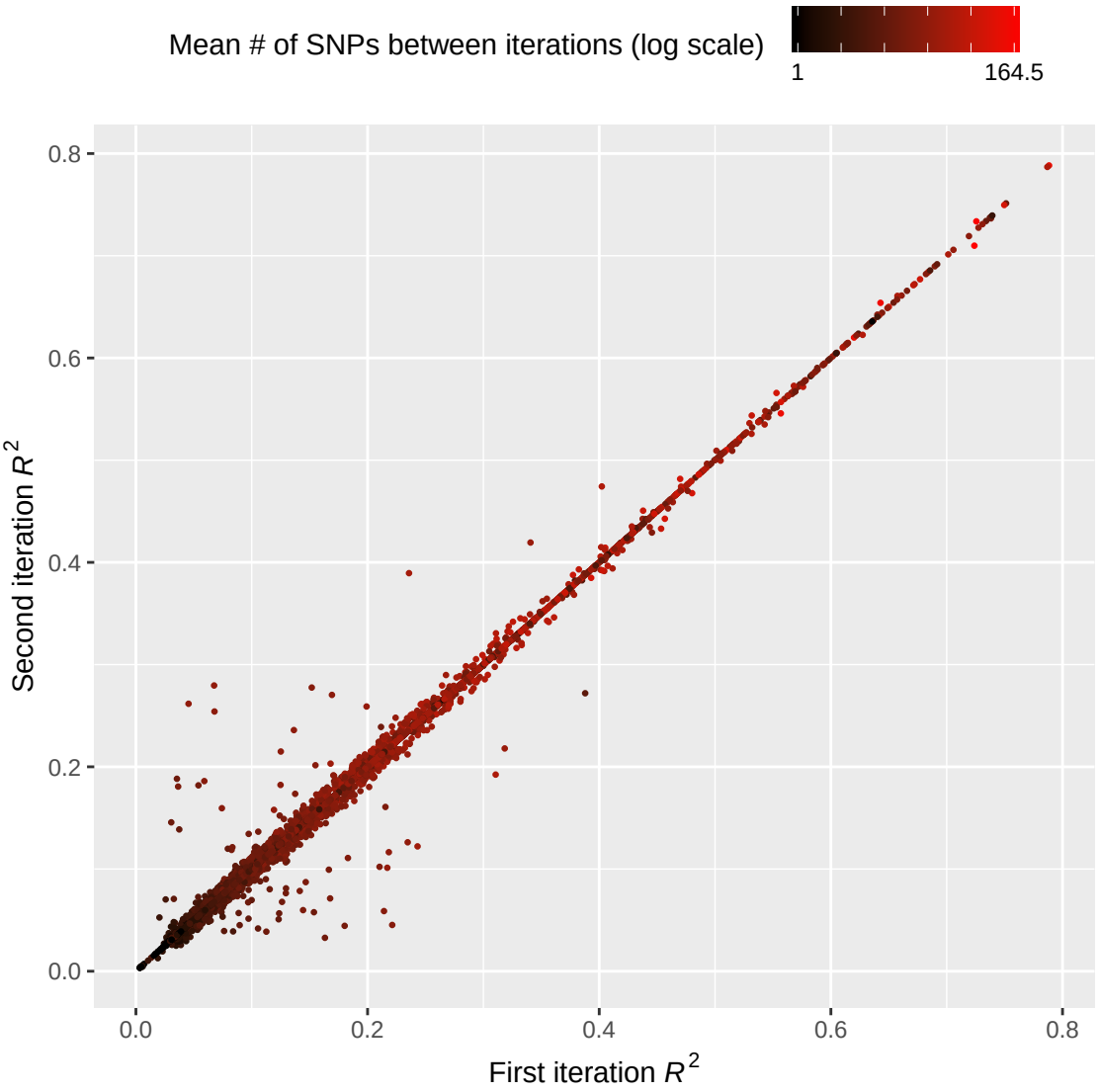


Figure 5.4: Model R^2 concordance between multi-iteration LASSO runs. Scatterplot showing the model R^2 concordance for 11,800 genes with at least one selected SNP in each of two independant identical runs of LASSO regression, each of which used the average mean-squared error across 200 cross-validated iterations to select the optimal value of λ. The colour represents the mean number of SNPs in each gene's model between the two iterations (log scale; black indicating a single SNP, red indicating the maximum observed number). R^2 values between runs were almost perfectly correlated ($r = 0.997$).

iterations does indeed reduce randomness in results. This reduction in randomness compared to using a single iteration of *glmnet* is important for reproducibility of results, thus for remaining analyses in this chapter, the first 200-iteration run of LASSO will be used (randomly chosen from the two independent runs). As the number of iterations used to determine the optimal λ increases, the replication rate would also increase; however this comes at considerable computational expense. Although the computation time was not explicitly recorded in this analysis, it was estimated that a single run of 200 iterations across all genes required more than one year of computation time (based on a single computing core).

Comparison with eQTLs

Of the 11,889 genes selected by the first 200-iteration run of LASSO regression, 11,686 (98.3%) had LASSO models that included the top eQTL SNP identified in section 5.3.1. The lack of inclusion of the top eQTL for the remaining 203 genes is due to the addition of the LASSO penalty factor to the simple linear model used in single eQTL analysis, which resulted in prioritisation of a different SNP over the top eQTL SNP for these genes. For each of these 203 genes, the LASSO-selected SNP with the strongest association in single eQTL analysis was ranked in the top 3% of the gene's eQTL associations, suggesting that LASSO regression still selected SNPs that were strongly associated with expression of these genes.

Of the 2,517 lncRNAs that were tested using LASSO regression, 1,581 (62.8%) had a final model with at least one SNP. This was only nominally different from the proportion of non-lncRNAs with at least one selected SNP (60.3%; $p = 0.018$). The proportion of lncRNAs with at least one SNP selected by LASSO was also not significantly different compared to the proportion of lncRNAs that were eGenes in the eQTL analysis (60.9%; $p = 0.16$). The median number of SNPs selected by LASSO per gene was 8 for both lncRNA and non-lncRNA genes. The median SNP distance from the TSS was 385.6 kilobases (compared to 80.9 kilobases for the top

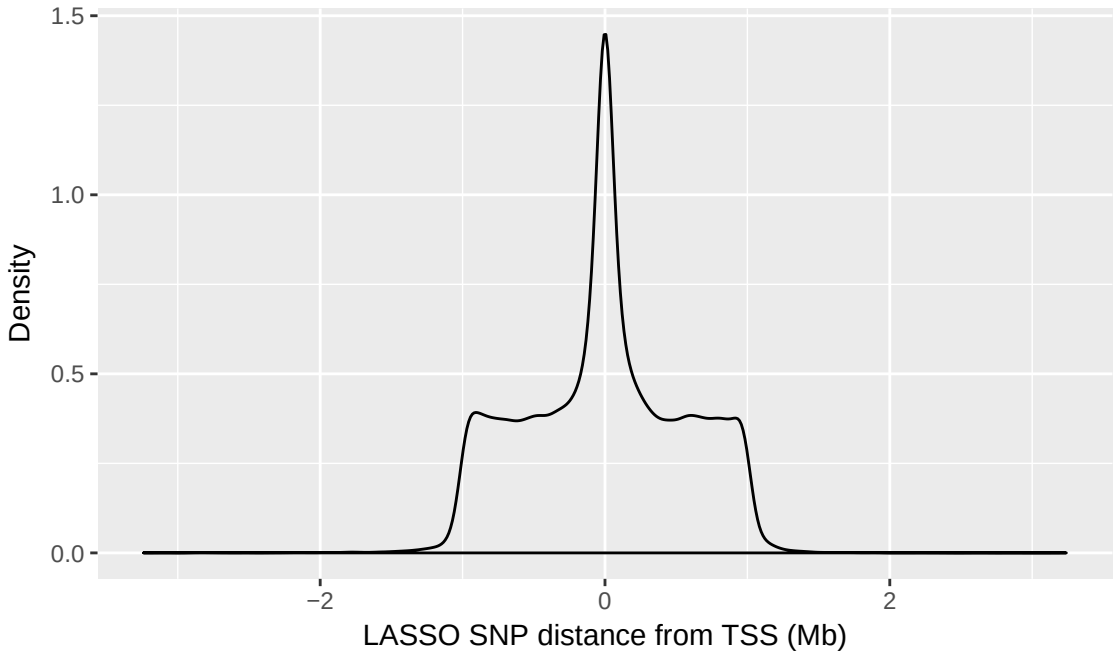


Figure 5.5: Distribution of distance to TSS for SNPs selected by LASSO. Density plot outlining the distance from the TSS of all SNPs selected in expression models by the iterative implementation of LASSO regression.

eQTL alone), although the mode of the distribution still falls near the TSS (Figure 5.5).

For 11,383/11,889 genes, the model from the iterative LASSO procedure explained the same or greater gene expression variation than the top independent eQTL alone. Of the remaining 506 genes, the models were approximately equal, with a maximum value of $R^2_{\text{Top eQTL}}/R^2_{\text{LASSO model}}$=1.004, again due to differences in modelling procedure (Figure 5.6). The median value of $R^2_{\text{LASSO model}}/R^2_{\text{Top eQTL}}$=1.97, with a maximum 20.8-fold increase to model R^2 when using LASSO regression compared to the top eQTL alone. This result is not surprising, since LASSO regression models inherently select the top eQTL (or a close proxy), and in the majority of models, select additional SNPs. This means that the top eQTL model is in most cases nested in the LASSO regression model, resulting in improved model fit after LASSO regression.

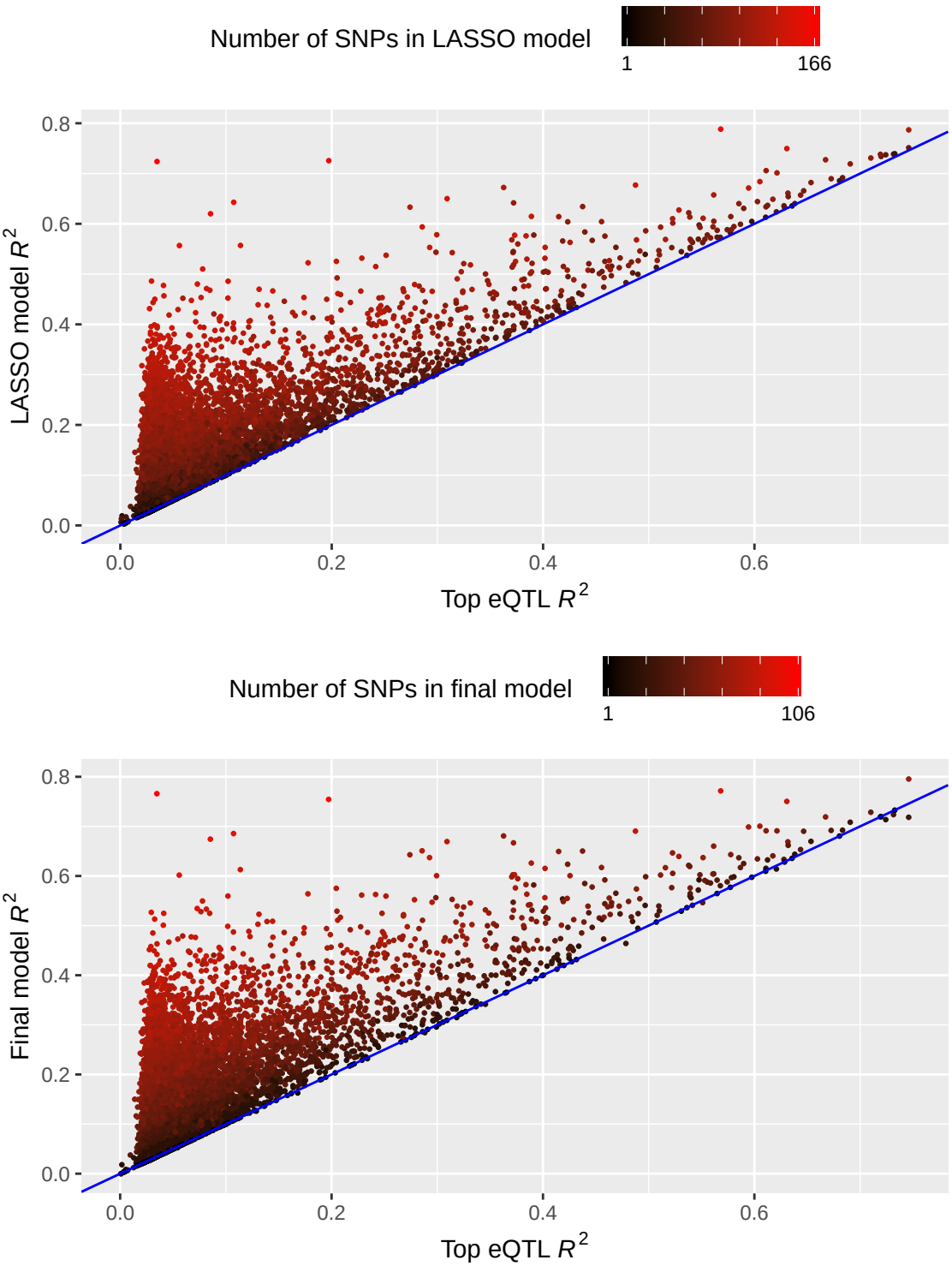


Figure 5.6: Model R^2 comparison between top eQTL and both LASSO regression models and final filtered models. Scatterplot showing the model R^2 per gene for the top eQTL alone (x-axis) against the model R^2 for each gene's corresponding LASSO model (y-axis, top plot), and final filtered multi-variant model (y-axis, bottom plot). Each point represents a gene, and blue lines have a slope of 1 and represent equality between the axes. The colour of dots represents the number of SNPs selected in the respective models (log scale).

Assessment of LD and further variant filtering

Given that LASSO regression results in models that improve over the top eQTL alone in terms of model R^2, I next assessed the efficacy of LASSO regression in selecting independent SNPs that all truly contribute to the respective models of gene expression, since it has been shown through simulation that even LASSO regression often performs imperfect selection of true variables in situations with correlated input covariates, and often selects variables that are not true predictors of selected outcomes.[291,292] Of the 10,972 genes that had models with more than one SNP selected by LASSO regression, 6,754 (61.6%) contained at least one pair of SNPs with LD $R^2 > 0.8$. This indicates that in these genes' models, there may be multiple SNPs included that represent the same signal. Although LASSO models are generally sparse, with a maximum of 166 SNPs selected for any one gene, this indicates that it should be possible to further reduce the SNP set for many genes, creating a more concise list of independent SNPs jointly associated with expression of each gene.

Thus, the following filtering process was implemented for each gene in order to remove SNPs that were the smallest contributors to the model's overall association with gene expression without significantly changing the model's overall fit:

1: SNPs selected by LASSO are modelled (simple linear regression) against the gene's expression (full model) with the model R^2 defined as R^2_{full}
2: The most significant SNP from the full model (smallest p-value) is modelled against the gene's expression with the model R^2 defined as R^2_{current}
3: **if** $R^2_{\text{current}} / R^2_{\text{full}} \geq 0.95$ **then**
4: The final filtered SNP list ($\text{SNPs}_{\text{final}}$) contains the top SNP only
5: **else**
6: The current list of SNPs ($\text{SNPs}_{\text{current}}$) is defined as the top SNP only
7: **while** $R^2_{\text{current}} / R^2_{\text{full}} < 0.95$ **do**
8: The next most significant SNP from the full model is added to $\text{SNPs}_{\text{current}}$

9: All SNPs in $\text{SNPs}_{\text{current}}$ are modelled against the gene's expression (simple linear model) with the model R^2 defined as R^2_{current}

10: **end while**

11: $\text{SNPs}_{\text{final}}$ is defined as $\text{SNPs}_{\text{current}}$

12: **end if**

After applying this filtering procedure, there was a median 22.2% reduction in the number of SNPs included per model, with the maximum SNPs selected for a single gene being reduced to 106 (compared to 166 in the unfiltered LASSO SNP sets) (Figure 5.7). After filtering, pairwise LD was tested between all SNPs in each model, and 1,707 genes (14.3%) contained at least one pair of SNPs with LD $R^2 > 0.8$. This constitutes a 74.7% reduction in the number of genes containing at least one pair of SNPs meeting this threshold when compared to LASSO SNP selection alone.

The final SNP coefficients for each gene were determined by modelling the final filtered set of SNPs for each gene against its expression using cross-validated ridge regression. Ridge regression was applied as a way to smooth and shrink the final coefficients for each gene's model. After filtering SNPs and re-modelling using ridge regression, the model R^2 increased compared to LASSO regression models for 11,033/11,889 genes (92.8%; Figure 5.7), and the R^2 comparisons between the final models and the top eQTLs alone were nearly identical to the corresponding comparisons between LASSO and eQTL models (Figure 5.6). The final models for all genes explained a median 12.8% (range 0.03% to 79.5%) of gene expression variation.

Using iterative LASSO regression to select an optimal penalty weight λ followed by further filtering of LASSO-selected SNPs and re-modelling using ridge regression has resulted in genetic models of gene expression with the following properties:

- Models are robust and largely reproducible.

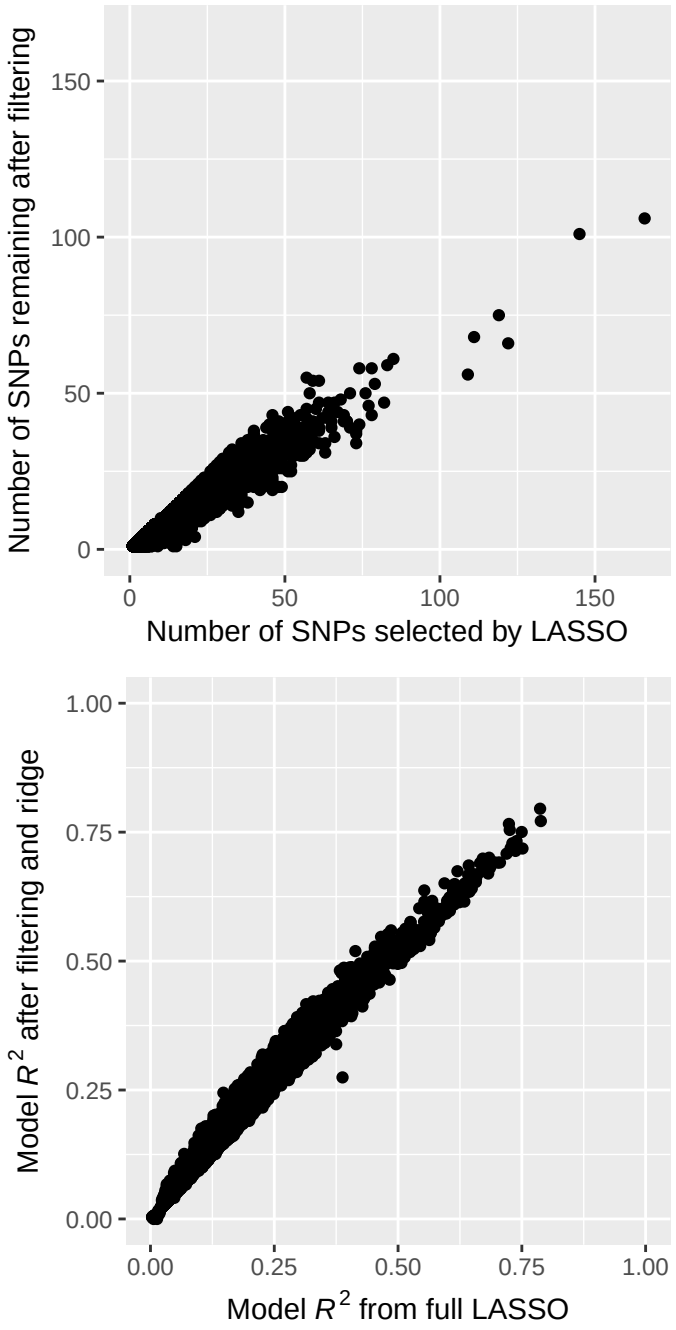


Figure 5.7: Results of LASSO SNP filtering and re-modelling. After filtering to remove SNPs with insignificant contributions to genetic models of gene expression selected using LASSO regression, the total number of SNPs per gene model was reduced by a median 22.2% (top plot). After taking the final SNP set for each gene and re-modelling the gene's expression using ridge regression, model R^2 increased for 92.8% of genes (bottom plot). For those that had decreased R^2, the decrease was generally negligible.

- Models contain fewer correlated SNPs compared to models using only LASSO regression.
- Models contain fewer SNPs with insignificant contribution to gene expression than LASSO-selected models.
- Models explain a greater proportion of expression variation than LASSO regression models in more than 90% of cases.

Results from both eQTL analysis and the multi-variant models of gene expression will be used in Chapter 6 to discern whether the effects of any T2D-associated genetic variants act through regulation of gene expression.

5.4 Discussion

This chapter has assessed the overall performance of single eQTLs for explaining gene expression variation, and has expanded on these to develop a modelling and filtering pipeline to explain a greater portion of each gene's expression variation using information from multiple SNPs in each gene's *cis* region. The main results of this chapter were as follows:

- As expected, joint multi-variant models of gene expression explain significantly more expression variation than the top eQTL for many genes.
- An iterative approach to cross-validated LASSO regression results in optimised models that are more reproducible and robust than models based on a single iteration.
- The presented pipeline for filtering SNPs selected by LASSO regression decreases the number of SNPs in most models and reduces effects of multicollinearity without significantly impacting predictive performance.

- Genetic associations with lncRNA expression follow the same patterns as non-lncRNAs in terms of proportion of eGenes, number of SNPs selected for multi-variant models, and expression variation explained by eQTLs or multi-variant models.

The first section of this chapter was a standard eQTL analysis across data from 426 individuals from the InsPIRE collaboration. Although this work does not differ significantly from the eQTL results presented in the core InsPIRE analysis, it was necessary to first prepare eQTL results using the same gene expression and SNP filtering procedures that were used for other analyses.

In the next section of this chapter, the extent to which the randomness of cross-validated LASSO modelling affected model results and reproducibility was examined. *glmnet* is a common implementation of LASSO and elastic net regression used in variant selection for many purposes. The utility of *glmnet* has been demonstrated for selecting SNPs in a genome-wide setting, using both simulated data and data from Fleckvieh bulls[291] and it is also the basis of gene expression modelling in the PrediXcan method.[156] It was demonstrated in this chapter that the differences between independent identical runs of cross-validated *glmnet* are extensive as suggested by its authors.[283] I performed two independent iterations of cross-validated LASSO regression with *glmnet*, and found that 2,588 genes had explanatory SNPs selected in only one of the iterations, and for those genes with at least one SNP selected in both iterations, only 34.3% had identical models. When performing the same analysis with the exception of selecting the optimal LASSO penalty weight by averaging 200 iterations of cross-validated LASSO regression, results were almost identical between runs.

This demonstrates the value of using an iterative approach when using *glmnet* in analysis. Failure to do so results in models that are dependent on the random data partition and are ultimately not optimised or reproducible. This raises concerns about the validity of the widely used gene models in PrediXcan and

S-PrediXcan analyses, since only a single iteration of the `cv.glmnet` function was used in their definition of expression models even in the most recent release of GTEx version 7 models (see footnote[†] for link to script). It is important to note, however, that doing many iterations (such as the 200 used in this chapter) will be dependent on the computing facilities available, as computation time increases linearly with increases to sample size or number of iterations, and at a greater than linear rate for more covariates.[283]

It was next observed that even after LASSO SNP selection, there was still multicollinearity within the selected SNPs (high pairwise LD for many pairs of selected SNPs). Although λ selection based on the minimum model mean-squared error has been previously shown to result in over-fitting,[291] suitable adjustments have not been considered in commonly used methods such as PrediXcan or S-PrediXcan.[156,163] The *glmnet* algorithm offers one solution, which is to select the largest value of λ that has mean-squared error within one standard deviation of the minimum mean-squared error. However, this method was previously shown to be too strict, selecting a maximum of $3/25$ true associated variants in a simulated GWAS dataset.[291] Thus, the pipeline proposed in this chapter provides a filtering protocol that is easy to implement, reduces the number of gene models with LD-paired SNPs ($R^2 > 0.8$), and through further modelling with ridge regression on the final selected SNPs, actually improves the overall model fit compared to LASSO in most cases, despite being based on a smaller number of explanatory SNPs.

Though modelling sets of SNPs that are in LD does not significantly alter the amount of variance in gene expression explained by a linear model, high multicollinearity in models can result in inaccurate coefficient estimates that will be highly variable between samples, and hypothesis testing on such coefficients will

[†]`https://github.com/hakyimlab/PredictDB_Pipeline_GTEx_v7/blob/101f6de5b1b552515c236c9c085bfb7e0a3842bc/model_training/scripts/gtex_v7_nested_cv_elnet.R`

have little power.[293] Thus, maintaining highly correlated SNPs in gene expression models ultimately means that the model coefficients are unstable, and interpretation and use in downstream analyses such as Mendelian randomisation will be unreliable and inaccurate. In fact, it has been shown that in summary-based Mendelian randomisation methods such as the inverse-variance weighted method commonly employed, variants used must be uncorrelated to make valid statistical inferences.[174,279] Thus, although variant filtering after LASSO selection adds a step to typical variant selection analysis, it is vital for interpreting SNP coefficients and identifying plausible causal biological mechanisms downstream.

Given the differences in the expression landscape of lncRNAs compared to messenger RNAs that was demonstrated in Chapter 3, throughout this chapter, various metrics were dissected to separate lncRNAs and non-lncRNAs and determine whether the genetic regulatory nature of lncRNAs also differs from other genes. It was observed that there were no discernible differences in these subsets of genes in terms of the proportion with significant genetic associations, the amount of expression variation explained by genetic variation (both single eQTLs and multi-SNP models), or the number of associated variants selected for the final gene expression models. Considering that lncRNAs are generally expressed at significantly lower levels than messenger RNAs,[186] and that identification of *cis*-eQTLs of low-expressed genes is more difficult than for genes with high expression,[294] this result is initially surprising. However, the difficulty in detecting eQTLs for lncRNAs in this study was likely mitigated by the RNA filtering protocol used, in that only lncRNAs reaching a strict expression threshold were included for analysis (Section 5.2.1). It is therefore possible that with less strict expression filtering, more low-expressed lncRNAs would have been tested for genetic associations, and a smaller proportion would have had significant effects.

This chapter has demonstrated that variable selection using LASSO regression

improves significantly on the gene expression variation explained by single eQTLs. It further offers a pipeline for improving the robustness, reproducibility, and quality of models selected using standard LASSO modelling procedures. The eQTLs and multi-variant models of gene expression defined in this chapter using RNA-seq from human islets will be used in Chapter 6 as a basis for identifying genes with potential causal mediation effects on T2D pathogenesis and glycaemic traits.

Chapter 6

Gene expression mediation of T2D and glycaemic trait associated genetic variation

6.1 Introduction

As discussed in Chapter 1, the majority of phenotype-associated variants identified in genome-wide association studies (GWASs) are in noncoding genomic regions. One common hypothesis is that some of these genetic associations with phenotype affect the phenotype through regulation of gene expression.[156,295,296] Identifying cases where a common set of genetic variants is associated with both a phenotype of interest and with the expression of a gene points to gene expression regulation as a probable functional mechanism through which the variants influence the phenotype. These prioritised cases can then be functionally tested to validate cellular processes.

6.1.1 Concordance between eQTLs and GWAS association variants

One way to identify genes that may act as mediators of genetic variant associations with a phenotype is by identifying variants that have significant associations with both phenotype and a gene's expression. If the same set of variants is associated with both phenotype and expression of a gene, it suggests that expression of the associated gene may mediate the variant effects on phenotype.[295] In the case of type 2 diabetes (T2D), it is also possible that even if GWAS-identified variants are associated with gene expression, the ultimate effect on T2D risk may be mediated through other phenotypes such as obesity or lipid levels.[47,297]

Two early studies, both using cadaver donor islets from the Nordic Islet Transplantation Program, tested for associations between gene expression and variants identified in T2D GWASs. The first of these, by Taneera *et al.*, used expression quantitative trait loci (eQTLs) from microarray RNA data of islets from 63 donors to identify five genes potentially regulated in *cis* by T2D-associated variants,[115] while a later analysis by Fadista *et al.* used eQTLs derived from RNA sequencing (RNA-seq) from islets of 89 individuals (average 38 million reads per sample) to identify five different T2D-associated variants with potential *cis* regulatory effects.[98] However, given the abundance of regulatory variants in the human genome,[89] it is possible that overlaps between eQTLs and GWAS variants may happen by chance, and may not represent true colocalised signals, implying that gene expression regulation may not be the mechanism through which these variants affect T2D risk.[298]

Studies have dealt with the issue of random coincidence of eQTLs and GWAS variants in a variety of ways. In a study by van de Bunt *et al.*, 218 GWAS-significant lead variants for T2D and glycaemic traits were selected and queried against eQTL results from deep RNA-seq of human islets from 118

individuals (average 72 million mapped reads per sample).[100] Through tests of linkage disequilibrium (LD) between GWAS and islet eQTL variants, followed by conditional analysis (depleting the effect of the GWAS variant in the eQTL study), fourteen loci tagging 19 independent GWAS variants were identified as having colocalised effects (effects where the eQTL and GWAS variants are thought to be tagging the same set of functional variants).[100] Similarly, Varshney *et al.* compiled a set of 214 GWAS-significant single nucleotide polymorphisms (SNPs) associated with T2D or glycaemic traits, then performed *cis* eQTL analysis with these variants using RNA-seq from human islets of 112 donors (81 from the study by Fadista *et al.* mentioned above[98] and 31 additional samples). For each eQTL association, they performed stepwise forward regression and ultimately identified 28 T2D- and glycaemic-trait-associated variants that may be colocalised with islet eQTLs for 25 unique significant eQTL genes (eGenes).[102]

In a recent study, Ottosson-Laakso *et al.* used RNA-seq from human islets of 191 donors (the largest study to date) to identify seventeen significant eQTLs (at thirteen unique loci) that were associated with T2D or glycaemic traits using simple overlapping of significant eQTLs and GWAS p-values;[101] however, this study did not assess causality or the potential of random signal overlap, since the authors' primary focus was not on genetics. The genes identified in each of these five islet studies are summarised in Table 6.1.

There have also been more sophisticated methods developed for identifying colocalised loci. Two methods that are widely used are the RTC method[298] and the coloc method.[299,300] RTC works under the same assumption that both van de Bunt *et al.* and Varshney *et al.* used: conditioning eQTL results on the causal GWAS variant will eliminate the eQTL signal at the locus if there is true colocalisation.

The coloc approach calculates posterior probabilities for each of five possible scenarios for a given locus to assess the likelihood of true colocalisation:[300]

- no eQTL or GWAS association

Table 6.1: Published genes with T2D- and/or glycaemic-trait-associated eQTL variants. Published genes where a significant islet eQTL coincided with a significant T2D- or glycaemic-trait-associated variant from independent GWASs.[98,100–102,115]

Gene	Taneera	Fadista	van de Bunt	Varshney	Ottosson-Laakso
CAMK1D	X		X	X	X
MTNR1B		X	X	X	X
ERAP2		X		X	X
STARD10			X	X	X
UBE2E2				X	X
GPSM1				X	X
PMS2P3				X	X
STAG3L1				X	X
FN3KRP					X
RP11-797D24.4					X
CDSN					X
HCG27					X
DNLZ					X
CARD9					X
LINC00261					X
RP11-331F4.4					X
SNX7					X
MADD	X		X	X	
HHEX	X				
KCNQ1	X				
GCKR	X				
AP3S2		X	X	X	
KCNK17		X		X	
ABO		X			
NKX6-3			X	X	
ADCY5			X	X	
FADS1			X	X	
DGKB			X		
ACP2			X		
WARS			X		
RBM6			X		
CTD-2260A17.2			X		
MGAT5			X		
ZMIZ1			X		
HMG20A				X	
KLHL42				X	
SIX3-AS1				X	
AC012354.6				X	
C2CD4B				X	
SIX3				X	
ABCB9				X	
VPS13C				X	
RP11-89K21.1				X	
GTF2IP1				X	
IGF2BP2				X	

- eQTL association but no GWAS association
- GWAS association but no eQTL association
- eQTL and GWAS association resulting from independent variants
- eQTL and GWAS association resulting from the same variant

The main limitation with both RTC and coloc is the inherent assumption that there can only be one causal variant at each locus, which Hormozdiari *et al.* have addressed with the eCAVIAR method.[295] In a proof-of-concept application of the method, eCAVIAR was used to identify 8 loci with high probability of colocalisation between variant associations with either fasting glucose or fasting proinsulin, and eQTLs in either Genotype-Tissue Expression Project (GTEx) tissues or pancreatic islets.[295]

The studies described above have all helped to identify potential islet-specific function of GWAS-identified associations with T2D and glycaemic traits. One consistent limitation across all of these studies is their reliance on the attainment of strict significance thresholds for both GWAS-identified variants and eQTLs. In the most recent publication by the GTEx Consortium, the discovery of eGenes was significantly affected by the tissue sample size, with no saturation of eGene discovery evident even at the highest sample sizes of 361 and 338 in skeletal muscle and whole blood respectively.[89] This indicates a significant limitation for eQTL discovery in the studies in islets, since the largest sample size used for eQTL discovery in any of the studies above was 191 in the analysis by Ottosson-Laakso *et al.*[101] Thus, these colocalisation efforts likely have limited power to detect effects, and even with a larger sample size in the Integrated network for systematic analysis of Pancreatic Islet RNA Expression (InsPIRE) collaboration, this is an important limitation to consider.

6.1.2 Multi-variant methods to identify effector genes

It has been shown that gene-level tests of association, which aggregate the association effects of multiple SNPs in a locus, can mitigate limitations of eQTL colocalisation efforts by reducing multiple testing burdens (increasing power), and identifying associations with small effects that may not be found in typical single-variant colocalisation analyses.[156,157] Given evidence that many eGenes are regulated by multiple independent *cis* variants[89,301] and that there are often many independent variants jointly contributing small effects to genetic models of gene expression (Chapter 5), gene-level tests seem an appropriate method for identifying potential effector transcripts for disease-associated genomic regions.

Not surprisingly, several methods have been developed in recent years to test for such associations. Three of the most commonly implemented methods are PrediXcan,[156] TWAS,[157] and S-PrediXcan[163] (previously known as MetaXcan).

The PrediXcan method first takes variant weights of association with gene expression (penalised regression association coefficients); then based on paired genotype and phenotype data from an independent study of interest, uses the penalised regression coefficients to 'impute' gene expression in the study dataset and associates this predicted expression with the phenotype. In the method's original publication, elastic net gene expression models from whole blood were used to predict gene expression from genetic data in the Wellcome Trust Case Control Consortium. Twenty-nine genes had imputed expression that was significantly associated with type 1 diabetes, including 21 within the extended major histocompatibility complex, which accounts for up to 50% of the familial risk of type 1 diabetes.[156,302] PrediXcan has three primary limitations:

- Model weights used in the majority of PrediXcan analyses (available from `predictdb.org`) are derived from elastic net regression models, despite issues discussed in Chapter 5 such as the multicollinearity introduced with raw penalised regression that makes calculated variant coefficients unreliable and

data-set-specific.

- The same elastic net model weights described in the point above have been derived using a single iteration of cross-validated penalised regression with the *glmnet* package in R, which was shown in Chapter 5 to produce results that are dependent on the random cross-validation data partition, and are not widely reproducible or robust.

- The PrediXcan method relies on availability of individual-level paired genotype and phenotype data from the outcome of interest, meaning that it cannot leverage results from high-powered GWASs from which only summary-level data is typically published.

The TWAS method uses genetic associations with gene expression from all *cis* variants for each gene, and calculates a linear combination of these coefficients weighted by corresponding GWAS effects as an aggregate gene-level test statistic.[157] This method uses a Bayesian linear mixed model to calculate variant effects on gene expression. Given that the majority of gene expression heritability explained by the *cis* region can be summarised using a small number of variants (Chapter 5), these models include many variants with no added benefit to gene expression models. Thus, similar to PrediXcan gene expression models, the models used in the TWAS method are overfit and include correlated covariates, resulting in less reliable analyses downstream.

One of the newest gene-level methods, S-PrediXcan, relies only on summary-level data, making it widely accessible for application across many expression studies and GWASs.[163] The S-PrediXcan method is an extension of PrediXcan, and estimates the same outcome as PrediXcan using only summary-level data from eQTL studies and GWASs.[163] To demonstrate the utility of the method, the authors of S-PrediXcan performed an analysis using gene expression models from GTEx tissues[89] and whole blood from the Depression

Genes and Networks study[303] alongside GWAS results for more than 100 traits.[163] All results from this large-scale analysis are available in an online database through `geno2pheno.org`. One concern with this database is that it uses the same genetic models of gene expression as PrediXcan, which have been derived using a single cross-validated iteration of elastic net regression. S-PrediXcan accounts for correlation between variants, which reduces the likelihood of 'double counting' the effects of correlated variants, but this does not account for the potential instability of coefficients derived from linear models with correlated covariates.[293]

6.1.3 Mendelian randomisation and its potential in gene expression studies

Mendelian randomisation (MR) is a method commonly employed to identify potential directional associations between pairs of phenotypes. MR analysis has been widely used across a vast spectrum of phenotypes to test for potential causal associations, and dozens of papers have been published describing its assumptions, limitations, potential applications, corrections, and tweaks, making it reliable, well-criticised, and robust across many applications.[171,173–175,279,304,305] It is therefore surprising that MR has not been widely used to assess the potential causal effects of gene expression regulation on phenotype and disease, considering that expression of a single gene is simply another measurable phenotype, and *cis* association models (as described in Chapter 5) can be viewed simply as 'small-scale' GWASs. Although this sort of MR application has been described as a potential method for identifying protein and RNA levels that causally influence phenotype and disease,[306] such '*cis* MR' analyses have not been widely performed.

MR is an application of traditional instrument variable (IV) analysis, where randomly assigned genotypes $\boldsymbol{G}$ are used as IVs to determine whether an exposure $\boldsymbol{X}$ has a causal effect on an outcome $\boldsymbol{Y}$. In traditional MR studies, the associations between $\boldsymbol{G}$ and $\boldsymbol{X}$ and between $\boldsymbol{G}$ and $\boldsymbol{Y}$ are typically derived from

the same samples, but more recent methods have been described to perform MR when these two associations are derived from independent datasets.[164,307] These two-sample MR (2SMR) methods are particularly applicable when considering gene expression's effect as an exposure $\boldsymbol{X}$ on an outcome $\boldsymbol{Y}$ (such as T2D), since large-scale studies measuring both of these factors are not abundant. Application of 2SMR is not complex, and several implementations exist, such as the *MendelianRandomization* package in R.[308] The more complex factor in all MR studies is the selection of a valid genetic instrument $\boldsymbol{G}$, which must meet the following requirements:[165]

- $\boldsymbol{G}$ must not be associated with any confounders of the relationship between $\boldsymbol{X}$ and $\boldsymbol{Y}$.
- $\boldsymbol{G}$ must be associated with the exposure $\boldsymbol{X}$.
- $\boldsymbol{G}$ must be independent of the outcome $\boldsymbol{Y}$ conditional on the effect of the exposure $\boldsymbol{X}$ and any confounders of the relationship between $\boldsymbol{X}$ and $\boldsymbol{Y}$.

The second of these assumptions is not generally of concern, as genetic variants are typically selected because of their observed associations with a risk factor. The SNP selection process described in Chapter 5 satisfies this assumption, since only SNPs that contribute significantly to gene expression models have been selected. It is important to note that paired datasets (eQTL and GWAS) in 2SMR should be from similar populations; otherwise, harmonisation of the two datasets in 2SMR can lead to inaccurate or biased outcomes given that both eQTL and GWAS associations may be population-specific or have differing magnitudes or directions in different populations.[172]

The other assumptions, both dealing with confounding effects, are not possible to explicitly prove, since confounding effects are generally not observed. Tests for confounding are generally performed after 2SMR to verify as comprehensively as possible that these assumptions have not been violated. Methods for assessing the

likelihood of confounding effects have been established, with the MR-Egger regression approach being the most common means for detecting violations of assumptions.[174] However, no single method is perfect, and the more extensive an analyst is in testing the assumptions of MR analysis, the more reliable the results will be. The importance of focussing on these assumptions has been summarised in a note of caution within the vignette of the *MendelianRandomization* R package:[308]

> ... it is important to remember that the difficult part of a Mendelian randomization analysis is not the computational method, but deciding what should go into the analysis: which risk factor, which outcome, and (most importantly) which genetic variants. Hopefully, the availability of these tools will enable less attention to be paid to the mechanics of the analysis, and more attention to these choices. The note of caution is that tools that make Mendelian randomization simple to perform run the risk of encouraging large numbers of speculative analyses to be performed in an unprincipled way. It is important that Mendelian randomization is not performed in a way that avoids critical thought. In releasing this package, the hope is that it will lead to more comprehensive and more reproducible causal inferences from Mendelian randomization, and not simply add more noise to the literature.

Therefore, using eQTL results and multi-variant gene expression models derived from Chapter 5, along with summary association data from large-scale GWAS meta-analyses (MAs) of T2D and glycaemic traits, this chapter has the following aims:

- Identify loci with evidence of concordant eQTL and phenotype effects.
- Assess the utility as IVs of SNPs selected for multi-variant models in Chapter 5.
- Perform 2SMR and S-PrediXcan analyses as gene-level methods to test for aggregate gene expression effects that influence T2D risk and glycaemic traits.
- Demonstrate the benefit of performing both single-variant colocalisation analyses and gene-based tests to comprehensively identify gene expression associations with T2D and glycaemic traits.

6.2 Methods

6.2.1 GWAS meta-analyses

Publicly available GWAS MA summary results for T2D, fasting glucose (FG), and fasting insulin (FI) were used for single-variant concordance, 2SMR, and S-PrediXcan analyses in this chapter. The three studies are summarised in Table 6.2.

A subset of 2,033,629 variants from the 4,860,212 variants described in section 5.2.2 that were both a) present in all three studies, and b) present in the InsPIRE data (section 2.2) with no missing values were used for analyses.

6.2.2 eQTL and GWAS concordance

Each MA described in section 6.2.1 was first subsetted to the complete set of variants with significant association p-values ($p < 5 \times 10^{-8}$). The top association signal (smallest p-value) was then selected from each 500 kilobase block. For these subsetted variants, pairwise LD was tested, and for pairs of variants with LD $R^2 > 0.8$, the most significantly associated variant was kept.

To identify variants with corresponding effects on gene expression, each of these significant variants was queried from the eQTL results presented in Chapter 5.3.1 to identify those that were also significant eQTLs at a false discovery rate (FDR) of 5%.

Since some of these overlaps may be due to chance,[100] I filtered to keep only those where the SNP was in high LD ($R^2 > 0.8$) with the lead eQTL SNP for the corresponding gene. The lead eQTL SNP was defined as the SNP that had the smallest eQTL association p-value for the respective gene within 11 kilobases of the lead phenotype-associated SNP. Though the distribution of haplotype block sizes is right-skewed, and in some cases blocks span nearly 400 kilobases, the 11 kilobase threshold was selected since Wang *et al.* observed this as the average LD

Table 6.2: Meta-analysis datasets used in Chapter 6

Consortium	Trait	Publication	# Studies	# of samples
DIAGRAM	T2D	Scott *et al.*, *Diabetes*, 2017[47]	18	26.7k cases 132.5k controls
MAGIC	FG	Manning *et al.*, *Nat Genet*, 2012[310]	29	58.1k (no T2D)
MAGIC	FI	Manning *et al.*, *Nat Genet*, 2012[310]	26	51.8k (no T2D)

block size in a population made up primarily of Europeans.[309]

6.2.3 Mendelian randomisation

All MR analyses were performed using the *MendelianRandomization* package version 0.2.0 in R,[308] except that MR analysis p-values were calculated manually due to a documented bug in version 0.2.0. The inverse-variance weighted (IVW) method[307] was implemented for all genes using the `mr_ivw` function, and MR-Egger regression[174] was performed for all genes with more than 2 variants in the final filtered regression models using the `mr_egger` function.

The IVW method is commonly implemented for MR analysis using multiple variants, and is conceptually the same as MA in GWAS settings: the coefficient of the risk factor's effect (gene expression in this case) on the outcome (trait of interest) is calculated as a weighted sum of the SNP coefficients on the trait of interest, with weights corresponding to the regression coefficients calculated in Chapter 5.[307] The test statistic, $\tilde{a}$ and its standard error $\mathrm{SE}(\tilde{a})$ are calculated for each gene as

$$\tilde{a} = \frac{\sum_{j=1}^{m} w_j \hat{\beta}_j s_j^{-2}}{\sum_{j=1}^{m} w_j^2 s_j^{-2}}$$

$$\mathrm{SE}(\tilde{a}) = \sqrt{\frac{1}{\sum_{j=1}^{m} w_j^2 s_j^{-2}}}$$

where w_j are the SNP coefficients for the m SNPs in the gene's final expression model, $\hat{\beta}_j$ are the corresponding beta values from the MA of interest, and s_j are the corresponding standard errors for the beta values from the MA of interest.[166] Assuming all m variants in the gene's model are uncorrelated, $\tilde{a}$ has a normal distribution with mean 0 and variance $\mathrm{SE}\,(\tilde{a})^2$.[166,305]

Complementing the IVW analysis is a test for heterogeneity of SNP effects, which tests whether SNPs are associated with the outcome with effect sizes that are proportional to the coefficients w_j. The test statistic used is Cochran's Q statistic, which when significant, suggests that at least one variant may have a pleiotropic effect. The Cochran Q statistic is calculated as follows:

$$Q = \sum_{j=1}^{m} w_j^2 s_j^{-2} \left(\frac{\hat{\beta}_j}{w_j} - \tilde{a} \right)^2$$

$$= \sum_{j=1}^{m} \left(\frac{\hat{\beta}_j - w_j \tilde{a}}{s_j} \right)^2$$

and follows a χ^2_{m-1} distribution under the null hypothesis of homogeneity of SNP effects (no pleiotropy).[304]

As a second test for unobserved pleiotropy, MR-Egger intercept tests were performed for each gene with more than two SNPs (a requirement for the method). MR-Egger regression is a method of MR analysis that, unlike the IVW method, includes an intercept term (allowing for non-zero intercepts). Although MR-Egger regression tests for a causal effect of the risk factor on the outcome, and also estimates the coefficient of the causal effect, estimates are often biased, low-powered, and have inflated false positive rates.[305] Instead of testing for causal inference using MR-Egger regression, it has become common to test only the intercept term from final regression models. Non-zero MR-Egger intercept terms indicate poor or invalid genetic instruments, and that there may be pleiotropic

effects that have not been accounted for. Thus, a significant non-zero MR-Egger intercept indicates that results from the IVW method are biased and unreliable.[305] MR-Egger regression was therefore used in addition to heterogeneity tests to assess the validity of each IVW test.

