Mechanisms of genome regulation in human islets and their role in the pathogenesis of type 2 diabetes

A thesis submitted to the University of Oxford



in candidature for the degree of Doctor of Philosophy

by

Martijn van de Bunt

Green Templeton College, University of Oxford

and

Oxford Centre for Diabetes, Endocrinology and Metabolism

Radcliffe Department of Medicine, University of Oxford

Michaelmas 2013

Declaration

The work in this thesis is the original work of the author. Experiments were conducted at the Oxford Centre for Diabetes Endocrinology and Metabolism, University of Oxford, under the supervision of Dr. Anna Gloyn and Prof. Mark McCarthy. Funding was provided by the Radcliffe Department of Medicine, University of Oxford.

The islet miRNA expression profile (chapter three) and its intersection with type 2 diabetes (T2D) genetics (section 6.3.2) has been published in *PLOS ONE*. The manuscript (van de Bunt *et al.,* 2013) is provided as an appendix.

Genotype calling for the expression profiling, expression quantitative trait and allele-specific expression analyses in islets (chapter four) was performed by Neil Robertson (Oxford). Islet functional elements (also used in chapter six) were generated in collaboration with Prof. Jorge Ferrer (Barcelona/London) and form part of a manuscript (Pasquali *et al.*, 2014) published in *Nature Genetics*, provided as an appendix. Chromatin signatures for other tissues were generated by Dr. Kyle Gaulton (Oxford).

Meta-analysis data used for the evaluation of fine-mapping approaches (chapter five) were obtained from the International Genetics of Ankylosing Spondylitis Consortium (Ankylosing spondylitis – analysis performed Adrian Cortes) and the GoT2D study (T2D – analysis performed by Clement Ma – also used in chapter six).

Whole-genome and -exome sequencing data used for the evaluation of intersecting islet annotations with T2D genetics data (chapter six) were generated by the GoT2D and T2D-GENES studies. Islet lncRNAs were generated in collaboration with Prof. Jorge Ferrer and the resulting manuscript (Morán *et al.*, 2012) published in *Cell Metabolism* is provided as an appendix.

This thesis is approximately 40,000 words.

Mechanisms of genome regulation in human islets and their role in the pathogenesis of type 2 diabetes

Martijn van de Bunt, Green Templeton College

Submitted in candidature for the degree of Doctor of Philosophy,

Michaelmas term 2013

Abstract

Genome-wide association studies (GWAS) have made substantial progress in implicating genomic regions in type 2 diabetes (T2D) susceptibility. Whilst attributing causal mechanisms to loci has proved non trivial, these studies have provided insights into the genetic architecture underlying the disease. GWAS findings indicate a causal role for gene regulatory processes, and suggest that pancreatic beta-cells play a pivotal role in mediating common T2D association. Work presented in this thesis therefore sought to generate novel regulatory annotations from human islets, and to assess whether T2D-associated loci can be accurately fine-mapped using statistical approaches, with the aim of improving understanding of causal mechanisms underlying these associations through integration of the two approaches.

Using small RNA sequencing in human islets and enriched beta-cell populations (both n=3) and mRNA sequencing in a large number of human islets (n=130), I increased the number of available human islet annotations. These studies identified high or islet-specific expression in many micro RNAs (miRNAs) without previously known roles in human islets. It also provided the largest study of quantitative trait loci (eQTLs) and allele-specific expression (ASE) in human islets to date, identifying significant eQTLs for 1,636 genes and significant ASE at 8,754 genes. There was enrichment of active islet chromatin, compared to other tissues, at the best eQTL variant for each gene, but also substantial sharing of significant eQTLs between islets and other tissues.

Simulations were used to assess the utility of fine-mapping approaches for refining common disease-associated loci to smaller intervals or sets of variants likely to include the causal variant. The results demonstrated that fine-mapping can indeed refine these loci to sets or intervals of a size more amenable to functional follow-up or focussed intersection with high quality annotations. Furthermore, using an approximated Bayesian approach, I was able to refine twenty-one of the known common T2D-associated loci.

Finally, using the newly generated annotations, I demonstrated enrichment of T2D association signal for regulatory RNA annotations (islet lncRNAs and miRNA target gene sets). I also identified examples in which these types of annotation overlap common and rare variation suggestive of a role in T2D pathogenesis. Using further islet annotations, I also uncovered potential causal mechanisms at four of the twenty-one fine-mapped common T2D loci.

These data therefore provide many novel islet regulatory annotations that can be intersected with T2D genetics, and provide a first example of how such an approach can lead to novel potential causal mechanisms underlying association loci.

Acknowledgements

You might have only arrived at page four of many that are to follow in the thesis, but this section was the last left to be written. It means the end of four years, which in January 2010 seemed such a long time away, has somehow suddenly been reached. After this page follow 40,000 words or so attempting to summarise all my (scientific) adventures and new experiences during this time. However, this page I will use to thank all the people who have made it such a great experience – people with whom I shared many happy moments and who kept me sane during times of stress.

Many thanks go out to my supervisors, Mark McCarthy and Anna Gloyn, for being patient and supportive along the way, and providing me with invaluable advice on the world of science. Thanks also go to Jorge Ferrer for making me feel at home during my visit to Barcelona, to Katharine Owen, for without your referral of my internship question to Anna back in 2006 I might have never arrived here, and to Kyle Gaulton for teaching me most of my coding skills and becoming a good friend in the process.

I would also like to thank all those people in Oxford who have made the past four (and in some cases more) years disappear so quickly. OCDEM and its coffee breaks have been the source of many laughs and new found knowledge. My Oxford experience would not have been the same without all its members present (Agata, Alex, Amy, Anne R., Carla, Chris, Fainia, Jana, Mandy, Matt S., Natasha, Neelam, Nicola, Saima, Søren) and past (Aparna, Bahram, Gaya, Laura, Matt R., Nick, Olly, Peta). Also thanks to my roommates (Juan, Loukas, Manny, Matthias, Patrick) and those upstairs (Andrew, Anubha, Neil, Will) at the WTCHG for providing infinite advice, excellent banter and many interesting conversations around the world. I count many of you as good friends.

Thank you, Mary, for being there (during the ups and downs of thesis writing) and for being senior editor of this thesis. Without you the number of “the”-s in it would have been substantially larger, and I would have been much more stressed.

Thank you.

Foppe, Silvia, Daphne, Michiel, Sicco, Roos, Herman, Manon, Mark, Chris – jullie bedank ik ook voor jullie steun en vriendschap. Ondanks dat ik Nederland verlaten heb, zijn jullie mij niet vergeten. Ik hoop dat de toekomst ons nog vele nieuwe avonturen gaat brengen, waar we allemaal uiteindelijk ook terecht komen in Nederland (of andere delen van de wereld).

Als laatste bedank ik jullie, mijn ouders en zus, omdat jullie er altijd voor me zijn en waren. Ik weet dat het niet altijd even makkelijk is om bij elkaar langs te komen met een zee ertussen, maar zonder jullie steun, aanmoedigingen en verhalen van thuis zou dit proefschrift er niet zijn gekomen.

Bedankt.

List of common abbreviations

(A)BF	(approximated) Bayes factor
ADP	Adenosine diphosphate
ASE	Allele-specific expression
ATP	Adenosine triphosphate
BMI	Body mass index
bp	Base pairs
ChIP-seq	Chromatin Immunoprecipitation sequencing
Chr	Chromosome
CEU	Utah Residents with Northern and Western European ancestry
DIAGRAM	Diabetes genetics replication and meta-analysis consortium
DNA	Deoxyribonucleic acid
ENCODE	Encyclopedia of DNA Elements (ENCODE) project
eQTL	Expression quantitative trait loci
gDNA	genomic DNA
FACS	fluorescence-activated cell sorting
GSIS	Glucose-stimulated insulin secretion
GWAS	Genome-wide association study
HOMA-B	Homeostatic model assessment of beta-cell function
HOMA-IR	Homeostatic model assessment of insulin resistance
Kb	Kilobase
LD	Linkage disequilibrium
lncRNA	long non-coding ribonucleic acid
MAF	Minor allele frequency
Mb	Megabase

miRNA	Micro ribonucleic acid
MODY	Maturity onset diabetes of the young
mRNA	Messenger ribonucleic acid
NCBI	National Centre for Biotechnology Information
OR	Odds ratio
OXCIT	Oxford centre for islet transplantation
PBS	Phosphate buffered saline
PCR	Polymerase chain reaction
qPCR	Quantitative polymerase chain reaction
RAF	Risk allele frequency
RIN	RNA integrity number
RNA	Ribonucleic acid
RNA-seq	Messenger ribonucleic acid sequencing
RT	Reverse transcriptase
T1D	Type 1 diabetes
T2D	Type 2 diabetes
T2D-GENES	Type 2 Diabetes Genetic Exploration by Next-generation sequencing in multi-Ethnic Samples study
UTR	Untranslated region
WHO	World health organisation
YRI	Yoruba in Ibadan, Nigera

Contents

CHAPTER1: Introduction

1.1 The impact of Type 2 diabetes on global health 2

1.2 The pathophysiology of T2D 2

1.2.1 Clinical presentation and diagnostic criteria..... 2

1.2.2 The aetiology of a complex disorder..... 4

1.2.2.a Obesity and insulin resistance 4

1.2.2.b Insulin secretion and functional beta-cell mass 6

1.3 Genetics and T2D 9

1.3.1 The heritability of T2D predisposition 9

1.3.2 Insights gained through studying T2D genetics..... 9

1.3.2.a Limited role for protein-coding variation 10

1.3.2.b Central role for pancreatic beta-cells in T2D predisposition.....14

1.4 Translating association signals into causal mechanisms 16

1.4.1 The challenges surrounding functional translation 16

1.4.2 Capturing the allele frequency spectrum 19

1.4.3 Refining association signals through fine-mapping approaches 21

1.4.4 Expression profiling and eQTL analysis 22

1.4.5 Epigenetic modulation 25

1.5 Thesis aims 29

CHAPTER2: General Methods

2.1 Human islet tissue bank 31

2.1.1 RNA extraction from human islets 31

2.1.2 DNA extraction from spleen for genotyping 32

2.1.3 DNA extraction from human islets for genotyping 32

2.2 Human islet RNA quality assessment 33

2.3 DNA quality assessment 34

2.3.1 PCR reaction components and cycling conditions 34

2.3.2 Agarose gel electrophoresis of PCR products 35

2.4 Quantitative PCR to estimate human islet purity 36

2.4.1 cDNA synthesis 36

2.4.2 Quantitative real-time PCR of the islet purity panel 36

2.5 Fluorescence-activated cell sorting (FACS) 38

2.5.1 Preparation of Krebs buffers for FACS 38

2.5.2 FACS enrichment of human beta-cells 39

2.6 Studies used for genetic analyses 40

2.6.1 Genetics of Type 2 Diabetes (GoT2D) study 41

2.6.2 Type 2 Diabetes Genetic Exploration by Next-generation sequencing in multi-Ethnic Samples (T2D-GENES) study 41

2.7 Bioinformatics and statistical methods 42

2.7.1 Determining overlap between variants and annotations 42

2.7.2 Variant type annotation with the Variant Effect Predictor tool 42

2.7.3 Statistical analyses 43

CHAPTER3: The miRNA profile of pancreatic islets and beta-cells

3.1 Introduction.....45

3.1.1 miRNAs are important regulatory RNAs..... 45

3.1.2 Dysregulation of miRNAs in disease..... 46

3.1.3 miRNAs in glucose homeostasis 51

3.2 Samples and Methods 52

3.2.1 Human pancreatic islets and beta-cells 52

3.2.2 Library preparation and sequencing 52

3.2.3 Raw sequence processing..... 52

3.2.4 Sequence mapping and quantification 53

3.2.5 Normalization 54

3.2.6 Tissue specific analysis 54

3.3 Results..... 57

3.3.1 miRNA expression in human pancreatic islets and beta-cells.....57

3.3.2 Tissue specificity of miRNA expression..... 63

3.4 Discussion 68

CHAPTER4: Transcriptome analysis of human pancreatic islets

4.1 Introduction 71

4.1.1 Transcriptome data can provide mechanistic insights at GWAS loci..... 71

4.1.2 RNA-seq and long non-coding RNAs.....77

4.1.3 Islet transcriptome analysis can uncover genes dysregulated in T2D..... 78

4.2 Methods.....79

4.2.1 Human pancreatic islets....79
4.2.2 Genotyping and imputation....80
4.2.3 Population stratification in the genotype data....81
4.2.4 RNA sequencing....84
4.2.5 mRNA quantification....84
4.2.6 Sample swap and outlier detection....84
4.2.7 Normalization of quantifications....85
4.2.8 QTL analysis....88
4.2.9 Allele-specific expression analysis....88
4.2.10 Gene set enrichment analysis and functional annotation....89

4.3 Results....91

4.3.1 Expression profile of human pancreatic islets....91
4.3.2 Effects of genetic variation on islet gene expression levels....96
4.3.3 Effects of genetic variation on allelic expression in islets....103

4.4 Discussion....108

CHAPTER5: Fine-mapping of common association loci

5.1 Introduction....111

5.1.1 Fine-mapping of GWAS loci: needs and difficulties....111
5.1.2 Fine-mapping strategies: trans-ethnic and Bayesian approaches....115

5.2 Methods....118

5.2.1 Bayesian approximation....118
5.2.2 Description of the simulations pipeline....119

5.3 Results 122

5.3.1 Establishing the usefulness and accuracy of fine-mapping in a Bayesian framework 122

5.3.2 Multi-ethnic study designs for fine-mapping 129

5.3.3 From simulations to empirical data 132

5.3.4 Application to T2D association data 137

5.4 Discussion 143

CHAPTER6: Intersecting islet annotations with T2D association data

6.1 Introduction 146

6.2 Methods 148

6.2.1 Enrichment of islet lncRNAs and miRNA targets in GWAS data 148

6.2.2 Variant testing of islet lncRNAs, miRNAs and miRNA targets in whole-genome sequencing data 150

6.2.3 Identify causal variants at fine-mapped loci using islet annotations 151

6.3 Results 153

6.3.1 Enrichment of common T2D association in islet annotations 153

6.3.2 Investigating T2D association of variation in islet regulatory RNA annotations across the allele-frequency spectrum 158

6.3.3 Using islet annotations to uncover potential causal mechanisms at fine-mapped common T2D association loci 166

6.4 Discussion 174

CHAPTER7: General Discussion.....177

APPENDICES

Tables.....186

Appendix Table 1 186

Appendix Table 2.....193

Appendix Table 3..... 204

References.....207

Published manuscripts..... 221

Chapter 1

Introduction

1.1 The impact of Type 2 diabetes on global health

Type 2 diabetes (T2D) is having a substantial impact on global health. Worldwide, T2D prevalence in the adult population has increased from approximately 8% in 1980 to 10% in 2011. More than 350 million people are currently affected[1] – a figure which is projected to increase to 550 million by 2030[2]. The disease has a major impact on global mortality and morbidity. For example, diabetic retinopathy, a common complication of T2D, is the leading cause of blindness in developed countries[2,3]. However, T2D is much more than a disease of the developed world. Eighty per cent of all affected individuals live in low- and middle-income countries[2], and 80% of the estimated 4.6 million disease-attributable deaths in 2011 occurred in these developing countries[2,3].

If no improvements in prevention are made, the number of people affected in the United Kingdom (UK) will rise from 3.4 million now to around 5.4 million in 2035[4]. In addition to the impact on quality of life, morbidity and mortality, this increase will have substantial economic implications. The cost burden for the NHS(National Health Service) in the same timeframe is predicted to rise (in real terms) from £8.8bn to £15.1bn, accounting for 17% of available budget by 2035[4]. Indirect costs, including working days lost through sickness or caregiving, are even higher – rising from £13bn to £20.4bn[4].

1.2 The pathophysiology of T2D

1.2.1 Clinical presentation and diagnostic criteria

Diabetes mellitus encompasses a group of metabolic disorders characterised by hyperglycaemia, underpinned by a variety of aetiologies. T2D is by far the most

abundant of this group, accounting for 90-95% of all diabetes incidence. The remaining cases are made up of type 1 diabetes (T1D) caused by auto-immune destruction of beta-cells (5-10%), maturity-onset diabetes of the young (MODY) associated with monogenetic defects in beta-cell function (1-2%), as well as several rarer, often syndromic, forms of diabetes.

The textbook clinical presentation suggests a patient with T2D presents as an obese individual with polydipsia (excessive thirst) and polyuria (excessive urine production/urination), often in combination with other symptoms of prolonged hyperglycaemia such as unexplained weight loss and fatigue. However, in reality a substantial proportion of affected individuals display very mild symptoms, which may go unnoticed for many years. Whilst remaining undiagnosed and untreated, affected individuals are already increasing their long-term risk of diabetic complications.

Diagnostic criteria for T2D in World Health Organisation (WHO) and American Diabetes Association (ADA) guidelines are based on three alternative measures of glycaemia: plasma glucose levels (either fasting >7 mmol/L or random >11.1 mmol/L), two hour plasma glucose concentration > 11.1 mmol/L after an oral glucose tolerance test (OGTT) and, more recently accepted, glycated haemoglobin (HbA1c) > 48 mmol/mol[5-7]. In the presence of symptomatic diabetes any of these findings is considered diagnostic, but if no clear symptoms are present then diabetes should be confirmed by repeat testing.

T2D is sometimes seen as a “mild” form of diabetes, since, unlike T1D, the glycaemic imbalance rarely results directly in life-threatening episodes. However, as outlined in the previous section, its long term complications make T2D a major cause of disability and premature death. Diabetic patients have a 2-3 times increased risk of macrovascular complications – including myocardial infarctions and stroke – compared to the population average[8]. Microvascular complications make T2D a

leading cause of kidney failure and visual disability, and put patients at significant risk of lower limb amputation. Long term follow-up trials, primarily the United Kingdom Prospective Diabetes Study (UKPDS), show that hyperglycaemia itself is one of the major causes of cardiovascular complications, highlighting the importance of achieving good glycaemic control[9,10].

1.2.2 The aetiology of a complex disorder

Despite decades of research into T2D, there are still many unknowns in its aetiology. It is characterised by relative pancreatic beta-cell insulin secretory insufficiency in the face of peripheral insulin resistance, without signs of autoimmune damage to the pancreatic islet[11]. The molecular mechanisms underlying these defects are not fully understood and the relative contributions of insulin secretory and sensitivity defects remain unclear. Given the diverse presentations of T2D, the contribution of defects is likely to differ between patients[5].

1.2.2.a Obesity and insulin resistance

Insulin plays a central role in the regulation of carbohydrate, fat and protein metabolism in the body. It is an anabolic hormone and has its effect by binding to insulin receptors on the cell surface. At basal levels, insulin reduces hepatic glucose production by inhibiting gluconeogenesis and lipolysis in adipose tissue. Higher postprandial insulin levels promote uptake of glucose from the bloodstream and its subsequent conversion into glycogen (glycogenesis) in liver and muscle, triglyceride formation (lipogenesis) in adipose tissue, and protein synthesis in a range of tissues.

Insulin resistant cells lose their ability to respond appropriately to circulating insulin, causing reduced glucose uptake, inappropriate gluconeogenesis and increased lipolysis[12]. Severe impairments of the insulin signalling pathway, such as those seen in monogenic insulin resistance disorders, give a powerful insight into the effects of

cellular disturbance of insulin sensitivity on maintaining normal glucose homeostasis. Monogenic insulin resistance is often caused by mutations in the insulin receptor gene (*INSR*)[13-15], but mutations in up to fifteen other genes, including the insulin signalling pathway gene *AKT2*[16], and *PPARγ*[17] and *LMNA*[18,19] – leading to partial lypodystrophy – have now also been reported. All these patients display marked hyperinsulinaemia, often have disturbed glycaemia and present with a range of associated features of varying severity, including growth retardation, hyperandrogenism, and lipodystrophy. Patients with *INSR* mutations have particularly high morbidity and mortality at an early age, with recessive frameshift and nonsense *INSR* mutations causing Donahue syndrome and frequently fatality in the first year of life.

Insulin resistance in T2D affects the two major sites of glucose disposal – liver and skeletal muscle - causing impaired suppression of hepatic glucose output and decreased insulin-stimulated glucose uptake and metabolism in skeletal muscle[20]. The mechanisms leading to insulin resistance in T2D are incompletely understood, but there is strong evidence associating it, at least partially, to obesity[11,20,21]. Both clinical and epidemiological observations demonstrate a strong association between obesity, insulin resistance and T2D[5,20]. The causal link between obesity and insulin resistance is further supported by the improvements in insulin sensitivity seen after weight reduction - be it achieved through dietary means or bariatric surgery – which can even lead to remission of T2D[20,22,23]. However, the link between obesity and insulin resistance is undoubtedly complex, as illustrated by patients with *PTEN* haploinsufficiency who were recently described to be constitutively more insulin sensitive while also showing increased levels of adiposity[24].

The pharmacological arsenal for treatment of T2D contains several classes of drugs which sensitise peripheral tissues to circulating insulin. Thiazolidinediones are specifically aimed at increasing insulin sensitivity and act through binding and activating one or more peroxisome proliferator-activated receptors (PPARs)[25]. These

nuclear regulatory proteins activate gene expression networks leading to enhanced free-fatty acid and glucose uptake. However, the concomitant increase in cardiovascular disease risk has led to concerns over the use of this type of drug[26]. Biguanides are another class of drugs affecting insulin sensitivity and contain the most commonly used antidiabetic agent; metformin. Most notably in the United Kingdom Prospective Diabetes Study (UKPDS), metformin has been shown to effectively reduce all-cause mortality and diabetes-related endpoints[9,27]. Whilst their effects are not completely understood, biguanides are thought to reduce hepatic gluconeogenesis and increase peripheral glucose uptake by activating AMP-activated protein kinase[28].

1.2.2.b Insulin secretion and functional beta-cell mass

Whilst insulin sensitivity clearly plays a role in the development of T2D, some individuals can display normal glucose tolerance in the face of marked insulin resistance[11]. This is because beta-cells can compensate for insulin resistance by increasing their insulin output[29]. Only when this compensatory mechanism fails will T2D develop.

Figure 1.1 shows the normal process of glucose-stimulated insulin secretion (GSIS). Glucose enters the beta-cell by facilitated diffusion through glucose transporters, of which GLUT1 is most abundant in humans[30]. Subsequently, intracellular glucose is phosphorylated by glucokinase (GCK) to form glucose-6-phosphate[31], which then undergoes glycolysis and finally generates ATP. The rise in ATP closes ATP-sensitive potassium (K_{ATP}) channels on the cell membrane by increasing the ATP:ADP ratio, thereby depolarising the beta-cell from its resting potential[32]. Membrane depolarisation in turn opens voltage-gated calcium channels, which triggers an influx of calcium and induces insulin exocytosis from secretory vesicles.

Monogenic disorders causing beta-cell dysfunction have provided valuable insights into the effect of disrupting these processes on insulin secretion in humans. Mutations in both genes encoding the two subunits of the beta-cell K_{ATP}-channel – *ABCC8*, encoding the sulfonylurea receptor 1 (SUR1) subunit, and *KCNJ11*, encoding the inwardly rectifying potassium channel Kir6.2 – are the most common causes of both permanent neonatal diabetes and the opposing phenotype of hyperinsulinaemic hypoglycaemia (HI), through activation and inactivation of the channel respectively[33-35].Similar opposite phenotypes are observed with mutations in the glucokinase gene, *GCK*. Heterozygous inactivating mutations, which reduce the affinity of glucokinase for glucose and set the threshold for GSIS at a slightly increased blood glucose level, are one of the commonest causes of MODY[36,37], and a rare cause of neonatal diabetes in the homozygous state[38]. Conversely, heterozygous activating *GCK* mutations, which lower the threshold for GSIS and result in insulin secretion at inappropriately low blood glucose levels, are a rare cause of HI[39].

These findings from monogenic forms of beta-cell dysfunction make it clear that protein such as Kir6.2 and SUR1 play pivotal and non-redundant roles in regulating GSIS. It also makes them excellent targets for pharmacologic intervention. Although development of GCK activators is hampered by non-pancreatic side effects, the K_{ATP}-channel is indeed the target of the sulphonylureas, a commonly used class of oral antidiabetic agents which increase insulin secretion from the beta-cell.

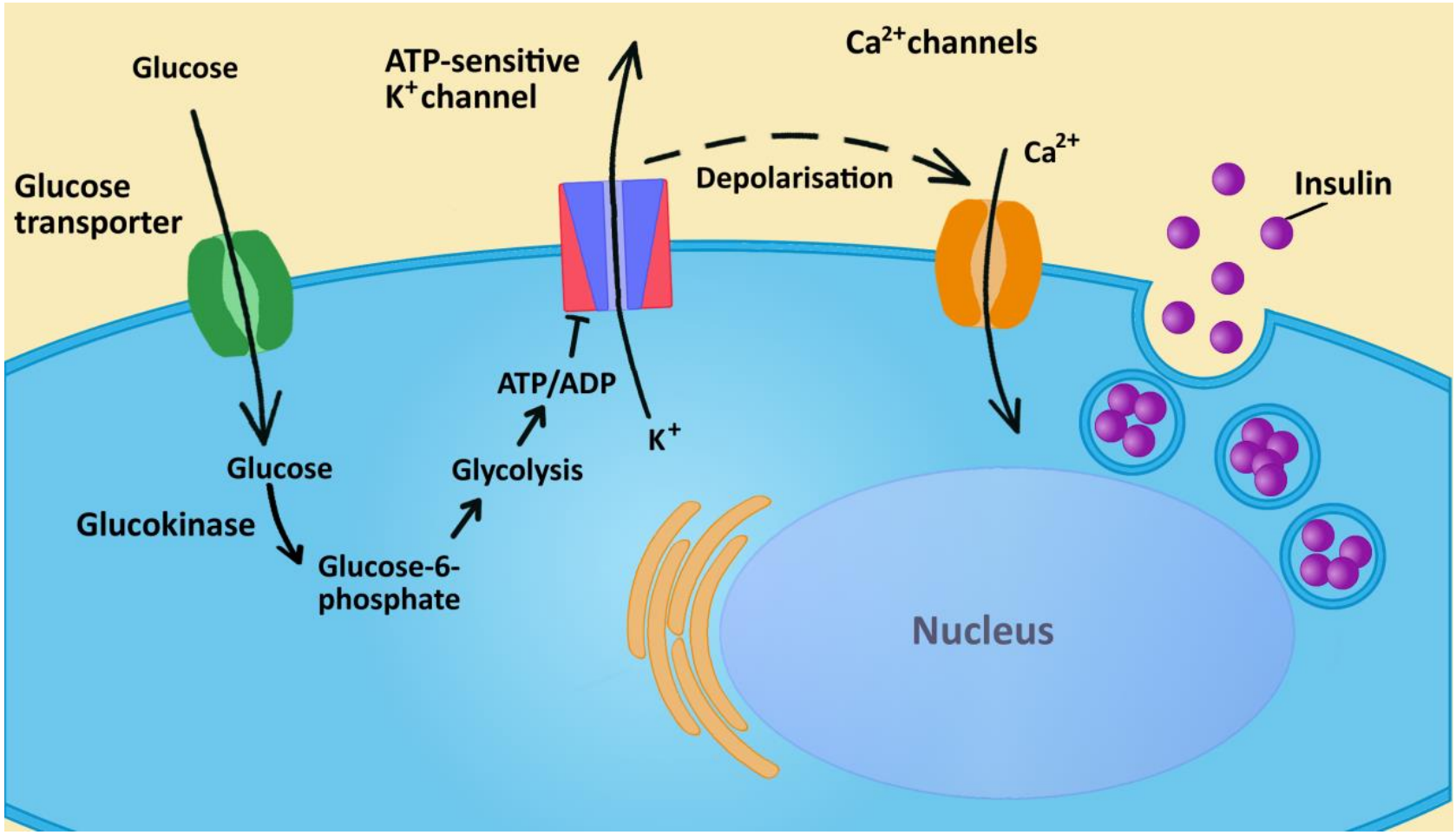


Figure 1.1: Insulin secretion from the pancreatic beta-cell. Glucose enters the beta-cell via glucose transporters. Once inside, glucokinase (GCK) the first enzyme in glycolysis, metabolises glucose to glucose-6-phosphate. Glycolysis then leads to an increase in the cellular ATP:ADP ratio. This causes the ATP sensitive potassium (K_{ATP}) channel to close, resulting in membrane depolarisation and subsequent opening of voltage gated calcium channels. The rise in intracellular calcium is the trigger for exocytosis of insulin secretory granules.

1.3 Genetics and T2D

1.3.1 The heritability of T2D predisposition

The dramatic increase in T2D incidence over the past few decades means that environmental changes must be the primary driver, since genetic selection for T2D susceptibly variants could not occur in such a short time-frame. However, it has long been recognised that an individual's risk for developing T2D is, in part, inherited. This means that whilst much of the global population is exposed to significant environmental changes, not everyone will be at equal risk of subsequently developing T2D. The genetic component of risk for developing a trait can be described as familial aggregation (e.g. the sibling relative risk) or as narrow-sense heritability (h^2) capturing the additive effects of parental alleles on the phenotypic variation of their offspring. Studies in twins show a higher concordance of T2D status in monozygotic than in dizygotic twin pairs[40,41], and family studies find the lifetime risk of developing T2D for individuals with an affected parent to be 40%, which is around four times higher than the average population risk[42]. The h^2 estimates of lifetime development of T2D derived from population cohort studies vary from 30 to 70%, and suggest a higher heritable component in individuals developing T2D before the age of 65[43,44]. The chance of an individual developing T2D therefore depends on both their exposure to environmental stressors, and on their inherited genetic constitution to deal adequately with this metabolic stress.

1.3.2 Insights gained through studying T2D genetics

Whilst the heritable nature of T2D susceptibility has been known for decades, substantial efforts from genome-wide linkage and candidate gene studies had until 2006 yielded only two (*KCNJ11*, *PPARG*) loci robustly associated with T2D

susceptibility[45,46]. Linkage analyses had proved powerful in discovering the high penetrance alleles underlying rare monogenic forms of non-autoimmune diabetes[36,47-53], but such alleles could not be found for common T2D. Since then, the arrival of genome-wide association studies (GWAS) has reshaped the genetic landscape. GWAS makes use of insights into haplotype structure gained through the HapMap project[54] to design genotype arrays which can capture much of the common variation within the human genome using a limited set of directly typed variants. These directly typed genotypes can then be expanded to a much larger set through imputation into haplotypes from the HapMap project. These GWAS, together with the large collaborative meta-analyses arising from them, have increased the number of loci that display genome-wide significant association for T2D to around 70 (Figure 1.2; Table 5.2)[55-64]. Although subsequent clinical translation of these signals has so far been slow and hampered by, for example, modest effect sizes and limited previous biological knowledge of the genes underlying association signals, several general characteristics of the genetic make-up of T2D susceptibility have been uncovered (Table 5.2). Several genes in loci associated with T2D are known drugs targets (including thiazolidinediones targeting *PPARG* and sulphonylureas *KCNJ11*), serving as proof-of-principle that, even though these loci only lead to small increases in the likelihood of developing T2D, GWAS findings can result in effective treatment options.

1.3.2.a Limited role for protein-coding variation

One clear characteristic of T2D GWAS loci, shared with GWAS findings for many other traits and diseases, is the tendency of the association signals to overlap non-protein-coding and poorly annotated sequence (Figure 1.2). In only a small handful of T2D associated loci (currently 6/75) is the index variant protein-coding, or in strong linkage disequilibrium (LD) with a protein-coding variant. Since the function of the non-protein-coding part of the genome is less well understood than the protein-coding

part, this tendency for GWAS variants to reside in non-coding sequence has hindered biological translation. However, current progress being made in characterising regulatory sequence is likely to improve understanding. Already large-scale efforts such as the Encyclopedia of DNA Elements (ENCODE) project, which characterises regulatory elements in a large number of cell-lines through a combination of next-generation sequencing approaches, has identified significant enrichment of functional regulatory variants amongst those associated with disease by GWAS[65].

Functional studies at specific GWAS loci have also provided evidence for causal regulatory variants. The 1p13 locus – the most strongly associated locus for low-density lipoprotein (LDL) cholesterol[66] and independently associated to coronary artery disease[67] and myocardial infarction[68] – is one where detailed investigation has uncovered a causal regulatory variant as well as the gene which this variant acts upon[69]. Through expression quantitative trait locus (eQTL) analysis of the GWAS index variant (rs646776) in three tissues relevant for lipid metabolism – liver, subcutaneous adipose tissue and omental adipose tissue – it could be established this variant was on a haplotype at which genotype status affected expression of three nearby genes (*CELSR2*, *PSRC1* and *SORT1*) in liver[69]. Furthermore, utilising the differences in LD between Europeans and African Americans, the 6.1kb region of maximal association, containing rs646776 and 15 other variants, could be reduced to a single most likely causal variant (rs12740374)[69]. Luciferase assays confirmed that rs12740374 was responsible for the difference in activity seen between the minor and major haplotype[69]. *In silico* binding prediction tools and electromobility shift assays demonstrated that this effect was due to a novel binding site for the liver-enriched transcription factor C/EBPα created by the minor allele[69]. Binding of C/EBPα at the minor allele was shown to reduce expression of *SORT1* in hepatocytes. Finally, when *SORT1* was knocked out in several rodent models, it resulted in an increase of LDL

cholesterol, and thus closing the loop between the variants and the effects observed for this locus in GWAS[66-69].

These results highlight the potential effects of functional regulatory variation on gene expression, and show that non-protein-coding variation has excellent candidacy for causality at GWAS loci. Functional studies of several association loci suggest that this might also hold true for GWAS findings for T2D and glucose-related traits[70-73].

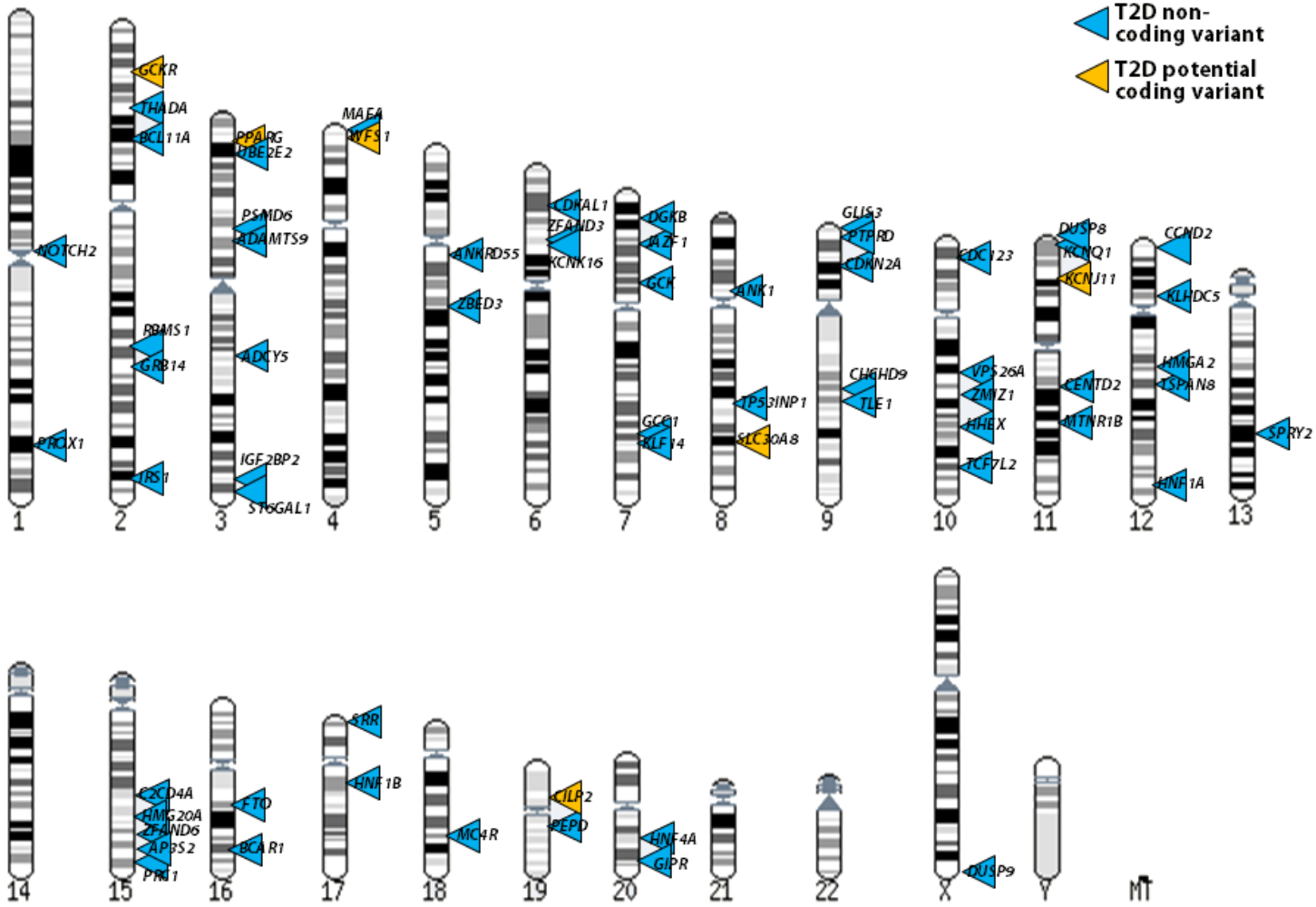


Figure 1.2: Loci currently associated with T2D. Each locus is labelled with the name of the closest or most biologically plausible gene, which in most cases is not proven to be causal. The colours denote whether the index variant at each locus is in strong ($r^2 > 0.8$) with a coding variant (orange) or non-coding (blue) [Adapted from McCarthy, *NEJM*, 2010[74]].

1.3.2.b Central role for pancreatic beta-cells in T2D predisposition

T2D association loci have also provided insights into the biological basis of the disease. By combining information on effect sizes and significance levels in GWAS data for measures of beta-cell function (HOMA-B) and insulin sensitivity (HOMA-IR) calculated from fasting glucose and insulin levels in ~37,000 non-diabetic individuals for the alleles implicated in T2D susceptibility, these data could shed a light on the relative importance of insulin secretory versus insulin sensitivity defects in the pathogenesis of T2D. This analysis showed that 15 of the 65 then associated loci also showed nominally significant association with HOMA-B, with only a few influencing insulin sensitivity[56,57,75].

Another line of evidence linking T2D association and beta-cell function is provided by the significant association now seen near many of the genes known to be involved in monogenic disorders affecting the beta-cell (*KCNJ11*, *HNF1A*, *HNF1B*, *WFS1*, *GCK*, and *HNF4A*). Of these, the *KCNJ11* locus, which was one of the first signals to be robustly associated to T2D[46,56], is also one of the few loci with a non-synonymous coding index variant (rs5219 – p.E23K) – although it has proved difficult to establish the functional mechanism of this variant (or indeed whether the causal variant is rs5219 or perhaps another missense variant in *ABCC8* [rs757110 – p. A1369S] which is in strong linkage disequilibrium)[76-79].

Together, current data suggest that most T2D association loci act through the beta-cell, directly or through changes in regulatory mechanisms such as incretins. It is possible that the case selection of T2D GWAS – enriched for patients with a stronger family history and a generally lower body mass index compared to the average T2D subject – has favoured the detection of loci affecting beta-cell function. However, BMI stratified re-analysis of current GWAS data has not provided much evidence for new loci influencing insulin sensitivity[80]. Knowing the central role for beta-cells in T2D

association is crucial to subsequent efforts to establish causal mechanisms, since much of the regulation of gene expression is tissue-specific. This is illustrated by the liver-specific transcription factor binding underlying the LDL-cholesterol association at the 1p13 locus discussed in section 1.3.2a (above)[69].

Another good example of the importance of knowing the tissue through which an association signal is acting is provided by an elegant study into the causal mechanism underlying the 6p21.1 region, associated with variation in size and number of erythrocytes (red blood cells)[81,82]. This association signal is located near the gene *CCND3*, and could be refined to two likely causal variants in close LD (rs9349205 and rs9394841) through multi-ethnic analysis. Using chromatin-immunoprecipitation (ChIP)-sequencing for a chromatin marker of enhancer regions, one of these variants (rs9349205) was shown to be located in an enhancer region specific to erythrocyte. Enhancer activity and erythrocyte-specificity of this element was confirmed by luciferase assays in the K562 erythrocyte cell line as well as several non-erythrocyte cell lines. The rs9349205 variant displayed allele-specific differences in luciferase activity, with significantly increased activity for the minor allele. Interaction between this enhancer element and the proximal promoter of the nearby *CCND3* gene in human adult erythroid progenitors was demonstrated using a chromatin conformation capture (3C) assay. Knockout of *CCND3* in rodent models, as well as in human and rodent erythrocyte progenitor cells, resulted in the production of fewer erythrocytes which were larger in size – a phenotype corresponding to the association in humans.

These results provide a likely functional mechanism for this association signal, and demonstrate how GWAS findings can lead to the discovery of novel biology. However, uncovering this mechanism and the one discussed in section 1.3.2.a was dependent upon conducting analyses in the relevant tissues, and so to uncover biology at other GWAS loci knowing the tissue in which to perform the studies will be vital.

1.4 Translating association signals into causal mechanisms

1.4.1 The challenges surrounding functional translation

Many association signals with an effect on T2D susceptibility have been identified, but translation of these association signals into molecular mechanisms, and subsequent opportunities for the development of novel clinical interventions, has proved much more challenging. As discussed in the previous section, several important obstacles stand in the way of translation. The small effect sizes of associated regions make it difficult to disentangle LD based on significance alone and to detect phenotypic differences when studying potentially causal variants in model systems. Furthermore, the tendency of association signals to lie in non-protein-coding sequence hinders our ability to design appropriate assays given incomplete understanding of non-coding sequence and the tendency of regulatory effects to be tissue-specific.

There are notable exceptions in which locus characteristics have facilitated functional follow-up. For example, the *TCF7L2* locus has a relatively large effect size[70,71], and the association signal at the *GCKR* locus encompasses coding variants located in a biologically relevant gene[67,68,83,84]. However, molecular mechanisms for the vast majority of signals have so far proved more elusive. While association loci are by convention labelled with the name of the nearest or most biologically plausible gene, this by no means indicates causality in absence of functional evidence. For example, despite being one of the earliest identified association signals and harbouring two genes with strong biological claim (*HHEX* and *IDE*) [69,7085,86], it is still unclear whether either of these genes (or both) might be causal in the locus on chromosome 10q23.33.

In another instance, detailed studies at the *KCNQ1* locus have uncovered strong support for a putative molecular mechanism through which one of the signals may be acting. At this imprinted locus, which contains four independent association signals (Table 5.2.), risk genotype status of the lead variants at one of the signals (rs2237895) was significantly correlated with changes in methylation patterns in both foetal and adult human islets[72,87]. DNA methylation changes alone may not result in gene expression differences[88]. However, due to the small effect size and limited availability of human islets, it was impossible to tie the changes in DNA methylation to a specific transcript which could mediate the phenotypic effect.

Most challenging, however, are the many loci mapping close to regions or genes without a known function. Here, gene function must be established or genomic annotations must be generated in relevant tissues, with substantial investment of both time and resources, before experiments to decipher the impact of T2D susceptibility variants can be designed.

One example of how progress can be made at such loci is illustrated by detailed studies of the T2D association signal near *TCF7L2*. This association signal - which to date has the largest effect size (odds ratio ~1.4) facilitating functional follow-up – was one of the earliest loci to be discovered and maps to an intron of *TCF7L2*[64]. At that time TCF7L2 (formerly TCF4) was known to be a ubiquitously expressed part of the Wnt signalling pathway (involved in the regulation of normal development and differentiation), but there was no literature on its involvement in diabetes. Seven years later the picture is quite different; searching for the combination of "TCF7L2 AND diabetes" in PubMed yields nearly 500 results. However, still the mechanisms of action of TCF7L2 are not yet known.

Through cellular approaches, rodent models and characterisation of human pancreatic islets, it is now known that modulation of *TCF7L2* expression levels has

effects on glucose-stimulated insulin secretion[89-91], incretin response[91], and beta-cell expansion in the pancreatic islet[89,91]. TCF7L2 may even have a wider impact on glucose metabolism, as demonstrated by a recent study suggesting that TCF7L2 is involved in hepatic glucose production as well as affecting pancreatic islet function[92]. However, the results of detailed studies on the impact of the T2D associated variants surrounding *TCF7L2* on metabolic parameters and islet physiology largely recapitulate the islet effects of TCF7L2, showing that these variants are associated with a blunted incretin effect[93], altered islet morphology[94], and reduced insulin secretion[93-95].

Through elegant studies looking at open chromatin (an evolutionarily conserved indicator of regulator activity) in DNA from human islets using Formaldehyde-Assisted Isolation of Regulatory Elements (FAIRE) sequencing, the strongest associated T2D variant in *TCF7L2* (rs7903146) was shown to be located in islet-selective open chromatin[70]. The authors also demonstrated that this open chromatin displayed strong enhancer activity, which was altered by the risk variant, and could then demonstrate allelic imbalance of the risk variant in islet FAIRE signals. These findings were replicated by another study using histone modifications, rather than open chromatin, to identify regulatory elements in islets[71]. Together, these results indicate that genetic variation at the *TCF7L2* locus affects T2D susceptibility by altering *cis* regulation and local chromatin structure in pancreatic islets. However, despite the indicative functional evidence and location of the transcript closest to the association locus, it is still not formally established that *TCF7L2* is the gene through which T2D association operates.

The expanding number of next-generation sequencing techniques able to interrogate genetic variation and genomic regulation provide new opportunities to tackle the challenges in translating genetics to functional mechanisms. The work in

this thesis aims to apply some of these developments to insights from T2D genetics using the strategies outlined in chapters 1.4.2 to 1.4.5.

1.4.2 Capturing the allele frequency spectrum

GWAS has been a powerful tool for identifying common (minor allele frequency [MAF] > 0.05) and, to a lesser extent, low frequency (MAF< 0.05 ≥ 0.005) variation involved in human traits and diseases. It relies on imputation to expand directly genotyped variants (ranging from 400k on older chips to up to 5M on those now available) to a more complete representation of the allele frequency spectrum. The process of imputation uses observed haplotypes in a reference panel (or multiple panels) to predict genotypes in the study samples. The HapMap project provided the first comprehensive survey of variation in diverse human populations, and has provided the scaffold for imputation of most GWAS to date[96]. It has been instrumental in the discovery of many GWAS loci, but is limited in sample size and, due to its reliance on array genotyping, the number of variants it can study. With publication of the first phase of the 1000 Genomes Project, which uses next-generation sequencing to characterise a larger number of samples, a panel with more detailed haplotype structure providing near complete coverage of the common and low frequency spectrum is now available for several ethnic groups[97]. Imputation into the 1000 Genomes panel should reduce the chance of missed association due to the absence of a variant in the imputation panel and allow for statistical fine-mapping of known associated regions (as will be discussed in the next section and chapter 5).

The discrepancy between total heritability estimates and the heritability explained by current GWAS findings for complex diseases such as T2D[98] is another reason for delving deeper into the allele frequency spectrum. This discrepancy could be explained, as proposed by the common disease-common variant hypothesis[99], by many more common variants of small effect sizes left to be discovered. For example,

by modelling the concordance between the *Z*-scores of independent nominally significant variants in two replication stages of GWAS meta-analysis data for T2D, it is estimated that – in the current data – at least another 488 T2D associated variants with small to modest effect sizes remain to be discovered[56]. However, variants outside the frequency range covered by GWAS also represent a potential explanation for filling the heritability gap.

Whereas most genetic variation is common, the largest number of genetic variants in an individual genome is found in the low (MAF< 0.05 ≥ 0.005) and rare (MAF < 0.005) allele frequency spectrum. Many Mendelian disorders caused by rare variants have been identified, but linkage and large-scale resequencing studies have found little evidence for involvement of such variants with large effect sizes in common human disease[100,101]. This does not exclude the possibility that low frequency or rare alleles contribute to T2D susceptibility, but means that such alleles will have small effect sizes and are thus unlikely to be discovered with current sample sizes of association studies. However, looking at such alleles in combination across genes or regulatory units – so called aggregate tests – will facilitate identification of such effects. Resequencing efforts focussing on candidate genes underlying common association loci have already identified aggregation of variants associated with T2D in, for example, the gene encoding melatonin receptor 1B (*MNTR1B*). This gene was found to harbour 36 rare variants (MAF <0.01) which together were significantly associated ($P = 1.64\times10^{-4}$) with T2D risk[102]. Functional studies established that twenty-six of these variants affected receptor function, with four resulting in complete loss of its melatonin binding and signalling capabilities.