6.2.4 S-PrediXcan

The S-PrediXcan method is an extension of the PrediXcan method that relies only on summary data as input.[163] In a manner similar to IVW 2SMR, S-PrediXcan is essentially a linear combination of MA z-scores (SNP associations with an outcome trait) weighted by coefficients from genetic models of expression, such as those described in Chapter 5. S-PrediXcan also includes a variant covariance term to account for correlation between variants.

For each gene with at least one variant in its final expression model, the S-PrediXcan test statistic, which is approximately standard normal, is calculated as

$$Z \approx \sum_{j=1}^{m} w_j \frac{\hat{\sigma}_j}{\hat{\sigma}} \frac{\hat{\beta}_j}{\text{SE}\left(\hat{\beta}_j\right)}$$

where w_j are the expression model coefficients for the m SNPs in the gene's expression model, $\hat{\sigma}_j$ are the standard error estimates of SNP genotype dosages, $\hat{\sigma}$ is the standard error estimate of the predicted expression of the gene, $\hat{\beta}_j$ are the MA association coefficients, and $\text{SE}\left(\hat{\beta}_j\right)$ are the corresponding MA standard errors.[163]

The test statistic was calculated for each gene using a custom R script. Final coefficients derived from the models described in section 6.3.2 were used as weights w_j, with corresponding $\hat{\beta}_j$ and $\text{SE}\left(\hat{\beta}_j\right)$ from MA summary statistics as described in section 6.2.1. $\hat{\sigma}_j^2$ were calculated as the simple sample variances of SNP genotype dosages, and $\hat{\sigma}^2$ was calculated as $\mathbf{W}^T \Gamma \mathbf{W}$, where $\mathbf{W}$ is the vector of expression coefficients for SNPs in the gene's model, and Γ is the simple covariance

matrix for the genotype dosages of all SNPs in the gene's model.[163]

6.3 Results

The aim of analyses in this chapter was to identify genes with expression effects in islets that may mediate genetic associations with T2D and glycaemic traits. For all analyses presented in this chapter, filtered and normalised RNA-seq data from islet samples of 426 individuals from the InsPIRE collaboration as described in section 5.2.1 was used. In total, 2,033,629 SNPs were used for analyses, representing a subset of the 4,860,212 variants described in section 5.2.2. This subset is the intersect of variants that were available in the three GWAS MAs described in section 6.2.1, which ensured that there would be no variants in multi-SNP gene expression models that were not available in all of the GWAS MAs.

6.3.1 eQTL and meta-analysis concordance

Similar to previous studies using RNA-seq from human islets,[98,101,115] I first identified T2D- and glycaemic-trait-associated loci that are also significant eQTLs in islets. The sample size of 426 individuals with RNA-seq in my analysis is more than double the size of the next largest similar published study by Ottosson-Laakso *et al.*,[101] providing greater power to detect association overlaps in my study. The variants with concordant genetic regulatory and phenotype effects identified here were largely derived for comparisons to results from multi-variant methods in the following section to ultimately identify robust T2D and glycaemic trait candidate effector genes.

First, to identify loci harbouring genetic variants that are associated with T2D-relevant phenotypes, I identified 36, 19, and 4 significant independent variants respectively in MAs for T2D, FG, and FI at the generally accepted $p < 5 \times 10^{-8}$

GWAS threshold.[311] Two of these signals were shared between FG and FI, and four separate signals were shared between FG and T2D (based on LD $R^2 > 0.8$).

These 59 variants were involved in a total of 1,081 eQTL tests described in section 5.3.1. After re-adjusting eQTL p-values from section 5.3.1 (to account for an FDR of 5% in the subset of eQTLs corresponding to the 2,033,629 SNPs present in the MAs), I found that there were 38 significant SNP-gene pairs involving trait-associated SNPs, representing 31 unique genes and 22 unique SNPs (21 independent based on LD testing). To identify the MA and eQTL overlaps that are most likely to be tagging the same functional variant, I filtered the 38 overlaps to keep only those where the variant was in strong LD ($R^2 > 0.8$) with the lead eQTL SNP for the corresponding gene (similar to the filtering method performed by van de Bunt *et al.*[100]).

After this LD filtering, there were 19 SNP-gene pairs remaining, representing 16 unique eGenes and 15 unique SNPs (14 independent; Figure 6.1 and Table 6.3). Of these 14 independent variants, 7 were associated with T2D alone, 3 with FG, 1 with FI, and 3 were associated with both T2D and FG (Table 6.3). Of the 16 unique eGenes identified in this analysis, 11 have been published in previous studies (Table 6.1): *ADCY5*, *UBE2E2*, *IGF2BP2*, *DGKB*, *CAMK1D*, *STARD10*, *HMG20A*, *STAG3L1*, *MTNR1B*, *MADD*, and *FADS1* (Figure 6.1).

Of the five unique eGene associations identified here that have not been previously published, four were for variants associated with T2D (*PLEKHA1*, *TCF7L2*, *HSD17B12*, and *RP11-613D13.5*) and one was associated with FI (*GTF2I*). Given the highly significant associations in both MA and eQTL data for some of these variants, it is somewhat surprising that some of these loci have not been identified in previous studies. In the case of rs2292626 for *PLEKHA1* (MA $p = 1.8 \times 10^{-12}$, eQTL $p = 3.21 \times 10^{-14}$; Figure 6.2), the lead MA variant was in near perfect LD with the lead eQTL variant ($R^2 = 0.99$). For *HSD17B12*, rs1061810 was the lead variant in both the MA and the eQTL analysis (MA

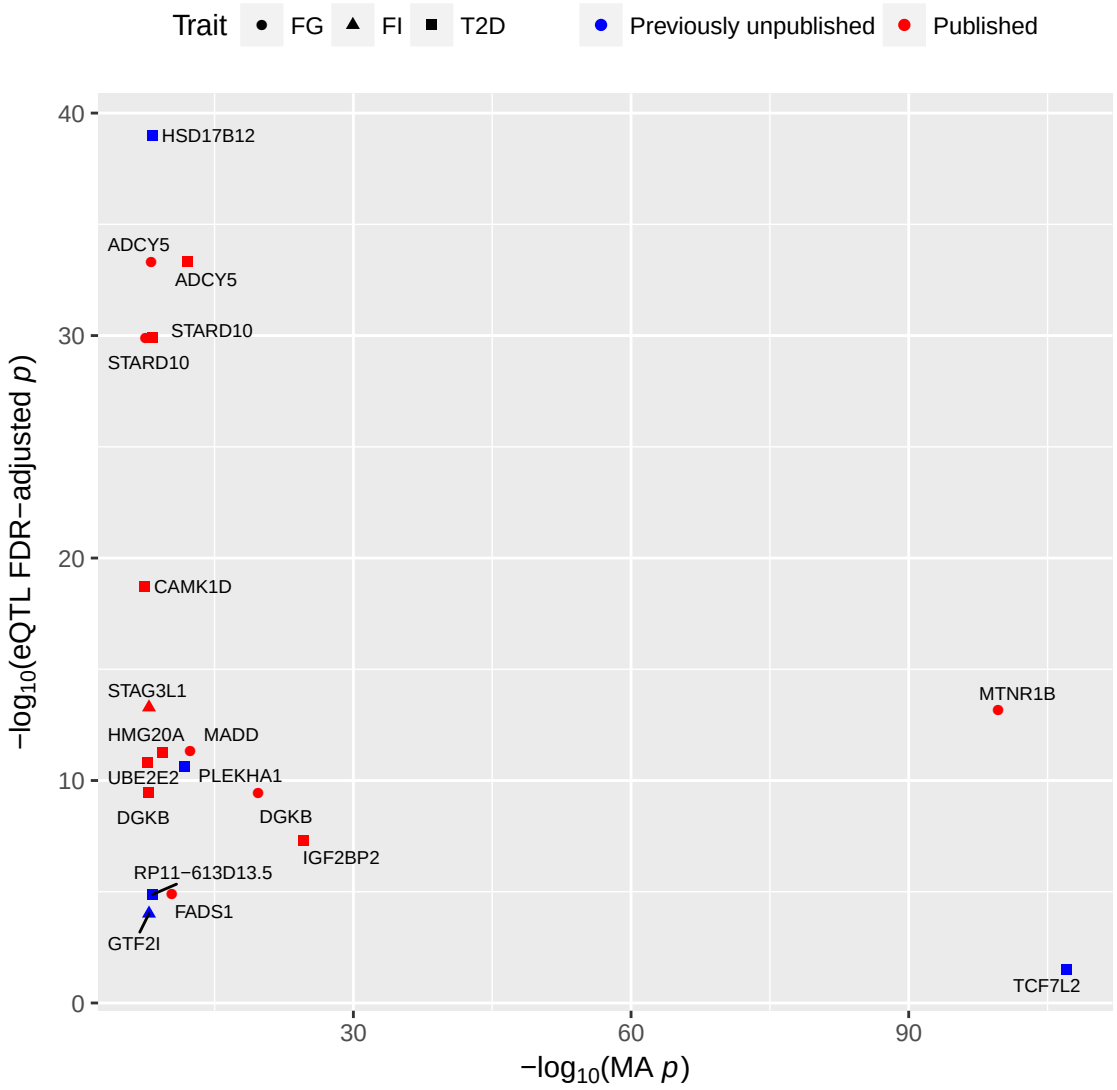


Figure 6.1: Comparison of MA and eQTL p-values for concordant signals. Plot showing $-\log_{10}$(FDR-adjusted p) from the eQTL analysis (y-axis) and corresponding MA (x-axis) for each significant SNP-gene pair in eQTL and MA concordance results. The shape of each point indicates the trait with which the SNP-gene pair is associated, and the colour indicates whether the eGene has been previously published as a potential effector transcript for T2D or glycaemic traits (red for published genes, blue for unpublished).

Table 6.3: eQTL and MA concordance results. All lead significant MA SNPs that tag significant lead eQTLs

SNP	**Chr:Pos**	**Trait**	**Trait** p	**RNA**	**eQTL** p
rs11708067	3:123065778	T2D	8.8e-13	*ADCY5*	1e-37
rs13094957	3:23457080	T2D	1.6e-08	*UBE2E2*	1.98e-14
rs4402960	3:185511687	T2D	2.7e-25	*IGF2BP2*	1.09e-10
rs2215383	7:15062983	T2D	1.4e-08	*DGKB*	5.4e-13
rs11257655	10:12307894	T2D	4e-08	*CAMK1D*	1.12e-22
rs2292626	10:124186714	T2D	1.8e-12	*PLEKHA1*	3.21e-14
rs7903146	10:114758349	T2D	9.3e-108	*TCF7L2*	0.000343
rs1061810	11:43877934	T2D	5.3e-09	*HSD17B12*	1.44e-43
rs11603334	11:72432985	T2D	5.3e-09	*STARD10*	3.19e-34
rs1061810	11:43877934	T2D	5.3e-09	*RP11-613D13.5*	5.08e-08
rs7178572	15:77747190	T2D	4.4e-10	*HMG20A*	6.78e-15
rs6963105	7:75097488	FI	1.2e-08	*STAG3L1*	5.12e-17
rs6963105	7:75097488	FI	1.2e-08	*GTF2I*	4.44e-07
rs11708067	3:123065778	FG	6.79e-09	*ADCY5*	1e-37
rs2191349	7:15064309	FG	1.98e-20	*DGKB*	5.63e-13
rs11603334	11:72432985	FG	2.92e-08	*STARD10*	3.19e-34
rs10830963	11:92708710	FG	2.2e-100	*MTNR1B*	7.12e-17
rs7944584	11:47336320	FG	4.38e-13	*MADD*	5.77e-15
rs174541	11:61565908	FG	4.08e-11	*FADS1*	4.94e-08

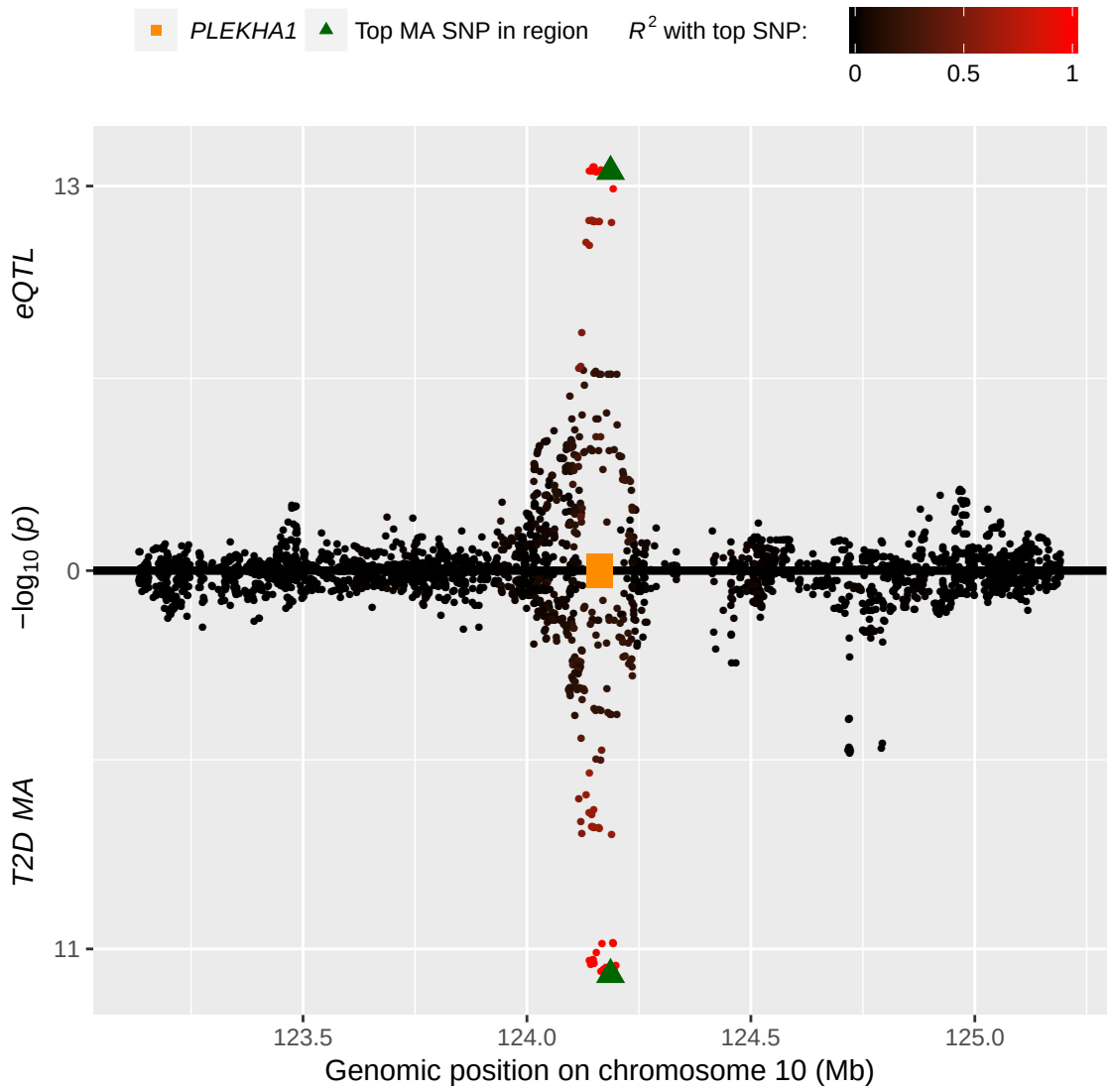


Figure 6.2: Concordance of rs2292626 effect on *PLEKHA1* expression and T2D. Miami plot showing all SNPs (points) within the *cis* region for *PLEKHA1* (x-axis). $-\log_{10} p$ values are shown for eQTL associations on the upward axis, and T2D MA associations on the downward axis. The lead variant, rs2292626, is shown as a triangle, with the red heat scale indicating LD with rs2292626.

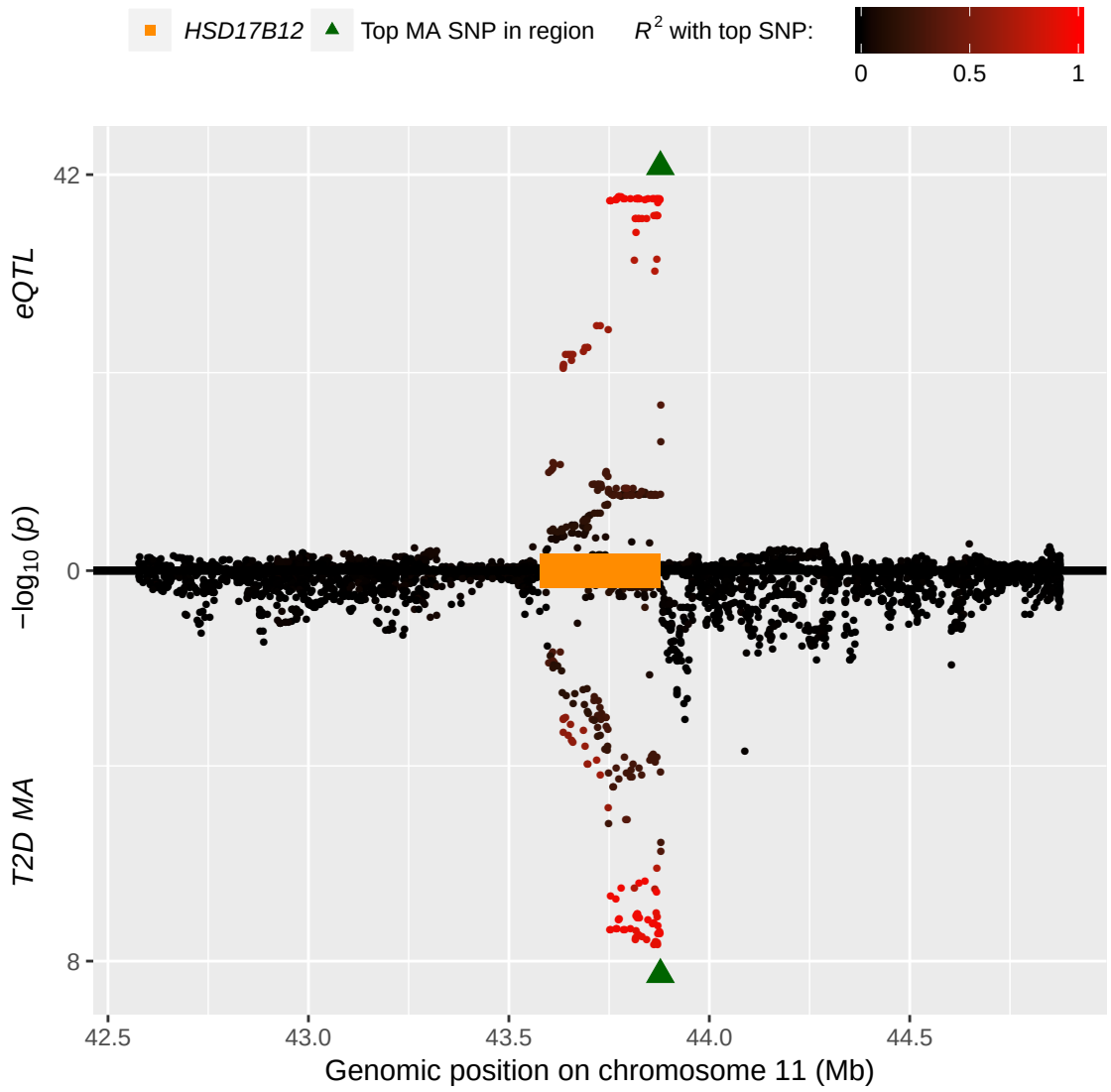


Figure 6.3: Concordance of rs1061810 effect on *HSD17B12* expression and T2D. Miami plot showing all SNPs (points) within the *cis* region for *HSD17B12* (x-axis). $-\log_{10} p$ values are shown for eQTL associations on the upward axis, and T2D MA associations on the downward axis. The lead variant, rs1061810, is shown as a triangle, with the red heat scale indicating LD with rs1061810.

$p = 5.3 \times 10^{-9}$, eqtl $p = 1.44 \times 10^{-43}$; Figure 6.3).

Amongst the 19 overall significant overlapping signals, one was an eQTL for a long noncoding RNA (lncRNA): the antisense lncRNA *RP11-613D13.5*. *RP11-613D13.5* will be discussed further in section 6.3.3.

The overall discrepancies in published studies and the data presented here are somewhat surprising, especially considering the published results presented in Table 6.1 all use islets that partially make up the InsPIRE collaboration data used here. These discrepancies are likely due primarily to differences in sample size and analysis pipelines. The largest study in Table 6.1 used islets from 191 individuals, while my analysis was based on eQTLs calculated from 426 individuals. Each of the five published studies in Table 6.1 uses a different method for identifying concordant signals (described in section 6.1), and each study based phenotype associations on different published phenotype datasets. Filtering pipelines and thresholds also vary between studies, with each study employing a different method of genotype filtering and imputation, RNA-seq mapping and filtering, and eQTL FDR calculation and cut-off.

Even the core analysis performed as part of the main InsPIRE collaboration analysis resulted in a different set of overlaps, which can largely be explained by differences in methodology. The core analysis for the main InsPIRE paper used exon-level eQTLs for the basis of their analysis, and also employed a stricter expression cut-off threshold than the one I used. The core InsPIRE analysis also required a probability of colocalisation greater than 80% based on results from *coloc*.[299,300] The identification of unpublished concordant MA and eQTL effects in this section — especially those with strong and distinct overlapping signals such as *PLEKHA1* and *HSD17B12* — suggests that results can differ significantly depending on the samples and methodology used.

Despite differences between my analysis and those previously published, the results presented here still replicated 11 loci for which concordant islet eQTL and

MA effects have been previously reported. This replication suggests that the most robust loci are detectable using a variety of methods. It is likely that as sample sizes increase, the reproducibility of signals (regardless of method) will also improve, such as was the case for differential expression (DE) analysis described in Chapter 4. These results suggest that in the meantime, it is important to assess a variety of analysis techniques and thresholds to make the most of the data at hand. In the following section, I have employed IVW 2SMR and S-PrediXcan testing to add a further level of analysis and leverage the information gained by the multi-variant gene expression models defined in Chapter 5. This complements the single-variant analysis presented here, and together with the single-variant analysis, allows for selection of robust eGenes that are relevant to T2D and glycaemic traits.

6.3.2 Two-sample Mendelian randomisation and S-PrediXcan

To further expand on loci identified with concordant MA and eQTL associations in section 6.3.1, two tests that rely only on association summary results were performed for each gene to identify those with genetically-regulated expression that may influence glycaemic traits and T2D risk: 2SMR IVW tests[166] and S-PrediXcan tests.[163] Both methods, unlike the single-variant overlaps assessed in section 6.3.1, combine the effects of multiple expression-associated variants for each gene and perform a single association test per gene (schematic diagram in Figure 6.4).

These gene-level tests reduce the multiple testing burden and leverage the joint multi-SNP gene expression models derived in Chapter 5, which were shown to be better predictors of gene expression than single eQTLs alone. Since these genetic models of gene expression are as good as or better than single eQTLs alone in nearly all cases (and therefore are generally better predictors of gene expression), genes identified using these methods should provide improved power to accurately

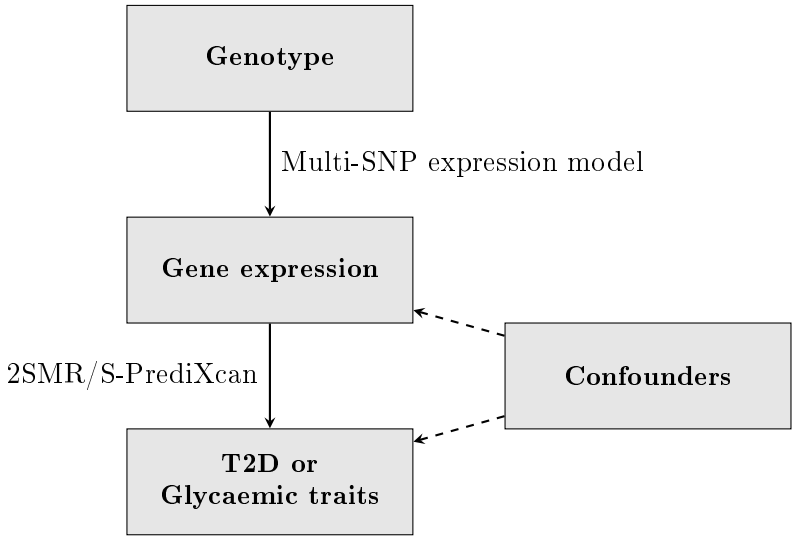


Figure 6.4: 2SMR and S-PrediXcan overview. For both 2SMR and S-PrediXcan analyses, the fundamental premise was to use multi-SNP models of gene expression derived in Chapter 5 to determine cases where the genetic component of gene expression is associated with T2D or glycaemic traits. For 2SMR analysis, MR-Egger intercept tests and explicit tests of SNP heterogeneity are used to assess the potential of confounding, while no direct tests are performed for S-PrediXcan models.

detect genes with expression associated with T2D risk and glycaemic traits, especially in cases where the expression models consist of moderate-effect SNPs that independently do not reach eQTL significance thresholds.[156] These analyses are thus intended to complement the single-variant associations, both to confirm the associations identified in section 6.3.1 and to identify further associations with modest SNP effects that were not identified in the single-variant overlap analysis.

For input in both 2SMR and S-PrediXcan analyses, multi-SNP expression models were derived using the same 2,033,629 input SNPs as for eQTL analysis in section 6.3.1. These models were derived using the same method as described in section 5.3.2, except that 500 iterations of `cv.glmnet` were used to select the optimal λ value for each gene, and 100 iterations were used to select the optimal λ in the final ridge regression models derived with the final filtered set of SNPs for each gene. These increases from 200 and 25 iterations respectively were to improve the robustness and reproducibility of models.

A total of 11,801 genes had a final model with at least one SNP included. There were a median 6 SNPs in the final models, with a maximum of 89 SNPs in a single model. There was a median improvement factor over the top eQTL per gene of $R^2_{\text{final}}/R^2_{\text{top eQTL}} = 2.3$, with a maximum improvement factor of 22.5 (Figure 6.5).

2SMR

Using the IVW method for 2SMR (section 6.2.3), the 11,801 genes with at least one SNP in the final expression model were tested for potential causal mediating effects on each of three traits: T2D, FG, and FI.

In addition to the IVW tests of association, three filtering methods were employed for each gene to assess the validity of each model's instruments and to filter out cases with evidence of unobserved pleiotropy: SNP heterogeneity tests (Cochran Q test), MR-Egger intercept tests, and expression model significance tests. The Cochran Q tests assess whether the SNPs in each gene's expression model have true effects on the outcome (T2D or glycaemic traits) that are proportional to the SNP effects on expression. A significant heterogeneity test suggests that SNPs do not have consistent effects on both expression and the outcome, indicating that at least one SNP may have a pleiotropic effect (thus violating IV assumptions). Similarly, a significant MR-Egger intercept test suggests that there is likely pleiotropy that has not been accounted for, and that the IVW test is therefore not suitable. Finally, to ensure that the genetic models of gene expression were strongly associated with expression, a model fit test was also assessed for each gene based on the F statistic from the model R^2 calculation described in section 5.2.4.

After filtering based on these three tests (FDR < 0.05), there were 100 significant gene associations that each had: i) a significant model fit F test, ii) a non-significant Cochran Q test for SNP heterogeneity, and iii) a non-significant MR-Egger regression intercept. All associations are shown in Figure 6.6 and

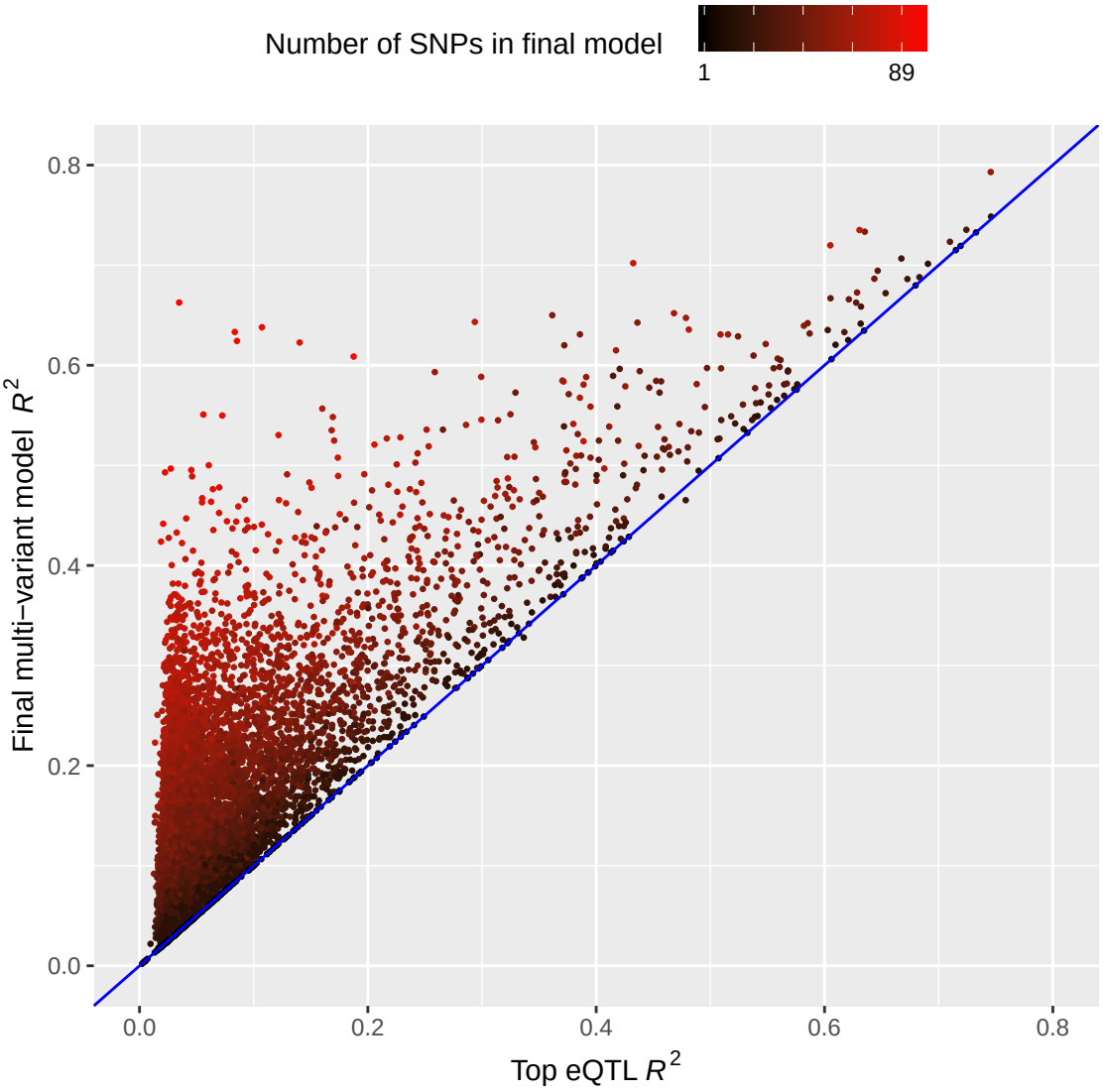


Figure 6.5: Model R^2 comparison between top eQTL and the final filtered penalised regression models. Scatterplot showing the model R^2 per gene for the top eQTL alone (x-axis) against the model R^2 for each gene's corresponding final ridge regression model (y-axis). Each point represents a gene, with the blue line having a slope of 1 and representing equality between the axes. The colour of each point indicates the number of SNPs in the final model (log scale).

summarised in Table A.7. Only six of these associations were for eGenes that have been previously published with concordant MA and eQTL effects in islets: *KLHL42* for T2D, *ACP2* for FG, and both *ADCY5* and *STARD10* for T2D and FG.

Five of the one hundred significant genes were also identified in the single-variant overlap analysis, all of which were amongst the top 15 most significant genes identified in 2SMR analysis. This indicates that there is likely a single strong concordant variant effect (or LD cluster) for those five genes. Although the majority of significant effects were based on models with five or less SNPs (58%), two significant genes — *PDE8B* and *ECE2* — had more than 20 SNPs in their models (24 and 22 respectively). Despite these genes having non-significant MR-Egger intercepts and non-significant tests for heterogeneity between SNP effects on gene expression and the respective traits, dissection and interpretation of these signals is complex (Figures 6.7 and 6.8 respectively).

For *PDE8B*, the overall IVW test statistic suggests that increased expression is associated with increased FG, and while 18/24 SNPs in the model are consistent with this effect (alternate allele has the same direction of effect for both expression and FG associations), the other SNPs in the model have opposite directions of effect. Therefore, even though the aggregate test statistic indicates that increased expression is associated with increased FG, the actual genetic architecture of *PDE8B* expression and FG associations is complex. A similar pattern is observable for *ECE2*, where only 17/22 modelled SNPs have consistent direction of effect (increased expression associated with T2D risk). In the cases of *PDE8B* and *ECE2*, the SNPs exhibiting opposite directions of effect have relatively small MA effect sizes compared to the SNPs with consistent directions of effect, thus they contribute minimally to the overall IVW test statistic estimate (section 6.2.3). This suggests that these SNPs, although associated with gene expression, do not contribute significantly to the respective trait outcomes, and although the

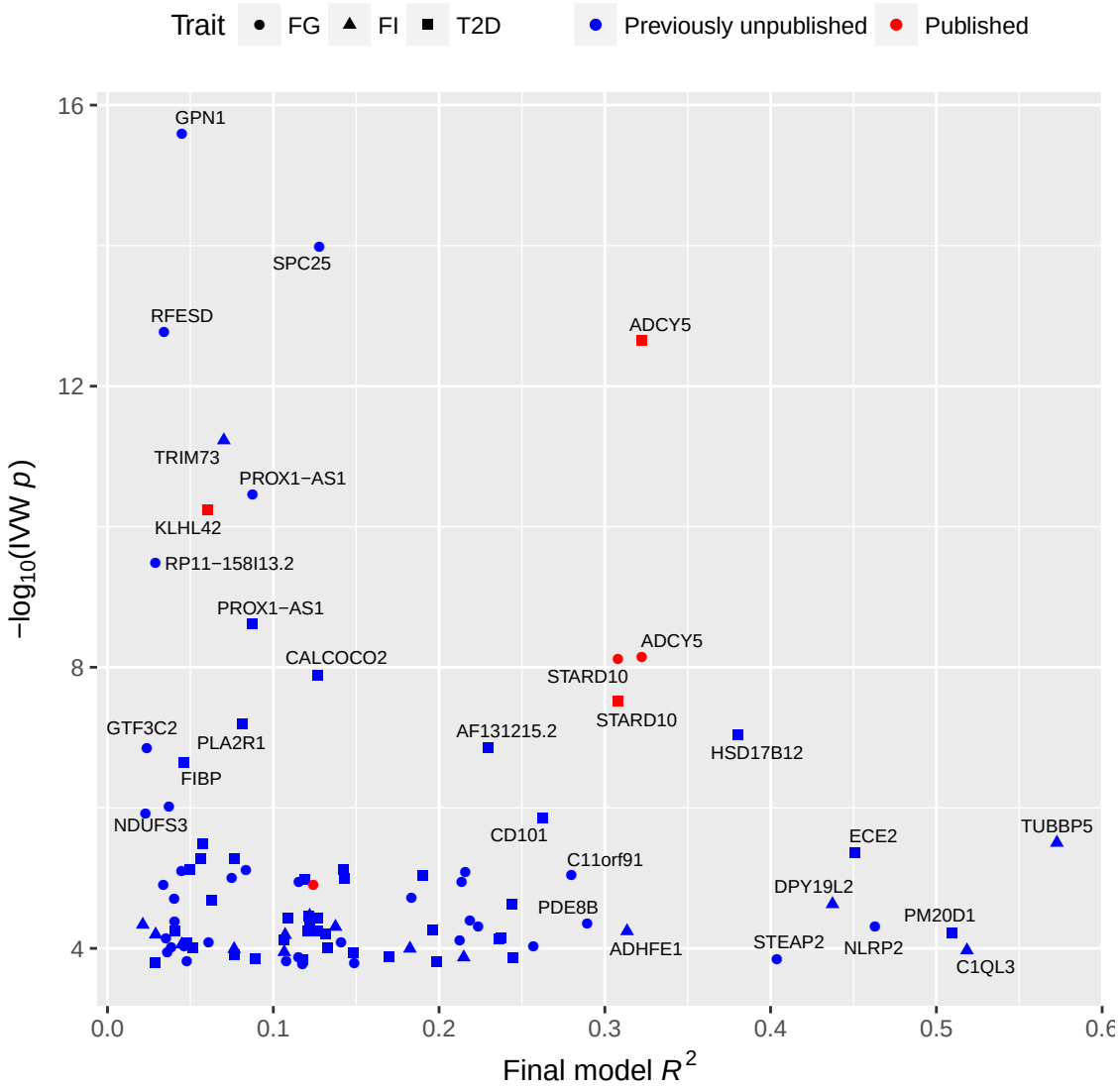


Figure 6.6: Results of 2SMR for gene expression effects on T2D, FG, and FI. Plot showing the final model R^2 for each gene (x-axis) and the $-\log_{10} p$ value for the IVW test statistic for each significant gene (y-axis). The shape of points shows the trait for which the association was significant, with the colour indicating whether the gene has been previously published as a potential islet effector transcript for T2D or glycaemic traits (red for published genes, blue for unpublished).

heterogeneity of SNP effects for these genes was not significant, there is evidently some heterogeneity at these loci.

There were 15 lncRNAs amongst the 100 genes identified using 2SMR: 5 antisense, 7 long intergenic noncoding RNAs, 2 processed transcripts, and 1 sense intronic gene. The proportion of lncRNAs that were significant using 2SMR (15.0%) was not significantly different from the proportion of lncRNAs in the full test set (13.3%; $p = 0.74$). The most robust of these lncRNA gene effects (those detected using multiple methods) will be discussed further in section 6.3.3.

S-PrediXcan

To further complement results from single-variant concordance analysis and 2SMR, and to assess robustness of associations between methods, S-PrediXcan was also implemented on all genes with at least one SNP in the final ridge regression model. After filtering to keep only genes with significant expression model R^2 (F-test, FDR < 0.05), 52 genes were significant across traits, 17 of which have been previously published with overlapping significant MA and islet eQTL effects (Figure 6.9; full list in Table A.8). There was also a high overlap between S-PrediXcan results and single-variant results presented in section 6.3.1, with 17/52 (32.7%) associations overlapping, compared to 5.0% of 2SMR results overlapping the single-variant analysis. The only genes that were detected in the single-variant overlap analysis but not using S-PrediXcan were *CAMK1D* and *GTF2I*, suggesting that in addition to identifying genes with modest concordant effects at multiple variants, S-PrediXcan also effectively captured the majority of genes with a single strong overlapping eQTL and MA peak. In the cases of *CAMK1D* and *GTF2I*, the multi-variant models of gene expression each contained multiple SNPs (19 and 3 respectively), despite there being only one clear significant MA association, suggesting that in addition to the single concordant variant observed at each locus in section 6.3.1, there are other variants contributing to the gene expression models

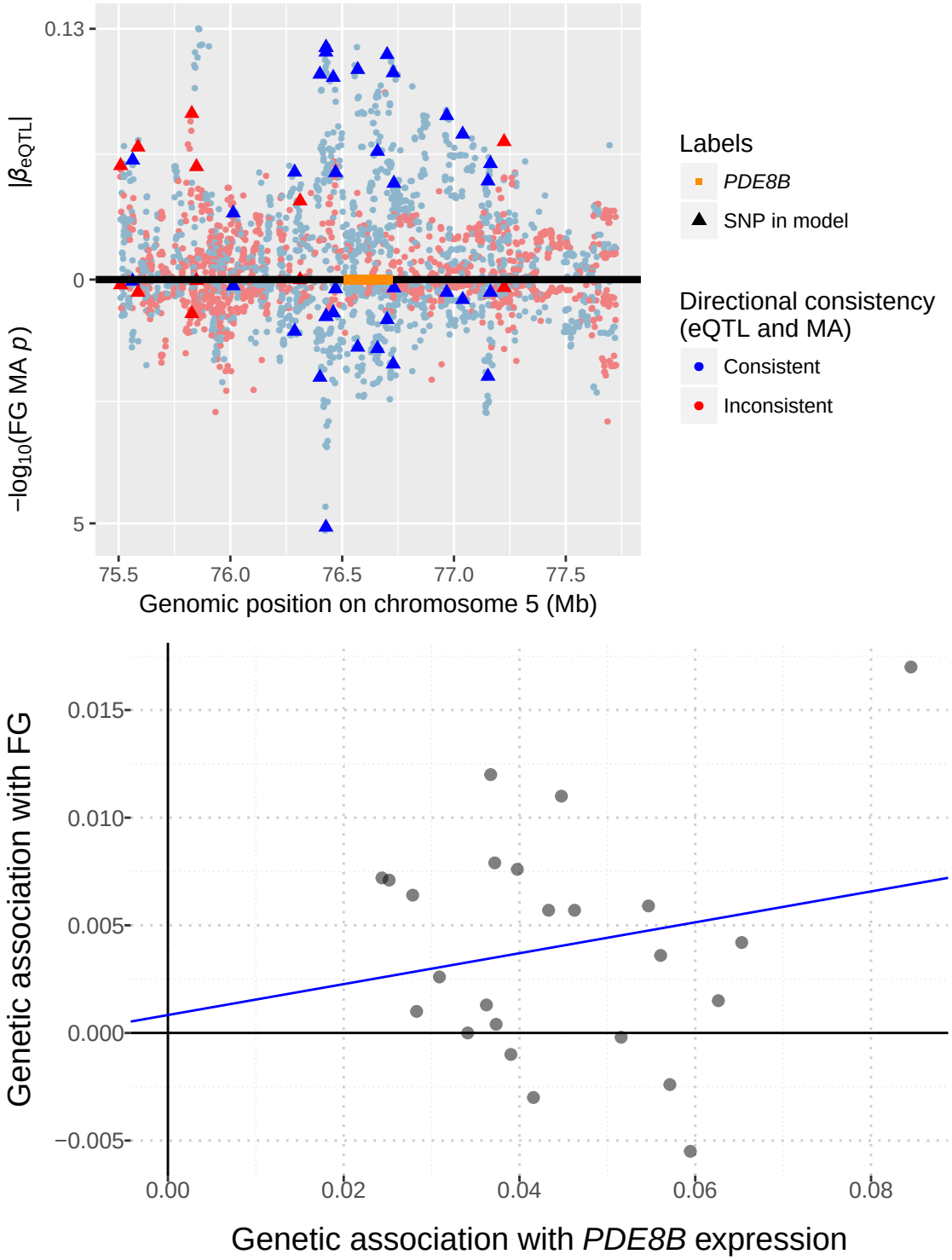


Figure 6.7: Comparison of eQTL and FG SNP effects at *PDE8B*. The top plot shows the absolute eQTL effect sizes (y-axis, upward) and the corresponding MA $-\log_{10} p$ values (y-axis, downward) for all *PDE8B cis* SNPs. Triangles indicate SNPs that were selected for the final filtered ridge regression model. Blue points represent SNPs that contributed to the IVW test statistic in a direction consistent with the total test statistic value. The bottom plot shows the MR-Egger estimate and intercept (blue line). The MR-Egger intercept did not significantly differ from zero.

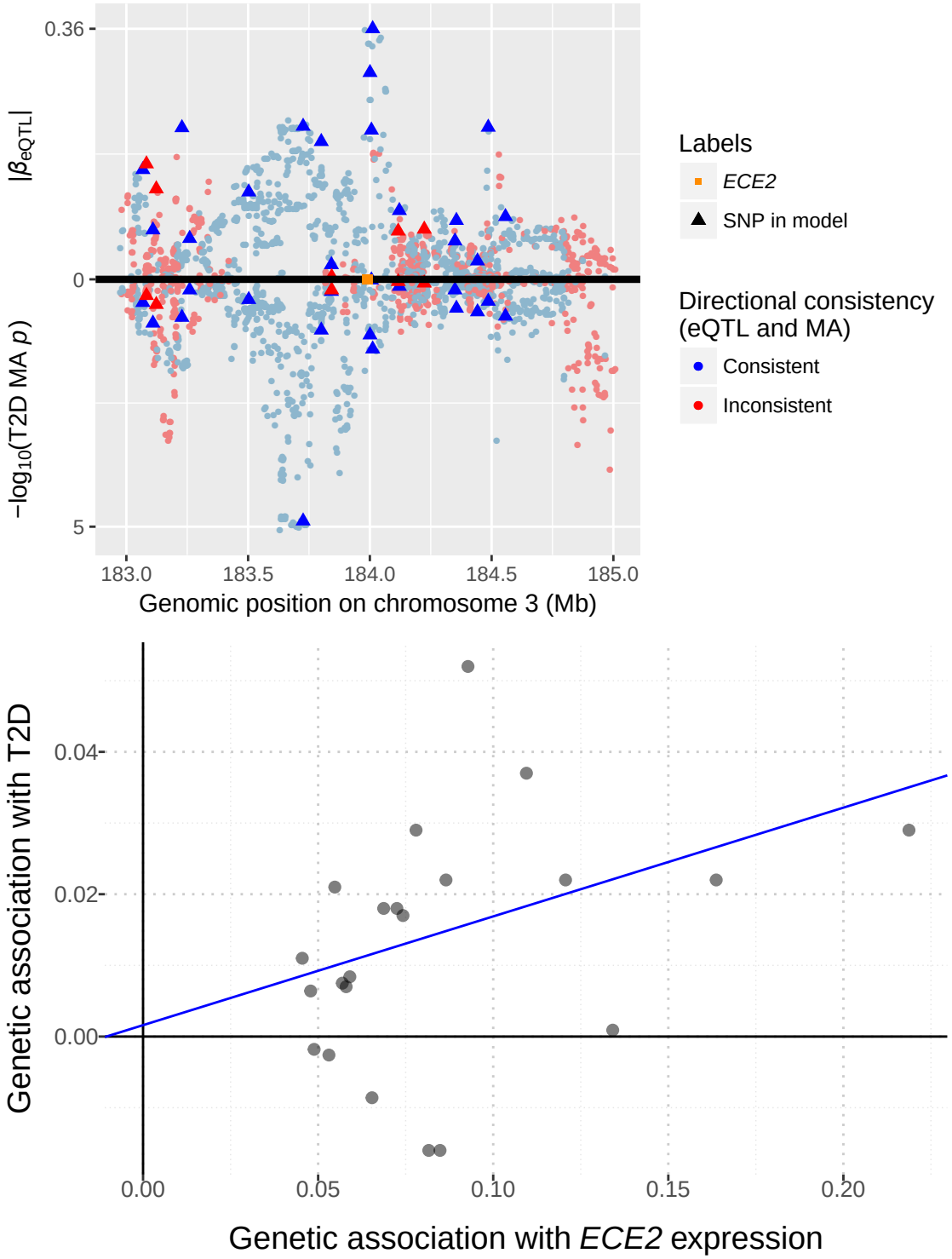


Figure 6.8: Comparison of eQTL and T2D SNP effects at *ECE2*. The top plot shows the absolute eQTL effect sizes (y-axis, upward) and the corresponding MA $-\log_{10} p$ values (y-axis, downward) for all *ECE2 cis* SNPs. Triangles indicate SNPs that were selected for the final filtered ridge regression model. Blue points represent SNPs that contributed to the IVW test statistic in a direction consistent with the total test statistic value. The bottom plot shows the MR-Egger estimate and intercept (blue line). The MR-Egger intercept did not significantly differ from zero.

that are not consistently associated with the relevant phenotypes.

Six of the fifty-two genes identified with S-PrediXcan were lncRNAs (11.5%), and this proportion was not significantly different from the proportion of lncRNAs in the whole test set ($p = 0.86$).

6.3.3 Integration with differential expression data

Since the islet T2D DE results presented in Chapter 4 have not been tested for causal influence on T2D risk, I next took the most robust gene associations from the above analyses, and intersected them with DE results from Chapter 4 to identify genes that are both i) differentially expressed between islets of T2D and normoglycaemic donors, and ii) have expression regulated by T2D- or glycaemic-trait-associated variants. Genes that are regulated by T2D-relevant variants but not differentially expressed in T2D donors may still influence diabetes risk through small, additive effects; similarly, those with DE but no evidence of genetic regulation by T2D-relevant variants may still alter T2D risk with expression regulated by another mechanism. However, it is hypothesised that the genes with robust evidence of regulation in islets by T2D-relevant variants that also exhibit significant DE in islets of T2D donors are more likely to have a causal impact on T2D risk compared to those genes exhibiting only one of these characteristics.

Given the overall differences in results between single-variant analysis, 2SMR, and S-PrediXcan, only those genes that were identified using at least 2/3 of the methods were examined further. This 'replication' requirement was used as a simple measure of robustness of results. Thirty-three genes were identified using at least two of the methods, with three genes identified using all methods: *ADCY5*, *HSD17B12*, and *STARD10* (Table 6.4). Gene ontology (GO) term enrichment analysis was performed for this set of 33 genes (as described in section 4.2.6) to assess the relevance of these genes to T2D and glucose homeostasis. Although no

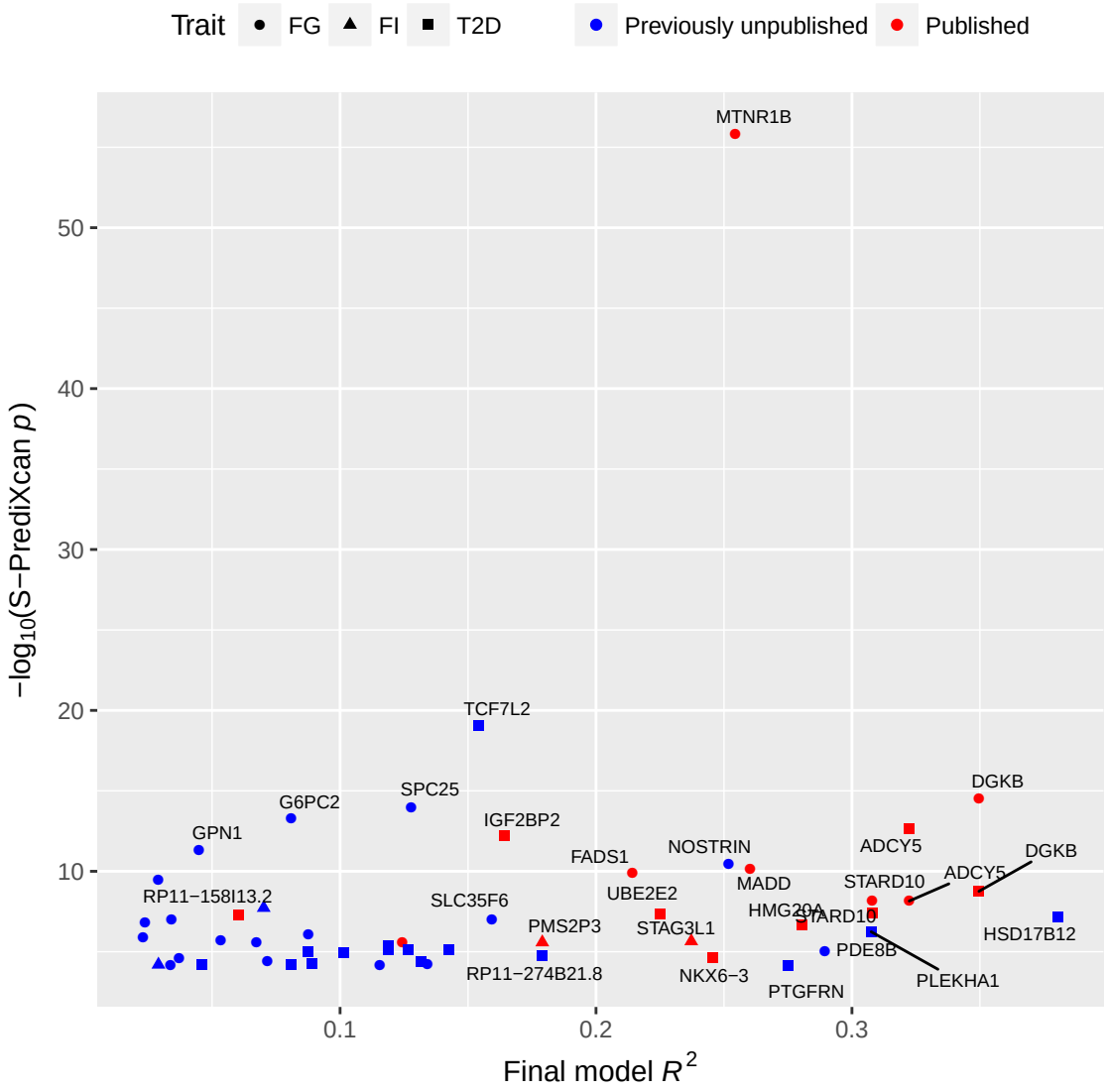


Figure 6.9: Results of S-PrediXcan for gene expression effects on T2D, FG, and FI. Plot showing the final model R^2 for each gene (x-axis) and the $-\log_{10} p$ value for the S-PrediXcan test statistic for each significant gene (y-axis). The shape of points shows the trait for which the association was significant, with the colour indicating whether the gene has been previously published as a potential islet effector transcript for T2D or glycaemic traits (red for published genes, blue for unpublished).

GO term was significantly over-represented after correcting for tests of 1,186 total terms, the terms 'regulation of hormone levels', 'regulation of insulin secretion', and 'glucose homeostasis' were amongst the six terms with the smallest nominal over-representation p-values. This supports the hypothesis that at least some of these robustly detected genes are functionally relevant to glucose homeostasis.

Of the 33 genes that replicated between at least two analysis methods above, 4 were differentially expressed in islets based on T2D status (section 4.3.2): *HSD17B12*, *RP11-613D13.5*, *ACP2*, and *PDE8B*. The potential functional relevance of these genes is described below. In addition, given the focus on lncRNAs throughout this thesis, the potential relevance of antisense lncRNA *PROX1-AS1*, which was significant in both 2SMR and S-PrediXcan analyses, is also described below. One other lncRNA — processed transcript *RP11-158I13.2* — was also identified with both 2SMR and S-PrediXcan, but has not been previously studied and has no known or proposed function, making interpretation difficult.

HSD17B12

It was observed in section 6.3.1 that both the eQTL analysis and the T2D MA had the same significant lead variant at the *HSD17B12* (hydroxysteroid 17-beta dehydrogenase 12 gene) region (rs1061810). The T2D risk allele reported by Scott *et al.* at rs1061810 (A allele) was associated with decreased expression of *HSD17B12*, and results from 2SMR and S-PrediXcan indicate that general reduced expression of *HSD17B12* is associated with increased T2D risk. The same risk allele reported by Scott *et al.* was also significantly associated with reduced expression of *HSD17B12* in all 48 tissues for which GTEx version 7 eQTL results have been published (`www.gtexportal.org`). In another study, Das and Sharma used quantitative reverse transcription polymerase chain reaction to show that non-diabetic rs1061810 risk-allele carriers had significantly lower *HSD17B12* adipose expression than non-risk-allele carriers.[312] HSD17B12 enzymes encoded by

Table 6.4: All genes identified as candidate mediators of SNP effects on T2D, FG, or FI using at least two methods. All genes that were identified as significant in at least two of the three analyses (single-variant concordance, 2SMR, and S-PrediXcan), indicating robustness of detection.

Gene	Concordance	2SMR	S-PrediXcan
ADCY5	X	X	X
HSD17B12	X	X	X
STARD10	X	X	X
DGKB	X		X
FADS1	X		X
HMG20A	X		X
IGF2BP2	X		X
MADD	X		X
MTNR1B	X		X
PLEKHA1	X		X
RP11-613D13.5	X		X
STAG3L1	X		X
TCF7L2	X		X
UBE2E2	X		X
ACP2		X	X
C1QL2		X	X
CALCOCO2		X	X
CENPA		X	X
FIBP		X	X
GPN1		X	X
GTF3C2		X	X
KLHL42		X	X
NDUFS3		X	X
PDE8B		X	X
PLA2R1		X	X
PROX1-AS1		X	X
RFESD		X	X
RP11-158I13.2		X	X
SF3B2		X	X
SPC25		X	X
TH		X	X
TRIM73		X	X
ZNF84		X	X

HSD17B12 can catalyse the conversion between 17-ketosteroids and 17β-hydroxysteroids (for example from androstenedione to testosterone), and have a known role in fatty acid synthesis and lipid metabolism.[313,314]

In a recent study by Ottosson-Laakso *et al.*, *HSD17B12* had significantly up-regulated expression in normoglycaemic islets after glucose stimulation ($\log_2$ fold-change $= 0.22$, $p_{\text{adjusted}} = 0.031$),[101] which is consistent with my finding that *HSD17B12* expression was significantly higher in islets from T2D donors compared to islets of normoglycaemic donors ($\log_2$ fold-change $= 0.29$, $p_{\text{adjust}} = 0.0055$). It was also previously observed that *Hsd17b12* is up-regulated in white adipose tissue of mice after treatment with either FGF21 or PPARγ, both of which promote glucose uptake and insulin sensitivity.[315,316]

The up-regulation of *HSD17B12* observed by Ottosson-Laakso *et al.* after acute glucose and the higher general expression of *HSD17B12* in islets of T2D donors support the hypothesis that it is involved in a mechanism triggered by elevated glucose. Given that the risk allele at rs1061810 is universally associated with decreased *HSD17B12* expression throughout many human tissues, the impairment of glucose-stimulated *HSD17B12* up-regulation in risk-allele carriers at rs1061810 may therefore be the mechanism through which the rs1061810 association with T2D is manifested. It is therefore possible that *HSD17B12* transcription, which is in part triggered by elevated glucose, is important in glucose homeostasis, and that its function is impaired for rs1061810 risk allele carriers through reduced expression. Even in a sub-sample of islets from 29 InsPIRE donors with T2D, the rs1061810 risk allele was still significantly associated with lower *HSD17B12* expression ($p = 4.80 \times 10^{-5}$; $n_{\text{CC}} = 20$, $n_{\text{AC}} = 7$, $n_{\text{AA}} = 2$), despite the significantly higher overall *HSD17B12* expression observed in the islets of T2D donors (Figure 6.10). Experimental work will be necessary to fully understand the functional relevance of *HSD17B12*.

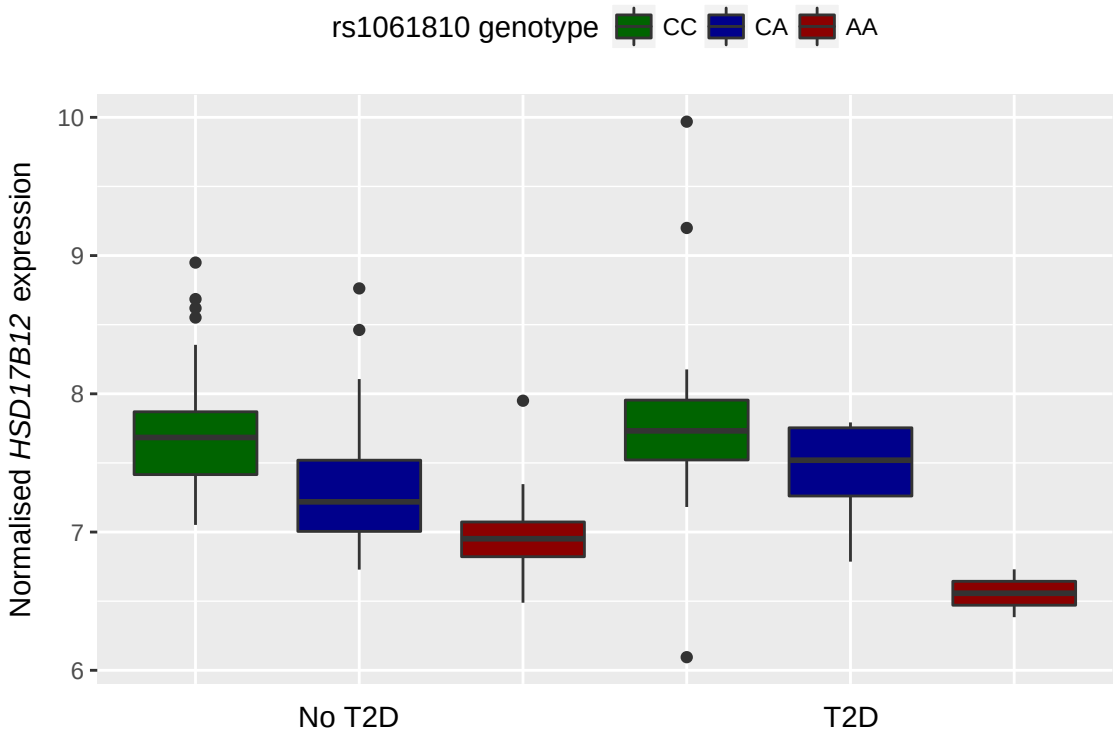


Figure 6.10: Expression of *HSD17B12* in islets stratified by T2D status and rs1061810 genotype. Distribution of normalised islet expression level of *HSD17B12* (y-axis) stratified by T2D status and genotype at rs1061810. Overall *HSD17B12* expression is significantly higher in islets from donors with T2D ($p_{\text{adjust}} = 0.0055$), and the T2D risk A allele at rs1061810 is associated with significant reduced expression using the whole sample ($p = 1.44 \times 10^{-43}$) and within T2D donors (4.80×10^{-5}).

RP11-613D13.5

The antisense lncRNA *RP11-613D13.5* overlaps *HSD17B12* and is transcribed from the opposite strand. The two genes also exhibit expression correlation ($r = 0.30$, $p = 5.25 \times 10^{-10}$). *RP11-613D13.5* was not assessed by Ottosson-Laakso *et al.* However, similar to *HSD17B12*, it was differentially expressed in islets of non-diabetic compared to T2D donors ($\log_2$ fold-change $= 0.33$, $p_{\text{adjusted}} = 0.0071$), rs1061810 was in high LD ($R^2 = 0.95$) with the lead eQTL variant for *RP11-613D13.5*, and results for each of the three methods in this chapter were directionally consistent (lower expression was associated with increased T2D risk). The T2D risk allele of rs1061810 was also significantly associated with reduced *RP11-613D13.5* expression in all 18 GTEx tissues for which the SNP-gene pair was significant. *HSD17B12* and *RP11-613D13.5* may therefore be involved in the same functional process. Given that both *RP11-613D13.5* and *HSD17B12* have nearly identical genetic regulatory signatures, knockdown experiments will be required to understand their transcriptional relationship.

ACP2

ACP2 (acid phosphatase 2, lysosomal gene) has been previously identified by van de Bunt *et al.* as having a lead eQTL variant in islets that overlaps a lead GWAS association. van de Bunt *et al.* hypothesised that the role of lysosomes in degradation of aging insulin granules may be a link between this gene and T2D,[100] although direct links between *ACP2* function and T2D have not been established. *ACP2* was identified as a potential mediator of FG-associated variants using 2SMR and S-PrediXcan (Figure 6.11), and DE analysis in Chapter 4 showed significantly higher *ACP2* expression in islets from donors with T2D ($\log_2$ fold-change $= 0.17$, $p_{\text{adjusted}} = 0.0039$). This is consistent with van de Bunt *et al.*, who found that *ACP2* expression was associated with increased fasting glucose and fasting proinsulin.[100]

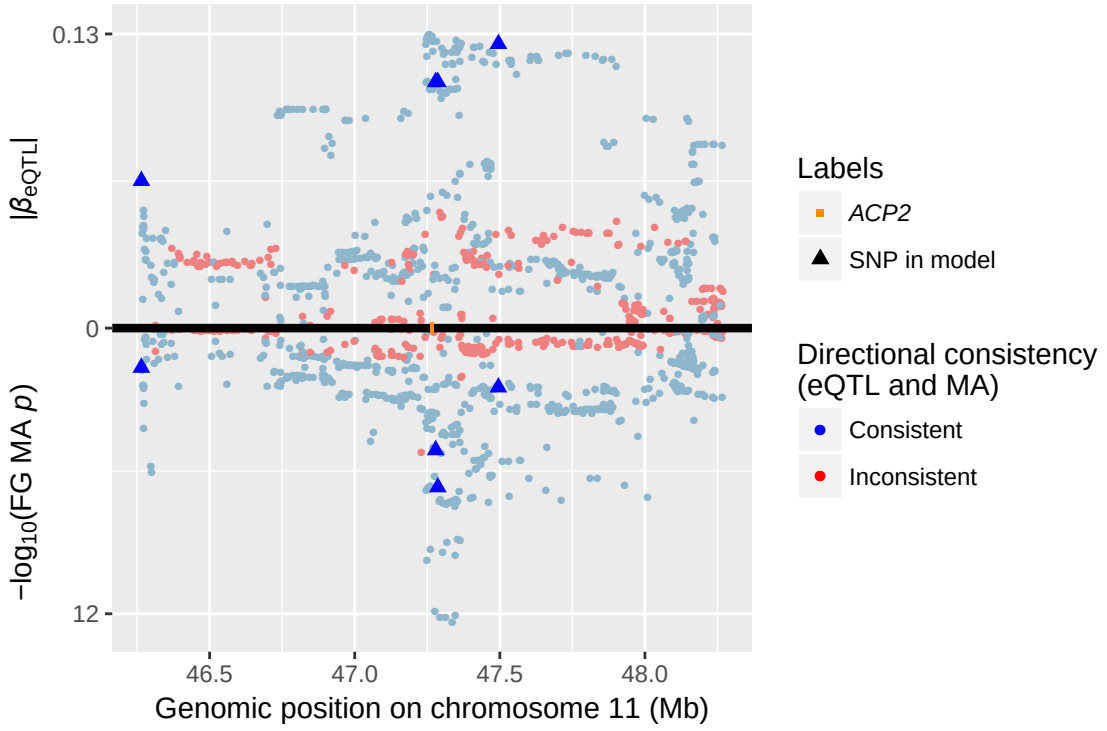


Figure 6.11: Comparison of eQTL and FG SNP effects at *ACP2*. The absolute eQTL effects (y-axis, upward) and the corresponding FG MA $-\log_{10} p$ values (y-axis, downward) for all *cis* SNPs to *ACP2*. Triangles indicate SNPs that were selected for the final filtered ridge regression model. Blue points represent SNPs that contributed to the IVW test statistic in a direction consistent with the total test statistic value. No SNPs selected for this multi-variant model were significant in the FG MA, but all four modelled SNPs had directionally consistent effects on both FG and *ACP2* expression.

PDE8B

The final candidate effector gene that was detected using more than one approach, and was also differentially expressed based on T2D status, was *PDE8B* (phosphodiesterase 8B gene), which has not been previously published as a potential T2D or FG effector gene. *PDE8B* encodes a cyclic nucleotide phosphodiesterase protein, which catalyzes hydrolosis of the second messenger cyclic 3'5'AMP.[317] No single *cis* SNP for *PDE8B* was genome-wide significant for FG; however, both 2SMR and S-PrediXcan found that the majority of variants selected for the multi-variant *PDE8B* expression model (18/24) had consistent corresponding effects in the FG MA, and that for the six SNPs with inconsistent directions of effect, the MA standard errors were large, thus giving the variants relatively low weights in the gene's overall test (Figure 6.7). Although studies have observed significant changes in glucose-dependent insulin secretion in rat insulinoma derived cells after *PDE8B* knockdown,[318,319] as well as changes in cyclic 3'5'AMP oscillations and loss of normal insulin secretion pulsatile pattern in insulin secreting mouse MIN6 cells when *PDE8B* was suppressed,[320] *PDE8B* function has not been experimentally validated in human cells.

PROX1-AS1

Although not differentially expressed by T2D status in islets, *PROX1-AS1* was detected as a potential T2D effector gene with both 2SMR and S-PrediXcan, and has been followed up given its status as a lncRNA. *PROX1-AS1* has been shown to harbour variants associated with osteonecrosis in the treatment of acute lymphoblastic leukemia in children,[321] but other than a nearby genome-wide significant variant associated with T2D (tagged by rs340874 in this analysis[47]), there have been no studies directly linking this lncRNA to T2D or glycaemic traits.

PROX1-AS1 is modestly correlated with expression of its paired sense gene, *PROX1* (prospero homeobox 1 gene; $r = 0.39$; Figure 6.12). *PROX1* knockdown

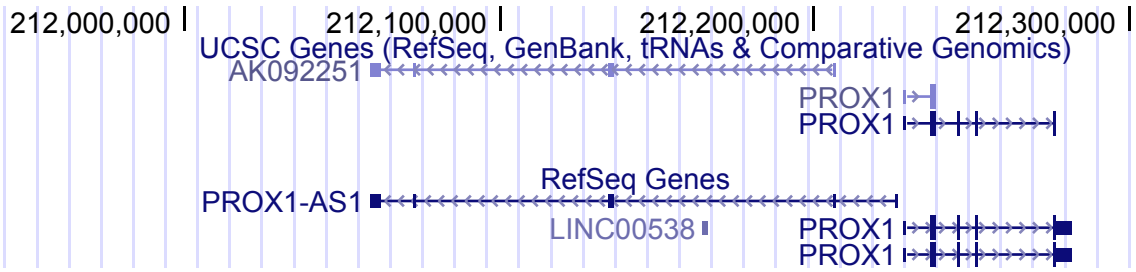


Figure 6.12: *PROX1-AS1* and proximal genes. Genomic location of *PROX1-AS1* and its proximal genes, including *PROX1*. Generated by the Human Genome Browser[270]

in INS-1E cells resulted in a nearly 2-fold decrease in glucose stimulated insulin secretion, and in the same study, it was shown that glucose stimulation (without knockdown) resulted in reduced messenger RNA levels, and increased protein levels.[322] Although *PROX1* was not differentially expressed in islets in Chapter 4, Ottosson-Laasko *et al.* found that *PROX1* had significantly lower expression in islets from donors with abnormal glucose tolerance compared to normoglycaemic islets, while also reporting no significant change in *PROX1* expression after glucose stimulation.[101] Given the modest correlation between *PROX1-AS1* and *PROX1*, the identification of *PROX1-AS1* as a potential effector transcript of T2D-associated variants using both 2SMR and S-PrediXcan (with no significant associations for *PROX1*; Figure 6.13), the clear functional relevance of *PROX1* in glucose-stimulated insulin secretion in rat INS-1E cells, and the fact that *PROX1* is differentially expressed between human islets with normal and abnormal glucose tolerance, *PROX1-AS1* is a candidate lncRNA regulator of *PROX1*, with expression dependent partially on *cis* genetic regulation. As is the case for other candidate genes identified, functional experiments, such as knocking down *PROX1-AS1* and assesssing both RNA levels and glucose-stimulated insulin response, would help classifty the functional relevance of this lncRNA.

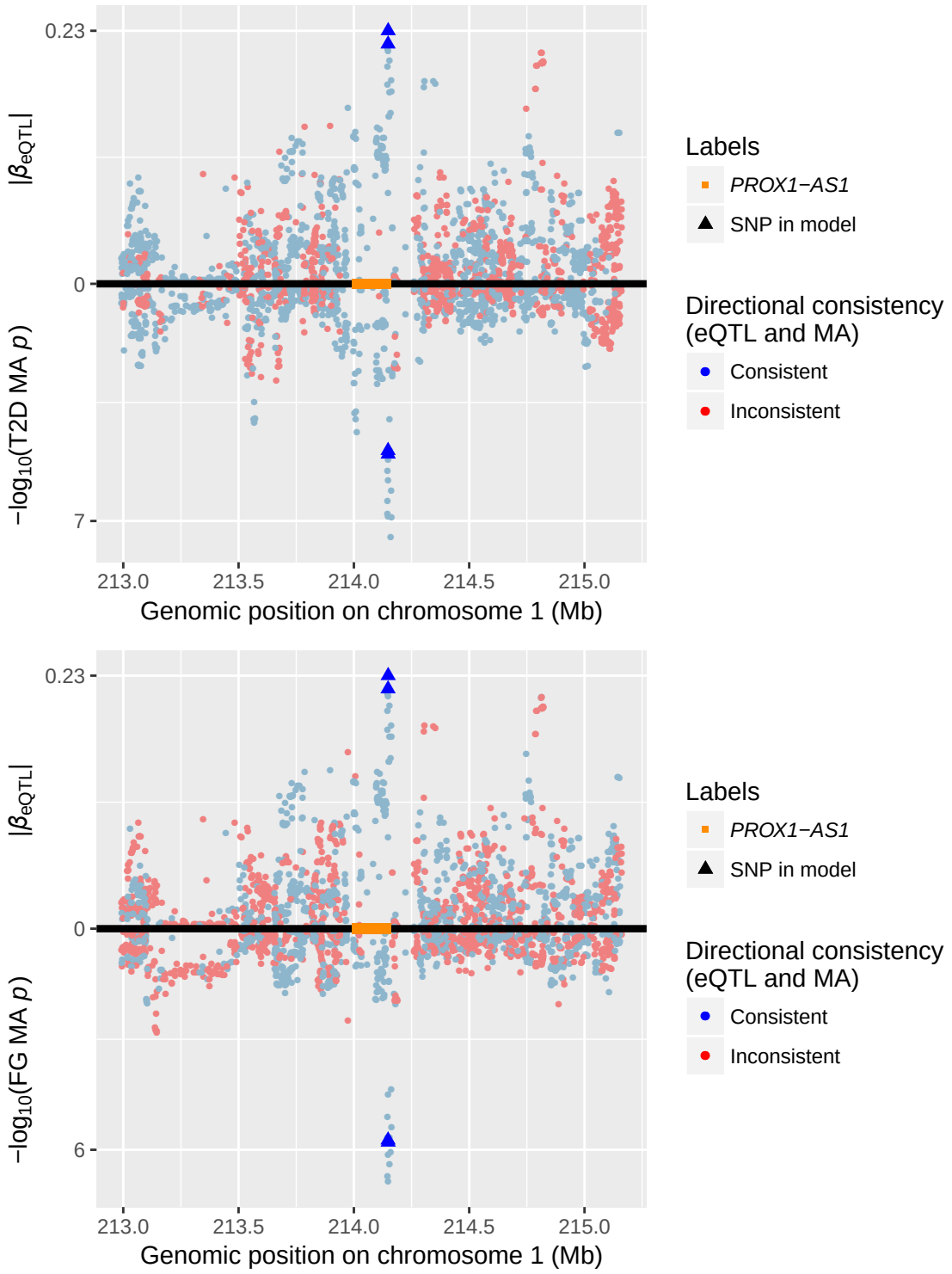


Figure 6.13: Comparison of eQTL and T2D/FG SNP effects at *PROX1-AS1*. The top plot shows the absolute eQTL effects (y-axis, upward) and the corresponding T2D MA $-\log_{10} p$ values (y-axis, downward) for all *cis* SNPs to *PROX1-AS1*, and the bottom plot shows corresponding data for the FG MA. Triangles indicate SNPs that were selected for the final filtered ridge regression model. Blue points represent SNPs that contributed to the IVW test statistic in a direction consistent with the total test statistic value.

CTB-12O2.1

The lncRNA *CTB-12O2.1* was identified in Chapter 4 as having a potential expression regulatory effect, and since no causal interpretation was possible in Chapter 4 based on expression data alone, it was queried in the results of this chapter. *CTB-12O2.1* did not have any significant islet eQTLs (minimum adjusted $p = 0.082$ for rs999556), and as a result, the multi-variant gene expression modelling process produced an intercept-only model as the best fit. This suggests that with the current sample set and variants used for analyses, there is little evidence for any genetic associations with *CTB-12O2.1* expression. This may change in future studies with larger sample sizes, increased sequencing resolution, or techniques that are further refined, but in this case, little information can be gained about this lncRNA using genetic data.

6.4 Discussion

This chapter has demonstrated how genetic models of gene expression (both single eQTLs and multi-variant models) can be used along with high-powered MAs to prioritise genes with expression that may impact T2D pathogenesis and glycaemic traits. The main results of this chapter were as follows:

- Simple eQTL and MA concordance identified 16 unique candidate effector eGenes in islets, 5 of which have not been previously published.
- 2SMR and S-PrediXcan analyses were able to extend the list of potential effector genes by using multi-variant expression models and leveraging the potential of small joint contributing effects.
- Thirty-three genes were significant in at least 2/3 analyses performed (single-variant concordance, 2SMR, and S-PrediXcan) and were nominally enriched for hormone- and insulin-associated GO terms.

- *HSD17B12* is significantly upregulated after glucose stimulation and has higher expression in islets of T2D donors compared to normoglycaemic islets. The T2D risk allele for rs1061810 is also associated with significantly reduced expression of *HSD17B12* in islets and 48 GTEx tissues, together suggesting potential functional relevance to glucose response. Its correlated antisense lncRNA *RP11-613D13.5* has nearly identical genetic regulatory architecture, suggesting that *RP11-613D13.5* and *HSD17B12* may be functionally related.

- The expression of antisense lncRNA *PROX1-AS1* was identified as a potential regulator of T2D-associated genetic variation using 2SMR and S-PrediXcan, and given the expression association with its sense coding gene *PROX1*, which has been functionally associated with insulin secretion, *PROX1-AS1* has been identified as a candidate functional regulatory lncRNA associated with glucose homeostasis.

The first analysis in this chapter looked for concordant effects between the eQTLs identified in Chapter 5 and GWAS MAs for T2D, FG, and FI. There has been no standard acceptable pipeline for performing such an analysis, with each of the five previous concordance studies in islets (Table 6.1) using a different method.[98,100–102,270] Given that the islet samples used in each of these studies make up a portion of the total islet samples from InsPIRE used in my analysis, it is not surprising that the majority (11/16) of the unique potential effector eGenes I identified have been published in these earlier studies.

Surprisingly, there were only nine genes from my eQTL concordance analysis that were also discovered in a colocalisation analysis performed by other members of the InsPIRE collaboration using the same genotype and RNA-seq data. There are a number of factors that likely influenced these differences:

- The InsPIRE core analysis used exon-level eQTLs compared to gene-level eQTLs used in this chapter.

- The core InsPIRE analysis used glycaemic trait SNP information from the ENGAGE MA,[323] instead of data from MAGIC[310] used in this chapter.
- Gene expression filtering was more strict in the InsPIRE analysis.

The differences in identified concordant islet eQTL and MA effects between my analysis and published studies indicate clearly that results are method- and sample-size dependent. It has been shown that even with sample sizes greater than 300, *cis* eQTL discovery is hampered by statistical power,[89] which has direct consequences in colocalisation analyses, which are dependent on identification of eQTLs in the tissue being studied. This sort of problem is not new in genomic studies, with a prime example being DE analysis, where in early studies with limited sample sizes, different methods yielded different results, but as study sample sizes have increased in general, results are no longer so significantly dependent on the choice of method.[225] Since there are only as many tests per tissue as there are expressed genes in DE analysis, powerful DE discovery requires smaller sample sizes than eQTL discovery, for which there are generally millions of tests in a single tissue. This suggests that compared to DE analyses, where results converge between methods with $n > 12$, many more samples will be required before convergence of concordance or colocalisation methods.[225] It is important to note that despite the concordant variants identified in this chapter tagging the local lead variant in both the MA and eQTL studies ($R^2 > 0.8$), it is possible that the variants identified here are not truly causal for both expression and phenotype effect, and may still represent chance overlaps.

Following single-variant analyses, I used multi-variant gene expression models derived using the methodology described in Chapter 5 as genetic instruments to perform 2SMR and S-PrediXcan analyses. These methods allowed for identification of genes for which there are multiple variants each contributing small effects to both gene expression and a phenotype of interest. Both multi-variant methods replicated a portion of significant genes from the single-variant analysis,

which suggests that for these genes (such as *HSD17B12* shown in Figure 6.3), there is a clear single block of correlated variants being picked up by the multi-variant methods, rather than the multi-SNP contributions hypothesised.

There were nineteen unique genes that had significant effects in both multi-variant methods, but not in the single-variant analysis. Some of these genes, such as *ACP2*, represent cases where multiple variants contribute joint corresponding associations with both gene expression and the outcome of interest. For *ACP2*, expression modelling selected four variants, all of which had consistent direction of effect for FG and *ACP2* expression (Figure 6.11). The most significant of these variants in the FG MA was just above the significance threshold ($p = 2.08 \times 10^{-7}$), and was thus not identified using single-variant concordance analysis. *ACP2* represents an example case where multi-variant gene-level test methods were able to identify a potential FG-associated effector gene that was not identified using single-variant concordance analysis.

It is important to note some limitations with multi-variant gene-level tests. As was the case for genes such as *PLEKHA1* and *HSD17B12* (Figures 6.2 and 6.3 respectively), significant associations using gene-level methods do not necessarily imply that there are truly multiple small signals contributing to the overall effects. Instead, these associations can sometimes be driven by a single variant. In other cases, results identified using multi-SNP methods also represent genes that may have been detected with single variant concordance analysis at reduced thresholds, such as was the case with *ACP2*, where one variant nearly reached the significance threshold for FG, and would have been identified by reducing the MA cut-off threshold.

In the case of *SPC25* (another gene identified using gene-level tests but not single-variant concordance), expression modelling selected only its lead eQTL SNP, which was significant in the FG MA; however, the top variant in the FG MA for that region (rs560887), which was used for single-variant concordance analysis, was

not a significant eQTL ($p = 0.09$). This is a case where differences in methodology — identifying the lead trait variant, then looking up eQTL results compared to identifying lead expression variants, then looking up trait associations — result in differing analysis outcomes. These examples together reiterate the importance of running different methodologies side-by-side to identify the most comprehensive list of candidate mediating genes as possible.

Despite testing for SNP heterogeneity and potential pleiotropic effects in the multi-variant gene-level analyses, it is not possible to entirely rule out pleiotropy. There may be an unobserved phenotype causing changes in both gene expression and the relevant outcome, rather than the proposed mechanism of gene expression being the causal influencer.

Given the theme of this thesis, I also focussed on robust lncRNA associations and their plausibility as regulatory genes in T2D risk and glycaemic traits. The lncRNA *PROX1-AS1* was identified using both multi-SNP methods. Although the *PROX1-AS1* RNA has not been previously associated with glycaemic traits or T2D, its corresponding sense gene, *PROX1*, has been experimentally validated as a relevant gene for insulin secretion, though the exact mechanism is not known.[322] The data presented here provide evidence that is suggestive of *PROX1-AS1* being the gene directly affected by T2D-associated genetic variation at that region, which may then regulate *PROX1* expression, function, and downstream effect on insulin secretion. Although *PROX1-AS1* was identified using both 2SMR and S-PrediXcan, the lead eQTL variant and the lead T2D MA variant in the region were not in high LD ($R^2 = 0.46$), so it is possible that the overlapping signals for eQTL and phenotype are not tagging the same causal variant (Figure 6.13). Experimental validation of *PROX1-AS1* would be required to answer these questions.

CTB-12O2.1, a lncRNA with potential functional relevance to the regulation of insulin secretion based on gene expression analysis in Chapter 4, had no significant

eQTL or multi-variant expression model. While this does not exclude the possibility that this lncRNA is a regulator, incorporation of genetic data does not provide further insight into its potential function or mechanism. Given the regulatory potential shown in Chapter 4, it still stands as a plausible candidate regulatory lncRNA in glycaemic traits and T2D; however, experimental validation would be necessary to test this hypothesis.

This chapter has demonstrated how both single-variant concordance analyses and multi-SNP gene-level tests can complement each other and provide a more comprehensive set of putative functional expressed genes. In demonstrating the complementary nature of these techniques, it was clearly shown that concordance (or colocalisation) and gene-level test results are still method- and sample-size-dependent, even with the relatively large sample of human islets used for analyses. Nevertheless, I was able to identify a set of 33 genes with robust detection (found using at least 2/3 discovery methods), including several biologically plausible candidate effector genes, such as *HSD17B12* and antisense lncRNA *PROX1-AS1*. Thus, by using the largest dataset of paired genotype and islet RNA-seq data that is available, this chapter has contributed an improved understanding of the regulatory potential of T2D-associated genetic variants, and has identified a robust set of genes that are potential mediators of these effects, and may be directly involved in glucose homeostasis and T2D risk.

Chapter 7

General discussion

The results presented in this thesis addressed four primary aims:

- Characterisation of differences in long noncoding RNA (lncRNA) and messenger RNA (mRNA) expression profiles, assessment of the downstream differences between RNA sequencing (RNA-seq) protocols, and assessment of the abundance of unannotated genes in human islets using total RNA-seq (Chapter 3).
- Identification of genes and lncRNA-enriched gene clusters that are involved in the regulation of insulin secretion and associated with type 2 diabetes (T2D) risk (Chapter 4).
- Demonstration of a pipeline for determining multi-variant models of gene expression that is more robust and sparse than common penalised regression approaches (Chapter 5).
- Identification of genes — particularly lncRNAs — that mediate genetic associations with T2D and glycaemic traits by using genetic models of gene expression and high-powered published meta-analyses (Chapter 6).

The first section of Chapter 3 described the general expression profile of lncRNAs in human tissues, and compared these profiles to those of mRNAs. This

analysis expanded on previous studies of the tissue-specificity of lncRNAs[186–188] in several ways: i) comparisons were made using a larger panel of tissues compared to previous studies; ii) tissue sample sizes were larger, allowing for more thorough quantification of genes; and iii) all tissues were from The Genotype-Tissue Expression Project (GTEx), and were therefore collected, prepared, and sequenced using strict, homogeneous procedures. This analysis ultimately found that significantly more lncRNAs than mRNAs are specific to a single tissue type, which highlighted the importance of performing lncRNA studies in functionally relevant tissues.

The remainder of Chapter 3 identified limitations of poly-A-tailed RNA selection (PA) on the detection and quantification of lncRNAs in islets, and further estimated the abundance of islet lncRNAs that are not annotated in the GENCODE reference transcriptome, which is used for RNA-seq quantifications in most studies. Using a sample of human islets from 38 individuals with paired PA and ribosomal RNA depletion (RD) preparations, the analysis showed that PA is less effective than RD at isolating lncRNAs, but with RD, a high portion of intronic reads along with extremely high expression levels of several noncoding RNAs (*RN7SL2*, *RN7SL1*, *RN7SK*, and *RPPH1*) severely dilutes the relative sequencing depth compared to PA. This finding is consistent with previous studies that observed a higher number of quantified lncRNAs and several aberrantly expressed noncoding RNAs after RD compared to PA (although in other tissue and cell types).[143,145] Given the high proportion of intergenic reads after RD, I then used RD RNA-seq from islets of 54 donors to identify transcribed regions in islets that are not annotated in GENCODE lncRNA definitions. I concluded from results in Chapter 3 that PA and GENCODE mapping — the protocols used for RNA-seq in Chapters 4-6 — do not capture all lncRNAs expressed in islets. However, the *de novo* lncRNA detection presented in Chapter 3 was compared to previous smaller islet lncRNA detection efforts,[103,137] and all studies resulted in

discordant findings (likely due to small sample sizes, differing islet purities, and different methodologies), making it difficult to determine which novel lncRNA definitions represented true islet expression. Given the small sample sizes of existing RD RNA-seq islet libraries and the sample- and method-dependent findings of existing *de novo* lncRNA detection efforts, existing large PA islet sample sets such as the one from the Integrated network for systematic analysis of Pancreatic Islet RNA Expression (InsPIRE) collaboration currently provide the best means of analysing lncRNA regulatory effects (albeit not exhaustively).

To identify genes that exhibited expression differences in tissue of donors with T2D, Chapter 4 first identified lncRNAs with significant differential expression (DE) between cases and controls using a panel of 48 tissues from GTEx and islet samples of 207 individuals from InsPIRE. I found that human islets had more differentially expressed genes than all GTEx tissues combined, though this may have been due in part to technical differences since islets from InsPIRE were procured and processed differently from GTEx tissues. However, given that nearly 40% of all significant DE effects in GTEx were in skeletal muscle (which is one of the key tissues in the development of T2D[261]), and that a previous islet DE study found a similar number of genes with DE based on glycated haemoglobin levels,[98] the number of genes associated with T2D in these tissues is plausible. Since most lncRNAs have no known functional mechanisms, to infer function of lncRNAs, co-expression analysis coupled with T2D-associated lncRNA enrichment was used to identify a cluster of co-expressed genes — including lncRNA *CTB-12O2.1* — that was present in only pancreas and islet and was enriched for genes involved in insulin-associated processes. *CTB-12O2.1* has not been previously studied for regulatory effects, but I found it to be the most significantly differentially expressed lncRNA in islets, and it had high expression correlation with its most proximal coding gene *GLRA1* ($r = 0.67$). *GLRA1* was downregulated in islets from T2D donors, which is consistent with two previous studies, and these studies

further showed that *GLRA1* is down-regulated in islets after acute glucose stimulation.[95,101] I thus proposed *CTB-12O2.1* as a candidate regulatory lncRNA, although using only RNA-seq data, it is not possible to ascertain whether expression changes that are associated with T2D are secondary effects caused by elevated glucose, or are changes that themselves causally influence T2D risk.[101] Knockdown of *CTB-12O2.1* followed by an assessment of resulting change to *GLRA1* expression (and expression of other co-expressed genes), along with tests of glucose-stimulated insulin secretion after knockdown of *CTB-12O2.1* would provide insight into whether this lncRNA causally influences insulin secretion, or if its expression change in islets from donors with T2D is secondary to T2D development.

Chapters 5 and 6 incorporated genetic information with RNA-seq data to identify genes that potentially mediate genetic variant effects on T2D and glycaemic traits. Though methods for calculating expression quantitative trait loci (eQTLs) have become widely standardised, methods that model gene expression jointly based on all *cis* regulatory variants vary.[156,157,159] In Chapter 5, I first showed that ordinary penalised regression as used in methods such as PrediXcan and S-PrediXcan[156,163] must be performed in an iterative way to produce replicable models (due to randomness of cross-validation). I further demonstrated that even after using LASSO regression, which imposes the strictest coefficient penalties, more than half of the gene models contained at least one pair of variants in high linkage disequilibrium. I presented a sparser variant selection pipeline that uses iterative LASSO modelling, further contribution-based variant filters, and finally ridge regression, and demonstrated that the final resulting models are not only more sparse than LASSO regression models, but they also explain a higher proportion of gene expression variation in most cases.

Analyses in Chapter 6 leveraged both eQTLs and multi-variant expression models established in Chapter 5, and used these as input for three separate

methods of identifying genes in islets that are candidate mediators of genetic variant effects on T2D and glycaemic traits. By using single-variant concordance analysis, inverse-variance weighted two-sample Mendelian Randomisation, and S-PrediXcan, I identified 33 genes with robust mediating potential. I hypothesised that genes with evidence of regulation by T2D-associated genetic variants that were also significantly differentially expressed in islets of donors with T2D were more likely than genes meeting only one such restriction to have high-impact causal influence on T2D risk. Four of the 33 genes with potential mediating effects — *HSD17B12*, *RP11-613D13.5*, *ACP2*, and *PDE8B* — also had significant DE based on T2D status in islets, and were thus prioritised for follow-up.