These finding demonstrate that integrated analyses of lower frequency variants can provide novel insights into T2D susceptibility and pathophysiology. Since these variants are of modest effect sizes, studying them requires sequencing of exomes or whole-genomes in large pedigrees or population cohorts. Even with the falling costs of

sequencing, performing this types of study represents a very significant investment of time and money. In T2D, two such studies (Genetics of Type 2 Diabetes [GoT2D] and Type 2 Diabetes Genetic Exploration by Next-generation sequencing in multi-Ethnic Samples [T2D-GENES] –described in chapter 2) are currently underway. In chapter 6, data from these studies will be used to establish the impact of low frequency and rare variants in regulatory sequence on T2D susceptibility.

1.4.3 Refining association signals through fine-mapping approaches

Much of the power of GWAS lies in the fact that only a proportion of common variants need to be genotyped in order to capture the majority of all common variants present in an individual. However, the LD upon which this approach relies makes it very difficult to disentangle the specific variants underlying an association signal. Functional studies testing every possible variant for causality are currently infeasible, given the costs and labour investments involved and the need to perform the experiments in the right tissue using the right assay for the variants' genomic function (both of which are often unknown). However, since there are many more regulatory active variants than variants acting on disease susceptibility, it is doubtful whether such functional testing of every variant in a locus without refinement of the association signal would in fact help uncover the causal one. Other approaches to home in on the causal variant(s) are therefore needed.

Association loci are most commonly defined by broad LD blocks flanked by recombination hotspots. Fine-mapping – the reduction of these large genomic regions containing many variants to a narrower region or smaller variant set – is the next step towards translation of association loci. Two complementary fine-mapping approaches currently being employed revolve around multi-ethnic study designs and high density genotyping of association loci. Multi-ethnic study designs harness the differences in

allele frequencies and LD patterns between different populations for fine-mapping shared association signals. One such study fine-mapping 58 loci associated with lipid traits in three diverse populations showed that this approach can indeed be powerful. Based on an assumption of shared causal variants, it refined the association signal at five of these loci, encompassing non-synonymous variants at two of them; a single, previously established functional[84], variant in *GCKR (GCKR*-P446L) and four variants, including *SLC12A4*-E4G, near *LCAT*[103].

High density genotyping, which captures nearly all variants at a locus, also allows refinement of the association signals – for example based on posterior probabilities of causality using Bayes' theorem. A recent publication by the Wellcome Trust Case Control Consortium demonstrated that this approach could successfully refine the association signal at 7 of the 14 tested loci. With the common and low-frequency allele spectrum having been nearly completely captured through the 1000 Genomes project and others, imputing existing genotyped samples into these panels should allow more successful statistical fine-mapping without the need to perform direct genotyping at all loci.

1.4.4 Expression profiling and eQTL analysis

Fine-mapping restricts the number of potentially causal variants, which when linked to genomic annotations can provide greater insights into functional mechanisms than "raw" GWAS loci. Gene expression data is one such layer of information important for the interpretation of variants. Technological advances have seen expression profiling move from Northern blot studies in the late 1970s, capturing only single transcripts, to the transcriptome-wide methods currently available. These techniques used to be microarray-based and relied on hybridisation of cDNA to reference probes. Now that next-generation sequencing has become more readily accessible and affordable, microarrays are giving way to RNA-sequencing (RNA-seq)

as the preferred method for transcriptome-wide analyses. Microarrays, unlike RNA-seq, rely on previous knowledge about existence of transcripts, have a limited detection range for transcript expression levels and give no information on genetic variation within transcripts[104]. RNA-seq therefore provides a much more detailed view of the transcriptome than could be obtained by microarray analysis and allows detailed interrogation of genetic variation within transcript through, for example, the study of allele-specific expression (ASE). However, when answering specific hypotheses requiring large numbers of samples, microarrays still have the advantage of being robust, high-throughput and relatively inexpensive, with probes that can be customised to tailor the experimental question[105].

The strength of RNA-seq, and transcriptome analysis, is illustrated by findings from the Encyclopedia of DNA Elements (ENCODE) consortium. By interrogating the transcriptome of many different cell-lines, they discovered over 42,000 potential novel transcripts[106]. Although these transcripts could be noise introduced through, for example, DNA contamination, at least 7,500 were multi-exonic and as such much more likely to be true novel transcripts[106]. Validation experiments did indeed confirm that 70-90% of the multi-exonic transcripts were genuine, with most of them displaying characteristics of long non-coding RNAs (lncRNAs)[106]. Similar efforts focusing on small RNAs also uncovered many unannotated transcripts and even identified novel classes of small RNAs[106,107]. Since cell-type specific expression patterns were the hallmark of most of the novel transcripts and most tissues involved in T2D are not well characterised, studying T2D-relevant tissues will be crucial for the discovery of tissue-specific regulatory mechanisms and biological mechanisms.

As well as providing the gene expression profile of a tissue, RNA-seq also captures genetic variation present in transcripts. Combining gene expression data with genotyping data alone allows detection of eQTLs, but the added information from

RNA-seq also enables the study of allele-specific expression (ASE). eQTLs describe the association between variants and transcript expression levels irrespective of the haplotype. They are most commonly study in the context of local (*cis*-eQTLs) effects on expression levels, since studying genome-wide interactions (*trans*-eQTLs) leads to power limitations related to very large multiple testing penalties. In comparison, ASE can pick up differences in allelic expression which might not affect total transcript expression, and also provide a measure of effect strength based on the ratio between alleles from heterozygous exonic variants present in the data. Since ASE analyses are performed on an intra-individual basis in data generated using identical technology, it also contains internal controls for many confounders.

Although very limited human islet or beta-cell eQTL, and no ASE, data is currently available[108], such data in other tissues have already furthered our understanding of some T2D-associated loci. A key example of this impact is provided by the signal near *KLF14*, a gene encoding the transcription factor Kruppel-like factor 14. The T2D analyses showed that this locus had an effect on insulin resistance[57], and in eQTL data from human adipose tissue the same set of variants associated with T2D susceptibility overlapped with a set of variants regulating *KLF14* expression in *cis*, as well as having significant *trans*-effects on a multitude of other adipose transcripts[109,110]. One caveat to this type of analysis is that coincidence of eQTL and GWAS signal in the same locus does not necessarily equate to a shared causal variant. However, in the case of *KLF14* not only did the eQTL and GWAS signals overlap in the same locus, but both signals also peaked at the same group of highly correlated variants ($r^2 > 0.96$) with the lead *cis*-eQTL variant in perfect LD ($r^2 = 1$) with the leadT2D variant[109,110]. These results do not exclude the possibility that the T2D signal is acting through a different (unstudied) tissue, in which the same variants might be an eQTL for a different gene, although this is unlikely given the other evidence linking this T2D signal to insulin sensitivity. As with the tissue-specificity of transcripts in ENCODE, this network of

cis- and *trans*-effects was limited to adipose tissue, again highlighting the need for performing studies in the appropriate tissue.

It would not have been possible to define the molecular mechanism underlying this locus without eQTL data, illustrating its strength in uncovering novel biology.

1.4.5 Epigenetic modulation

Next-generation sequencing is not only useful in directly studying RNA and DNA, but can also be used to study a large number of epigenetic modulators on a genome-wide scale (Figure 1.3). Methods such as whole-genome bisulphite and ChIP sequencing enable global DNA methylation profiling and large-scale identification of regulatory elements (through their chromatin modification signatures) or binding sites for transcription factors of interest. Integrating results from transcriptomic studies, such as eQTL analyses, with epigenetic data, creates an even richer set of annotations for establishing underlying molecular mechanisms at T2D association loci.

Last year, the ENCODE project published the largest catalogue of epigenomic maps to date. It combined a vast quantity of information comprising accessible chromatin (125 cell types), profiles of 12 different histone modifications (46 cell types) and 119 transcription factors (72 cell types), DNA methylation information (82 cell types), RNA expression levels (15 cell lines), and chromatin interaction maps (4 cell lines)[111]. By profiling DNase I hypersensitive sites (DHS) across 125 cell-lines, ENCODE uncovered ~2.9 million regions of accessible chromatin, which, when integrated with all the other information, provide important insight into the complexities of gene regulatory networks[112]. It shows that a single promoter often interacts with > 20 distal regulatory elements, and that these regulatory elements themselves can bind to more than one promoter[112]. Like the novel transcripts

described in section 1.4.4[106], a large proportion (approximately 30%) of the DHS uncovered by ENCODE were specific to a single tissue, with only ~0.1% of chromatin accessible across all 125 tissues[112].

With so many novel genomic annotations discovered, ENCODE then looked at the overlap between these and GWAS associated loci to determine whether improved annotations could reveal more potentially causal variants. Overall, of the 4724 distinct lead variants in the GWAS catalogue, 44.8% overlapped any ENCODE annotation present in GENCODE or RegulomeDB[65]. The percentage of loci overlapping annotations increased to 80% when including variants in strong LD ($r^2 > 0.8$) with the lead variant, compared to using the lead variants alone[65]. This analysis again highlights how few GWAS loci overlap coding variants; when using only CEU LD patterns, which assumes all loci have been discovered in European populations, just 16% of all 4724 LD regions encompassed such a variant[65].

Not only does epigenetic profiling allow large scale comparison of variation in regulatory active regions between tissues, it can also be instrumental in understanding the development and biology of a single tissue. For example, investigation of the distribution of chromatin modified by trimethylation of histone H3 lysine 27 (H3K27me3), which marks Polycomb-dependent repression, in tissues sharing a common developmental origin in mice provided novel insights into the origins of beta-cells[113]. Unlike the developmentally more related acinar cells, the beta-cell H3K27me3 and transcriptional profile most closely resembled that of ectodermal-derived neural tissues[113]. Subsequent studies showed that this indeed resulted from the selective removal of Polycomb-dependent repression at neural regulatory genes as the pancreatic progenitors went down the route to their endocrine fate[114]. Similarly, studies of Polycomb-repression in a differentiation model deriving pancreatic endocrine cells from human embryonic stem cells observed dynamic changes in Polycomb-repression patterns[115]. While at the end of the *in vitro* differentiation step

the cells exhibited several aberrant epigenomic programs, further differentiation by implanting these cells in mice made their profiles increasingly resemble that of true endocrine cells.

These examples demonstrate that epigenomic data is useful both for discovering molecular mechanisms underlying GWAS data and for identifying biology underlying processes such as cellular fate and cancer development. As highlighted by ENCODE data, expression is often regulated at a tissue-specific level, so studying those cell types most relevant to the disease or trait of interest is crucial to translate findings from genetic studies. As ENCODE contains very few T2D relevant cell types, more detailed studies are needed to provide the most accurate annotations with the best chance of identifying causal variation at GWAS loci.

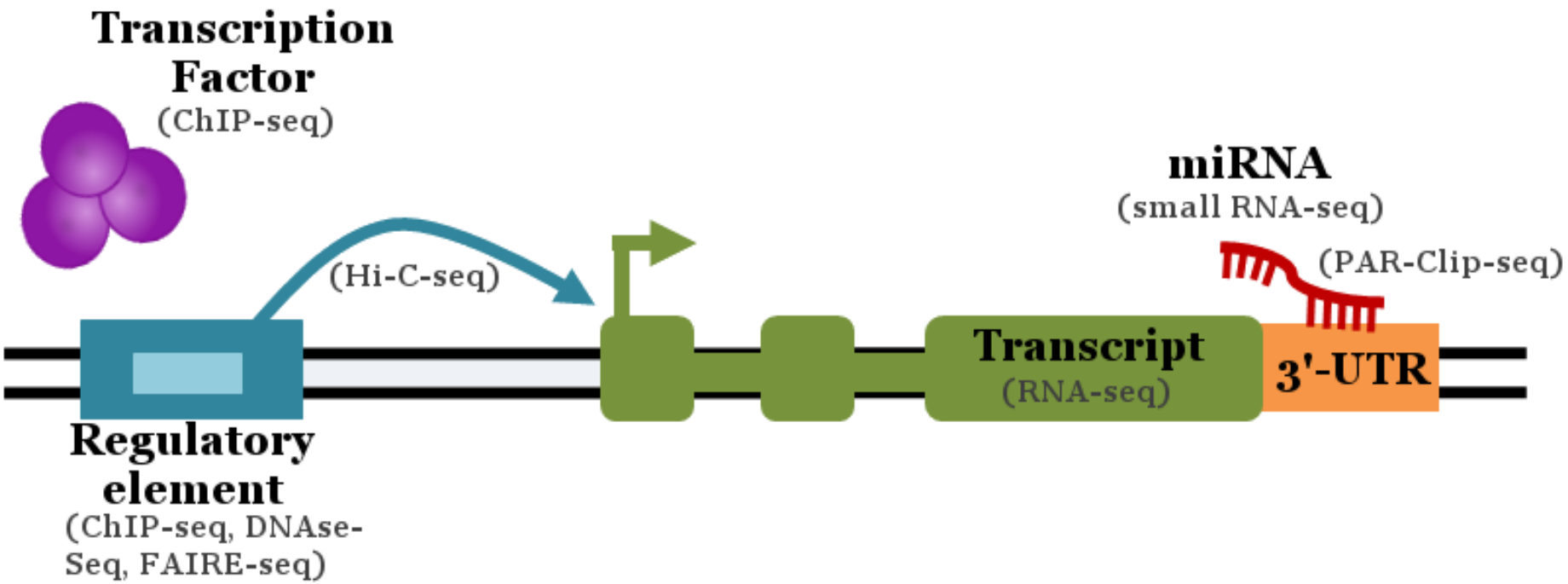


Figure 1.3: Schematic representation of gene regulatory processes accessible to next-generation sequencing approaches. The text in brackets denotes exemplar technologies used to profile each regulatory process.

1.5 Thesis aims

This introduction has outlined the substantial challenges that currently hinder efforts to uncover causal variants underlying T2D susceptibility loci, and recent advances made in (epi)genomics that might help overcome some of them. Work presented in this thesis aims to advance understanding of T2D-associated loci by integrating genetic data with genomic data generated from human pancreatic islets. More specifically, this work aims:

- to further define the regulatory machinery of human islets by collecting data on mRNA and miRNA transcripts through RNA-sequencing, characterising transcripts present in human islets and identifying regulatory sequence through eQTL and ASE analyses (chapters three and four).

- to determine the extent to which fine-mapping studies, using samples from both single and multiple ancestral groups, can resolve association signals to a smaller set of variants or tighter genomic interval containing the causal variant. This work entails extensive simulations and is described in chapter five.

- to combine genetic and genomic data and establish whether this aids identification of causal variants underlying T2D susceptibility loci. For this, genomics data generated on human islets and genetics data from previously published T2D GWAS, as well as unpublished whole-genome and exome-sequencing studies, are integrated to look at both single-variant association and association burden across the identified islet regulatory units (chapter six).

Chapter 2

General Methods

2.1 Human islet tissue bank

RNA was extracted from human adult islets to quantify miRNA (chapter three) and mRNA (chapter four) expression. All islet preparations were extracted with full ethical consent from cadaveric donors. Details of sample origin and donor characteristics are provided within the relevant experimental chapters.

2.1.1 RNA extraction from human islets

RNA was extracted from human islets using the guanidinium-thiocyanate-phenol-chloroform method[116]. For islets isolated at the Oxford Centre for Islet Transplantation (OXCIT), islets cells in suspension were transferred to a 15 mL centrifuge tube after arrival and gently centrifuged at 80-90 x g for 3 minutes at 4˚C to pellet. To clean remaining media from the islets, the supernatant was removed and three wash-steps were performed by adding one mL of 1x phosphate buffered saline (PBS) (Sigma Aldrich, Gillingham, UK) and centrifuging at 80-90 x g for three minutes each time. After the final cleaning step 1 mL Trizol (Ambion, Warrington, UK) was added to the cells. The cells were lysed by pipetting immediately to ensure rapid inhibition of RNase activity and incubated at room temperature for ten minutes. Lysates were then transferred to clean 1.5 mL RNase-free centrifuge tubes (Applied Biosystems, Warrington, UK). For islet preparations isolated in Edmonton, this initial stage was performed locally by members of the MacDonald lab and Trizol fractions were shipped to Oxford for further processing.

Two hundred µL chloroform (FisherScientific, Loughborough, UK) was added to each tube, before vigorous shaking for 30 seconds to begin organic and aqueous phase separation. Samples were incubated at room temperature for 5min and spun at 12,000 x g for 30 minutes at 4˚C to complete phase separation. The aqueous phase containing the RNA was transferred to a clean 1.5ml RNase-free tube by pipette, and 500µl isopropanol (Fisher Scientific, Loughborough, UK) was added to precipitate the RNA. For DNA extraction (see

section 2.3) the remaining organic and DNA phases were stored at -20˚C. The RNA solution was incubated for 5min at room temperature with repeated inversion and stored overnight at -20˚C. The following day, RNA was pelleted by centrifugation at 12,000 x g for 50 min (4˚C) and supernatant was carefully removed. The pellet was washed in 1 ml 75% ethanol (Sigma Aldrich, Gillingham, UK) before centrifugation at 12,000 x g for 30min. This procedure was repeated with another 1ml fresh 75% ethanol. After the final ethanol wash was removed, the RNA pellet was allowed to air-dry for 10min. To re-suspend the RNA, a minimum of 20µl RNase-free water (more as necessary for complete re-suspension) was added to each sample. After determining sample quality (see section 2.2), samples were stored at -80˚C for future use.

2.1.2 DNA extraction from spleen for genotyping

To allow genotyping without depleting DNA stocks obtained from human islets, OXCIT obtained spleen samples from human islet donors where possible. Prior to extraction, spleen tissue was stored at 4°C. DNA was extracted using an automated Maxwell 16 system with a Tissue DNA Purification Kit (both Promega, Madison, USA) according to manufacturer's instructions. Briefly, spleen samples were lysed by shearing and genomic DNA was captured on proprietary MagneSil Paramegnetic Particles. These particles were wash several times, before being released into 500-1000 µL elution buffer. Eluted stock DNA was stored at -20°C. Successful DNA extraction was confirmed by PCR with a validated primer set and agarose gel check as in section 2.3.

2.1.3 DNA extraction from human islets for genotyping

Where no spleen sample was available, DNA was extracted from human islets using the fraction remaining after phenol-chloroform extraction of RNA described in section 2.1.1. To precipitate the DNA, 300µl pure (100%) ethanol was added to the thawed solution. This mixture was incubated at room temperature for a minimum of 30min with frequent inversion. DNA was then pelleted by centrifugation at 4000 x g for 5min (4°C), and the

supernatant removed. The pellet was washed with 0.1M trisodium citrate (Sigma Aldrich) in 10% ethanol and left at room temperature for 30min, with occasional inversion. After centrifugation (4000 x g, 5min, 4°C), this process was repeated and followed by another wash step with 75% ethanol. After the final wash step, pellets were air-dried for 10min to remove residual ethanol and re-suspended in a minimum of 100 µL 8mM NaOH (Sigma Aldrich). Re-suspension was aided by mixing at 400 x g for 1hr on a plate mixer heated to 50°C (Eppendorf, Cambridge, UK). Successful DNA extraction was confirmed by PCR with a validated primer set and agarose gel check as described in section 2.3. Extracted DNA was stored at -20°C.

2.2 Human islet RNA quality assessment

RNA quality was assessed through spectrophotometric and Agilent BioAnalyser (Agilent, Wokingham, UK) analysis, to determine purity and integrity respectively.

To assess RNA purity and concentration, a ThermoScientific NanoDrop™ spectrophotometer (ThermoScientific, Wilmington, USA) was set to the RNA-40 programme and blanked with nuclease-free water (1.5 µL). After blanking, 1 µL of each sample was added to the pedestal and absorbance measured at wavelengths of 230, 260 and 280nm. RNA purity was determined through 260/280 and 260 /230 ratios; RNA was classified as pure if 260/280 ratios fell between 1.7-2.1 and 260/230 ratios between 1.5-2.2.

RNA was run on an Agilent 2100 Bioanalyser using a total eukaryotic RNA 6000 nano chip (both Agilent) to determine RNA integrity. Chips were prepared according to manufacturer's instructions using 1µl sample. Peaks were called and quantified for each sample based on the fluorescence emitted by dye bound to the RNA when samples migrating through the capillary pass a laser. An RNA Integrity Number (RIN score) was derived from the size and location of the ribosomal peaks according to a proprietary algorithm (figure 2.1),

where 0 indicates a completely degraded sample and 10 indicates perfect integrity. Samples with RIN scores > 5 were used for subsequent experiments.

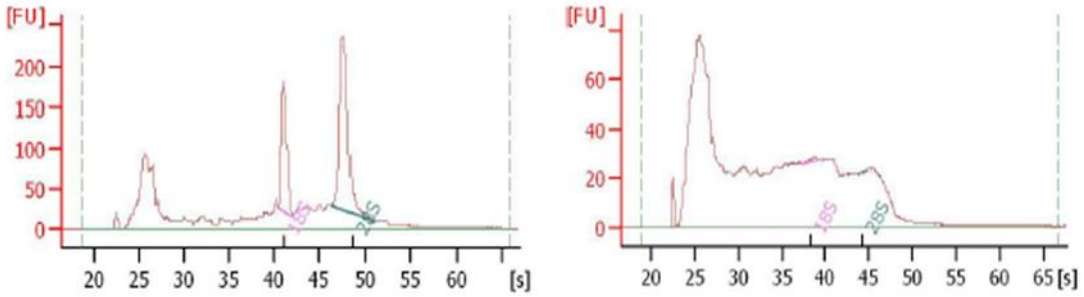


Figure 2.1: RNA integrity testing using the Agilent Bioanalyser system. The left panel shows a Bioanalyser trace of good quality RNA (RIN score = 8.1) with identifiable miRNA peak - passing through the laser at around 25seconds - followed by clear 18S and 28S ribosomal RNA peaks at 41 and 48 seconds respectively. The right panel shows a trace with considerable degradation - most RNAs have been broken down to smaller sizes and ribosomal peaks are barely evident (RIN score = 2.3, RNA considered unusable).

2.3 DNA quality assessment

DNA was amplified by polymerase chain reaction (PCR) as a quality control prior to genotyping on an Illumina Omni2.5 + exome genotyping array (chapter four).

2.3.1 PCR reaction components and cycling conditions

PCR was used to amplify gDNA and cDNA to produce quantities sufficient for gel electrophoresis. Unless otherwise stated, standard reaction mixes were used as detailed in table 2.1 – with each reaction containing 20ng DNA. Reaction volumes quoted allow for the addition of 1 µL DNA plus 11.5 µL of reagent mix (Applied Biosystems). The volume of distilled water was adjusted to compensate for variation in DNA concentrations. The primers used for all reactions were targeted to Glucokinase (*GCK*) exon 2.

Reagent	Volume in 12.5µl reaction (µL)	Final reaction concentration
10 x PCR buffer II	1.25	n/a
Magnesium Chloride (25mM)	0.75	1.5mM
dNTPs (2.5mM each)	0.50	0.1mM (each)
Forward primer (5mM)	0.50	0.2mM
Reverse primer (5mM)	0.50	0.2mM
Amplitaq Gold	0.05	0.25U
dH_2O	7.95	n/a

Table 2.1: PCR reaction components: Buffer, Magnesium Chloride, dNTPs and Amplitaq gold all supplied by Applied Biosystems. Volumes provided for n=1 standard 12.5µl reaction.

Samples were briefly centrifuged before amplification on a PTC-225 Peltier Thermal-Cycler (MJ Research, Waltham, Massachusetts, USA) according to cycling conditions as detailed in table 2.2.

Step	Action	Temp (◦C)	Time
1	Polymerase activation	95	12min
2	DNA melting	94	30s
3	Primer annealing	58	30s
4	Synthesis & Extension	72	1min
5	Steps 2-4 repeat	35 times	
6	Final extension	72	10min

Table 2.2: Standard PCR cycling conditions.

2.3.2 Agarose gel electrophoresis of PCR products

The success and specificity of all PCR reactions was confirmed via agarose gel electrophoresis of the PCR product. A 2% gel was made by dissolving agarose (Sigma Aldrich, Gillingham, UK) in 1X Tris/Borate/EDTA (TBE) (NBS Biologicals, Huntingdon, UK using an 800W microwave for 2min. After cooling the agarose under a running tap for 5-10 minutes, the intercalating agent Ethidium Bromide (Sigma Aldrich) was added to produce a final 0.002% solution. Each gel contained 3g agarose, 150mL TBE and 3µl ethidium bromide. Gels were set in an electrophoresis tank of appropriate size containing combs (both Appleton Woods, Birmingham, UK) for loading of the DNA samples. Once set, gels were covered with 1xTBE.

PCR product (5µL) plus 5µL Orange G loading buffer (4g sucrose in 10ml dH20 and 20mg Orange G [Sigma Aldrich]) was added to each sample well. Finally, 4µL (0.2µg) 100bp DNA ladder (New England BioLabs, Hitchin, UK) was added to determine fragment size of the PCR products. The gel was run for 40 minutes at 120V. Gels were visualised under Ultra Violet (UV) light on a Gel Doc 2000 illuminator system using Quantity One v4.3.1 software (both BioRad, Hemel Hempstead, UK).

2.4 Quantitative PCR to estimate human islet purity

To obtain an indication of islet sample purity before miRNA (chapter three) or mRNA (chapter four) sequencing, the expression of a panel of endocrine and exocrine makers was established using quantitative PCR.

2.4.1 cDNA synthesis

Single stranded cDNA was generated from RNA using the QuantiTect reverse transcription (RT) kit (Qiagen, Crawley, UK), which contains Omniscript and Sensiscript reverse transcriptases and random primers. To eliminate residual gDNA contamination, 50 ng RNA in 12 µL distilled water was mixed with 2 µL gDNA wipeout buffer and incubated at 42°C for 2min. Mastermix (4 µL 5x RT buffer, 1 µL RT random primer mix, and 1 µL reverse transcriptase) was added to each sample and mixed by pipetting. Sample were briefly centrifuged and incubated at 42°C for 30min – during which the reverse transcription reaction took place – followed by 95°C for 3min to inactivate the enzyme.

2.4.2 Quantitative real-time PCR of the islet purity panel

Quantitative real-time PCR of the islet purity panel (Table 2.3) was performed in a reaction volume of 10 µL, comprising 4 µL diluted cDNA, 5 µL gene expression mastermix (Applied Biosystems), 0.5 µL assay (detailed in Table 2.3) and 0.5 µL dH2O. Samples were amplified in triplicate with all six assays on one plate. Thermal cycling and fluorescence

detection was performed on an ABI7900 HT using the following amplification programme: 50˚C for 2min, 95˚C for 10min followed by 40 cycles of 95˚C for 15s and 60˚C for 1 minute.

	Gene	Assay ID
Endocrine Markers	Insulin (*INS*)	Hs00355773_m1
	Glucagon (*GCG*)	Hs00174967_m1
	Somatostatin (*SST*)	Hs00356144_m1
Exocrine Markers	Chymotrypsin (*CTRB*)	Hs00200713_m1
	Pancreatic Lipase (*PNLIP*)	Hs00609591_m1
	Amylase, alpha-2A (*AMY2A*)	Hs00420710_g1
Housekeeper Markers	peptidylprolyl isomerase A (*PPIA*)	Hs99999904_m1
	hypoxanthine-guanine phosphoribosylytransferase 1 (*HPRT1*)	Hs99999909_m1

Table 2.3: Gene expression assays for islet purity analysis. All probes are inventoried expression assays fluorescently labelled with FAM (Applied Biosystems)

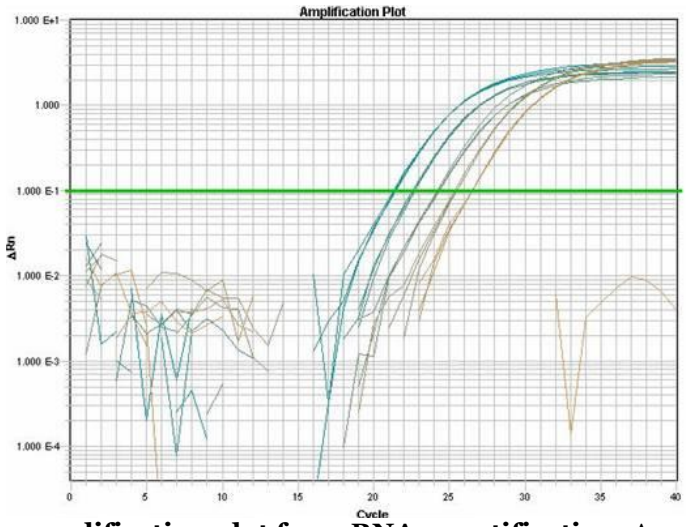


Figure 2.2: example amplification plot for mRNA quantification. Amplification plot of a typical experiment, with cycle number (x-axis) plotted against a logarithmic scale of fluorescence intensity (y-axis). The threshold value (green line) from which CTs are generated is set in the exponential phase of the reaction.

Samples were amplified on an Applied Biosystems 7900HT machine, which took fluorescence readings after every amplification cycle. The results were analysed using SDSv2.3 software (Applied Biosystems). Output values were comparative threshold (CT) scores: the cycle number at which fluorescence levels crossed a critical threshold (figure 2.2),

with threshold levels set to be in the exponential phase of amplification. CT values for target genes were compared to the mean of the housekeeper genes for that sample (ΔCT). The resulting value was linearised, using the formula $2^{-\Delta CT}$. From these data, purity was determined by calculating the proportion of overall expression which came from the endocrine markers. Based on previous experiments in our laboratory, a cut-off of 0.6 was chosen for sample inclusion (which approximately corresponds to dithazone-staining purity > 50%) for the mRNA sequencing experiments (chapter four).

2.5 Fluorescence-activated cell sorting (FACS)

For the comparison between islet and beta-cell miRNA profiles reported in chapter three, fluorescence-activated cell sorting (FACS) was performed according to methods described in Parnaud *et al.*[117], which are described in detail below.

2.5.1 Preparation of Krebs buffers for FACS

Prior to the experiment, 4X Bicarbonate Hepes Buffer was prepared using 0.1M Hepes (Sigma Aldrich), 0.02M Sodium Chloride [NaCl] (Sigma Aldrich) and 0.02M Sodium Bicarbonate [NaHCO3] (FisherScientific). A mixed salts buffer was also prepared using 0.6M NaCl, 0.02M Potassium Chloride [KCl] (Sigma Aldrich), 0.005M Calcium Chloride Dehydrate [CaCl2.H20] (AppliChem, Lancaster, UK), 0.006M Potassium Dihydrogen Phosphate [KH2PO4] (Sigma Aldrich) and 0.006M Magnesium Sulphate [MgSO4.7.H20] (Sigma Aldrich). Modified Krebs-Ringer bicarbonate HEPES buffer (pH7.4) was prepared using 80mL dH20, 37.5 mL 4X Bicarbonate Hepes Buffer and 30mL mixed salts buffer with 2.8mM glucose (Sigma Aldrich). 10mL of this Krebs buffer was transferred to a separate beaker, after which 700 mg bovine serum albumin (BSA [Serological Corporation, California, US]) was added to the remaining 140 mL to generate 0.5% BSA-Krebs solution. To generate a 0.25% BSA-Krebs solution, 10 mL 0.5% BSA-Krebs solution was added to the remaining 10 mL Krebs buffer without BSA. Both buffers were incubated at 37°C before FACS.

2.5.2 FACS enrichment of human beta-cells

Following overnight culture in an incubator at 37˚C, islets were transferred to a 50ml non-stick falcon tube (Becton Dickinson, Oxford, UK), pre-coated with CMRL 1066 media supplemented 10% fetal calf serum, 2mM L-glutamine, 10mM Hepes buffer, gentamycin, and penicillin/streptomycin. Islets were gently centrifuged at 100 x g for 1 min to pellet and re-suspended in 1 mL fresh media before transferral to a pre-coated 15ml non-stick falcon tube (Becton Dickinson). The pellets were washes three times by adding 14 mL Tampon Noir (0.15M Sodium Chloride [NaCl], 0.008M Sodium Phosphate Dibasic Dihydrate [Na2HPO4.2H20], 0.003M Potassium Chloride [KCl], 0.0015M Potassium Dihydrogen Phosphate [KH2PO4] [all Sigma Aldrich]) followed by centrifugation at 100 x g for 1 minute. After each wash the supernatant was removed and the pellet dislodged when adding the Tampon Noir to ensure thorough washing. After the final wash step, 1 mL Accutase (PAA Laboratories, Yeovil, UK) was added to the disperse the islets into single cells. Samples were incubated in a water bath (37°C) for 10 min while gently pipetting every 30 seconds to ensure agitation and dispersal of the pellet. The reaction was stopped by adding 0.5% BSA-Krebs solution. Samples were centrifuged at 300 x g for 5 min and another wash in 0.5% BSA-Krebs solution was performed. After this, the pellet was re-suspended in 1mL 0.5% BSA-Krebs solution.

Two FACS tubes (Becton Dickinson) were coated with 0.5% BSA-Krebs solution – one for staining and one control tube. To the control tube 200µl 0.5% BSA-Krebs solution and 10 µL re-suspended cells were added. The remaining 990 µL cells were used for the beta-cell sorting. Samples were labelled with three different markers –an anti-ductal antibody (AcNCA19-9) and a DNA-binding agent (7-aminoactinomycin D [7AAD]), which allows the removal of dead and ductal cells, as well as a Zinc-specific dye (Newport Green) which selects for the Zinc containing beta-cells. First, 40 µL of the anti-ductal antibody AcNCA19-9 (Novacastra, Newcastle Upon Tyne, UK) was added. Both tubes were incubated at 4°C for 1 hour, re-suspending the pellet every 10 min. Following incubation, the cells were washed

twice with 0.5% BSA-Krebs solution with centrifugation at 800 x g for 5 minutes. The pellet was again re-suspended in 1 mL 0.5% BSA-Krebs solution and 5µL of secondary antibody (anti-mouse Alexa 633 antibody [Invitrogen, Paisley, UK] was added. Both sample and control tube were incubated at 4°C for another hour, re-suspending the pellet every 10 min.

The sample was washed again twice, now using the 0.25% BSA-Krebs solution followed by an 800 x g spin for 5 minutes. The pellet was re-suspended in 750 µL 0.25% BSA-Krebs solution and 750 µL Newport Green solution (15 µL pre-warmed Pluritonic X, 2 µL Newport green [both Invitrogen] in 1 mL 0.25% BSA-Krebs solution). Both sample and control tube were incubated for 30 minutes at 37°C, re-suspending every 10 minutes. This was followed by another wash in 0.5% BSA-Krebs solution, after which the pellet was resuspended in 1 mL 0.5% BSA-Krebs solution. For the final round of staining, 10µL 7AAD (Invitrogen) was added to the sample tube and 2µl to the control tube – the control cells were stained to allow separation of the spectra for 7AAD and Alexa-633 – and both tubes were incubated at 4°C for 10 minutes. Both sample and control were washed a final time in 0.5% BSA-Krebs solution and resuspended in 1ml 0.5% BSA-Krebs solution before straining through a 100µm filter to ensure only single cells were added to the FAC sorter. Sorting was performed using a Beckman Coulter 7 – colour MoFlo high speed cell sorter (Beckman Coulter, High Wycombe, UK) at the Jenner Institute (University of Oxford, Oxford, UK). Cells were sorted on the basis of low ductal marker expression, low 7-AAD expression and high Newport Green staining using pre-specified gating parameters.

2.6 Studies used for genetic analyses

Information on the variants identified in two large sequencing studies using T2D cases and normoglycaemic controls were used to refine known common T2D association signals in chapter five. These data were also used to interrogate the overlap between T2D association data and islet annotations in chapter six.

2.6.1 Genetics of Type 2 Diabetes (GoT2D) study

The GoT2D study consists of 2,657 individuals of Central and Northern Europe ancestry comprising 1,326 T2D cases and 1,331 controls. Samples were recruited as part of the FUSION study (n=979)[118], British Diabetic Association Warren 2 repository (n=322)[119], St. Thomas' UK Adult Twin Registry (n=322)[120], KORA-gen (n=205)[121], and the Finnish [Botnia] and Swedish [Malmo] prospective cohorts (n=829)[122,123]. To interrogate the full allele frequency spectrum, these samples were studied using low-pass (4x) whole-genome sequencing, high coverage (>30x) exome-sequencing, and high density genotyping arrays (Illumina Omni 2.5 and HumanExome BeadChip).

2.6.2 Type 2 Diabetes Genetic Exploration by Next-generation sequencing in multi-Ethnic Samples (T2D-GENES) study

The T2D-GENES study used in this thesis comprises 10,283 individuals spanning five population ancestries, of which 5,147 have been diagnosed with T2D and 5,118 are normoglycaemic controls (T2D status was unavailable for 18 samples). All individuals were exome-sequenced at high depth (>30x – details on contributing study in Table 2.4).

Ancestry	Individuals	Contributing studies
African-American	2,086	– Jackson Heart Study[124] – Individuals recruited from North Carolina, South Carolina, Georgia, Tennessee, or Virginia[125]
South Asian	2,219	– London Life Sciences Prospective Population Study[126,127] – Singapore Indian Eye Study[128]
East Asian	2,165	– Korean Association REsource[129] – Singapore Diabetes Cohort Study – Singapore Prospective Study Program [130-133]
Hispanic	1,959	– San Antonio Family Heart Study[134] – San Antonio Family Diabetes/Gallbladder Study[135] – Veterans Administration Genetic Epidemiology Study[136] – Family Investigation of Nephropathy and Diabetes[137], San Antonio component – Individuals from Starr County, Texas[138]
European	1,854	– Metabolic Syndrome in Men study[139] – Ashkenazi individuals recruited from the metropolitan New York region[140]

Table 2.4: description of the study samples used in the T2D-GENES study

2.7 Bioinformatics and statistical methods

2.7.1 Determining overlap between variants and annotations

The overlap between (islet) annotations and eQTL variants in chapter four, as well as fine-mapped variants in chapter six, was determined using the "*intersect*" function of BEDTools version 2.16.2[141]. Islet annotations used to determine overlap were the miRNAs expressed in human islets and their TargetScan version 6.2[142] predicted mRNA targets as described in chapter three, mRNAs found to be expressed in human islets in chapter four, and islet chromatin signatures from Pasquali *et al.*[143]. Further chromatin signatures were derived by Dr. Kyle Gaulton (Wellcome Trust Centre for Human Genetics, University of Oxford, Oxford, UK) from publicly available data on nine cell types from the ENCODE project (GM12878 cells [lymphoblastoid cell line], H1 human embryonic stem cells, K562 cells [immortalised myelogenous leukemia line], HUVEC [human umbilical vein endothelial cells], Hep G2 cells [immortalised hepatocellular carcinoma line], HSMM [skeletal muscle myoblasts], HMEC [mammary epithelial cells], NHEK [epidermal keratinocytes], NHLF [lung fibroblasts])[111] as well as hASC pre-adipocyte cells[144] and used to determine potential enrichment of islet annotations in chapters four and six.

2.7.2 Variant type annotation with the Variant Effect Predictor tool

Effects of variants of interest were annotated using the Variant Effect Predictor tool[145] on default settings, with Ensembl release 73[146] as the database providing information on genes, transcripts and protein sequence. This tool provides annotations for variants overlapping splicing features ("*splice_donor_variant*", "*splice_acceptor_variant*", "*splice_region_variant*"), start or stop codons ("*stop_gained*", "*stop_lost*", "*initiator_codon_variant*", "*incomplete_terminal_codon_variant*", "*stop_retained_variant*"), coding sequence ("*frameshift_variant*", "*missense_variant*",

"*synonymous_variant*", "*coding_sequence_variant*"), untranslated regions ("*5_prime_UTR_variant*", "*3_prime_UTR_variant*"), non-coding sequence ("*mature_miRNA_variant*", "*non_coding_exon_variant*", "*nc_transcript_variant*", "*intron_variant*", "*NMD_transcript_variant*"), and annotations for variants not overlapping any known features ("*upstream_gene_variant*","*downstream_gene_variant*", "*intergenic_variant*"). For most analyses, the most deleterious annotation per gene was used for each variant.

2.7.3 Statistical analyses

Fisher's exact test, binomial test, Spearman *rho* coefficient test, and hierarchical clustering were all implemented in the R statistical package version 2.15.1 (http://www.r-project.org)[147]. Permutations to establish empirical *P*-value cut-offs in chapters three and four were performed using custom Perl scripts. Unless otherwise specified, statistical significance was defined as P-value < 0.05.

Details of chapter-specific statistical methods are provided within relevant chapters.

Chapter 3

The miRNA profile of pancreatic islets and beta-cells

3.1 Introduction

The vast majority of T2D association loci identified by GWAS map to the non-protein-coding genome[56]. Accurate interpretation of the functional impact of identified variation therefore relies upon an understanding of the processes involved in regulation of transcription. Since most transcriptional regulation acts in a tissue-specific manner, it is important to focus on those regulatory processes that are active in the cells and tissues implicated in T2D pathogenesis[111].

3.1.1 miRNAs are important regulatory RNAs

miRNAs are short (~22nt) non-coding RNA regulators which have attracted significant interest in recent years. Since the description of the first miRNA, lin-4, in C. Elegans in 1993[148], miRNAs have been extensively studied. miRNA precursors (pri-miRNAs) are transcribed by polymerase II or spliced from introns of protein-coding genes[149]. In the nucleus, these pri-miRNA hairpins are cleaved into ~70 nucleotide pre-miRNAs by Drosha, a member of the RNase III enzyme family[149]. After transport to the cytoplasm, the pre-miRNA is further processed by Dicer, another RNase III family member, to the mature miRNA product and loaded into a miRNA-induced silencing complex (miRISC) (Figure 3.1)[149]. In humans, the number of identified mature miRNAs, as catalogued by miRBase[150], now stands at nearly 2,600.

After loading into the miRISC, miRNAs can bind to their target mRNAs according to complementarity. This binding regulates the target transcript by either translational repression or mRNA destabilization through deadenylation (Figure 3.2), although the specific contribution of these two pathways is not yet determined[151]. The non-coding nature of miRNAs means that no protein is available to facilitate an easy functional follow-up. However, several techniques have now been developed to globally profile miRNA-mRNA binding (e.g. PAR-CLiP [Photoactivatable-Ribonucleoside-Enhanced Crosslinking and

Immunoprecipitation][152] and HITS-CLIP [High-throughput sequencing of RNA isolated by crosslinking immunoprecipitation][153]). These techniques crosslink paired miRNA-mRNA complexes with protein components of the miRISC, followed by immunoprecipitation of one of the miRISC proteins (generally Argonaute). The immunoprecipitated RNA is subsequently sequenced and potential miRNA-mRNA pairs are predicted from the resulting data. However, since the number of currently reported profiles is low, determining which genes are regulated by specific miRNAs still relies heavily on computational predictions. For example, using the Targetscan v6.2 algorithm, which predicts targets based on conserved matches to bases 2-7 of the mature miRNA sequence, miRNAs are currently predicted to target 60% of all known protein-coding genes, with each miRNA affecting on average 350 genes[142]. The limited availability of transcriptome-wide miRNA target profiling data means that establishing the accuracy of these computational approaches is currently impossible.

Regardless of these unanswered questions, the pivotal role of miRNAs throughout vertebrate development is well established. Several animal models have demonstrated that miRNAs are required for normal embryogenesis and cellular differentiation. In zebrafish, *Dicer1* mutants lacking both maternal and zygotic Dicer activity show defects in morphogenesis, organogenesis and germ layer formation[154,155], whilst abolishing Dgcr8 (a miRISC subunit) function in mouse embryonic stem (ES) cells demonstrates that miRNAs are needed for silencing ES cell pluripotency[156]. In adults, miRNAs are also instrumental to normal cellular functions. Knocking out *Dicer1* in adult mice through an inducible Cre recombinase results in rapid development of defects of the intestine, bone marrow, spleen and thymus – and is fatal within 10 days[157].

3.1.2 Dysregulation of miRNAs in disease

Given these crucial roles for miRNAs, it is not surprising that dysregulated miRNA expression or changes in miRNA binding can lead to disease. For example, miRNAs have been implicated in the different stages of cancer development and progression. Global miRNA analysis of several human cancers shows a broad reduction in mature miRNA

expression levels[158]. Subsequent global knockdown of miRNA biogenesis through conditional deletion of *Dicer1* in mouse lung adenocarcinoma cells demonstrated that such global reduction in mature miRNAs promotes tumorigenesis[159]. These data suggest that miRNAs can play a role as tumour suppressors.

Conversely, specific miRNAs can also act as oncogenes. One of the established oncogenic miRNAs is the *mir-17–92* cluster at chromosome 13q31. This cluster of miRNAs, and most of the mature miRNAs to which it gives rise, is upregulated in many different types of cancer[160]. The non-coding RNA *C13orf25* containing the *mir-17–92* cluster was shown to be under transcriptional control of another known oncogene, MYC[161]. Overexpression of both *Myc* and the *mir 17-92* cluster leads to a significant increase in tumour development in murine haematopoietic stem cells compared to overexpression of *Myc* alone[162]. The oncogenic potential of the *mir 17-92* cluster cannot be attributed to a single miRNA alone, but rather seems to result from the orchestrated actions of multiple miRNAs in the cluster[162].

miRNA dysregulation has also been implicated in cardiovascular and neurological disorders, including heart failure and amyotrophic lateral sclerosis[163,164], as well as syndromic diseases such as Rett's syndrome[165]. Not only changes in miRNA expression levels, but also altered miRNA binding at specific mRNA targets caused by single nucleotide variants, can result in disease. This has now been demonstrated for a wide range of diseases ranging from unrestricted muscular growth and hypertension to Crohn's disease and Tourette's syndrome[166-169]. The Crohn's disease *IGRM* locus provides a good example of how studying miRNAs can provide insights at disease associated loci. Variants across this locus (including a 20kb deletion) have been associated with inflammatory bowel disease in general, as well as Crohn's disease in particular[170,171]. Both the 20kb deletion and the current lead variant for inflammatory bowel disease (rs11741861)[170] at this locus are in complete linkage disequilibrium (LD) with a synonymous variant within the *IRGM* gene itself (rs10065172). The autophagic protein IGRM plays an important role in the innate immune response against intracellular pathogens[172]. While rs10065172 does not affect IRGM protein

function, it does overlap a predicted binding site for the miR-196 family. Brest *et al.* demonstrated that miR-196 is normally upregulated in the intestinal epithelia of individuals with Crohn's disease, and this upregulation leads to a decrease in *IRGM* expression[168]. However, the risk allele for Crohn's disease abolishes the miR-196 binding site, resulting in a loss of *IGRM* expression regulation[168]. In carriers of the risk allele, the autophagic response initiated by IGRM is therefore no longer regulated by miR-196[168]. In turn, this dysregulation is likely to affect the intestinal inflammation seen in Crohn's disease.

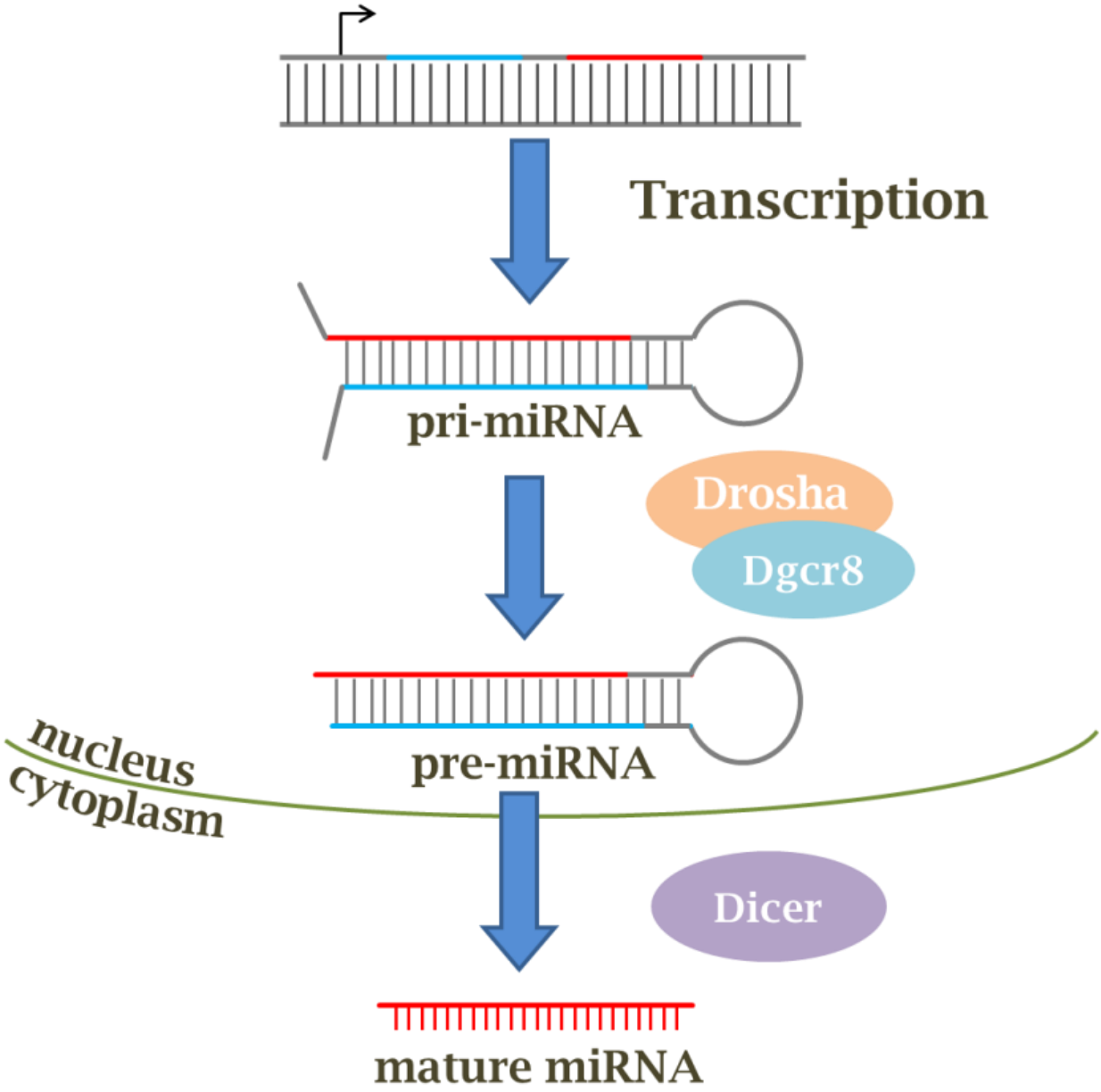


Figure 3.1: miRNA biogenesis. After transcription by RNA polymerase II, the primary miRNA is processed into a shorter precursor hairpin by Drosha. This hairpin is transported from the nucleus through a nuclear transport receptor, cleaved into mature miRNA by Dicer and finally loaded into the miRISC.

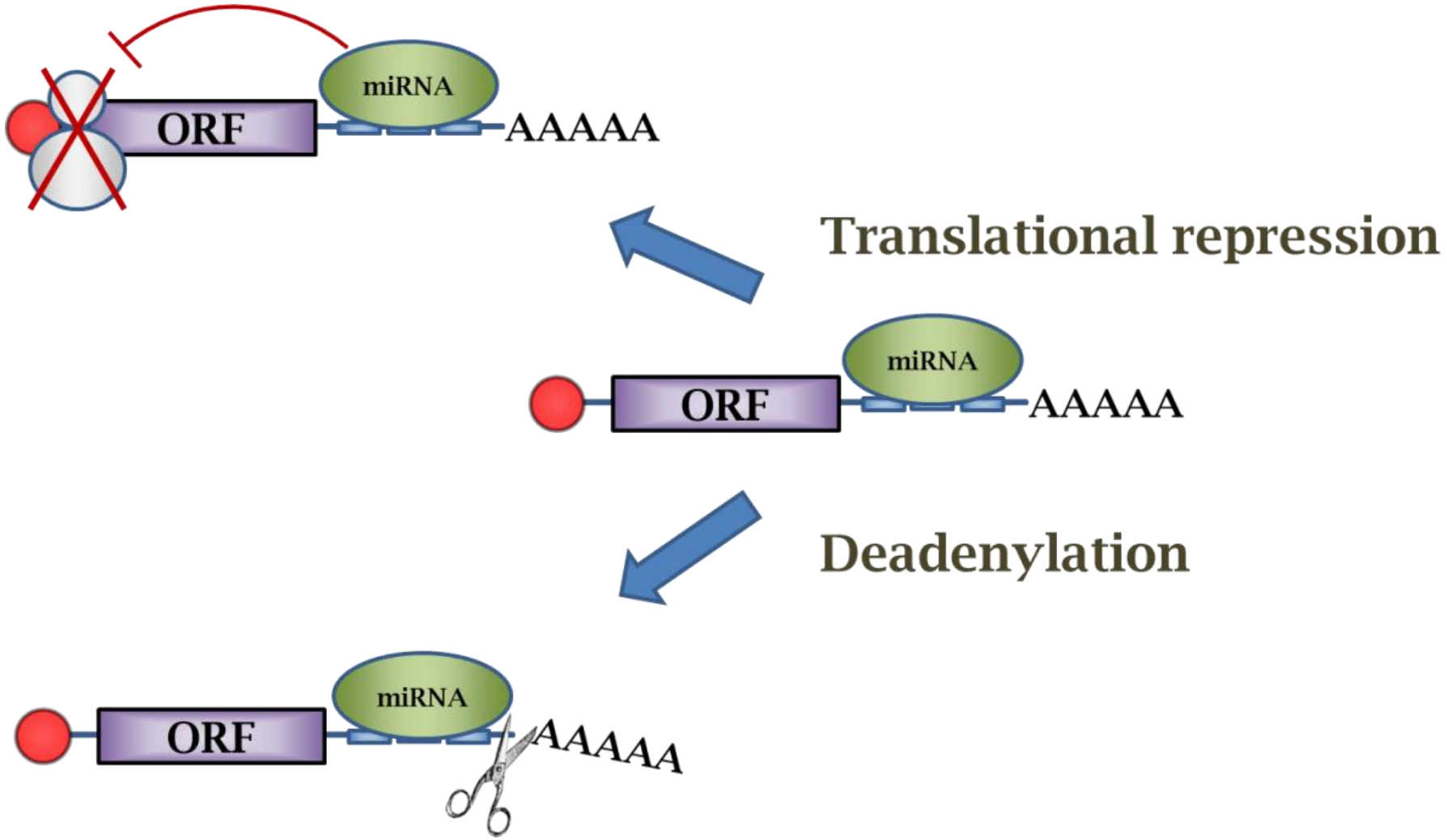


Figure 3.2: miRNAs action. After binding, miRNAs regulate their mRNA targets by either translational repression through stalling of the ribosome complex or mRNA destabilisation through deadenylation.

3.1.3 miRNAs in glucose homeostasis

One of the first miRNAs to be linked to a phenotype was the islet miRNA miR-375. A study by Poy et al. in 2004 showed that miR-375 is involved in regulation of glucose-stimulated insulin secretion (GSIS) in the murine insulin-secreting cell-line MIN6[173]. These changes were mediated through its effects on target mRNAs such as *Mtpn*, a protein involved in vesicle transport[173]. Since then, several other miRNAs with potential roles in insulin secretion have been identified, primarily in rodents. For example, miR-124a and miR-96 negatively regulate insulin secretion by acting on genes involved in the exocytosis pathway[174,175], while increased expression of another miRNA, miR-33a, reduces GSIS by modulating islet cholesterol homeostasis[176]. More recently, miRNAs have also been shown to be involved in maintaining glucose homeostasis through regulation of insulin sensitivity. In mouse models of obesity (and in some cases also in obese human individuals), miR-143, miR-103, miR-107, miR-181, and miR-802 were all found to be upregulated and, upon further investigation, shown to impair hepatic glucose sensitivity[177-180]. By investigating potential targets of these miRNAs, these studies uncovered key proteins mediating effects on hepatic insulin sensitivity. Some of these identified proteins, including sirtuin-1 and caveolin-1[178,179], had previously described roles in this pathway, whereas the role of others (e.g. Hnf1b[180]) was more unexpected.

Given their established roles in other diseases, as well as in islet physiology and insulin sensitivity, miRNAs and miRNA regulation have excellent candidature for a role in T2D. As outlined in section 1.3.2, many of the variants robustly associated with T2D are thought to affect beta-cell function. Since islet physiology and architecture show significant differences between rodents and humans[181,182], and with few suitable human beta-cell lines available[183], the genomic characterisation of primary human islet preparations provides an important opportunity to develop functional annotations that can support biological inference at T2D association signals.

In this chapter I set out to define the miRNA profile of primary human islets and enriched human beta-cell preparations to provide such functional annotations and increase our knowledge of miRNA expression in human islets and beta-cells.

3.2 Samples and Methods

3.2.1 Human pancreatic islets and beta-cells

Human islets were obtained (with research consent) from the Oxford DRWF Human Islet Isolation Facility (n=4) and through existing collaborations with Barcelona (n=2) from deceased donors of European descent. Clinical characteristics of the donors can be found in Table 3.1. High islet purity was established both by dithizone labelling as part of the islet isolation facility's quality control process, and through comparative quantification of endocrine and exocrine markers by qRT-PCR as described in section 2.4.

The beta-cell enriched preparation was collected through fluorescence-activated cell sorting of human islet preparations as described in section 2.5. RNA was extracted from all samples as described in section 2.1.1 and checked for integrity as described in section 2.2.

3.2.2 Library preparation and sequencing

Only samples with a total RNA yield in excess of 1ug and a quality score (RIN) > 7 were selected for library preparation. Libraries were prepared and sequenced at the High-Throughput Genomics group (Wellcome Trust Centre for Human Genetics, University of Oxford, Oxford, UK) using Illumina v1.5 (2 islets and 1 beta-cell sample) or TruSeq v1 (2 beta-cell and 1 islet sample) small library preparation protocols, and sequenced on one lane each using single-ended 50 basepair reads on Illumina GAIIx and HiSeq2000 platforms respectively.

3.2.3 Raw sequence processing

Raw data were obtained in Fastq format, and pre-processed to remove residual 3'-adaptor sequences using the fastx_clipper function from the FASTX-toolkit (http://hannonlab.cshl.edu/fastx_toolkit/index.html). Sequences less than 16 bases after adaptor stripping, and reads containing primarily N bases, were removed. The read lengths of the remaining reads show enrichment of a peak around 21 bases (Figure 3.3), which corresponds to the expected size of miRNAs.

3.2.4 Sequence mapping and quantification

Since ambiguities in read mapping, together with frequent contamination by large numbers of adaptor dimers, pose significant challenges to accurate quantification of small RNAs, a three step alignment approach was used. Reads were aligned using novoalign v. 2.07.11 (http://www.novocraft.com, parameters -h 60 60 -t30 -s -m -l 16 -R 0 -r A 30), which is aware of sequence ambiguities, to three consecutive references:

1. Contaminating sequences of adaptors, linkers, adaptor-linker, and adaptor-tag combinations.

2. Full-length miRBase v18 hairpin sequences, combined with other known and predicted human small RNA sequences present in Ensembl v63[150 184].

3. Human reference genome (NCBI build 37).

Reads mapping to contaminating sequences were excluded. The remaining reads were prioritised in order miRNA > ncRNA > genome, and alignments with the smallest edit distance were selected. miRNA hairpin aligning reads were further split out into mature miRNAs, star miRNAs and those reads mapping only to the hairpin by using the respective coordinates on the hairpin from miRBase v18 extended by 3 bases either side of the mature and star sequences.

Finally, expression levels of each ncRNA were quantified by counting the aligning number of reads. In case of multiple mapping, where a read maps between k alternative

sequences in one reference, 1/k was assigned to the count of each. Those RNA sequences observed less than 100 times across all samples were excluded. This cut-off was chosen since the reproducibility of ncRNA expression levels across samples fell below 100, and it only excluded <5% of all mapped reads from the analyses.

3.2.5 Normalization

Data were normalized to be able to compare read counts between samples, which differ in read depth. An inflation factor *i* was calculated for each library *l* using a method proposed by Anders and Huber[185]. This was determined as the median inflation factor across all genes *g* using the following formula:

$$il = \underset{g}{\text{median}}\left(\frac{\mathrm{n}_{gl}}{\underset{l}{\mathrm{GM}}(\mathrm{n}_{gl})}\right)$$

where n_{gl} is the read count each gene *g* in that library *l*, and GM is the geometric mean. The data were then $\log_2$-transformed $[\log_2(\frac{\mathrm{n}_{gl}}{il})]$ to account for heteroskedastic distribution. These values were used in all subsequent analyses.

3.2.6 Tissue specificity analysis

To determine whether miRNAs showed enriched expression in human islets, islet profiles were compared against publicly available human small RNA sequencing data using Illumina technology. Data from B-cells [186], liver[187], pigment cells [188], pooled thymocytes, bone marrow, CD34+ progenitor cells[189], skin[190], lung, kidney, skeletal muscle, heart, pancreas, frontal orbital gyrus, spleen, and liver tissue[191], as well as data from adipose tissue available through collaboration with the MuTHER consortium[192], were downloaded, realigned and normalized together with the islet and beta-cell samples. To minimise background noise (for example due to PCR artefacts), only those miRNAs observed at least 1000 times across the normalized values of all samples (n=367) were taken forward for tissue specific analysis.

Subsequently, a tissue specificity score, defined as the expression of an miRNA in a tissue divided by the sum of its expression in all tissues, was calculated for each miRNA in each tissue. The resulting fractional expression is an indicator of the proportion of that miRNA's expression which is attributable to a given tissue. Since this approach would be limited in a set of tissues containing very similar profiles, the results from islets and beta-cells were combined and the correlation between the different profiles was calculated. The resulting matrix was subjected to single-linkage hierarchical clustering in R version 2.15.1 using hclust[147]. To create profiles as unrelated as possible, tissues with a Euclidian distance < 0.25 were averaged along their branch. For each of the remaining 11 profiles, tissue specific scores were calculated. Tissue-specificity was defined by a score > 0.5, which, using a permutation-based approach comprising 1 million permutations of the observed data, amounted to a false-discovery rate of 8%.

	Isl1	**Isl2**	**Isl3**	**Beta1**	**Beta2**	**Beta3**
Gender	Male	Female	Female	Male	Female	Female
Age (years)	59	57	32	41	51	59
Cause of death	Ischemic brain injury	Ischemic brain injury	Trauma	Meningioma	N/A	N/A
BMI (kg/m^2)	24	36	27	29	31	29
Pre-mortem diagnosis of diabetes	No	No	No	No	No	No
Cold ischaemia (hours)	6.4	8.5	8	8.5	5	7.5
Islet purity (%)	94	75	80	80	90	70
QC+ reads	14,000,000	648,000	68,000,000	3,000,000	38,000,000	26,000,000

Table 3.1: Clinical characteristics of human islet donors. N/A denotes information not available.

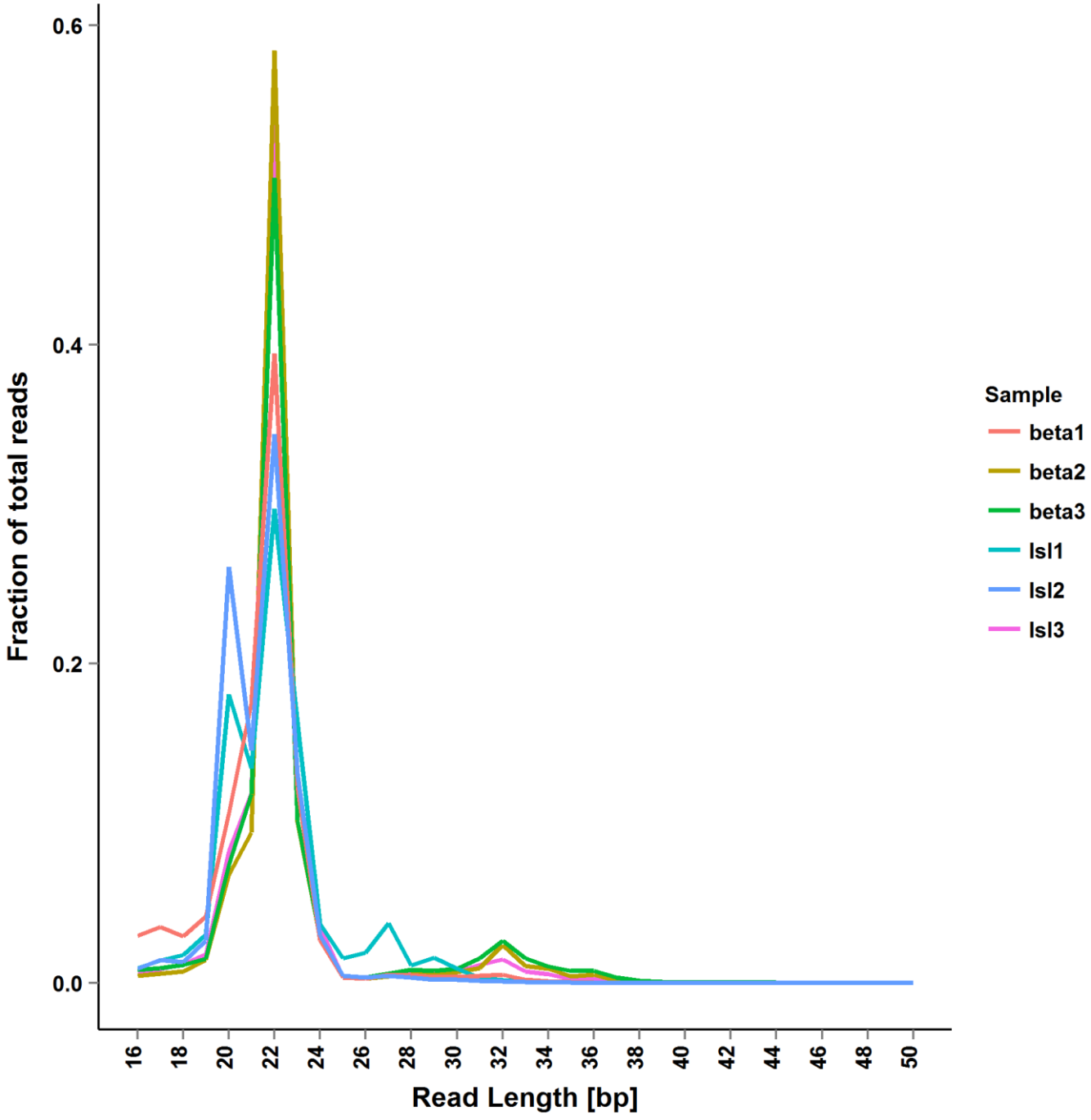


Figure 3.3: Read length distribution of trimmed reads. The fraction of total reads per sample (y-axis) is plotted by read length after adapter removal (x-axis) Remaining adapter sequences were removed from the 50 basepair reads using Fastx Clipper tool, requiring a minimum length of 16 basepairs after trimming. A clear peak is present at the expected length (20-22 basepairs) of mature miRNA transcripts.

3.3 Results

3.3.1 miRNA expression in human pancreatic islets and beta-cells

The data presented in this chapter were generated using human islets from six donors acquired with ethical consent. To allow comparison between islet and beta-cell profiles, fluorescence-activated cell sorting (FACS) was used for three of the samples to select for beta-cells. In line with previous reports on this method[117], the proportion of beta-cells in the preparation following FACS ranged between 90-95%, and, as such, these will subsequently be referred to as "enriched beta-cell" preparations. For consistency, all six samples (3 islet and 3 enriched beta-cells; see Table 3.1 for clinical characteristics) were sequenced as 50 base pair single reads on Illumina platforms (either GAIIx or Hiseq2000). Due to improvements in sequencing technology over time, there was an appreciable increase in read depth during the period of sample collection; 648,000, 14 million, 68 million for islets and 3 million, 38 million, 26 million for enriched beta-cells respectively (Table 3.1). In total, after quality control, 83 million reads mapping to small RNA species were acquired across the three islet samples and 68 million across the enriched beta-cell samples, providing ample sequencing depth for the analyses presented in this chapter.

The size selection step during library preparation and alignment process should favour detection of miRNAs over other small non-coding RNA species such as small nucleolar RNAs (snoRNAs) - which are generally longer. Indeed almost all reads (average 92%) mapped to mature miRNA sequences annotated in miRBase v18[150], with the remainder being primarily snoRNAs (average 2.4%) plus a variety of the other non-coding species in Ensembl v63[184] (Figure 3.4).

Since the number of annotated human miRNAs is still expanding and newly discovered miRNAs are often cell type specific, the miRDeep2 package – a probabilistic method for

discovering miRNAs from small RNA sequencing data using the predicted secondary structure of potential miRNA precursors[193] – was used at default settings to uncover novel, unannotated miRNAs in the sequencing data. No novel miRNAs were identified using this method, which could indicate that there are indeed no novel miRNAs present in human islets or, due to the limited number of samples, there were not enough observations for any novel miRNA to allow its detection.

There was good reproducibility across the three samples of each type (Spearman *rho* ranging from 0.73-0.97; Figure 3.5). These expression profiles were of medium complexity with an average of 38 miRNAs responsible for 90% of the total miRNA aligned reads in islets, and 21 in enriched beta-cells. Three hundred and sixty six known mature miRNAs were expressed above background levels (defined as more than 100 combined reads across the three samples of the same origin) in islets, and 346 in enriched beta-cells (Figure 3.6; Appendix Table 1). Between these islets and beta-cells, 384 unique miRNAs were identified, of which 85% (n=328) were shared between groups (Appendix Table 1). All but one expressed miRNA was present in at least two of the three samples of each type, and the most abundant miRNAs showed consistently high expression across the samples.

Several of the most highly expressed miRNAs in human islets (including miR-375, miR-7-5p, miR-148a-3p, miR-26a-5p and miR-127-3p) have been previously found to be expressed in rodent or human islets[173,194,195], although the physiological function of most is yet to be determined. Other miRNAs highly expressed in the islet data presented in this chapter, such as miR-143-3p, have been shown to regulate insulin sensitivity in other tissues, but have not previously been implicated with islet-expression[177]. Finally, the analysis uncovers a substantial set of miRNAs (e.g. miR-27b-3p, miR-192-5p) for which expression in islets has not been previously described. Islet expression of these miRNAs has been confirmed by studies published after the analyses in this chapter were published[196,197].

Comparison of the miRNA expression profiles of islets and enriched beta-cells reveals a strong correlation between the two (Figure 3.6). Most of the other abundant miRNAs show

comparable levels of expression between islets and beta-cells, although this observation is limited in power by having only three samples for each of the populations - and derivation of islets and beta-cells from different donors. However, it also highlights some interesting (although not statistically significant) differences. Amongst the most highly expressed miRNAs, miR-375 shows substantially higher abundance in enriched beta-cells than islets (42% vs 27%). On the basis that an average islet contains ~50% beta-cells[182], and, given an estimated 90% purity of the FACS-enriched beta-cell preparations, the ~1.5 fold enrichment of miR-375 in beta-cells compared to islets indicates that, in human as in rodents[173], this miRNA is predominantly expressed in beta-cells. This is supported by a recent study[197], which compared the miRNA expression profiles of sorted human alpha- and beta-cells using quantitative RT-PCR, and found a significant enrichment of miR-375 in beta-cell compared to alpha-cells. Of the ten most significantly enriched beta-cell miRNAs in the study, five, including the two previously implicated with beta-cell function (miR-127 and miR-204), also show an increase in beta-cell expression profiles in the data presented in this chapter.

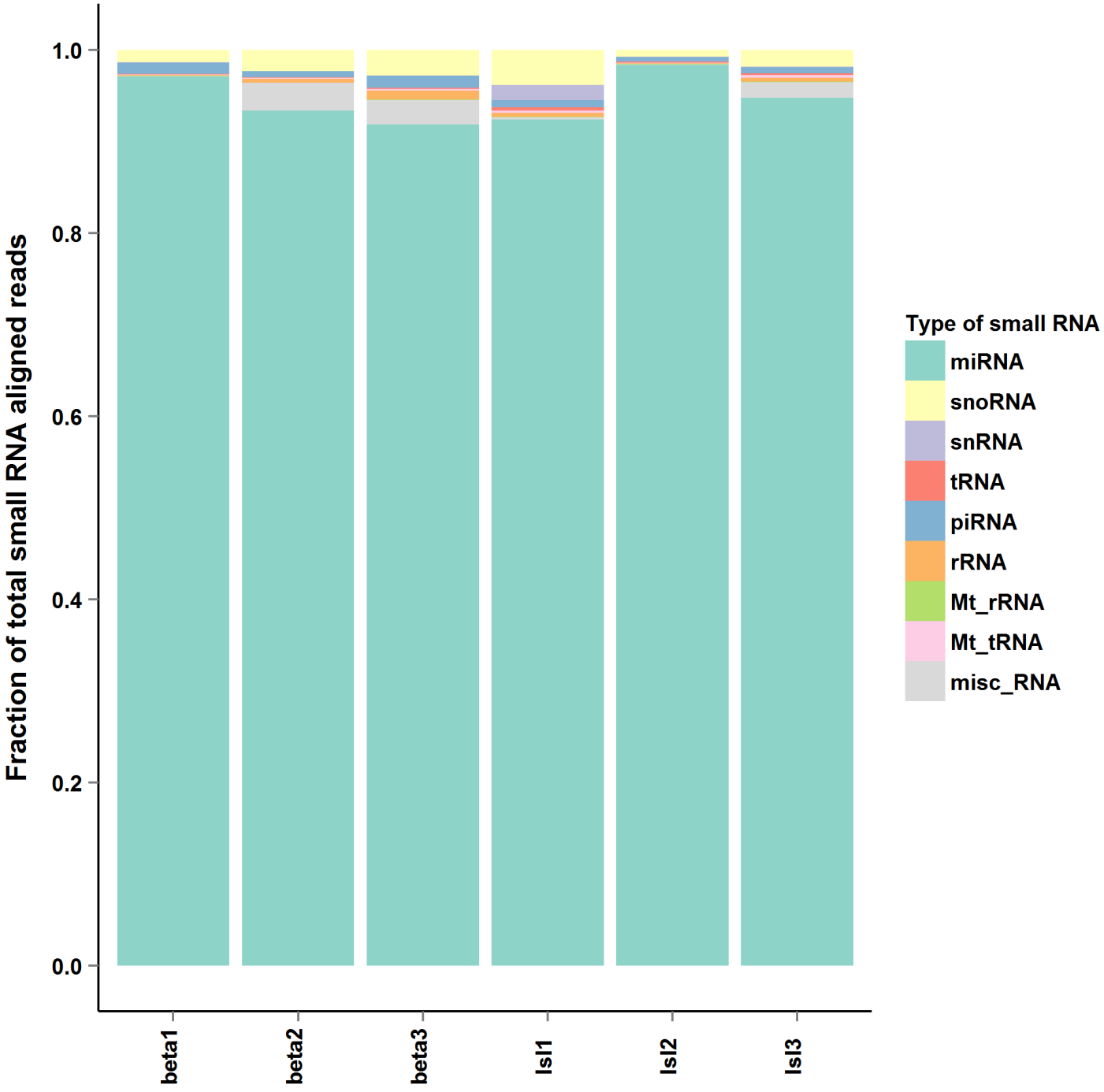


Figure 3.4: Expression of the different types of small RNAs in islet and beta-cell samples. For each of the samples, the fraction of total small RNA aligned reads for each type of small RNA annotated by the pipeline (y-axis) is plotted as a stacked histogram. It is evident that the vast majority of small RNA aligning reads map to miRNAs.

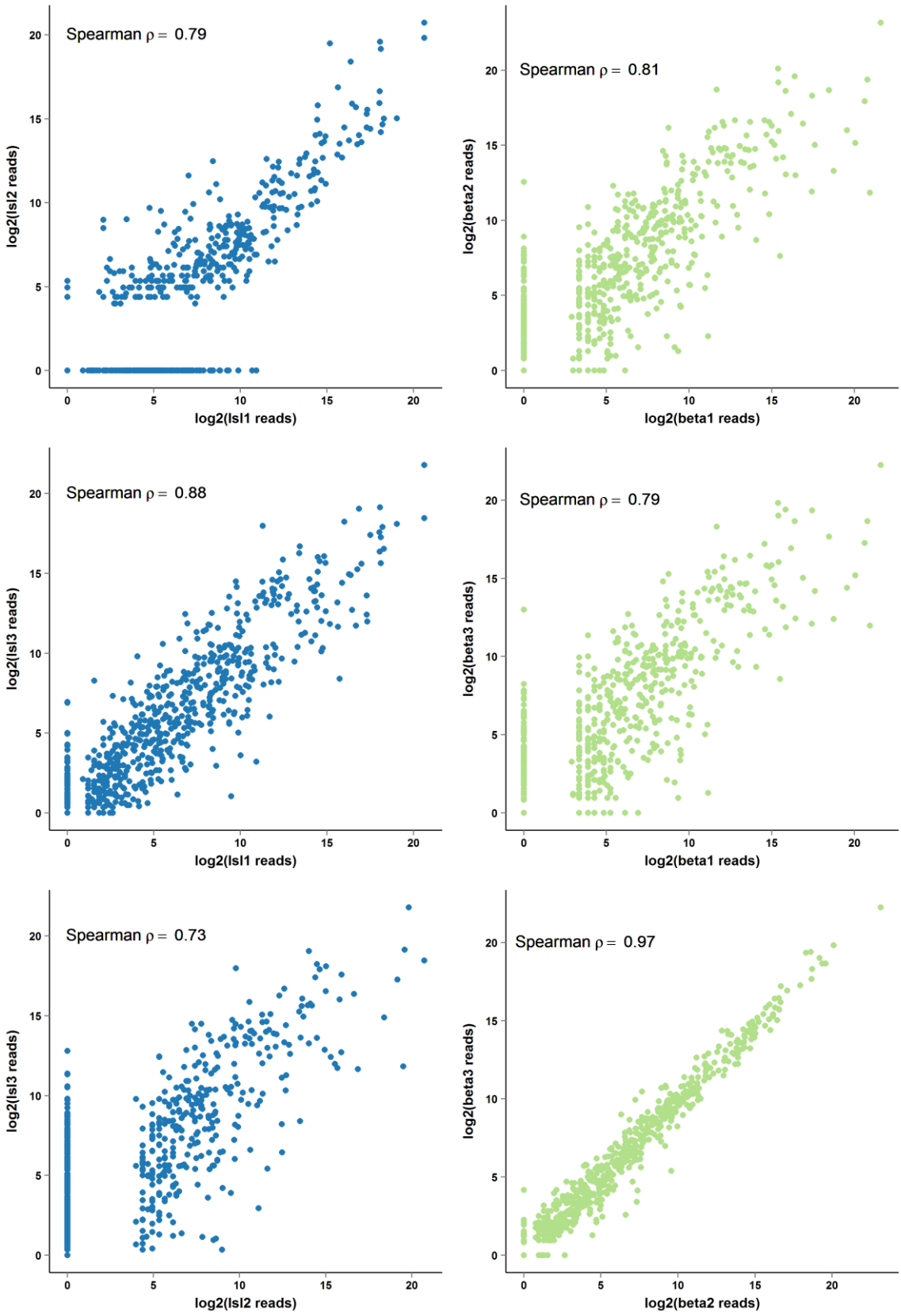


Figure 3.5: Correlation between miRNA expression profiles of sequenced samples. Scatterplots of the expression values in $\log_2$(reads) for each of the whole islet samples (blue dots) and FACS enriched beta-cell samples (green dots). The correlation between the samples (as measured by Spearman *rho*) is annotated on each plot.

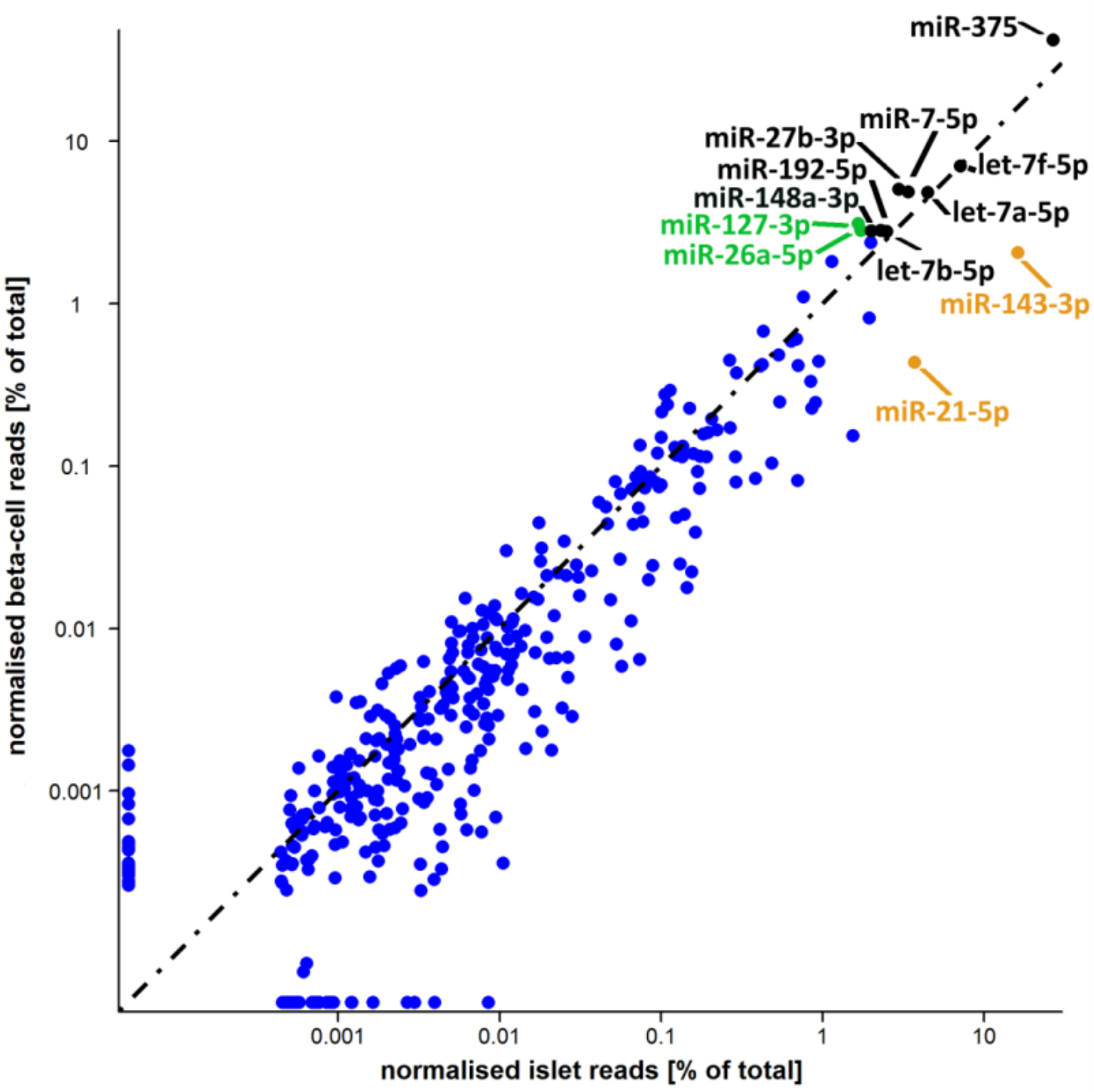


Figure 3.6: Islet and beta-cell miRNA profiles. The top 10 most abundant miRNAs in islets and beta-cells are annotated, with colours denoting those shared between the two top 10s (black), in islet top 10 only (orange), and in beta-cell top 10 only (green). There was good correlation between the two profiles ($r^2 = 0.78$). The black dashed line represents equality.

3.3.2 Tissue specificity of miRNA expression

To establish whether any miRNA found to be expressed in islets and beta-cells shows tissue-specific enrichment in expression levels, miRNA expression profiles were compared against equivalent, publicly available, next-generation sequencing-derived miRNA data from other human tissues, including B-cells[186], liver[187], pigment cells[188], pooled thymocytes, bone marrow, CD34+ progenitor cells[189], skin[190], lung, kidney, skeletal muscle, heart, whole pancreas, frontal orbital gyrus, spleen, liver tissue[191], and adipose tissue[192] present in the NCBI sequence read archive. After downloading, raw reads were realigned and normalized in parallel with islet and beta-cell samples (see sections 3.2.3, 3.2.4, and 3.2.5). A tissue specificity score, defined as the expression of a miRNA in the reference tissue divided by the sum of its expression in all other tissues (described in section 3.2.6) was derived from the combined results, with a score >0.5 considered to indicate tissue-specificity. Such a score indicates that half or more of all observed reads for a given miRNA were derived from a single tissue.

Forty of the 367 islet miRNAs were identified as islet-specific (Appendix Table 2). Equivalent specificity analyses using other tissues as reference found a mean of only 10 enriched miRNAs per tissue (ranging from 1 [in spleen, kidney and lung] to 31 [orbital gyrus] ; Appendix Table 2). By permuting the data, the false-discovery rate of this analysis at the 0.5 cut-off was estimated at 8%. Inevitably, these comparisons are influenced by the type and number of tissues for which data were available; for some there were much similar data available (e.g. the lymphoid cell types) while others, such as orbital gyrus, had very few related cell types present. Notwithstanding these issues, both the number of islet-specific miRNAs detected, and the hierarchical clustering, suggests that the miRNA profile of human islets is relatively distinct from the other available tissues (Figure 3.7).

The top 10 most islet-specific miRNAs (Figure 3.8), all of which have a specificity score >0.8, comprise a mixed selection of miRNAs previously reported to be islet-enriched and miRNAs not previously identified in islets at all. The findings presented in this chapter

replicate six out of the nine previously reported islet-specific miRNAs in humans and rodents (miR-184, miR-183-5p, miR-7-5p, miR-127-3p, miR-375, and miR-493-5p)[194,198]. Of these, miR-375 is an established islet miRNA with roles in both insulin secretion and the maintenance of normal alpha- and beta-cell mass[173,199], whilst miR-7-5p is thought to play a role during embryonic development of endocrine cells [198,200]. There is also published evidence for a role in insulin biosynthesis and secretion for miR-184, miR-182-5p and miR-127-3p, although for miR-184 and miR-127-3p this is restricted to a statistical correlation between islet expression levels and glucose-stimulated insulin secretion[194,195]. For other islet-specific miRNA transcripts identified through the tissue specificity analysis in this chapter, including miR-409-5p and miR-183-5p, the high degree of islet-specificity may point to novel roles in the development and maintenance of islet cellular phenotypes.

Primary miRNA sequences tend to cluster in genomic intervals spanning <10 kb, and it is thought that members of these clusters are transcribed together[192,201]. Several islet expressed miRNAs also show evidence for co-transcription. Such examples are found amongst miRNAs showing the greatest absolute abundance in islets, as well as amongst those that were most islet-specific. Two of the most islet-specific transcripts, miR-182 and miR-183, originate from the same cluster on chromosome 7q32.2 - whilst two others (miR-136 and miR-127) cluster together on chromosome 14q32.2. Across samples, expression levels of miRNAs within these clusters are highly correlated in both islets and enriched beta-cells (r^2 >0.6), suggesting that, like the oncomir *mir-17-92* cluster in section 3.1.2, they might act in unison to regulate islet function.

A recent publication by Kameswaran *et al.* shows this is indeed the case for the cluster on chromosome 14q32.2[196]. By comparing the miRNA profiles of three non-diabetic and four T2D donors, they identify this cluster as being downregulated in T2D[196]. By extending their analysis to ten T2D and fourteen normoglycaemic islet preparations using qPCR, they provide further evidence of significant downregulation in T2D of eight of the twelve miRNAs in this cluster that were studied[196]. Consistent with the findings presented in this chapter as

well as other published data[197], this miRNA cluster showed a beta-cell specific expression pattern – the long noncoding *MEG3* RNA containing the miRNA cluster showed around 20 fold higher expression in beta-cells compared to alpha-cells[196]. Furthermore, global profiling of miRNA-mRNA binding using HITS-CLIP identified 1,784 genes targeted by the miRNAs in the cluster[196]. Targets included genes located inside T2D association loci, such as *TP53INP1,* as well as islet-related genes, such as *INS-IGF2* and *IAPP*[196].

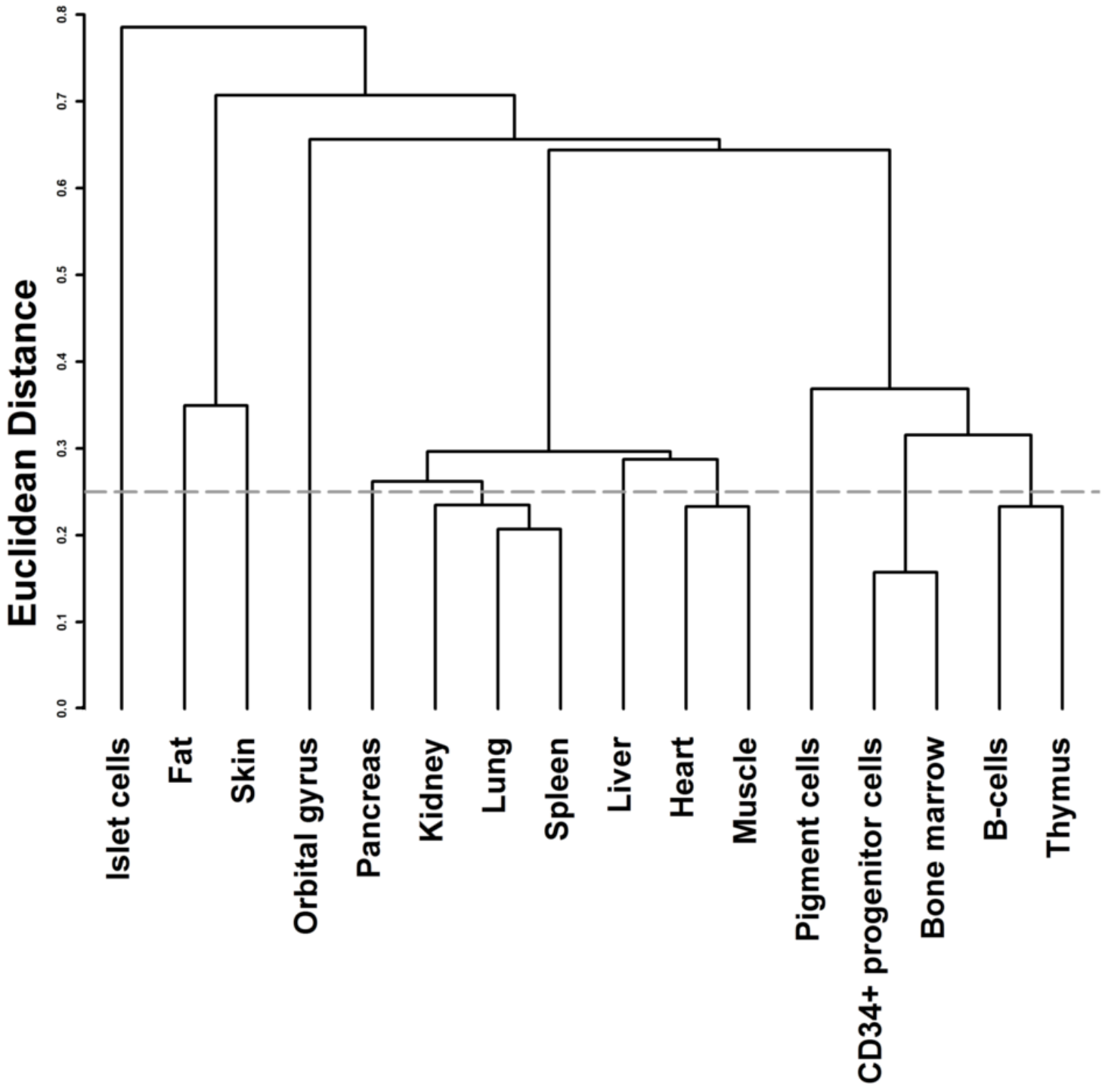


Figure 3.7: Hierarchical clustering of miRNA profiles. Results for the single-linkage hierarchical clustering of inter-tissue miRNA profile correlations for all available tissue. The dashed grey line indicates a Euclidean distance of 0.25.

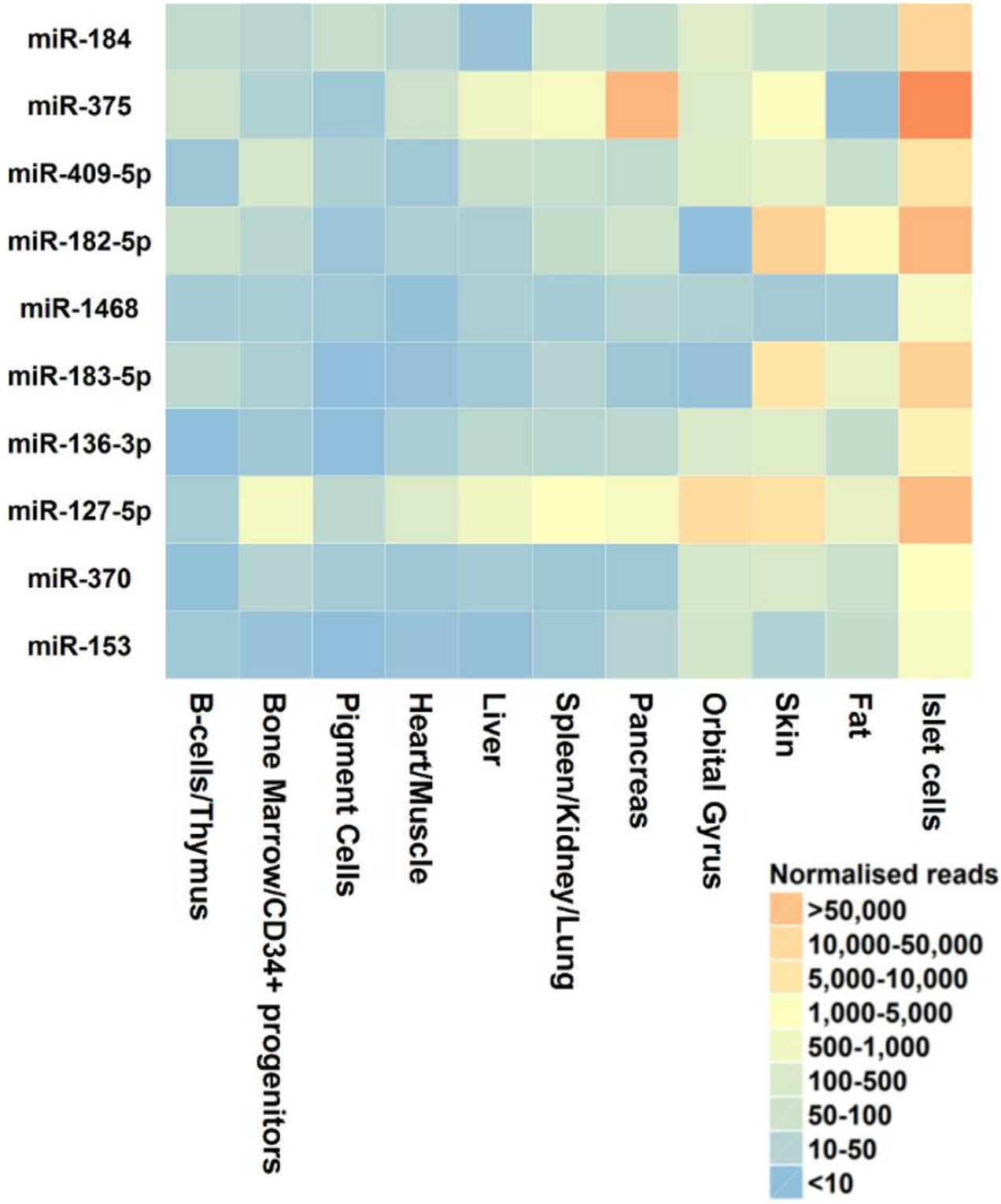


Figure 3.8: Islet-enriched miRNA. The top 10 most tissue specific islet miRNAs displayed in descending order. The colours indicate the normalized expression levels of these miRNAs across the different profiles used in the analysis.

3.4 Discussion

This chapter has described a catalogue of miRNAs expressed in human pancreatic islets and beta-cells generated to provide improved non-coding annotations and gain a detailed insight into the miRNA regulatory landscape of these endocrine cells.

The data described in this chapter can be used to corroborate data from rodent models or cell lines. For example, the abundance of miR-375 in the human islet miRNA profile provides valuable support for a critical role in human pancreatic beta-cell physiology, in line with its well-established role in rodent islet biology. As discussed in section 3.1.3, knockdown or over-expression of miR-375 in a clonal rodent beta-cell line (MIN6) influences GSIS[173]. Further studies in rodents have established *in vivo* effects of miR-375 on glucose metabolism: miR-375 null mice are hyperglycaemic and have a reduced beta-cell mass[199]. miR-375 knockdown in obese ob/ob mice has even more pronounced effects on glycaemia, resulting in a severe diabetic phenotype[199]. Having established that miR-375 is also abundantly expressed in human islets through the results shown in this chapter, further studies are warranted to define the contribution of miR-375 to the pathogenesis of T2D – perhaps studying the role of this miRNA in compensatory proliferative mechanisms during metabolic stress.