By assessing the findings from Chapter 6 alongside those of several publications,[89,101,312,315,316] I determined that *HSD17B12* and its antisense lncRNA *RP11-613D13.5* tend to be elevated in response to increased blood glucose, and that T2D risk allele carriers at rs1061810 have ubiquitous lower expression of *HSDS17B12* across the 49 human tissue types tested when compared to non-risk-allele carriers. The association between the rs1061810 T2D risk allele and lower *HSD17B12* expression was still present in islets after stratifying by T2D status. These genes have thus been proposed as candidate mediators of the rs1061810 effect on T2D; however, given the global scale of eQTL associations, it is possible that the gene expression association is secondary to T2D development rather than causal. Since the eQTL association between rs1061810 and *HSD17B12* is widespread across all tested tissues and in both healthy donors and those with T2D, it is also possible that the same variant affects T2D risk and *HSD17B12* expression through independent mechanisms, and the expression changes to *HSD17B12* may therefore be independent of the rs1061810 association with T2D risk. Functional experiments are necessary to delineate these effects. I also identified a second lncRNA, *PROX1-AS1*, that has been previously unstudied, but may be a regulator of its paired sense gene *PROX1*, which has high expression

correlation with *PROX1-AS1* in islets. *PROX1* has been functionally associated with insulin secretion in rat INS-1E cells,[322] and was prioritised by Ottosson-Laakso *et al.* as one of 92 genes most likely to have a causal influence on T2D development based on DE between islets of hyperglycaemic and normoglycaemic donors, lack of expression change after acute glucose exposure in healthy islets, and significant correlation with insulin secretion.[101] Given the expression regulation effect of the T2D-associated variant rs340874 on *PROX1-AS1* (but not *PROX1*), the expression correlation between *PROX1* and *PROX1-AS1*, and the findings by Ottosson-Laakso *et al.*, *PROX1-AS1* has been proposed as a candidate regulator of *PROX1* expression, and may ultimately have a role in the development of T2D.

This thesis has effectively identified candidate functional lncRNAs and mRNAs that may be of direct relevance to glucose homeostasis and T2D pathogenesis. However, these *in silico* discoveries do not serve as confirmation of true causal effects. Instead, the results presented can be used to prioritise genes and lncRNAs, and hypothesise functional mechanisms that can be followed up and proven or disproven experimentally. For many analyses in my thesis, lncRNA knockdown or overexpression experiments followed by RNA-seq or quantitative reverse transcription polymerase chain reaction would provide insight into causal transcriptional relationships between lncRNAs (such as *CTB-12O2.1*, *RP11-613D13.5*, and *PROX1-AS1*) and their corresponding candidate target genes. Given the hypothesised relevance to glucose homeostasis of the majority of genes identified throughout this thesis, measures of expression changes after acute glucose stimulation, and experimental testing for changes to glucose-stimulated insulin secretion in response to perturbed expression of these lncRNAs would further inform the potential functional mechanisms through which they act.[95,101,103,137,251,324,325]

The conclusion in Chapter 3 that total RNA-seq is required to exhaustively

sequence and annotate lncRNAs in islets raises a conundrum for the design of future experiments, since it was also shown that RNA-seq after RD is inefficient, with approximately 50% of all resulting reads mapping to either introns, or to one of four highly abundant noncoding RNAs. This significant reduction in relative sequencing depth means that to maintain a similar mRNA sequencing depth when compared to PA preparation, samples must be sequenced at nearly twice the depth. Although current islet RD RNA-seq datasets are relatively small, merging existing datasets would allow for improved lncRNA detection and annotation. For example, combining the samples presented in Chapter 3 with those published by Akerman *et al.*[103] would bring the overall sample size to nearly 100 individuals, allowing for improved precision in *de novo* lncRNA detection. Since most current studies are primarily focussed on mRNAs and require cost-effective sequencing for large sample sets, RD RNA-seq sample sizes are not increasing at the same rate as PA datasets; however, as further datasets are released, more comprehensive lncRNA annotation and downstream analysis will be possible.

Most of the analyses throughout this thesis were based on RNA-seq from human pancreatic islets, with sample sizes ranging from 38 individuals in the comparison between PA and RD sequencing preparations in Chapter 3, to 426 individuals in detection of effector genes at T2D- and glycaemic-trait-associated loci in Chapter 6. While the islet RNA-seq datasets used throughout Chapters 4-6 had more biological replicates than any other individual islet dataset currently available, the islet sample sizes are significantly smaller than sample sizes in RNA-seq studies of other more easily accessible tissue types such as blood, with eQTL meta-analyses in whole blood now reaching sample sizes above 5,000.[89,149,301,326] It was shown in the most recent GTEx Consortium core publication that eQTL discovery has not yet plateaued with sample sizes in the hundreds, and it is therefore likely that there are further gene expression effects in islets that are currently undetectable.[89]

For RNA-seq studies in whole islets, it is also important to consider purity and beta-cell content. Since islet samples tend to contain varying proportions of exocrine cells[98] and the cellular make-up of islets varies between individuals, it can be difficult to dissect which expression effects are a result of exocrine cells, beta-cells, alpha-cells, or delta-cells.[327] Pure human pancreatic islets typically consist of approximately 50% beta-cells, 40% alpha-cells, 10% delta-cells, and a small proportion of pancreatic-polypeptide cells,[327] thus even in pure islets, it is difficult to determine whether observed gene expression effects or mechanisms are due to a single cell-type, or occur broadly throughout islets. Although some datasets of cell-type-specific RNA-seq exist for beta-cells and alpha-cells, the sample sizes are significantly smaller than existing whole islet RNA-seq datasets (which are themselves small), making powerful eQTL and expression regulation analysis difficult in sorted islet cells.[96,328] While rapid improvements to the depth, throughput, and sample quality requirements of single-cell sequencing will partially mitigate this problem in the future,[329,330] the number of cell-type-specific libraries created from whole human islets is still limited by availability of samples, since only a subset of the already small islet sample sets available are of sufficient quality and mass to perform cell-type-specific RNA-seq experiments.

It is clear that as sample sizes and sequencing depth increase, and as islet-specific lncRNA annotations improve, the power and resolution of analysis methods presented throughout this thesis are likely to improve. However, given the existing limitations in transcriptomic analysis of pancreatic islets, the analyses presented throughout have effectively leveraged the largest paired genotype and islet RNA-seq dataset available and have incorporated public RNA-seq data from a wide spectrum of human tissues in GTEx. These data have not only expanded the general understanding of lncRNA expression, detection, and tissue specificity, but have also identified several lncRNAs and mRNAs that may be of functional relevance to the development of T2D. The results presented have demonstrated the

ways in which public large-scale datasets such as GTEx and large-scale meta-analyses can be used in tandem with focussed islet-specific data to prioritise plausible functional mechanisms influencing T2D pathogenesis. This dissertation will be important for guiding future data collection, refining methodologies, and designing experimental studies in the realm of T2D transcriptomics, and further application in ever-improving datasets of the methods and analyses throughout, along with experimental validation of findings, will provide insight into actionable mechanisms and therapeutic targets to tackle the ever-growing social and economic burden of T2D.

Appendices

Table A.1: Overall expression summary for GTEx tissues. Table outlining the number of samples, total genes expressed, and overall number of unique genes expressed in each tissue.
$^{\alpha,\beta}$Number of unique lncRNA and coding genes expressed

Tissue	n	**genes**	**unique**	**lncRNA**$^{\alpha}$	**coding**$^{\beta}$
Adipose Subcutaneous	444	18952	13	11	2
Adipose Visceral	355	19084	20	10	4
Adrenal Gland	190	18299	55	31	10
Artery Aorta	299	17835	12	8	0
Artery Coronary	173	18047	4	4	0
Artery Tibial	442	17944	16	10	3
Brain Amygdala	100	18542	8	3	3
Brain Anterior Cingulate	122	18431	4	3	1
Brain Caudate	162	19111	7	3	0
Brain Cerebellar Hemisphere	138	18153	26	9	4
Brain Cerebellum	175	18523	27	13	3
Brain Cortex	159	18578	4	3	0
Brain Frontal Cortex	129	18225	4	2	1
Brain Hippocampus	125	18953	16	10	3
Brain Hypothalamus	122	19150	26	8	14
Brain Nucleus Accumbens	148	19090	17	10	2
Brain Putamen	126	18579	6	5	1
Brain Spinal Cord	92	18016	17	13	0

Table A.1: Overall expression summary for GTEx tissues. Table outlining the number of samples, total genes expressed, and overall number of unique genes expressed in each tissue.
$^{\alpha,\beta}$Number of unique lncRNA and coding genes expressed

Tissue	n	**genes**	**unique**	**lncRNA**$^{\alpha}$	**coding**$^{\beta}$
Brain Substantia Nigra	88	18180	2	1	0
Breast Mammary	291	19726	41	21	8
Cells EBV Transformed Lymphocytes	130	15750	111	59	17
Cells Transformed Fibroblasts	345	16362	63	39	10
Colon Sigmoid	233	18532	10	7	2
Colon Transverse	274	19468	44	28	4
Esophagus Gastroesophageal Junction	244	18225	1	1	0
Esophagus Mucosa	408	18661	66	33	8
Esophagus Muscularis	370	18370	3	2	1
Heart Atrial Appendage	297	18442	31	16	6
Heart Left Ventricle	303	18273	32	16	3
Liver	176	17605	191	103	24
Lung	427	20171	91	54	14
Muscle Skeletal	566	17147	123	55	15
Nerve Tibial	416	19283	68	37	10
Pancreas	248	17308	54	19	10
Pituitary	184	20015	155	82	32
Salivary Gland Minor	97	18198	21	13	7
Skin Not Sun Exposed	389	19408	40	15	9
Skin Sun Exposed	475	19407	22	6	2
Small Intestine	137	19438	31	15	7
Spleen	162	18595	62	27	4
Stomach	262	19250	47	16	12
Thyroid	447	20229	142	79	13
Whole Blood	409	17114	226	66	34

Table A.2: Summary of PA and RD paired libraries. This table shows a comparison of basic read mapping statistics for each sample using both sequencing methods.

Sample	Method	Total reads (Million)	Mapped reads (Million)	rRNA reads	Exonic reads (%)	Intronic reads (%)	Intergenic reads (%)
Sample 1	RD	186.4	172.3 (92.5%)	74	62.2	31.8	6
	PA	127	118.1 (93%)	70	81.6	13.9	4.5
Sample 2	RD	208.6	195.6 (93.8%)	102	64.8	28.6	6.7
	PA	150.1	133.8 (89.1%)	46	83.7	11.8	4.5
Sample 3	RD	230.4	220.5 (95.7%)	166	57.6	36.7	5.7
	PA	104.5	91.1 (87.2%)	90	83.1	13.6	3.3
Sample 4	RD	213	204.2 (95.9%)	142	50.9	44.6	4.5
	PA	136.1	122.6 (90.1%)	150	82.4	14	3.6
Sample 5	RD	195.8	188.3 (96.2%)	166	56.1	38.2	5.6
	PA	117.6	108.1 (92%)	180	87.9	8.9	3.3
Sample 6	RD	221.6	212.9 (96.1%)	164	60.5	33.8	5.7
	PA	138.5	126.8 (91.5%)	124	85.1	11.1	3.8
Sample 7	RD	164.6	155.9 (94.7%)	160	70.1	25.5	4.4
	PA	155.3	144.7 (93.2%)	194	88.1	7.1	4.8

Table A.2: Summary of PA and RD paired libraries. This table shows a comparison of basic read mapping statistics for each sample using both sequencing methods.

Sample	**Method**	**Total reads (Million)**	**Mapped reads (Million)**	**rRNA reads**	**Exonic reads (%)**	**Intronic reads (%)**	**Intergenic reads (%)**
Sample 8	RD	145.3	138.5 (95.3%)	58	83.5	13.3	3.2
	PA	98.3	91.9 (93.5%)	50	71.1	5.5	23.4
Sample 9	RD	162.3	154.8 (95.4%)	82	63.8	31.7	4.5
	PA	106	98.9 (93.3%)	64	61.8	12	26.2
Sample 10	RD	156.9	151.7 (96.7%)	98	54.2	40.2	5.6
	PA	120.9	115.4 (95.4%)	66	77.3	13	9.7
Sample 11	RD	210.9	200.7 (95.1%)	106	69.9	25.9	4.2
	PA	125.1	114.3 (91.4%)	34	72.9	9.4	17.8
Sample 12	RD	154.7	145.5 (94%)	84	71.2	24.4	4.4
	PA	140.2	128.3 (91.6%)	98	90.6	6.8	2.5
Sample 13	RD	217.7	208 (95.5%)	92	69.4	26.8	3.8
	PA	149.6	138.7 (92.7%)	144	88.8	8.5	2.8
Sample 14	RD	151.4	146.3 (96.6%)	116	77.4	19.2	3.5
	PA	129.2	120.2 (93%)	84	90.1	7.2	2.7

Table A.2: Summary of PA and RD paired libraries. This table shows a comparison of basic read mapping statistics for each sample using both sequencing methods.

Sample	**Method**	**Total reads (Million)**	**Mapped reads (Million)**	**rRNA reads**	**Exonic reads (%)**	**Intronic reads (%)**	**Intergenic reads (%)**
Sample 15	RD	202.7	193.6 (95.5%)	98	73.5	22.2	4.3
	PA	134.5	124.5 (92.5%)	100	88.7	8.1	3.2
Sample 16	RD	156.2	147.2 (94.2%)	144	56.8	34.9	8.3
	PA	116.3	110.5 (95.1%)	46	89.4	7.9	2.7
Sample 17	RD	160.3	147.8 (92.2%)	104	67.9	27.1	5.1
	PA	125.5	117.7 (93.8%)	90	90.2	7.3	2.5
Sample 18	RD	158.8	149.4 (94.1%)	182	52	38.5	9.6
	PA	117.7	110.4 (93.8%)	58	89.3	7.7	3
Sample 19	RD	189.7	186.9 (98.5%)	48	12.8	4.5	82.7
	PA	132	124 (93.9%)	86	93.7	4.4	1.9
Sample 20	RD	161.3	154.9 (96.1%)	204	47.3	46.9	5.8
	PA	105.5	97.4 (92.3%)	46	93.2	4.9	1.9
Sample 21	RD	174.5	168.9 (96.8%)	216	51.1	42.5	6.4
	PA	111	102.9 (92.7%)	94	91.8	5.4	2.8

Table A.2: Summary of PA and RD paired libraries. This table shows a comparison of basic read mapping statistics for each sample using both sequencing methods.

Sample	**Method**	**Total reads (Million)**	**Mapped reads (Million)**	**rRNA reads**	**Exonic reads (%)**	**Intronic reads (%)**	**Intergenic reads (%)**
Sample 22	RD	159.9	154.2 (96.4%)	58	70.4	24.9	4.8
	PA	118.6	109 (92%)	62	91.6	6.3	2.1
Sample 23	RD	151.5	145.3 (95.9%)	66	63.7	31.1	5.2
	PA	110.9	103.7 (93.5%)	60	94	4	2.1
Sample 24	RD	157.1	147.7 (94%)	76	74.2	20.9	4.9
	PA	134	126 (94%)	60	89.9	7.6	2.5
Sample 25	RD	159	147.3 (92.6%)	76	65.5	30.1	4.4
	PA	102.2	93.2 (91.2%)	58	89.7	8	2.4
Sample 26	RD	165.9	156.9 (94.6%)	148	53.1	42.4	4.5
	PA	135.5	125.3 (92.5%)	110	85.9	11.5	2.6
Sample 27	RD	161.6	153 (94.7%)	116	58.7	34.7	6.5
	PA	107.7	99.6 (92.5%)	38	89.7	7.4	2.9
Sample 28	RD	160.1	150.7 (94.2%)	86	61	33.3	5.7
	PA	140.5	129.3 (92%)	78	87	9.9	3.1

Table A.2: Summary of PA and RD paired libraries. This table shows a comparison of basic read mapping statistics for each sample using both sequencing methods.

Sample	**Method**	**Total reads (Million)**	**Mapped reads (Million)**	**rRNA reads**	**Exonic reads (%)**	**Intronic reads (%)**	**Intergenic reads (%)**
Sample 29	RD	238.9	229.5 (96.1%)	204	52.6	41.9	5.5
	PA	53.1	48.5 (91.2%)	92	85.3	11	3.7
Sample 30	RD	161.3	155.4 (96.3%)	100	57.8	36.6	5.6
	PA	117.9	109.3 (92.7%)	92	72.2	11.8	16
Sample 31	RD	167	159.6 (95.6%)	70	84.1	12.4	3.4
	PA	103.3	95 (92%)	84	91	6.4	2.6
Sample 32	RD	156.8	151.7 (96.7%)	24	92.8	4.7	2.5
	PA	120.4	105.5 (87.6%)	18	96.1	1.6	2.4
Sample 33	RD	165.1	160.3 (97.1%)	20	86.7	9	4.3
	PA	146.3	128.6 (87.9%)	22	94.1	2.6	3.3
Sample 34	RD	170.6	165.1 (96.7%)	28	92.5	4.1	3.4
	PA	126.1	110.8 (87.8%)	22	94.5	1.7	3.8
Sample 35	RD	167.4	160.8 (96%)	44	88.9	7	4.1
	PA	140.9	123.1 (87.3%)	20	93.3	3.5	3.3

Table A.2: Summary of PA and RD paired libraries. This table shows a comparison of basic read mapping statistics for each sample using both sequencing methods.

Sample	**Method**	**Total reads (Million)**	**Mapped reads (Million)**	**rRNA reads**	**Exonic reads (%)**	**Intronic reads (%)**	**Intergenic reads (%)**
Sample 36	RD	138.9	134.1 (96.5%)	62	87.5	8.5	4
	PA	151.9	135 (88.9%)	34	92.6	4	3.4
Sample 37	RD	166.4	159.8 (96.1%)	134	68.9	26	5
	PA	120.5	109.1 (90.5%)	34	90.6	5.6	3.8
Sample 38	RD	159.5	154 (96.6%)	54	78.9	16.8	4.3
	PA	128.5	116 (90.3%)	46	87	6.4	6.6

Table A.3: Overrepresented reads in contaminated RD islet sample. Table showing all overrepresented sequences (> 0.3% of reads) from both read ends of paired-end reads from the contaminated RD islet sample. The best BLAST[212] matches for all reads are strains of *pseudomonas*.

Sequence	% reads	End
CTCCGCTGTCGGTACTACATGCTTAGCCGTGTCTATTGAGTTAATCCGCAGCCGCCCGACGGGCAGGGTGCTTTG	1.6	1
GTACTCCGCTGTCGGTACTACATGCTTAGCCGTGTCTATTGAGTTAATCCGCAGCCGCCCGACGGGCAGGGTGCT	0.6	1
CCGCTGTCGGTACTACATGCTTAGCCGTGTCTATTGAGTTAATCCGCAGCCGCCCGACGGGCAGGGTGCTTTGGA	0.5	1
GCCGTGTCTATTGAGTTAATCCGCAGCCGCCCGACGGGCAGGGTGCTTTGGACGAGTTGGGTAAGTTTTAGTCGC	0.5	1
CCCCAACGTGCTACACGACGATCCTGTTCTGTATGACAATCACTTCGGGTTTACAGGCATCCCCTAGTGATTGCT	0.4	1
GTTCTGTATGACAATCACTTCGGGTTTACAGGCATCCCCTAGTGATTGCTGGAGCCGAAGCTACCAGGAGAGCGG	0.4	1
GTCGGTACTACATGCTTAGCCGTGTCTATTGAGTTAATCCGCAGCCGCCCGACGGGCAGGGTGCTTTGGACGAGT	0.3	1
TAGGATTCGACGCCGGTGATGAAACTCTAGGTGCATGCCGAGTTGGTAACAGAACTCGTAAATCCACTGTTGCAA	0.6	2
TGAAACTCTAGGTGCATGCCGAGTTGGTAACAGAACTCGTAAATCCACTGTTGCAACTTTCTATAGTTGCCAATG	0.4	2
ATTCGACGCCGGTGATGAAACTCTAGGTGCATGCCGAGTTGGTAACAGAACTCGTAAATCCACTGTTGCAACTTT	0.4	2
TTAGGATTCGACGCCGGTGATGAAACTCTAGGTGCATGCCGAGTTGGTAACAGAACTCGTAAATCCACTGTTGCA	0.4	2
GGATTCGACGCCGGTGATGAAACTCTAGGTGCATGCCGAGTTGGTAACAGAACTCGTAAATCCACTGTTGCAACT	0.3	2

Table A.4: GTEx tissues and sample make-up used in Chapter 4. Table outlining the total number of samples, as well as the number and percentage with T2D. This represents a subset of GTEx tissues with $n > 50$, and only includes samples with available T2D status.

Tissue	n	n **with T2D (%)**
Adipose Subcutaneous	442	92 (20.8)
Adipose Visceral	354	68 (19.2)
Adrenal Gland	190	37 (19.5)
Artery Aorta	297	46 (15.5)
Artery Coronary	173	29 (16.8)
Artery Tibial	441	91 (20.6)
Brain Amygdala	100	24 (24)
Brain Anterior Cingulate	122	26 (21.3)
Brain Caudate	162	43 (26.5)
Brain Cerebellar Hemisphere	138	33 (23.9)
Brain Cerebellum	174	31 (17.8)
Brain Cortex	158	37 (23.4)
Brain Frontal Cortex	129	31 (24)
Brain Hippocampus	125	32 (25.6)
Brain Hypothalamus	122	30 (24.6)
Brain Nucleus Accumbens	148	38 (25.7)
Brain Putamen	126	25 (19.8)
Brain Spinal Cord	92	19 (20.7)
Brain Substantia Nigra	88	17 (19.3)
Breast Mammary	291	46 (15.8)
Cells EBV Transformed Lymphocytes	130	22 (16.9)
Cells Transformed Fibroblasts	344	74 (21.5)
Colon Sigmoid	233	39 (16.7)
Colon Transverse	273	51 (18.7)
Esophagus Gastroesophageal Junction	243	43 (17.7)

Table A.4: GTEx tissues and sample make-up used in Chapter 4. Table outlining the total number of samples, as well as the number and percentage with T2D. This represents a subset of GTEx tissues with $n > 50$, and only includes samples with available T2D status.

Tissue	n	n **with T2D (%)**
Esophagus Mucosa	408	74 (18.1)
Esophagus Muscularis	368	64 (17.4)
Heart Atrial Appendage	295	67 (22.7)
Heart Left Ventricle	301	53 (17.6)
Liver	176	28 (15.9)
Lung	426	86 (20.2)
Muscle Skeletal	565	121 (21.4)
Nerve Tibial	416	83 (20)
Ovary	134	25 (18.7)
Pancreas	246	40 (16.3)
Pituitary	184	39 (21.2)
Prostate	152	25 (16.4)
Salivary Gland Minor	97	14 (14.4)
Skin Not Sun Exposed	387	77 (19.9)
Skin Sun Exposed	474	101 (21.3)
Small Intestine	137	22 (16.1)
Spleen	162	26 (16)
Stomach	262	40 (15.3)
Testis	257	52 (20.2)
Thyroid	445	95 (21.3)
Uterus	110	19 (17.3)
Vagina	116	20 (17.2)
Whole Blood	408	89 (21.8)

Table A.5: Co-expression modules enriched for T2D-associated lncRNAs. Each row represents a gene co-expression module from GTEx that is enriched for lncRNA T2D DE p-values.

Tissue:Module	**Coding** n	**lncRNA** n	**other** n
Adipose Visceral:1	26	5	2
Adrenal Gland:1	211	34	12
Artery Aorta:1	43	12	0
Artery Coronary:1	294	13	4
Artery Coronary:2	70	11	3
Artery Tibial:1	114	14	1
Artery Tibial:2	62	1	0
Brain Anterior Cingulate:1	353	70	19
Brain Anterior Cingulate:2	208	38	5
Brain Anterior Cingulate:3	44	12	4
Brain Caudate:1	173	29	5
Brain Cerebellar Hemisphere:1	494	79	23
Brain Cerebellum:1	24	57	4
Brain Cerebellum:2	665	44	8
Brain Frontal Cortex:1	9	21	0
Brain Frontal Cortex:2	48	9	0
Brain Hypothalamus:1	249	57	8
Brain Spinal Cord:1	345	113	31
Breast Mammary:1	117	8	1
Breast Mammary:2	175	3	0
Colon Transverse:1	51	97	8
Esophagus Mucosa:1	452	44	3
Nerve Tibial:1	248	25	5
Nerve Tibial:2	221	10	2
Pancreas:1	49	5	0

Table A.5: Co-expression modules enriched for T2D-associated lncRNAs. Each row represents a gene co-expression module from GTEx that is enriched for lncRNA T2D DE p-values.

Tissue:Module	**Coding** n	**lncRNA** n	**other** n
Pituitary:1	259	25	9
Prostate:1	45	11	0
Skin Not Sun Exposed:1	226	20	2
Skin Not Sun Exposed:2	69	11	0
Small Intestine:1	59	121	4
Small Intestine:2	210	10	0
Testis:1	665	433	85
Testis:2	108	28	6
Thyroid:1	36	7	2

Table A.6: Pancreas module enriched GO terms. Full list of GO terms in the pancreas module enriched for lncRNA T2D DE p-values. The corrected p-value is the GO term's hypergeometric over-representation p-value after Benjamini-Hochberg correction.

GO term	**GO ID**	**Corrected p**
cell-cell signaling	GO:0007267	1.32e-05
regulation of hormone levels	GO:0010817	1.79e-05
signal release	GO:0023061	6.95e-05
peptide hormone secretion	GO:0030072	9.62e-05
peptide secretion	GO:0002790	9.62e-05
single organism signaling	GO:0044700	9.62e-05
signaling	GO:0023052	9.62e-05
cell communication	GO:0007154	9.62e-05
peptide transport	GO:0015833	0.000104
amide transport	GO:0042886	0.000145
hormone secretion	GO:0046879	0.000145
hormone transport	GO:0009914	0.00015
regulation of peptide hormone secretion	GO:0090276	0.00015
regulation of peptide secretion	GO:0002791	0.000163
regulation of peptide transport	GO:0090087	0.000165
regulation of hormone secretion	GO:0046883	0.000407
regulation of biological quality	GO:0065008	0.000772
chemical synaptic transmission	GO:0007268	0.000777
anterograde trans-synaptic signaling	GO:0098916	0.000777
synaptic signaling	GO:0099536	0.000777
trans-synaptic signaling	GO:0099537	0.000777
insulin secretion	GO:0030073	0.00171
sensory perception of pain	GO:0019233	0.00262
secretion by cell	GO:0032940	0.00297
nitrogen compound transport	GO:0071705	0.00699

Table A.6: Pancreas module enriched GO terms. Full list of GO terms in the pancreas module enriched for lncRNA T2D DE p-values. The corrected p-value is the GO term's hypergeometric over-representation p-value after Benjamini-Hochberg correction.

GO term	**GO ID**	**Corrected p**
transport	GO:0006810	0.0071
secretion	GO:0046903	0.00754
regulation of insulin secretion	GO:0050796	0.00791
single-organism localization	GO:1902578	0.00999
establishment of localization	GO:0051234	0.0102
single-organism transport	GO:0044765	0.0103
regulation of transport	GO:0051049	0.0103
protein secretion	GO:0009306	0.0154
neurological system process	GO:0050877	0.0193
regulation of ion transport	GO:0043269	0.0193
sensory perception	GO:0007600	0.0226
regulation of protein secretion	GO:0050708	0.0279
feeding behavior	GO:0007631	0.0279
negative regulation of acute inflammatory response	GO:0002674	0.0279
localization	GO:0051179	0.0283
eating behavior	GO:0042755	0.0313
G-protein coupled receptor signaling pathway	GO:0007186	0.039
regulation of secretion by cell	GO:1903530	0.0434

Table A.7: Full list of significant 2SMR results. All significant results for T2D, FG, and FI using the IVW 2SMR method.

Gene	Chr:Pos	Model SNPs	Model R^2	Trait	p
GPN1	2:27851114-27874375	2	0.04	FG	2.58e-16
SPC25	2:169690642-169769881	1	0.13	FG	1.03e-14
RFESD	5:94982458-95020477	2	0.03	FG	1.71e-13
ADCY5	3:123001143-123168605	1	0.32	T2D	2.24e-13
TRIM73	7:75024337-75040279	3	0.07	FI	5.88e-12
PROX1-AS1	1:213991094-214160973	2	0.09	FG	3.5e-11
KLHL42	12:27932953-27955973	2	0.06	T2D	5.77e-11
RP11-158I13.2	2:27805913-27858041	1	0.03	FG	3.26e-10
PROX1-AS1	1:213991094-214160973	2	0.09	T2D	2.39e-09
ADCY5	3:123001143-123168605	1	0.32	FG	7.06e-09
STARD10	11:72465774-72504726	2	0.31	FG	7.74e-09
CALCOCO2	17:46908350-46943884	9	0.13	T2D	1.3e-08
STARD10	11:72465774-72504726	2	0.31	T2D	3.06e-08
PLA2R1	2:160788519-160919121	5	0.08	T2D	6.39e-08
HSD17B12	11:43577986-43878167	2	0.38	T2D	9.22e-08
AF131215.2	8:10965298-10967236	9	0.23	T2D	1.4e-07
GTF3C2	2:27548716-27579868	1	0.02	FG	1.42e-07
FIBP	11:65651212-65656010	2	0.05	T2D	2.24e-07
NDUFS3	11:47586888-47606114	3	0.04	FG	9.75e-07
CENPA	2:26987157-27023935	1	0.02	FG	1.19e-06
CD101	1:117544382-117579167	11	0.26	T2D	1.38e-06
TUBBP5	9:141044565-141071821	5	0.57	FI	3.14e-06
GATA4	8:11534468-11617511	4	0.06	T2D	3.28e-06
ECE2	3:183967438-184010819	22	0.45	T2D	4.38e-06
RPL39L	3:186838736-186898696	2	0.08	T2D	5.25e-06
RBM43	2:152104454-152118393	2	0.06	T2D	5.32e-06
TH	11:2185159-2193107	2	0.14	T2D	7.48e-06
EDEM2	20:33703167-33865928	2	0.05	T2D	7.59e-06
RP11-48B3.4	8:81453535-81455339	5	0.08	FG	7.79e-06
LINC00543	13:28527663-28529507	2	0.04	FG	8.05e-06

Table A.7: Full list of significant 2SMR results. All significant results for T2D, FG, and FI using the IVW 2SMR method.

Gene	Chr:Pos	Model SNPs	Model R^2	Trait	p
FAM86JP	3:125635463-125648867	6	0.22	FG	8.23e-06
C11orf91	11:33719807-33722347	5	0.28	FG	9.13e-06
CAB39L	13:49882786-50018262	7	0.19	T2D	9.2e-06
SLC2A1	1:43391052-43424530	2	0.07	FG	9.88e-06
EEF1E1	6:8073593-8102811	11	0.14	T2D	1.02e-05
ITFG3	16:284545-319942	6	0.12	T2D	1.05e-05
OVGP1	1:111956936-111970399	15	0.21	FG	1.14e-05
LINC00839	10:42970991-42990784	3	0.12	FG	1.14e-05
ACP2	11:47260853-47270457	4	0.12	FG	1.24e-05
ZNF84	12:133613878-133639885	2	0.03	FG	1.27e-05
PRMT7	16:68344877-68392466	2	0.18	FG	1.93e-05
DFNB59	2:179316163-179326117	2	0.04	FG	1.95e-05
NXPH3	17:47653220-47661189	5	0.06	T2D	2.05e-05
DPY19L2	12:63952693-64062719	10	0.44	FI	2.34e-05
RNF6	13:26706253-26796791	7	0.24	T2D	2.35e-05
NEK6	9:127019885-127115586	3	0.12	T2D	3.42e-05
AC011897.1	2:14775215-14790933	11	0.12	FI	3.42e-05
PSTPIP1	15:77285700-77329673	2	0.13	T2D	3.68e-05
AF186192.5	8:145925738-145933902	6	0.11	T2D	3.69e-05
ADSS	1:244571796-244615436	15	0.22	FG	4.05e-05
ATP6V0D1	16:67471917-67515140	3	0.04	FG	4.12e-05
LAP3	4:17578815-17609595	7	0.12	T2D	4.48e-05
PDE8B	5:76506274-76725632	24	0.29	FG	4.49e-05
RP11-344N10.5	10:74857802-74861055	3	0.02	FI	4.59e-05
RP11-379B8.1	6:82554902-82643197	6	0.14	FI	4.92e-05
AC068039.4	2:172629992-172632354	11	0.22	FG	4.94e-05
NLRP2	19:55464498-55512510	17	0.46	FG	4.95e-05
CTTNBP2	7:117350705-117514193	16	0.2	T2D	5.48e-05
LRRC37A15P	4:103648431-103651878	17	0.12	T2D	5.57e-05
PRIM1	12:57125380-57146157	4	0.13	T2D	5.65e-05

Table A.7: Full list of significant 2SMR results. All significant results for T2D, FG, and FI using the IVW 2SMR method.

Gene	Chr:Pos	Model SNPs	Model R^2	Trait	p
DOC2A	16:30016830-30034591	2	0.04	T2D	5.67e-05
ADHFE1	8:67342420-67383836	18	0.31	FI	5.68e-05
PM20D1	1:205797150-205819260	17	0.51	T2D	6.1e-05
C1QL2	2:119913819-119916465	8	0.13	T2D	6.18e-05
RP11-158I13.2	2:27805913-27858041	1	0.03	FI	6.33e-05
LGR4	11:27387508-27494322	10	0.11	FI	6.43e-05
SNX32	11:65601112-65624367	4	0.24	T2D	7.17e-05
SPINK4	9:33218363-33248565	2	0.04	FG	7.3e-05
NUS1	6:117996665-118031803	11	0.24	T2D	7.35e-05
RP4-614O4.12	20:33789550-33791091	5	0.11	T2D	7.67e-05
SLC43A2	17:1472561-1532180	8	0.21	FG	7.7e-05
EIF2S2	20:32676104-32700138	5	0.06	FG	8.16e-05
SKAP1	17:46210802-46507637	11	0.14	FG	8.23e-05
CTSW	11:65647280-65651212	1	0.05	T2D	8.38e-05
GPN1	2:27851114-27874375	2	0.04	FI	8.77e-05
CCNYL2	10:42903616-42967688	11	0.26	FG	9.36e-05
KIAA1161	9:34366668-34376851	5	0.05	FG	9.46e-05
CEP41	7:130033612-130082274	2	0.04	FG	9.65e-05
INO80E	16:30006615-30017114	6	0.13	T2D	9.76e-05
CDKAL1	6:20534688-21232635	2	0.05	T2D	9.8e-05
ZNF583	19:56909335-56947404	8	0.18	FI	0.000101
DES	2:220283099-220291461	4	0.08	FI	0.000102
C1QL3	10:16555742-16564004	9	0.52	FI	0.000107
ACN9	7:96745902-96811075	7	0.11	FI	0.000113
RPPH1	14:20811207-20811844	2	0.04	FG	0.000114
ATP5I	4:666225-668127	16	0.15	T2D	0.000117
RP11-488C13.5	14:77248083-77253067	3	0.08	T2D	0.000124
LRRC2	3:46556913-46621589	15	0.17	T2D	0.000132
ABCC9	12:21950335-22094336	14	0.24	T2D	0.000134
SFTPD	10:81697496-81742370	10	0.21	FI	0.000134

Table A.7: Full list of significant 2SMR results. All significant results for T2D, FG, and FI using the IVW 2SMR method.

Gene	**Chr:Pos**	**Model SNPs**	**Model R^2**	**Trait**	p
MCF2L-AS1	13:113621798-113623138	4	0.12	FG	0.000135
SF3B2	11:65818200-65836779	5	0.09	T2D	0.000138
STEAP2	7:89796904-89867451	10	0.4	FG	0.000145
ITFG3	16:299546-319942	6	0.12	T2D	0.000147
DYNLT1	6:159057506-159065771	5	0.11	FG	0.000151
CD226	18:67498394-67629039	2	0.05	FG	0.000154
BTN2A3P	6:26421619-26432611	20	0.2	T2D	0.000155
SOX10	22:38366693-38383429	2	0.03	T2D	0.000161
GTPBP8	3:112709765-112733907	7	0.15	FG	0.000163
IST1	16:71879899-71962913	6	0.12	FG	0.00017

Table A.8: Full list of significant S-PrediXcan results. All significant results for T2D, FG, and FI using the S-PrediXcan method.

Gene	**Chr:Pos**	**Model SNPs**	**Model R^2**	**Trait**	p
MTNR1B	11:92702886-92718232	10	0.25	FG	1.52e-56
TCF7L2	10:114710009-114927437	10	0.15	T2D	8.38e-20
DGKB	7:14184674-15014402	16	0.35	FG	2.88e-15
SPC25	2:169690642-169769881	1	0.13	FG	1.03e-14
G6PC2	2:169757750-169766505	2	0.08	FG	5.14e-14
ADCY5	3:123001143-123168605	1	0.32	T2D	2.24e-13
IGF2BP2	3:185361527-185542844	4	0.16	T2D	6.12e-13
GPN1	2:27851114-27874375	2	0.04	FG	4.8e-12
NOSTRIN	2:169643049-169722024	20	0.25	FG	3.47e-11
MADD	11:47290712-47351582	6	0.26	FG	7.22e-11
FADS1	11:61567099-61596790	15	0.21	FG	1.27e-10
RP11-158I13.2	2:27805913-27858041	1	0.03	FG	3.26e-10
DGKB	7:14184674-15014402	16	0.35	T2D	1.76e-09
STARD10	11:72465774-72504726	2	0.31	FG	6.2e-09
ADCY5	3:123001143-123168605	1	0.32	FG	7.06e-09
TRIM73	7:75024337-75040279	3	0.07	FI	1.87e-08
STARD10	11:72465774-72504726	2	0.31	T2D	3.9e-08
UBE2E2	3:23244511-23633284	9	0.23	T2D	4.27e-08
KLHL42	12:27932953-27955973	2	0.06	T2D	5.42e-08
HSD17B12	11:43577986-43878167	2	0.38	T2D	6.61e-08
SLC35F6	2:26987152-27004099	14	0.16	FG	9.43e-08
RFESD	5:94982458-95020477	2	0.03	FG	1e-07
GTF3C2	2:27548716-27579868	1	0.02	FG	1.42e-07
HMG20A	15:77712754-77777949	10	0.28	T2D	2.16e-07
PLEKHA1	10:124134212-124191867	20	0.31	T2D	5.76e-07
PROX1-AS1	1:213991094-214160973	2	0.09	FG	8.76e-07
CENPA	2:26987157-27023935	1	0.02	FG	1.19e-06
FNDC4	2:27714750-27718112	4	0.05	FG	2.06e-06
STAG3L1	7:74988448-75024657	5	0.24	FI	2.23e-06
AC226150.4	12:133787246-133787772	5	0.07	FG	2.39e-06

Table A.8: Full list of significant S-PrediXcan results. All significant results for T2D, FG, and FI using the S-PrediXcan method.

Gene	**Chr:Pos**	**Model SNPs**	**Model R^2**	**Trait**	p
PMS2P3	7:75132254-75157478	11	0.18	FI	2.6e-06
ACP2	11:47260853-47270457	4	0.12	FG	2.75e-06
MARK2P9	10:94178424-94180632	8	0.12	T2D	4.25e-06
TH	11:2185159-2193107	2	0.14	T2D	7.59e-06
HTRA1	10:124221041-124274424	9	0.12	T2D	7.85e-06
CALCOCO2	17:46908350-46943884	9	0.13	T2D	7.88e-06
PDE8B	5:76506274-76725632	24	0.29	FG	9.64e-06
PROX1-AS1	1:213991094-214160973	2	0.09	T2D	9.89e-06
RP11-613D13.5	11:43851259-43902276	2	0.1	T2D	1.17e-05
RP11-274B21.8	7:128327342-128327926	14	0.18	T2D	1.71e-05
NKX6-3	8:41502697-41508855	9	0.25	T2D	2.17e-05
NDUFS3	11:47586888-47606114	3	0.04	FG	2.41e-05
DDB2	11:47236493-47260767	6	0.07	FG	3.89e-05
C1QL2	2:119913819-119916465	8	0.13	T2D	3.94e-05
YKT6	7:44240567-44253893	6	0.13	FG	5.44e-05
SF3B2	11:65818200-65836779	5	0.09	T2D	5.52e-05
PLA2R1	2:160788519-160919121	5	0.08	T2D	5.87e-05
FIBP	11:65651212-65656010	2	0.05	T2D	6.29e-05
RP11-158I13.2	2:27805913-27858041	1	0.03	FI	6.33e-05
EDF1	9:139756571-139760738	7	0.12	FG	6.74e-05
ZNF84	12:133613878-133639885	2	0.03	FG	7.06e-05
PTGFRN	1:117452679-117532980	6	0.27	T2D	7.07e-05

Bibliography

[1] World Health Organization. Global report on diabetes. Report, 2016.

[2] International Diabetes Federation. IDF diabetes atlas, 8th edn. Brussels, Belgium: International Diabetes Federation. Report, 2017.

[3] World Health Organization. Use of Glycated Haemoglobin (HbA1c) in the Diagnosis of Diabetes Mellitus. Report, 2011.

[4] Alberti, K.G.M.M. and Zimmet, P.Z. Definition, diagnosis and classification of diabetes mellitus and its complications. Part 1: diagnosis and classification of diabetes mellitus. Provisional report of a WHO Consultation. *Diabetic Medicine* **15**(7): 539–553, 1998.

[5] World Health Organization. Definition and diagnosis of diabetes mellitus and intermediate hyperglycaemia. Report, 2006.

[6] Diabetes UK. Facts and stats. Report, 2016.

[7] Zoungas, S., Woodward, M., Li, Q., Cooper, M.E., Hamet, P., Harrap, S., Heller, S., Marre, M., Patel, A., Poulter, N., Williams, B., and Chalmers, J. Impact of age, age at diagnosis and duration of diabetes on the risk of macrovascular and microvascular complications and death in type 2 diabetes. *Diabetologia* **57**(12): 2465–2474, 2014.

[8] Taylor, R. Type 2 Diabetes. *Etiology and reversibility* **36**(4): 1047–1055, 2013.

[9] Tabák, A.G., Jokela, M., Akbaraly, T.N., Brunner, E.J., Kivimäki, M., and Witte, D.R. Trajectories of glycaemia, insulin sensitivity, and insulin secretion before diagnosis of type 2 diabetes: an analysis from the Whitehall II study. *The Lancet* **373**(9682): 2215–2221, 2009.

[10] Kahn, S.E. The relative contributions of insulin resistance and beta-cell dysfunction to the pathophysiology of Type 2 diabetes. *Diabetologia* **46**(1): 3–19, 2003.

[11] Petersen, K.F., Dufour, S., Savage, D.B., Bilz, S., Solomon, G., Yonemitsu, S., Cline, G.W., Befroy, D., Zemany, L., Kahn, B.B., Papademetris, X., Rothman, D.L., and Shulman, G.I. The role of skeletal muscle insulin resistance in the pathogenesis of the metabolic syndrome. *Proceedings of the National Academy of Sciences* **104**(31): 12587–12594, 2007.

[12] Perseghin, G., Lattuada, G., De Cobelli, F., Esposito, A., Costantino, F., Canu, T., Scifo, P., De Taddeo, F., Maffi, P., Secchi, A., Del Maschio, A., and Luzi, L. Reduced intrahepatic fat content is associated with increased whole-body lipid

oxidation in patients with type 1 diabetes. *Diabetologia* **48**(12): 2615–2621, 2005.

[13] Elbein, S.C., Hasstedt, S.J., Wegner, K., and Kahn, S.E. Heritability of Pancreatic β-Cell Function among Nondiabetic Members of Caucasian Familial Type 2 Diabetic Kindreds. *The Journal of Clinical Endocrinology & Metabolism* **84**(4): 1398–1403, 1999.

[14] Watanabe, R.M., Valle, T., Hauser, E.R., Ghosh, S., Eriksson, J., Kohtamäki, K., Ehnholm, C., Tuomilehto, J., Collins, F.S., Bergman, R.N., and Boehnke, M. Familiality of Quantitative Metabolic Traits in Finnish Families with Non-Insulin-Dependent Diabetes mellitus. *Human Heredity* **49**(3): 159–168, 1999.

[15] Walford, G.A., Gustafsson, S., Rybin, D., Stančáková, A., Chen, H., Liu, C.T., Hong, J., Jensen, R.A., Rice, K., Morris, A.P., Mägi, R., Tönjes, A., Prokopenko, I., Kleber, M.E., Delgado, G., Silbernagel, G., Jackson, A.U., Appel, E.V., Grarup, N., Lewis, J.P., Montasser, M.E., Landenvall, C., Staiger, H., Luan, J., Frayling, T.M., Weedon, M.N., Xie, W., Morcillo, S., Martínez-Larrad, M.T., Biggs, M.L., Ida Chen, Y.D., Corbaton-Anchuelo, A., Færch, K., Zumaquero, J.M., Goodarzi, M.O., Kizer, J., Koistinen, H.A., Leong, A., Lind, L., Lindgren, C., Machicao, F., Manning, A.K., Martín-Núñez, G.M., Rojo-Martínez, G., Rotter, J.I., Siscovick, D.S., Zmuda, J.M., Zhang, Z., Serrano-Rios, M., Smith, U., Soriguer, F., Hansen, T., Jørgensen, T.J., Linnenberg, A., Pedersen, O., Walker, M., Langenberg, C., Scott, R.A., Wareham, N.J., Fritsche, A., Häring, H.U., Stefan, N., Groop, L., O'Connell, J.R., Boehnke, M., Bergman, R.N., Collins, F.S., Mohlke, K.L., Tuomilehto, J., März, W., Kovacs, P., Stumvoll, M., Psaty, B.M., Kuusisto, J., Laakso, M., Meigs, J.B., Dupuis, J., Ingelsson, E., and Florez, J.C. Genome-wide association study of the modified Stumvoll Insulin Sensitivity Index identifies *BCL2* and *FAM19A2* as novel insulin sensitivity loci. *Diabetes* , 2016.

[16] Goodarzi, M.O., Langefeld, C.D., Xiang, A.H., Chen, Y.D.I., Guo, X., Hanley, A.J.G., Raffel, L.J., Kandeel, F., Buchanan, T.A., Norris, J.M., Fingerlin, T.E., Lorenzo, C., Rewers, M.J., Haffner, S.M., Bowden, D.W., Rich, S.S., Bergman, R.N., Rotter, J.I., Watanabe, R.M., and Wagenknecht, L.E. Insulin Sensitivity and Insulin Clearance are Heritable and Have Strong Genetic Correlation in Mexican Americans. *Obesity (Silver Spring, Md.)* **22**(4): 1157–1164, 2014.