To uncover novel miRNAs potentially involved in normal islet physiology, focusing on those identified by the data presented in this chapter to be highly abundant in islets or highly islet-specific might be an interesting avenue. Indeed, for many of the miRNAs that are most abundantly expressed in human islets and/or show evidence of islet-enriched expression, there is little, if any, existing data concerning their effects on islet development, function or dysregulation in T2D. The highly abundant let-7 family is one example of such miRNAs warranting further functional investigation. Two recent studies have demonstrated that knockdown or overexpression of let-7, a family of miRNAs that is very highly conserved throughout the animal kingdom, alters murine insulin sensitivity in a range of tissues[202,203].

One of these studies also showed reduced insulin secretion during an intraperitoneal glucose tolerance test in the let-7 overexpressing mice, hinting at a role for let-7 in glucose-stimulated insulin secretion[203]. However, the effects of altered let-7 expression on pancreatic islet physiology were not further characterised in these studies.

A final example illustrating the usefulness and relevance of the islet miRNA profile comes from a recent collaboration between me and Dr. Anna Gloyn in Oxford and Dr. Matthew Poy at the Max Delbrueck Center for Molecular Medicine in Berlin, Germany. His group investigated the role of miR-184 – showing the highest level of human islet-enrichment of all the miRNAs studied in this chapter – in islets from a large number of transgenic mouse models. These studies have identified miR-184 as a regulator of compensatory beta-cell proliferation during insulin resistance by modulating the expression of a key component of the miRISC, Argonaute2[204]. These observations could be extended to humans using islets from the Oxford Islet Biobank and from Dr. Jorge Ferrer in Barcelona, Spain, in which miR-184 was found to be decreased in islets from T2D donors – mimicking the findings in the rodent models[204]. This result, combined with studies such as those into the miRNA cluster on chromosome 14q32.2 described in section 3.3.2[196], demonstrate that islet-specific miRNAs are highly enriched for functionally relevant miRNA.

In conclusion, the data presented in this chapter can serve as a valuable resource for translating rodent findings to humans, identifying miRNAs which could be involved in normal islet physiology and beta-cell function, and can also provide a reference for the study of miRNA-mediated abnormalities in T2D. The overlap between islet-expressed miRNAs, their mRNA target sites and the findings from T2D GWAS will be explored in chapter six.

Chapter 4

Transcriptome analysis of human pancreatic islets

4.1 Introduction

4.1.1 Transcriptome data can provide mechanistic insights at GWAS loci

Data derived from transcriptome profiles can shed light on the mechanisms through which variants identified from genome-wide association studies (GWAS) might be exercising their effects. Through expression quantitative trait loci (eQTL) and allele-specific expression (ASE) analysis, amongst others, variants can be linked to expression of specific genes. Since eQTLs are also often cell-type specific[205], such analysis might allow identification of both the gene(s) and tissue(s) underlying association signals. The potential success of this approach has been highlighted by several recent studies. *Cis*-eQTL analyses in three tissues (subcutaneous adipose tissue, lymphoblastoid cell lines [LCLs] and skin) from around 850 mono- and dizygotic twins, conducted as part of the MuTHER (Multiple Tissue Human Expression Resource) project, identified substantial overlap between *cis*-eQTL signals and GWAS associated variants present in the NHGRI Catalog of Published Genome-Wide Association Studies[206]. In 22-26% of the overlapping loci, there was strong evidence for coincidence of the eQTL and trait association signal[206]. – around twice as much coincidence as expected by chance[175]. The majority (~60%) of the eQTLs coinciding with GWAS signal were tissue-specific, with only 15% discovered in all three tissues[206]. Another recent study used RNA-sequencing to study eQTLs in LCLs from around 900 individuals as part of the 1000 Genomes Project[207]. Similarly to Grundberg *et al.*, this study also identified a significant (albeit more modest) enrichment of eQTLs amongst GWAS variants compared to matched null variants, as well as an overrepresentation of GWAS variants amongst the top eQTL variants[207]. Strictly speaking, overlap between an eQTL and association signal does not prove that the eQTL gene is causal. However, this is very likely to be the case. Studies such as these therefore demonstrate that eQTL analysis, when performed in appropriate tissues, can highlight likely causal mechanisms underlying GWAS loci.

The number of eQTL analyses performed in tissues critical in T2D pathogenesis is currently low, but those conducted are already having an impact on our understanding of T2D pathogenesis. Consider, for example, the T2D-association signal near the gene encoding KLF14, the transcription factor Kruppel-like factor 14. This particular locus has been clearly shown to exert its T2D influence through a primary effect on insulin resistance, rather than insulin secretion[56]. In human adipose tissue samples, it was possible to show that the same set of T2D-associated variants was regulating *KLF14* transcription in *cis*[109] and a wider network of adipose transcripts in *trans*[110]. These eQTL analyses therefore serve to define the molecular machinery through which these particular variants act. It is worth noting that the patterns of *cis*- and *trans*- transcript regulation appear to be entirely restricted to subcutaneous adipose tissue, illustrating once again the need for studies to be performed in specific cells of interest.

Whilst most of the published studies on gene expression in human islets (summarised in Table 4.1) are too small for formal eQTL analyses, they have been able to show that, for example, the vast majority of genes implicated in T2D through GWAS are expressed in pancreatic beta-cells[108]. ASE, which uses the ratio between reads from heterozygous exonic variants as a marker to assay differential transcript expression between the two alleles of a haplotype, can be detected from RNA-seq even in smaller studies. This is because background noise in ASE analyses is lower than in eQTL studies – with the alternative alleles serving as intra-sample controls of one another. ASE can serve as a functional assay on the strength of a genetic variant's effect on expression levels and needs, at best, only one heterozygous individual to detect allelic imbalance. However, ASE, and even more so eQTL analyses, are generally better powered to identify common regulatory variation. In sorted beta-cell populations, such analyses for example showed significant ASE for the rs5215 missense variant [p.E23K] in *KCNJ11*[108], which is robustly associated with T2D[56 46]. The significant ASE identified at rs5215 could indicate that the mRNA transcript containing the rs5215 risk allele is less stable, leading to reduced protein levels. However, since the effect of this missense variant (or another non-synonymous variant [*ABCC8* p. A1369S] in strong LD)

on protein function is thought to underlie the T2D association signal[76-78,208], further follow up and replication will be required.

ASE analyses can also be informative in deconvoluting mechanisms of susceptibility at imprinted loci. This approach has provided novel insights into potential causal mechanisms at the *KCNQ1* locus, at the imprinted domain on chromosome 11p15.5. Variation at this locus confers T2D risk only when maternally inherited[109], and subsequent studies identified significant changes in methylation status depending on the genotype of the T2D risk variant rs2237895[72,87]. By studying the expression patterns of the eight genes located within the imprinted region (*KCNQ1*, *KCNQ1OT1*, *CDKN1C*, *PHLDA2, SLC22A18*, *SLC22A18AS*, *KCNQ1DN* and *TRPM5*) in foetal pancreas and adult islets, Travers *et al.* could reduce the number of genes likely to be involved to three (*KCNQ1*, *KCNQ1OT1*, and *CDKN1C*)[72]. These three genes showed monoallelic expression, consistent with the observed effects of parent-of-origin on T2D-risk, in either the foetal tissue (*KCNQ1* and *KCNQ1OT1*) or at both time points (*CDKN1C*)[72].

In the case of T2D, ASE and eQTL studies should not only be performed in whole human islets, but also in the various cellular constituents of the islets. While sample size for such studies will be even more challenging than for whole islets, several groups are starting to accumulate expression data gathered from these specific cell-types[108,209,210].

Authors	Year	Ref	Method	Samples used for transcriptome profiling	Key Findings
Nica *et al.*	2013	[108]	RNA-seq	Whole islet, beta-cell and non-beta-cell fractions from 11 non-diabetic (ND) donors [not all sample types were available for all donors]	Identifies around 3000 genes differentially expressed in beta-cells compared to 18 other tissues Substantial differences between beta-cell and non-beta-cell fractions Finds evidence for allele-specific expression at T2D associated loci
Bugliani *et al.*	2013	[211]	Microarray	7 ND and 6 T2D islets	Identifies 1180 differentially expressed genes Ubiquitin–proteasome pathway consistently altered in T2D islets
Bramswig *et al.*	2013	[210]	RNA-seq	Alpha, beta and exocrine cells from 6 ND islets	Many alpha-cell genes display chromatin signatures corresponding to status of poised for activation Treatment of islets with a histone methyltransferase inhibitor results in colocalisation of alpha and beta-cell markers
Morán *et al.*	2012	[209]	RNA-seq	6 ND islets and 2 ND beta-cells	Identifies many novel, often islet-specific lncRNAs Functional studies implicates these lncRNAs in islet physiology and development Dysregulation of several islet lncRNAs in T2D
Eizirik *et al.*	2012	[212]	RNA-seq	5 ND islets	Catalogue of gene expression, and splice variants, in human islets Most genes implicated in T1D through association studies are expressed in islets

Mahdi *et al.*	2012	[213]	Microarray	38 ND and 10 T2D islets	Identifies clusters of genes associated with T2D and reduced insulin secretion Identifies SFRP4 as potential new biomarker for islet dysfunction in T2D
Taneera *et al.*	2012	[214]	Microarray	54 ND and 9 T2D islets	Integration of genetic variation with gene expression levels in islets and functional annotations identifies genes potentially involved in T2D pathogenesis
Negi *et al.*	2012	[215]	Microarray	Laser-captured beta-cells from 8 whole pancreases, 8 freshly isolated islets, 5 3-day cultured islets	Inflammatory response genes and pancreatic progenitor transcription factors are upregulated during islet isolation and culture
Dorrell *et al.*	2011	[216]	Microarray	Alpha, beta, acinar and ductal cells from 5 ND islets	Supports roles of known endocrine transcription factors MAFB in beta-cells, and ARX, IRX2 in alpha-cells. Identifies *HOPX* , *HBAC9* and *BATF2* as highly differentially expressed between different islet subpopulations
Martens *et al.*	2011	[217]	Microarray	10 ND beta-cells	Beta-cell gene pathways often of neuronal or mucosal origin ~ 15% of these pathways beta-cell specific
Dominguez *et al.*	2011	[218]	Microarray	7 ND and 6 T2D islets	Specific analysis of class II PI3K isoforms suggestes downregulation of PI3K-C2α in T2D islets
Marselli *et al.*	2010	[219]	Microarray	Beta-cells from 10 ND and 10 T2D donors	Comparison of T2D and ND islets identifies differential expression of genes linked to glucotoxicity, islet regeneration, and oxidative stress

Kutlu *et al.*	2009	[220]	RNA-seq	2 ND islet preparations	Proof-of-concept study provides global expression profile of human islets using 17bp RNA-seq
Marselli *et al.*	2008	[221]	Microarray	3 ND islet and beta-cells	Isolated islets show differential expression of pro-inflammatory genes, and genes related to hypoxia and oxidative stress

Table 4.1: Summary of all published studies using mRNA profiling in human islets.

4.1.2 RNA-seq and long non-coding RNAs

Next-generation sequencing technologies allow more detailed interrogation of the transcriptional and epigenetic landscape of a cell population, enabling the discovery of novel transcripts in regions that were previously unannotated[106,107]. Many of these novel transcripts are non-coding, have multiple expressed isoforms and show tissue-specific expression profiles[106] – making them less amenable to techniques such as microarray analysis.

Long non-coding RNAs (lncRNAs), one class of non-coding transcripts, have attracted considerable scientific interest in recent years. Like mRNAs, they are often transcribed by RNA polymerase II and can by poly-adenylated. However, unlike protein-coding genes, frequently very little is known about their function. It has been shown that lncRNAs can regulate other genes both in *cis* by activating transcription[222], as well as in *trans* by, for example, interacting with protein complexes[223]. Through GWAS, lncRNAs have been implicated in many diseases. One well-described example is a region on chromosome 9p21, encompassing *CDKN2B* antisense RNA 1 (*CDKN2B-AS1*, also known as *ANRIL* [antisense lincRNA of the *INK4* locus]), which has been implicated in a range of traits ranging from intracranial aneurysms to cancer and T2D[224].

A recent study, of which I was an author, focused on characterisation of the lncRNAs transcribed in several human islet and beta-cell preparations[209]. By intersecting islet transcriptomic data with ChIP-sequencing analysis of the chromatin modifications which help to define the structures of active genes, 1128 lncRNA genes were found to be expressed in human islets. My part in this publication centred on the intersection of these islet lncRNAs with human T2D genetic data, which will be further described in section 6.3.2. Many of the identified lncRNAs displayed tissue-specific expression and have genomic properties consistent with a role in islets: they were found in close proximity to genes known to play important roles in the development and maintenance of islet cellular phenotypes including *MAFB*, *FOXA2*, *HNF1A* and *PDX1*, and often mapped adjacent to islet-specific

open chromatin domains[70,209]. Physiologically, these islet-specific lncRNAs tended to be upregulated during *in vivo* and *in vitro* differentiation of pancreatic endocrine cells, with some also up-regulated in response to high glucose, suggesting that they could be involved in adaptive islet-cell responses. Preliminary functional studies into the modes-of-action of these islet-active transcripts demonstrated that inhibition of one of them led to selective reduction of *GLIS3* expression levels[209]. *GLIS3* encodes an important islet transcription factor mutated in a monogenic form of diabetes (OMIM #610199) and implicated in the pathogenesis of both T1D and T2D by GWAS[56,225].

These data demonstrate that islet lncRNAs have regulatory functions, and suggest their upregulation during hyperglycaemia and pancreas development might be causal rather than reactive[209].

4.1.3 Islet transcriptome analysis can uncover genes dysregulated in T2D

RNA-seq data can also be used to compare the islet transcriptomes of normoglycaemic and T2D donors. Dysregulated genes or gene networks identified through such analyses could provide novel biomarkers for islet dysfunction, or drug targets to combat T2D. However, since T2D donors are generally not considered for the transplantation programmes which are the primary source of human islets, the availability of T2D islets is very limited.

Nonetheless, in a recent study Mahdi *et al.* demonstrated that even at low numbers, comparison of islets from T2D and normoglycaemic donors can lead to pertinent findings[213]. The study authors used microarray analysis to profile the islet transcriptome of 10 cadaveric donors with T2D and 38 without. Utilising co-expression network analysis on these expression data, a cluster of genes which showed higher expression in diabetic islets was identified. Furthermore, expression of this gene cluster was correlated with increased HbA1c and reduced glucose-stimulated insulin secretion. One gene in the cluster, *SFRP4*, encoding

secreted frizzled-related protein 4, was of particular interest – it showed a high degree of connectivity with the other genes in the cluster and was highly correlated with all three traits (T2D, HbA1c, and glucose-stimulated insulin secretion).

SFRP4 is a protein involved in extracellular regulation of the Wingless (Wnt) pathway, which, through genes such as *TCF7L2*, has already been linked to T2D pathogenesis[64,226]. In follow-up functional studies in human and rodent islets, SFRP4 was indeed shown to influence islet function[213]. Increased levels of SFRP4 resulted in lower Ca^{2+} influx - most likely by affecting the expression of Ca^{2+} channels – with a subsequent reduction of insulin exocytosis. Since SFRP4 was found to be secreted from islets, it was also evaluated as a potential circulating biomarker for islet dysfunction[213]. These studies showed significant elevation of SFRP4 levels in the serum of T2D patients as compared to normoglycaemic controls. Moreover, elevated levels of SFRP4 were correlated with future development of T2D in individuals without a current clinical diagnosis. However, these findings do not necessarily indicate a causal role for SFRP4 in the pathogenesis of T2D. In fact, its responsiveness to interleukins indicates that the observed expression changes are likely to be secondary to inflammatory stimuli. Nonetheless, while larger validation studies are needed to establish the true potential of SFRP4 as a clinical biomarker, this example clearly demonstrates how human islet transcriptome analysis can establish new hypotheses with potential clinical impact.

This introduction has described many of the applications for which islet RNA-seq data can be used. The main aim of the work described in the rest of this chapter is to perform the largest study to-date of the effects of genetic variation on islet gene expression through ASE and eQTL analyses.

4.2 Methods

4.2.1 Human pancreatic islets

Human islets were obtained (with research consent) from the Oxford DRWF Human Islet Isolation Facility (n=47) and through existing collaborations with Dr. Patrick MacDonald at the Alberta Diabetes Institute, Edmonton, Canada (n=89) from deceased donors. These 136 islet preparations can be further divided into two groups; samples that were freshly extracted (Oxford n=47; Edmonton n=18) and historical samples from Edmonton which had been stored in liquid nitrogen for up to twenty years (n=71).

Frozen samples were processed in Edmonton before shipment by members of Patrick MacDonald's lab (James Lyon, Dr. Jocelyn Manning Fox). First, the samples were almost thawed in a 37˚C water bath before transferral to 4˚C ice slush and centrifugation at 1000 rpm for 1 minute. The samples were then left on ice for 29 minutes, before a final 20 minute thawing step at room temperature during which 10 mL Medium M199 (containing 2.2g/L Sodium bicarbonate and L-glutamine, and supplemented with 10%FBS, 5mM HEPES, 50 U/mL Penicillin, and 50 µg/mL streptomycin) was added. Once thawed, the samples were washed twice with CMRL medium and cultured overnight at 37˚C. They were then hand-picked and frozen in 1ml Trizol by members of the MacDonald lab before shipping to Oxford. Functional studies performed in the MacDonald lab confirmed the thawed islets had retained their glucose-stimulated insulin secretory capacity even after twenty years in liquid nitrogen.In Oxford, RNA and DNA was extracted from all samples as described in section 2.1 and checked for integrity as described in sections 2.2 and 2.3.

4.2.2 Genotyping and imputation

In total, 136 samples were genotyped on the Illumina Omni 2.5 + exome genotyping array. This array contains ~ 2.5 million variants with MAF > 0.025 in 1000 Genomes populations, as well as ~240,000 low-frequency coding variants. It was chosen because the probe density provides an excellent scaffold for imputation, whilst the low-frequency coding variants would allow ASE studies even if imputation of these markers failed. Samples were prepared according to the Illumina Infinium protocol and run on the Illumina iScan platform at the Oxford Genomics Centre (Wellcome Trust Centre for Human Genetics,

University of Oxford, Oxford, UK). Genotypes were called by Neil Robertson (Wellcome Trust Centre for Human Genetics) with Illumina GenCall software using the standard Illumina Cluster file and default genotype calling cut-offs. One sample failed genotyping with a call rate < 90% (Figure 4.1A).

The 135 samples successfully genotyped were imputed to increase the number of variants that could be used for QTLs analyses. Variants with a call rate < 99% (Figure 4.1B) and MAF < 0.01, as well as those deviating from Hardy–Weinberg equilibrium (p-value < 0.0001 - Figure 4.1C), were filtered out before imputation – leaving 1,323,351 variants. Haplotypes were inferred from these genotype data using Shapeit[227]. Genotypes were imputed into the phased haplotypes using IMPUTE2[228] with the entire 1000 Genomes Phase 1 v3 release[97] as the reference panel. For the ASE analysis described in section 4.2.9, genotypes were called from the imputed data, whereby all genotype calls with a posterior probability < 0.9 were set to missing.

4.2.3 Population stratification in the genotype data

To determine ethnicity of the samples, an identity-by-descent (IBS) matrix of genotype data of independent variants overlapping between HapMap and the directly genotyped variants in the islet samples was calculated in Plink[229]. The first two principal components (PCs) of this matrix showed clear clustering into populations, and identified nine samples which did not cluster with the HapMap Utah residents with Northern and Western European ancestry (CEU) samples (Figure 4.2A). These samples were excluded from QTL analyses, but retained for ASE studies. Principal component analysis (PCA) of all directly genotyped autosomal variants with MAF > 0.05 of the samples clustering with HapMap CEU individuals using EIGENSTRAT[230] identified two additional outliers (Figure 4.2B), which were also excluded from the QTL analyses. To correct for remaining population stratification in the individuals of European descent, the first two principal components from the data when excluding these two samples were used as covariates in the QTL analyses (Figure 4.2.C).

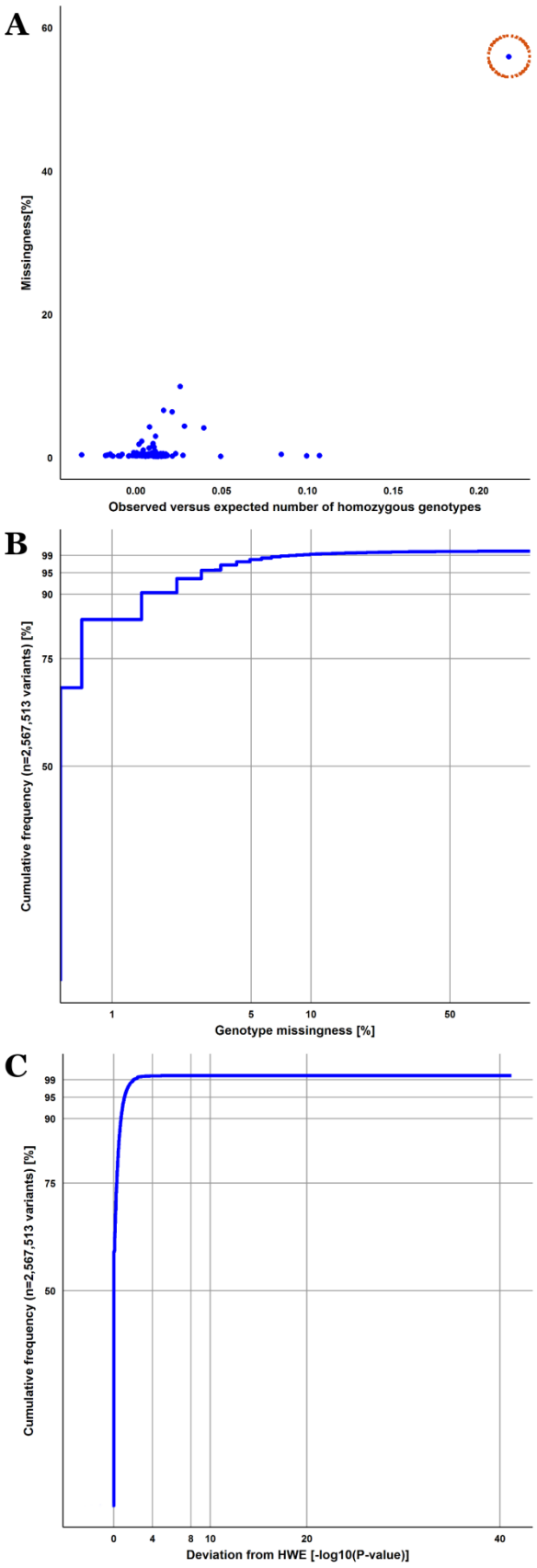


Figure 4.1: Omni 2.5 + exome genotype QC. Panel A shows heterozygosity (x-axis) versus missingness (y-axis) per sample. One sample is a clear outlier (circled in orange) and was removed in subsequent analyses. Panel B and C show the cumulative frequency distribution for the genotype missingness rate (B) and deviation from Hardy-Weinberg equilibrium (C) per variant.

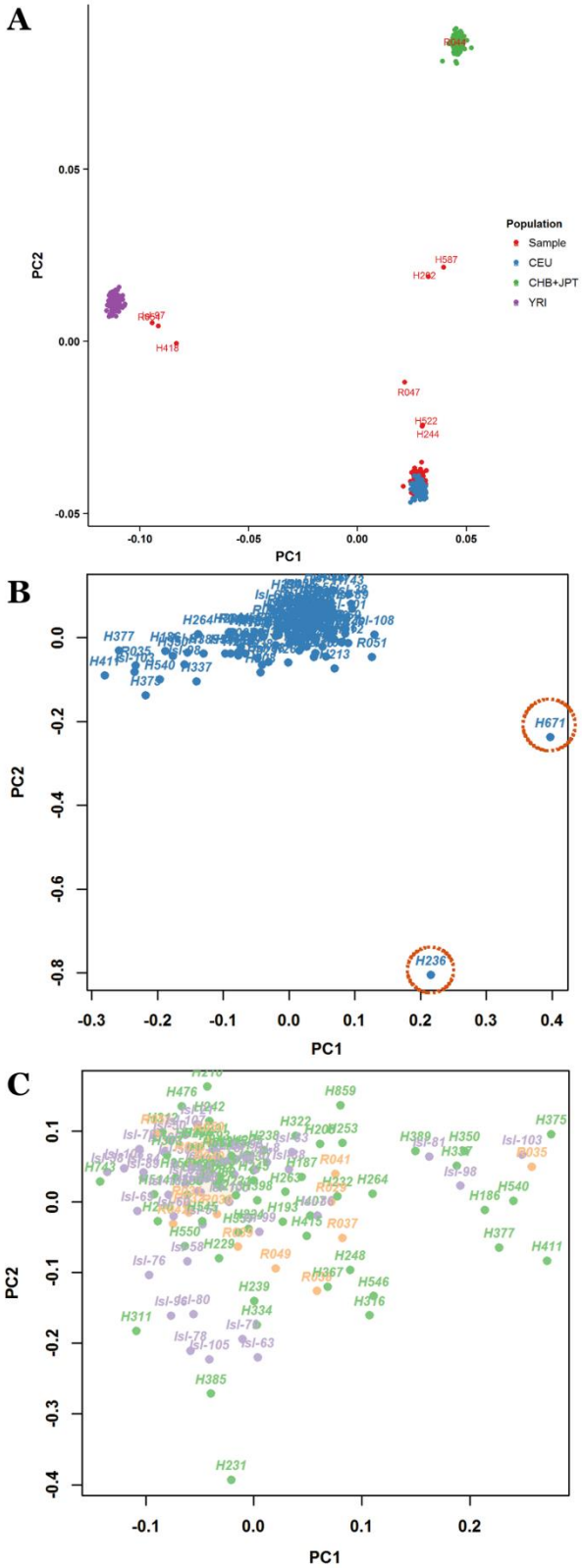


Figure 4.2: Population stratification analyses. Panel A shows the IBS clustering of the islet samples (red) with HapMap samples with European (CEU - blue), African (YRI – purple) and East Asian (CHB+JPT – green) ancestry. The nine labelled islet samples did not cluster with the CEU individuals. Panel B shows another two outliers (circled in orange) in the PCA of the European ancestry islet genotype data. Panel C shows the first two PCs with the two outliers removed, which shows remaining stratification within these samples. Colours denote origin of sample; fresh samples from Oxford (purple) and Edmonton (orange), and legacy Edmonton samples stored in liquid nitrogen (green).

4.2.4 RNA sequencing

All samples used for library preparation had a total RNA yield in excess of 1 μg and a quality score (RIN) > 5. Libraries were prepared at the High-Throughput Genomics group (Wellcome Trust Centre for Human Genetics, University of Oxford, Oxford, UK) using NEBNext ultra directional RNA library prep kit for Illumina with custom 8bp indexes[231]. Libraries were multiplexed (3 samples per lane), clustered using TruSeq PE Cluster Kit v3, and sequenced using Illumina TruSeq v3 chemistry at 100 basepair reads on the Illumina HiSeq2000 platform. Samples were mapped with TopHat2[232] on default settings using GENCODE v17[233] as transcriptome reference and the GRCh37 assembly as genome reference. The sequenced data was required to contain at least 10M mapped and properly paired reads after filtering for mapping quality.

4.2.5 mRNA quantification

Transcripts were quantified using Flux Capacitor[234] with GENCODE v17[233] as the reference for transcript structures. Flux Capacitor maps reads to splice graphs produced from the reference annotation. The proportion of reads attributable to each isoform at exons shared between different isoforms of one gene is estimated based on the density function of reads along the isoform. Only read pairs completely included within the exon boundaries of a transcript and paired in the expected orientation were considered. Expression levels of overlapping transcripts (e.g. different isoforms of the same gene) were estimated by normalising the read counts to sequence depth and exonic transcript size (reads per kilobase per million [RPKM] mapped reads). Gene expression was quantified for protein-coding and lncRNA genes by summating the RPKMs for the individual transcripts per gene.

4.2.6 Sample swap and outlier detection

Genotypes were called from the mRNA-seq data – described in section 4.2.9 – and compared to the array genotypes to detect sample swaps. This identified one sample without

a corresponding DNA sample and one duplicate sample, which were excluded from subsequent analyses.

To further estimate sample quality, Spearman rank correlations between all samples were calculated using transcript RPKMs. Only transcripts present in > 90% of samples were included (n=18,519). From these data, the median correlation of one sample against all the other samples was determined. In parallel, PCA was performed on transcript expression values to further establish consistency between samples. From these analyses, three additional samples were removed due to low correlation (spearman *rho* < 0.8) with other samples (Figure 4.3A). One of these samples (Isl-36) was also an outlier in PCA, which identified clustering in line with the three different sample sources (Oxford, Edmonton freshly isolated and Edmonton legacy frozen samples; Figure 4.3B).

4.2.7 Normalization of quantifications

Expression quantification from RNA-seq data can be confounded by technical biases – including GC content – which reduces the power of most downstream analyses. Correcting for such confounders requires different approaches depending on the study question. When comparing the effects of local genetic variation between individuals on the expression of a single gene (*cis*-QTLs), it is possible to improve eQTL discovery by correcting out synthetic covariates detected in the expression data by PCA or related methods such as PEER[235]. PEER normalisation was therefore applied to gene quantification of the 119 islet preparations from individuals of European descent. PEER was run on the 18,519 consistently expressed genes in those 119 samples, correcting for ten hidden covariates with the mean expression of each gene added as an additional covariate. The mean was then added back to the residuals from this regression model. PEER corrected samples showed a marked increase in inter-sample correlation (Figure 4.4A) and on PCA clustering based on samples origin (Oxford, Edmonton fresh, Edmonton frozen) had completely disappeared (Figure 4.4B). The values were standard normalised before QTL analyses to prevent false-positive associations due to outliers.

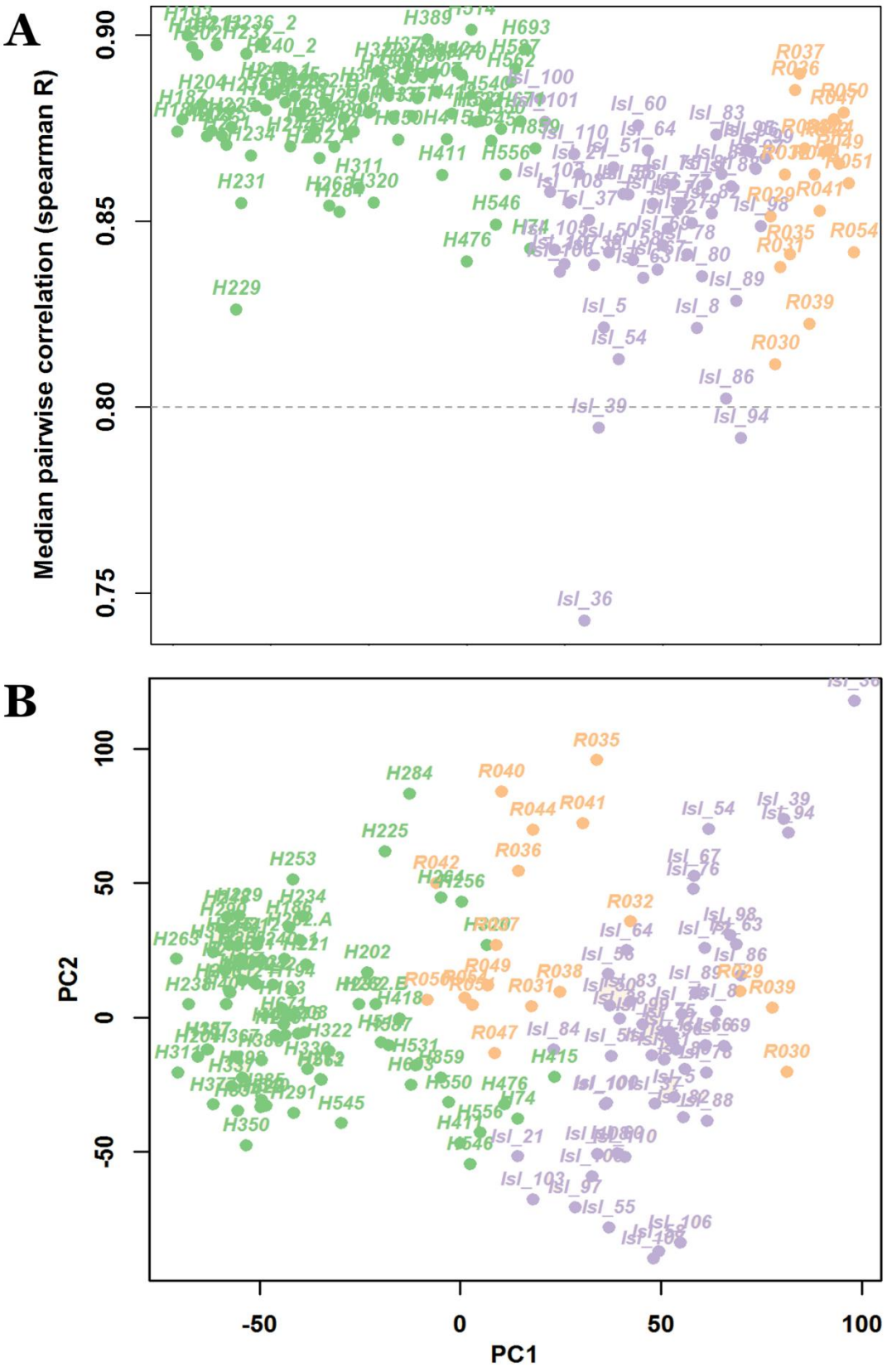


Figure 4.3: Gene expression QC. Panel A shows median spearman correlation between the labelled sample and all other samples studied. Samples with a spearman rho < 0.8 (dashed line) were excluded. Panel B displays the first two PCs from the gene expression data for the 133 samples that have been successfully genotyped. Aside from Isl-36 it shows no clear outlier, but identifies clustering by sample origin.

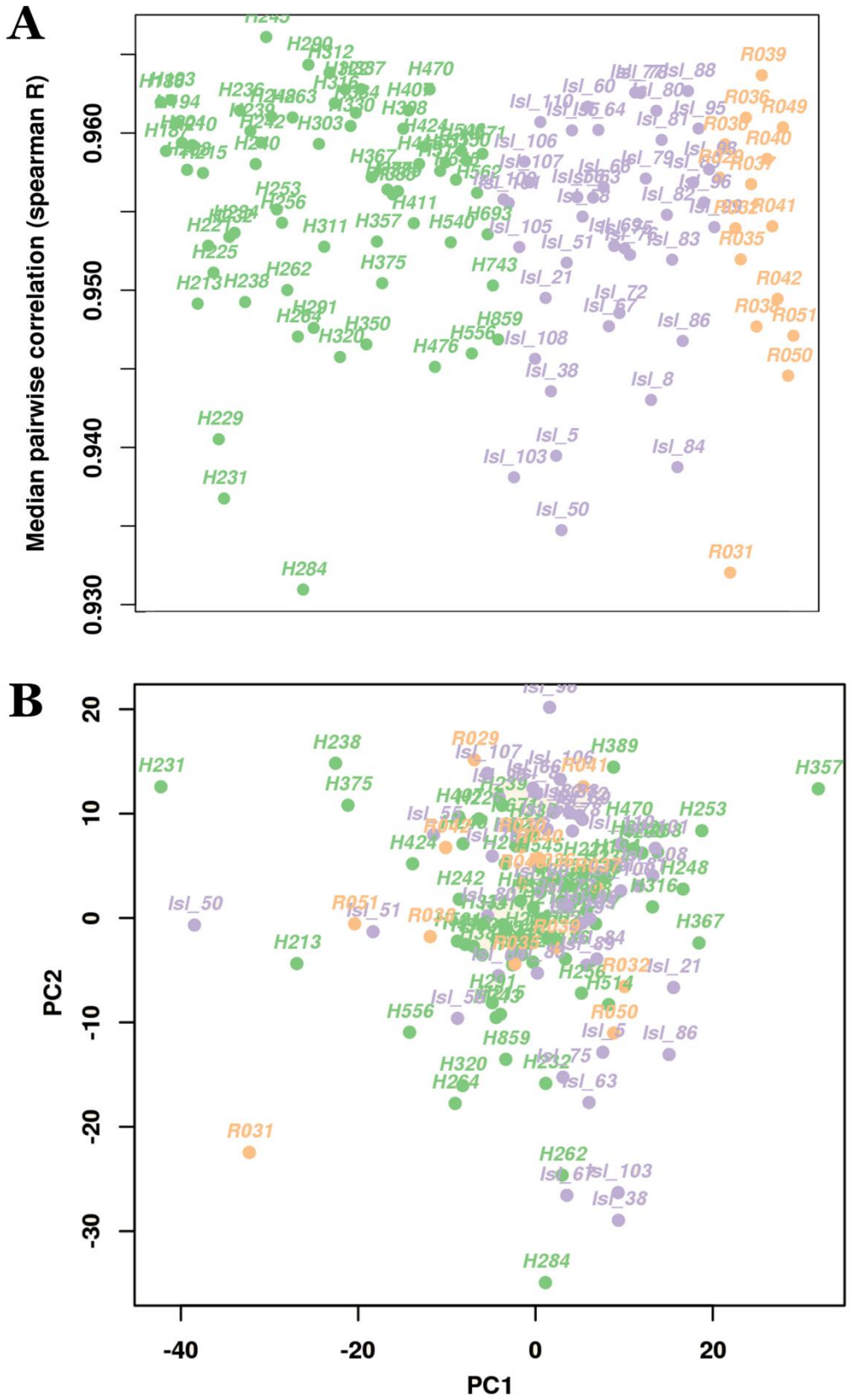


Figure 4.4: Sample correlation after PEER normalisation. Panel A shows median spearman correlation between the labelled sample and all other samples studied. The median correlation increased from between 0.8-0.9 to around 0.95 after normalisation. Panel B displays the first two PCs from the PEER normalised gene expression data for the 121 samples. The clustering by sample origin has disappeared and most samples cluster homogeneously.

4.2.8 QTL analysis

Expression QTL analysis was performed using SNPTEST v2.5ß[236], including only variants with MAF > 0.05 located within 1Mb flanking regions of the transcription start site (TSS) for the tested gene. The variants were tested under an additive model using a missing data likelihood score test to account for uncertainty in the imputation. After testing, variants with an information measure < 0.4 were excluded from further analyses. To establish the *P*-value threshold, the expression values of all genes were randomly permuted 20 times and tested for association, with the best *P*-value per permutation retained. The global 5%FDR *P*-value threshold was derived from these ~360,000 best *P*-values and used as the significance threshold.

4.2.9 Allele-specific expression analysis

For the ASE analysis, all single nucleotide variant sites overlapping protein-coding genes and lncRNAs were extracted from the imputation reference panel. Alleles at these sites were called from the RNA-seq data using the samtools mpileup function including only reads with a base quality and mapping quality > 10. Since ASE analysis is done on a per sample basis, those samples excluded from the eQTL analysis as population outliers (n=11) could be included. For each of the 130 samples, only sites with a heterozygous genotype from imputation with both alleles observed in the RNA-seq data (to verify that the genotype was truly heterozygote) and with coverage > 16 reads were included in the analysis.

Next, for each sample an expected average reference-alternative ratio was calculated across the sites to correct for mapping bias and differences in GC content between individuals. Sites with > 30 reads were downsampled to exactly 30 reads to prevent them exerting a disproportional effect on the expected ratios. As evident in Figure 4.5, there was clear evidence of nucleotide bias favouring G and C. When testing all sites per individual for the presence of allele-specific expression using a binomial test, the observed reference-alternative ratio was therefore used as the expected ratio.

4.2.10 Gene set enrichment analysis and functional annotation

Enrichment of gene set annotations were performed using the online Database for Annotation, Visualization and Integrated Discovery (DAVID) v6.7[237] in the context of the whole human genome. Default annotation categories were used and enrichment results sorted by significance of Benjamini corrected p-values[238].

The Variant Effect Predictor[145] (VEP version 73; http://www.ensembl.org/info/docs/variation/vep/vep_script.html) tool from Ensembl was used to generate annotations for the allele-specific expression variants as described in section 2.7.2. The annotation for the transcript with the highest predicted functional impact was selected for each variant.

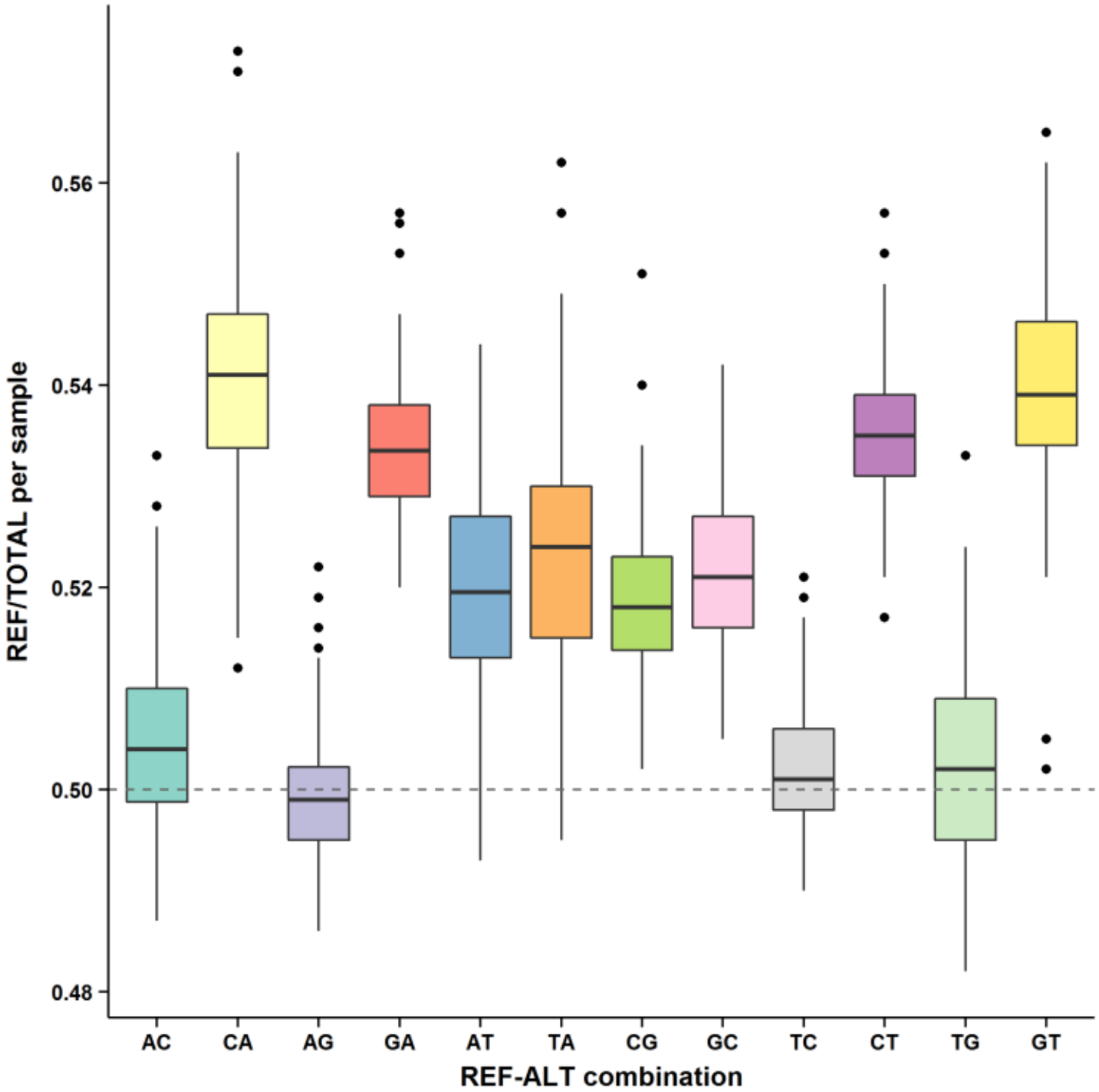


Figure 4.5: Ratios of reference allele over all alleles per site. Boxplots of the ratios per individual show clear evidence of GC bias and a good reproducibility of ratios across samples.

4.3 Results

4.3.1 Expression profile of human pancreatic islets

After quality control, high quality RNA-sequencing data was available for 130 human islet preparations. The median number of raw reads obtained per sample was 124 million, with a median of 58 million reads (47%) mapping to exonic features in GENCODE version 17[233] (after all filters were applied). All samples passed the minimum threshold of 10 million well-mapped reads, and the data is relatively uniformly distributed with only a few outliers (Figure 4.6A-B).

In these 130 samples, a median of 23,430 protein-coding and lncRNAs genes were detected per sample, of which 18,519 are found in > 90% of the samples (Figure 4.6C). The consistently expressed genes are present as 137,290 alternative isoforms with evidence of expression (RPKM > 0.1) in at least one sample. In line with previously observations in a large RNA-seq dataset on LCLs generated from 1000 Genomes individuals[207], around 50% of the known isoforms of a gene are predicted to be expressed at any one time (Figure 4.7A). The major isoform accounts for at least 30% of the total quantified gene expression even with 60 known isoforms (Figure 4.7B), which is consistent with the notion that not all isoforms are equally biologically important and indicates that the isoform quantifications are likely to be of good quality.

The majority of the genes found to be expressed in the islet data (83%) are protein-coding, with a smaller percentage (17%) being lncRNAs. The most abundant genes show significant enrichment for functional annotations relating to the pancreas and its high metabolic activity (Figure 4.8A). *INS* was by far the most abundantly transcribed gene, comprising on average 9.1% of the total transcriptome per sample (Figure 4.8B). This was 10-fold higher than the next most abundant gene, *GCG*, which makes up around 0.9%. Not surprisingly given islet isolation methods and differences in sample origin, there is evidence

of substantial (exocrine) contamination. Compared to Nica *et al.* data, there is a lower contribution of *INS* expression to the total transcriptome (9.1% versus 21.0%)[108], and exocrine genes *PRSS1* and *CTRB1* (encoding respectively trypsin 1 and chymotrypsinogen B1), together with the inflammatory gene *IL8* (encoding interleukin-8), are found amongst the top 10 most abundant genes (Figure 4.8B). However, the high expression of *INS* and *GCG* indicates that there is a clear endocrine signature in the islet data – an indication further confirmed by overlap of five of the ten most abundant transcripts in the RNA-seq data (*INS*, *GCG*, *TTR*, *REG1A*, and *PRSS1* – Figure 4.8B) with those most abundant in the profiles of published RNA-seq data on a smaller number of islets[108]. There also is a strong correlation between the median islet profile and Nica *et al.*[108] beta-cell profile (Spearman *rho* = 0.83), the raw data of which was reprocessed according to the methods used in this chapter. Overall, this confirms that the data is generally of endocrine origin and well suited for eQTL and ASE discovery.

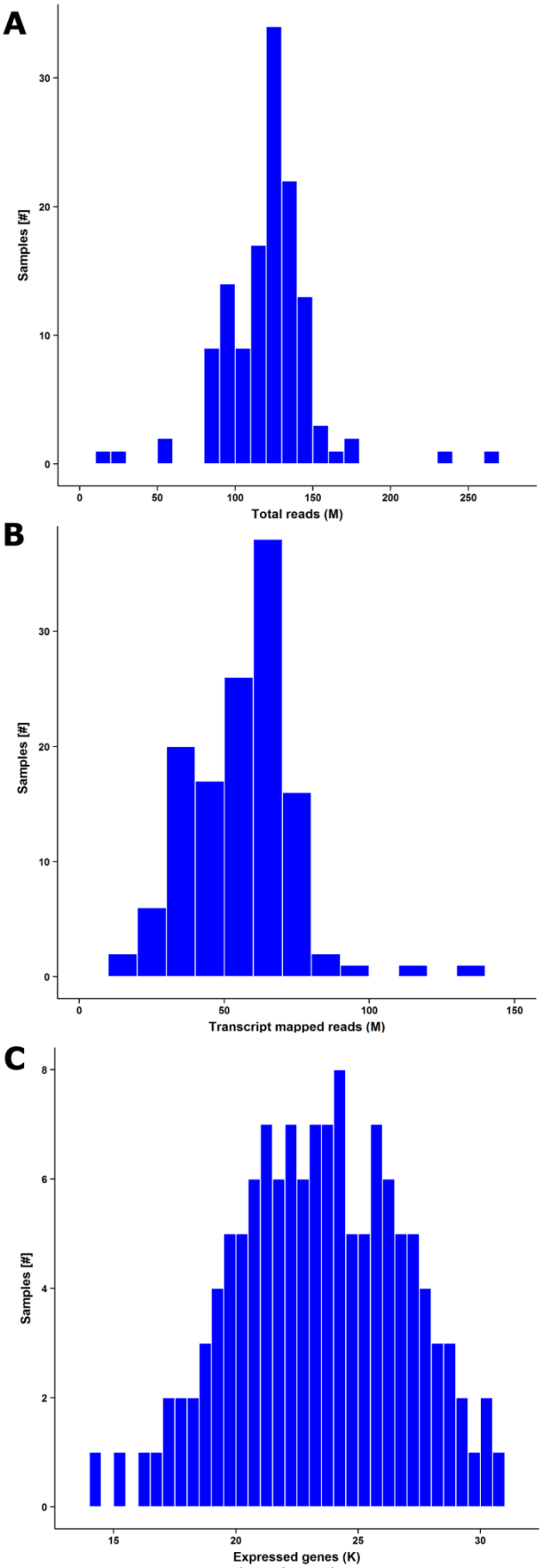


Figure 4.6: Read and gene count distribution. Histograms of the per sample total raw reads counts (A), number of reads mapping to GENCODE version 17 transcripts after filtering (B), and expressed gene counts (C).

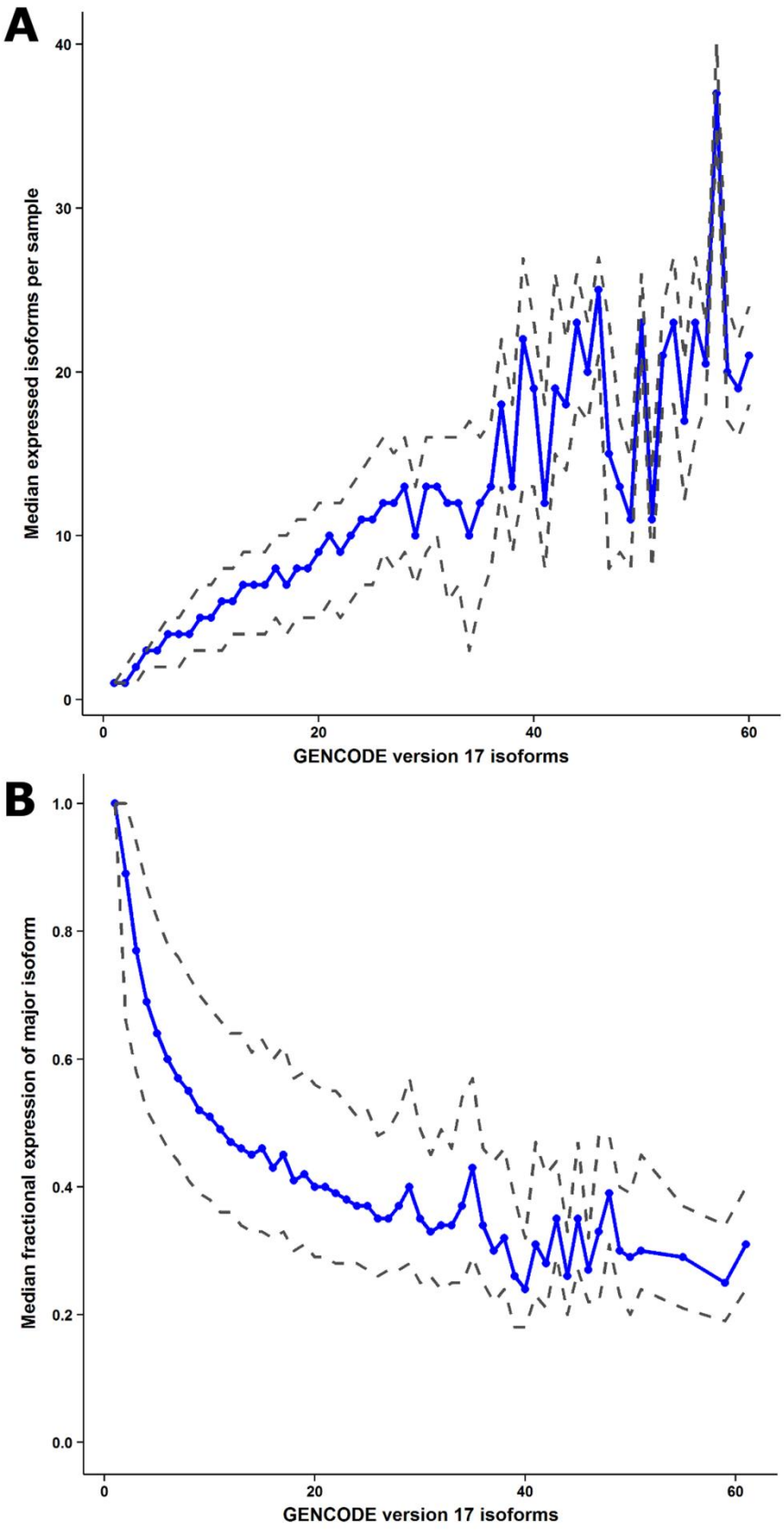


Figure 4.7: Isoform quantification metric. The number of isoforms expressed (RPKM > 0.1) (A), and contribution of the major isoform to the total gene expression (B) per number of isoforms annotated in GENCODE version 17. The blue points represent the median with the dashed lines indicating the 1st and 3rd quartiles.

A

Term	Benjamini P-value	Count	Fold Enrichment
Translational elongation	3.60E-23	22	32.03
Ribosome	4.80E-22	18	49.40
Ribosomal protein	7.35E-18	20	21.32
Cytosolic ribosome	3.16E-17	17	29.81
Protein biosynthesis	1.35E-16	19	20.25
Structural constituent of ribosome	1.86E-16	20	17.97
Ribosome	3.65E-16	18	19.13
Ribosome	1.70E-15	21	13.87
Ribonucleoprotein	6.99E-15	20	14.36
Ribosomal subunit	2.07E-14	17	18.86
Ribonucleoprotein complex	2.58E-14	27	7.45
Translation	1.17E-13	23	10.22
Cytosolic part	2.27E-13	17	15.88
Pancreas	5.57E-13	9	100.18
Acetylation	5.85E-13	45	3.42
Structural molecule activity	1.18E-10	25	5.95
Cytosol	6.39E-10	34	3.63
Signal	5.75E-08	42	2.59
Cytosolic large ribosomal subunit	1.76E-07	8	29.90
Cytosolic small ribosomal subunit	2.25E-07	8	28.40

B

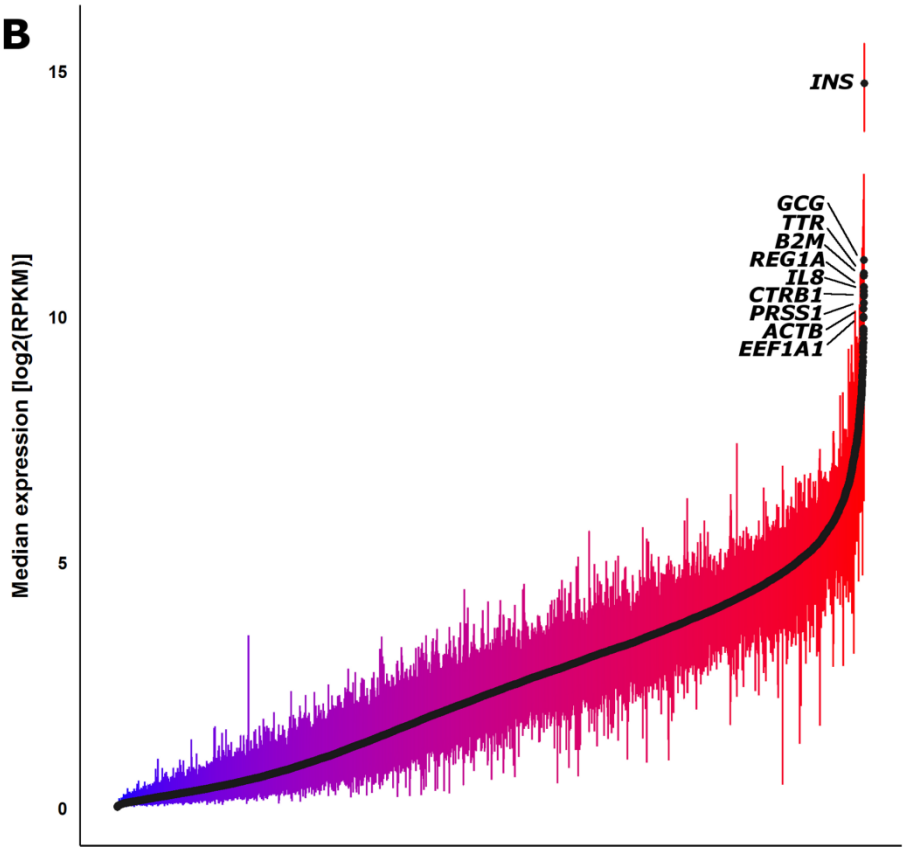


Figure 4.8: Characteristics of islet expressed genes. Panel A shows the top 20 most enriched functional annotations (using DAVID[237]) of the 100 most abundant genes in the islet expression profile with terms relating to protein synthesis (purple) and pancreas/cell signalling (green) highlighted . Panel B shows the distribution of the median expression (black dots) with interquartile range (lines coloured by expression rank ranging from blue [lowest expressed] to red [highest expressed]) per gene.

4.3.2 Effects of genetic variation on islet gene expression levels

eQTL analysis of the 18,519 consistently expressed protein-coding genes and lncRNAs uncovered 1,636 genes with significant (*P*-value < $2.7x10^{-5}$, corresponding to a global false discovery rate [FDR] of 5%) association between genotype and gene expression levels. While the islet preparations were of varying purity, PEER normalisation corrected for the majority of purity effects. Before normalisation, there was strong significant correlation between a measure of islet purity derived from the sequencing data (the sum of major endocrine gene expression levels [*INS* + *GCG* + *SST*] divided by the sum of both major endo- and exocrine gene expression levels [*INS* + *GCG* + *SST* + *PNLIP* + *CPB1* + *PRSS3*]) and the first (Spearman *rho* = 0.67, *P*-value < $2.2x10^{-16}$) and second (Spearman *rho* = -0.37, *P*-value = $3.1x10^{-5}$) of the twenty principal components derived from the per sample gene expression profiles. After PEER normalisation with ten hidden covariates, only a small significant correlation of the purity measure and the second principal component remained (Spearman *rho* = 0.26, *P*-value = $3.9x10^{-3}$). Most of the significant eQTL variants were located within 150-200kb of the TSS (Figure 4.9), which is consistent with previous eQTL analyses[205,239,240] and ENCODE data showing that the majority of chromatin accessible by DNAse I is located within 250kb of the TSS[112].

There is a statistically significant enrichment of lncRNAs (1.48 fold enrichment over protein-coding genes, Fisher's *P*-value $1.1x10^{-9}$) amongst the 1,636 eQTL genes– 372 lncRNAs and 1,264 protein-coding genes with an eQTL out of 3,173 and 15,346 expressed genes respectively. This feature is even more striking when taking into account greater eQTL visibility for highly expressed genes: lncRNAs are significantly underrepresented amongst the 10% (n=1852) most abundant islet genes (n=40 lncRNAs – Fisher's *P*-value < $.2.2x10^{-16}$). Some of the lncRNAs with significant eQTLs are located near genes with known functions in pancreatic islets, including *SLC2A1-AS1* and *GLIS3-AS1*, located near the facilitative glucose transporter GLUT1 and the transcription factor Gli-similar 3 (GLIS3) respectively (Figure

4.10A-B). *GLUT1* is the most highly expressed glucose transporter in human islets and beta-cells[30], whereas GLIS3 plays a critical role in pancreas development[241,242]. The gene encoding *GLIS3* is located in an association locus for both T1D and T2D[56,225], and mutation in *GLIS3* give rise to a syndrome characterised by, amongst other symptoms, neonatal diabetes[241]. Given the functionality of other lncRNAs in islets[209], these eQTL lncRNAs might therefore be involved in regulation of the nearby islet genes. The overlap between T2D genetics and islet eQTLs will be further explored in chapter six.

Barring the generic term "polymorphism" – indicating that there is at least one alternative protein not leading to disease caused by sequence variation – there was no significant enrichment (Benjamini *P*-value < 0.05) of the default functional annotations in DAVID[237] amongst the significant eQTL genes. To investigate the potential function of the variation underlying the eQTLs, functional elements defined from ChIP-seq data for chromatin marks in human islets[143] was used. These data were generated as part of a collaborative effort with Prof. Jorge Ferrer's lab (Barcelona, Spain and Imperial College London, London, UK), involving several researchers from Oxford (Dr. Kyle Gaulton, Dr. Anna Gloyn, Prof. Mark McCarthy and – in a minor capacity – me). Functional elements were based on four markers of active chromatin (mono- and trimethylation of histone H3 lysine 4 [H3K4me1 and H3K4me3], trimethylation of histone H3 lysine 36 [H3K36me3], and acetylation of histone H3 lysine 27 [H3K27ac]) and two of silenced/repressed chromatin (trimethylation of histone H3 lysine 27 [H3K27me3] and the transcription factor CCCTC-binding factor [CTCF]) all derived from the same islet preparation. The overlap between islet functional elements and the most significant variants was determined. This showed that 39% (n=638) of the 1,636 index variants overlapped an islet annotation, divided amongst marks of active chromatin (active promoters n=83; enhancer elements n=205; transcribed sequence n=172) and inactive chromatin (poised promoters n=16; CTCF-bound chromatin n=13; repressed chromatin n=150). However, due to LD and power limitations, it is not be possible to determine whether the index variant is indeed the variant giving rise to the eQTL.

The overlap between islet annotations is therefore likely to be an underestimation of the true overlap.

By comparing the islet annotations to those derived from ten publicly available cell-line data using identical methods (chromatin annotations generated by Dr. K.J. Gaulton, [Wellcome Trust Centre for Human Genetics, University of Oxford, Oxford, UK]), it is possible to look at enrichment of different categories of islet annotation amongst the eQTLs. For the significant eQTLs (again using the index variant as proxy for the signal), there is a significant reduction compared to the same annotation in the ten cell-lines in overlap of islet repressive elements (1.19 fold reduction, Fisher's *P*-value 0.03), whereas there is no significant enrichment for enhancer regions (Fisher's *P*-value 0.47) or active promoter regions (Fisher's *P*-value 0.86).

The size of the study described in this chapter is primarily powered to detect strong eQTL signals. Such signals tend to be located near the TSS – in line with the distribution of the islet data described in Figure 4.9 – and shared between tissues[205,206,239,243]. Using publicly available data from the MuTHER (Multiple Tissue Human Expression Resource) project[206], it was possible to look at tissue sharing between the islet eQTLs and those of subcutaneous adipose tissue, skin and LCLs. These data were generated using gene arrays and had genotype densities based on HapMap phase 2. As such, about half (54% - n=891) of the significant islet eQTL genes and 11% (n = 6,475) of all significant variants could be used in the comparison. Since the eQTL analysis was performed in variants with a MAF > 5%, these HapMap variants should still provide adequate proxies for the majority of the signals identified. Using *P*-values for the variants overlapping between all datasets, it was possible to estimate the proportion of true null hypotheses amongst all tests (π_0). From this the proportion of significant replicating signals across tissues could be determined ($\pi_1 = 1 - \pi_0$ – Figure 4.11)[244]. These analyses indeed showed substantial sharing of detected islet eQTLs between tissues (π_1 = 0.58-0.69). Since a large proportion of the active islet chromatin used to determine enrichment will not be islet-specific, the shared nature of islet eQTLs across

tissues is likely to explain why active islet chromatin marks are not more enriched across the significant islet eQTLs compared to active chromatin marks in other tissues.

Given the study size, many more true islet signals not currently reaching study-wide significance should be present in the data. Such weaker signals are more likely to be tissue-specific[205,206,239,243], and, as such, should show a higher enrichment of islet-active chromatin. The analysis was therefore extended to include all best variants per gene, regardless of their significance. This does identify significant enrichment for active promoters (1.15 fold increase, Fisher's *P*-value 8.3×10^{-3}) and enhancer elements (1.06 fold increase, Fisher's *P*-value 8.0×10^{-3}) compared to the same annotation in the other ten cell-lines, accompanied by an even stronger reduction in the overlap of repressive elements (1.40 fold reduction, Fisher's *P*-value $< 2.2 \times 10^{-16}$) than seen in the significant eQTLs only. These findings suggest many more islet eQTL remain to be discovered by increasing sample size.

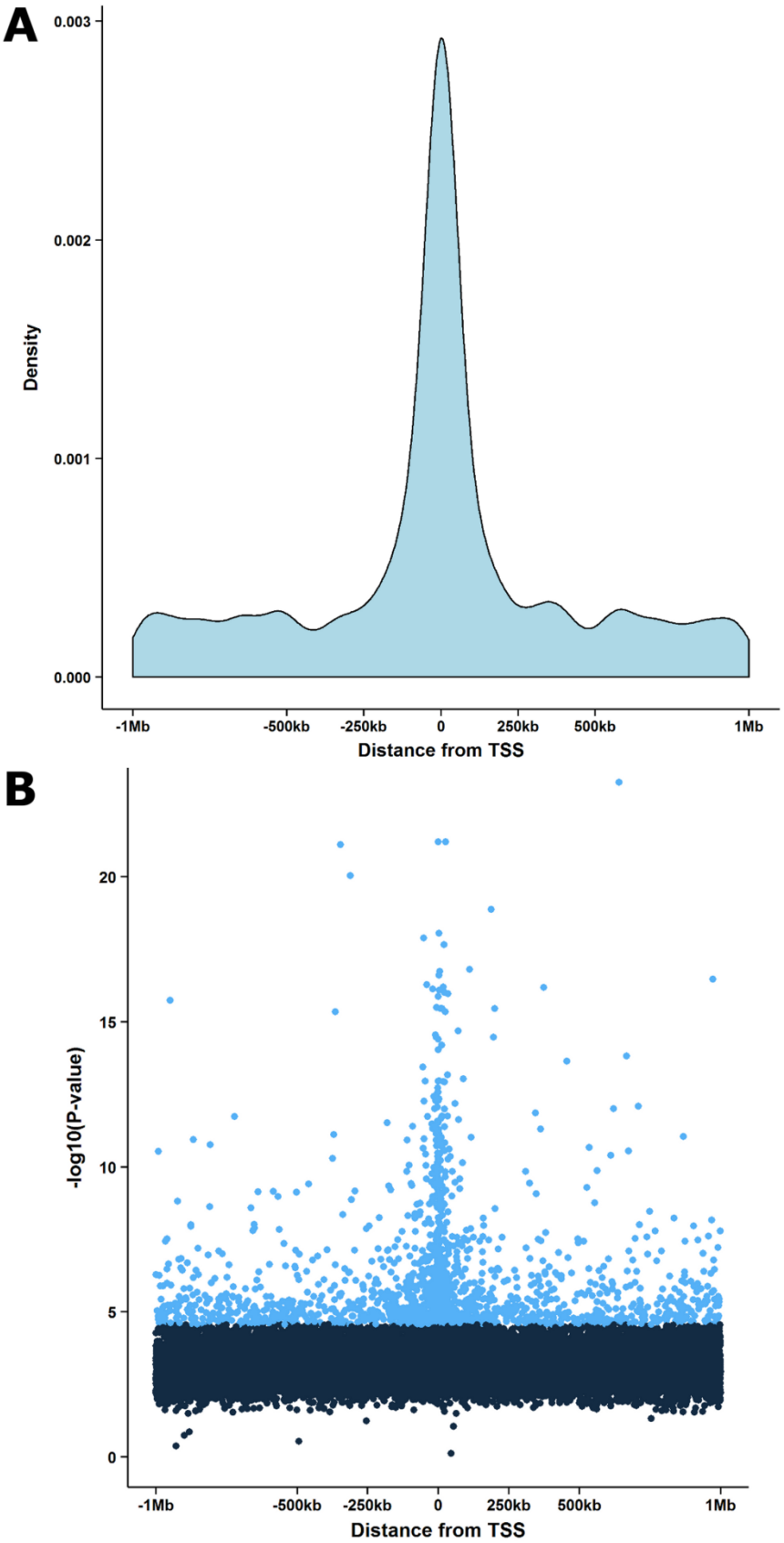


Figure 4.9: Distribution of best eQTL variants per gene. The top panel (A) shows the data density of best eQTL variants per gene for significant genes relative to the distance to the TSS. The bottom panel (B) shows the distance from the TSS for the gene on the x-axis and the strength of association with expression levels on the y-axis with each dot denoting a variant. Colours indicate whether variants achieved global significance (*P*-value < 2.7×10^{-5}, light blue) or not (dark blue).

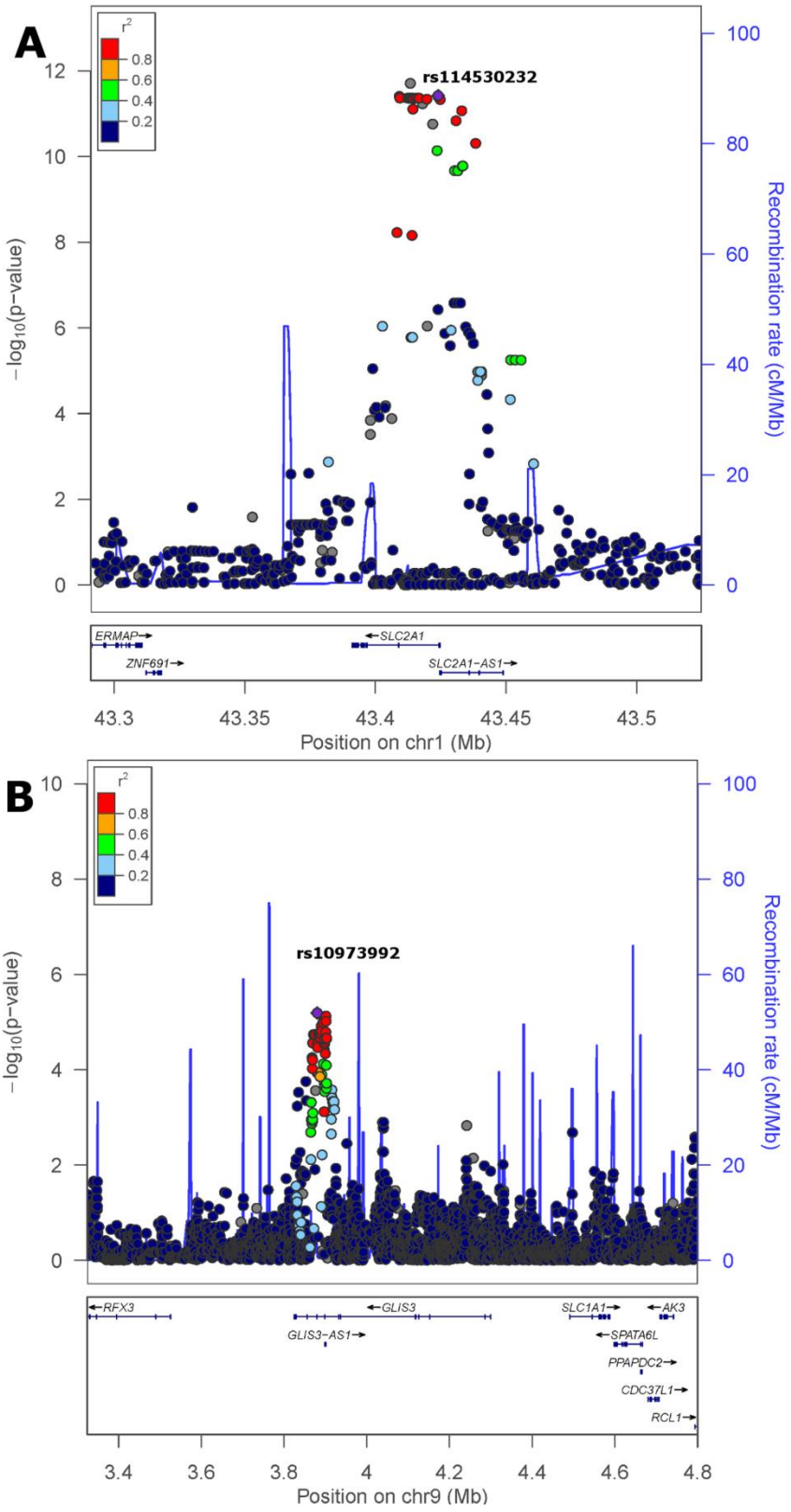


Figure 4.10: Significant islet eQTLs for lncRNAs near genes with known islet functions. Local plots showing the strength of association between variants and expression levels of *SLC2A1-AS1* (A) and *GLIS3-AS1* (B). The left y-axis denotes the strength of the association, while the right x-axis shows the recombination rate. Each point denotes a variant and the colours signify the linkage disequilibrium between the variant and the index variant (in purple). Since the most significant variant for *SLC2A1-AS1* was not present in the data used for plotting, LD is calculated for the next most significant variant.

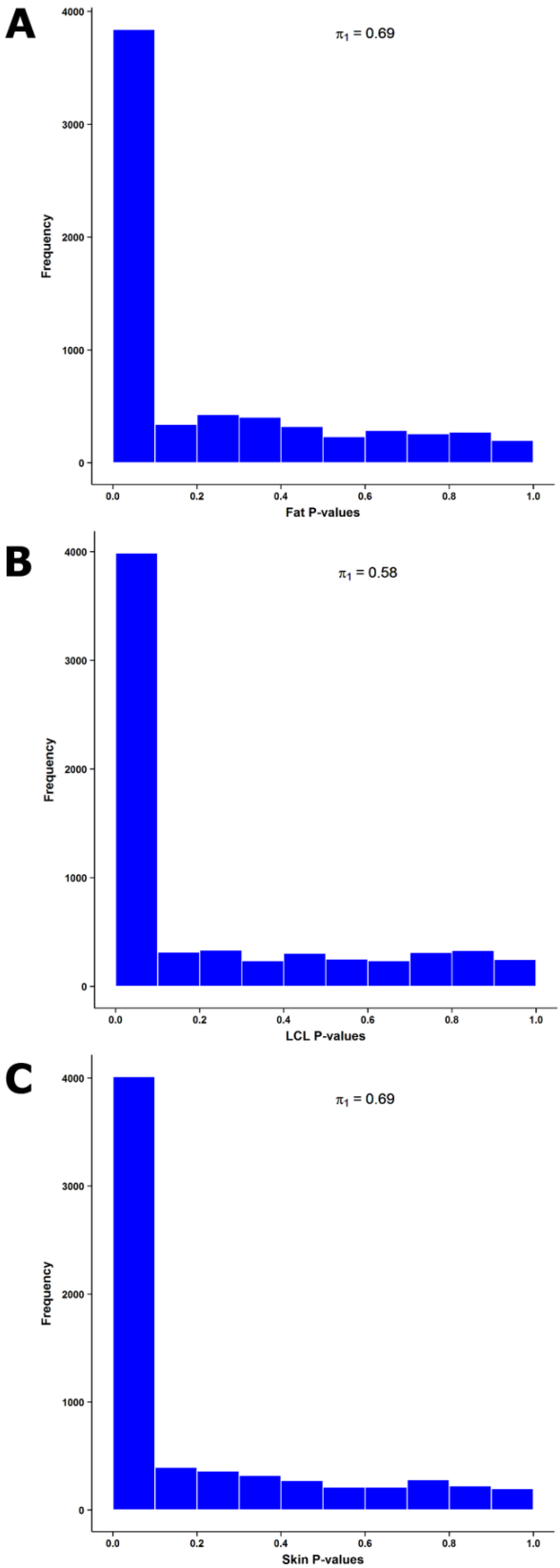


Figure 4.11: *P*-value distribution of *cis*-eQTLs identified in human islets across other tissues. Histograms of the eQTL *P*-values in subcutaneous adipose tissue (A), lymphoblastoid cell lines (B) and skin (C) part of the MuTHER study.

4.3.3 Effects of genetic variation on allelic expression in islets

In addition to eQTL analyses, RNA-seq allows identification and quantification of regulatory variation through consideration of haplotype specific transcript expression. Even when an eQTL cannot be detected, for example owing to low allele frequency, allele-specific expression (ASE) in heterozygous individuals can still uncover the regulatory effects of genetic variation. For each individual, all heterozygous variants overlapping protein-coding and lncRNAs transcripts with good coverage (> 16 reads) in the RNA-seq data (mean coverage of included sites 76 reads per sample-Figure 4.12A-B) were tested for ASE (median 6,456 sites per sample – Figure 4.12C). On average, 8.8% of the tested sites per individual show significant (*P*-value < 0.005) ASE (median 501 sites – Figure 4.12D) - corresponding to a median of 556 genes per individual out of a median of 7,126 tested (Figure 4.12E-F). In total, this means that 47.3% (n=8,754) of all tested genes (n=18,519) contain at least one significant ASE variant. As might be expected, there is reasonable correlation between coverage and the fraction of ASE sites per total heterozygous sites per individual (Spearman *rho* 0.51 – Figure 4.12B). No samples show evidence of increased or reduced heterozygosity in the genotype data, suggesting that the very high fraction of ASE sites in several individuals might point to epigenetic effects underlying at least a proportion of the detected variants. Indeed a substantial number of genes (42.3%) with a significant islet eQTL also contain at least one significant ASE site in one individual. However, this does not represent a significant enrichment in overlap between the two, which is perhaps not surprising given that ASE also captures mechanisms such as splicing which do not necessarily impact upon total gene expression levels.

To facilitate comparison of ASE across individuals, sites with coverage >30 reads were downsampled to exactly 30 reads and tested for ASE. Most sites with significant ASE in at least one individual have less than 10 heterozygous samples in the 130, corresponding to a MAF < 0.05, and the strongest ASE effects are also found at the rarer sites (Figure 4.13A). Using the fraction of samples with ASE over total heterozygous samples at all sites covered in

> 2 individuals as a proxy for the allele frequency of the causal variant giving rise to the ASE, the effect is even stronger since nearly all of the sites are significant in less than 10% of heterozygous individuals (Figure 4.13B).The results from this analysis therefore indicate that most of the regulatory variation identified by ASE is rare, which is in line with recent data on ASE from lymphoblastoid cell lines for 462 individuals generated as part of the 1000 Genomes Project[207], and perhaps expected given the numerical abundance of rare variants in the genome.

Nearly half (48.5% – 13,507 variants) of the 27,855 sites with significant ASE in at least one sample are located in 3'-untranslated regions, with the other major annotations being synonymous (16.0% – 4,460 variants), intronic (13.4% – 3,742 variants), and missense variants (12.1% – 3,364 variants). The substantial proportion of intronic variants which exhibit significant ASE suggests that many of the causal variants underlying the ASE lead to alternative splicing, since high coverage of intronic sequence is unusual in RNA-seq data. Around 1.1% of the significant variants are predicted to cause loss-of-function, creating premature stop codons or affecting splicing. One such variant potentially interesting in the islet context is rs10190751, which shows significant ASE in 28 individuals. The variant is located in a splice acceptor site of *CFLAR,* encoding CASP8 and FADD-Like Apoptosis Regulator protein. CFLAR is an inhibitor of the Fas death cell receptor pathway[245], and has been shown to divert the apoptotic pathway triggered by Fas into a signal for proliferation[246,247]. CFLAR expression is reduced in human islets from T2D donors and in islets from normoglycaemic donors exposed to high glucose concentrations, which results in an increase of apoptotic cells[247]. Conversely, upregulation of CFLAR leads to a reduction in apoptosis and increase in proliferation when exposed to high levels of glucose[247]. At low glucose concentration, overexpression of CFLAR by transfection of stimulation with transforming growth factor beta (TGF-β) has no effect on proliferation[247]. However, stimulation of the Fas-CFLAR pathway with low concentrations of interleukin-1β, which is increased in hyperglycaemia, does result in increased proliferation[248]. This effect can be blocked with a small interfering RNA targeting CFLAR[248]. Since not all isoforms of the

protein are thought to have the same effects on the Fas-CFLAR pathway[249], identifying a variant likely to affect splicing might offer a naturally occurring system through which these differences can be studied.

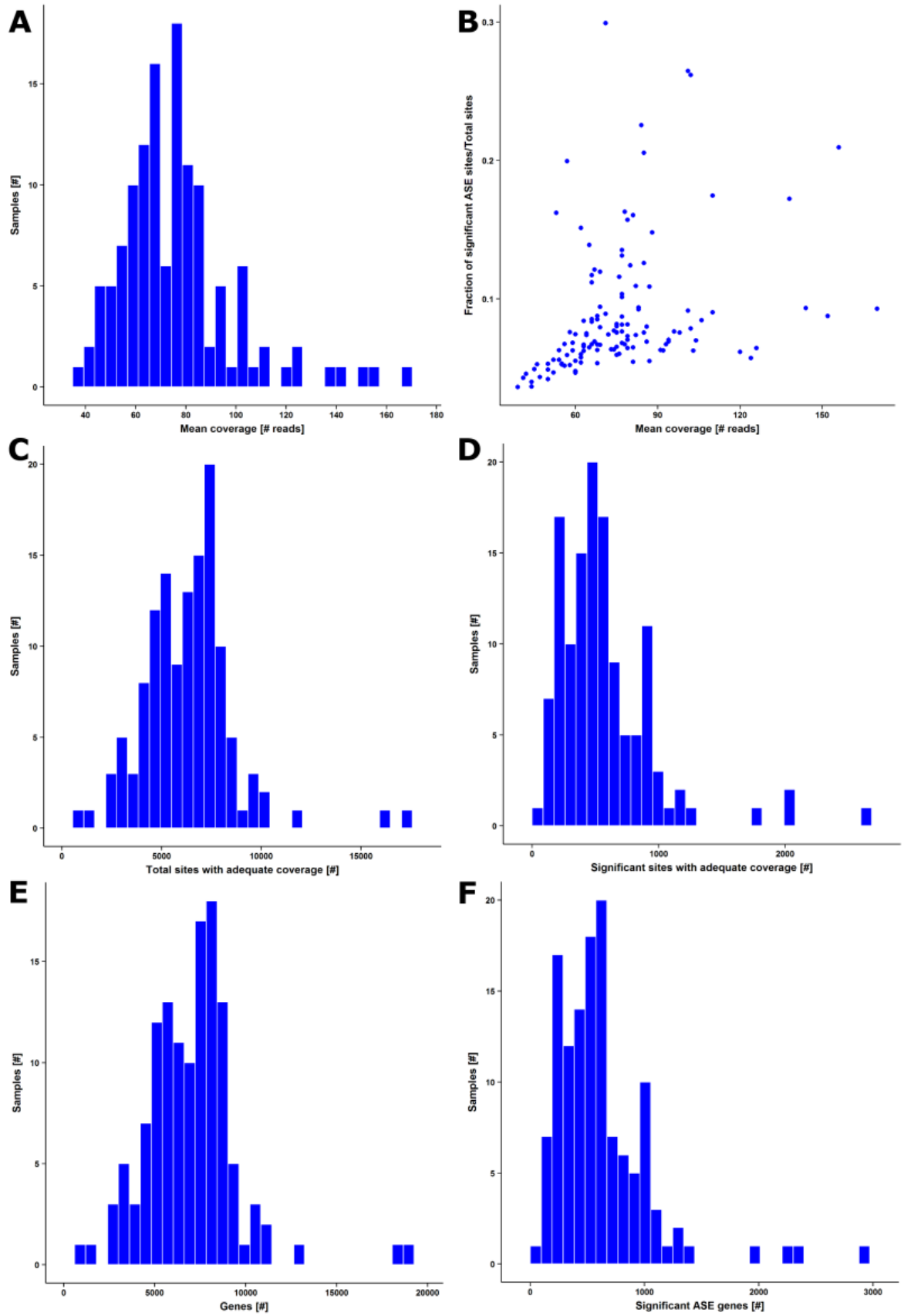


Figure 4.12: Characteristics of sites used for ASE analysis. Statistics on the mean coverage per site (A) and correlation between coverage and fraction of total sites that shows significant ASE (B), as well as the total number of sites (C) or genes (E) per individual tested for ASE with the number of significant sites (D) or genes (F).

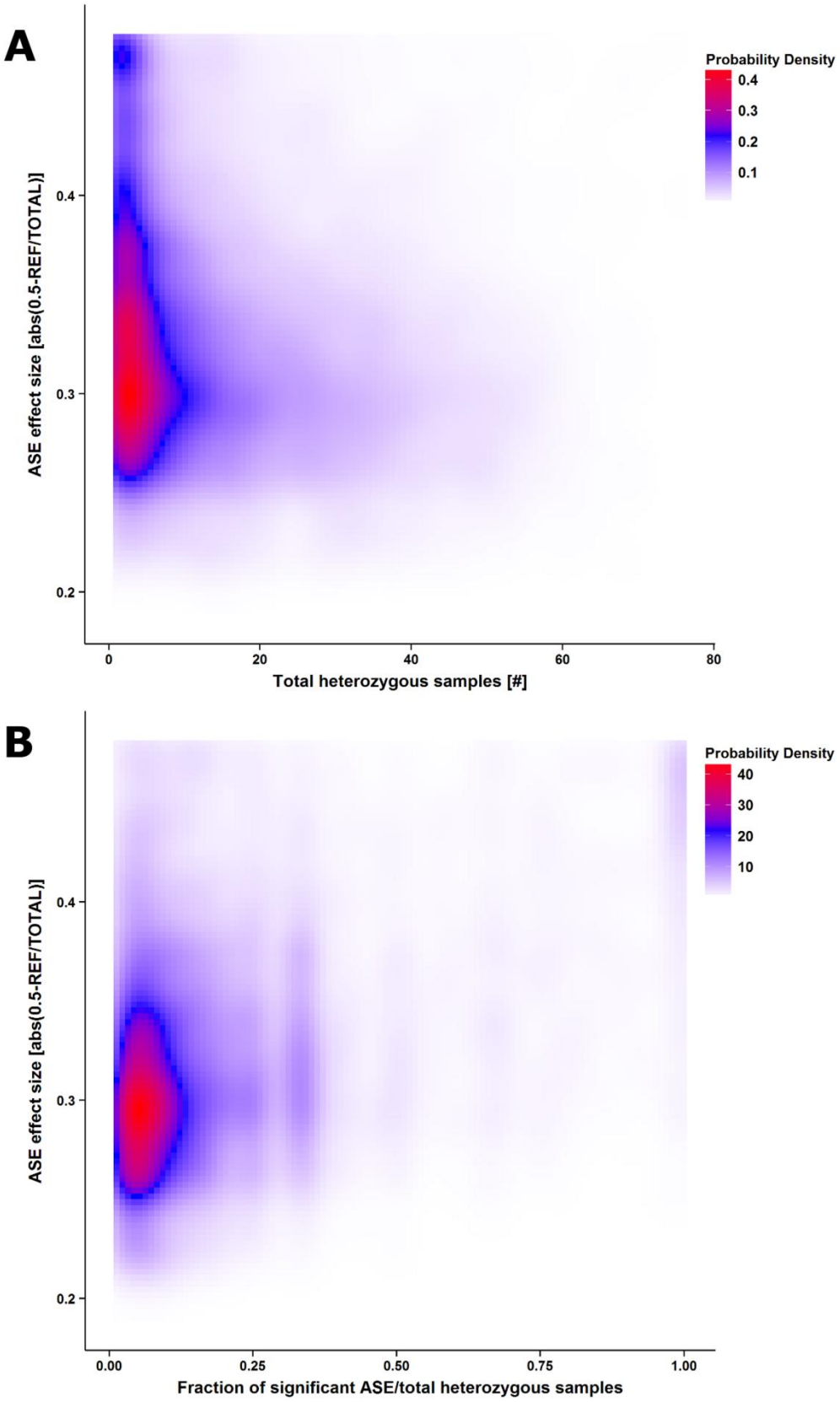


Figure 4.13: Relationship between ASE frequency and effect. Smoothed density plots of the ASE effect size (y-axis) versus the total number of heterozygous samples (A) and fraction of significant ASE sites over total number of heterozygotes (B). ASE is clearly an event associated with rare alleles, and has a modest impact on the allelic balance.

4.4 Discussion

This chapter has described the first and largest to-date catalogue of gene expression in human islets, and explored the effects of genetic variation on gene expression and allele usage. Overall, the gene expression profiles were of good quality and enriched for endocrine genes, with *INS* and *GCG* being the two most abundant transcripts and a strong correlation between the islet and previously published beta-cell[108] profiles. There was some evidence of contamination by exocrine and inflammatory cell types. Since only very impure samples were excluded before data generation, and both the freeze/thawing of many samples and the islet isolation process itself can be seen as severe trauma to the cells, the presence of such contaminants is unsurprising. However, normalisation using hidden covariates almost entirely corrected for these difference when studying *cis*-acting QTLs.

By interrogating the impact of genetic variation on islet gene expression, the data reported in this chapter uncovers a wide range of regulatory variation. In line with published data, the significant eQTL genes uncovered in an islet dataset of this size are primarily shared between tissues[205,206,239,243]. However, the enrichment of islet regulatory elements, in comparison with regulatory elements from different cell-lines seen for eQTL variant across the *P*-value range, suggests that there are many true islet regulatory variants present in the currently non-significant part of the data. By combining data from different research groups, larger future analyses will provide the statistical power to identify more of these tissue-specific islet eQTLs. The enrichment of lncRNAs genes amongst the significant eQTL genes suggests that the expression of these transcripts, which often show a very cell-line specific expression pattern[106,209], is even more influenced by genetic variation than protein-coding genes. Several of the lncRNAs with significant eQTLs were found near genes with known islet function. A recent study by Morán *et al.* hinted at important functional roles for lncRNAs in islet physiology[209], which makes these lncRNAs with significant eQTL signal in islets attractive targets for further functional studies. Allele-specific expression analysis also

uncovers widespread regulatory variation outside the range detectable through eQTL analyses. Most of the variants detected through these analyses are rare, and, when compared across individuals, are likely to be frequently caused by even rarer events.

The current approach was based on gene expression levels quantified using Flux Capacitor[234], which was used to determine isoform expression levels as well as total gene expression. However, whilst similarly quantified isoform expression levels have been used in other QTL studies[207] and the global isoform profile was of good quality, individual-level predictions in the islet data were too noisy to perform such analysis. Since the published data was generated from LCLs[207], the less homogeneous cell population and probable larger variation in RNA quality associated with islet data might help explain this difference in usability of the isoform quantifications. Future studies using raw reads overlapping exons therefore offer a promising approach to studying such effects on a transcriptome-wide scale. The data can also be used to study *trans*-regulatory effects, or to inform islet co-expression networks. While such analyses will require a different approach to eQTL analysis, which studies only local effects, and require careful consideration of sample inclusion criteria and normalisation strategies given potential contaminants, they are very likely to uncover new islet biology.

In conclusion, the data presented in this chapter can serve as an important reference for interrogating genetic variation identified through T2D GWAS in the context of islet regulatory effects. I will explore the islet regulatory variation in the context of common T2D association loci in section 6.3.3. By investigating islet gene expression in the context of functional data, it can also help to uncover novel gene expression networks involved in islet physiology.

Chapter 5

Fine-mapping of common association loci

5.1 Introduction

The previous two chapters have outlined efforts to increase understanding, and genomic annotation, of islet regulatory sequence. However, whilst genome-wide association studies (GWAS) have successfully identified hundreds of genomic regions that play a role in over five hundred diseases and traits, determining causal variants has proved much more difficult. Current strategies following GWAS findings try to refine the location of causal variants as far as possible, and then overlap these refined locations with regulatory annotations. However, whilst there are many competing theories, no detailed studies have been performed to determine how well fine-mapping actually works. Before GWAS data can be properly connected to annotations such as those generated in chapters three and four, it is therefore necessary to understand how well the true location of causal variants within GWAS loci can be refined.

In this chapter, I will describe the need for statistical refinement of association loci to identify causal variants, and take advantage of 1000 Genomes data to explore the utility of such methods (using Bayesian and trans-ethnic fine-mapping as exemplar approaches) in achieving this refinement. Finally, I will aim to refine known common T2D association data by applying the Bayesian fine-mapping approach to genome-wide meta-analysis data from the GoT2D study.

5.1.1 Fine-mapping of GWAS loci: needs and difficulties

Since causal variants underlying GWAS association loci are mostly unknown, the span of associated regions is generally defined by broad LD blocks flanked by recombination hotspots. The resulting genomic regions can span up to 1Mb, include hundreds of variants and incorporate many annotated features (see, for example, figure 5.1). Much progress is currently being made in improving tissue and cell-line specific annotations for regulatory elements through efforts such as ENCODE[111] and the work described in chapters 3 and 4 of

this thesis. However, without means to confidently reduce association intervals from broad LD blocks to more manageably sized regions or numbers of variants, even accurate functional annotations from relevant tissues are likely to struggle in uncovering causal variants.

Statistical fine-mapping can be used to home in on causal variant(s) at GWAS loci. Recent studies have shed light on many of the issues which previously hampered fine-mapping efforts. For example, uncertainty existed about the frequency of causal variants. The reproducibility of association signals across different ethnic populations, including those for T2D[250], suggested that causal variants were to be found in the common (minor allele frequency [MAF] > 0.05) allele frequency spectrum. However, some argued the common variants were, in fact, "synthetic" association signals reflecting underlying rare causal variants with large effect sizes[251]. According to this argument, rare causal variants not interrogated by GWAS would create "synthetic" association signals at the common variants with which they are in closest LD. Since then, much data to improve our understanding of association loci using a variety of study designs – large scale GWAS [56,252,253], fine-mapping using high density local genotyping[252-254], re-sequencing[100], trans-ethnic fine-mapping[103], and simulation studies into genetic architecture of complex traits[255,256] – have become available. GWAS and fine-mapping data clearly point to underlying causal variants being largely common, whilst trans-ethnic refinement identifies consistency of association signals across ethnic groups both of which argue against rare causal alleles. Furthermore, modelling the genetic architecture underlying GWAS data identifies a long tail of association loci for complex traits: if many of these loci were driven by rare variation, familial aggregation would be much higher than observed. The findings from all these studies therefore largely refute the "synthetic" association theory.

With the likely (common) allele frequency of causal variants underlying GWAS signals established, another requirement for successful refinement of association loci is complete coverage of that part of the allele frequency spectrum in which the causal variants reside.

Otherwise an unobserved variant might inadvertently be excluded from the fine-mapped region. Whilst instrumental for the GWAS imputation which facilitated the discovery of so many association loci, the available data from HapMap did not provide such complete coverage. However, the recent arrival of whole-genome sequencing data from consortia such as the 1000 Genomes project[97] means that coverage of the common and low frequency (MAF > 0.01) allele spectrum across ancestry groups is now almost complete.

Two of the major obstacles to successful fine-mapping have therefore been overcome. However, it remains to be established how LD might limit fine-mapping efforts. The next section will outline two potential strategies to deal with this potential issue.

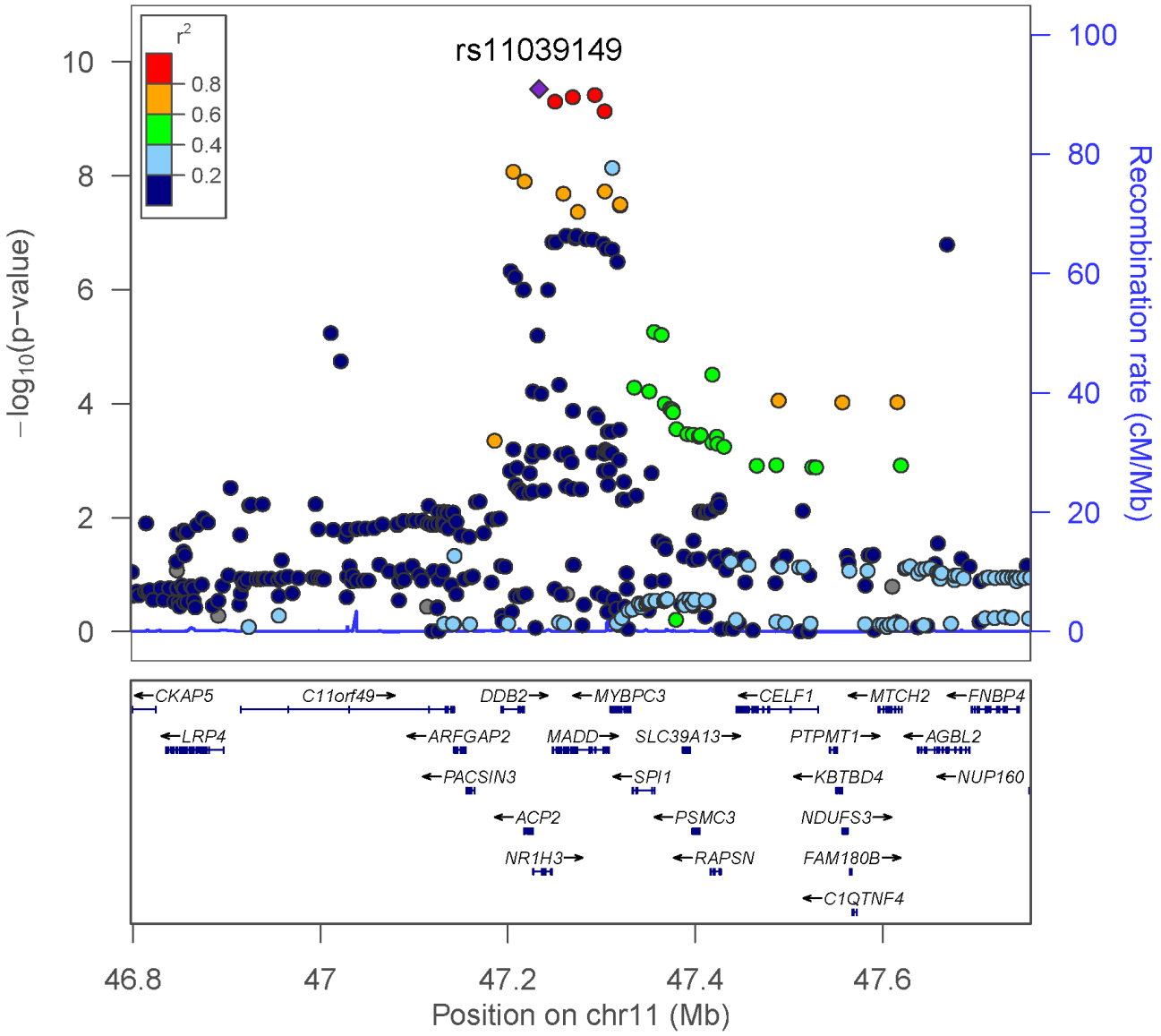


Figure 5.1: *MADD* locus associated with fasting glucose levels. The *MADD* fasting glucose locus is an instructive example of a locus at which refinement of the association signal will be needed to identify causal genes. Although a clear association signal is present, there is no delineation through recombination (blue line corresponding to the right y-axis) – meaning that LD (as denoted in r^2 by the colours of the dots) spans a large interval containing more than 10 different genes.