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

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

[19] Pessin, J. and Kwon, H. Adipokines Mediate Inflammation and Insulin Resistance. *Frontiers in Endocrinology* **4**(71), 2013.

[20] Hegele, R.A. Monogenic forms of insulin resistance: apertures that expose the common metabolic syndrome. *Trends in Endocrinology & Metabolism* **14**(8): 371–377, 2003.

[21] Ardon, O., Procter, M., Tvrdik, T., Longo, N., and Mao, R. Sequencing analysis of insulin receptor defects and detection of two novel mutations in *INSR* gene. *Molecular Genetics and Metabolism Reports* **1**(Supplement C): 71–84, 2014.

[22] Scott, R.A., Lagou, V., Welch, R.P., Wheeler, E., Montasser, M.E., Luan, J., Mägi, R., Strawbridge, R.J., Rehnberg, E., Gustafsson, S., Kanoni, S., Rasmussen-Torvik, L.J., Yengo, L., Lecoeur, C., Shungin, D., Sanna, S., Sidore, C., Johnson, P.C.D., Jukema, J.W., Johnson, T., Mahajan, A., Verweij, N., Thorleifsson, G., Hottenga, J.J., Shah, S., Smith, A.V., Sennblad, B., Gieger, C., Salo, P., Perola, M., Timpson, N.J., Evans, D.M., Pourcain, B.S., Wu, Y., Andrews, J.S., Hui, J., Bielak, L.F., Zhao, W., Horikoshi, M., Navarro, P., Isaacs, A., O'Connell, J.R., Stirrups, K., Vitart, V., Hayward, C., Esko, T., Mihailov, E., Fraser, R.M., Fall, T., Voight, B.F., Raychaudhuri, S., Chen, H., Lindgren, C.M., Morris, A.P., Rayner, N.W., Robertson, N., Rybin, D., Liu, C.T., Beckmann, J.S., Willems, S.M., Chines, P.S., Jackson, A.U., Kang, H.M., Stringham, H.M., Song, K., Tanaka, T., Peden, J.F., Goel, A., Hicks, A.A., An, P., Müller-Nurasyid, M., Franco-Cereceda, A., Folkersen, L., Marullo, L., Jansen, H., Oldehinkel, A.J., Bruinenberg, M., Pankow, J.S., North, K.E., Forouhi, N.G., Loos, R.J.F., Edkins, S., Varga, T.V., Hallmans, G., Oksa, H., Antonella, M., Nagaraja, R., Trompet, S., Ford, I., Bakker, S.J.L., Kong, A., Kumari, M., Gigante, B., Herder, C., Munroe, P.B., Caulfield, M., Antti, J., Mangino, M., Small, K., Miljkovic, I., Liu, Y., Atalay, M., Kiess, W., James, A.L., Rivadeneira, F., Uitterlinden, A.G., Palmer, C.N.A., Doney, A.S.F., Willemsen, G., Smit, J.H., Campbell, S., Polasek, O., Bonnycastle, L.L., Hercberg, S., Dimitriou, M., Bolton, J.L., Fowkes, G.R., Kovacs, P., Lindström, J., Zemunik, T., Bandinelli, S., Wild, S.H., Basart, H.V., Rathmann, W., Grallert, H., Diabetes Genetics Replication Meta-analysis Consortium, Maerz, W., Kleber, M.E., Boehm, B.O., Peters, A., Pramstaller, P.P., Province, M.A., Borecki, I.B., Hastie, N.D., Rudan, I., Campbell, H., Watkins, H., Farrall, M., Stumvoll, M., Ferrucci, L., Waterworth, D.M., Bergman, R.N., Collins, F.S., Tuomilehto, J., Watanabe, R.M., de Geus, E.J.C., Penninx, B.W., Hofman, A., Oostra, B.A., Psaty, B.M., Vollenweider, P., Wilson, J.F., Wright, A.F., Hovingh, G.K., Metspalu, A., Uusitupa, M., Magnusson, P.K.E., Kyvik, K.O., Kaprio, J., Price, J.F., Dedoussis, G.V., Deloukas, P., Meneton, P., Lind, L., Boehnke, M., Shuldiner, A.R., van Duijn, C.M., Morris, A.D., Toenjes, A., Peyser, P.A., Beilby, J.P., Körner, A., Kuusisto, J., Laakso, M., Bornstein, S.R., Schwarz, P.E.H., Lakka, T.A., Rauramaa, R., Adair, L.S., Smith, G.D., Spector, T.D., Illig, T., de Faire, U., Hamsten, A., Gudnason, V., Kivimaki, M., Hingorani, A., Keinanen-Kiukaanniemi, S.M., Saaristo, T.E., Boomsma, D.I., Stefansson, K., van der Harst, P., Dupuis, J., Pedersen, N.L., Sattar, N., Harris, T.B., Cucca, F., Ripatti, S., Salomaa, V., Mohlke, K.L., Balkau, B., Froguel, P., Pouta, A., Jarvelin, M.R., Wareham, N.J., Bouatia-Naji, N., McCarthy, M.I., Franks, P.W., Meigs, J.B., Teslovich, T.M., Florez, J.C., Langenberg, C.,

Ingelsson, E., Prokopenko, I., and Barroso, I. Large-scale association analyses identify new loci influencing glycemic traits and provide insight into the underlying biological pathways. *Nature Genetics* **44**: 991, 2012.

[23] Scott, R.A., Fall, T., Pasko, D., Barker, A., Sharp, S.J., Arriola, L., Balkau, B., Barricarte, A., Barroso, I., Boeing, H., Clavel-Chapelon, F., Crowe, F.L., Dekker, J.M., Fagherazzi, G., Ferrannini, E., Forouhi, N.G., Franks, P.W., Gavrila, D., Giedraitis, V., Grioni, S., Groop, L.C., Kaaks, R., Key, T.J., Kühn, T., Lotta, L.A., Nilsson, P.M., Overvad, K., Palli, D., Panico, S., Quirós, J.R., Rolandsson, O., Roswall, N., Sacerdote, C., Sala, N., Sánchez, M.J., Schulze, M.B., Siddiq, A., Slimani, N., Sluijs, I., Spijkerman, A.M., Tjonneland, A., Tumino, R., van der A, D.L., Yaghootkar, H., McCarthy, M.I., Semple, R.K., Riboli, E., Walker, M., Ingelsson, E., Frayling, T.M., Savage, D.B., Langenberg, C., and Wareham, N.J. Common Genetic Variants Highlight the Role of Insulin Resistance and Body Fat Distribution in Type 2 Diabetes, Independent of Obesity. *Diabetes* **63**(12): 4378–4387, 2014.

[24] Lotta, L.A., Gulati, P., Day, F.R., Payne, F., Ongen, H., van de Bunt, M., Gaulton, K.J., Eicher, J.D., Sharp, S.J., Luan, J., De Lucia Rolfe, E., Stewart, I.D., Wheeler, E., Willems, S.M., Adams, C., Yaghootkar, H., E.PIC-InterAct Consortium, Cambridge FPLD1 Consortium, Forouhi, N.G., Khaw, K.T., Johnson, A.D., Semple, R.K., Frayling, T., Perry, J.R.B., Dermitzakis, E., McCarthy, M.I., Barroso, I., Wareham, N.J., Savage, D.B., Langenberg, C., O'Rahilly, S., and Scott, R.A. Integrative genomic analysis implicates limited peripheral adipose storage capacity in the pathogenesis of human insulin resistance. *Nature Genetics* **49**: 17, 2016.

[25] Weir, G.C. and Bonner-Weir, S. Five Stages of Evolving Beta-Cell Dysfunction During Progression to Diabetes. *Diabetes* **53**(suppl 3): S16–S21, 2004.

[26] Griscelli, F., Feraud, O., Ernault, T., Oudrihri, N., Turhan, A.G., Bonnefond, A., Froguel, P., and Bennaceur-Griscelli, A. Generation of an induced pluripotent stem cell (iPSC) line from a patient with maturity-onset diabetes of the young type 13 (MODY13) with a the potassium inwardly-rectifying channel, subfamily J, member 11 (KCNJ11) mutation. *Stem Cell Research* **23**: 178–181, 2017.

[27] De Franco, E., Flanagan, S.E., Houghton, J.A.L., Allen, H.L., Mackay, D.J.G., Temple, I.K., Ellard, S., and Hattersley, A.T. The effect of early, comprehensive genomic testing on clinical care in neonatal diabetes: an international cohort study. *The Lancet* **386**(9997): 957–963, 2015.

[28] Gloyn, A.L., Pearson, E.R., Antcliff, J.F., Proks, P., Bruining, G.J., Slingerland, A.S., Howard, N., Srinivasan, S., Silva, J.M., Molnes, J., Edghill, E.L., Frayling, T.M., Temple, I.K., Mackay, D., Shield, J.P., Sumnik, Z., van Rhijn, A., Wales, J.K., Clark, P., Gorman, S., Aisenberg, J., Ellard, S., Njølstad, P.R., Ashcroft, F.M., and Hattersley, A.T. Activating Mutations in the Gene Encoding the ATP-Sensitive Potassium-Channel Subunit Kir6.2 and Permanent Neonatal Diabetes. *New England Journal of Medicine* **350**(18): 1838–1849, 2004.

[29] Bonnefond, A., Philippe, J., Durand, E., Dechaume, A., Huyvaert, M.,

Montagne, L., Marre, M., Balkau, B., Fajardy, I., Vambergue, A., Vatin, V., Delplanque, J., Le Guilcher, D., De Graeve, F., Lecoeur, C., Sand, O., Vaxillaire, M., and Froguel, P. Whole-Exome Sequencing and High Throughput Genotyping Identified *KCNJ11* as the Thirteenth MODY Gene. *PLOS ONE* **7**(6): e37423, 2012.

[30] Proks, P., Arnold, A.L., Bruining, J., Girard, C., Flanagan, S.E., Larkin, B., Colclough, K., Hattersley, A.T., Ashcroft, F.M., and Ellard, S. A heterozygous activating mutation in the sulphonylurea receptor SUR1 (*ABCC8*) causes neonatal diabetes. *Human Molecular Genetics* **15**(11): 1793–1800, 2006.

[31] Babenko, A.P., Polak, M., Cavé, H., Busiah, K., Czernichow, P., Scharfmann, R., Bryan, J., Aguilar-Bryan, L., Vaxillaire, M., and Froguel, P. Activating Mutations in the *ABCC8* Gene in Neonatal Diabetes Mellitus. *New England Journal of Medicine* **355**(5): 456–466, 2006.

[32] Pearson, E.R., Flechtner, I., Njølstad, P.R., Malecki, M.T., Flanagan, S.E., Larkin, B., Ashcroft, F.M., Klimes, I., Codner, E., Iotova, V., Slingerland, A.S., Shield, J., Robert, J.J., Holst, J.J., Clark, P.M., Ellard, S., Søvik, O., Polak, M., and Hattersley, A.T. Switching from Insulin to Oral Sulfonylureas in Patients with Diabetes Due to Kir6.2 Mutations. *New England Journal of Medicine* **355**(5): 467–477, 2006.

[33] Bonnefond, A. and Froguel, P. Rare and Common Genetic Events in Type 2 Diabetes: What Should Biologists Know? *Cell Metabolism* **21**(3): 357–368, 2015.

[34] Shields, B.M., Hicks, S., Shepherd, M.H., Colclough, K., Hattersley, A.T., and Ellard, S. Maturity-onset diabetes of the young (MODY): how many cases are we missing? *Diabetologia* **53**(12): 2504–2508, 2010.

[35] Froguel, P., Vaxillaire, M., Sun, F., Velho, G., Zouali, H., Butel, M.O., Lesage, S., Vionnet, N., Clément, K., Fougerousse, F., Tanizawa, Y., Weissenbach, J., Beckmann, J.S., Lathrop, G.M., Passa, P., Permutt, M.A., and Cohen, D. Close linkage of glucokinase locus on chromosome 7p to early-onset non-insulin-dependent diabetes mellitus. *Nature* **356**: 162, 1992.

[36] Byrne, M.M., Sturis, J., Fajans, S.S., Ortiz, F.J., Stoltz, A., Stoffel, M., Smith, M.J., Bell, G.I., Halter, J.B., and Polonsky, K.S. Altered Insulin Secretory Responses to Glucose in Subjects with a Mutation in the *MODY1* Gene on Chromosome 20. *Diabetes* **44**(6): 699–704, 1995.

[37] Njølstad, P.R., Søvik, O., Cuesta-Muñoz, A., Bjørkhaug, L., Massa, O., Barbetti, F., Undlien, D.E., Shiota, C., Magnuson, M.A., Molven, A., Matschinsky, F.M., and Bell, G.I. Neonatal Diabetes Mellitus Due to Complete Glucokinase Deficiency. *New England Journal of Medicine* **344**(21): 1588–1592, 2001.

[38] Glaser, B., Kesavan, P., Heyman, M., Davis, E., Cuesta, A., Buchs, A., Stanley, C.A., Thornton, P.S., Permutt, M.A., Matschinsky, F.M., and Herold, K.C. Familial Hyperinsulinism Caused by an Activating Glucokinase Mutation. *New England Journal of Medicine* **338**(4): 226–230, 1998.

[39] Prokopenko, I., Poon, W., Mägi, R., Prasad B, R., Salehi, S.A., Almgren, P., Osmark, P., Bouatia-Naji, N., Wierup, N., Fall, T., Stančáková, A., Barker, A., Lagou, V., Osmond, C., Xie, W., Lahti, J., Jackson, A.U., Cheng, Y.C., Liu, J., O'Connell, J.R., Blomstedt, P.A., Fadista, J., Alkayyali, S., Dayeh, T., Ahlqvist, E., Taneera, J., Lecoeur, C., Kumar, A., Hansson, O., Hansson, K., Voight, B.F., Kang, H.M., Levy-Marchal, C., Vatin, V., Palotie, A., Syvänen, A.C., Mari, A., Weedon, M.N., Loos, R.J.F., Ong, K.K., Nilsson, P., Isomaa, B., Tuomi, T., Wareham, N.J., Stumvoll, M., Widen, E., Lakka, T.A., Langenberg, C., Tönjes, A., Rauramaa, R., Kuusisto, J., Frayling, T.M., Froguel, P., Walker, M., Eriksson, J.G., Ling, C., Kovacs, P., Ingelsson, E., McCarthy, M.I., Shuldiner, A.R., Silver, K.D., Laakso, M., Groop, L., and Lyssenko, V. A Central Role for *GRB10* in Regulation of Islet Function in Man. *PLOS Genetics* **10**(4): e1004235, 2014.

[40] Sladek, R., Rocheleau, G., Rung, J., Dina, C., Shen, L., Serre, D., Boutin, P., Vincent, D., Belisle, A., Hadjadj, S., Balkau, B., Heude, B., Charpentier, G., Hudson, T.J., Montpetit, A., Pshezhetsky, A.V., Prentki, M., Posner, B.I., Balding, D.J., Meyre, D., Polychronakos, C., and Froguel, P. A genome-wide association study identifies novel risk loci for type 2 diabetes. *Nature* **445**: 881, 2007.

[41] Scott, L.J., Mohlke, K.L., Bonnycastle, L.L., Willer, C.J., Li, Y., Duren, W.L., Erdos, M.R., Stringham, H.M., Chines, P.S., Jackson, A.U., Prokunina-Olsson, L., Ding, C.J., Swift, A.J., Narisu, N., Hu, T., Pruim, R., Xiao, R., Li, X.Y., Conneely, K.N., Riebow, N.L., Sprau, A.G., Tong, M., White, P.P., Hetrick, K.N., Barnhart, M.W., Bark, C.W., Goldstein, J.L., Watkins, L., Xiang, F., Saramies, J., Buchanan, T.A., Watanabe, R.M., Valle, T.T., Kinnunen, L., Abecasis, G.R., Pugh, E.W., Doheny, K.F., Bergman, R.N., Tuomilehto, J., Collins, F.S., and Boehnke, M. A Genome-Wide Association Study of Type 2 Diabetes in Finns Detects Multiple Susceptibility Variants. *Science* **316**(5829): 1341–1345, 2007.

[42] Zeggini, E., Weedon, M.N., Lindgren, C.M., Frayling, T.M., Elliott, K.S., Lango, H., Timpson, N.J., Perry, J.R.B., Rayner, N.W., Freathy, R.M., Barrett, J.C., Shields, B., Morris, A.P., Ellard, S., Groves, C.J., Harries, L.W., Marchini, J.L., Owen, K.R., Knight, B., Cardon, L.R., Walker, M., Hitman, G.A., Morris, A.D., Doney, A.S.F., McCarthy, M.I., and Hattersley, A.T. Replication of Genome-Wide Association Signals in UK Samples Reveals Risk Loci for Type 2 Diabetes. *Science* **316**(5829): 1336–1341, 2007.

[43] Saxena, R., Voight, B.F., Lyssenko, V., Burtt, N.P., de Bakker, P.I., Chen, H., Roix, J.J., Kathiresan, S., Hirschhorn, J.N., Daly, M.J., Hughes, T.E., Groop, L., Altshuler, D., Almgren, P., Florez, J.C., Meyer, J., Ardlie, K., Bengtsson, K., Isomaa, B., Lettre, G., Lindblad, U., Lyon, H.N., Melander, O., Newton-Cheh, C., Nilsson, P., Orho-Melander, M., Råstam, L., Speliotes, E.K., Taskinen, M.R., Tuomi, T., Guiducci, C., Berglund, A., Carlson, J., Gianniny, L., Hackett, R., Hall, L., Holmkvist, J., Laurila, E., Sjögren, M., Sterner, M., Surti, A., Svensson, M., Svensson, M., Tewhey, R., Blumenstiel, B., Parkin, M., DeFelice, M., Barry, R., Brodeur, W., Camarata, J., Chia, N., Fava, M.,

Gibbons, J., Handsaker, B., Healy, C., Nguyen, K., Gates, C., Sougnez, C., Gage, D., Nizzari, M., Gabriel, S.B., Chirn, G.W., Ma, Q., Parikh, H., Richardson, D., Ricke, D., and Purcell, S. Genome-Wide Association Analysis Identifies Loci for Type 2 Diabetes and Triglyceride Levels. *Science* , 2007.

[44] The Genomes Project Consortium. A global reference for human genetic variation. *Nature* **526**(7571): 68–74, 2015.

[45] Fuchsberger, C., Flannick, J., Teslovich, T.M., Mahajan, A., Agarwala, V., Gaulton, K.J., Ma, C., Fontanillas, P., Moutsianas, L., McCarthy, D.J., Rivas, M.A., Perry, J.R.B., Sim, X., Blackwell, T.W., Robertson, N.R., Rayner, N.W., Cingolani, P., Locke, A.E., Tajes, J.F., Highland, H.M., Dupuis, J., Chines, P.S., Lindgren, C.M., Hartl, C., Jackson, A.U., Chen, H., Huyghe, J.R., van de Bunt, M., Pearson, R.D., Kumar, A., Müller-Nurasyid, M., Grarup, N., Stringham, H.M., Gamazon, E.R., Lee, J., Chen, Y., Scott, R.A., Below, J.E., Chen, P., Huang, J., Go, M.J., Stitzel, M.L., Pasko, D., Parker, S.C.J., Varga, T.V., Green, T., Beer, N.L., Day-Williams, A.G., Ferreira, T., Fingerlin, T., Horikoshi, M., Hu, C., Huh, I., Ikram, M.K., Kim, B.J., Kim, Y., Kim, Y.J., Kwon, M.S., Lee, J., Lee, S., Lin, K.H., Maxwell, T.J., Nagai, Y., Wang, X., Welch, R.P., Yoon, J., Zhang, W., Barzilai, N., Voight, B.F., Han, B.G., Jenkinson, C.P., Kuulasmaa, T., Kuusisto, J., Manning, A., Ng, M.C.Y., Palmer, N.D., Balkau, B., Stančáková, A., Abboud, H.E., Boeing, H., Giedraitis, V., Prabhakaran, D., Gottesman, O., Scott, J., Carey, J., Kwan, P., Grant, G., Smith, J.D., Neale, B.M., Purcell, S., Butterworth, A.S., Howson, J.M.M., Lee, H.M., Lu, Y., Kwak, S.H., Zhao, W., Danesh, J., Lam, V.K.L., Park, K.S., Saleheen, D., So, W.Y., Tam, C.H.T., Afzal, U., Aguilar, D., Arya, R., Aung, T., Chan, E., Navarro, C., Cheng, C.Y., Palli, D., Correa, A., Curran, J.E., Rybin, D., Farook, V.S., Fowler, S.P., Freedman, B.I., Griswold, M., Hale, D.E., Hicks, P.J., Khor, C.C., Kumar, S., Lehne, B., Thuillier, D., Lim, W.Y., Liu, J., van der Schouw, Y.T., Loh, M., Musani, S.K., Puppala, S., Scott, W.R., Yengo, L., Tan, S.T., Taylor Jr, H.A., Thameem, F., Wilson, G., Wong, T.Y., Njølstad, P.R., Levy, J.C., Mangino, M., Bonnycastle, L.L., Schwarzmayr, T., Fadista, J., Surdulescu, G.L., Herder, C., Groves, C.J., Wieland, T., Bork-Jensen, J., Brandslund, I., Christensen, C., Koistinen, H.A., Doney, A.S.F., Kinnunen, L., Esko, T., Farmer, A.J., Hakaste, L., Hodgkiss, D., Kravic, J., Lyssenko, V., Hollensted, M., Jørgensen, M.E., Jørgensen, T., Ladenvall, C., Justesen, J.M., Käräjämäki, A., Kriebel, J., Rathmann, W., Lannfelt, L., Lauritzen, T., Narisu, N., Linneberg, A., Melander, O., Milani, L., Neville, M., Orho-Melander, M., Qi, L., Qi, Q., Roden, M., Rolandsson, O., Swift, A., Rosengren, A.H., Stirrups, K., Wood, A.R., Mihailov, E., Blancher, C., Carneiro, M.O., Maguire, J., Poplin, R., Shakir, K., Fennell, T., DePristo, M., Hrabé de Angelis, M., Deloukas, P., Gjesing, A.P., Jun, G., Nilsson, P., Murphy, J., Onofrio, R., Thorand, B., Hansen, T., Meisinger, C., Hu, F.B., Isomaa, B., Karpe, F., Liang, L., Peters, A., Huth, C., O'Rahilly, S.P., Palmer, C.N.A., Pedersen, O., Rauramaa, R., Tuomilehto, J., Salomaa, V., Watanabe, R.M., Syvänen, A.C., Bergman, R.N., Bharadwaj, D., Bottinger, E.P., Cho, Y.S., Chandak, G.R., Chan, J.C.N., Chia, K.S., Daly, M.J., Ebrahim, S.B., Langenberg, C., Elliott,

P., Jablonski, K.A., Lehman, D.M., Jia, W., Ma, R.C.W., Pollin, T.I., Sandhu, M., Tandon, N., Froguel, P., Barroso, I., Teo, Y.Y., Zeggini, E., Loos, R.J.F., Small, K.S., Ried, J.S., DeFronzo, R.A., Grallert, H., Glaser, B., Metspalu, A., Wareham, N.J., Walker, M., Banks, E., Gieger, C., Ingelsson, E., Im, H.K., Illig, T., Franks, P.W., Buck, G., Trakalo, J., Buck, D., Prokopenko, I., Mägi, R., Lind, L., Farjoun, Y., Owen, K.R., Gloyn, A.L., Strauch, K., Tuomi, T., Kooner, J.S., Lee, J.Y., Park, T., Donnelly, P., Morris, A.D., Hattersley, A.T., Bowden, D.W., Collins, F.S., Atzmon, G., Chambers, J.C., Spector, T.D., Laakso, M., Strom, T.M., Bell, G.I., Blangero, J., Duggirala, R., Tai, E.S., McVean, G., Hanis, C.L., Wilson, J.G., Seielstad, M., Frayling, T.M., Meigs, J.B., Cox, N.J., Sladek, R., Lander, E.S., Gabriel, S., Burtt, N.P., Mohlke, K.L., Meitinger, T., Groop, L., Abecasis, G., Florez, J.C., Scott, L.J., Morris, A.P., Kang, H.M., Boehnke, M., Altshuler, D., and McCarthy, M.I. The genetic architecture of type 2 diabetes. *Nature* **536**: 41, 2016.

[46] Mahajan, A., Taliun, D., Thurner, M., Robertson, N.R., Torres, J.M., Rayner, N.W., Steinthorsdottir, V., Scott, R.A., Grarup, N., Cook, J.P., Schmidt, E.M., Wuttke, M., Sarnowski, C., Mägi, R., Nano, J., Gieger, C., Trompet, S., Lecoeur, C., Preuss, M., Prins, B.P., Guo, X., Bielak, L.F., Bennett, A.J., Bork-Jensen, J., Brummett, C.M., Canouil, M., Eckardt, K.U., Fischer, K., Kardia, S.L., Kronenberg, F., Läll, K., Liu, C.T., Locke, A.E., Luan, J., Ntalla, I., Nylander, V., Schönherr, S., Schurmann, C., Yengo, L., Bottinger, E.P., Brandslund, I., Christensen, C., Dedoussis, G., Florez, J.C., Ford, I., Franco, O.H., Frayling, T.M., Giedraitis, V., Hackinger, S., Hattersley, A.T., Herder, C., Ikram, M.A., Ingelsson, M., Jørgensen, M.E., Jørgensen, T., Kriebel, J., Kuusisto, J., Ligthart, S., Lindgren, C.M., Linneberg, A., Lyssenko, V., Mamakou, V., Meitinger, T., Mohlke, K.L., Morris, A.D., Nadkarni, G., Pankow, J.S., Peters, A., Sattar, N., Stančáková, A., Strauch, K., Taylor, K.D., Thorand, B., Thorleifsson, G., Thorsteinsdottir, U., Tuomilehto, J., Witte, D.R., Dupuis, J., Peyser, P.A., Zeggini, E., Loos, R.J.F., Froguel, P., Ingelsson, E., Lind, L., Groop, L., Laakso, M., Collins, F.S., Jukema, J.W., Palmer, C.N.A., Grallert, H., Metspalu, A., Dehghan, A., Köttgen, A., Abecasis, G., Meigs, J.B., Rotter, J.I., Marchini, J., Pedersen, O., Hansen, T., Langenberg, C., Wareham, N.J., Stefansson, K., Gloyn, A.L., Morris, A.P., Boehnke, M., and McCarthy, M.I. Fine-mapping of an expanded set of type 2 diabetes loci to single-variant resolution using high-density imputation and islet-specific epigenome maps. *bioRxiv* , 2018.

[47] Scott, R.A., Scott, L.J., Mägi, R., Marullo, L., Gaulton, K.J., Kaakinen, M., Pervjakova, N., Pers, T.H., Johnson, A.D., Eicher, J.D., Jackson, A.U., Ferreira, T., Lee, Y., Ma, C., Steinthorsdottir, V., Thorleifsson, G., Qi, L., Van Zuydam, N.R., Mahajan, A., Chen, H., Almgren, P., Voight, B.F., Grallert, H., Müller-Nurasyid, M., Ried, J.S., Rayner, N.W., Robertson, N., Karssen, L.C., van Leeuwen, E.M., Willems, S.M., Fuchsberger, C., Kwan, P., Teslovich, T.M., Chanda, P., Li, M., Lu, Y., Dina, C., Thuillier, D., Yengo, L., Jiang, L., Sparso, T., Kestler, H.A., Chheda, H., Eisele, L., Gustafsson, S., Frånberg, M., Strawbridge, R.J., Benediktsson, R., Hreidarsson, A.B., Kong, A., Sigurõsson,

G., Kerrison, N.D., Luan, J., Liang, L., Meitinger, T., Roden, M., Thorand, B., Esko, T.o., Mihailov, E., Fox, C., Liu, C.T., Rybin, D., Isomaa, B., Lyssenko, V., Tuomi, T., Couper, D.J., Pankow, J.S., Grarup, N., Have, C.T., Jørgensen, M.E., Jørgensen, T., Linneberg, A., Cornelis, M.C., van Dam, R.M., Hunter, D.J., Kraft, P., Sun, Q., Edkins, S., Owen, K.R., Perry, J.R., Wood, A.R., Zeggini, E., Tajes-Fernandes, J., Abecasis, G.R., Bonnycastle, L.L., Chines, P.S., Stringham, H.M., Koistinen, H.A., Kinnunen, L., Sennblad, B., Mühleisen, T.W., Nöthen, M.M., Pechlivanis, S., Baldassarre, D., Gertow, K., Humphries, S.E., Tremoli, E., Klopp, N., Meyer, J., Steinbach, G., Wennauer, R., Eriksson, J.G., Männistö, S., Peltonen, L., Tikkanen, E., Charpentier, G., Eury, E., Lobbens, S., Gigante, B., Leander, K., McLeod, O., Bottinger, E.P., Gottesman, O., Ruderfer, D., Blüher, M., Kovacs, P., Tonjes, A., Maruthur, N.M., Scapoli, C., Erbel, R., Jöckel, K.H., Moebus, S., de Faire, U., Hamsten, A., Stumvoll, M., Deloukas, P., Donnelly, P.J., Frayling, T.M., Hattersley, A.T., Ripatti, S., Salomaa, V., Pedersen, N.L., Boehm, B.O., Bergman, R.N., Collins, F.S., Mohlke, K.L., Tuomilehto, J., Hansen, T., Pedersen, O., Barroso, I., Lannfelt, L., Ingelsson, E., Lind, L., Lindgren, C.M., Cauchi, S., Froguel, P., Loos, R.J., Balkau, B., Boeing, H., Franks, P.W., Barricarte Gurrea, A., Palli, D., van der Schouw, Y.T., Altshuler, D., Groop, L.C., Langenberg, C., Wareham, N.J., Sijbrands, E., van Duijn, C.M., Florez, J.C., Meigs, J.B., Boerwinkle, E., Gieger, C., Strauch, K., Metspalu, A., Morris, A.D., Palmer, C.N., Hu, F.B., Thorsteinsdottir, U., Stefansson, K., Dupuis, J., Morris, A.P., Boehnke, M., McCarthy, M.I., and Prokopenko, I. An Expanded Genome-Wide Association Study of Type 2 Diabetes in Europeans. *Diabetes* **66**(11): 2888–2902, 2017.

[48] Albrechtsen, A., Grarup, N., Li, Y., Sparsø, T., Tian, G., Cao, H., Jiang, T., Kim, S.Y., Korneliussen, T., Li, Q., Nie, C., Wu, R., Skotte, L., Morris, A.P., Ladenvall, C., Cauchi, S., Stančáková, A., Andersen, G., Astrup, A., Banasik, K., Bennett, A.J., Bolund, L., Charpentier, G., Chen, Y., Dekker, J.M., Doney, A.S.F., Dorkhan, M., Forsen, T., Frayling, T.M., Groves, C.J., Gui, Y., Hallmans, G., Hattersley, A.T., He, K., Hitman, G.A., Holmkvist, J., Huang, S., Jiang, H., Jin, X., Justesen, J.M., Kristiansen, K., Kuusisto, J., Lajer, M., Lantieri, O., Li, W., Liang, H., Liao, Q., Liu, X., Ma, T., Ma, X., Manijak, M.P., Marre, M., Mokrosiński, J., Morris, A.D., Mu, B., Nielsen, A.A., Nijpels, G., Nilsson, P., Palmer, C.N.A., Rayner, N.W., Renström, F., Ribel-Madsen, R., Robertson, N., Rolandsson, O., Rossing, P., Schwartz, T.W., Slagboom, P.E., Sterner, M., D.E.S.I.R. Study Group, Tang, M., Tarnow, L., The Diagram Consortium, Tuomi, T., van't Riet, E., van Leeuwen, N., Varga, T.V., Vestmar, M.A., Walker, M., Wang, B., Wang, Y., Wu, H., Xi, F., Yengo, L., Yu, C., Zhang, X., Zhang, J., Zhang, Q., Zhang, W., Zheng, H., Zhou, Y., Altshuler, D., 't Hart, L.M., Franks, P.W., Balkau, B., Froguel, P., McCarthy, M.I., Laakso, M., Groop, L., Christensen, C., Brandslund, I., Lauritzen, T., Witte, D.R., Linneberg, A., Jørgensen, T., Hansen, T., Wang, J., Nielsen, R., and Pedersen, O. Exome sequencing-driven discovery of coding polymorphisms associated with common metabolic phenotypes. *Diabetologia* **56**(2): 298–310, 2013.

[49] Cook, J.P. and Morris, A.P. Multi-ethnic genome-wide association study identifies novel locus for type 2 diabetes susceptibility. *European Journal Of Human Genetics* **24**: 1175, 2016.

[50] Dupuis, J., Langenberg, C., Prokopenko, I., Saxena, R., Soranzo, N., Jackson, A.U., Wheeler, E., Glazer, N.L., Bouatia-Naji, N., Gloyn, A.L., Lindgren, C.M., Mägi, R., Morris, A.P., Randall, J., Johnson, T., Elliott, P., Rybin, D., Thorleifsson, G., Steinthorsdottir, V., Henneman, P., Grallert, H., Dehghan, A., Hottenga, J.J., Franklin, C.S., Navarro, P., Song, K., Goel, A., Perry, J.R.B., Egan, J.M., Lajunen, T., Grarup, N., Sparsø, T., Doney, A., Voight, B.F., Stringham, H.M., Li, M., Kanoni, S., Shrader, P., Cavalcanti-Proença, C., Kumari, M., Qi, L., Timpson, N.J., Gieger, C., Zabena, C., Rocheleau, G., Ingelsson, E., An, P., O'Connell, J., Luan, J., Elliott, A., McCarroll, S.A., Payne, F., Roccasecca, R.M., Pattou, F., Sethupathy, P., Ardlie, K., Ariyurek, Y., Balkau, B., Barter, P., Beilby, J.P., Ben-Shlomo, Y., Benediktsson, R., Bennett, A.J., Bergmann, S., Bochud, M., Boerwinkle, E., Bonnefond, A., Bonnycastle, L.L., Borch-Johnsen, K., Böttcher, Y., Brunner, E., Bumpstead, S.J., Charpentier, G., Chen, Y.D.I., Chines, P., Clarke, R., Coin, L.J.M., Cooper, M.N., Cornelis, M., Crawford, G., Crisponi, L., Day, I.N.M., de Geus, E.J.C., Delplanque, J., Dina, C., Erdos, M.R., Fedson, A.C., Fischer-Rosinsky, A., Forouhi, N.G., Fox, C.S., Frants, R., Franzosi, M.G., Galan, P., Goodarzi, M.O., Graessler, J., Groves, C.J., Grundy, S., Gwilliam, R., Gyllensten, U., Hadjadj, S., Hallmans, G., Hammond, N., Han, X., Hartikainen, A.L., Hassanali, N., Hayward, C., Heath, S.C., Hercberg, S., Herder, C., Hicks, A.A., Hillman, D.R., Hingorani, A.D., Hofman, A., Hui, J., Hung, J., Isomaa, B., Johnson, P.R.V., Jørgensen, T., Jula, A., Kaakinen, M., Kaprio, J., Kesaniemi, Y.A., Kivimaki, M., Knight, B., Koskinen, S., Kovacs, P., Kyvik, K.O., Lathrop, G.M., Lawlor, D.A., Le Bacquer, O., Lecoeur, C., Li, Y., Lyssenko, V., Mahley, R., Mangino, M., Manning, A.K., Martínez-Larrad, M.T., McAteer, J.B., McCulloch, L.J., McPherson, R., Meisinger, C., Melzer, D., Meyre, D., Mitchell, B.D., Morken, M.A., Mukherjee, S., Naitza, S., Narisu, N., Neville, M.J., Oostra, B.A., Orrù, M., Pakyz, R., Palmer, C.N.A., Paolisso, G., Pattaro, C., Pearson, D., Peden, J.F., Pedersen, N.L., Perola, M., Pfeiffer, A.F.H., Pichler, I., Polasek, O., Posthuma, D., Potter, S.C., Pouta, A., Province, M.A., Psaty, B.M., Rathmann, W., Rayner, N.W., Rice, K., Ripatti, S., Rivadeneira, F., Roden, M., Rolandsson, O., Sandbaek, A., Sandhu, M., Sanna, S., Sayer, A.A., Scheet, P., Scott, L.J., Seedorf, U., Sharp, S.J., Shields, B., Sigurðsson, G., Sijbrands, E.J.G., Silveira, A., Simpson, L., Singleton, A., Smith, N.L., Sovio, U., Swift, A., Syddall, H., Syvänen, A.C., Tanaka, T., Thorand, B., Tichet, J., Tönjes, A., Tuomi, T., Uitterlinden, A.G., van Dijk, K.W., van Hoek, M., Varma, D., Visvikis-Siest, S., Vitart, V., Vogelzangs, N., Waeber, G., Wagner, P.J., Walley, A., Walters, G.B., Ward, K.L., Watkins, H., Weedon, M.N., Wild, S.H., Willemsen, G., Witteman, J.C.M., Yarnell, J.W.G., Zeggini, E., Zelenika, D., Zethelius, B., Zhai, G., Zhao, J.H., Zillikens, M.C., Diagram Consortium, Giant Consortium, Global BPgen Consortium, Borecki, I.B., Loos, R.J.F., Meneton, P., Magnusson, P.K.E., Nathan, D.M., Williams, G.H., Hattersley, A.T., Silander, K., Salomaa, V., Smith, G.D., Bornstein, S.R.,

Schwarz, P., Spranger, J., Karpe, F., Shuldiner, A.R., Cooper, C., Dedoussis, G.V., Serrano-Ríos, M., Morris, A.D., Lind, L., Palmer, L.J., Hu, F.B., Franks, P.W., Ebrahim, S., Marmot, M., Kao, W.H.L., Pankow, J.S., Sampson, M.J., Kuusisto, J., Laakso, M., Hansen, T., Pedersen, O., Pramstaller, P.P., Wichmann, H.E., Illig, T., Rudan, I., Wright, A.F., Stumvoll, M., Campbell, H., Wilson, J.F., Hamsten, A., Procardis Consortium, and The Magic Investigators. New genetic loci implicated in fasting glucose homeostasis and their impact on type 2 diabetes risk. *Nature Genetics* **42**: 105, 2010.

[51] Kong, A., Steinthorsdottir, V., Masson, G., Thorleifsson, G., Sulem, P., Besenbacher, S., Jonasdottir, A., Sigurdsson, A., Kristinsson, K.T., Jonasdottir, A., Frigge, M.L., Gylfason, A., Olason, P.I., Gudjonsson, S.A., Sverrisson, S., Stacey, S.N., Sigurgeirsson, B., Benediktsdottir, K.R., Sigurdsson, H., Jonsson, T., Benediktsson, R., Olafsson, J.H., Johannsson, O.T., Hreidarsson, A.B., Sigurdsson, G., Diagram Consortium, Ferguson-Smith, A.C., Gudbjartsson, D.F., Thorsteinsdottir, U., and Stefansson, K. Parental origin of sequence variants associated with complex diseases. *Nature* **462**: 868, 2009.

[52] Mahajan, A., Go, M.J., Zhang, W., Below, J.E., Gaulton, K.J., Ferreira, T., Horikoshi, M., Johnson, A.D., Ng, M.C.Y., Prokopenko, I., Saleheen, D., Wang, X., Zeggini, E., Abecasis, G.R., Adair, L.S., Almgren, P., Atalay, M., Aung, T., Baldassarre, D., Balkau, B., Bao, Y., Barnett, A.H., Barroso, I., Basit, A., Been, L.F., Beilby, J., Bell, G.I., Benediktsson, R., Bergman, R.N., Boehm, B.O., Boerwinkle, E., Bonnycastle, L.L., Burtt, N., Cai, Q., Campbell, H., Carey, J., Cauchi, S., Caulfield, M., Chan, J.C.N., Chang, L.C., Chang, T.J., Chang, Y.C., Charpentier, G., Chen, C.H., Chen, H., Chen, Y.T., Chia, K.S., Chidambaram, M., Chines, P.S., Cho, N.H., Cho, Y.M., Chuang, L.M., Collins, F.S., Cornelis, M.C., Couper, D.J., Crenshaw, A.T., van Dam, R.M., Danesh, J., Das, D., de Faire, U., Dedoussis, G., Deloukas, P., Dimas, A.S., Dina, C., Doney, A.S.F., Donnelly, P.J., Dorkhan, M., van Duijn, C., Dupuis, J., Edkins, S., Elliott, P., Emilsson, V., Erbel, R., Eriksson, J.G., Escobedo, J., Esko, T., Eury, E., Florez, J.C., Fontanillas, P., Forouhi, N.G., Forsen, T., Fox, C., Fraser, R.M., Frayling, T.M., Froguel, P., Frossard, P., Gao, Y., Gertow, K., Gieger, C., Gigante, B., Grallert, H., Grant, G.B., Groop, L.C., Groves, C.J., Grundberg, E., Guiducci, C., Hamsten, A., Han, B.G., Hara, K., Hassanali, N., Hattersley, A.T., Hayward, C., Hedman, A.K., Herder, C., Hofman, A., Holmen, O.L., Hovingh, K., Hreidarsson, A.B., Hu, C., Hu, F.B., Hui, J., Humphries, S.E., Hunt, S.E., Hunter, D.J., Hveem, K., Hydrie, Z.I., Ikegami, H., Illig, T., Ingelsson, E., Islam, M., Isomaa, B., Jackson, A.U., Jafar, T., James, A., Jia, W., Jockel, K.H., Jonsson, A., Jowett, J.B.M., Kadowaki, T., Kang, H.M., Kanoni, S., Kao, W.H.L., Kathiresan, S., Kato, N., Katulanda, P., Keinanen-Kiukaanniemi, S.M., Kelly, A.M., Khan, H., Khaw, K.T., Khor, C.C., Kim, H.L., Kim, S., Kim, Y.J., Kinnunen, L., Klopp, N., Kong, A., Korpi-Hyovalti, E., Kowlessur, S., Kraft, P., Kravic, J., Kristensen, M.M., Krithika, S., Kumar, A., Kumate, J., Kuusisto, J., Kwak, S.H., Laakso, M., Lagou, V., Lakka, T.A., Langenberg, C., Langford, C., Lawrence, R., Leander, K., Lee, J.M., Lee, N.R., Li, M., Li, X., Li, Y., Liang, J., Liju, S., Lim, W.Y.,

Lind, L., Lindgren, C.M., Lindholm, E., Liu, C.T., Liu, J.J., Lobbens, S., Long, J., Loos, R.J.F., Lu, W., Luan, J., Lyssenko, V., Ma, R.C.W., Maeda, S., Magi, R., Mannisto, S., Matthews, D.R., Meigs, J.B., Melander, O., Metspalu, A., Meyer, J., Mirza, G., Mihailov, E., Moebus, S., Mohan, V., Mohlke, K.L., Morris, A.D., Muhleisen, T.W., Muller-Nurasyid, M., Musk, B., Nakamura, J., Nakashima, E., Navarro, P., Ng, P.K., Nica, A.C., Nilsson, P.M., Njolstad, I., Nothen, M.M., Ohnaka, K., Ong, T.H., Owen, K.R., Palmer, C.N.A., Pankow, J.S., Park, K.S., Parkin, M., Pechlivanis, S., Pedersen, N.L., Peltonen, L., Perry, J.R.B., Peters, A., Pinidiyapathirage, J.M., Platou, C.G.P., Potter, S., Price, J.F., Qi, L., Radha, V., Rallidis, L., Rasheed, A., Rathmann, W., Rauramaa, R., Raychaudhuri, S., Rayner, N.W., Rees, S.D., Rehnberg, E., Ripatti, S., Robertson, N., Roden, M., Rossin, E.J., Rudan, I., Rybin, D., Saaristo, T.E., Salomaa, V., Saltevo, J., Samuel, M., Sanghera, D.K., Saramies, J., Scott, J., Scott, L.J., Scott, R.A., Segre, A.V., Sehmi, J., Sennblad, B., Shah, N., Shah, S., Shera, A.S., Shu, X.O., Shuldiner, A.R., Sigursson, G., Sijbrands, E., Silveira, A., Sim, X., Sivapalaratnam, S., Small, K.S., So, W.Y., Stancakova, A., Stefansson, K., Steinbach, G., Steinthorsdottir, V., Stirrups, K., Strawbridge, R.J., Stringham, H.M., Sun, Q., Suo, C., Syvanen, A.C., Takayanagi, R., Takeuchi, F., Tay, W.T., Teslovich, T.M., Thorand, B., Thorleifsson, G., Thorsteinsdottir, U., Tikkanen, E., Trakalo, J., Tremoli, E., Trip, M.D., Tsai, F.J., Tuomi, T., Tuomilehto, J., Uitterlinden, A.G., Valladares-Salgado, A., Vedantam, S., Veglia, F., Voight, B.F., Wang, C., Wareham, N.J., Wennauer, R., Wickremasinghe, A.R., Wilsgaard, T., Wilson, J.F., Wiltshire, S., Winckler, W., Wong, T.Y., Wood, A.R., Wu, J.Y., Wu, Y., Yamamoto, K., Yamauchi, T., Yang, M., Yengo, L., Yokota, M., Young, R., Zabaneh, D., Zhang, F., Zhang, R., Zheng, W., Zimmet, P.Z., Altshuler, D., Bowden, D.W., Cho, Y.S., Cox, N.J., Cruz, M., Hanis, C.L., Kooner, J., Lee, J.Y., Seielstad, M., Teo, Y.Y., Boehnke, M., Parra, E.J., Chambers, J.C., Tai, E.S., McCarthy, M.I., Morris, A.P., Diabetes Genetics Replication Meta-analysis Consortium, Asian Genetic Epidemiology Network Type 2 Diabetes Consortium, South Asian Type 2 Diabetes Consortium, Mexican American Type 2 Diabetes Consortium, and generation sequencing in multi Ethnic Samples Consortium, T. Genome-wide trans-ancestry meta-analysis provides insight into the genetic architecture of type 2 diabetes susceptibility. *Nat Genet* **46**(3): 234–244, 2014.