5.1.2 Fine-mapping strategies: trans-ethnic and Bayesian approaches

LD blocks often span large intervals, and the correlation between genotypes within them makes it challenging to distinguish true causality between variants in high LD. As mentioned in section 1.4.3, several strategies have been proposed to overcome the problem of LD. For example, the differences in haplotype structure and LD between populations can be used to refine association signals – something often referred to as trans-ethnic fine-mapping. Trans-ethnic fine-mapping will work if there is overlap of association signal and little allelic heterogeneity between ethnic groups at a given GWAS locus. Provided this holds true, the differences in haplotype structure and extension of LD across a region in a set of populations can be used to refine the original signal to a smaller interval. The utility of trans-ethnic approaches for fine-mapping susceptibility genes was demonstrated in the late 1980s: by investigating haplotype differences for genes encoding class II antigens of the major histocompatibility complex (HLA-DQ and HLA-DR) in individuals of Caucasian and African ancestry, Todd *et al.* were able to refine this established Type 1 diabetes susceptibility locus[257].

Recent analyses using multi-ethnic GWAS data have shown that this trans-ethnic approach is also able to fine-map association signals. Wu *et al.* performed a fine-mapping study of 58 known loci associated with lipid levels in three different populations (African Americans, East Asians and Europeans)[103]. Using a *P*-value threshold of 10^{-4}, 38 of the 58 previously identified genome-wide significant loci also showed significant study-wide association. While per-locus fine-mapping statistics – such as the number of overlapping variants in high LD with the meta-analysis index variant across the ancestry groups – were not reported, the study identified five loci which could be refined successfully to between one and four variants using differences in LD structure between the populations.

Trans-ethnic fine-mapping of known loci for body mass index (BMI) combining African American and European LD had similar success[258]. Eight of the 21 loci previously reported in

European populations showed significant association in the African American cohort at a *P*-value threshold of 5.8×10^{-5} and were used for fine-mapping. The mean number of variants in LD ($r^2 > 0.5$) with the index variant, and the mean interval spanned by these variants, were both reduced using a trans-ethnic fine-mapping approach (from 55 to 16 variants and 86kb to 27 kb, respectively). Only one of the eight loci was resilient to fine-mapping – in Europeans this locus already spanned the smallest region and included only five variants.

A Bayesian approach to fine-mapping association loci is another strategy to overcome the problems posed by LD. Such an approach allows comparison between variants within an association locus based on their posterior probability of causality given all other variants in that locus. Through this method, it is possible to define credible sets of variants or credible intervals defined by a posterior probability threshold (e.g. 99%) for association. Provided the assumptions regarding the number of causal variants when calculating the posterior probabilities are correct, and enough association signal is present in a given locus, Bayesian approaches can thus resolve LD beyond a blanket threshold across all loci.

Several studies have been published showing that a Bayesian approach is an attractive solution for fine-mapping. Frequentist alternatives to fine-mapping are also available, which typically involve stepwise logistic regression conditioning on association of each variant individually. Such approaches have been successful in refining association signals[103], and uncovering secondary signals at many association loci[56,57,66,82,253,259]. Both approaches will be hampered by strong LD and small effect sizes. However, whereas the frequentist approach treats refinement of each locus as identical, the Bayesian alternative provides a probability that each refinement will be correct given all association signal present. This means that at a locus with 100 variants just below the significance threshold of the last variant included in a fine-mapped set by stepwise conditional analysis, the Bayesian-derived posterior probability of this set containing the causal variant would be much different from a locus that contains only a single variant with a similar significance level to the last included variant. Even though the association signal might be stronger in the former example, the latter will

therefore be a better candidate for functional follow-up. Furthermore, Bayesian analyses allow the inclusion of prior information, which make them more suited to scenarios in which risk is the property of the haplotype (e.g. several variants in phase affecting binding of a transcription factor).

There are several published studies showing the applicability of Bayesian fine-mapping approaches. Maller *et al.* performed high density genotyping of 14 loci implicated in three different diseases (T2D, coronary artery disease [CAD], and Graves' disease) and applied a full Bayesian approach to refine the association signals[254]. Using 95% likely credible sets, this study effectively fine-mapped half the loci, and at three of these seven loci more than 50% of the posterior probability for causality rested on a single variant (*TCF7L2* and *CDKN2A/B* for T2D, and the *CTLA4* locus for Graves' disease). The seven loci for which the fine-mapping was unsuccessful also showed the weakest association signals, which meant the sample size of the study was too small to resolve regional LD and refine the signal.

A recent study by the International Multiple Sclerosis Genetics Consortium also successfully used Bayesian fine-mapping to refine association signals for multiple sclerosis[252]. Of the 66 association loci with sufficiently high density genotype data available, around 50% were refined to 90% credible sets containing less than 50 variants and around 33% to 90% credible intervals spanning less than 100 kb. Similar to findings from Maller *et al.*, several of the refined loci contained single variants with a large (>50%) posterior probability for causality (at *TNFSF14*, *IL2RA*, *TNFRSF1A*, *IL12A* and *STAT4*).

The previous studies use a full Bayesian approach to refine association loci. This requires the specification of a prior distribution for all unknown parameters, which is computationally expensive and necessitates access to raw genotype data. Alternatively, Bayes factors can be approximated from a "likelihood" of association derived from the summary statistics of extensive publicly available GWAS data combined with a normal prior on the log relative risk[260]. In both approaches, posterior probabilities of the likelihood of causality can then be derived (with the incorporation of prior weights on certain types of variants where

appropriate) from the resulting Bayes factors for each variant within a region with evidence for association.

Fine-mapping GWAS loci to localise causal variants will be instrumental in defining mechanisms underlying association signals. Since the accuracy of fine-mapping is unknown, simulations were used to evaluate how often fine-mapping approaches (using approximated Bayesian fine-mapping and trans-ethnic study designs) can successfully refine association loci.

5.2 Methods

5.2.1 Bayesian approximation

While privacy concerns mean that individual-level genotype data for GWAS are generally unavailable to researchers, large amounts of summary statistics are available online. Summary data do not contain enough information to calculate Bayes factors directly, but the information contained within can be used in a Bayesian framework by instead approximating Bayes factors as proposed by Wakefield in 2007[260]. This null hypothesis is a model of no association, whereas under the alternative model there is association. The approximation assumes that the likelihood distribution for association is summarised by the regression parameters with the prior for association centred on 0 (which corresponds to the null of no association) and a variance dependent on W, which describes the strength of association conditional on its existence. The value for W is set at 0.4, which equates to a 95% belief that the relative risk corresponding to departure from the null model is less than 1.5.

This results in an equation for the approximated Bayes Factor given as

$$\mathrm{ABF} = \frac{1}{\sqrt{1-r}} \exp\left(-\frac{Z^2}{2} r\right)$$

where Z is the Z-statistic describing the strength of association derived from the regression and r a shrinkage factor

$$r = \frac{W}{V + W}$$

defining the ratio of prior variance to total variance. The resulting Bayes factor is dependent the effect size through Z and the power of the study through the variance of the effect V.

5.2.2 Description of the simulations pipeline

The design of the simulations pipeline used to establish the ability of the approximated Bayesian approach (with and without trans-ethnic study populations) to accurately locate the causal variant and effectively refine the number of credible variants and size of the credible locus is outlined in figure 5.2. Using phased haplotypes from 85 CEU individuals (1000G Project Phase 1, release 3[97]) containing only single nucleotide variants, 1 Mb autosomal genomic regions, selected at random from the genome, were simulated using HAPGEN2[261]. This algorithm was selected because it maintains realistic LD patterns by resampling known haplotypes (see example in figure 5.3). The regions each contained a single causal allele of varying risk allele frequency (RAF; RAF 0.5, 0.2, 0.1, and 0.05) and phenotypic effect (odds ratio [OR] 1.5, 1.2 and 1, where 1 corresponds to the null model of no association). For each RAF/OR parameter setting, genotype data for ten studies of 1k cases versus 1k controls were generated for one thousand regions.

As a "gold standard" – i.e. accurate sequencing of the whole region – association testing was performed directly on the simulated haplotypes as described below. To better reflect a GWAS setting, the variants simulated for each study were downsampled to an Illumina HumanOmni2.5 BeadChip array (unless mentioned otherwise these data are used for the results presented in section 5.3). The downsampling was done with and without 5% random variant failure per study, where 5% random failure represents effects such as DNA quality differences and probe failure. The downsampled genotypes were imputed into the 1000G Phase 1, release 3, reference panel using IMPUTE2[262]. Only variants with an INFO score > 0.4 were included in the association analysis.

For each of the ten studies at each locus, association testing was performed by means of logistic regression under an additive model, implemented in SNPTEST, with association summary statistics for all ten studies aggregated by fixed-effects meta-analysis using GWAMA[263]. From the GWAS summary data, Bayes factors for each variant were approximated as described in section 5.2.1. These approximated Bayes factors (ABFs) were used to derive posterior probabilities for association for all variants in each of the simulated regions.

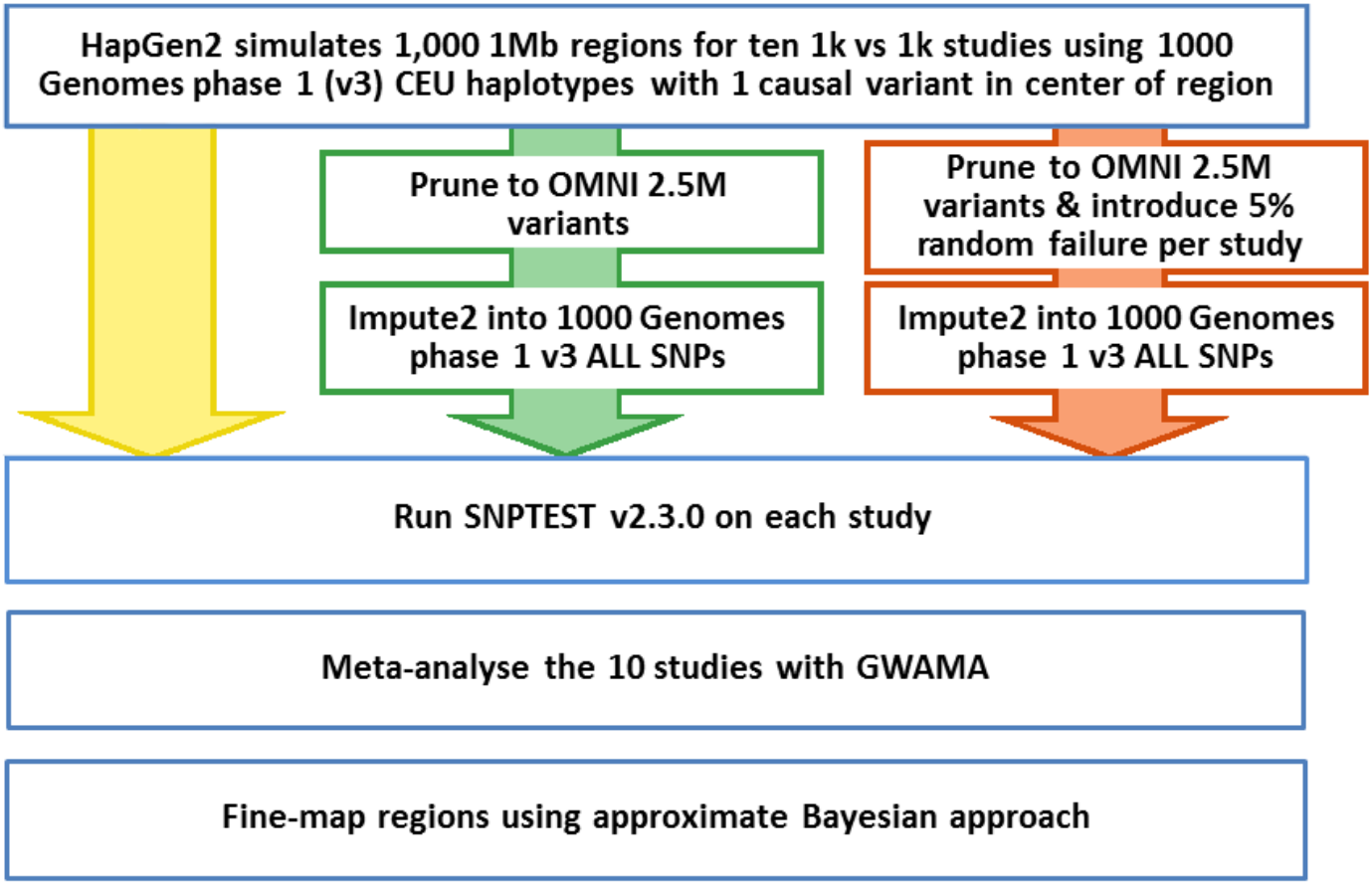


Figure 5.2: Schematic representation of the simulation pipeline. Description of the programs and datasets used at each step with colours denoting the three arms of the pipeline representing sequence-like genotypes (yellow), perfect GWAS-like data (green) and GWAS-like data with experimental variation (orange).

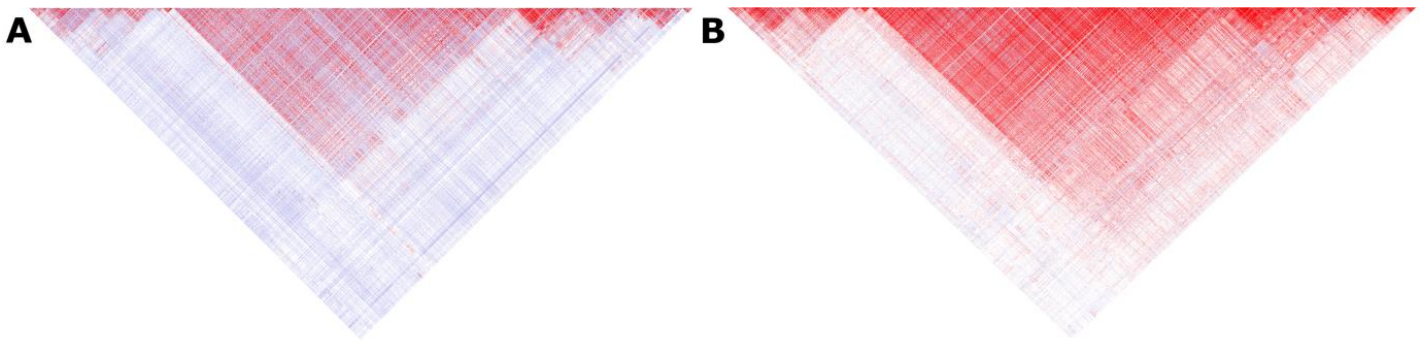


Figure 5.3: example LD plots of 1Mb simulated region with HapGen2. The figure shows a representative region on chr2:53,698,430- 54,698,430 with the original LD in the 85 CEU individuals in 1000 Genomes phase 1 data (A) and the simulated LD in 1000 samples using HapGen2 (B). Sampling of 1000 genotypes from an original pool of 85 leads to an overall increase in LD between variants, but the overall pattern of LD from the original data remains.

5.3 Results

5.3.1 Establishing the usefulness and accuracy of fine-mapping in a Bayesian framework

Firstly, to allow correct interpretation of the data, the calibration of the Bayesian approach was evaluated (i.e. credible sets had appropriate coverage of the causal variant under a model of association). Credible sets containing all variants with a cumulative posterior probability of 95% and 99% were derived from the data for each study replicate after meta-analysis. Credible genomic intervals for containing the causal variant were defined using the most 5' and 3' variant within the credible sets. Since the causal variant was known, inclusion or exclusion (coverage) of the causal variant in these credible sets and intervals could be established. The coverage ranged between 95.0-98.0% and 98.0-99.9% for the 95 and 99% credible sets respectively (Figure 5.4). The credible genomic intervals were more likely still than the credible sets to contain the causal variant, with 98.0-99.6% and 99.3-100% coverage for the 95 % and 99% intervals, respectively (Figure 5.4).

Even when introducing random 5% variant failure per study, potentially eliminating the causal variant or important variants for imputation in the scaffold, coverage of the causal variant did not drop dramatically (93.0-97.0% and 98.0-99.5% for respectively 95% and 99% credible sets; Figure 5.4). An overall genotype failure rate of 0.5% (reflecting 5% per study with ten studies included in the GWAS) is unrealistically high for modern arrays, and in empirical data the genotype failure follows a non-random distribution, so the 5% random failure scenario represents a conservative estimate for coverage in real data.

The data reported above establishes that the credible sets are well calibrated for coverage. The next question to answer is whether there is in fact a need for fine-mapping at all: if the simulations showed near complete overlap between the most significantly

associated variant and the causal variant, further analyses aiming to identify causal mechanisms could focus on the lead variant alone.

The overall power of the GWAS scenario to detect any genome-wide significant association ($\log_{10}(ABF)>6$[264] – corresponding to a prior odds of association of one variant in a million) at the simulated loci – is high across all conditions (minimum power 79% for effect allele with RAF 0.1 and OR 1.2). In the "gold standard" scenario with the most power (RAF 0.5 and OR 1.5), 79% of the simulations accurately identify the causal variant as the most associated (Figure 5.5a-b). However, when the risk allele is less common (RAF 0.1) and the phenotypic effect is more in line with most GWAS hits (OR 1.2), the probability of accurately identifying the causal variant by focussing on the most associated variant falls to 30% (Figure 5.5a-b). This demonstrates that, rather than reporting the lead variant, more elaborate approaches to fine-mapping are needed to accurately identify causal variants from GWAS results.

Next, effectiveness of the Bayesian approach for refining the number of credible variants and/or the size of the credible interval was determined. The results demonstrate that the median number of variants contained within the credible sets varies with the strength of the association signal. When there was strong association signal - e.g. at RAF 0.5 at OR 1.5 - the median 95% credible set contained only a single variant and the median 99% set two, with 95% of all sets containing <18 and <23 variants for the 95% and 99% credible sets respectively (Figure 5.6a-b). In the least powerful scenario tested (RAF 0.1 at OR 1.2), the median set still only contained 15 variants at 95% and 24 variants at 99% posterior probability cut-offs, but the range was much larger, with 95% of all sets now containing <189 and <1,545 variants (Figure 5.6a-b). The credible intervals show a similar pattern, with a range of median sizes [95% of all sets] from <8.8kb [<116kb] for RAF 0.5 variants to <51kb [<1Mb] for RAF 0.1 variants with an OR of 1.2 at 99% likelihood (Figure 5.6c-d).

These results clearly demonstrate the ability of the Bayesian approach to accurately identify likely sets and intervals containing the causal variant. However, it could still be the case that applying a blanket r^2 threshold to capture all variants in strong LD with the most associated variant might be just as informative, rendering the Bayesian approach unnecessary. To assess this possibility, a set of variants in strong LD ($r^2 > 0.8$) with the most significant variant and the genomic interval covered by these variants were calculated for each simulation. The size of these sets and intervals, as well as the number of times the causal variant was included in them, could then be compared against the results from Bayesian fine-mapping. This comparison shows that in all but two scenarios (RAF 0.1 and 0.2 at OR 1.2), the Bayesian approach at 95% and 99% likelihood of causality outperforms r^2 in ability to refine the number of credible variants and size of the credible region (Figure 5.7). In terms of inclusion of causal variant, the 99% credible sets and intervals are most accurate (99.3% and 99.8% respectively), with the r^2 approach and the 95% credible sets showing similar but lower overall accuracy (97.6% versus 96.8% for the credible sets and 98.2% versus 98.4% for the credible intervals respectively)(Figure 5.7). However, accuracy of the r^2 approach is much more susceptible to a drop in association signal than Bayesian fine-mapping. Whilst in the scenario of strong association signal (RAF 0.5, OR 1.5) the causal variant is included in 100% of the likely sets and intervals, this falls to 90.4% and 92.5% respectively in the scenario with least signal (RAF 0.1, OR 1.2) (Figure 5.7). In comparison, the accuracy of the 99% and 95% credible sets drops from 99.8% to 98.2% and 97.4% to 95.2% respectively, whereas the accuracy of the 99% and 95% credible intervals drops from 100% to 99.4% and 98.7% to 98.2% respectively (Figure 5.7). These results therefore show that the Bayesian approach provides, overall, better refinement of association signals and is less susceptible to changes in the strength of association than a blanket r^2 approach.

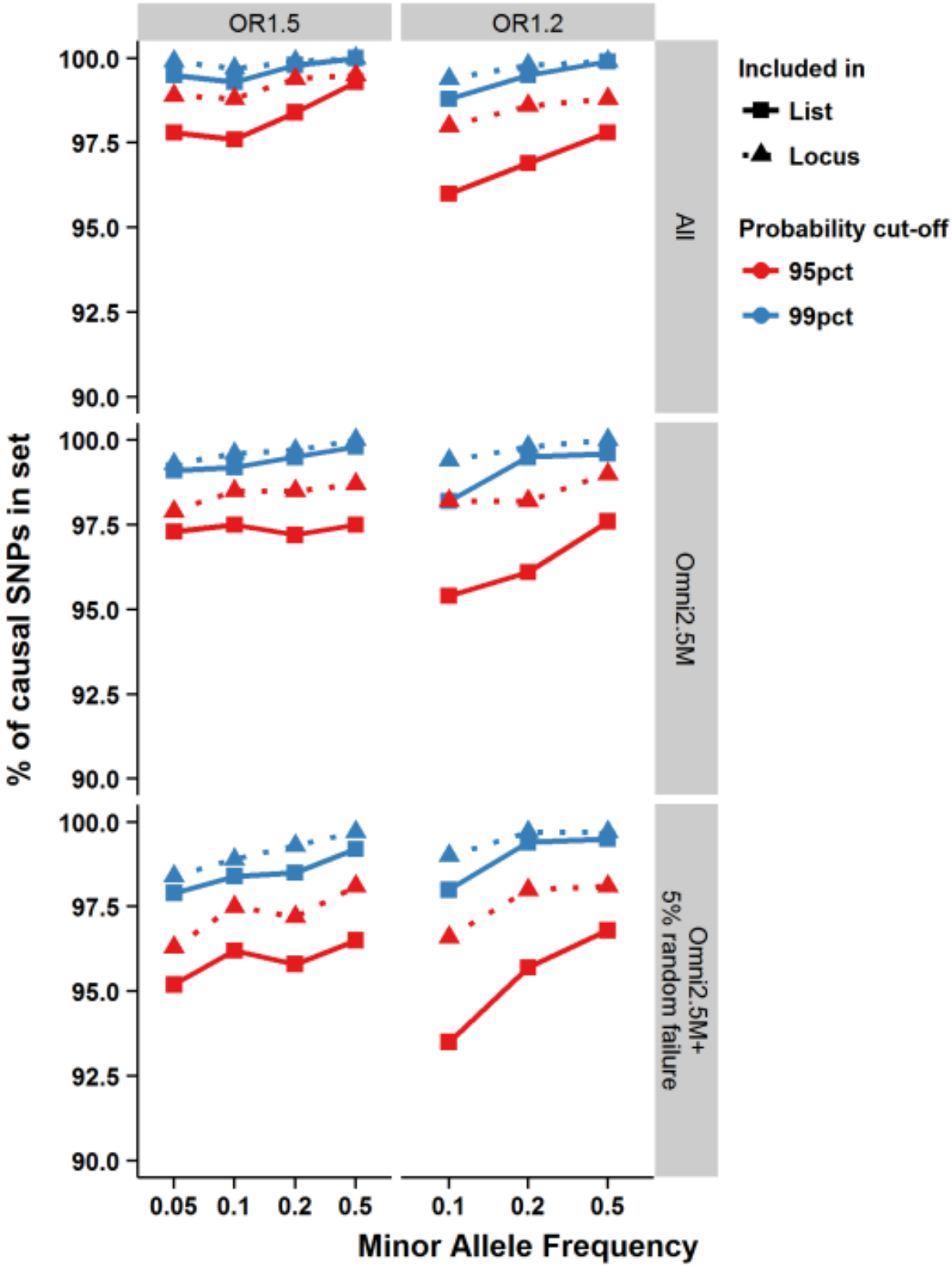


Figure 5.4: coverage of the causal variant in the credible sets and intervals. Percentage of simulations with the causal variant included in the credible set (y-axis) by RAF of the causal variant (x-axis). The figure is split according the OR (horizontal) and arm of the simulation (vertical). Colours denote probability cut-off used for the credible sets and shapes denote whether it is the credible list of variants (squares) or interval (triangles).

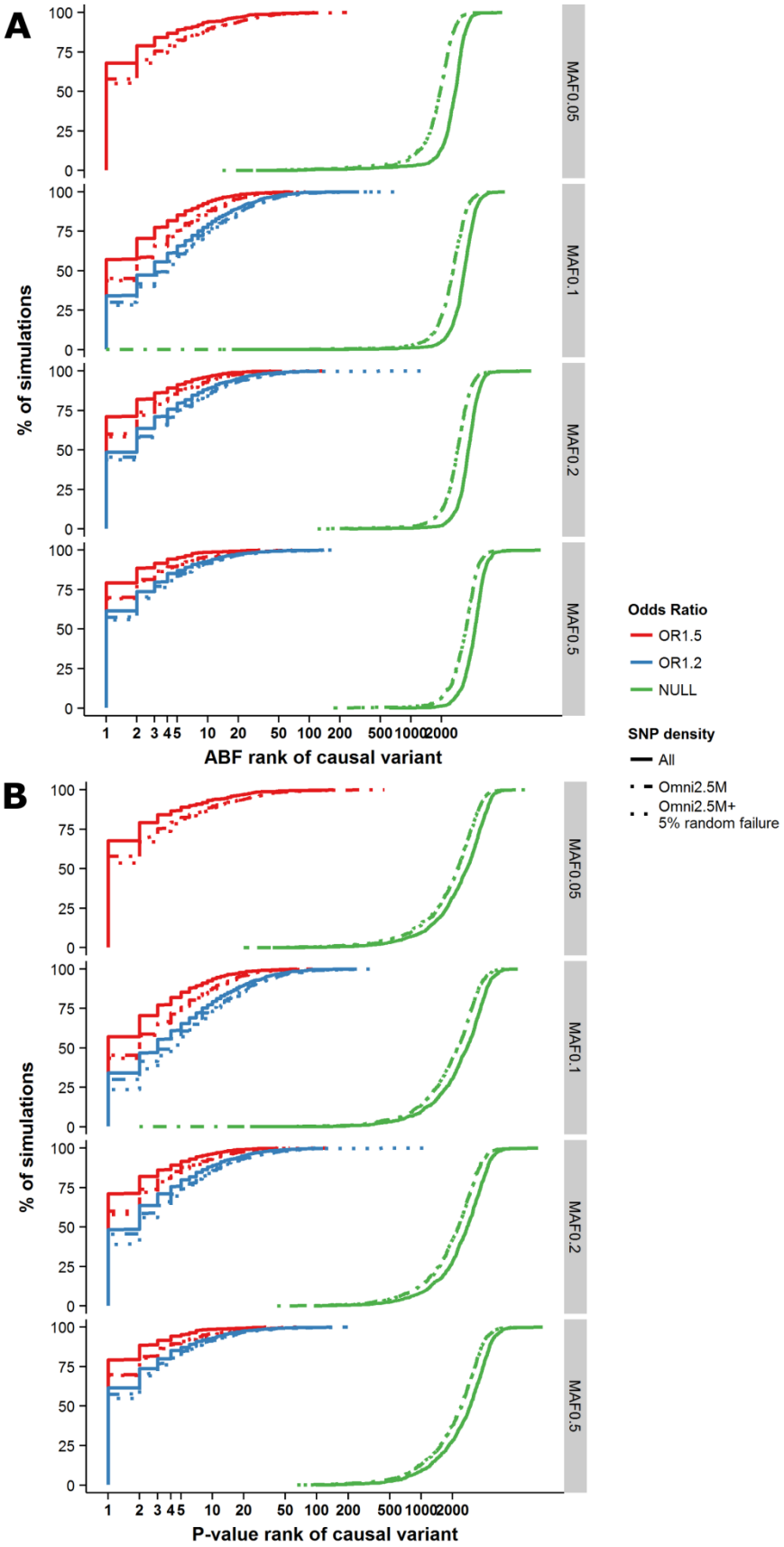


Figure 5.5: ranking of the causal variant. Cumulative percentage of simulations (y-axis) with decreasing ranking of the causal variant amongst all variants in the regions (x-axis) based on ABF (A) and – for comparison purposes – *P*-value (B).

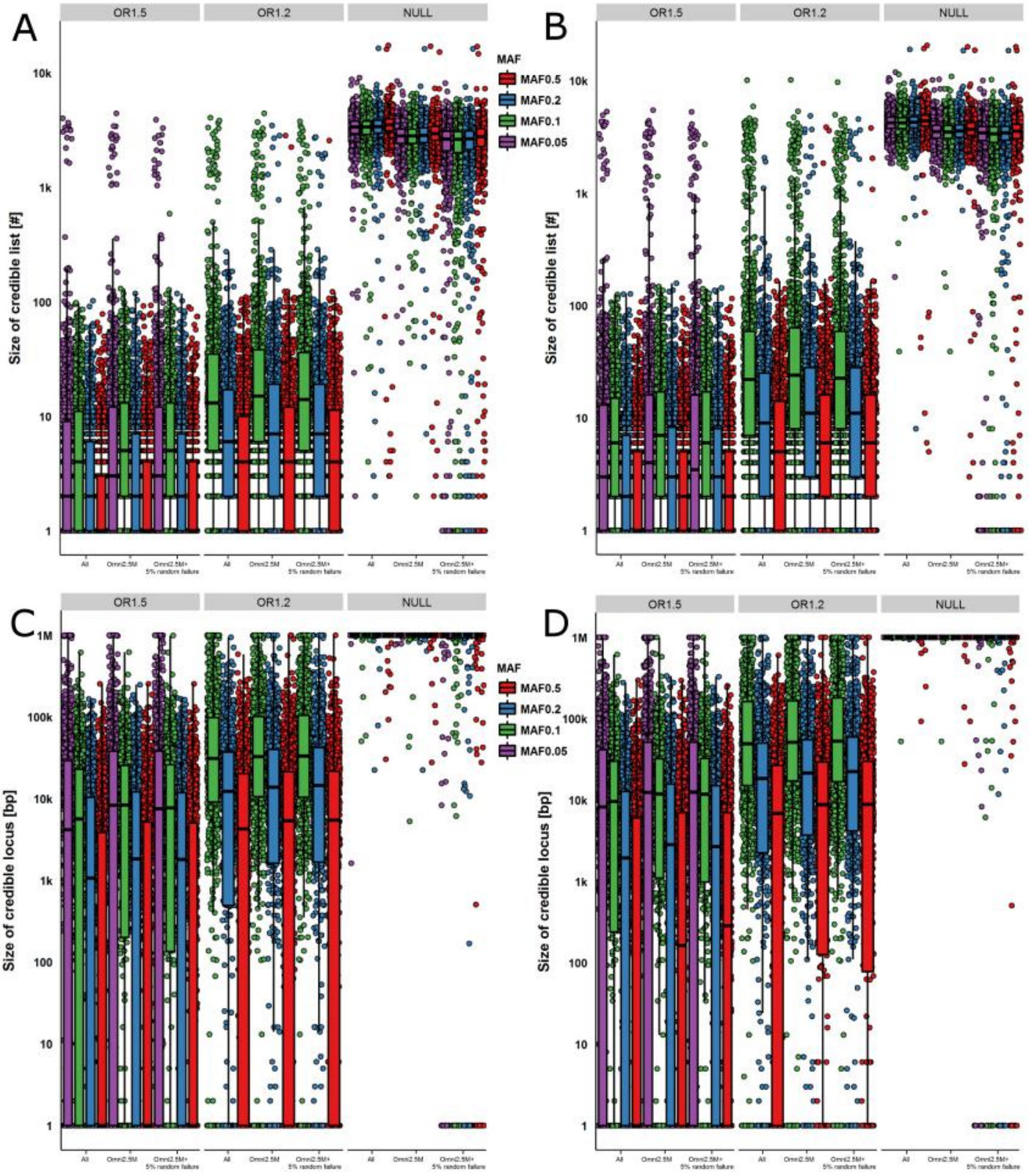


Figure 5.6: refinement of the credible sets and intervals. The y-axis shows the number of variants in the 95% (left panel) and 99% (right panel) likely credible sets (a-b) and, similarly, size of the credible intervals (c-d). The boxplots show the median and interquartile range of the simulations, while each point denotes a single simulated study. The data is split according to the arm of the simulations on the x-axis. The colour of the boxplots/points denotes the RAF of the simulated causal variant, while each panel is split by the effect size along the horizontal plane.

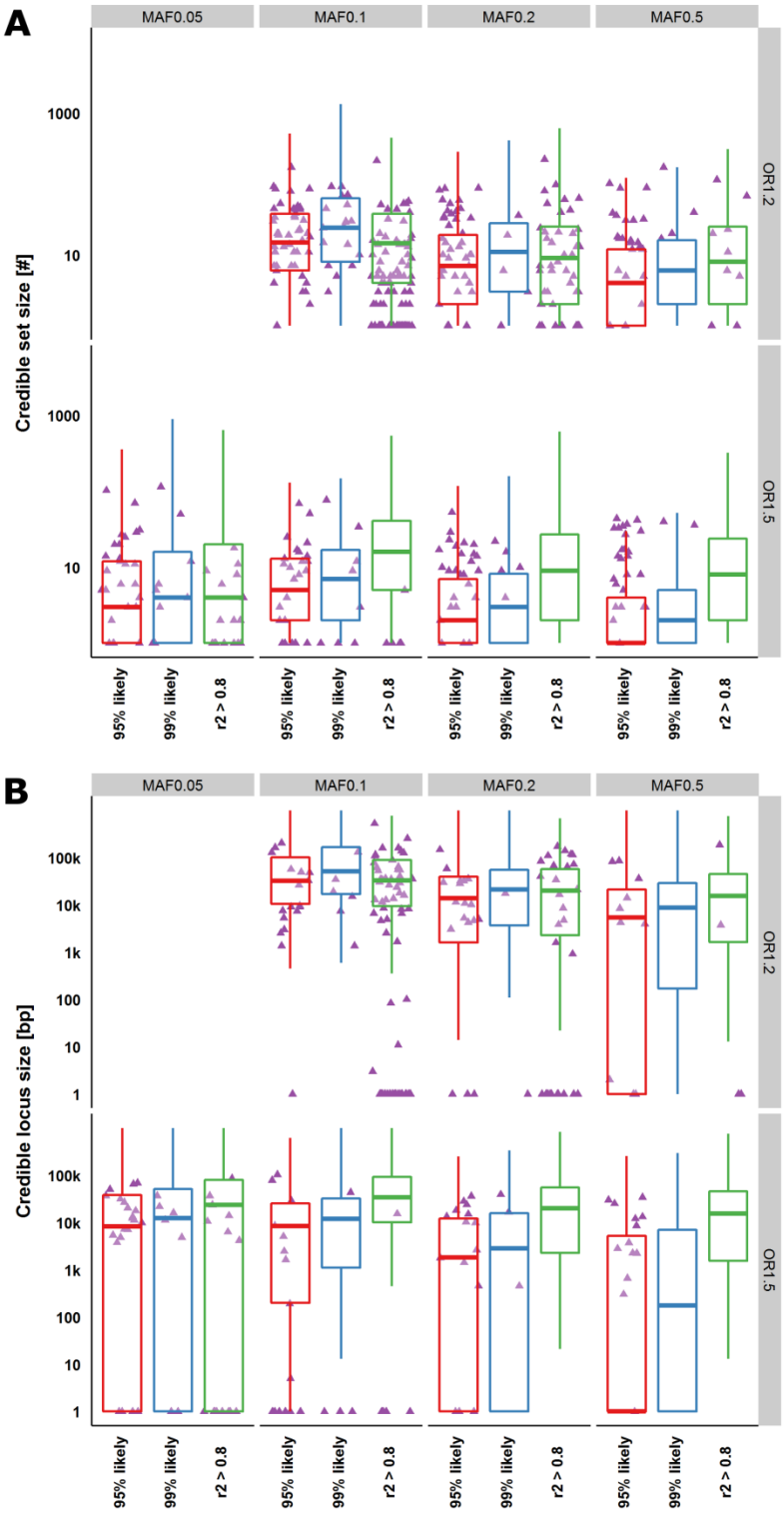


Figure 5.7: comparison between Bayesian fine-mapping and r^2 threshold approaches. The y-axis shows the number of variants (panel A) and size of the credible intervals (panel B) for the 95% credible, 99% credible and $r^2 > 0.8$ sets (x-axis). The panels are split according to RAF along the horizontal plane and effect size along the vertical plane. The colour of the boxplots also denotes the type of set (95% credible, 99% credible and $r^2 > 0.8$). The boxplots show the median and interquartile range of the simulated data, with each purple triangle denoting a simulation where the causal variant is not included in the credible set of locus. To allow easier interpretation, triangles that would overlap are plotted next to each other.

5.3.2 Multi-ethnic study designs for fine-mapping

The previous section demonstrates that association loci are amenable to refinement through Bayesian fine-mapping. Another commonly used approach – trans-ethnic fine-mapping – assumes that adding samples of different ethnic populations will refine association signals by disrupting local linkage disequilibrium, provided there is a shared causal variant between the ancestry groups. To investigate its effects on Bayesian fine-mapping, 1, 3 and 5 populations simulated using CEU haplotypes were replaced by ones simulated using Yoruban (YRI) haplotypes. The causal variant had an OR of 1.2 and the RAF (0.1 and 0.2) was derived from the CEU population and assumed to be shared between these populations. The results showed that in both the lower power RAF 0.1 and the more powerful RAF 0.2 scenario, trans-ethnic fine-mapping resulted in an improvement of resolution compared to a CEU only study. Replacing five CEU populations for YRI increased the likelihood of the most significant variant also being causal by 1.7 (27% to 46%) at RAF 0.1 and by 1.3 (45% to 60%) at RAF 0.2 (Figure 5.8). This increase in the power to correctly identify a causal variant with OR 1.2 in a multi-ethnic design is similar to a single population study variant with an OR 1.5. The credible sets also reflected increased resolution – the median 99% likely sets were reduced to 58% (24 to 14 variants) and 55% (11 to 6 variants) of the CEU size for RAF 0.1 and 0.2 respectively, and the 99% likely intervals to 64% (50kb to 32kb) and 51% (22kb to 11kb).

An increase in fine-mapping resolution correlated to the number of YRI populations added might have been expected, because this should result in shorter LD blocks in more of the simulated loci. However, adding 5 YRI panels did not result in increased fine-mapping resolution over 3 YRI panels (Figure 5.8). This result could indicate that fine-mapping has reached saturation, but, since the RAF of the causal variant is derived from the CEU population without taking into account its frequency in the YRI population, it is more likely to be driven by a reduction in power when the causal variants are in fact not observed at this frequency in the YRI population. Indeed, whereas only ~1% of all simulated CEU cohorts

have missing data for the causal variant, the same statistic stands at 11% for the YRI cohorts. This results in a reduction in power to detect genome-wide association dependent on the number of YRI populations added into the GWAS – 80%, 77%, 73%, and 61% for 1, 3, 5 and 10 YRI populations, respectively, at RAF0.1 and 96%, 94%, 85%, and 69%, respectively, at RAF0.2. The ABFs and posterior probabilities will be affected by this drop through their part-dependency on the power of the study. Notwithstanding this minor limitation of the simulation framework, these results highlight the usefulness of multi-ethnic design for fine-mapping purposes.

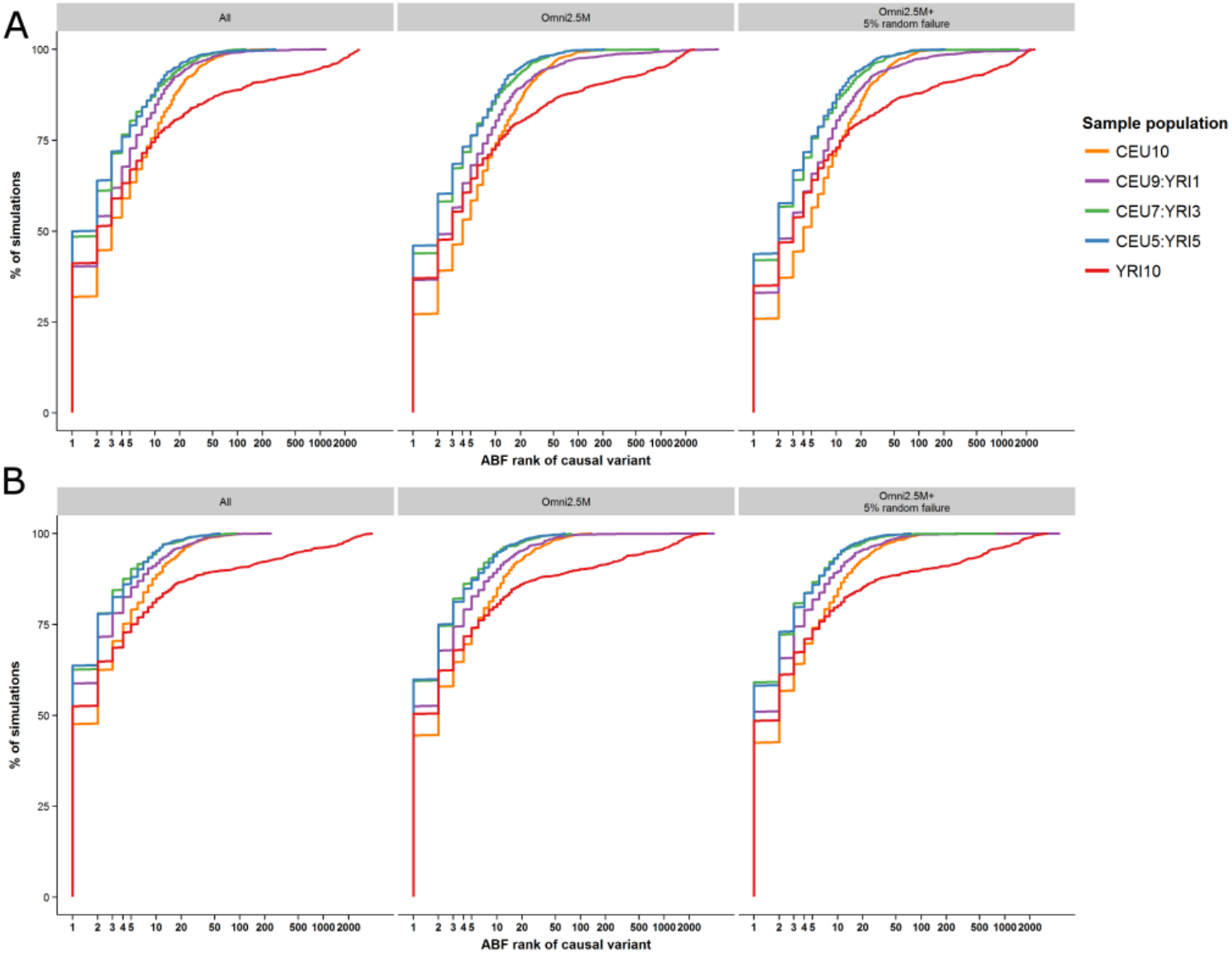


Figure 5.8: ranking of the causal variant using trans-ethnic fine-mapping. Cumulative percentage of simulations (y-axis) with decreasing ranking of the causal variant amongst all variants in the regions (x-axis) based with a causal variant RAF of 0.1 (A) and 0.2 (B). The panels are split along the horizontal by arm of the simulations.

5.3.3 From simulations to empirical data

Finally, data on 26 non-MHC loci genotyped at high density and imputed into a 1000 Genomes reference panel for 9,049 cases and 13,607 controls of European origin from a published GWAS into ankylosing spondylitis[253] were acquired through collaboration with the International Genetics of Ankylosing Spondylitis Consortium. These data were used to establish whether the effectiveness of fine-mapping observed in the simulations could be replicated in empirical data. The genotype density and sample size were in line with the simulated datasets. The median MAF (0.36 [range 0.06-0.50]) and OR (1.15 [1.09-1.68]) of the most significant variants in these loci would result in power to detect association in line with the (lower end of the) simulations (Table 5.1). Fine-mapping resulted in median credible sets containing 20 [range 1-259] and 38 [2-1,113] variants, and credible intervals spanning 36kb [1 base-820kb] and 56kb [1.4kb-858kb] for the 95% and 99% likelihood cut-offs respectively (Table 5.1). Of the 26 loci, 15 were reduced from their original size of several 100kb, containing hundreds of variants, to either less than 30 credible variants or a <50kb credible interval at 99% likelihood (Table 5.1). This number is in line with the 64% similarly fine-mapped loci achieved with simulations of a causal variant of RAF 0.1 and OR 1.2, especially when taking into account the higher power to detect genome-wide significant association in the simulations (79% versus 71% in the empirical data).

Consistent with meta-analysis data from Cortes *et al.*[253], the reported index variant for each locus is included in all but two of the 99% credible sets and all but one of the 99% credible intervals (Table 5.1). Of the two variants that were not included in the 99% credible sets, one failed imputation but would be included when using the 99% credible interval (Table 5.1). At the other locus, most of the posterior probability rested on two variants in strong LD (r^2 = 0.9) with the reported Cortes *et al.* index variant rather than the index variant itself (Table 5.1). These two variants (rs34236350 and rs3749171) fall in a 1.4kb interval which overlaps the last intron and exon of *GPR35*, an orphan G protein-coupled receptor (Figure 5.9). One of these two most likely variants, rs3749171, is a missense variant

(threonine to methionine) in all three annotated transcripts of *GPR35*, and is also associated with another set of auto-immune diseases (inflammatory bowel disorders)[170]. This suggests that rs3749171, rather than the Cortes *et al.* index variant at this locus, might be mediating the association signal. It also shows that refinement is able to identify functionally relevant variants. Another example of the efficacy of refinement can be found at the *FCGR2A* locus, which is fine-mapped to 6 variants spanning 11kb of the promoter region and first 3 exons of *FCGR2A*, an immunoglobulin Fc receptor gene found on the surface of many immune response cells. This 99% credible set also contains a missense variant (rs1801274) associated with other immune-related disorders (Figur5e 5.9)[170,253].

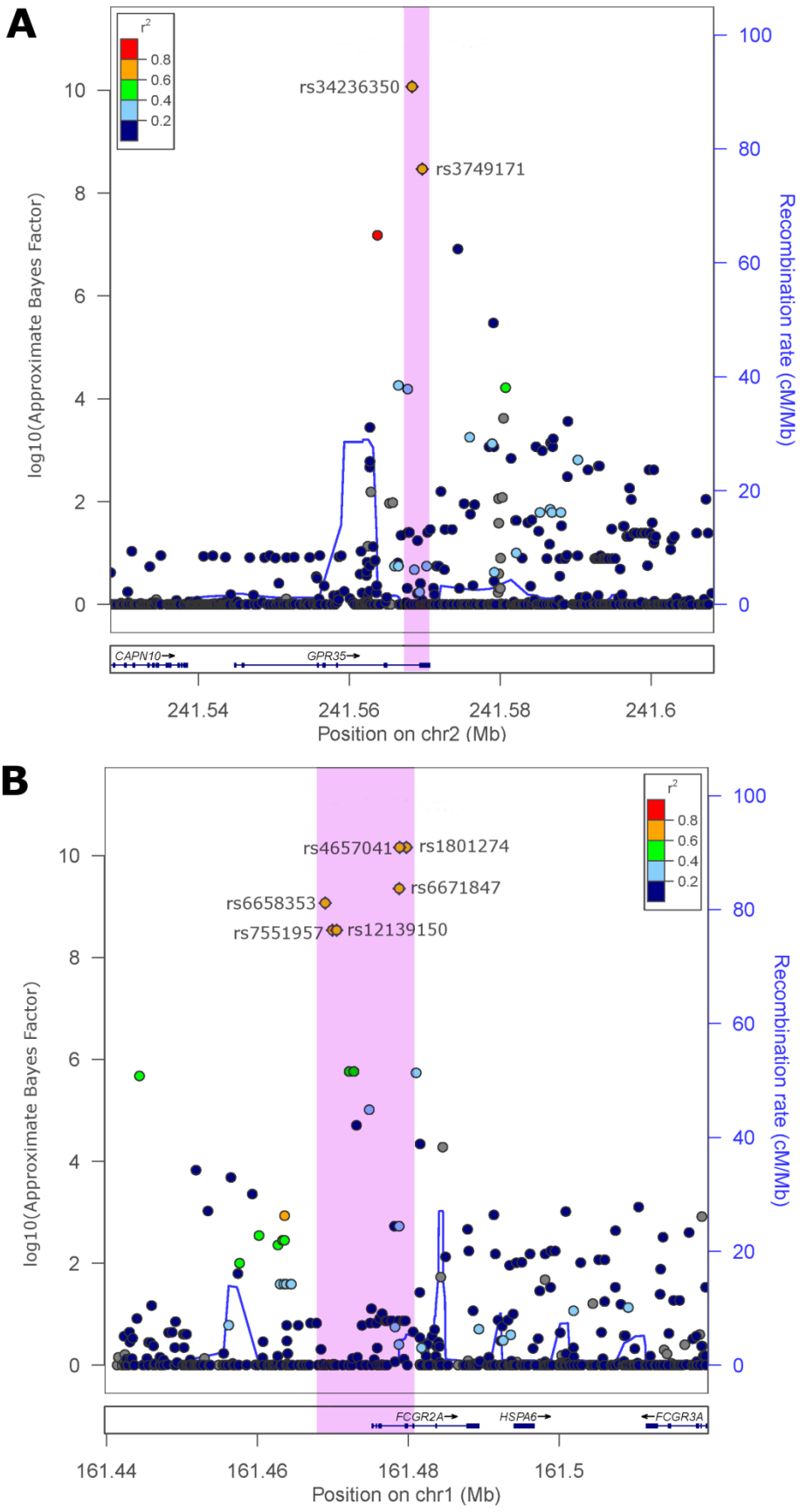


Figure 5.9: fine-mapping the AS regions. *GPR35* (A) and *FCGR2A* (B) association loci for AS, where the 99% credible interval (pink) and credible sets (gold dots) could be fine-mapped to a small region containing few variant

Locus	Most significant ABF variant	MAF	OR	log_{10} (ABF)	95% credible set [#]	99% credible set [#]	95% credible locus [bp]	99% credible locus [bp]	Index variant of primary signal in Cortes *et al.* [253]	Cortes *et al.* index variant in 99% credible set
ERAP1	rs2032890	0.28	1.38	44.82	12	12	5,157	5,157	rs30187	Yes
2p15	rs4672505	0.39	1.31	37.57	4	5	9,241	11,369	rs6759298	Yes
IL23R	rs11581607	0.06	1.68	27.06	9	11	13,432	17,037	rs11209026	Yes
KIF21B	rs296520	0.27	1.22	16.14	10	10	7,367	7,367	rs41299637	Yes
21q22	rs4817986	0.25	1.19	12.19	8	11	3,462	4,361	rs2836883	Yes
RUNX3	rs10751775	0.48	1.15	11.06	43	48	14,774	16,050	rs6600247	Yes
IL6R	rs12133641	0.40	1.16	10.70	44	53	90,060	101,676	rs4129267	Yes
FCGR2A	rs1801274	0.49	1.15	10.16	4	6	10,692	10,692	rs1801274	Yes
GPR35	rs34236350	0.19	1.19	10.07	1	2	1	1,367	rs4676410	No - Cortes *et al.* variant has r^2 0.9 with rs34236350 but falls outside the 99% credible set and interval
NKX2-3	rs11190127	0.35	1.18	10.04	5	24	16,954	24,075	rs11190133	Yes
NPEPPS	rs4239162	0.50	1.15	9.95	113	141	247,336	247,336	rs9901869	Yes
TNFRSF1A	rs1800693	0.39	1.14	8.10	6	12	9,107	51,070	rs1860545	Yes
IL12B	rs6556416	0.31	1.14	7.61	12	14	51,413	104,349	rs6556416	Yes
IL27	rs34836	0.47	1.13	7.32	153	211	412,070	576,700	rs75301646	No - Cortes *et al.* variant not imputed. Within 99% credible interval
TYK2	rs74179925	0.21	1.16	7.07	21	42	275,321	344,221	rs35164067	Yes
CARD9	rs3923827	0.49	1.12	6.46	46	58	60,954	61,477	rs1128905	Yes
NOS2	rs2779273	0.38	1.12	6.16	45	48	37,520	37,520	rs2531875	Yes
ZMIZ1	rs1108618	0.41	1.12	5.90	19	34	34,451	34,949	rs1250550	Yes

GPR65	rs12434101	0.09	1.19	5.11	17	22	62,525	76,527	rs11624293	Yes
SH2B3	rs11065898	0.22	1.13	5.10	146	258	364,621	543,538	rs11065898	Yes
BACH2	rs9451298	0.29	1.11	4.77	10	13	27,354	27,354	rs17765610	Yes
UBE2L3	rs390387	0.19	1.16	4.56	93	100	87,614	87,614	rs2283790	Yes
IL7R	rs7717955	0.26	1.12	4.38	73	150	150,373	340,775	rs11742270	Yes
IL1R1	rs4851529	0.38	1.10	4.16	295	1,113	819,681	857,929	rs4851529	Yes
UBE2E3	rs12620692	0.43	1.10	4.00	45	144	87,466	491,492	rs12615545	Yes
ICOSLG	rs2838519	0.40	1.09	3.24	36	315	27,772	176,848	rs7282490	Yes

Table 5.1: Detailed characteristics of the 26 fine-mapping loci for AS from Cortes *et al.*[253]

5.3.4 Application to T2D association data

Having established the utility of Bayesian fine-mapping in the previous three sections, this approach can now be applied to common T2D association loci. The original publications identifying these loci were all based on imputation of HapMap 2 variants, which do not provide the resolution required for accurate fine-mapping. However, with low-pass whole-genome sequencing, high-coverage whole-exome sequencing and high density array genotyping in 2,657 individuals, the GoT2D study (described in section 2.6.1) provides an excellent panel for imputation with a larger sample size than the current Phase I release of the 1000 Genomes project. As well as its larger absolute sample size, GoT2D also contains only individuals of Northern European descent, which should result in even further increased resolution in the common and low-frequency allele spectrum compared to the 1000 Genomes panel[265]. As part of the GoT2D project, thirteen studies consisting of around 45,000 individuals (~12,000 T2D cases and ~33,000 normoglycaemic controls) were imputed into the panel derived from the 2,657 sequenced individuals (described in section 2.6.1). The Bayesian fine-mapping approach was applied to the T2D association meta-analysis results (meta-analysis results generated by Clement Ma, University of Michigan, as part of the GoT2D study), which combined the association for the GoT2D sequenced data with the thirteen imputed datasets.

Bayesian fine-mapping of these data results in 99% credible intervals and 99% credible variants sets both less than 10% of the original locus size and variant set in 21 of the 69 previously identified loci (30%; Figure 5.10; Table 5.2). Virtually all of the fine-mapped loci had originally been discovered in European ancestry populations (Table 5.2). Given the European ancestry of individuals in the GoT2D meta-analysis it is perhaps not surprising that only one of the 17 loci discovered in non-European populations – a locus near *KCNQ1* (rs2237892) with a very strong reported odds ratio (1.4) – is amenable to fine-mapping in this dataset (Table 5.2).

With the exception of the *PPARG* and *BCAR1* loci, the remaining loci that cannot be fine-mapped have non-genome-wide significant association signal in these data (Figure 5.10; Table 5.2). Since the majority of T2D association signals have been discovered in meta-analysis three to four times larger than the dataset available for the fine-mapping, and often relied on stage 2 replication through direct genotyping of most significant variants in even larger sample sets to reach genome-wide significance, this is almost certainly due to insufficient power to detect association rather than absence of signal at these loci. Given the clear reduction in number of likely variants in those loci with suggestive association ($\log_{10}(ABF)$ between 4 – 6), applying this fine-mapping approach to larger datasets imputed in high-density genotype panels will allow LD to be resolved and thus more clearly fine-map the remaining loci.

An important caveat to this analysis is the definition of successful fine-mapping. Here fine-mapping is defined in a relative sense – i.e. a reduction by 90% of the locus size and variant set – which is primarily aimed at allowing improved overlap with genomic annotations such as those generated in chapters three and four. However, when using these data to select loci for follow-up in a laboratory setting, the absolute number of variants left to study after fine-mapping will be more important than the relative reduction in variants or locus size. In that context, the *WFS1* locus deemed here to be successfully fine-mapped would not be classified as such, given that it still contains 129 variants to follow-up. Conversely, the *BCAR1* locus, still spanning more than 10% of the original 1Mb interval after fine-mapping and therefore deemed to have been unsuccessfully refined, would be considered successful since it has only 15 variants left to study.

In chapter six, the twenty one successfully fine-mapped loci will be further investigated in the context of islet regulatory annotations.

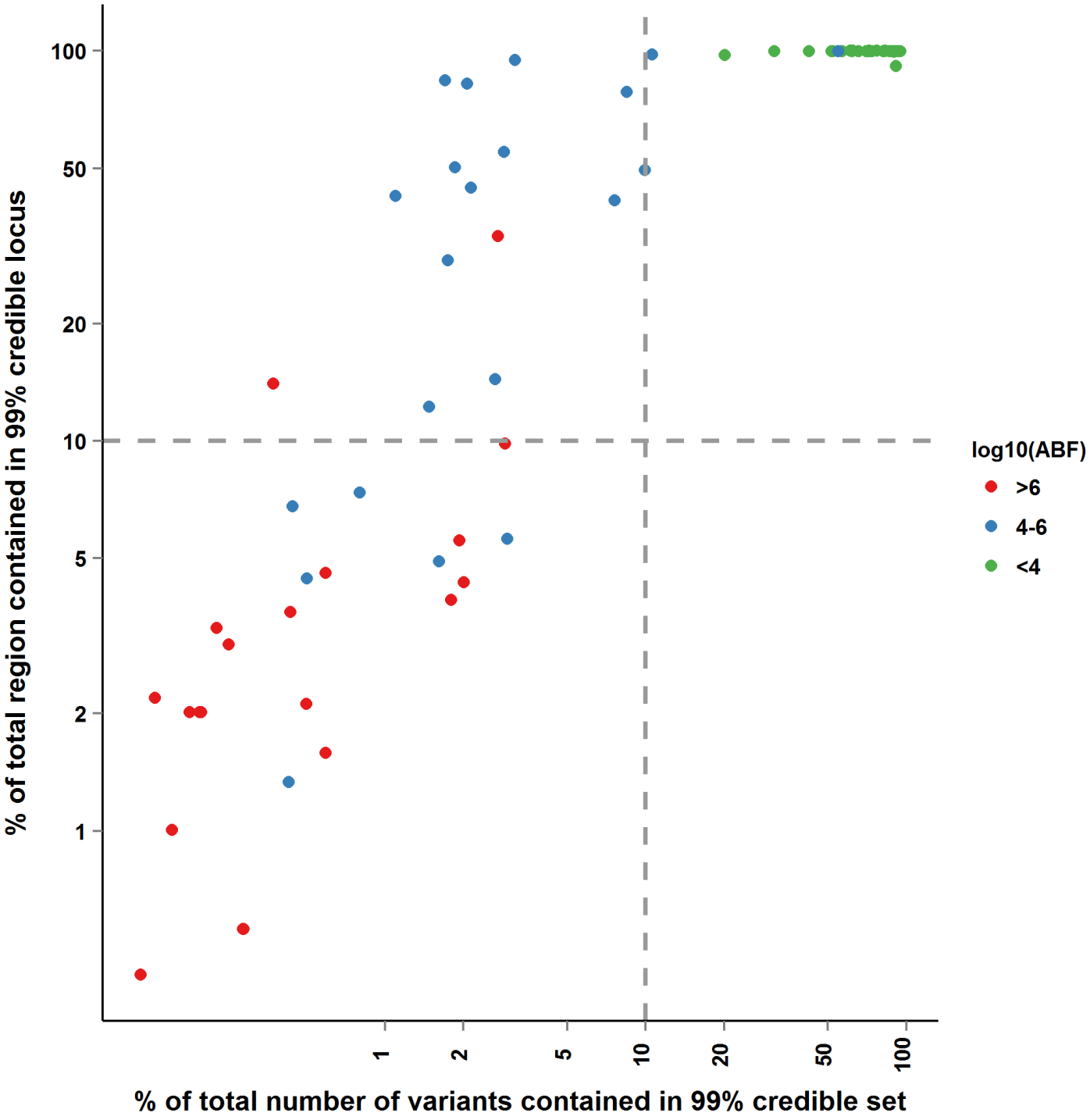


Figure 5.10: fine-mapping common T2D associated regions. Plotted is the reduction of the 1Mb region surrounding the most strongly associated variant (y-axis) against the reduction in the number of variants from all those within the 1Mb region (x-axis). Colours denote the strength of association signal in the GoT2D meta-analysis results.

Locus	Chromo-some	Index variant	RAF	OR	Population	Variants in 99% credible set (in 1Mb interval before fine-mapping)	99% credible locus (all starting from a 1Mb interval) [basepairs]	Most significant variant	BF most significant variant	Posterior probability most significant variant
TCF7L2	10	rs7903146	0.31	1.40	European	3 (2,600)	4,279	rs7903146	5.11E+87	0.90
CDKN2A/B	9	rs10811661	0.84	1.18	European	9 (3,146)	5,610	rs7018475	2.10E+10	0.63
		rs944801	0.59	1.08	European					
		rs7018475	0.26	0.74	European					
ZBED3	5	rs6878122	0.32	1.10	European	5 (3,280)	10,056	rs7708285	9.28E+06	0.25
SPRY2	13	rs1359790	0.73	1.08	European	11 (2,579)	13,340	rs1327315	1.21E+05	0.13
ZMIZ1	10	rs12571751	0.55	1.08	European	22 (3,728)	15,849	rs703969	2.60E+06	0.22
KCNQ1	11	rs231362	0.51	1.08	European	7 (3,597)	20,134	rs2237895	1.24E+09	0.67
		rs163184	0.49	1.09	European					
		rs2237892	0.65	1.40	East Asian					
		rs231361	0.25	1.09	European					
BCL11A	2	rs243088	0.47	1.07	European	14 (2,813)	21,140	rs243020	2.47E+07	0.21
CCND2	12	rs11063069	0.20	1.12	European men	4 (3,053)	21,936	rs4238013	3.04E+07	0.71
CDKAL1	6	rs7756992	0.27	1.17	European	8 (3,187)	30,073	rs7756992	3.32E+25	0.51
SLC30A8	8	rs3802177	0.71	1.14	European	6 (2,662)	33,133	rs3802177	1.94E+11	0.37
HHEX-IDE	10	rs1111875	0.55	1.11	European	12 (2,771)	36,414	rs2185756	5.85E+17	0.17
IGF2BP2	3	rs4402960	0.29	1.13	European	44 (2,452)	39,163	rs71320321	1.53E+20	0.05
FTO	16	rs9936385	0.42	1.13	European	62 (3,093)	43,402	rs11642015	3.07E+15	0.13
TLE1	9	rs2796441	0.58	1.07	European	13 (2,595)	44,370	rs2796441	9.53E+05	0.73
ADCY5	3	rs11717195	0.82	1.11	European	15 (2,536)	45,802	rs11708067	2.54E+08	0.60
KLHDC5	12	rs10842994	0.78	1.10	European	54 (3,358)	49,088	rs3751239	8.65E+05	0.35
HMGA2	12	rs2261181	0.11	1.13	European	44 (2,283)	55,590	rs10878334	1.45E+06	0.13
WFS1	4	rs4458523	0.63	1.10	European	129 (4,375)	56,075	rs1046316	7.73E+05	0.07
JAZF1	7	rs849135	0.48	1.11	European	14 (3,162)	67,869	rs1513272	4.78E+05	0.15
HNF1B	17	rs11651052	0.46	1.10	European	16 (2,005)	73,714	rs7405696	1.79E+05	0.25

IRS1	2	rs2943640	0.63	1.10	European	78 (2,706)	98,273	rs2972143	3.22E+11	0.07
PRC1	15	rs12899811	0.32	1.08	European	52 (3,520)	122,387	rs4544218	2.87E+05	0.14
BCAR1	16	rs7202877	0.90	1.12	European	15 (4,020)	140,163	rs55993634	5.85E+07	0.49
DGKB	7	rs17168486	0.17	1.11	European	102# (3,670)	144,039#	rs13220985	3.18E+04	0.05
GIPR	19	rs8108269	0.28	1.10	European women	52 (2,989)	290,510	rs4803850	1.65E+05	0.37
PPARG	3	rs1801282	0.88	1.13	European	87 (3,212)	334,953	rs11709077	4.17E+07	0.15
THADA	2	rs10203174	0.91	1.14	European	264 (3,465)	413,952	rs111239033	3.44E+04	0.03
ST6GAL1	3	rs16861329	0.75	1.08	South Asian	39 (3,546)	424,197	rs59322089	3.65E+04	0.14
UBE2E2	3	rs1496653	0.77	1.09	European	72 (3,362)	445,867	rs60079501	9.03E+04	0.11
HMG20A	15	rs7177055	0.70	1.08	European	281 (2,823)	494,957	rs12904938	2.06E+04	0.07
ARAP1	11	rs1552224	0.83	1.14	European	39 (2,097)	503,476	rs74333814	8.03E+04	0.12
DGKB	7	rs6960043	0.48	1.06	European	105 (3,856)	550,998	rs13220985	3.18E+04	0.05
C2CD4A/B	15	rs7163757	0.60	1.18	East Asian	267 (3,158)	783,449	rs8025877	2.78E+04	0.05
DUSP8	11	rs2334499	0.43	1.08	European	84 (4,061)	823,414	rs10770141	1.10E+05	0.18
CILP2	19	rs10401969	0.07	1.13	European	39 (2,297)	840,556	rs2228603	1.10E+05	0.81
NOTCH2	1	rs10923931	0.11	1.08	European	1,196 (1,310)	915,616	rs72695180	1.02E+02	0.10
TSPAN8-LGR5	12	rs7955901	0.43	1.07	European	107 (3,386)	945,974	rs4760790	2.03E+05	0.16
MC4R	18	rs12970134	0.26	1.08	European	680 (3,371)	976,490	rs12955983	2.25E+03	0.02
ANK1	8	rs516946	0.79	1.09	European	332 (3,134)	978,399	rs13262861	2.30E+04	0.25
GCC1-HNF4A	7	rs6467136	0.78	1.11	East Asian	2,114 (2,355)	996,218	rs1557644	5.95E+01	0.03
R3HDML-HNF4A	20	rs6017317	0.45	1.09	East Asian	1,481 (2,836)	996,750	rs12625671	3.71E+03	0.19
CDC123-CAMK1D	10	rs12779790	0.22	1.11	European	2,086 (3,802)	998,376	rs61848346	1.20E+04	0.55
KCNJ11	11	rs5215	0.34	1.09	European	1,219 (2,875)	998,399	rs5213	5.20E+03	0.17
TP53INP1	8	rs896854	0.46	1.06	European	2,405 (3,334)	998,624	rs13270004	7.39E+02	0.07
VPS26A	10	rs1802295	0.26	1.07	South Asian	1,048 (3,359)	998,850	rs3847343	3.46E+03	0.09
MAEA	4	rs6815464	0.58	1.13	East Asian	2,725 (3,882)	998,875	rs1466216	8.57E+02	0.06
MTNR1B	11	rs10830963	0.29	1.10	European	1,614 (2,845)	998,885	rs10830963	7.32E+03	0.44

BCL2	18	rs12454712	0.59*	1.09	Multi-ethnic	2,376 (2,719)	998,921	rs146794594	2.61E+02	0.09
PROX1	1	rs2075423	0.52	1.07	European	2,391 (3,235)	998,925	rs61818896	7.94E+02	0.08
HNF1A	12	rs7957197	0.80	1.07	European	2,120 (3,428)	998,929	rs61953351	5.82E+03	0.28
RBM43-RND3	2	rs7560163	0.89	1.33	African American	2,625 (3,206)	999,080	rs1485131	2.66E+02	0.04
KCNK16	6	rs1535500	0.43	1.08	East Asian	2,910 (3,527)	999,121	rs10305420	1.49E+03	0.25
CHCHD9	9	rs13292136	0.92	1.11	European	2,508 (2,755)	999,133	rs117810756	1.51E+02	0.07
PSMD6	3	rs831571	0.62	1.09	East Asian	2,735 (3,016)	999,157	rs114659438	1.04E+02	0.04
SRR	17	rs391300	0.74	0.28	East Asian	2,780 (2,920)	999,215	rs181959632	3.05E+01	0.03
KLF14	7	rs972283	0.55	1.07	European	1,723 (2,625)	999,222	rs117806325	2.04E+03	0.19
AP3S2	15	rs2028299	0.31	1.08	South Asian	2,506 (3,067)	999,227	rs11637921	2.39E+02	0.05
GCKR	2	rs780094	0.59	1.06	European	1,487 (1,721)	999,247	rs144828013	2.24E+02	0.11
HNF4A	20	rs4812829	0.29	1.09	South Asian	1,448 (2,801)	999,286	rs12625671	3.71E+03	0.20
PEPD	19	rs3786897	0.57	1.10	East Asian	3,606 (4,004)	999,367	rs77423258	2.12E+02	0.06
ZFAND3	6	rs9470794	0.30	1.12	East Asian	2,838 (3,015)	999,399	rs80161177	4.85E+01	0.03
GRB14	2	rs13389219	0.59	1.07	European	2,299 (2,497)	999,501	rs13024606	7.69E+01	0.05
PTPRD	9	rs17584499	0.12	1.57	East Asian	4,157 (4,443)	999,541	rs62534654	5.14E+01	0.02
ZFAND6	15	rs11634397	0.66	1.06	European	3,161 (3,495)	999,585	rs10083546	2.42E+02	0.08
GLIS3	9	rs7041847	0.52	1.10	East Asian	3,242 (4,510)	999,650	rs112270926	3.46E+03	0.26
ANKRD55	5	rs459193	0.71	1.08	European	2,680 (3,478)	999,667	rs3843467	9.38E+02	0.11
RBMS1	2	rs7593730	0.79	1.10	European	1,993 (3,198)	999,705	rs6710938	2.48E+03	0.14
GCK	7	rs4607517	0.18	1.07	European	2,086 (2,513)	999,759	rs2008428	4.97E+02	0.13
ADAMTS9	3	rs6795735	0.59	1.08	European	2,195 (3,597)	999,843	rs13096392	1.52E+03	0.10
DUSP9	X	rs5945326	0.74	1.27	European	NA	NA	NA	NA	NA

Table 5.2: fine-mapping of common T2D association loci. . Results are ordered by size of the 99% likely causal interval. The name of each locus represents closest or best regional candidate gene and labelled variant represents strongest association; neither are necessarily causal. RAF = risk allele frequency; frequency of the associated allele, based on 1000 Genomes phase 1 information for the listed population with the exception of the South Asian population where the reported RAF has been used. Those loci successfully fine-mapped are highlighted in green. OR = odds ratio; per-allele change in T2D risk. BF = approximated Bayes Factor *RAF averaged across multiple populations used in trans-ethnic meta-analysis. ♯Since the signal picked up at this locus is the primary signal for rs6960043, this locus is artificially fine-mapped by the window boundary.

5.4 Discussion

In conclusion, simulated and empirical data analyses demonstrate that common association signals can be effectively fine-mapped (both in terms of relative reduction of the locus size as well as the absolute number of variants remaining) using a Bayesian approach and, provided there is a shared causal allele, multi-ethnic study designs. Such trans-ethnic studies have a modest beneficial effect on detecting the causal variant by significance alone, but add considerable resolution to fine-mapping when using Bayesian fine-mapping – nearly halving the credible sets and intervals found using single population GWAS. When applied to T2D association meta-analysis data, this approach results in a reduction of the association signal in locus size and variant set of > 90% in 30% of all reported loci (described in section 5.3.4).

Knowing the causal variant(s) at a given locus will be crucial in gaining broader mechanistic insights into T2D pathogenesis. Without it enrichment for transcription factor binding sites or other regulatory information important for pathway analyses will primarily need to be determined in broad regions, leading to the inclusion of many false-positive findings. For stronger association signals in particular, fine-mapping can narrow the set of variants for interrogation down to a number which is tractable in a laboratory setting. However, even less strong signals are often reduced to less than 30 variants and/or genomic intervals <50kb. This will greatly reduce the number of genomic annotations overlapping the association signal and help to prioritise specific signals, as well as aid the selection of appropriate functional assays for follow-up.

The reduced number of variants to be tested after fine-mapping should allow better calibrated investigation of enrichment across functional elements and pathways than previous analyses, which relied mainly on the lead variant within a signal and all variants in close LD. Since enrichment analyses are probabilistic, and as such allow for a certain degree of error in the assignment of causation, this means they primarily gain statistical power

through fine-mapping. However, for functional studies, inclusion of the causal variant at a given locus is critical, rendering adequate fine-mapping much more important. The continuing development of high-throughput screening systems means that functional follow-up of variant sets of the size resulting from fine-mapping will be feasible in the near future.

With more and more genomic data becoming available through efforts such as ENCODE[111], the NIH Roadmap Epigenomics[266], and the Genotype-Tissue Expression project (GTEx)[267], as well as smaller projects focused on specific tissues, it will become increasingly possible to look for general enrichment characteristics of association signals. These analyses will show whether association signals tend to affect tissue-specific elements or act primarily on functional elements present in many tissues. It will also uncover which types of annotation are often found in association intervals, and, conversely, which types are depleted for signal. The advantage of fine-mapping in a Bayesian framework is that, when such prior information from external data resources becomes clear, it can be included to give greater weight to variant annotations. This, in turn, should allow for even more fine-tuned refinement of association signals.

In chapter six, I will determine whether the refined T2D loci overlap with the islet annotations generated in chapters three and four, or others which are publicly available[70,71].

Chapter 6

Intersecting islet annotations with T2D association data

6.1 Introduction

Identification of causal variant or mechanisms underlying T2D association signals identified through genome-wide association studies (GWAS) has not kept pace with the discovery of new associated loci. The ability to identify such causal variants and mechanisms relies partly on refining association loci to more manageable genomic intervals or sets of variants – the utility of statistical fine-mapping approaches was explored in chapter five and resulted in refinement for twenty one T2D associated loci.

Another major limitation in interpreting GWAS findings has been the absence of high quality annotations for a large proportion of the non-coding genome in which association loci reside. Efforts such as the ENCODE[111] and NIH Roadmap Epigenomics[266] projects have now provided detailed annotations for many different cell types and tissues, but not including most of the tissues relevant for T2D pathogenesis. Analysis of T2D association loci in context of their effects on beta-cell function (as measured by HOMA-B), and the prevalence of such loci near genes involved in monogenic forms of beta-cell dysfunction, suggests that beta-cells play a pivotal role in mediating common T2D association signals[56,57]. The work presented in chapters three and four, together with recent efforts to define functional islet annotations by other groups including ourselves[143,268], now enable integration of T2D association signals with many such islet annotations.

There are many lines of evidence suggesting that information on functional elements in appropriate cell types will lead to a better understanding of association signals – both in terms of enrichment characteristics of all GWAS loci combined, and the ability to identify causal transcripts or variant(s) at specific association signals. Enrichment of T2D associated variants across functional elements helps uncover important information about the mechanistic architecture of T2D. Large scale studies performed in non-T2D-relevant cell lines (and sometimes tissues), have identified general enrichment of variants present in the GWAS catalogue amongst the top-ranking expression quantitative locus (eQTL) variants[207],

as well as variants present in functional chromatin[65]. Whilst no eQTL studies have currently been published in human islets, recent efforts to define islet regulatory elements have identified enrichment of T2D associated variants amongst active enhancer elements[143,268].

Approaches combining different types of regulatory annotation have also been instrumental in finding causal mechanisms. Identification of such mechanisms will improve the power of enrichment analyses to detect shared properties between association loci, and will, for example, help to justify drug discovery efforts for those genes involved. eQTL information has been instrumental in identifying *KLF14* as the causal gene underlying the T2D[57] and high-density lipoprotein cholesterol[66] associated locus on chromosome 7q32.3[109], and in uncovering the network of genes regulated by the encoded transcription factor[110]. Functional elements discovered through sequencing approaches have been the means by which the likely causal molecular mechanism – altered islet enhancer activity – at the *TCF7L2* locus has been discovered[70,71], although this enhancer has still to be definitively linked to *TCF7L2* expression itself.

With a large quantity of islet annotations now available, many different approaches attempting to integrate these annotations with T2D genetics data are underway. Whilst I have worked on several projects generating different islet annotations (miRNAs [chapter three][269], islet long non-coding RNAs (lncRNAs)[209], eQTL and ASE variants [chapter four], and islet regulatory elements[143]), the work in this chapter will concentrate on those I characterised in most detail – the islet non-coding RNA annotations. Islet miRNA annotations were generated as described in chapter three, and lncRNAs expressed in human islets were identified in collaboration with Prof. Jorge Ferrer using a combination of RNA-seq and ChIP-seq (annotations generated by Ignasi Morán [Barcelona, Spain])[209]. Using these annotations, I will firstly use enrichment analysis to explore whether islet regulatory RNAs might mediate common T2D association signal. Secondly, making use of my approval to work within the context of whole-genome sequencing data generated as part of the GoT2D study, I will look across the allele frequency spectrum for evidence of specific islet lncRNAs

and miRNA involvement in T2D risk. Finally, as proof-of-principle for the utility of integrating different islet annotations and fine-mapping approaches to determine likely causal mechanisms, I will extend the types of islet annotations studied and explore the overlap between these annotations and the twenty one fine-mapped T2D association loci from section 5.3.4.

6.2 Methods

6.2.1 Enrichment of islet lncRNAs and miRNA targets in GWAS data

First I asked whether islet regulatory RNAs were involved in mediating T2D association. MAGENTA[270] was used to determine the enrichment of T2D association signal in large-scale genome-wide association meta-analysis data[57] across islet miRNA targets and lncRNAs. MAGENTA uses the per variant association *P*-values in meta-analysis summary statistics to obtain calibrated *P*-values of per-transcript evidence for trait association, correcting for differences in transcript size, number of variants, and patterns of linkage disequilibrium which would otherwise confound enrichment analyses.

Annotation type	Data used	Data source
Islet miRNAs	miRBase v19[150] coordinates	http://www.mirbase.org
Islet lncRNAs	Morán *et al.* coordinates[209]	http://www.betacellregulation.net/images/stories/lab/data
Predicted islet miRNA targets	Coordinates of Targetscan 6.2[142] conserved predictions	http://www.targetscan.org
	Coordinates of miRanda predictions with good mirSVR scores (release August 2010)[271]	http://www.microrna.org/microrna/getDownloads.do
	Coordinates of miRDB (version 4)[272] predictions	http://mirdb.org/miRDB/download.html

Table 6.1: Details of miRNA and lncRNAs annotations used in the MAGENTA enrichment analyses.

BEDTools version 2.16.2[141] (described in section 2.7.1) was used to identify variants overlapping islet miRNA and lncRNA annotations (Table 6.1). Since the available association data was based on human genome hg18, genomic coordinates were, where appropriate, converted to this genome build using the UCSC liftOver tool with "*-minMatch = 1*".

For MAGENTA analysis, all islet lncRNAs (n=1,128) were tested as a single group, and split into two subgroups based on their distance from the nearest annotated feature ("antisense only" [n=761] and "intergenic only" [n=367] – corresponding to those lncRNAs located < 1kb from and antisense to know transcripts, and those > 1kb from any annotated coding feature respectively). Sets of predicted mRNA targets for each individual miRNA were constructed from all three prediction algorithms (TargetScan 6.2, miRanda, miRDB).

Trait	Consortium	Data Source
T2D[57]	DIAGRAM (Diabetes Genetics Replication and Meta-analysis) Consortium	Available through collaboration
Fasting glucose levels[55]	MAGIC (Meta-Analyses of Glucose and Insulin-related traits) Consortium	ftp://ftp.sanger.ac.uk/pub/magic/
HOMA-B[55]		
HOMA-IR[55]		
Body mass index[273]	GIANT (Genetic Investigation of ANthropometric Traits) Consortium	http://www.broadinstitute.org/collaboration/giant/index.php/GIANT_consortium_data_files
Waist-hip ratio[274]		
Height[259]		
High-density lipoprotein[66]	Global Lipids Genetics Consortium	http://www.sph.umich.edu/csg/abecasis/public/lipids2010

Table 6.2: Details of GWAS datasets used in MAGENTA enrichment analyses.

Islet miRNA target sets were compared against one another, whilst enrichment for islet lncRNAs was compared between T2D and other traits. For islet lncRNAs, therefore, similar GWAS data on fasting glucose levels, HOMA-B, HOMA-IR[55], height[259], waist-hip ratio[274],

high-density lipoprotein[66] and body mass index [273] were included (Table 6.2). The choice of traits was defined by those for which summary genome wide association data were available either through collaboration or via public domain deposition.

For these miRNA target sets, MAGENTA was run using pre-defined settings using 100,000 permutations of the *P*-value calculation, and the 75th percentile cut-off *P*-values for establishing enrichment. To correct for multiple testing, *q*-values, which represent the expected proportion of false positives incurred when calling a particular gene set significant, were calculated using the *qvalue* command from the package "*qvalue*" in R version 2.15.1[147] with the setting "*robust = TRUE*". Those gene sets with *P*-values <0.01 and *q*-values <0.1 were deemed significant.

Since MAGENTA by default compares the association signal in the set of interest against a null distribution formed from coding genes, the program was modified to instead compare the lncRNA set against random fragments of the mappable genome. For each of the lncRNA sets (all islet lncRNAs, antisense only, intergenic only), three independent sets of 100,000 random fragments, matched for the respective distributions of lncRNA size and distance from protein-coding genes, were generated using BEDTools version 2.16.1[141] and custom Perl scripts. For each of the lncRNA sets, an enrichment *P*-value was computed by comparing the enrichment score for the set of lncRNAs to that from a null distribution of 10,000 randomly sampled sets of random genome fragments (each set containing the same number of fragments as present in the respective lncRNA test set). The 95th percentile of all association *P*-values, which were corrected for size, number of variants, linkage disequilibrium properties and distance from protein-coding genes, was used as the lncRNA enrichment cut-off.

6.2.2 Variant testing of islet lncRNAs, miRNAs and miRNA targets in whole-genome sequencing data

To determine whether individual islet regulatory RNAs showed evidence for a role in T2D susceptibility, I looked at annotations in light of variation identified by GoT2D whole-genome sequencing data. Islet miRNA precursor and predicted target sequences and islet lncRNAs were acquired from the sources described in the previous section (Table 6.1). Only miRNA targets predicted by the algorithm with medium stringency (TargetScan 6.2[142]) were used, representing a compromise between inclusion of too much neutral variation and exclusion of too much true biology. Exonic coordinates of the islet lncRNAs, available as described in the previous section, were converted to the current human genome build (hg19) using the UCSC liftOver tool with "*-minMatch = 0.95*". All 1,573 islet lncRNA exons converted successfully – four exons contained one mismatch and one exon contained two mismatches between the builds.

Rare variant (MAF < 0.01) testing was performed on GoT2D whole-genome sequencing data (described in section 2.6.1) using the optimised version of the SNP-set (Sequence) Kernel Association Test (SKAT-O) test[275] implemented in the EPACTS (Efficient and Parallelizable Association Container Toolbox) pipeline (available from http://www.sph.umich.edu/csg/kang/epacts/). Annotations (islet lncRNAs, islet miRNAs and Targetscan 6.2 predicted target genes of islet miRNAs – described in Table 6.2) were only considered for testing where they overlapped more than one variant with a minor allele frequency (MAF) < 0.01. Gender, the first two principal components derived from the array genotyping data (to account for population structure) and a date-based filter (to account for study-specific technical bias caused by a change in sequencing technology) were included as covariates in the analysis.

6.2.3 Identify causal variants at fine-mapped loci using islet annotations

To identify potential causal mechanisms at the twenty one common T2D association loci fine-mapped in chapter five, the 99% credible intervals and variants were intersected with islet annotations (Figure 6.1). Effects of fine-mapped variants on genes, transcripts and

protein sequence were annotated using the Variant Effect Predictor tool[145] and Ensembl release 74 as described in section 2.7.2. Overlap between variants and islet regulatory data (Figure 6.1 and section 2.7.1) was determined using BEDTools version 2.16.2[141] as described in section 2.7.1.

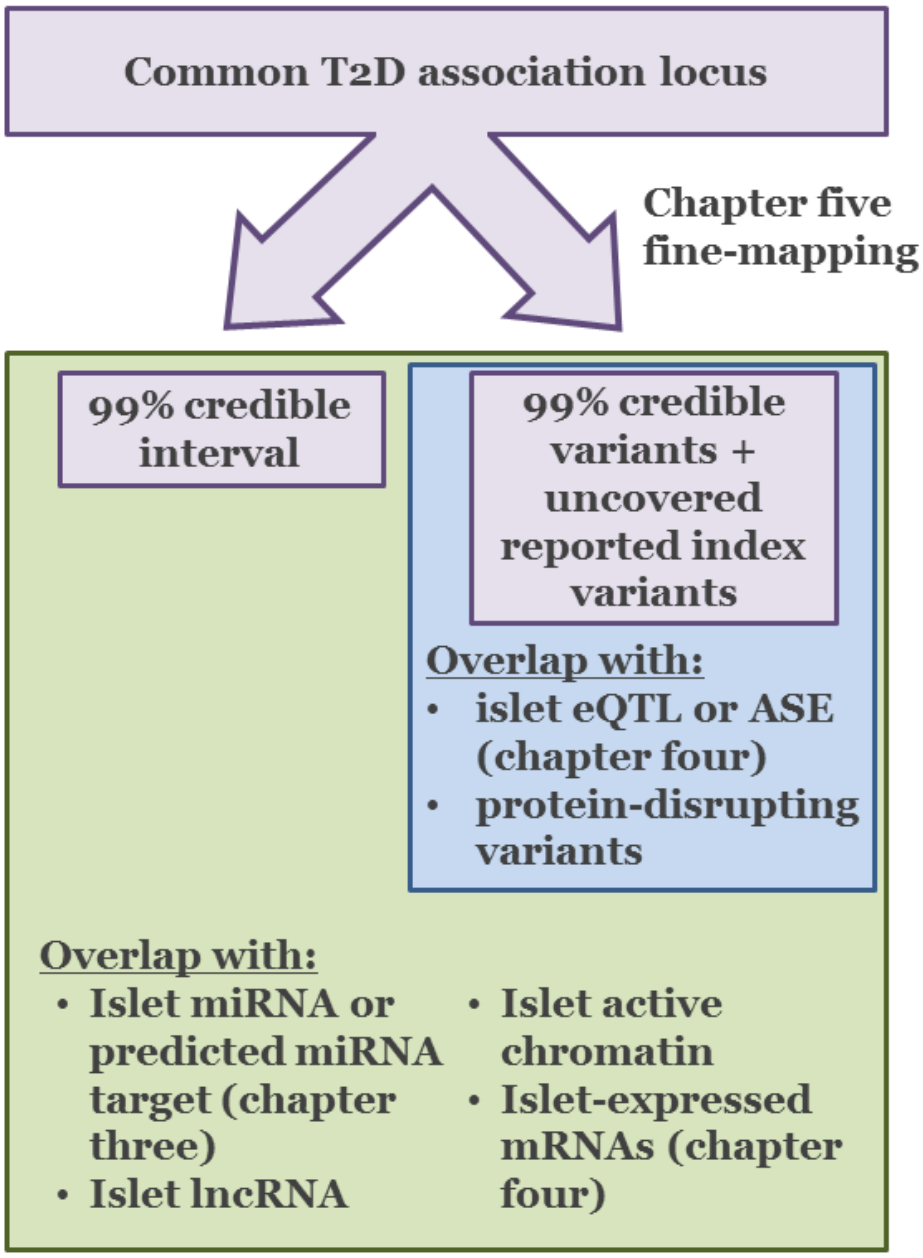


Figure 6.1: Schematic representation of the approach to identifying potential causal mechanisms at fine-mapped T2D association loci. A potential causal mechanism deemed to have been identified when the 99% credible interval overlapped islet annotations and the 99% credible variants also overlapped nominally significant islet eQTLs, showed significant ASE, or were annotated as protein-disrupting variants.

6.3 Results

6.3.1 Enrichment of common T2D association in islet annotations

Having established the miRNA profile of human islets in chapter three, and, through collaboration with Prof. Jorge Ferrer's group in Barcelona, identified novel islet lncRNAs[209], it was possible to investigate global effects of such annotations on common T2D association signals. Several of these islet lncRNAs were shown to be functionally active in human islets[209], and were found near loci implicated in monogenic syndromes of beta-cell dysfunction disorders as well as near common T2D association loci[209]. For example, islet antisense lncRNAs were found within loci in which mutations are known to cause neonatal diabetes (*ABCC8*/*KCNJ11*)[33,34], pancreas agenesis (*GATA6*)[276] and MODY (*HNF1A*)[51].

To evaluate the potential involvement of the newly identified islet lncRNAs in T2D association, I used MAGENTA to test whether the set of islet lncRNAs (n=1,128) were enriched for trait associations in genome-wide data compared to randomly-selected non-coding sequence. The analysis was applied to available data sets for T2D[57] and related continuous glycaemic measures[55], as well as 5 non-islet related control phenotypes[259,273,274] (Table 6.2). The results demonstrate that islet lncRNA transcripts are, as a group, nominally enriched for association with both T2D (*P*-value =0.02) and fasting glucose (*P*-value =0.01), with enrichment for the control phenotypes height (*P*-value =0.02) and waist:hip ratio (*P*-value =0.001 – Table 6.3). This enrichment in association signal for anthropometric traits might seem surprising, but it is worth remembering that only 45% of these islet-expressed lncRNAs show an islet-specific expression patterns (when compared to sixteen other human tissues)[209]. Furthermore, the significantly associated loci for waist:hip ratio and height implicate developmental processes in the genetic architecture underlying association for both traits[259,274]. Numerous islet lncRNAs were upregulated during differentiation using an

endocrine differentiation protocol[209], suggesting that such lncRNAs could well play a role in developmental processes (including those in other tissues).

Trait	All islet lncRNAs (n=1,128)	Intergenic islet lncRNAs (n=367)	Antisense islet lncRNAs (n=761)
Type 2 Diabetes	0.02	0.03	0.16
Fasting Glucose	0.01	0.003	0.20
HOMA-B	0.29	0.53	0.19
HOMA-IR	0.52	0.46	0.44
Body Mass Index	0.24	0.46	0.07
Height	0.02	0.06	0.04
Waist-Hip Ratio	0.001	0.26	0.004

Table 6.3: MAGENTA enrichment analysis of islet lncRNAs in genetic association data. Enrichment *P*-values for each of the lncRNAs sets compared to matched genomic regions in the different datasets. *P*-values are displayed as averages of three randomly generated genomic backgrounds, with the 95th percentile of all observed association scores used as enrichment cut-off.

Antisense lncRNAs are thought to primarily regulate the transcription of nearby transcripts in *cis*[277-279]. Intergenic lncRNAs are a rather more heterogeneous group, with, for example, *HOTAIR* interacting with proteins to regulate transcription of the *HOXD* locus in *trans*[280] and another well-known example, *XIST*, involved in X chromosome inactivation and localisation[281,282]. Islet lncRNAs were therefore split by type ("intergenic" versus "antisense") and each group examined for enrichment independently. Enrichment of association signal was identified for intergenic lncRNAs for T2D and fasting glucose (T2D P = 0.03; fasting glucose P =0.003 – Table 6.3). This result is consistent with the characteristics of intergenic lncRNAs, which show significantly more islet-specific expression (55%) than antisense lncRNAs (40% – Fisher's P-value = 7.5×10^{-6}).

Futhermore, islet intergenic lncRNAs have previously been shown to have functional relevance. For example, two such lncRNAs show glucose-dependent expression patterns[209]. Although these expression changes might be reactive, functional studies of another intergenic lncRNA, *HI-LNC25* – a multi-exonic transcript located in a 1.6Mb gene desert on

chromosome 20 – provide evidence to link lncRNA expression changes with biological mechanisms. Through an RNAi-screen in the human beta-cell line EndoC-β-H1[183], *GLIS3*, an important islet transcription factor located on chromosome 9, was identified as a potential regulatory target of *HI-LNC25*[209]. Knockdown experiments of *HI-LNC25* identified a specific and reproducible concomitant reduction of *GLIS3* mRNA levels, an effect which was sustained during several days in culture[209]. The functional relevance of the intergenic islet lncRNAs, combined with their enrichment of T2D association signal, makes them excellent candidates to play a role in mediating T2D susceptibility.

In a similar vein to the lncRNAs enrichment analysis, MAGENTA was used to determine whether the predicted target gene sets of islet-expressed miRNAs were enriched for evidence of association with T2D. Three alternative target prediction algorithms (TargetScan, miRanda and miRDB)[142,271,272] – which often differed in target genes predictions for individual miRNAs – were used to determine the robustness of the predictions. The degree of T2D association signal enrichment was determined in the target genes predicted by each method alone and by their overlap (Figure 6.2). This analysis identified significant enrichment (*P*-values <0.01 and *q*-values <0.1) across islet miRNA targeted gene sets with nearly all of the (combinations of) prediction algorithms used (Appendix Table 3). When treating each algorithm individually, none of the target gene sets predicted by the most stringent algorithm (miRDB) showed enrichment of T2D association signal, whereas nine target sets predicted by the medium stringency TargetScan, and fifty one of those predicted by the least conservative MiRanda algorithm, were significantly enriched (Figure 6.2; Appendix Table 3). In the set of targets predicted by all three different algorithms, significant enrichment was also found across the predicted target genes for three islet-expressed miRNAs (miR-17-5p, *P*-value = 0.004, *q*-value = 0.09; miR-93-5p, *P*-value = 0.005, *q*-value = 0.09; miR-20a-5p, *P*-value = 0.009, *q*-value = 0.09 – Figure 6.2; Appendix Table 3). The target gene sets for these three miRNAs show near complete overlap with one another (miR-93-5p has 465 such predicted targets whereas miR-17-5p and miR-20a-5p have 476). All three miRNAs are part of the highly conserved miR-17~92 family, transcribed from three

miRNA clusters on chromosomes 7, 13, and X[283]. Sequence conservation analysis of *mir-17*, *mir-93* and *mir-20* shows that they are ancient paralogues[283] and helps to explain the shared target predictions. As discussed in section 3.1.2, the miR-17~92 family are known to act as oncogenes[162] and are involved in normal vertebrate differentiation and embryogenesis[284,285]. Given that the miR-17~92 family is involved in cell cycle regulation, which is thought to be a key process underlying T2D association, this could well be the mechanism through which these miRNAs are underpinning the observed enrichment of T2D association signal.

Enrichment analyses show that islet lncRNAs, as well as several islet miRNA target gene sets, are globally enriched for T2D association signal. Functional evidence also suggests that islet lncRNAs and the targets of enriched islet miRNA gene sets play biologically relevant roles in processes such as cellular differentiation. The work presented in this chapter therefore highlight new classes of genomic elements which should be interrogated to dissect the functional aetiology of T2D susceptibility.