[53] Mahajan, A., Wessel, J., Willems, S.M., Zhao, W., Robertson, N.R., Chu, A.Y., Gan, W., Kitajima, H., Taliun, D., Rayner, N.W., Guo, X., Lu, Y., Li, M., Jensen, R.A., Hu, Y., Huo, S., Lohman, K.K., Zhang, W., Cook, J.P., Prins, B.P., Flannick, J., Grarup, N., Trubetskoy, V.V., Kravic, J., Kim, Y.J., Rybin, D.V., Yaghootkar, H., Müller-Nurasyid, M., Meidtner, K., Li-Gao, R., Varga, T.V., Marten, J., Li, J., Smith, A.V., An, P., Ligthart, S., Gustafsson, S., Malerba, G., Demirkan, A., Tajes, J.F., Steinthorsdottir, V., Wuttke, M., Lecoeur, C., Preuss, M., Bielak, L.F., Graff, M., Highland, H.M., Justice, A.E., Liu, D.J., Marouli, E., Peloso, G.M., Warren, H.R., Afaq, S., Afzal, S., Ahlqvist, E., Almgren, P., Amin, N., Bang, L.B., Bertoni, A.G., Bombieri, C., Bork-Jensen, J., Brandslund, I., Brody, J.A., Burtt, N.P., Canouil, M., Chen,

Y.D.I., Cho, Y.S., Christensen, C., Eastwood, S.V., Eckardt, K.U., Fischer, K., Gambaro, G., Giedraitis, V., Grove, M.L., de Haan, H.G., Hackinger, S., Hai, Y., Han, S., Tybjærg-Hansen, A., Hivert, M.F., Isomaa, B., Jäger, S., Jørgensen, M.E., Jørgensen, T., Käräjämäki, A., Kim, B.J., Kim, S.S., Koistinen, H.A., Kovacs, P., Kriebel, J., Kronenberg, F., Läll, K., Lange, L.A., Lee, J.J., Lehne, B., Li, H., Lin, K.H., Linneberg, A., Liu, C.T., Liu, J., Loh, M., Mägi, R., Mamakou, V., McKean-Cowdin, R., Nadkarni, G., Neville, M., Nielsen, S.F., Ntalla, I., Peyser, P.A., Rathmann, W., Rice, K., Rich, S.S., Rode, L., Rolandsson, O., Schönherr, S., Selvin, E., Small, K.S., Stančáková, A., Surendran, P., Taylor, K.D., Teslovich, T.M., Thorand, B., Thorleifsson, G., Tin, A., Tönjes, A., Varbo, A., Witte, D.R., Wood, A.R., Yajnik, P., Yao, J., Yengo, L., Young, R., Amouyel, P., Boeing, H., Boerwinkle, E., Bottinger, E.P., Chowdhury, R., Collins, F.S., Dedoussis, G., Dehghan, A., Deloukas, P., Ferrario, M.M., Ferrières, J., Florez, J.C., Frossard, P., Gudnason, V., Harris, T.B., Heckbert, S.R., Howson, J.M.M., Ingelsson, M., Kathiresan, S., Kee, F., Kuusisto, J., Langenberg, C., Launer, L.J., Lindgren, C.M., Männistö, S., Meitinger, T., Melander, O., Mohlke, K.L., Moitry, M., Morris, A.D., Murray, A.D., de Mutsert, R., Orho-Melander, M., Owen, K.R., Perola, M., Peters, A., Province, M.A., Rasheed, A., Ridker, P.M., Rivadineira, F., Rosendaal, F.R., Rosengren, A.H., Salomaa, V., Sheu, W.H.H., Sladek, R., Smith, B.H., Strauch, K., Uitterlinden, A.G., Varma, R., Willer, C.J., Blüher, M., Butterworth, A.S., Chambers, J.C., Chasman, D.I., Danesh, J., van Duijn, C., Dupuis, J., Franco, O.H., Franks, P.W., Froguel, P., Grallert, H., Groop, L., Han, B.G., Hansen, T., Hattersley, A.T., Hayward, C., Ingelsson, E., Kardia, S.L.R., Karpe, F., Kooner, J.S., Köttgen, A., Kuulasmaa, K., Laakso, M., Lin, X., Lind, L., Liu, Y., Loos, R.J.F., Marchini, J., Metspalu, A., Mook-Kanamori, D., Nordestgaard, B.G., Palmer, C.N.A., Pankow, J.S., Pedersen, O., Psaty, B.M., Rauramaa, R., Sattar, N., Schulze, M.B., Soranzo, N., Spector, T.D., Stefansson, K., Stumvoll, M., Thorsteinsdottir, U., Tuomi, T., Tuomilehto, J., Wareham, N.J., Wilson, J.G., Zeggini, E., Scott, R.A., Barroso, I., Frayling, T.M., Goodarzi, M.O., Meigs, J.B., Boehnke, M., Saleheen, D., Morris, A.P., Rotter, J.I., and McCarthy, M.I. Refining the accuracy of validated target identification through coding variant fine-mapping in type 2 diabetes. *Nature Genetics* **50**(4): 559–571, 2018.

[54] Morris, A.P., Voight, B.F., Teslovich, T.M., Ferreira, T., Segrè, A.V., Steinthorsdottir, V., Strawbridge, R.J., Khan, H., Grallert, H., Mahajan, A., Prokopenko, I., Kang, H.M., Dina, C., Esko, T., Fraser, R.M., Kanoni, S., Kumar, A., Lagou, V., Langenberg, C., Luan, J., Lindgren, C.M., Müller-Nurasyid, M., Pechlivanis, S., Rayner, N.W., Scott, L.J., Wiltshire, S., Yengo, L., Kinnunen, L., Rossin, E.J., Raychaudhuri, S., Johnson, A.D., Dimas, A.S., Loos, R.J.F., Vedantam, S., Chen, H., Florez, J.C., Fox, C., Liu, C.T., Rybin, D., Couper, D.J., Kao, W.H.L., Li, M., Cornelis, M.C., Kraft, P., Sun, Q., van Dam, R.M., Stringham, H.M., Chines, P.S., Fischer, K., Fontanillas, P., Holmen, O.L., Hunt, S.E., Jackson, A.U., Kong, A., Lawrence, R., Meyer, J., Perry, J.R.B., Platou, C.G.P., Potter, S., Rehnberg, E., Robertson, N., Sivapalaratnam, S., Stančáková, A., Stirrups, K., Thorleifsson, G., Tikkanen, E., Wood, A.R., Almgren, P., Atalay, M., Benediktsson, R., Bonnycastle, L.L.,

Burtt, N., Carey, J., Charpentier, G., Crenshaw, A.T., Doney, A.S.F., Dorkhan, M., Edkins, S., Emilsson, V., Eury, E., Forsen, T., Gertow, K., Gigante, B., Grant, G.B., Groves, C.J., Guiducci, C., Herder, C., Hreidarsson, A.B., Hui, J., James, A., Jonsson, A., Rathmann, W., Klopp, N., Kravic, J., Krjutškov, K., Langford, C., Leander, K., Lindholm, E., Lobbens, S., Männistö, S., Mirza, G., Mühleisen, T.W., Musk, B., Parkin, M., Rallidis, L., Saramies, J., Sennblad, B., Shah, S., Sigurõsson, G., Silveira, A., Steinbach, G., Thorand, B., Trakalo, J., Veglia, F., Wennauer, R., Winckler, W., Zabaneh, D., Campbell, H., van Duijn, C., Uitterlinden, A.G., Hofman, A., Sijbrands, E., Abecasis, G.R., Owen, K.R., Zeggini, E., Trip, M.D., Forouhi, N.G., Syvänen, A.C., Eriksson, J.G., Peltonen, L., Nöthen, M.M., Balkau, B., Palmer, C.N.A., Lyssenko, V., Tuomi, T., Isomaa, B., Hunter, D.J., Qi, L., Wellcome Trust Case Control Consortium, Magic Investigators, Giant Consortium, AGEN-T2D Consortium, SAT2D Consortium, Shuldiner, A.R., Roden, M., Barroso, I., Wilsgaard, T., Beilby, J., Hovingh, K., Price, J.F., Wilson, J.F., Rauramaa, R., Lakka, T.A., Lind, L., Dedoussis, G., Njølstad, I., Pedersen, N.L., Khaw, K.T., Wareham, N.J., Keinanen-Kiukaanniemi, S.M., Saaristo, T.E., Korpi-Hyövälti, E., Saltevo, J., Laakso, M., Kuusisto, J., Metspalu, A., Collins, F.S., Mohlke, K.L., Bergman, R.N., Tuomilehto, J., Boehm, B.O., Gieger, C., Hveem, K., Cauchi, S., Froguel, P., Baldassarre, D., Tremoli, E., Humphries, S.E., Saleheen, D., Danesh, J., Ingelsson, E., Ripatti, S., Salomaa, V., Erbel, R., Jöckel, K.H., Moebus, S., Peters, A., Illig, T., de Faire, U., Hamsten, A., Morris, A.D., Donnelly, P.J., Frayling, T.M., Hattersley, A.T., Boerwinkle, E., Melander, O., Kathiresan, S., Nilsson, P.M., Deloukas, P., Thorsteinsdottir, U., Groop, L.C., Stefansson, K., Hu, F., Pankow, J.S., Dupuis, J., Meigs, J.B., Altshuler, D., Boehnke, M., McCarthy, M.I., and The Diagram Consortium. Large-scale association analysis provides insights into the genetic architecture and pathophysiology of type 2 diabetes. *Nature genetics* **44**(9): 981–990, 2012.

[55] Qi, L., Cornelis, M.C., Kraft, P., Stanya, K.J., Linda Kao, W.H., Pankow, J.S., Dupuis, J., Florez, J.C., Fox, C.S., Paré, G., Sun, Q., Girman, C.J., Laurie, C.C., Mirel, D.B., Manolio, T.A., Chasman, D.I., Boerwinkle, E., Ridker, P.M., Hunter, D.J., Meigs, J.B., Lee, C.H., Meta-Analysis of Glucose and Insulin-related traits Consortium, Diabetes Genetics Replication Meta-analysis Consortium, van Dam, R.M., and Hu, F.B. Genetic variants at 2q24 are associated with susceptibility to type 2 diabetes. *Human Molecular Genetics* **19**(13): 2706–2715, 2010.

[56] Rung, J., Cauchi, S., Albrechtsen, A., Shen, L., Rocheleau, G., Cavalcanti-Proenca, C., Bacot, F., Balkau, B., Belisle, A., Borch-Johnsen, K., Charpentier, G., Dina, C., Durand, E., Elliott, P., Hadjadj, S., Jarvelin, M.R., Laitinen, J., Lauritzen, T., Marre, M., Mazur, A., Meyre, D., Montpetit, A., Pisinger, C., Posner, B., Poulsen, P., Pouta, A., Prentki, M., Ribel-Madsen, R., Ruokonen, A., Sandbaek, A., Serre, D., Tichet, J., Vaxillaire, M., Wojtaszewski, J.F.P., Vaag, A., Hansen, T., Polychronakos, C., Pedersen, O., Froguel, P., and Sladek, R. Genetic variant near *IRS1* is associated with type 2 diabetes, insulin resistance and hyperinsulinemia. *Nat Genet* **41**(10): 1110–1115, 2009.

[57] Saxena, R., Elbers, C., Guo, Y., Peter, I., Gaunt, T., Mega, J., Lanktree, M., Tare, A., Castillo, B., Li, Y., Johnson, T., Bruinenberg, M., Gilbert-Diamond, D., Rajagopalan, R., Voight, B., Balasubramanyam, A., Barnard, J., Bauer, F., Baumert, J., Bhangale, T., Böhm, B., Braund, P., Burton, P., Chandrupatla, H., Clarke, R., Cooper-DeHoff, R., Crook, E., Davey-Smith, G., Day, I., de Boer, A., de Groot, M.H., Drenos, F., Ferguson, J., Fox, C., Furlong, C., Gibson, Q., Gieger, C., Gilhuijs-Pederson, L., Glessner, J., Goel, A., Gong, Y., Grant, S.A., Grobbee, D., Hastie, C., Humphries, S., Kim, C., Kivimaki, M., Kleber, M., Meisinger, C., Kumari, M., Langaee, T., Lawlor, D., Li, M., Lobmeyer, M., Maitland-van der Zee, A.H., Meijs, M.L., Molony, C., Morrow, D., Murugesan, G., Musani, S., Nelson, C., Newhouse, S., O'Connell, J., Padmanabhan, S., Palmen, J., Patel, S., Pepine, C., Pettinger, M., Price, T., Rafelt, S., Ranchalis, J., Rasheed, A., Rosenthal, E., Ruczinski, I., Shah, S., Shen, H., Silbernagel, G., Smith, E., Spijkerman, A.M., Stanton, A., Steffes, M., Thorand, B., Trip, M., van der Harst, P., van der A, D., van Iperen, E.A., van Setten, J., van Vliet-Ostaptchouk, J., Verweij, N., Wolffenbuttel, B.R., Young, T., Zafarmand, M.H., Zmuda, J., Boehnke, M., Altshuler, D., McCarthy, M., Kao, W.H.L., Pankow, J., Cappola, T., Sever, P., Poulter, N., Caulfield, M., Dominiczak, A., Shields, D., Bhatt, D.L., Zhang, L., Curtis, S., Danesh, J., Casas, J., van der Schouw, Y., Onland-Moret, N.C., Doevendans, P., Dorn, G.W.I., Farrall, M., FitzGerald, G., Hamsten, A., Hegele, R., Hingorani, A., Hofker, M., Huggins, G., Illig, T., Jarvik, G., Johnson, J., Klungel, O., Knowler, W., Koenig, W., März, W., Meigs, J., Melander, O., Munroe, P., Mitchell, B., Bielinski, S., Rader, D., Reilly, M., Rich, S., Rotter, J., Saleheen, D., Samani, N., Schadt, E., Shuldiner, A., Silverstein, R., Kottke-Marchant, K., Talmud, P., Watkins, H., Asselbergs, F.W., de Bakker, P.W., McCaffery, J., Wijmenga, C., Sabatine, M., Wilson, J., Reiner, A., Bowden, D., Hakonarson, H., Siscovick, D., and Keating, B. Large-Scale Gene-Centric Meta-Analysis across 39 Studies Identifies Type 2 Diabetes Loci. *The American Journal of Human Genetics* **90**(3): 410–425, 2012.

[58] Shu, X.O., Long, J., Cai, Q., Qi, L., Xiang, Y.B., Cho, Y.S., Tai, E.S., Li, X., Lin, X., Chow, W.H., Go, M.J., Seielstad, M., Bao, W., Li, H., Cornelis, M.C., Yu, K., Wen, W., Shi, J., Han, B.G., Sim, X.L., Liu, L., Qi, Q., Kim, H.L., Ng, D.P.K., Lee, J.Y., Kim, Y.J., Li, C., Gao, Y.T., Zheng, W., and Hu, F.B. Identification of New Genetic Risk Variants for Type 2 Diabetes. *PLOS Genetics* **6**(9): e1001127, 2010.

[59] Steinthorsdottir, V., Thorleifsson, G., Sulem, P., Helgason, H., Grarup, N., Sigurdsson, A., Helgadottir, H.T., Johannsdottir, H., Magnusson, O.T., Gudjonsson, S.A., Justesen, J.M., Harder, M.N., Jørgensen, M.E., Christensen, C., Brandslund, I., Sandbæk, A., Lauritzen, T., Vestergaard, H., Linneberg, A., Jørgensen, T., Hansen, T., Daneshpour, M.S., Fallah, M.S., Hreidarsson, A.B., Sigurdsson, G., Azizi, F., Benediktsson, R., Masson, G., Helgason, A., Kong, A., Gudbjartsson, D.F., Pedersen, O., Thorsteinsdottir, U., and Stefansson, K. Identification of low-frequency and rare sequence variants associated with elevated or reduced risk of type 2 diabetes. *Nature Genetics* **46**: 294, 2014.

[60] Voight, B.F., Scott, L.J., Steinthorsdottir, V., Morris, A.P., Dina, C., Welch, R.P., Zeggini, E., Huth, C., Aulchenko, Y.S., Thorleifsson, G., McCulloch, L.J., Ferreira, T., Grallert, H., Amin, N., Wu, G., Willer, C.J., Raychaudhuri, S., McCarroll, S.A., Langenberg, C., Hofmann, O.M., Dupuis, J., Qi, L., Segrè, A.V., van Hoek, M., Navarro, P., Ardlie, K., Balkau, B., Benediktsson, R., Bennett, A.J., Blagieva, R., Boerwinkle, E., Bonnycastle, L.L., Boström, K.B., Bravenboer, B., Bumpstead, S., Burtt, N.P., Charpentier, G., Chines, P.S., Cornelis, M., Couper, D.J., Crawford, G., Doney, A.S.F., Elliott, K.S., Elliott, A.L., Erdos, M.R., Fox, C.S., Franklin, C.S., Ganser, M., Gieger, C., Grarup, N., Green, T., Griffin, S., Groves, C.J., Guiducci, C., Hadjadj, S., Hassanali, N., Herder, C., Isomaa, B., Jackson, A.U., Johnson, P.R.V., Jørgensen, T., Kao, W.H.L., Klopp, N., Kong, A., Kraft, P., Kuusisto, J., Lauritzen, T., Li, M., Lieverse, A., Lindgren, C.M., Lyssenko, V., Marre, M., Meitinger, T., Midthjell, K., Morken, M.A., Narisu, N., Nilsson, P., Owen, K.R., Payne, F., Perry, J.R.B., Petersen, A.K., Platou, C., Proença, C., Prokopenko, I., Rathmann, W., Rayner, N.W., Robertson, N.R., Rocheleau, G., Roden, M., Sampson, M.J., Saxena, R., Shields, B.M., Shrader, P., Sigurdsson, G., Sparsø, T., Strassburger, K., Stringham, H.M., Sun, Q., Swift, A.J., Thorand, B., Tichet, J., Tuomi, T., van Dam, R.M., van Haeften, T.W., van Herpt, T., van Vliet-Ostaptchouk, J.V., Walters, G.B., Weedon, M.N., Wijmenga, C., Witteman, J., The Magic Investigators, The Giant Consortium, Bergman, R.N., Cauchi, S., Collins, F.S., Gloyn, A.L., Gyllensten, U., Hansen, T., Hide, W.A., Hitman, G.A., Hofman, A., Hunter, D.J., Hveem, K., Laakso, M., Mohlke, K.L., Morris, A.D., Palmer, C.N.A., Pramstaller, P.P., Rudan, I., Sijbrands, E., Stein, L.D., Tuomilehto, J., Uitterlinden, A., Walker, M., Wareham, N.J., Watanabe, R.M., Abecasis, G.R., Boehm, B.O., Campbell, H., Daly, M.J., Hattersley, A.T., Hu, F.B., Meigs, J.B., Pankow, J.S., Pedersen, O., Wichmann, H.E., Barroso, I., Florez, J.C., Frayling, T.M., Groop, L., Sladek, R., Thorsteinsdottir, U., Wilson, J.F., Illig, T., Froguel, P., van Duijn, C.M., Stefansson, K., Altshuler, D., Boehnke, M., and McCarthy, M.I. Twelve type 2 diabetes susceptibility loci identified through large-scale association analysis. *Nature Genetics* **42**: 579, 2010.

[61] Zeggini, E., Scott, L.J., Saxena, R., Voight, B.F., Marchini, J.L., Hu, T., de Bakker, P.I.W., Abecasis, G.R., Almgren, P., Andersen, G., Ardlie, K., Boström, K.B., Bergman, R.N., Bonnycastle, L.L., Borch-Johnsen, K., Burtt, N.P., Chen, H., Chines, P.S., Daly, M.J., Deodhar, P., Ding, C.J., Doney, A.S.F., Duren, W.L., Elliott, K.S., Erdos, M.R., Frayling, T.M., Freathy, R.M., Gianniny, L., Grallert, H., Grarup, N., Groves, C.J., Guiducci, C., Hansen, T., Herder, C., Hitman, G.A., Hughes, T.E., Isomaa, B., Jackson, A.U., Jørgensen, T., Kong, A., Kubalanza, K., Kuruvilla, F.G., Kuusisto, J., Langenberg, C., Lango, H., Lauritzen, T., Li, Y., Lindgren, C.M., Lyssenko, V., Marvelle, A.F., Meisinger, C., Midthjell, K., Mohlke, K.L., Morken, M.A., Morris, A.D., Narisu, N., Nilsson, P., Owen, K.R., Palmer, C.N.A., Payne, F., Perry, J.R.B., Pettersen, E., Platou, C., Prokopenko, I., Qi, L., Qin, L., Rayner, N.W., Rees, M., Roix, J.J., Sandbæk, A., Shields, B., Sjögren, M., Steinthorsdottir, V., Stringham, H.M., Swift, A.J., Thorleifsson, G., Thorsteinsdottir, U., Timpson,

N.J., Tuomi, T., Tuomilehto, J., Walker, M., Watanabe, R.M., Weedon, M.N., Willer, C.J., Wellcome Trust Case Control Consortium, Illig, T., Hveem, K., Hu, F.B., Laakso, M., Stefansson, K., Pedersen, O., Wareham, N.J., Barroso, I., Hattersley, A.T., Collins, F.S., Groop, L., McCarthy, M.I., Boehnke, M., and Altshuler, D. Meta-analysis of genome-wide association data and large-scale replication identifies additional susceptibility loci for type 2 diabetes. *Nature Genetics* **40**: 638, 2008.

[62] Cho, Y.S., Chen, C.H., Hu, C., Long, J., Hee Ong, R.T., Sim, X., Takeuchi, F., Wu, Y., Go, M.J., Yamauchi, T., Chang, Y.C., Kwak, S.H., Ma, R.C.W., Yamamoto, K., Adair, L.S., Aung, T., Cai, Q., Chang, L.C., Chen, Y.T., Gao, Y., Hu, F.B., Kim, H.L., Kim, S., Kim, Y.J., Lee, J.J.M., Lee, N.R., Li, Y., Liu, J.J., Lu, W., Nakamura, J., Nakashima, E., Ng, D.P.K., Tay, W.T., Tsai, F.J., Wong, T.Y., Yokota, M., Zheng, W., Zhang, R., Wang, C., So, W.Y., Ohnaka, K., Ikegami, H., Hara, K., Cho, Y.M., Cho, N.H., Chang, T.J., Bao, Y., Hedman, A., Morris, A.P., McCarthy, M.I., Diagram Consortium, MuTher Consortium, Takayanagi, R., Park, K.S., Jia, W., Chuang, L.M., Chan, J.C.N., Maeda, S., Kadowaki, T., Lee, J.Y., Wu, J.Y., Teo, Y.Y., Tai, E.S., Shu, X.O., Mohlke, K.L., Kato, N., Han, B.G., and Seielstad, M. Meta-analysis of genome-wide association studies identifies eight new loci for type 2 diabetes in east Asians. *Nature Genetics* **44**: 67, 2011.

[63] Hanson, R.L., Muller, Y.L., Kobes, S., Guo, T., Bian, L., Ossowski, V., Wiedrich, K., Sutherland, J., Wiedrich, C., Mahkee, D., Huang, K., Abdussamad, M., Traurig, M., Weil, E.J., Nelson, R.G., Bennett, P.H., Knowler, W.C., Bogardus, C., and Baier, L.J. A Genome-Wide Association Study in American Indians Implicates *DNER* as a Susceptibility Locus for Type 2 Diabetes. *Diabetes* **63**(1): 369–376, 2014.

[64] Hara, K., Fujita, H., Johnson, T.A., Yamauchi, T., Yasuda, K., Horikoshi, M., Peng, C., Hu, C., Ma, R.C.W., Imamura, M., Iwata, M., Tsunoda, T., Morizono, T., Shojima, N., So, W.Y., Leung, T.F., Kwan, P., Zhang, R., Wang, J., Yu, W., Maegawa, H., Hirose, H., Diagram consortium, Kaku, K., Ito, C., Watada, H., Tanaka, Y., Tobe, K., Kashiwagi, A., Kawamori, R., Jia, W., Chan, J.C.N., Teo, Y.Y., Shyong, T.E., Kamatani, N., Kubo, M., Maeda, S., and Kadowaki, T. Genome-wide association study identifies three novel loci for type 2 diabetes. *Human Molecular Genetics* **23**(1): 239–246, 2014.

[65] Imamura, M., Maeda, S., Yamauchi, T., Hara, K., Yasuda, K., Morizono, T., Takahashi, A., Horikoshi, M., Nakamura, M., Fujita, H., Tsunoda, T., Kubo, M., Watada, H., Maegawa, H., Okada-Iwabu, M., Iwabu, M., Shojima, N., Ohshige, T., Omori, S., Iwata, M., Hirose, H., Kaku, K., Ito, C., Tanaka, Y., Tobe, K., Kashiwagi, A., Kawamori, R., Kasuga, M., Kamatani, N., Diabetes Genetics Replication Meta-analysis Consortium, Nakamura, Y., and Kadowaki, T. A single-nucleotide polymorphism in *ANK1* is associated with susceptibility to type 2 diabetes in Japanese populations. *Human Molecular Genetics* **21**(13): 3042–3049, 2012.

[66] Imamura, M., Takahashi, A., Yamauchi, T., Hara, K., Yasuda, K., Grarup, N.,

Zhao, W., Wang, X., Huerta-Chagoya, A., Hu, C., Moon, S., Long, J., Kwak, S.H., Rasheed, A., Saxena, R., Ma, R.C.W., Okada, Y., Iwata, M., Hosoe, J., Shojima, N., Iwasaki, M., Fujita, H., Suzuki, K., Danesh, J., Jørgensen, T., Jørgensen, M.E., Witte, D.R., Brandslund, I., Christensen, C., Hansen, T., Mercader, J.M., Flannick, J., Moreno-Macías, H., Burtt, N.P., Zhang, R., Kim, Y.J., Zheng, W., Singh, J.R., Tam, C.H.T., Hirose, H., Maegawa, H., Ito, C., Kaku, K., Watada, H., Tanaka, Y., Tobe, K., Kawamori, R., Kubo, M., Cho, Y.S., Chan, J.C.N., Sanghera, D., Frossard, P., Park, K.S., Shu, X.O., Kim, B.J., Florez, J.C., Tusié-Luna, T., Jia, W., Tai, E.S., Pedersen, O., Saleheen, D., Maeda, S., and Kadowaki, T. Genome-wide association studies in the Japanese population identify seven novel loci for type 2 diabetes. *Nature Communications* **7**: 10531, 2016.

[67] Kooner, J.S., Saleheen, D., Sim, X., Sehmi, J., Zhang, W., Frossard, P., Been, L.F., Chia, K.S., Dimas, A.S., Hassanali, N., Jafar, T., Jowett, J.B.M., Li, X., Radha, V., Rees, S.D., Takeuchi, F., Young, R., Aung, T., Basit, A., Chidambaram, M., Das, D., Grundberg, E., Hedman, r.K., Hydrie, Z.I., Islam, M., Khor, C.C., Kowlessur, S., Kristensen, M.M., Liju, S., Lim, W.Y., Matthews, D.R., Liu, J., Morris, A.P., Nica, A.C., Pinidiyapathirage, J.M., Prokopenko, I., Rasheed, A., Samuel, M., Shah, N., Shera, A.S., Small, K.S., Suo, C., Wickremasinghe, A.R., Wong, T.Y., Yang, M., Zhang, F., Diagram Consortium, MuTher Consortium, Abecasis, G.R., Barnett, A.H., Caulfield, M., Deloukas, P., Frayling, T.M., Froguel, P., Kato, N., Katulanda, P., Kelly, M.A., Liang, J., Mohan, V., Sanghera, D.K., Scott, J., Seielstad, M., Zimmet, P.Z., Elliott, P., Teo, Y.Y., McCarthy, M.I., Danesh, J., Tai, E.S., and Chambers, J.C. Genome-wide association study in individuals of South Asian ancestry identifies six new type 2 diabetes susceptibility loci. *Nature Genetics* **43**: 984, 2011.

[68] Li, H., Gan, W., Lu, L., Dong, X., Han, X., Hu, C., Yang, Z., Sun, L., Bao, W., Li, P., He, M., Sun, L., Wang, Y., Zhu, J., Ning, Q., Tang, Y., Zhang, R., Wen, J., Wang, D., Zhu, X., Guo, K., Zuo, X., Guo, X., Yang, H., Zhou, X., Zhang, X., Qi, L., Loos, R.J., Hu, F.B., Wu, T., Liu, Y., Liu, L., Yang, Z., Hu, R., Jia, W., Ji, L., Li, Y., and Lin, X. A Genome-Wide Association Study Identifies *GRK5* and *RASGRP1* as Type 2 Diabetes Loci in Chinese Hans. *Diabetes* **62**(1): 291–298, 2013.

[69] Ma, R.C.W., Hu, C., Tam, C.H., Zhang, R., Kwan, P., Leung, T.F., Thomas, G.N., Go, M.J., Hara, K., Sim, X., Ho, J.S.K., Wang, C., Li, H., Lu, L., Wang, Y., Li, J.W., Wang, Y., Lam, V.K.L., Wang, J., Yu, W., Kim, Y.J., Ng, D.P., Fujita, H., Panoutsopoulou, K., Day-Williams, A.G., Lee, H.M., Ng, A.C.W., Fang, Y.J., Kong, A.P.S., Jiang, F., Ma, X., Hou, X., Tang, S., Lu, J., Yamauchi, T., Tsui, S.K.W., Woo, J., Leung, P.C., Zhang, X., Tang, N.L.S., Sy, H.Y., Liu, J., Wong, T.Y., Lee, J.Y., Maeda, S., Xu, G., Cherny, S.S., Chan, T.F., Ng, M.C.Y., Xiang, K., Morris, A.P., Keildson, S., Hu, R., Ji, L., Lin, X., Cho, Y.S., Kadowaki, T., Tai, E.S., Zeggini, E., McCarthy, M.I., Hon, K.L., Baum, L., Tomlinson, B., So, W.Y., Bao, Y., Chan, J.C.N., and Jia, W. Genome-wide association study in a Chinese population identifies a

susceptibility locus for type 2 diabetes at 7q32 near *PAX4*. *Diabetologia* **56**(6): 1291–1305, 2013.

[70] Moltke, I., Grarup, N., Jørgensen, M.E., Bjerregaard, P., Treebak, J.T., Fumagalli, M., Korneliussen, T.S., Andersen, M.A., Nielsen, T.S., Krarup, N.T., Gjesing, A.P., Zierath, J.R., Linneberg, A., Wu, X., Sun, G., Jin, X., Al-Aama, J., Wang, J., Borch-Johnsen, K., Pedersen, O., Nielsen, R., Albrechtsen, A., and Hansen, T. A common Greenlandic *TBC1D4* variant confers muscle insulin resistance and type 2 diabetes. *Nature* **512**: 190, 2014.

[71] Ng, M.C.Y., Shriner, D., Chen, B.H., Li, J., Chen, W.M., Guo, X., Liu, J., Bielinski, S.J., Yanek, L.R., Nalls, M.A., Comeau, M.E., Rasmussen-Torvik, L.J., Jensen, R.A., Evans, D.S., Sun, Y.V., An, P., Patel, S.R., Lu, Y., Long, J., Armstrong, L.L., Wagenknecht, L., Yang, L., Snively, B.M., Palmer, N.D., Mudgal, P., Langefeld, C.D., Keene, K.L., Freedman, B.I., Mychaleckyj, J.C., Nayak, U., Raffel, L.J., Goodarzi, M.O., Chen, Y.D.I., Taylor, Herman A., J., Correa, A., Sims, M., Couper, D., Pankow, J.S., Boerwinkle, E., Adeyemo, A., Doumatey, A., Chen, G., Mathias, R.A., Vaidya, D., Singleton, A.B., Zonderman, A.B., Igo, Robert P., J., Sedor, J.R., The Find Consortium, Kabagambe, E.K., Siscovick, D.S., McKnight, B., Rice, K., Liu, Y., Hsueh, W.C., Zhao, W., Bielak, L.F., Kraja, A., Province, M.A., Bottinger, E.P., Gottesman, O., Cai, Q., Zheng, W., Blot, W.J., Lowe, W.L., Pacheco, J.A., Crawford, D.C., The eMERGE Consortium, The DIAGRAM Consortium, Grundberg, E., The MuTHER Consortium, Rich, S.S., Hayes, M.G., Shu, X.O., Loos, R.J.F., Borecki, I.B., Peyser, P.A., Cummings, S.R., Psaty, B.M., Fornage, M., Iyengar, S.K., Evans, M.K., Becker, D.M., Kao, W.H.L., Wilson, J.G., Rotter, J.I., Sale, M.M., Liu, S., Rotimi, C.N., Bowden, D.W., and The MEta-analysis of type DIabetes in African Americans Consortium. Meta-Analysis of Genome-Wide Association Studies in African Americans Provides Insights into the Genetic Architecture of Type 2 Diabetes. *PLOS Genetics* **10**(8): e1004517, 2014.

[72] Palmer, N.D., McDonough, C.W., Hicks, P.J., Roh, B.H., Wing, M.R., An, S.S., Hester, J.M., Cooke, J.N., Bostrom, M.A., Rudock, M.E., Talbert, M.E., Lewis, J.P., Diagram Consortium, Magic Investigators, Ferrara, A., Lu, L., Ziegler, J.T., Sale, M.M., Divers, J., Shriner, D., Adeyemo, A., Rotimi, C.N., Ng, M.C.Y., Langefeld, C.D., Freedman, B.I., and Bowden, D.W. A Genome-Wide Association Search for Type 2 Diabetes Genes in African Americans. *PLOS ONE* **7**(1): e29202, 2012.

[73] Perry, J.R.B., Voight, B.F., Yengo, L., Amin, N., Dupuis, J., Ganser, M., Grallert, H., Navarro, P., Li, M., Qi, L., Steinthorsdottir, V., Scott, R.A., Almgren, P., Arking, D.E., Aulchenko, Y., Balkau, B., Benediktsson, R., Bergman, R.N., Boerwinkle, E., Bonnycastle, L., Burtt, N.P., Campbell, H., Charpentier, G., Collins, F.S., Gieger, C., Green, T., Hadjadj, S., Hattersley, A.T., Herder, C., Hofman, A., Johnson, A.D., Kottgen, A., Kraft, P., Labrune, Y., Langenberg, C., Manning, A.K., Mohlke, K.L., Morris, A.P., Oostra, B., Pankow, J., Petersen, A.K., Pramstaller, P.P., Prokopenko, I., Rathmann, W., Rayner, W., Roden, M., Rudan, I., Rybin, D., Scott, L.J., Sigurdsson, G.,

Sladek, R., Thorleifsson, G., Thorsteinsdottir, U., Tuomilehto, J., Uitterlinden, A.G., Vivequin, S., Weedon, M.N., Wright, A.F., Magic Investigators, Diagram Consortium, Giant Consortium, Hu, F.B., Illig, T., Kao, L., Meigs, J.B., Wilson, J.F., Stefansson, K., van Duijn, C., Altschuler, D., Morris, A.D., Boehnke, M., McCarthy, M.I., Froguel, P., Palmer, C.N.A., Wareham, N.J., Groop, L., Frayling, T.M., and Cauchi, S. Stratifying Type 2 Diabetes Cases by BMI Identifies Genetic Risk Variants in *LAMA1* and Enrichment for Risk Variants in Lean Compared to Obese Cases. *PLOS Genetics* **8**(5): e1002741, 2012.

[74] Saxena, R., Saleheen, D., Been, L.F., Garavito, M.L., Braun, T., Bjonnes, A., Young, R., Ho, W.K., Rasheed, A., Frossard, P., Sim, X., Hassanali, N., Radha, V., Chidambaram, M., Liju, S., Rees, S.D., Ng, D.P.K., Wong, T.Y., Yamauchi, T., Hara, K., Tanaka, Y., Hirose, H., McCarthy, M.I., Morris, A.P., Basit, A., Barnett, A.H., Katulanda, P., Matthews, D., Mohan, V., Wander, G.S., Singh, J.R., Mehra, N.K., Ralhan, S., Kamboh, M.I., Mulvihill, J.J., Maegawa, H., Tobe, K., Maeda, S., Cho, Y.S., Tai, E.S., Kelly, M.A., Chambers, J.C., Kooner, J.S., Kadowaki, T., Deloukas, P., Rader, D.J., Danesh, J., and Sanghera, D.K. Genome-Wide Association Study Identifies a Novel Locus Contributing to Type 2 Diabetes Susceptibility in Sikhs of Punjabi Origin From India. *Diabetes* **62**(5): 1746–1755, 2013.

[75] Tabassum, R., Chauhan, G., Dwivedi, O.P., Mahajan, A., Jaiswal, A., Kaur, I., Bandesh, K., Singh, T., Mathai, B.J., Pandey, Y., Chidambaram, M., Sharma, A., Chavali, S., Sengupta, S., Ramakrishnan, L., Venkatesh, P., Aggarwal, S.K., Ghosh, S., Prabhakaran, D., Srinath, R.K., Saxena, M., Banerjee, M., Mathur, S., Bhansali, A., Shah, V.N., Madhu, S.V., Marwaha, R.K., Basu, A., Scaria, V., McCarthy, M.I., DIAGRAM, INDICO, Venkatesan, R., Mohan, V., Tandon, N., and Bharadwaj, D. Genome-Wide Association Study for Type 2 Diabetes in Indians Identifies a New Susceptibility Locus at 2q21. *Diabetes* **62**(3): 977–986, 2013.

[76] The Sigma Type Diabetes Consortium. Association of a low-frequency variant in *HNF1A* with type 2 diabetes in a latino population. *JAMA* **311**(22): 2305–2314, 2014.

[77] Tsai, F.J., Yang, C.F., Chen, C.C., Chuang, L.M., Lu, C.H., Chang, C.T., Wang, T.Y., Chen, R.H., Shiu, C.F., Liu, Y.M., Chang, C.C., Chen, P., Chen, C.H., Fann, C.S.J., Chen, Y.T., and Wu, J.Y. A Genome-Wide Association Study Identifies Susceptibility Variants for Type 2 Diabetes in Han Chinese. *PLOS Genetics* **6**(2): e1000847, 2010.

[78] Yamauchi, T., Hara, K., Maeda, S., Yasuda, K., Takahashi, A., Horikoshi, M., Nakamura, M., Fujita, H., Grarup, N., Cauchi, S., Ng, D.P.K., Ma, R.C.W., Tsunoda, T., Kubo, M., Watada, H., Maegawa, H., Okada-Iwabu, M., Iwabu, M., Shojima, N., Shin, H.D., Andersen, G., Witte, D.R., Jørgensen, T., Lauritzen, T., Sandbæk, A., Hansen, T., Ohshige, T., Omori, S., Saito, I., Kaku, K., Hirose, H., So, W.Y., Beury, D., Chan, J.C.N., Park, K.S., Tai, E.S., Ito, C., Tanaka, Y., Kashiwagi, A., Kawamori, R., Kasuga, M., Froguel, P., Pedersen,

O., Kamatani, N., Nakamura, Y., and Kadowaki, T. A genome-wide association study in the Japanese population identifies susceptibility loci for type 2 diabetes at *UBE2E2* and *C2CD4A-C2CD4B*. *Nature Genetics* **42**: 864, 2010.

[79] Zhao, W., Rasheed, A., Tikkanen, E., Lee, J.J., Butterworth, A.S., Howson, J.M.M., Assimes, T.L., Chowdhury, R., Orho-Melander, M., Damrauer, S., Small, A., Asma, S., Imamura, M., Yamauch, T., Chambers, J.C., Chen, P., Sapkota, B.R., Shah, N., Jabeen, S., Surendran, P., Lu, Y., Zhang, W., Imran, A., Abbas, S., Majeed, F., Trindade, K., Qamar, N., Mallick, N.H., Yaqoob, Z., Saghir, T., Rizvi, S.N.H., Memon, A., Rasheed, S.Z., Memon, F.u.R., Mehmood, K., Ahmed, N., Qureshi, I.H., Tanveer us, S., Iqbal, W., Malik, U., Mehra, N., Kuo, J.Z., Sheu, W.H.H., Guo, X., Hsiung, C.A., Juang, J.M.J., Taylor, K.D., Hung, Y.J., Lee, W.J., Quertermous, T., Lee, I.T., Hsu, C.C., Bottinger, E.P., Ralhan, S., Teo, Y.Y., Wang, T.D., Alam, D.S., Di Angelantonio, E., Epstein, S., Nielsen, S.F., Nordestgaard, B.G., Tybjaerg-Hansen, A., Young, R., CHD Exome+ Consortium, Benn, M., Frikke-Schmidt, R., Kamstrup, P.R., EPIC-CVD Consortium, EPIC-Interact Consortium, Michigan Biobank, Jukema, J.W., Sattar, N., Smit, R., Chung, R.H., Liang, K.W., Anand, S., Sanghera, D.K., Ripatti, S., Loos, R.J.F., Kooner, J.S., Tai, E.S., Rotter, J.I., Chen, Y.D.I., Frossard, P., Maeda, S., Kadowaki, T., Reilly, M., Pare, G., Melander, O., Salomaa, V., Rader, D.J., Danesh, J., Voight, B.F., and Saleheen, D. Identification of new susceptibility loci for type 2 diabetes and shared etiological pathways with coronary heart disease. *Nature Genetics* **49**: 1450, 2017.

[80] Ali, O. Genetics of type 2 diabetes. *World Journal of Diabetes* **4**(4): 114–123, 2013.

[81] Meigs, J.B., Cupples, L.A., and Wilson, P.W. Parental transmission of type 2 diabetes: the Framingham Offspring Study. *Diabetes* **49**(12): 2201–2207, 2000.

[82] Willemsen, G., Ward, K.J., Bell, C.G., Christensen, K., Bowden, J., Dalgård, C., Harris, J.R., Kaprio, J., Lyle, R., Magnusson, P.K.E., Mather, K.A., Ordoňana, J.R., Perez-Riquelme, F., Pedersen, N.L., Pietiläinen, K.H., Sachdev, P.S., Boomsma, D.I., and Spector, T. The Concordance and Heritability of Type 2 Diabetes in 34,166 Twin Pairs From International Twin Registers: The Discordant Twin (DISCOTWIN) Consortium. *Twin Research and Human Genetics* **18**(6): 762–771, 2015.

[83] McClellan, J. and King, M.C. Genetic Heterogeneity in Human Disease. *Cell* **141**(2): 210–217, 2010.

[84] Prasad, R.B. and Groop, L. Genetics of Type 2 Diabetes — Pitfalls and Possibilities. *Genes* **6**(1): 87–123, 2015.

[85] Dimas, A.S., Lagou, V., Barker, A., Knowles, J.W., Mägi, R., Hivert, M.F., Benazzo, A., Rybin, D., Jackson, A.U., Stringham, H.M., Song, C., Fischer-Rosinsky, A., Boesgaard, T.W., Grarup, N., Abbasi, F.A., Assimes, T.L., Hao, K., Yang, X., Lecoeur, C., Barroso, I., Bonnycastle, L.L., Böttcher, Y., Bumpstead, S., Chines, P.S., Erdos, M.R., Graessler, J., Kovacs, P., Morken, M.A., Narisu, N., Payne, F., Stancakova, A., Swift, A.J., Tönjes, A., Bornstein,

S.R., Cauchi, S., Froguel, P., Meyre, D., Schwarz, P.E., Häring, H.U., Smith, U., Boehnke, M., Bergman, R.N., Collins, F.S., Mohlke, K.L., Tuomilehto, J., Quertemous, T., Lind, L., Hansen, T., Pedersen, O., Walker, M., Pfeiffer, A.F., Spranger, J., Stumvoll, M., Meigs, J.B., Wareham, N.J., Kuusisto, J., Laakso, M., Langenberg, C., Dupuis, J., Watanabe, R.M., Florez, J.C., Ingelsson, E., McCarthy, M.I., and Prokopenko, I. Impact of Type 2 Diabetes Susceptibility Variants on Quantitative Glycemic Traits Reveals Mechanistic Heterogeneity. *Diabetes* **63**(6): 2158–2171, 2014.

[86] Wood, A.R., Jonsson, A., Jackson, A.U., Wang, N., van Leewen, N., Palmer, N.D., Kobes, S., Deelen, J., Boquete-Vilarino, L., Paananen, J., Stančáková, A., Boomsma, D.I., de Geus, E.J., Eekhoff, E.M., Fritsche, A., Kramer, M., Nijpels, G., Simonis-Bik, A., van Haeften, T.W., Mahajan, A., Boehnke, M., Bergman, R.N., Tuomilehto, J., Collins, F.S., Mohlke, K.L., Banasik, K., Groves, C.J., McCarthy, M.I., Pearson, E.R., Natali, A., Mari, A., Buchanan, T.A., Taylor, K.D., Xiang, A.H., Gjesing, A.P., Grarup, N., Eiberg, H., Pedersen, O., Chen, Y.D., Laakso, M., Norris, J.M., Smith, U., Wagenknecht, L.E., Baier, L., Bowden, D.W., Hansen, T., Walker, M., Watanabe, R.M., 't Hart, L.M., Hanson, R.L., and Frayling, T.M. A Genome-Wide Association Study of IVGTT-Based Measures of First Phase Insulin Secretion Refines the Underlying Physiology of Type 2 Diabetes Variants. *Diabetes* , 2017.

[87] Bernstein, B.E., Stamatoyannopoulos, J.A., Costello, J.F., Ren, B., Milosavljevic, A., Meissner, A., Kellis, M., Marra, M.A., Beaudet, A.L., Ecker, J.R., Farnham, P.J., Hirst, M., Lander, E.S., Mikkelsen, T.S., and Thomson, J.A. The NIH Roadmap Epigenomics Mapping Consortium. *Nature Biotechnology* **28**: 1045, 2010.

[88] The Encode Project Consortium. An integrated encyclopedia of DNA elements in the human genome. *Nature* **489**: 57, 2012.

[89] GTEx Consortium, Battle, A., Brown, C.D., Engelhardt, B.E., and Montgomery, S.B. Genetic effects on gene expression across human tissues. *Nature* **550**(7675): 204–213, 2017.

[90] Ionescu-Tirgoviste, C., Gagniuc, P.A., Gubceac, E., Mardare, L., Popescu, I., Dima, S., and Militaru, M. A 3D map of the islet routes throughout the healthy human pancreas. *Scientific Reports* **5**: 14634, 2015.

[91] Gaulton, K.J., Nammo, T., Pasquali, L., Simon, J.M., Giresi, P.G., Fogarty, M.P., Panhuis, T.M., Mieczkowski, P., Secchi, A., Bosco, D., Berney, T., Montanya, E., Mohlke, K.L., Lieb, J.D., and Ferrer, J. A map of open chromatin in human pancreatic islets. *Nature Genetics* **42**: 255, 2010.

[92] Pasquali, L., Gaulton, K.J., Rodríguez-Seguí, S.A., Mularoni, L., Miguel-Escalada, I., Akerman, I., Tena, J.J., Morán, I., Gómez-Marín, C., van de Bunt, M., Ponsa-Cobas, J., Castro, N., Nammo, T., Cebola, I., García-Hurtado, J., Maestro, M.A., Pattou, F., Piemonti, L., Berney, T., Gloyn, A.L., Ravassard, P., Skarmeta, J.L.G., Müller, F., McCarthy, M.I., and Ferrer, J. Pancreatic islet enhancer clusters enriched in type 2 diabetes risk-associated variants. *Nature Genetics* **46**: 136, 2014.

[93] Gaulton, K.J., Ferreira, T., Lee, Y., Raimondo, A., Mägi, R., Reschen, M.E., Mahajan, A., Locke, A., William Rayner, N., Robertson, N., Scott, R.A., Prokopenko, I., Scott, L.J., Green, T., Sparso, T., Thuillier, D., Yengo, L., Grallert, H., Wahl, S., Frånberg, M., Strawbridge, R.J., Kestler, H., Chheda, H., Eisele, L., Gustafsson, S., Steinthorsdottir, V., Thorleifsson, G., Qi, L., Karssen, L.C., van Leeuwen, E.M., Willems, S.M., Li, M., Chen, H., Fuchsberger, C., Kwan, P., Ma, C., Linderman, M., Lu, Y., Thomsen, S.K., Rundle, J.K., Beer, N.L., van de Bunt, M., Chalisey, A., Kang, H.M., Voight, B.F., Abecasis, G.R., Almgren, P., Baldassarre, D., Balkau, B., Benediktsson, R., Blüher, M., Boeing, H., Bonnycastle, L.L., Bottinger, E.P., Burtt, N.P., Carey, J., Charpentier, G., Chines, P.S., Cornelis, M.C., Couper, D.J., Crenshaw, A.T., van Dam, R.M., Doney, A.S.F., Dorkhan, M., Edkins, S., Eriksson, J.G., Esko, T., Eury, E., Fadista, J., Flannick, J., Fontanillas, P., Fox, C., Franks, P.W., Gertow, K., Gieger, C., Gigante, B., Gottesman, O., Grant, G.B., Grarup, N., Groves, C.J., Hassinen, M., Have, C.T., Herder, C., Holmen, O.L., Hreidarsson, A.B., Humphries, S.E., Hunter, D.J., Jackson, A.U., Jonsson, A., Jørgensen, M.E., Jørgensen, T., Kao, W.H.L., Kerrison, N.D., Kinnunen, L., Klopp, N., Kong, A., Kovacs, P., Kraft, P., Kravic, J., Langford, C., Leander, K., Liang, L., Lichtner, P., Lindgren, C.M., Lindholm, E., Linneberg, A., Liu, C.T., Lobbens, S., Luan, J., Lyssenko, V., Männistö, S., McLeod, O., Meyer, J., Mihailov, E., Mirza, G., Mühleisen, T.W., Müller-Nurasyid, M., Navarro, C., Nöthen, M.M., Oskolkov, N.N., Owen, K.R., Palli, D., Pechlivanis, S., Peltonen, L., Perry, J.R.B., Platou, C.G.P., Roden, M., Ruderfer, D., Rybin, D., van der Schouw, Y.T., Sennblad, B., Sigurðsson, G., Stančáková, A., Steinbach, G., Storm, P., Strauch, K., Stringham, H.M., Sun, Q., Thorand, B., Tikkanen, E., Tonjes, A., Trakalo, J., Tremoli, E., Tuomi, T., Wennauer, R., Wiltshire, S., Wood, A.R., Zeggini, E., Dunham, I., Birney, E., Pasquali, L., Ferrer, J., Loos, R.J.F., Dupuis, J., Florez, J.C., Boerwinkle, E., Pankow, J.S., van Duijn, C., Sijbrands, E., Meigs, J.B., Hu, F.B., Thorsteinsdottir, U., Stefansson, K., Lakka, T.A., Rauramaa, R., Stumvoll, M., Pedersen, N.L., Lind, L., Keinanen-Kiukaanniemi, S.M., Korpi-Hyövälti, E., Saaristo, T.E., Saltevo, J., Kuusisto, J., Laakso, M., Metspalu, A., Erbel, R., Jöcke, K.H., Moebus, S., Ripatti, S., Salomaa, V., Ingelsson, E., Boehm, B.O., Bergman, R.N., Collins, F.S., Mohlke, K.L., Koistinen, H., Tuomilehto, J., Hveem, K., Njølstad, I., Deloukas, P., Donnelly, P.J., Frayling, T.M., Hattersley, A.T., de Faire, U., Hamsten, A., Illig, T., Peters, A., Cauchi, S., Sladek, R., Froguel, P., Hansen, T., Pedersen, O., Morris, A.D., Palmer, C.N.A., Kathiresan, S., Melander, O., Nilsson, P.M., Groop, L.C., Barroso, I., Langenberg, C., Wareham, N.J., O'Callaghan, C.A., Gloyn, A.L., Altshuler, D., Boehnke, M., Teslovich, T.M., McCarthy, M.I., Morris, A.P., and The DIAbetes Genetics Replication Meta-analysis Consortium. Genetic fine mapping and genomic annotation defines causal mechanisms at type 2 diabetes susceptibility loci. *Nature Genetics* **47**: 1415, 2015.

[94] Thurner, M., van de Bunt, M., Torres, J.M., Mahajan, A., Nylander, V., Bennett, A.J., Gaulton, K.J., Barrett, A., Burrows, C., Bell, C.G., Lowe, R., Beck, S., Rakyan, V.K., Gloyn, A.L., and McCarthy, M.I. Integration of human pancreatic islet genomic data refines regulatory mechanisms at Type 2 Diabetes

susceptibility loci. *eLife* **7**: e31977, 2018.

[95] Taneera, J., Fadista, J., Ahlqvist, E., Atac, D., Ottosson-Laakso, E., Wollheim, C.B., and Groop, L. Identification of novel genes for glucose metabolism based upon expression pattern in human islets and effect on insulin secretion and glycemia. *Human Molecular Genetics* **24**(7): 1945–1955, 2015.

[96] Nica, A.C., Ongen, H., Irminger, J.C., Bosco, D., Berney, T., Antonarakis, S.E., Halban, P.A., and Dermitzakis, E.T. Cell-type, allelic, and genetic signatures in the human pancreatic beta cell transcriptome. *Genome Research* **23**(9): 1554–1562, 2013.

[97] Parker, S.C.J., Stitzel, M.L., Taylor, D.L., Orozco, J.M., Erdos, M.R., Akiyama, J.A., van Bueren, K.L., Chines, P.S., Narisu, N., NISC Comparative Sequencing Program, Black, B.L., Visel, A., Pennacchio, L.A., and Collins, F.S. Chromatin stretch enhancer states drive cell-specific gene regulation and harbor human disease risk variants. *Proceedings of the National Academy of Sciences* **110**(44): 17921–17926, 2013.

[98] Fadista, J., Vikman, P., Laakso, E.O., Mollet, I.G., Esguerra, J.L., Taneera, J., Storm, P., Osmark, P., Ladenvall, C., Prasad, R.B., Hansson, K.B., Finotello, F., Uvebrant, K., Ofori, J.K., Di Camillo, B., Krus, U., Cilio, C.M., Hansson, O., Eliasson, L., Rosengren, A.H., Renström, E., Wollheim, C.B., and Groop, L. Global genomic and transcriptomic analysis of human pancreatic islets reveals novel genes influencing glucose metabolism. *Proceedings of the National Academy of Sciences* **111**(38): 13924–13929, 2014.

[99] Olsson, A.H., Volkov, P., Bacos, K., Dayeh, T., Hall, E., Nilsson, E.A., Ladenvall, C., Rönn, T., and Ling, C. Genome-Wide Associations between Genetic and Epigenetic Variation Influence mRNA Expression and Insulin Secretion in Human Pancreatic Islets. *PLOS Genetics* **10**(11): e1004735, 2014.

[100] van de Bunt, M., Manning Fox, J.E., Dai, X., Barrett, A., Grey, C., Li, L., Bennett, A.J., Johnson, P.R., Rajotte, R.V., Gaulton, K.J., Dermitzakis, E.T., MacDonald, P.E., McCarthy, M.I., and Gloyn, A.L. 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, 2015.

[101] Ottosson-Laakso, E., Krus, U., Storm, P., Prasad, R.B., Oskolkov, N., Ahlqvist, E., Fadista, J., Hansson, O., Groop, L., and Vikman, P. Glucose-induced Changes in Gene Expression in Human Pancreatic Islets — Causes or Consequences of Chronic Hyperglycemia. *Diabetes* , 2017.

[102] Varshney, A., Scott, L.J., Welch, R.P., Erdos, M.R., Chines, P.S., Narisu, N., Albanus, R.D., Orchard, P., Wolford, B.N., Kursawe, R., Vadlamudi, S., Cannon, M.E., Didion, J.P., Hensley, J., Kirilusha, A., Bonnycastle, L.L., Taylor, D.L., Watanabe, R., Mohlke, K.L., Boehnke, M., Collins, F.S., Parker, S.C.J., and Stitzel, M.L. Genetic regulatory signatures underlying islet gene expression and type 2 diabetes. *Proceedings of the National Academy of Sciences* **114**(9): 2301–2306, 2017.

[103] Akerman, I., Tu, Z., Beucher, A., Rolando, D.M.Y., Sauty-Colace, C., Benazra, M., Nakic, N., Yang, J., Wang, H., Pasquali, L., Moran, I., Garcia-Hurtado, J., Castro, N., Gonzalez-Franco, R., Stewart, A.F., Bonner, C., Piemonti, L., Berney, T., Groop, L., Kerr-Conte, J., Pattou, F., Argmann, C., Schadt, E., Ravassard, P., and Ferrer, J. Human Pancreatic β Cell lncRNAs Control Cell-Specific Regulatory Networks. *Cell Metabolism* **25**(2): 400–411, 2017.

[104] Fagerberg, L., Hallström, B.M., Oksvold, P., Kampf, C., Djureinovic, D., Odeberg, J., Habuka, M., Tahmasebpoor, S., Danielsson, A., Edlund, K., Asplund, A., Sjöstedt, E., Lundberg, E., Szigyarto, C.A.K., Skogs, M., Takanen, J.O., Berling, H., Tegel, H., Mulder, J., Nilsson, P., Schwenk, J.M., Lindskog, C., Danielsson, F., Mardinoglu, A., Sivertsson, s., von Feilitzen, K., Forsberg, M., Zwahlen, M., Olsson, I., Navani, S., Huss, M., Nielsen, J., Ponten, F., and Uhlén, M. Analysis of the Human Tissue-specific Expression by Genome-wide Integration of Transcriptomics and Antibody-based Proteomics. *Molecular & Cellular Proteomics* **13**(2): 397–406, 2014.

[105] Ernst, J., Kheradpour, P., Mikkelsen, T.S., Shoresh, N., Ward, L.D., Epstein, C.B., Zhang, X., Wang, L., Issner, R., Coyne, M., Ku, M., Durham, T., Kellis, M., and Bernstein, B.E. Mapping and analysis of chromatin state dynamics in nine human cell types. *Nature* **473**: 43, 2011.

[106] Scott, L.J., Erdos, M.R., Huyghe, J.R., Welch, R.P., Beck, A.T., Wolford, B.N., Chines, P.S., Didion, J.P., Narisu, N., Stringham, H.M., Taylor, D.L., Jackson, A.U., Vadlamudi, S., Bonnycastle, L.L., Kinnunen, L., Saramies, J., Sundvall, J., Albanus, R.D., Kiseleva, A., Hensley, J., Crawford, G.E., Jiang, H., Wen, X., Watanabe, R.M., Lakka, T.A., Mohlke, K.L., Laakso, M., Tuomilehto, J., Koistinen, H.A., Boehnke, M., Collins, F.S., and Parker, S.C.J. The genetic regulatory signature of type 2 diabetes in human skeletal muscle. *Nature Communications* **7**: 11764, 2016.

[107] Emilsson, V., Thorleifsson, G., Zhang, B., Leonardson, A.S., Zink, F., Zhu, J., Carlson, S., Helgason, A., Walters, G.B., Gunnarsdottir, S., Mouy, M., Steinthorsdottir, V., Eiriksdottir, G.H., Bjornsdottir, G., Reynisdottir, I., Gudbjartsson, D., Helgadottir, A., Jonasdottir, A., Jonasdottir, A., Styrkarsdottir, U., Gretarsdottir, S., Magnusson, K.P., Stefansson, H., Fossdal, R., Kristjansson, K., Gislason, H.G., Stefansson, T., Leifsson, B.G., Thorsteinsdottir, U., Lamb, J.R., Gulcher, J.R., Reitman, M.L., Kong, A., Schadt, E.E., and Stefansson, K. Genetics of gene expression and its effect on disease. *Nature* **452**: 423, 2008.

[108] Beer, N.L. and McCarthy, M.I. Genome-Wide Association Studies in Type 2 Diabetes. *Frontiers in Diabetes* **23**: 1–13, 2014.

[109] The MuTHER Consortium. Identification of an imprinted master *trans* regulator at the *KLF14* locus related to multiple metabolic phenotypes. *Nature Genetics* **43**: 561, 2011.

[110] Segerstolpe, A., Palasantza, A., Eliasson, P., Andersson, E.M., Andréasson, A.C., Sun, X., Picelli, S., Sabirsh, A., Clausen, M., Bjursell, M.K., Smith, D., Kasper, M., Ämmälä, C., and Sandberg, R. Single-Cell Transcriptome Profiling

of Human Pancreatic Islets in Health and Type 2 Diabetes. *Cell Metabolism* **24**(4): 593–607, 2016.

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

[112] Muraro, M., Dharmadhikari, G., Grün, D., Groen, N., Dielen, T., Jansen, E., van Gurp, L., Engelse, M., Carlotti, F., de Koning, E.P., and van Oudenaarden, A. A Single-Cell Transcriptome Atlas of the Human Pancreas. *Cell Systems* **3**(4): 385–394.e3, 2016.

[113] Blodgett, D.M., Nowosielska, A., Afik, S., Pechhold, S., Cura, A.J., Kennedy, N.J., Kim, S., Kucukural, A., Davis, R.J., Kent, S.C., Greiner, D.L., Garber, M.G., Harlan, D.M., and diIorio, P. Novel Observations From Next-Generation RNA Sequencing of Highly Purified Human Adult and Fetal Islet Cell Subsets. *Diabetes* **64**(9): 3172–3181, 2015.

[114] Dorrell, C., Schug, J., Canaday, P.S., Russ, H.A., Tarlow, B.D., Grompe, M.T., Horton, T., Hebrok, M., Streeter, P.R., Kaestner, K.H., and Grompe, M. Human islets contain four distinct subtypes of β cells. *Nature Communications* **7**: 11756, 2016.

[115] Taneera, J., Lang, S., Sharma, A., Fadista, J., Zhou, Y., Ahlqvist, E., Jonsson, A., Lyssenko, V., Vikman, P., Hansson, O., Parikh, H., Korsgren, O., Soni, A., Krus, U., Zhang, E., Jing, X.J., Esguerra, J.L.S., Wollheim, C.B., Salehi, A., Rosengren, A., Renström, E., and Groop, L. A Systems Genetics Approach Identifies Genes and Pathways for Type 2 Diabetes in Human Islets. *Cell Metabolism* **16**(1): 122–134, 2012.

[116] Locke, J.M., Hysenaj, G., Wood, A.R., Weedon, M.N., and Harries, L.W. Targeted allelic expression profiling in human islets identifies *cis*-regulatory effects for multiple variants identified by type 2 diabetes genome-wide association studies. *Diabetes* , 2014.

[117] Ohno, S. So Much "Junk" DNA in our Genome. *Brookhaven Symposia in Biology* **23**: 366–370, 1972.

[118] Kung, J.T.Y., Colognori, D., and Lee, J.T. Long Noncoding RNAs: Past, Present, and Future. *Genetics* **193**(3): 651–669, 2013.

[119] Margulies, M., Egholm, M., Altman, W.E., Attiya, S., Bader, J.S., Bemben, L.A., Berka, J., Braverman, M.S., Chen, Y.J., Chen, Z., Dewell, S.B., Du, L., Fierro, J.M., Gomes, X.V., Godwin, B.C., He, W., Helgesen, S., Ho, C.H., Irzyk, G.P., Jando, S.C., Alenquer, M.L.I., Jarvie, T.P., Jirage, K.B., Kim, J.B., Knight, J.R., Lanza, J.R., Leamon, J.H., Lefkowitz, S.M., Lei, M., Li, J., Lohman, K.L., Lu, H., Makhijani, V.B., McDade, K.E., McKenna, M.P., Myers, E.W., Nickerson, E., Nobile, J.R., Plant, R., Puc, B.P., Ronan, M.T., Roth, G.T., Sarkis, G.J., Simons, J.F., Simpson, J.W., Srinivasan, M., Tartaro, K.R., Tomasz, A., Vogt, K.A., Volkmer, G.A., Wang, S.H., Wang, Y., Weiner, M.P., Yu, P., Begley, R.F., and Rothberg, J.M. Genome sequencing in

microfabricated high-density picolitre reactors. *Nature* **437**: 376, 2005.

[120] International Human Genome Sequencing Consortium. Initial sequencing and analysis of the human genome. *Nature* **409**: 860, 2001.

[121] Weber, A.P. Discovering New Biology through Sequencing of RNA. *Plant Physiology* **169**(3): 1524–1531, 2015.

[122] Maxmen, A. RNA: The genome's rising stars. *Nature* **496**: 127–129, 2013.

[123] Penny, G.D., Kay, G.F., Sheardown, S.A., Rastan, S., and Brockdorff, N. Requirement for *Xist* in X chromosome inactivation. *Nature* **379**(6561): 131–137, 1996.

[124] Tripathi, R., Chakraborty, P., and Varadwaj, P.K. Unraveling long non-coding RNAs through analysis of high-throughput RNA-sequencing data. *Non-coding RNA Research* **2**(2): 111–118, 2017.

[125] Liu, S.J., Horlbeck, M.A., Cho, S.W., Birk, H.S., Malatesta, M., He, D., Attenello, F.J., Villalta, J.E., Cho, M.Y., Chen, Y., Mandegar, M.A., Olvera, M.P., Gilbert, L.A., Conklin, B.R., Chang, H.Y., Weissman, J.S., and Lim, D.A. CRISPRi-based genome-scale identification of functional long noncoding RNA loci in human cells. *Science* **355**(6320), 2017.

[126] Harrow, J., Frankish, A., Gonzalez, J.M., Tapanari, E., Diekhans, M., Kokocinski, F., Aken, B.L., Barrell, D., Zadissa, A., Searle, S., Barnes, I., Bignell, A., Boychenko, V., Hunt, T., Kay, M., Mukherjee, G., Rajan, J., Despacio-Reyes, G., Saunders, G., Steward, C., Harte, R., Lin, M., Howald, C., Tanzer, A., Derrien, T., Chrast, J., Walters, N., Balasubramanian, S., Pei, B., Tress, M., Rodriguez, J.M., Ezkurdia, I., van Baren, J., Brent, M., Haussler, D., Kellis, M., Valencia, A., Reymond, A., Gerstein, M., Guigó, R., and Hubbard, T.J. GENCODE: The reference human genome annotation for The ENCODE Project. *Genome Research* **22**(9): 1760–1774, 2012.

[127] Ma, L., Bajic, V.B., and Zhang, Z. On the classification of long non-coding RNAs. *RNA Biology* **10**(6): 924–933, 2013.

[128] Kowalczyk, M.S., Higgs, D.R., and Gingeras, T.R. Molecular biology: RNA discrimination. *Nature* **482**(7385): 310–311, 2012.

[129] Ji, P., Diederichs, S., Wang, W., Böing, S., Metzger, R., Schneider, P.M., Tidow, N., Brandt, B., Buerger, H., Bulk, E., Thomas, M., Berdel, W.E., Serve, H., and Müller-Tidow, C. MALAT-1, a novel noncoding RNA, and thymosin $\beta 4$ predict metastasis and survival in early-stage non-small cell lung cancer. *Oncogene* **22**: 8031, 2003.

[130] Gupta, R.A., Shah, N., Wang, K.C., Kim, J., Horlings, H.M., Wong, D.J., Tsai, M.C., Hung, T., Argani, P., Rinn, J.L., Wang, Y., Brzoska, P., Kong, B., Li, R., West, R.B., van de Vijver, M.J., Sukumar, S., and Chang, H.Y. Long non-coding RNA *HOTAIR* reprograms chromatin state to promote cancer metastasis. *Nature* **464**: 1071, 2010.

[131] Iyer, M.K., Niknafs, Y.S., Malik, R., Singhal, U., Sahu, A., Hosono, Y., Barrette, T.R., Prensner, J.R., Evans, J.R., Zhao, S., Poliakov, A., Cao, X.,

Dhanasekaran, S.M., Wu, Y.M., Robinson, D.R., Beer, D.G., Feng, F.Y., Iyer, H.K., and Chinnaiyan, A.M. The landscape of long noncoding RNAs in the human transcriptome. *Nat Genet* **47**(3): 199–208, 2015.

[132] Wapinski, O. and Chang, H.Y. Long noncoding RNAs and human disease. *Trends in Cell Biology* **21**(6): 354–361, 2011.

[133] Peng, Y. and Calin, G.A. Crucial role of non-coding RNAs in disease. *Cancer Letters* **420**: 127–128, 2018.

[134] Pullen, T.J. and Rutter, G.A. Roles of lncRNAs in Pancreatic Beta Cell Identity and Diabetes Susceptibility. *Frontiers in Genetics* **5**, 2014.

[135] Djebali, S., Davis, C.A., Merkel, A., Dobin, A., Lassmann, T., Mortazavi, A.M., Tanzer, A., Lagarde, J., Lin, W., Schlesinger, F., Xue, C., Marinov, G.K., Khatun, J., Williams, B.A., Zaleski, C., Rozowsky, J., Röder, M., Kokocinski, F., Abdelhamid, R.F., Alioto, T., Antoshechkin, I., Baer, M.T., Bar, N.S., Batut, P., Bell, K., Bell, I., Chakrabortty, S., Chen, X., Chrast, J., Curado, J., Derrien, T., Drenkow, J., Dumais, E., Dumais, J., Duttagupta, R., Falconnet, E., Fastuca, M., Fejes-Toth, K., Ferreira, P., Foissac, S., Fullwood, M.J., Gao, H., Gonzalez, D., Gordon, A., Gunawardena, H., Howald, C., Jha, S., Johnson, R., Kapranov, P., King, B., Kingswood, C., Luo, O.J., Park, E., Persaud, K., Preall, J.B., Ribeca, P., Risk, B., Robyr, D., Sammeth, M., Schaffer, L., See, L.H., Shahab, A., Skancke, J., Suzuki, A.M., Takahashi, H., Tilgner, H., Trout, D., Walters, N., Wang, H., Wrobel, J., Yu, Y., Ruan, X., Hayashizaki, Y., Harrow, J., Gerstein, M., Hubbard, T., Reymond, A., Antonarakis, S.E., Hannon, G., Giddings, M.C., Ruan, Y., Wold, B., Carninci, P., Guigó, R., and Gingeras, T.R. Landscape of transcription in human cells. *Nature* **489**(7414): 101–108, 2012.

[136] DiStefano, J.K. Beyond the Protein-Coding Sequence: Noncoding RNAs in the Pathogenesis of Type 2 Diabetes. *The Review of Diabetic Studies : RDS* **12**(3-4): 260–276, 2015.

[137] Morán, I., Akerman, l., van de Bunt, M., Xie, R., Benazra, M., Nammo, T., Arnes, L., Nakić, N., García-Hurtado, J., Rodríguez-Seguí, S., Pasquali, L., Sauty-Colace, C., Beucher, A., Scharfmann, R., van Arensbergen, J., Johnson, P., Berry, A., Lee, C., Harkins, T., Gmyr, V., Pattou, F., Kerr-Conte, J., Piemonti, L., Berney, T., Hanley, N., Gloyn, A., Sussel, L., Langman, L., Brayman, K., Sander, M., McCarthy, M., Ravassard, P., and Ferrer, J. Human β Cell Transcriptome Analysis Uncovers lncRNAs That Are Tissue-Specific, Dynamically Regulated, and Abnormally Expressed in Type 2 Diabetes. *Cell Metabolism* **16**(4): 435–448, 2012.

[138] Senée, V., Chelala, C., Duchatelet, S., Feng, D., Blanc, H., Cossec, J.C., Charon, C., Nicolino, M., Boileau, P., Cavener, D.R., Bougnères, P., Taha, D., and Julier, C. Mutations in *GLIS3* are responsible for a rare syndrome with neonatal diabetes mellitus and congenital hypothyroidism. *Nature Genetics* **38**: 682, 2006.

[139] Wen, X. and Yang, Y. Emerging roles of GLIS3 in neonatal diabetes, type 1 and

type 2 diabetes. *Journal of Molecular Endocrinology* **58**(2): R73–R85, 2017.

[140] Li, S., Tighe, S.W., Nicolet, C.M., Grove, D., Levy, S., Farmerie, W., Viale, A., Wright, C., Schweitzer, P.A., Gao, Y., Kim, D., Boland, J., Hicks, B., Kim, R., Chhangawala, S., Jafari, N., Raghavachari, N., Gandara, J., Garcia-Reyero, N., Hendrickson, C., Roberson, D., Rosenfeld, J.A., Smith, T., Underwood, J.G., Wang, M., Zumbo, P., Baldwin, D.A., Grills, G.S., and Mason, C.E. Multi-platform assessment of transcriptome profiling using RNA-seq in the ABRF next-generation sequencing study. *Nature Biotechnology* **32**: 915, 2014.

[141] Kumar, A., Kankainen, M., Parsons, A., Kallioniemi, O., Mattila, P., and Heckman, C.A. The impact of RNA sequence library construction protocols on transcriptomic profiling of leukemia. *BMC Genomics* **18**(1): 629, 2017.

[142] Kukurba, K.R. and Montgomery, S.B. RNA Sequencing and Analysis. *Cold Spring Harbor protocols* **2015**(11): 951–969, 2015.

[143] Sultan, M., Amstislavskiy, V., Risch, T., Schuette, M., Dökel, S., Ralser, M., Balzereit, D., Lehrach, H., and Yaspo, M.L. Influence of RNA extraction methods and library selection schemes on RNA-seq data. *BMC Genomics* **15**(1): 675, 2014.

[144] Zhao, W., He, X., Hoadley, K.A., Parker, J.S., Hayes, D.N., and Perou, C.M. Comparison of RNA-Seq by poly (A) capture, ribosomal RNA depletion, and DNA microarray for expression profiling. *BMC Genomics* **15**(1): 419, 2014.

[145] Zhao, S., Zhang, Y., Gamini, R., Zhang, B., and von Schack, D. Evaluation of two main RNA-seq approaches for gene quantification in clinical RNA sequencing: polyA+ selection versus rRNA depletion. *Scientific Reports* **8**(1): 4781, 2018.

[146] Guo, Y., Zhao, S., Sheng, Q., Guo, M., Lehmann, B., Pietenpol, J., Samuels, D.C., and Shyr, Y. RNA-seq by Total RNA Library Identifies Additional RNAs Compared to Poly(A) RNA Library. *BioMed Research International* **2015**: 862130, 2015.

[147] Kissopoulou, A., Jonasson, J., Lindahl, T.L., and Osman, A. Next Generation Sequencing Analysis of Human Platelet PolyA+ mRNAs and rRNA-Depleted Total RNA. *PLOS ONE* **8**(12): e81809, 2013.

[148] Sun, Z., Asmann, Y.W., Nair, A., Zhang, Y., Wang, L., Kalari, K.R., Bhagwate, A.V., Baker, T.R., Carr, J.M., Kocher, J.P.A., Perez, E.A., and Thompson, E.A. Impact of Library Preparation on Downstream Analysis and Interpretation of RNA-Seq Data: Comparison between Illumina PolyA and NuGEN Ovation Protocol. *PLOS ONE* **8**(8): e71745, 2013.

[149] Lappalainen, T., Sammeth, M., Friedländer, M.R., 't Hoen, P.A.C., Monlong, J., Rivas, M.A., Gonzàlez-Porta, M., Kurbatova, N., Griebel, T., Ferreira, P.G., Barann, M., Wieland, T., Greger, L., van Iterson, M., Almlöf, J., Ribeca, P., Pulyakhina, I., Esser, D., Giger, T., Tikhonov, A., Sultan, M., Bertier, G., MacArthur, D.G., Lek, M., Lizano, E., Buermans, H.P.J., Padioleau, I., Schwarzmayr, T., Karlberg, O., Ongen, H., Kilpinen, H., Beltran, S., Gut, M., Kahlem, K., Amstislavskiy, V., Stegle, O., Pirinen, M., Montgomery, S.B.,

Donnelly, P., McCarthy, M.I., Flicek, P., Strom, T.M., The Geuvadis Consortium, Lehrach, H., Schreiber, S., Sudbrak, R., Carracedo, A., Antonarakis, S.E., Häsler, R., Syvänen, A.C., van Ommen, G.J., Brazma, A., Meitinger, T., Rosenstiel, P., Guigó, R., Gut, I.G., Estivill, X., and Dermitzakis, E.T. Transcriptome and genome sequencing uncovers functional variation in humans. *Nature* **501**: 506, 2013.

[150] Nilsson, E., Jansson, P.A., Perfilyev, A., Volkov, P., Pedersen, M., Svensson, M.K., Poulsen, P., Ribel-Madsen, R., Pedersen, N.L., Almgren, P., Fadista, J., Rönn, T., Klarlund Pedersen, B., Scheele, C., Vaag, A., and Ling, C. 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, 2014.

[151] Wang, X., Chang, X., Zhang, P., Fan, L., Zhou, T., and Sun, K. Aberrant Expression of Long Non-Coding RNAs in Newly Diagnosed Type 2 Diabetes Indicates Potential Roles in Chronic Inflammation and Insulin Resistance. *Cellular Physiology and Biochemistry* **43**(6): 2367–2378, 2017.

[152] Gao, Y., Wu, F., Zhou, J., Yan, L., Jurczak, M.J., Lee, H.Y., Yang, L., Mueller, M., Zhou, X.B., Dandolo, L., Szendroedi, J., Roden, M., Flannery, C., Taylor, H., Carmichael, G.G., Shulman, G.I., and Huang, Y. The H19/let-7 double-negative feedback loop contributes to glucose metabolism in muscle cells. *Nucleic Acids Research* **42**(22): 13799–13811, 2014.

[153] Kallen, A., Zhou, X.B., Xu, J., Qiao, C., Ma, J., Yan, L., Lu, L., Liu, C., Yi, J.S., Zhang, H., Min, W., Bennett, A., Gregory, R., Ding, Y., and Huang, Y. The Imprinted H19 LncRNA Antagonizes Let-7 MicroRNAs. *Molecular Cell* **52**(1): 101–112, 2013.

[154] Leti, F. and DiStefano, J. Long Noncoding RNAs as Diagnostic and Therapeutic Targets in Type 2 Diabetes and Related Complications. *Genes* **8**(8): 207, 2017.

[155] van Bakel, H., Nislow, C., Blencowe, B.J., and Hughes, T.R. Most "Dark Matter" Transcripts Are Associated With Known Genes. *PLOS Biology* **8**(5): e1000371, 2010.

[156] Gamazon, E.R., Wheeler, H.E., Shah, K.P., Mozaffari, S.V., Aquino-Michaels, K., Carroll, R.J., Eyler, A.E., Denny, J.C., GTEx Consortium, Nicolae, D.L., Cox, N.J., and Im, H.K. A gene-based association method for mapping traits using reference transcriptome data. *Nat Genet* **47**(9): 1091–1098, 2015.

[157] Gusev, A., Ko, A., Shi, H., Bhatia, G., Chung, W., Penninx, B.W.J.H., Jansen, R., de Geus, E.J.C., Boomsma, D.I., Wright, F.A., Sullivan, P.F., Nikkola, E., Alvarez, M., Civelek, M., Lusis, A.J., Lehtimaki, T., Raitoharju, E., Kahonen, M., Seppala, I., Raitakari, O.T., Kuusisto, J., Laakso, M., Price, A.L., Pajukanta, P., and Pasaniuc, B. Integrative approaches for large-scale transcriptome-wide association studies. *Nat Genet* **48**(3): 245–252, 2016.

[158] Jansen, R., Hottenga, J.J., Nivard, M.G., Abdellaoui, A., Laport, B., de Geus, E.J., Wright, F.A., Penninx, B.W.J.H., and Boomsma, D.I. Conditional eQTL

analysis reveals allelic heterogeneity of gene expression. *Human Molecular Genetics* **26**(8): 1444–1451, 2017.

[159] Zhu, Z., Zhang, F., Hu, H., Bakshi, A., Robinson, M.R., Powell, J.E., Montgomery, G.W., Goddard, M.E., Wray, N.R., Visscher, P.M., and Yang, J. Integration of summary data from GWAS and eQTL studies predicts complex trait gene targets. *Nature Genetics* **48**: 481, 2016.

[160] Zou, H. and Hastie, T. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **67**(2): 301–320, 2005.

[161] Hoerl, A.E. and Kennard, R.W. Ridge Regression: Biased Estimation for Nonorthogonal Problems. *Technometrics* **12**(1): 55–67, 1970.

[162] Tibshirani, R. Regression Shrinkage and Selection via the Lasso. *Journal of the Royal Statistical Society. Series B (Methodological)* **58**(1): 267–288, 1996.

[163] Barbeira, A.N., Dickinson, S.P., Bonazzola, R., Zheng, J., Wheeler, H.E., Torres, J.M., Torstenson, E.S., Shah, K.P., Garcia, T., Edwards, T., Stahl, E.A., Huckins, L.M., GTEx Consortium, Nicolae, D., Cox, N.J., and Im, H.K. Exploring the phenotypic consequences of tissue specific gene expression variation inferred from GWAS summary statistics. *Nature Communications* **9**(1825), 2018.

[164] Pierce, B.L. and Burgess, S. Efficient Design for Mendelian Randomization Studies: Subsample and 2-Sample Instrumental Variable Estimators. *American Journal of Epidemiology* **178**(7): 1177–1184, 2013.

[165] Didelez, V. and Sheehan, N. Mendelian randomization as an instrumental variable approach to causal inference. *Statistical Methods in Medical Research* **16**(4): 309–330, 2007.

[166] Johnson, T. Efficient Calculation for Multi-SNP Genetic Risk Scores. *American Society of Human Genetics Annual Meeting* , 2012.

[167] Lousdal, M.L. An introduction to instrumental variable assumptions, validation and estimation. *Emerging Themes in Epidemiology* **15**: 1, 2018.

[168] Corbin, L.J., Richmond, R.C., Wade, K.H., Burgess, S., Bowden, J., Smith, G.D., and Timpson, N.J. Body mass index as a modifiable risk factor for type 2 diabetes: Refining and understanding causal estimates using Mendelian randomisation. *Diabetes* , 2016.

[169] Wang, T., Huang, T., Li, Y., Zheng, Y., Manson, J.E., Hu, F.B., and Qi, L. Low birthweight and risk of type 2 diabetes: a Mendelian randomisation study. *Diabetologia* **59**(9): 1920–1927, 2016.

[170] Lawlor, D.A. Commentary: Two-sample Mendelian randomization: opportunities and challenges. *International Journal of Epidemiology* **45**(3): 908–915, 2016.

[171] Burgess, S., Scott, R.A., Timpson, N.J., Davey Smith, G., and Thompson, S.G. Using published data in Mendelian randomization: a blueprint for efficient identification of causal risk factors. *European Journal of Epidemiology* **30**(7):

543–552, 2015.

[172] Hartwig, F.P., Davies, N.M., Hemani, G., and Davey Smith, G. Two-sample Mendelian randomization: avoiding the downsides of a powerful, widely applicable but potentially fallible technique. *International Journal of Epidemiology* **45**(6): 1717–1726, 2016.

[173] Bowden, J., Del Greco M, F., Minelli, C., Davey Smith, G., Sheehan, N., and Thompson, J. A framework for the investigation of pleiotropy in two-sample summary data Mendelian randomization. *Statistics in Medicine* **36**(11): 1783–1802, 2017.

[174] Bowden, J., Davey Smith, G., and Burgess, S. Mendelian randomization with invalid instruments: effect estimation and bias detection through Egger regression. *International Journal of Epidemiology* **44**(2): 512–525, 2015.

[175] Burgess, S., Davies, N.M., and Thompson, S.G. Bias due to participant overlap in two-sample Mendelian randomization. *Genetic Epidemiology* **40**(7): 597–608, 2016.

[176] Delaneau, O., Marchini, J., and Zagury, J.F. A linear complexity phasing method for thousands of genomes. *Nature Methods* **9**: 179, 2011.

[177] Howie, B.N., Donnelly, P., and Marchini, J. A Flexible and Accurate Genotype Imputation Method for the Next Generation of Genome-Wide Association Studies. *PLOS Genetics* **5**(6): e1000529, 2009.

[178] Marco-Sola, S., Sammeth, M., Guigó, R., and Ribeca, P. The GEM mapper: fast, accurate and versatile alignment by filtration. *Nature Methods* **9**: 1185, 2012.

[179] Manning Fox, J.E., Lyon, J., Dai, X.Q., Wright, R.C., Hayward, J., van de Bunt, M., Kin, T., Shapiro, A.M.J., McCarthy, M.I., Gloyn, A.L., Ungrin, M.D., Lakey, J.R., Kneteman, N.M., Warnock, G.L., Korbutt, G.S., Rajotte, R.V., and MacDonald, P.E. Human islet function following 20 years of cryogenic biobanking. *Diabetologia* **58**(7): 1503–1512, 2015.

[180] Kin, T. and Shapiro, J. Partial dorsal agenesis accompanied with circumportal pancreas in a donor for islet transplantation. *Islets* **2**(3): 146–148, 2010.

[181] Dobin, A., Davis, C.A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., Batut, P., Chaisson, M., and Gingeras, T.R. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29**(1): 15–21, 2013.

[182] Liao, Y., Smyth, G.K., and Shi, W. featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. *Bioinformatics* **30**(7): 923–930, 2014.

[183] O'Leary, N.A., Wright, M.W., Brister, J.R., Ciufo, S., Haddad, D., McVeigh, R., Rajput, B., Robbertse, B., Smith-White, B., Ako-Adjei, D., Astashyn, A., Badretdin, A., Bao, Y., Blinkova, O., Brover, V., Chetvernin, V., Choi, J., Cox, E., Ermolaeva, O., Farrell, C.M., Goldfarb, T., Gupta, T., Haft, D., Hatcher, E., Hlavina, W., Joardar, V.S., Kodali, V.K., Li, W., Maglott, D., Masterson, P., McGarvey, K.M., Murphy, M.R., O'Neill, K., Pujar, S., Rangwala, S.H.,

Rausch, D., Riddick, L.D., Schoch, C., Shkeda, A., Storz, S.S., Sun, H., Thibaud-Nissen, F., Tolstoy, I., Tully, R.E., Vatsan, A.R., Wallin, C., Webb, D., Wu, W., Landrum, M.J., Kimchi, A., Tatusova, T., DiCuccio, M., Kitts, P., Murphy, T.D., and Pruitt, K.D. Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. *Nucleic Acids Research* **44**(D1): D733–D745, 2016.

[184] Frankish, A., Uszczynska, B., Ritchie, G.R.S., Gonzalez, J.M., Pervouchine, D., Petryszak, R., Mudge, J.M., Fonseca, N., Brazma, A., Guigo, R., and Harrow, J. Comparison of GENCODE and RefSeq gene annotation and the impact of reference geneset on variant effect prediction. *BMC Genomics* **16**(8): S2, 2015.

[185] Hon, C.C., Ramilowski, J.A., Harshbarger, J., Bertin, N., Rackham, O.J.L., Gough, J., Denisenko, E., Schmeier, S., Poulsen, T.M., Severin, J., Lizio, M., Kawaji, H., Kasukawa, T., Itoh, M., Burroughs, A.M., Noma, S., Djebali, S., Alam, T., Medvedeva, Y.A., Testa, A.C., Lipovich, L., Yip, C.W., Abugessaisa, I., Mendez, M., Hasegawa, A., Tang, D., Lassmann, T., Heutink, P., Babina, M., Wells, C.A., Kojima, S., Nakamura, Y., Suzuki, H., Daub, C.O., de Hoon, M.J.L., Arner, E., Hayashizaki, Y., Carninci, P., and Forrest, A.R.R. An atlas of human long non-coding RNAs with accurate 5' ends. *Nature* **543**, 2017.

[186] Jiang, C., Li, Y., Zhao, Z., Lu, J., Chen, H., Ding, N., Wang, G., Xu, J., and Li, X. Identifying and functionally characterizing tissue-specific and ubiquitously expressed human lncRNAs. *Oncotarget* **7**(6): 7120–7133, 2016.

[187] Derrien, T., Johnson, R., Bussotti, G., Tanzer, A., Djebali, S., Tilgner, H., Guernec, G., Martin, D., Merkel, A., Knowles, D.G., Lagarde, J., Veeravalli, L., Ruan, X., Ruan, Y., Lassmann, T., Carninci, P., Brown, J.B., Lipovich, L., Gonzalez, J.M., Thomas, M., Davis, C.A., Shiekhattar, R., Gingeras, T.R., Hubbard, T.J., Notredame, C., Harrow, J., and Guigó, R. The GENCODE v7 catalog of human long noncoding RNAs: Analysis of their gene structure, evolution, and expression. *Genome Research* **22**(9): 1775–1789, 2012.

[188] Melé, M., Ferreira, P.G., Reverter, F., DeLuca, D.S., Monlong, J., Sammeth, M., Young, T.R., Goldmann, J.M., Pervouchine, D.D., Sullivan, T.J., Johnson, R., Segrè, A.V., Djebali, S., Niarchou, A., The GTEx Consortium, Wright, F.A., Lappalainen, T., Calvo, M., Getz, G., Dermitzakis, E.T., Ardlie, K.G., and Guigó, R. The human transcriptome across tissues and individuals. *Science* **348**(6235): 660–665, 2015.

[189] Cabili, M.N., Trapnell, C., Goff, L., Koziol, M., Tazon-Vega, B., Regev, A., and Rinn, J.L. Integrative annotation of human large intergenic noncoding RNAs reveals global properties and specific subclasses. *Genes & Development* , 2011.

[190] Kornienko, A.E., Dotter, C.P., Guenzl, P.M., Gisslinger, H., Gisslinger, B., Cleary, C., Kralovics, R., Pauler, F.M., and Barlow, D.P. Long non-coding RNAs display higher natural expression variation than protein-coding genes in healthy humans. *Genome Biology* **17**(1): 14, 2016.

[191] Yang, L., Duff, M.O., Graveley, B.R., Carmichael, G.G., and Chen, L.L. Genomewide characterization of non-polyadenylated RNAs. *Genome Biology*

12(2): R16, 2011.

[192] Kashi, K., Henderson, L., Bonetti, A., and Carninci, P. Discovery and functional analysis of lncRNAs: Methodologies to investigate an uncharacterized transcriptome. *Biochimica et Biophysica Acta (BBA) - Gene Regulatory Mechanisms* **1859**(1): 3–15, 2016.

[193] Sun, Z., Nair, A., Chen, X., Prodduturi, N., Wang, J., and Kocher, J.P. UClncR: Ultrafast and comprehensive long non-coding RNA detection from RNA-seq. *Scientific Reports* **7**(1): 14196, 2017.

[194] Grabherr, M.G., Haas, B.J., Yassour, M., Levin, J.Z., Thompson, D.A., Amit, I., Adiconis, X., Fan, L., Raychowdhury, R., Zeng, Q., Chen, Z., Mauceli, E., Hacohen, N., Gnirke, A., Rhind, N., di Palma, F., Birren, B.W., Nusbaum, C., Lindblad-Toh, K., Friedman, N., and Regev, A. Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nature Biotechnology* **29**: 644, 2011.

[195] Schulz, M.H., Zerbino, D.R., Vingron, M., and Birney, E. *Oases*: robust *de novo* RNA-seq assembly across the dynamic range of expression levels. *Bioinformatics* **28**, 2012.

[196] Trapnell, C., Williams, B.A., Pertea, G., Mortazavi, A., Kwan, G., van Baren, M.J., Salzberg, S.L., Wold, B.J., and Pachter, L. Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nature Biotechnology* **28**: 511, 2010.

[197] Pertea, M., Pertea, G.M., Antonescu, C.M., Chang, T.C., Mendell, J.T., and Salzberg, S.L. StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. *Nature Biotechnology* **33**: 290, 2015.

[198] Mayne, B.T., Bianco-Miotto, T., Buckberry, S., Breen, J., Clifton, V., Shoubridge, C., and Roberts, C.T. Large Scale Gene Expression Meta-Analysis Reveals Tissue-Specific, Sex-Biased Gene Expression in Humans. *Frontiers in Genetics* **7**: 183, 2016.

[199] Gershoni, M. and Pietrokovski, S. The landscape of sex-differential transcriptome and its consequent selection in human adults. *BMC Biology* **15**(1): 7, 2017.

[200] Robinson, M.D. and Oshlack, A. A scaling normalization method for differential expression analysis of RNA-seq data. *Genome Biology* **11**(3): R25, 2010.

[201] Robinson, M.D., McCarthy, D.J., and Smyth, G.K. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* **26**(1): 139–140, 2010.

[202] Ritchie, M.E., Phipson, B., Wu, D., Hu, Y., Law, C.W., Shi, W., and Smyth, G.K. *limma* powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Research* , 2015.

[203] Leeks, J.T., Johnson, W.E., Parker, H.S., Fertig, E.J., Jaffe, A.E., Storey, J.D., Zhang, Y., and Torres, L.C. sva: Surrogate Variable Analysis , 2016.

[204] Leek, J.T. and Storey, J.D. Capturing Heterogeneity in Gene Expression

Studies by Surrogate Variable Analysis. *PLOS Genetics* **3**(9): e161, 2007.

[205] Hoen, P.A.C., Friedländer, M.R., Almlöf, J., Sammeth, M., Pulyakhina, I., Anvar, S.Y., Laros, J.F.J., Buermans, H.P.J., Karlberg, O., Brännvall, M., The Geuvadis Consortium, den Dunnen, J.T., van Ommen, G.J.B., Gut, I.G., Guigó, R., Estivill, X., Syvänen, A.C., Dermitzakis, E.T., and Lappalainen, T. Reproducibility of high-throughput mRNA and small RNA sequencing across laboratories. *Nature Biotechnology* **31**: 1015, 2013.

[206] Villegas, V.E. and Zaphiropoulos, P.G. Neighboring Gene Regulation by Antisense Long Non-Coding RNAs. *International Journal of Molecular Sciences* **16**(2): 3251–3266, 2015.

[207] Chaisson, M.J. and Tesler, G. Mapping single molecule sequencing reads using basic local alignment with successive refinement (BLASR): application and theory. *BMC Bioinformatics* **13**(1): 238, 2012.

[208] Zhao, H., Chen, M., Lind, S.B., and Pettersson, U. Distinct temporal changes in host cell lncRNA expression during the course of an adenovirus infection. *Virology* **492**(Supplement C): 242–250, 2016.

[209] Cruceanu, C., Tan, P.P.C., Rogic, S., Lopez, J.P., Torres-Platas, S.G., Gigek, C.O., Alda, M., Rouleau, G.A., Pavlidis, P., and Turecki, G. Transcriptome Sequencing of the Anterior Cingulate in Bipolar Disorder: Dysregulation of G Protein-Coupled Receptors. *American Journal of Psychiatry* **172**(11): 1131–1140, 2015.

[210] Canella, D., Praz, V., Reina, J.H., Cousin, P., and Hernandez, N. Defining the RNA polymerase III transcriptome: Genome-wide localization of the RNA polymerase III transcription machinery in human cells. *Genome Research* **20**(6): 710–721, 2010.

[211] Peterlin, B.M., Brogie, J.E., and Price, D.H. 7SK snRNA: a noncoding RNA that plays a major role in regulating eukaryotic transcription. *Wiley Interdisciplinary Reviews: RNA* **3**(1): 92–103, 2012.

[212] Johnson, M., Zaretskaya, I., Raytselis, Y., Merezhuk, Y., McGinnis, S., and Madden, T.L. NCBI BLAST: a better web interface. *Nucleic Acids Research* **36**(Web Server issue): W5–W9, 2008.

[213] Winsor, G.L., Griffiths, E.J., Lo, R., Dhillon, B.K., Shay, J.A., and Brinkman, F.L. Enhanced annotations and features for comparing thousands of *Pseudomonas* genomes in the Pseudomonas genome database. *Nucleic Acids Research* **44**(D1): D646–D653, 2016.

[214] Régnier, P. and Marujo, P.E. Polyadenylation and Degradation of RNA in Prokaryotes. *Madame Curie Bioscience Database [Internet]* , 2013.

[215] Carvalhais, V., Delgado-Rastrollo, M., Melo, L.D.R., and Cerca, N. Controlled RNA contamination and degradation and its impact on qPCR gene expression in *S. epidermidis* biofilms. *Journal of Microbiological Methods* **95**(2): 195–200, 2013.

[216] Conesa, A., Madrigal, P., Tarazona, S., Gomez-Cabrero, D., Cervera, A.,

McPherson, A., Szcze'sniak, M.W., Gaffney, D.J., Elo, L.L., Zhang, X., and Mortazavi, A. A survey of best practices for RNA-seq data analysis. *Genome Biology* **17**: 13, 2016.

[217] Khachane, A.N. and Harrison, P.M. Mining Mammalian Transcript Data for Functional Long Non-Coding RNAs. *PLOS ONE* **5**(4): e10316, 2010.

[218] Ramsköld, D., Wang, E.T., Burge, C.B., and Sandberg, R. An Abundance of Ubiquitously Expressed Genes Revealed by Tissue Transcriptome Sequence Data. *PLOS Computational Biology* **5**(12): e1000598, 2009.

[219] Zhu, J., Chen, G., Zhu, S., Li, S., Wen, Z., Bin, L., Zheng, Y., and Shi, L. Identification of Tissue-Specific Protein-Coding and Noncoding Transcripts across 14 Human Tissues Using RNA-seq. *Scientific Reports* **6**: 28400, 2016.

[220] Stegle, O., Parts, L., Piipari, M., Winn, J., and Durbin, R. Using probabilistic estimation of expression residuals (PEER) to obtain increased power and interpretability of gene expression analyses. *Nature protocols* **7**(3): 500–507, 2012.

[221] Soneson, C. and Delorenzi, M. A comparison of methods for differential expression analysis of RNA-seq data. *BMC Bioinformatics* **14**(1): 91, 2013.

[222] Anders, S. and Huber, W. Differential expression analysis for sequence count data. *Genome Biology* **11**(10): R106, 2010.

[223] Tarazona, S., García-Alcalde, F., Dopazo, J., Ferrer, A., and Conesa, A. Differential expression in RNA-seq: A matter of depth. *Genome Research* **21**(12): 2213–2223, 2011.

[224] Jeanmougin, M., de Reynies, A., Marisa, L., Paccard, C., Nuel, G., and Guedj, M. Should We Abandon the t-Test in the Analysis of Gene Expression Microarray Data: A Comparison of Variance Modeling Strategies. *PLOS ONE* **5**(9): e12336, 2010.

[225] Schurch, N.J., Schofield, P., Gierliński, M., Cole, C., Sherstnev, A., Singh, V., Wrobel, N., Gharbi, K., Simpson, G.G., Owen-Hughes, T., Blaxter, M., and Barton, G.J. How many biological replicates are needed in an RNA-seq experiment and which differential expression tool should you use? *RNA* **22**(6): 839–851, 2016.

[226] Li, C., Liang, G., Yao, W., Sui, J., Shen, X., Zhang, Y., Ma, S., Ye, Y., Zhang, Z., Zhang, W., Yin, L., and Pu, Y. Differential expression profiles of long non-coding RNAs reveal potential biomarkers for identification of human gastric cancer. *Oncology Reports* **35**: 1529–1540, 2016.

[227] Peng, F., Wang, R., Zhang, Y., Zhao, Z., Zhou, W., Chang, Z., Liang, H., Zhao, W., Qi, L., Guo, Z., and Gu, Y. Differential expression analysis at the individual level reveals a lncRNA prognostic signature for lung adenocarcinoma. *Molecular Cancer* **16**(1): 98, 2017.

[228] He, J., Ou, Q., Liu, C., Shi, L., Zhao, C., Xu, Y., Kong, S.K., Loo, J.F.C., Li, B., and Gu, D. Differential expression of long non-coding RNAs in patients with tuberculosis infection. *Tuberculosis* **107**(Supplement C): 73–79, 2017.

[229] Li, L., Cong, Y., Gao, X., Wang, Y., and Lin, P. Differential expression profiles of long non-coding RNAs as potential biomarkers for the early diagnosis of acute myocardial infarction. *Oncotarget* **8**(51): 88613–88621, 2017.

[230] Trypsteen, W., Mohammadi, P., Van Hecke, C., Mestdagh, P., Lefever, S., Saeys, Y., De Bleser, P., Vandesompele, J., Ciuffi, A., Vandekerckhove, L., and De Spiegelaere, W. Differential expression of lncRNAs during the HIV replication cycle: an underestimated layer in the HIV-host interplay. *Scientific Reports* **6**: 36111, 2016.

[231] van Dam, S., Võsa, U., van der Graaf, A., Franke, L., and de Magalhães, J.a.P. Gene co-expression analysis for functional classification and gene-disease predictions. *Briefings in Bioinformatics* bbw139–bbw139, 2017.

[232] Tzfadia, O., Diels, T., De Meyer, S., Vandepoele, K., Aharoni, A., and Van de Peer, Y. CoExpNetViz: Comparative Co-Expression Networks Construction and Visualization Tool. *Frontiers in Plant Science* **6**(1194), 2016.

[233] Ho, R., Sances, S., Gowing, G., Amoroso, M.W., O'Rourke, J.G., Sahabian, A., Wichterle, H., Baloh, R.H., Sareen, D., and Svendsen, C.N. ALS disrupts spinal motor neuron maturation and aging pathways within gene co-expression networks. *Nature Neuroscience* **19**: 1256, 2016.

[234] Ala, U., Piro, R.M., Grassi, E., Damasco, C., Silengo, L., Oti, M., Provero, P., and Di Cunto, F. Prediction of Human Disease Genes by Human-Mouse Conserved Coexpression Analysis. *PLOS Computational Biology* **4**(3): e1000043, 2008.

[235] Langfelder, P. and Horvath, S. WGCNA: an R package for weighted correlation network analysis. *BMC Bioinformatics* **9**(1): 559, 2008.

[236] Vandepoele, K., Quimbaya, M., Casneuf, T., De Veylder, L., and Van de Peer, Y. Unraveling Transcriptional Control in Arabidopsis Using *cis*-Regulatory Elements and Coexpression Networks. *Plant Physiology* **150**(2): 535–546, 2009.

[237] Langfelder, P., Zhang, B., and Horvath, S. Defining clusters from a hierarchical cluster tree: the Dynamic Tree Cut package for R. *Bioinformatics* **24**(5): 719–720, 2008.

[238] D'Haeseleer, P. How does gene expression clustering work? *Nature Biotechnology* **23**: 1499, 2005.

[239] Zhao, Y., Karypis, G., and Fayyad, U. Hierarchical Clustering Algorithms for Document Datasets. *Data Mining and Knowledge Discovery* **10**(2): 141–168, 2005.

[240] Hartigan, J.A. and Wong, M.A. Algorithm AS 136: A K-Means Clustering Algorithm. *Journal of the Royal Statistical Society. Series C (Applied Statistics)* **28**(1): 100–108, 1979.

[241] Gupta, S., Ellis, S.E., Ashar, F.N., Moes, A., Bader, J.S., Zhan, J., West, A.B., and Arking, D.E. Transcriptome analysis reveals dysregulation of innate immune response genes and neuronal activity-dependent genes in autism. *Nature Communications* **5**: 5748, 2014.

[242] Ainsworth, H.F. and Cordell, H.J. Using gene expression data to identify causal pathways between genotype and phenotype in a complex disease: application to Genetic Analysis Workshop 19. *BMC Proceedings* **10**(Suppl 7): 79–84, 2016.

[243] The Gene Ontology Consortium, Ashburner, M., Ball, C.A., Blake, J.A., Botstein, D., Butler, H., Cherry, J.M., Davis, A.P., Dolinski, K., Dwight, S.S., Eppig, J.T., Harris, M.A., Hill, D.P., Issel-Tarver, L., Kasarskis, A., Lewis, S., Matese, J.C., Richardson, J.E., Ringwald, M., Rubin, G.M., and Sherlock, G. Gene Ontology: tool for the unification of biology. *Nature genetics* **25**(1): 25–29, 2000.

[244] The Gene Ontology Consortium. Expansion of the Gene Ontology knowledgebase and resources. *Nucleic Acids Research* **45**(D1): D331–D338, 2017.

[245] Kanehisa, M., Sato, Y., Kawashima, M., Furumichi, M., and Tanabe, M. KEGG as a reference resource for gene and protein annotation. *Nucleic Acids Research* **44**(D1): D457–D462, 2016.

[246] Huang, D.W., Sherman, B.T., and Lempicki, R.A. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nature Protocols* **4**: 44, 2008.

[247] Huang, D.W., Sherman, B.T., and Lempicki, R.A. Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. *Nucleic Acids Research* **37**(1): 1–13, 2009.

[248] Mi, H., Huang, X., Muruganujan, A., Tang, H., Mills, C., Kang, D., and Thomas, P.D. PANTHER version 11: expanded annotation data from Gene Ontology and Reactome pathways, and data analysis tool enhancements. *Nucleic Acids Research* **45**(D1): D183–D189, 2017.

[249] Falcon, S. and Gentleman, R. Using GOstats to test gene lists for GO term association. *Bioinformatics* **23**(2): 257–258, 2007.

[250] Kogelman, L.J.A., Cirera, S., Zhernakova, D.V., Fredholm, M., Franke, L., and Kadarmideen, H.N. Identification of co-expression gene networks, regulatory genes and pathways for obesity based on adipose tissue RNA Sequencing in a porcine model. *BMC Medical Genomics* **7**: 57–57, 2014.

[251] Axelsson, A.S., Mahdi, T., Nenonen, H.A., Singh, T., Hänzelmann, S., Wendt, A., Bagge, A., Reinbothe, T.M., Millstein, J., Yang, X., Zhang, B., Gusmao, E.G., Shu, L., Szabat, M., Tang, Y., Wang, J., Salö, S., Eliasson, L., Artner, I., Fex, M., Johnson, J.D., Wollheim, C.B., Derry, J.M.J., Mecham, B., Spégel, P., Mulder, H., Costa, I.G., Zhang, E., and Rosengren, A.H. Sox5 regulates beta-cell phenotype and is reduced in type 2 diabetes. *Nature Communications* **8**: 15652, 2017.

[252] Klomp, J.A. and Furge, K.A. Genome-wide matching of genes to cellular roles using guilt-by-association models derived from single sample analysis. *BMC Research Notes* **5**(1): 370, 2012.

[253] Gillis, J. and Pavlidis, P. "Guilt by Association" Is the Exception Rather Than

the Rule in Gene Networks. *PLOS Computational Biology* **8**(3): e1002444, 2012.

[254] McDermott-Roe, C., Leleu, M., Rowe, G.C., Palygin, O., Bukowy, J.D., Kuo, J., Rech, M., Hermans-Beijnsberger, S., Schaefer, S., Adami, E., Creemers, E.E., Heinig, M., Schroen, B., Arany, Z., Petretto, E., and Geurts, A.M. Transcriptome-wide co-expression analysis identifies *LRRC2* as a novel mediator of mitochondrial and cardiac function. *PLOS ONE* **12**(2): e0170458, 2017.

[255] Cali-Daylan, A.E. and Dincer, P. Gene co-expression network analysis of dysferlinopathy: Altered cellular processes and functional prediction of *TOR1AIP1*, a novel muscular dystrophy gene. *Neuromuscular Disorders* **27**(3): 269–277, 2017.

[256] van Dam, S., Cordeiro, R., Craig, T., van Dam, J., Wood, S.H., and de Magalhães, J.a.P. GeneFriends: An online co-expression analysis tool to identify novel gene targets for aging and complex diseases. *BMC Genomics* **13**(1): 535, 2012.

[257] Whitlock, M.C. Combining probability from independent tests: the weighted Z-method is superior to Fisher's approach. *Journal of Evolutionary Biology* **18**(5): 1368–1373, 2005.

[258] Li, Q., Hu, J., Ding, J., and Zheng, G. Fisher's method of combining dependent statistics using generalizations of the gamma distribution with applications to genetic pleiotropic associations. *Biostatistics (Oxford, England)* **15**(2): 284–295, 2014.

[259] Carlson, M. org.Hs.eg.db: Genome wide annotation for Human , 2016.

[260] Timmons, J.A., Szkop, K.J., and Gallagher, I.J. Multiple sources of bias confound functional enrichment analysis of global -omics data. *Genome Biology* **16**(1): 186, 2015.

[261] Björnholm, M. and Zierath, J. Insulin signal transduction in human skeletal muscle: identifying the defects in Type II diabetes. *Biochemical Society Transactions* **33**(2): 354–357, 2005.

[262] Mootha, V.K., Lindgren, C.M., Eriksson, K.F., Subramanian, A., Sihag, S., Lehar, J., Puigserver, P., Carlsson, E., Ridderstråle, M., Laurila, E., Houstis, N., Daly, M.J., Patterson, N., Mesirov, J.P., Golub, T.R., Tamayo, P., Spiegelman, B., Lander, E.S., Hirschhorn, J.N., Altshuler, D., and Groop, L.C. PGC-1α-responsive genes involved in oxidative phosphorylation are coordinately downregulated in human diabetes. *Nature Genetics* **34**: 267, 2003.

[263] Patti, M.E., Butte, A.J., Crunkhorn, S., Cusi, K., Berria, R., Kashyap, S., Miyazaki, Y., Kohane, I., Costello, M., Saccone, R., Landaker, E.J., Goldfine, A.B., Mun, E., DeFronzo, R., Finlayson, J., Kahn, C.R., and Mandarino, L.J. Coordinated reduction of genes of oxidative metabolism in humans with insulin resistance and diabetes: Potential role of *PGC1* and *NRF1*. *Proceedings of the National Academy of Sciences* **100**(14): 8466–8471, 2003.

[264] Bork-Jensen, J., Scheele, C., Christophersen, D.V., Nilsson, E., Friedrichsen, M., Fernandez-Twinn, D.S., Grunnet, L.G., Litman, T., Holmstrøm, K., Vind,

B., Højlund, K., Beck-Nielsen, H., Wojtaszewski, J., Ozanne, S.E., Pedersen, B.K., Poulsen, P., and Vaag, A. Glucose tolerance is associated with differential expression of microRNAs in skeletal muscle: results from studies of twins with and without type 2 diabetes. *Diabetologia* **58**(2): 363–373, 2015.

[265] Su, J., Ekman, C., Oskolkov, N., Lahti, L., Ström, K., Brazma, A., Groop, L., Rung, J., and Hansson, O. A novel atlas of gene expression in human skeletal muscle reveals molecular changes associated with aging. *Skeletal Muscle* **5**(1): 35, 2015.

[266] Gong, R., Chen, M., Zhang, C., Chen, M., and Li, H. A comparison of gene expression profiles in patients with coronary artery disease, type 2 diabetes, and their coexisting conditions. *Diagnostic Pathology* **12**(1): 44, 2017.

[267] Stretch, C., Khan, S., Asgarian, N., Eisner, R., Vaisipour, S., Damaraju, S., Graham, K., Bathe, O.F., Steed, H., Greiner, R., and Baracos, V.E. Effects of Sample Size on Differential Gene Expression, Rank Order and Prediction Accuracy of a Gene Signature. *PLoS ONE* **8**(6): e65380, 2013.

[268] Ching, T., Huang, S., and Garmire, L.X. Power analysis and sample size estimation for RNA-Seq differential expression. *RNA* **20**(11): 1684–1696, 2014.

[269] Vance, K.W. and Ponting, C.P. Transcriptional regulatory functions of nuclear long noncoding RNAs. *Trends in Genetics* **30**(8): 348–355, 2014.

[270] Kent, W.J., Sugnet, C.W., Furey, T.S., Roskin, K.M., Pringle, T.H., Zahler, A.M., Haussler, and David. The Human Genome Browser at UCSC. *Genome Research* **12**(6): 996–1006, 2002.

[271] Cerf, M.E. Beta cell dynamics: beta cell replenishment, beta cell compensation and diabetes. *Endocrine* **44**(2): 303–311, 2013.

[272] Marco, A., Konikoff, C., Karr, T.L., and Kumar, S. Relationship between gene co-expression and sharing of transcription factor binding sites in *Drosophila melanogaster*. *Bioinformatics* **25**(19): 2473–2477, 2009.

[273] MacArthur, J., Bowler, E., Cerezo, M., Gil, L., Hall, P., Hastings, E., Junkins, H., McMahon, A., Milano, A., Morales, J., Pendlington, Z.M., Welter, D., Burdett, T., Hindorff, L., Flicek, P., Cunningham, F., and Parkinson, H. The new NHGRI-EBI Catalog of published genome-wide association studies (GWAS Catalog). *Nucleic Acids Research* **45**(D1): D896–D901, 2017.

[274] Li, X., Hastie, A.T., Hawkins, G.A., Moore, W.C., Ampleford, E.J., Milosevic, J., Li, H., Busse, W.W., Erzurum, S.C., Kaminski, N., Wenzel, S.E., Meyers, D.A., and Bleecker, E.R. eQTL of bronchial epithelial cells and bronchial alveolar lavage deciphers GWAS-identified asthma genes. *Allergy* **70**(10): 1309–1318, 2015.

[275] Raj, T., Rothamel, K., Mostafavi, S., Ye, C., Lee, M.N., Replogle, J.M., Feng, T., Lee, M., Asinovski, N., Frohlich, I., Imboywa, S., Von Korff, A., Okada, Y., Patsopoulos, N.A., Davis, S., McCabe, C., Paik, H.I., Srivastava, G.P., Raychaudhuri, S., Hafler, D.A., Koller, D., Regev, A., Hacohen, N., Mathis, D., Benoist, C., Stranger, B.E., and De Jager, P.L. Polarization of the Effects of

Autoimmune and Neurodegenerative Risk Alleles in Leukocytes. *Science* **344**(6183): 519–523, 2014.

[276] Zeng, P., Wang, T., and Huang, S. *Cis*-SNPs Set Testing and PrediXcan Analysis for Gene Expression Data using Linear Mixed Models. *Scientific Reports* **7**(1): 15237, 2017.

[277] Xu, Z., Wu, C., Wei, P., and Pan, W. A Powerful Framework for Integrating eQTL and GWAS Summary Data. *Genetics* , 2017.

[278] Torres, J.M., Barbeira, A.N., Bonazzola, R., Morris, A.P., Shah, K.P., Wheeler, H.E., Bell, G., Cox, N.J., and Im, H.K. Integrative cross tissue analysis of gene expression identifies novel type 2 diabetes genes. *bioRxiv* , 2017.

[279] Burgess, S., Butterworth, A., and Thompson, S.G. Mendelian Randomization Analysis With Multiple Genetic Variants Using Summarized Data. *Genetic Epidemiology* **37**(7): 658–665, 2013.

[280] Liu, W. and Li, Q. An Efficient Elastic Net with Regression Coefficients Method for Variable Selection of Spectrum Data. *PLOS ONE* **12**(2): e0171122, 2017.

[281] Wheeler, H.E., Shah, K.P., Brenner, J., Garcia, T., Aquino-Michaels, K., GTEx Consortium, Cox, N.J., Nicolae, D.L., and Im, H.K. Survey of the Heritability and Sparse Architecture of Gene Expression Traits across Human Tissues. *PLoS Genetics* **12**(11): e1006423, 2016.

[282] Shabalin, A.A. Matrix eQTL: ultra fast eQTL analysis via large matrix operations. *Bioinformatics* **28**(10): 1353–1358, 2012.

[283] Friedman, J., Hastie, T., and Tibshirani, R. Regularization Paths for Generalized Linear Models via Coordinate Descent. *Journal of statistical software* **33**(1): 1–22, 2010.

[284] Warnes, G., Gorjanc, G., Leisch, F., and Man, M. genetics: Population Genetics , 2013.

[285] Obeidat, M., Hao, K., Bossé, Y., Nickle, D.C., Nie, Y., Postma, D.S., Laviolette, M., Sandford, A.J., Daley, D.D., Hogg, J.C., Elliott, W.M., Fishbane, N., Timens, W., Hysi, P.G., Kaprio, J., Wilson, J.F., Hui, J., Rawal, R., Schulz, H., Stubbe, B., Hayward, C., Polasek, O., Järvelin, M.R., Zhao, J.H., Jarvis, D., Kähönen, M., Franceschini, N., North, K.E., Loth, D.W., Brusselle, G.G., Smith, A.V., Gudnason, V., Bartz, T.M., Wilk, J.B., O'Connor, G.T., Cassano, P.A., Tang, W., Wain, L.V., Artigas, M.S., Gharib, S.A., Strachan, D.P., Sin, D.D., Tobin, M.D., London, S.J., Hall, I.P., and Paré, P.D. Molecular mechanisms underlying variations in lung function: a systems genetics analysis. *The Lancet Respiratory Medicine* **3**(10): 782–795, 2015.

[286] Häsler, R., Venkatesh, G., Tan, Q., Flachsbart, F., Sinha, A., Rosenstiel, P., Lieb, W., Schreiber, S., Christensen, K., Christiansen, L., and Nebel, A. Genetic interplay between human longevity and metabolic pathways — a large-scale eQTL study. *Aging Cell* **16**(4): 716–725, 2017.

[287] Fagny, M., Paulson, J.N., Kuijjer, M.L., Sonawane, A.R., Chen, C.Y., Lopes-Ramos, C.M., Glass, K., Quackenbush, J., and Platig, J. Exploring

regulation in tissues with eQTL networks. *Proceedings of the National Academy of Sciences* **114**(37): E7841–E7850, 2017.

[288] Brown, C.D., Mangravite, L.M., and Engelhardt, B.E. Integrative Modeling of eQTLs and *Cis*-Regulatory Elements Suggests Mechanisms Underlying Cell Type Specificity of eQTLs. *PLOS Genetics* **9**(8): e1003649, 2013.

[289] Wright, F.A., Sullivan, P.F., Brooks, A.I., Zou, F., Sun, W., Xia, K., Madar, V., Jansen, R., Chung, W., Zhou, Y.H., Abdellaoui, A., Batista, S., Butler, C., Chen, G., Chen, T.H., D'Ambrosio, D., Gallins, P., Ha, M.J., Hottenga, J.J., Huang, S., Kattenberg, M., Kochar, J., Middeldorp, C.M., Qu, A., Shabalin, A., Tischfield, J., Todd, L., Tzeng, J.Y., van Grootheest, G., Vink, J.M., Wang, Q., Wang, W., Wang, W., Willemsen, G., Smit, J.H., de Geus, E.J., Yin, Z., Penninx, B.W.J.H., and Boomsma, D.I. Heritability and genomics of gene expression in peripheral blood. *Nature Genetics* **46**: 430, 2014.

[290] Goldinger, A., Henders, A.K., McRae, A.F., Martin, N.G., Gibson, G., Montgomery, G.W., Visscher, P.M., and Powell, J.E. Genetic and Nongenetic Variation Revealed for the Principal Components of Human Gene Expression. *Genetics* **195**(3): 1117–1128, 2013.

[291] Waldmann, P., Mészáros, G., Gredler, B., Fuerst, C., and Sölkner, J. Evaluation of the lasso and the elastic net in genome-wide association studies. *Frontiers in Genetics* **4**: 270, 2013.

[292] Lu, F. and Petkova, E. A comparative study of variable selection methods in the context of developing psychiatric screening instruments. *Statistics in medicine* **33**(3): 401–421, 2014.

[293] Goldberger, A.S. A course in econometrics. *Harvard University Press* , 1991.

[294] Kumar, V., Westra, H.J., Karjalainen, J., Zhernakova, D.V., Esko, T., Hrdlickova, B., Almeida, R., Zhernakova, A., Reinmaa, E., Võsa, U., Hofker, M.H., Fehrmann, R.S.N., Fu, J., Withoff, S., Metspalu, A., Franke, L., and Wijmenga, C. Human Disease-Associated Genetic Variation Impacts Large Intergenic Non-Coding RNA Expression. *PLoS Genetics* **9**(1): e1003201, 2013.

[295] Hormozdiari, F., van de Bunt, M., Segrè, A., Li, X., Joo, J., Bilow, M., Sul, J., Sankararaman, S., Pasaniuc, B., and Eskin, E. Colocalization of GWAS and eQTL Signals Detects Target Genes. *The American Journal of Human Genetics* **99**(6): 1245–1260, 2016.

[296] Huang, Y.T., Liang, L., Moffatt, M.F., Cookson, W.O.C.M., and Lin, X. iGWAS: Integrative Genome-Wide Association Studies of Genetic and Genomic Data for Disease Susceptibility Using Mediation Analysis. *Genetic Epidemiology* **39**(5): 347–356, 2015.

[297] Karaderi, T., Drong, A.W., and Lindgren, C.M. Insights into the Genetic Susceptibility to Type 2 Diabetes from Genome-Wide Association Studies of Obesity-Related Traits. *Current Diabetes Reports* **15**: 83, 2015.

[298] Nica, A.C., Montgomery, S.B., Dimas, A.S., Stranger, B.E., Beazley, C., Barroso, I., and Dermitzakis, E.T. Candidate Causal Regulatory Effects by

Integration of Expression QTLs with Complex Trait Genetic Associations. *PLOS Genetics* **6**(4): e1000895, 2010.

[299] Wallace, C., Rotival, M., Cooper, J.D., Rice, C.M., Yang, J.H.M., McNeill, M., Smyth, D.J., Niblett, D., Cambien, F., Tiret, L., Todd, J.A., Clayton, D.G., and Blankenberg, S. Statistical colocalization of monocyte gene expression and genetic risk variants for type 1 diabetes. *Human Molecular Genetics* **21**(12): 2815–2824, 2012.

[300] Giambartolomei, C., Vukcevic, D., Schadt, E.E., Franke, L., Hingorani, A.D., Wallace, C., and Plagnol, V. Bayesian Test for Colocalisation between Pairs of Genetic Association Studies Using Summary Statistics. *PLOS Genetics* **10**(5): e1004383, 2014.

[301] Joehanes, R., Zhang, X., Huan, T., Yao, C., Ying, S., Nguyen, Q.T., Demirkale, C.Y., Feolo, M.L., Sharopova, N.R., Sturcke, A., Schäffer, A.A., Heard-Costa, N., Chen, H., Liu, P.c., Wang, R., Woodhouse, K.A., Tanriverdi, K., Freedman, J.E., Raghavachari, N., Dupuis, J., Johnson, A.D., O'Donnell, C.J., Levy, D., and Munson, P.J. Integrated genome-wide analysis of expression quantitative trait loci aids interpretation of genomic association studies. *Genome Biology* **18**(1): 16, 2017.

[302] Noble, J.A. and Valdes, A.M. Genetics of the HLA Region in the Prediction of Type 1 Diabetes. *Current diabetes reports* **11**(6): 533–542, 2011.

[303] Battle, A., Mostafavi, S., Zhu, X., Potash, J.B., Weissman, M.M., McCormick, C., Haudenschild, C.D., Beckman, K.B., Shi, J., Mei, R., Urban, A.E., Montgomery, S.B., Levinson, D.F., and Koller, D. Characterizing the genetic basis of transcriptome diversity through RNA-sequencing of 922 individuals. *Genome Research* **24**(1): 14–24, 2014.

[304] Burgess, S., Bowden, J., Fall, T., Ingelsson, E., and Thompson, S.G. Sensitivity Analyses for Robust Causal Inference from Mendelian Randomization Analyses with Multiple Genetic Variants. *Epidemiology* **28**(1): 30–42, 2017.

[305] Burgess, S. and Thompson, S.G. Interpreting findings from Mendelian randomization using the MR-Egger method. *European Journal of Epidemiology* **32**(5): 377–389, 2017.

[306] Swerdlow, D.I., Kuchenbaecker, K.B., Shah, S., Sofat, R., Holmes, M.V., White, J., Mindell, J.S., Kivimaki, M., Brunner, E.J., Whittaker, J.C., Casas, J.P., and Hingorani, A.D. Selecting instruments for Mendelian randomization in the wake of genome-wide association studies. *International Journal of Epidemiology* **45**(5): 1600–1616, 2016.

[307] The International Consortium for Blood Pressure Genome-Wide Association Studies. Genetic variants in novel pathways influence blood pressure and cardiovascular disease risk. *Nature* **478**: 103, 2011.

[308] Yavorska, O. MendelianRandomization: Mendelian Randomization Package , 2016.

[309] Wang, Z., Hildesheim, A., Wang, S.S., Herrero, R., Gonzalez, P., Burdette, L.,

Hutchinson, A., Thomas, G., Chanock, S.J., and Yu, K. Genetic Admixture and Population Substructure in Guanacaste Costa Rica. *PLOS ONE* **5**(10): e13336, 2010.

[310] Manning, A.K., Hivert, M.F., Scott, R.A., Grimsby, J.L., Bouatia-Naji, N., Chen, H., Rybin, D., Liu, C.T., Bielak, L.F., Prokopenko, I., Amin, N., Barnes, D., Cadby, G., Hottenga, J.J., Ingelsson, E., Jackson, A.U., Johnson, T., Kanoni, S., Ladenvall, C., Lagou, V., Lahti, J., Lecoeur, C., Liu, Y., Martinez-Larrad, M.T., Montasser, M.E., Navarro, P., Perry, J.R.B., Rasmussen-Torvik, L.J., Salo, P., Sattar, N., Shungin, D., Strawbridge, R.J., Tanaka, T., van Duijn, C.M., An, P., de Andrade, M., Andrews, J.S., Aspelund, T., Atalay, M., Aulchenko, Y., Balkau, B., Bandinelli, S., Beckmann, J.S., Beilby, J.P., Bellis, C., Bergman, R.N., Blangero, J., Boban, M., Boehnke, M., Boerwinkle, E., Bonnycastle, L.L., Boomsma, D.I., Borecki, I.B., Bottcher, Y., Bouchard, C., Brunner, E., Budimir, D., Campbell, H., Carlson, O., Chines, P.S., Clarke, R., Collins, F.S., Corbaton-Anchuelo, A., Couper, D., de Faire, U., Dedoussis, G.V., Deloukas, P., Dimitriou, M., Egan, J.M., Eiriksdottir, G., Erdos, M.R., Eriksson, J.G., Eury, E., Ferrucci, L., Ford, I., Forouhi, N.G., Fox, C.S., Franzosi, M.G., Franks, P.W., Frayling, T.M., Froguel, P., Galan, P., de Geus, E., Gigante, B., Glazer, N.L., Goel, A., Groop, L., Gudnason, V., Hallmans, G., Hamsten, A., Hansson, O., Harris, T.B., Hayward, C., Heath, S., Hercberg, S., Hicks, A.A., Hingorani, A., Hofman, A., Hui, J., Hung, J., Jarvelin, M.R., Jhun, M.A., Johnson, P.C.D., Jukema, J.W., Jula, A., Kao, W.H., Kaprio, J., Kardia, S.L.R., Keinanen-Kiukaanniemi, S., Kivimaki, M., Kolcic, I., Kovacs, P., Kumari, M., Kuusisto, J., Kyvik, K.O., Laakso, M., Lakka, T., Lannfelt, L., Lathrop, G.M., Launer, L.J., Leander, K., Li, G., Lind, L., Lindstrom, J., Lobbens, S., Loos, R.J.F., Luan, J., Lyssenko, V., Magi, R., Magnusson, P.K.E., Marmot, M., Meneton, P., Mohlke, K.L., Mooser, V., Morken, M.A., Miljkovic, I., Narisu, N., O'Connell, J., Ong, K.K., Oostra, B.A., Palmer, L.J., Palotie, A., Pankow, J.S., Peden, J.F., Pedersen, N.L., Pehlic, M., Peltonen, L., Penninx, B., Pericic, M., Perola, M., Perusse, L., Peyser, P.A., Polasek, O., Pramstaller, P.P., Province, M.A., Raikkonen, K., Rauramaa, R., Rehnberg, E., Rice, K., Rotter, J.I., Rudan, I., Ruokonen, A., Saaristo, T., Sabater-Lleal, M., Salomaa, V., Savage, D.B., Saxena, R., Schwarz, P., Seedorf, U., Sennblad, B., Serrano-Rios, M., Shuldiner, A.R., Sijbrands, E.J.G., Siscovick, D.S., Smit, J.H., Small, K.S., Smith, N.L., Smith, A.V., Stancakova, A., Stirrups, K., Stumvoll, M., Sun, Y.V., Swift, A.J., Tonjes, A., Tuomilehto, J., Trompet, S., Uitterlinden, A.G., Uusitupa, M., Vikstrom, M., Vitart, V., Vohl, M.C., Voight, B.F., Vollenweider, P., Waeber, G., Waterworth, D.M., Watkins, H., Wheeler, E., Widen, E., Wild, S.H., Willems, S.M., Willemsen, G., Wilson, J.F., Witteman, J.C.M., Wright, A.F., Yaghootkar, H., Zelenika, D., Zemunik, T., Zgaga, L., Wareham, N.J., McCarthy, M.I., Barroso, I., Watanabe, R.M., Florez, J.C., Dupuis, J., Meigs, J.B., and Langenberg, C. A genome-wide approach accounting for body mass index identifies genetic variants influencing fasting glycemic traits and insulin resistance. *Nat Genet* **44**(6): 659–669, 2012.

[311] Fadista, J., Manning, A.K., Florez, J.C., and Groop, L. The (in)famous GWAS

P-value threshold revisited and updated for low-frequency variants. *European Journal of Human Genetics* **24**(8): 1202–1205, 2016.

[312] Das, S.K. and Sharma, N.K. Expression quantitative trait analyses to identify causal genetic variants for type 2 diabetes susceptibility. *World Journal of Diabetes* **5**(2): 97–114, 2014.

[313] Kemiläinen, H., Adam, M., Mäki-Jouppila, J., Damdimopoulou, P., Damdimopoulos, A.E., Kere, J., Hovatta, O., Laajala, T.D., Aittokallio, T., Adamski, J., Ryberg, H., Ohlsson, C., Strauss, L., and Poutanen, M. The Hydroxysteroid (17β) Dehydrogenase Family Gene HSD17B12 Is Involved in the Prostaglandin Synthesis Pathway, the Ovarian Function, and Regulation of Fertility. *Endocrinology* **157**(10): 3719–3730, 2016.

[314] Moon, Y.A. and Horton, J.D. Identification of Two Mammalian Reductases Involved in the Two-carbon Fatty Acyl Elongation Cascade. *Journal of Biological Chemistry* **278**(9): 7335–7343, 2003.

[315] Muise, E.S., Azzolina, B., Kuo, D.W., El-Sherbeini, M., Tan, Y., Yuan, X., Mu, J., Thompson, J.R., Berger, J.P., and Wong, K.K. Adipose Fibroblast Growth Factor 21 Is Up-Regulated by Peroxisome Proliferator-Activated Receptor γ and Altered Metabolic States. *Molecular Pharmacology* **74**(2): 403–412, 2008.

[316] Muise, E.S., Souza, S., Chi, A., Tan, Y., Zhao, X., Liu, F., Dallas-Yang, Q., Wu, M., Sarr, T., Zhu, L., Guo, H., Li, Z., Li, W., Hu, W., Jiang, G., Paweletz, C.P., Hendrickson, R.C., Thompson, J.R., Mu, J., Berger, J.P., and Mehmet, H. Downstream Signaling Pathways in Mouse Adipose Tissues Following Acute In Vivo Administration of Fibroblast Growth Factor 21. *PLoS ONE* **8**(9): e73011, 2013.

[317] NCBI Resource Coordinators. Database Resources of the National Center for Biotechnology Information. *Nucleic Acids Research* **45**(D1): D12–D17, 2017.

[318] Waddleton, D., Wu, W., Feng, Y., Thompson, C., Wu, M., Zhou, Y.P., Howard, A., Thornberry, N., Li, J., and Mancini, J.A. Phosphodiesterase 3 and 4 comprise the major cAMP metabolizing enzymes responsible for insulin secretion in INS-1 (832/13) cells and rat islets. *Biochemical Pharmacology* **76**(7): 884–893, 2008.

[319] Dov, A., Abramovitch, E., Warwar, N., and Nesher, R. Diminished Phosphodiesterase-8B Potentiates Biphasic Insulin Response to Glucose. *Endocrinology* **149**(2): 741–748, 2008.

[320] Tian, G., Sågetorp, J., Xu, Y., Shuai, H., Degerman, E., and Tengholm, A. Role of phosphodiesterases in the shaping of sub-plasma-membrane cAMP oscillations and pulsatile insulin secretion. *Journal of Cell Science* **125**(21): 5084–5095, 2012.

[321] Karol, S.E., Mattano, L.A., Yang, W., Maloney, K.W., Smith, C., Liu, C., Ramsey, L.B., Fernandez, C.A., Chang, T.Y., Neale, G., Cheng, C., Mardis, E., Fulton, R., Scheet, P., San Lucas, F.A., Larsen, E.C., Loh, M.L., Raetz, E.A., Hunger, S.P., Devidas, M., and Relling, M.V. Genetic risk factors for the development of osteonecrosis in children under age 10 treated for acute

lymphoblastic leukemia. *Blood* **127**(5): 558–564, 2016.

[322] Lecompte, S., Pasquetti, G., Hermant, X., Grenier-Boley, B., Gonzalez-Gross, M., De Henauw, S., Molnar, D., Stehle, P., Béghin, L., Moreno, L.A., Amouyel, P., Dallongeville, J., and Meirhaeghe, A. Genetic and Molecular Insights Into the Role of *PROX1* in Glucose Metabolism. *Diabetes* **62**(5): 1738–1745, 2013.

[323] Horikoshi, M., Mägi, R., van de Bunt, M., Surakka, I., Sarin, A.P., Mahajan, A., Marullo, L., Thorleifsson, G., Hägg, S., Hottenga, J.J., Ladenvall, C., Ried, J.S., Winkler, T.W., Willems, S.M., Pervjakova, N., Esko, T.o., Beekman, M., Nelson, C.P., Willenborg, C., Wiltshire, S., Ferreira, T., Fernandez, J., Gaulton, K.J., Steinthorsdottir, V., Hamsten, A., Magnusson, P.K.E., Willemsen, G., Milaneschi, Y., Robertson, N.R., Groves, C.J., Bennett, A.J., Lehtimäki, T., Viikari, J.S., Rung, J., Lyssenko, V., Perola, M., Heid, I.M., Herder, C., Grallert, H., Müller-Nurasyid, M., Roden, M., Hypponen, E., Isaacs, A., van Leeuwen, E.M., Karssen, L.C., Mihailov, E., Houwing-Duistermaat, J.J., de Craen, A.J.M., Deelen, J., Havulinna, A.S., Blades, M., Hengstenberg, C., Erdmann, J., Schunkert, H., Kaprio, J., Tobin, M.D., Samani, N.J., Lind, L., Salomaa, V., Lindgren, C.M., Slagboom, P.E., Metspalu, A., van Duijn, C.M., Eriksson, J.G., Peters, A., Gieger, C., Jula, A., Groop, L., Raitakari, O.T., Power, C., Penninx, B.W.J.H., de Geus, E., Smit, J.H., Boomsma, D.I., Pedersen, N.L., Ingelsson, E., Thorsteinsdottir, U., Stefansson, K., Ripatti, S., Prokopenko, I., McCarthy, M.I., Morris, A.P., and Engage Consortium. Discovery and Fine-Mapping of Glycaemic and Obesity-Related Trait Loci Using High-Density Imputation. *PLOS Genetics* **11**(7): e1005230, 2015.

[324] Thomsen, S.K., Ceroni, A., van de Bunt, M., Burrows, C., Barrett, A., Scharfmann, R., Ebner, D., McCarthy, M.I., and Gloyn, A.L. Systematic Functional Characterization of Candidate Causal Genes for Type 2 Diabetes Risk Variants. *Diabetes* , 2016.

[325] Mellado-Gil, J.M., Fuente-Martín, E., Lorenzo, P.I., Cobo-Vuilleumier, N., López-Noriega, L., Martín-Montalvo, A., Gómez, I.d.G.H., Ceballos-Chávez, M., Gómez-Jaramillo, L., Campos-Caro, A., Romero-Zerbo, S.Y., Rodríguez-Comas, J., Servitja, J.M., Rojo-Martinez, G., Hmadcha, A., Soria, B., Bugliani, M., Marchetti, P., Bérmudez-Silva, F.J., Reyes, J.C., Aguilar-Diosdado, M., and Gauthier, B.R. The type 2 diabetes-associated HMG20A gene is mandatory for islet beta cell functional maturity. *Cell Death & Disease* **9**(3): 279, 2018.

[326] Westra, H.J., Peters, M.J., Esko, T.o., Yaghootkar, H., Schurmann, C., Kettunen, J., Christiansen, M.W., Fairfax, B.P., Schramm, K., Powell, J.E., Zhernakova, A., Zhernakova, D.V., Veldink, J.H., Van den Berg, L.H., Karjalainen, J., Withoff, S., Uitterlinden, A.G., Hofman, A., Rivadeneira, F., 't Hoen, P.A.C., Reinmaa, E., Fischer, K., Nelis, M., Milani, L., Melzer, D., Ferrucci, L., Singleton, A.B., Hernandez, D.G., Nalls, M.A., Homuth, G., Nauck, M., Radke, D., Völker, U., Perola, M., Salomaa, V., Brody, J., Suchy-Dicey, A., Gharib, S.A., Enquobahrie, D.A., Lumley, T., Montgomery, G.W., Makino, S., Prokisch, H., Herder, C., Roden, M., Grallert, H., Meitinger, T., Strauch, K., Li, Y., Jansen, R.C., Visscher, P.M., Knight, J.C., Psaty, B.M.,

Ripatti, S., Teumer, A., Frayling, T.M., Metspalu, A., van Meurs, J.B.J., and Franke, L. Systematic identification of trans eQTLs as putative drivers of known disease associations. *Nature Genetics* **45**: 1238, 2013.

[327] Steiner, D.J., Kim, A., Miller, K., and Hara, M. Pancreatic islet plasticity: Interspecies comparison of islet architecture and composition. *Islets* **2**(3): 135–145, 2010.

[328] Bramswig, N.C., Everett, L.J., Schug, J., Dorrell, C., Liu, C., Luo, Y., Streeter, P.R., Naji, A., Grompe, M., and Kaestner, K.H. Epigenomic plasticity enables human pancreatic α to β cell reprogramming. *The Journal of Clinical Investigation* **123**(3): 1275–1284, 2013.

[329] Haque, A., Engel, J., Teichmann, S.A., and Lönnberg, T. A practical guide to single-cell RNA-sequencing for biomedical research and clinical applications. *Genome Medicine* **9**(1): 75, 2017.

[330] Svensson, V., Vento-Tormo, R., and Teichmann, S.A. Exponential scaling of single-cell RNA-seq in the past decade. *Nature Protocols* **13**: 599, 2018.