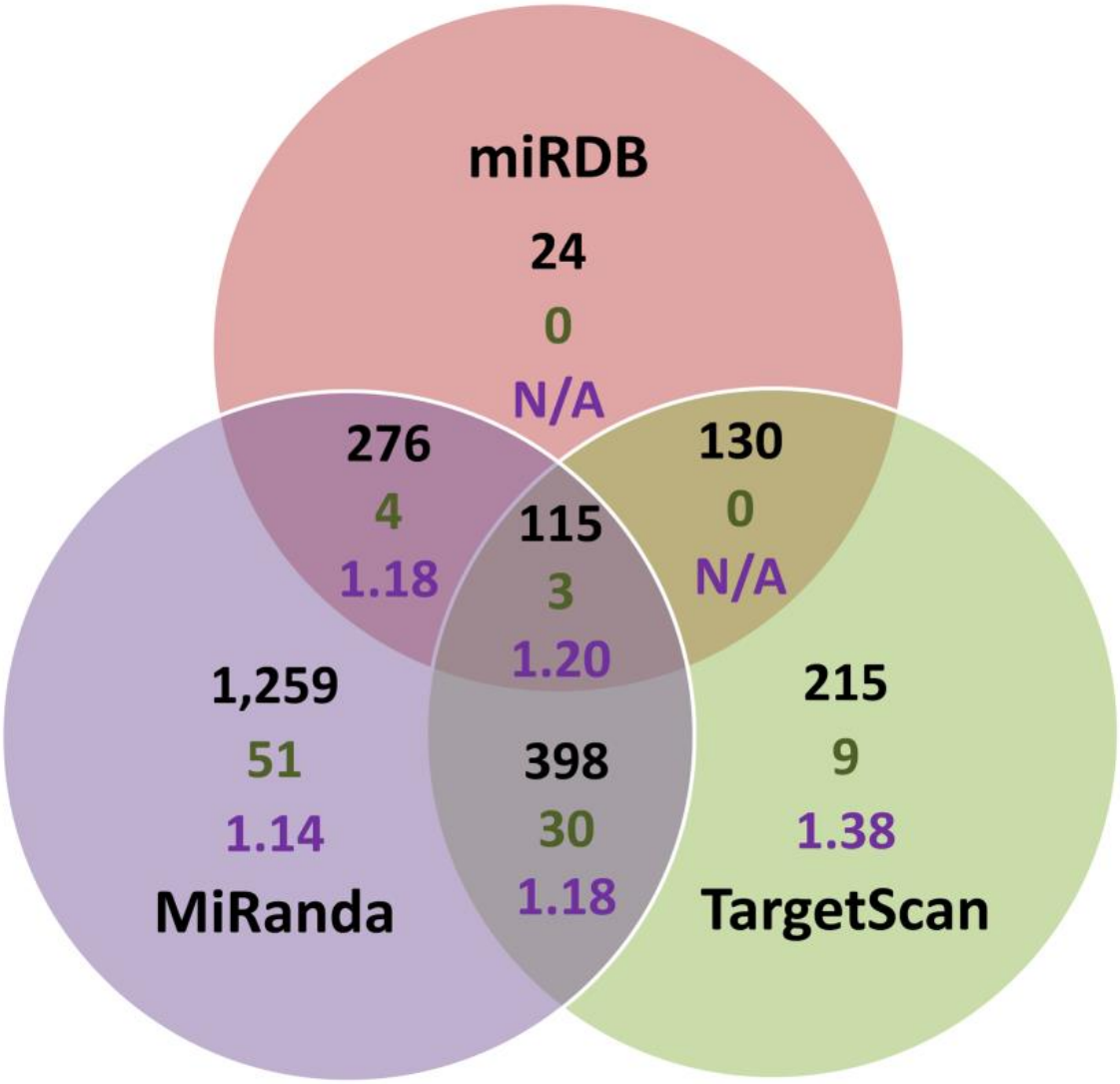


Figure 6.2: Results for the T2D association signal enrichment analysis in miRNA target gene sets. The Venn diagram represents the sets of miRNA target genes for each miRNA as predicted by miRDB, miRanda and TargetScan alone, or the overlap of these methods. Annotated are the median number of genes in each set (black text), the number of significantly (P <0.001, q-value <0.1) enriched gene sets (green text), and the median enrichment of the significantly enriched gene sets (purple text).

6.3.2 Investigating T2D association of variation in islet regulatory RNA annotations across the allele-frequency spectrum

Given the enrichment of common T2D signal across miRNA targets and lncRNAs, the localisation of islet lncRNAs close to T2D association signals and genes implicated in monogenic beta-cell dysfunction, and the important role of miRNAs in gene regulation, rare variation in these regulatory RNAs could be involved in T2D susceptibility. The whole-genome sequencing of 1,326 T2D cases and 1,331 controls as part of the GoT2D study provided a first opportunity to explore this possibility.

In the whole-genome sequencing data, 222 variants were found to overlap 99 precursors for miRNA expressed in human islets or beta-cells. Only nine of those 222 variants overlapped the mature miRNA sequence itself (in miR-323a-3p, miR-369-5p, miR-628-5p and miR-770-5p), the others lying in the rest of the miRNA hairpin. The 39,103 predicted target sites for islet expressed miRNAs contained 1,516 variants, whereas 1,059 of the 1,128 islet lncRNAs overlapped 32,713 variants. Nearly all (95%) of the observed variation across miRNA precursors was rare (minor allele frequency < 0.01). miRNA target sites also showed a high fraction of rare variation (86%), with lncRNAs containing the smallest proportion of rare variants (71%). This corroborates published data on the conservation of these types of sequence, showing extensive conservation of miRNA hairpin sequence[286,287], with miRNA targets sites and islet lncRNAs being more conserved than random 3'-UTR sequence and non-coding sequence respectively[209,288].

To determine whether any of the 34,451 variants overlapping miRNA and lncRNAs showed evidence of single variant association, statistics for each were extracted from the GoT2D meta-analysis data described in section 5.3.4 (which includes only variants with an observed minor allele count amenable to single variant association testing [> 3 counts]). Common variation overlapping islet lncRNAs and predicted miRNA target sequence showed a clear inflation from the expected *P*-value distribution across all variants (Figure 6.3), which

is in line with the T2D association signal enrichment for these types of annotation identified in section 6.3.2. In the low frequency and rare allele frequency spectrum, the association results show little deviation from the expected (Figure 6.3). The miRNAs themselves are particularly devoid of any signal, with no single variant reaching nominal significance (*P*-value < 0.01).

Although the most significant common association signals for both islet lncRNAs and miRNA targets are not genome-wide significant, they both have a *P*-value < $5x10^{-5}$ – making them good candidates for replication. When looking at these two common variant loci in more detail, the association signal overlapping the lncRNAs *HI-LNC1086* remains of interest, whereas the signal overlapping the miRNA target (for miR-128 at *KLHL18*) is part of a very large LD block overlapping many genes, making it very hard to determine whether this overlap is of functional relevance (Figure 6.4). *HI-LNC1086* is likely to be a splice variant of *COX10-AS1*, and widely expressed[209]. The nearby gene *COX10* encodes an important nuclear factor required for the biogenesis of the final enzyme (complex IV) in the mitochondrial respiratory chain[289]. Mutations in *COX10* result in complex, often fatal, disorders affecting many different organ systems, and reported symptoms include hypoglycaemia[290-293]. The coding sequence of *COX10* contains no genome-wide significant variation associated with T2D or measures of glycaemia[294]. However, potential regulatory effects of this antisense transcript have not been studied and could therefore make this lncRNAs an excellent candidate for involvement in T2D pathogenesis.

Next, the contribution of combined rare variation per miRNA precursor, miRNA target sites and lncRNAs as a group to T2D susceptibility was tested. Predicted miRNA target sites were grouped by TargetScan miRNA family (n=103 with overlapping variants). Since testing for an excess of rare variants in sequence features requires more than one observed variant per group, 34 miRNA precursors, 1 miRNA family and 26 lncRNAs were excluded from the analysis. Results of these analyses show that none of the individual groups tested reach a *P*-value withstanding any kind of multiple testing correction (Figure 6.5A,C,E). Furthermore,

the *P*-value distribution for none of the three different classes (miRNAs, predicted miRNA targets, and lncRNAs) deviates from the expected (Figure 6.5A,C,E). The number of variants per group in the miRNA target and lncRNAs classes (Figure 6.5D,F) should make this test well powered to detect aggregation of rare variation, provided the ratio of causal to neutral variants is high and these variants exert an appreciable effect size. However, miRNA transcripts contain a median of only two variants (and four at maximum – Figure 6.5B), which is too few for these types of tests to exclude a potential role for miRNA transcripts.

As part of the GoT2D and T2D-GENES studies (described in section 2.6), whole exome sequencing was performed on 6,447 individuals with T2D and6,346 normoglycaemic controls. Whilst the vast majority of the sequence captured was protein-coding, several arbitrarily selected miRNA transcripts were also included. As such, results from the original whole-genome sequenced samples could be compared against results from this almost fivefold larger dataset for 46 miRNA transcripts. Given the inclusion of different ancestry groups and higher coverage in the whole-exome sequencing compared to the whole-genome data, a greater than linear increase of rare variants to sample size might be expected. However, the number of rare variants in fact exhibited a less than linear increase to sample size (fivefold increase in sample size versus a median 2.5 fold increase in rare variation – Figure 6.6A). This lower-than-expected increase in rare variation is likely to reflect the strong conservation of this type of sequence.

There is very little correlation between the gene-based results (which are nearly all non-significant in both analyses) for all but one miRNA transcript (Figure 6.6B). This miRNA transcript (*mir-152*) has only two variants in the GoT2D data, but six in the 13,000 exomes, and the *P*-value decreases from 0.05 to 0.02 between these two datasets. Whilst *mir-152* is not significantly associated with T2D susceptibility when accounting for multiple testing, the results show clear consistency between the two datasets. TargetScan 6.2 predicts conserved targets for the mature miRNA (miR-152) derived from this transcript at 698 genes. miR-152 has been shown to be involved in cell cycle control[295], and its expression is often silenced by

DNA hypermethylation in cancer[295,296]. Expression levels were also found to be significantly upregulated in serum of T1D patients compared to non-diabetic controls[297], indicating that this miRNA might be glucose-dependent. Whilst these results are not in any way conclusive, miR-152 may therefore be of interest for further functional follow-up.

In conclusion, the analyses in this section do not uncover evidence for widespread involvement of rare variation in islet lncRNAs or miRNA-related annotations in T2D susceptibility. However they do identify several potentially interesting genes for follow-up, which includes suggestive evidence of T2D association for common variation in the lncRNA *HI-LNC1086* near *COX10* and consistent nominally significant association with rare variation across *mir-152*. These regulatory RNAs may therefore represent novel transcripts involved in T2D pathogenesis, and serve to highlight the importance of these types of annotations in uncovering such potential new biology.

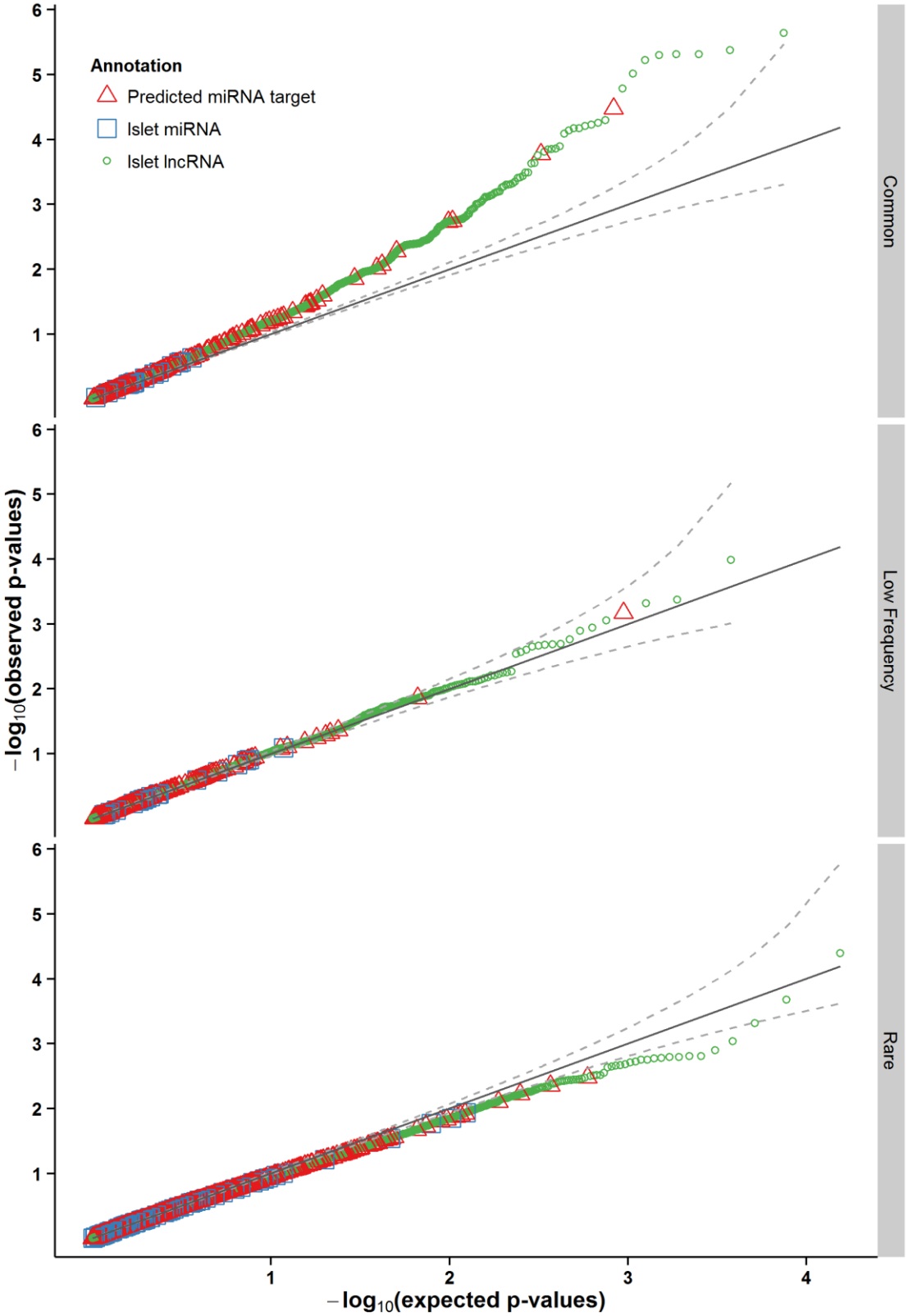


Figure 6.3: Single variant association results for variants overlapping islet miRNAs and lncRNAs annotations. Quantile-quantile plots of the expected *P*-value distribution (x-axis) versus the observed *P*-value distribution (y-axis) split by allele frequency along the vertical. The three different annotation categories are denoted as green open circles (islet lncRNAs), red open triangles (islet miRNA targets) and blue open squares (islet miRNAs).

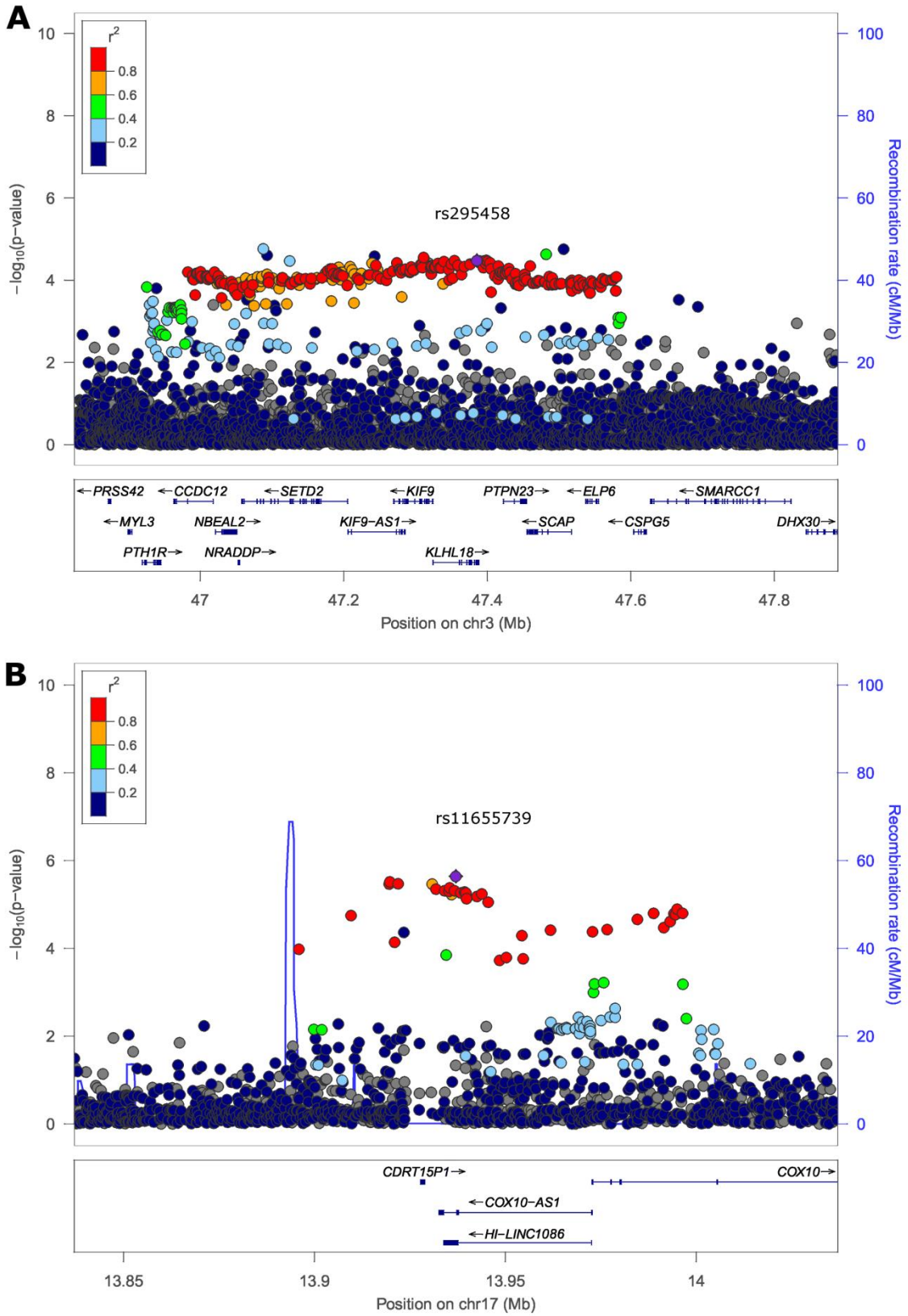


Figure 6.4: Detailed regional plots of the most significant common islet lncRNAs and miRNA target variants. The panels show T2D association *P*-values (left y-axis) in the GoT2D meta-analysis data for all analysed variant near rs295458 overlapping a predicted binding site for miR-128 in the *KLHL18* 3'-UTR (A), and near rs11655739 overlapping the islet lncRNAs *HI-LINC1086* (both variants denoted as purple triangles). The colours indicate LD relationships. Gene annotations are present at the bottom of each plot.

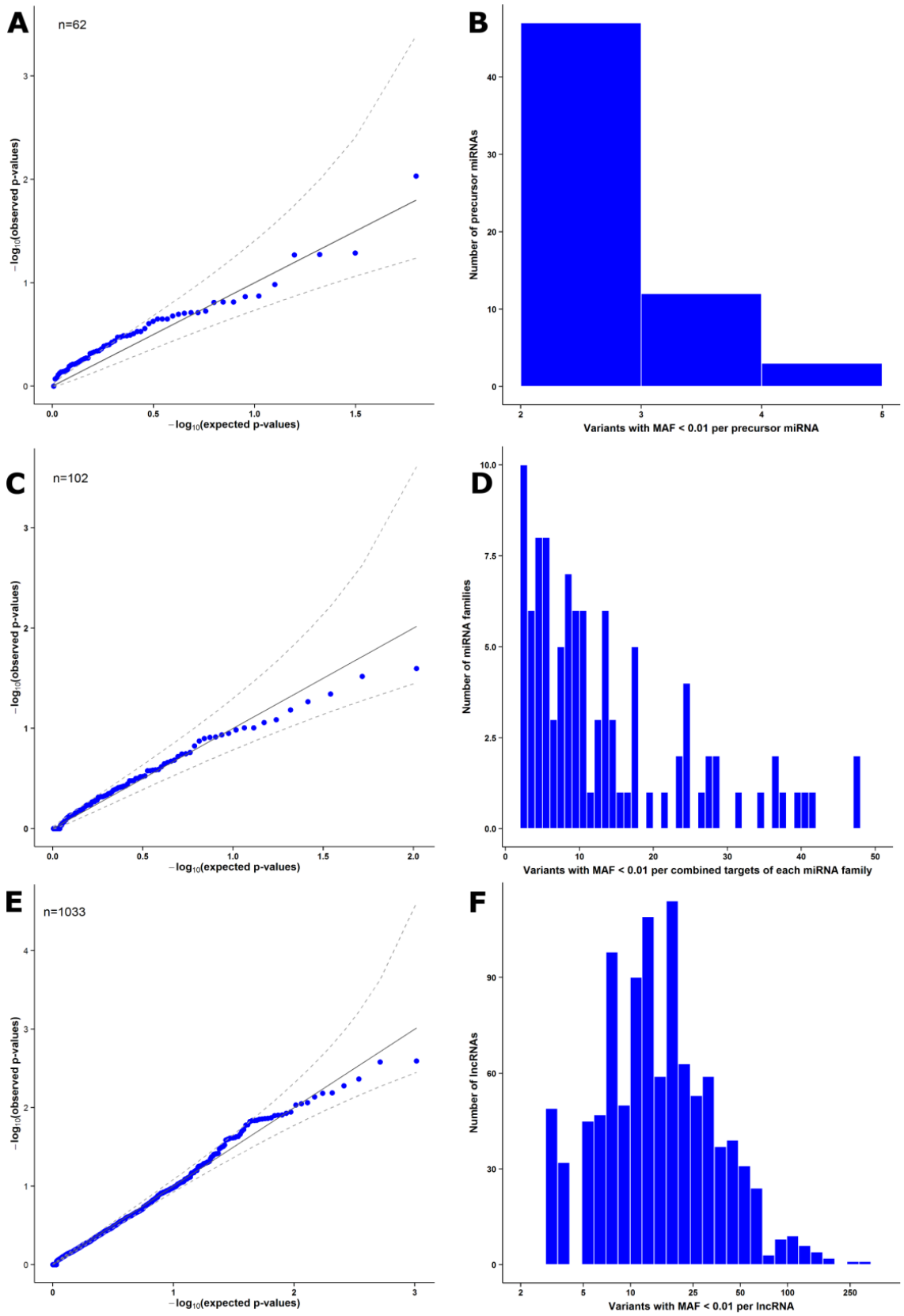


Figure 6.5: Rare variant grouping tests of islet miRNAs and lncRNAs. For islet miRNA precursors (A,B) and their predicted miRNA targets grouped by miRNA family (C,D), as well as islet lncRNAs (E,F), the left hand panel depicts the quantile-quantile plot of the expected *P*-value distribution (x-axis) versus the observed *P*-value distribution (y-axis). The right hand panels show the histograms of the number of rare variants per tested element for that class.

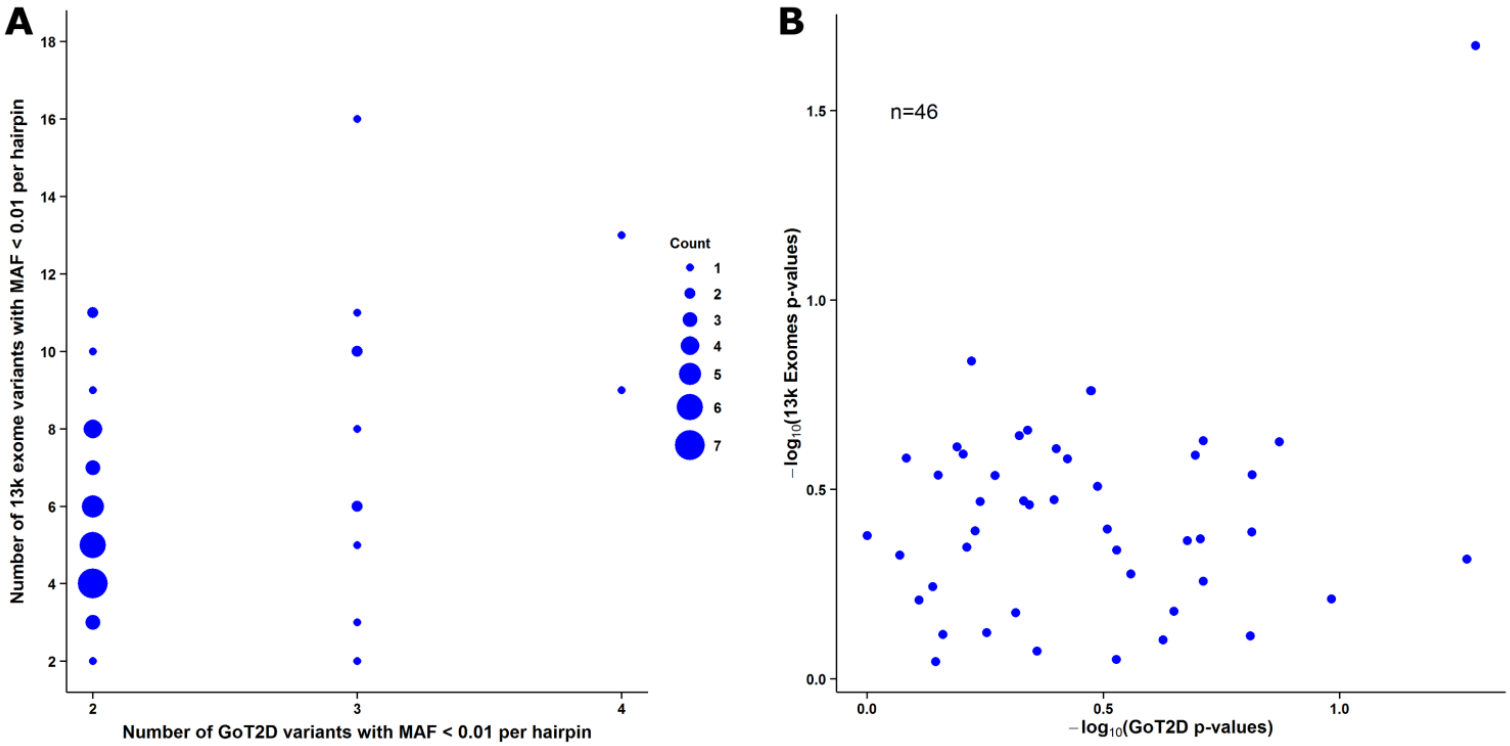


Figure 6.6: miRNA precursor rare variation grouping test comparison between GoT2D whole-genome sequenced samples and GoT2D+T2D-GENES whole-exome sequenced dataset. Panel A shows the increase in rare variants per miRNA precursor in the 13,000 exomes (y-axis) versus the 2,700 whole-genomes (y-axis) with the size of each point denoting the number of overlapping observations. Panel B shows the comparison of burden test *P*-values between the two datasets, with mir-152 (top right-hand corner) the clear outlier in terms of correlation.

6.3.3 Using islet annotations to uncover potential causal mechanisms at fine-mapped common T2D association loci

The previous two sections have focussed on islet regulatory RNA annotations. However, many other types of islet annotations are now available, and many alternative approaches to integrate these annotations with T2D genetics data are underway. To provide evidence for the potential success of such efforts, I will expand the number of islet annotations studied with the functional islet annotations generated in chapters three and four, information on islet regulatory elements[143](described in section 2.7.1), and variant type annotations generated with the Variant Effect Predictor tool (described in section 2.7.2). The twenty one T2D association loci that could successfully be refined in chapter five are used as exemplar regions for this kind of data integration.

At each of the fine-mapped loci, the available islet annotations were overlapped with variants in the 99% credible sets (as described in section 5.3.4) with the addition of all reported genome-wide significant variants not included in those sets (i.e. reported secondary signals). By far the most common annotation category for the fine-mapped variants is "active islet enhancer element" (derived from Pasquali *et al.*[143] chromatin signatures), indicating that all but two of the association loci contain chromatin active in adult islets (Figure 6.7). Information on eQTLs in human islets (as described in chapter four) was used to investigate whether 99% credible variants overlapping active islet chromatin showed evidence for regulatory effects of local gene expression. Whilst none of the eQTL variants reached the global 5% false-discovery rate significance cut-off (P-value < 2.7×10^{-5}) in the islet dataset, there was evidence of nominally significant eQTLs (P-value < 0.01) at five of the loci containing active islet chromatin (Figure 6.8A-E). At four of these five loci, the overlapping 99% credible variants coincided with the peak variants for the eQTL signal (Figure 6.8B-E) – the most significant eQTL variant was either the 99% credible variant inside the islet regulatory element (n=3), or in close linkage disequilibrium (LD) with the overlapping variant (n=1). The three genes for which the eQTL signal overlaps with an islet regulatory

element, as well as T2D association, are *ZMIZ1* in the *ZMIZ1* locus, *SLC22A18AS* in the *KCNQ1* locus and *PUS10* at the *BCL11A* locus (Figure 6.8B-D). At *ADCY5,* in the *ADCY5* locus, the top eQTL variant is in close LD with the islet regulatory variant (r^2 ~ 0.84 in European populations – Figure 6.8E).

Of these four regions, the *KCNQ1* locus has been most extensively studied. The region around the primary signal on chromosome 11p15.5 is known to be imprinted and has a well-established roll in Beckwith-Wiedemann syndrome, which is often associated with hypoglycaemia[298]. While methylation at the T2D risk variant rs2237895 is significantly correlated with genotype status[72,87], *SLC22A18AS* is biallelicly expressed in both foetal and adult human islets – suggesting that the observed methylation changes at the risk variant are unlikely to act through this gene[72]. Finding an eQTL for *SLC22A18AS,* overlapping islet active chromatin, at another variant within this locus (rs163184) might lead to reconsideration of a potential role for this gene, and underlines the complexity of defining causal mechanisms underlying association signals.

Two other genes – *ADCY5*, encoding adenylate cyclase 5, and *ZMIZ1*, which encodes Zinc Finger MIZ Domain-Containing Protein 1 – both have good functional claims for a role in islets. The T2D association signal at the *ADCY5* locus is robustly associated with changes in several measures of glycaemia[55,80,299], as well as beta-cell function as measured by HOMA-B[55]. In the data described in chapter four, the 99% credible variants for the same signal overlap a nominally significant islet eQTL for *ADCY5*. Adenylate cyclases play an important part in mediation of G-protein coupled receptor signalling by generating cyclic adenosine monophosphate (cAMP) upon receptor binding. In the beta-cell, incretin hormones (glucagon-like peptide-1 [GLP1] and glucose-dependent insulinotropic polypeptide [GIP]) potentiate insulin secretion by acting through this pathway[300]. These findings would therefore suggest that lower levels of adenylate cyclase might result in a lower incretin effect, which for adenyl cyclase 8 was indeed shown to be the case in both rodent and human islets[301]. In line with the lower incretin effect observed for reduced adenylate cyclase levels,

the allele associated with increased fasting plasma glucose and increased risk of developing T2D shows a nominally significant reduction in *ADCY5* levels (P =3.2x10^{-4}, β = -0.54 for the best eQTL variant) in the islet data.

ZMIZ1 is not a well characterised gene, but has some interesting features. It belongs to the protein inhibitor of activated STAT (PIAS) family, which was originally identified as a transcriptional regulator of the JAK–STAT pathway involved in cytokine response[302]. However, it was later discovered that ZMIZ1 plays a role in post-translational regulation of proteins through sumoylation[303]. Sumoylation, in turn, was found to modulate insulin exocytosis in rodent and human beta-cells: blocking the sumoylation process results in increased insulin exocytosis, whereas enhanced sumoylation inhibits insulin secretion[304]. Similarly to the *ADCY5* locus, the risk allele for *ZMIZ1* affects gene expression in the expected direction and leads to increased *ZMIZ1* expression (P =1.2x10^{-3}, β = 0.40).

The final gene with an eQTL overlapping an islet regulatory element and T2D association is *PUS10*, encoding pseudouridylate synthase 10, at the *BCL11A* locus. The link between pseudouridylate synthases, which are involved in posttranslational modification of RNA molecules and tRNA biogenesis[305], and islet function is not clear. However, PUS10, unlike most other members of pseudouridylate synthase families, also plays a role in regulating TRAIL (TNF-related apoptosis-inducing ligand) induced apoptosis[306]. This pathway is expressed in human islets, and as such PUS10 may protect islets against cellular stressors.

At a further six of the twenty one loci, the fine-mapped regions overlapped genes themselves. A direct measure of their impact is therefore available in the form of the allele-specific expression data – provided there is sufficient coverage of the variants. Two of these loci (*WFS1* and *SLC30A8* – Figure 6.9) do indeed have adequate coverage of overlapping variants (> 16 reads) in most heterozygous islet samples. The analysis shows that both genes have several sites displaying significant allele-specific expression (P < 0.01) in multiple individuals.

In *SLC30A8,* one 3'-UTR variant (rs3802177) shows a striking predominance of the reference allele (Figure 6.4A). Rs3802177 overlaps miRanda predicted target sites for several islet expressed miRNAs[269]. Disruption of these target sites could be responsible for the observed effect, although it is difficult to see how loss of miRNA binding would result in reduced allelic expression. The allelic imbalance at rs3802177 is therefore more likely to be a consequence of alternative splicing, resulting in an isoform with a shorter 3'-UTR region not containing the variant. Based on GENCODE version 17 annotations, this also implies a change in SLC30A8 protein structure. Most of the research at this locus has focussed on the T2D associated missense variant rs13266634 (p.W325R), which is thought to be causal[62]. Detailed studies in rodent islets have found conflicting effects of *Slc30a8* deletion on glucose homeostasis, with some suggesting impaired glucose tolerance[307,308] and others finding no changes[309]. In humans, the effects of loss of function coding mutations in *SLC30A8* on T2D risk suggest that haploinsufficiency protects against T2D[310]. The rs3802177 variant, which itself has been reported as the most significantly associated T2D variant in several studies[57,80], might offer an alternative causal mechanism at this locus and help explain the difficulties linking W325R to the association signal.

The allelic imbalance in *WFS1* is more generalised, with many variants affected in a subset of individuals (Figure 6.4B). *WFS1* has a strong claim for mediating the association signal, since mutations within it give rise to an autosomal dominant syndrome causing (amongst other symptoms) diabetes[311,312]. The associated region spans a large LD block and contains many potentially regulatory variants. A myriad of mechanisms have been suggested to underlie T2D association at *WFS1* – there have been studies demonstrating altered methylation patterns of the T2D associated variant (rs1801214) depending on risk haplotype[87], there are associated variants overlapping predicted binding sites for islet-expressed miRNAs[269], islet-specific long non-coding RNAs (lncRNAs)[209] and missense variants, and reported haplotypic effects on islet enhancer activity[71]. However, none of these findings alone has been sufficient to explain the functional mechanism. Identifying significant allele-specific expression (*P*-value < 0.01) in more than one individual at seven

WFS1 variants now provides evidence of *cis*-regulatory effects at this locus. However, while these data help to solidify the link between the T2D association signal and *WFS1*, finding the exact causal mechanism will remain challenging given the very strong LD block at this locus.

In conclusion, the proof-of-principle analyses described in this section demonstrate that combining refined association loci with extensive high quality annotations can provide many new potential avenues for functional follow-up. Whilst one type of annotation alone not often provides sufficiently substantial evidence, the combination of more than one type (e.g. nominally significant islet eQTL combined with active islet chromatin) can add weight to mechanistic hypotheses.

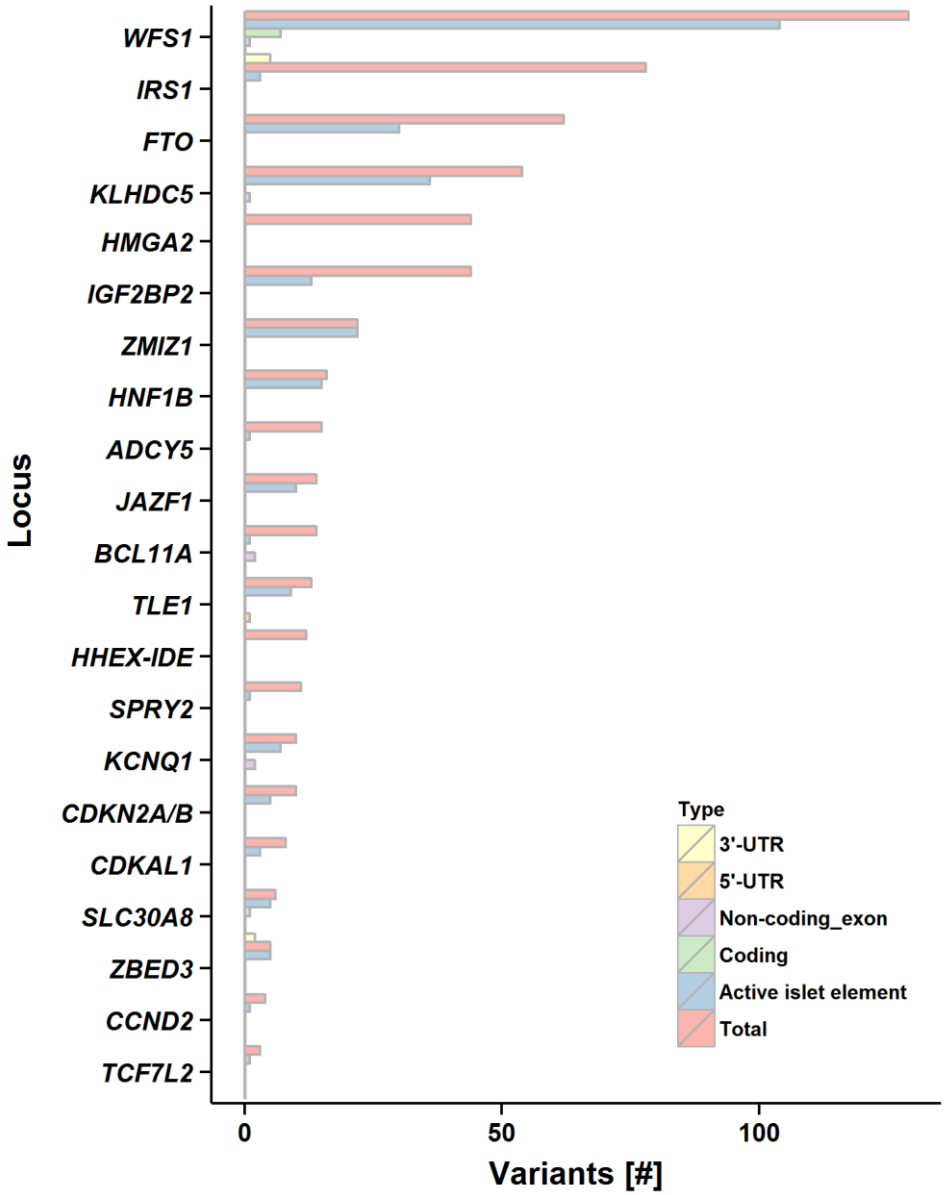


Figure 6.7: annotation of variants contained in the 99% likely causal sets. For each of the successfully fine-mapped T2D associated loci (y-axis) the number of variants contained in each functional category is plotted (x-axis).

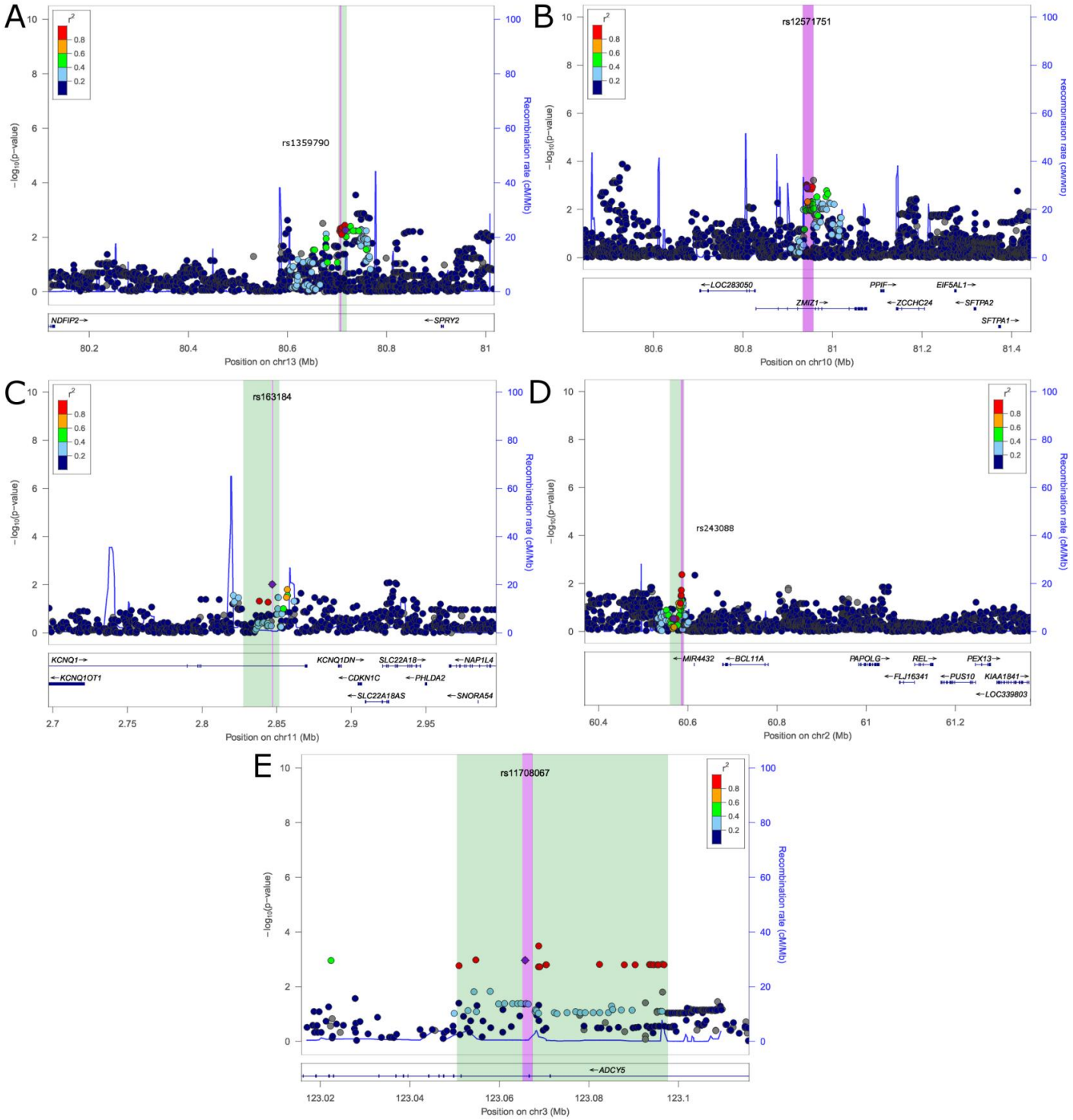


Figure 6.8: Fine-mapped T2D loci with overlap of 99% credible causal variants with islet active enhancers and nominally significant eQTLs. The panels show eQTL *P*-values (left y-axis) for all analysed variant for *NDFIP2* (*SPRY2* locus - A), *ZMIZ1* (*ZMIZ1* locus - B), *SLC22A18AS* (*KCNQ1* locus - C), *PUS10* (*BCL11A* locus - D) and *ACDY5* (*ADCY5* locus - E). Previously reported T2D association variants for each locus are indicated as purple dots, whereas the other colours indicate LD relationships. Gene annotations are present at the bottom of each plot, with the pink and green highlighted regions indicating the islet enhancer coordinates and 99% credible interval respectively. At *ZMIZ1* both regions completely overlap, hence the green region is not visible.

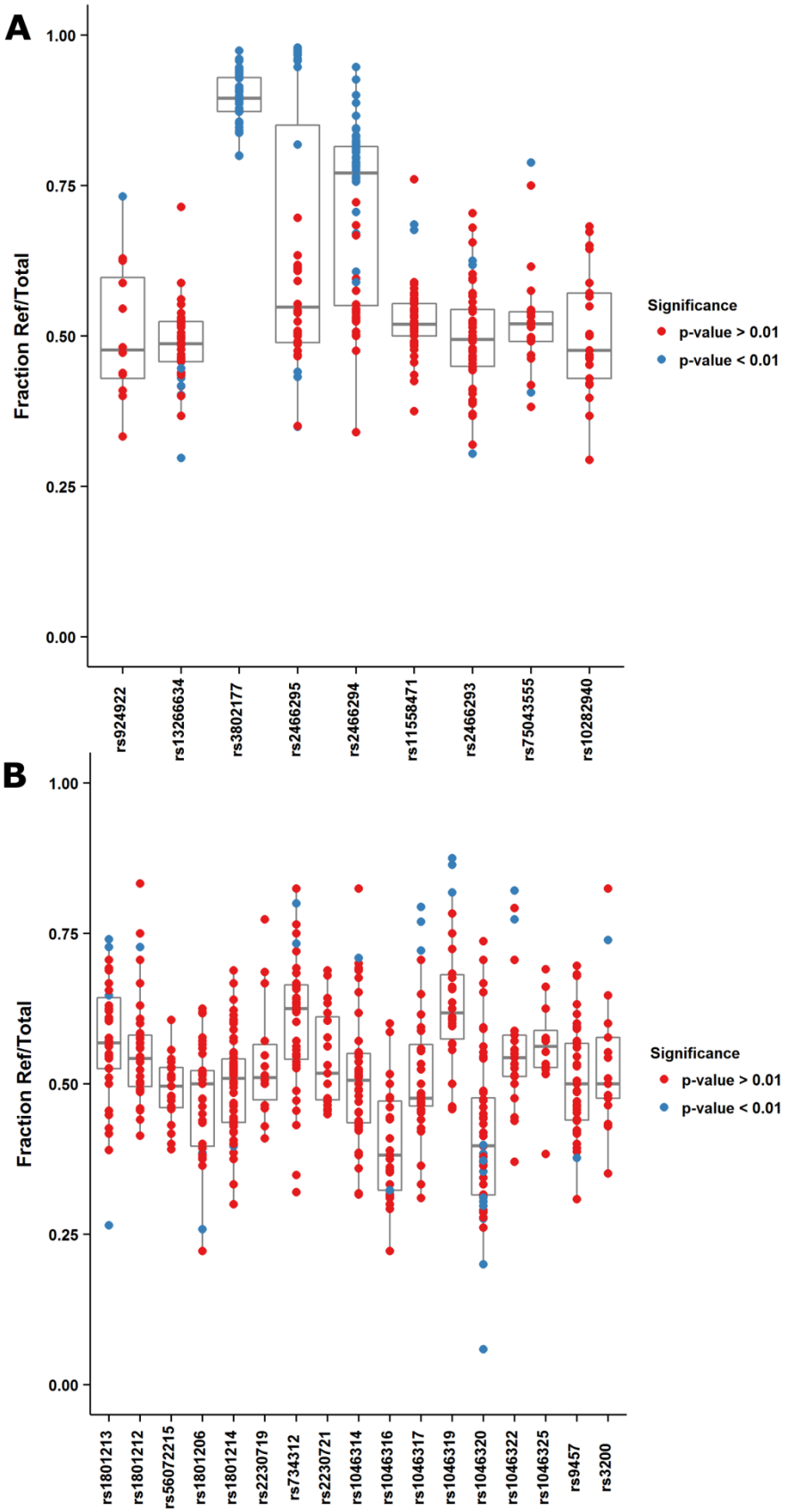


Figure 6.9: allele-specific expression results for all expressed variants in *SLC30A8* (A) and *WFS1* (B). The fraction of the reference allele over total allele counts for each individual (y-axis) is plotted per variant (x-axis). Boxplots show the median and interquartile range of the data. Colours denote whether variants showed significant ASE (blue) or not (red).

6.4 Discussion

Using novel islet annotations generated for expressed miRNAs and, through collaboration, lncRNAs, the work presented in this chapter shows that high quality islet annotations can provide novel insights into T2D pathogenesis. Enrichment analysis demonstrates that both islet lncRNAs and predicted miRNA target genes of islet-expressed miRNAs are enriched for T2D association signal. Furthermore, one of the novel islet lncRNAs shows suggestive association with T2D susceptibility in meta-analysis data from the GoT2D study. Whilst this finding will require replication to provide conclusive genome-wide significant evidence connecting this lncRNA to T2D susceptibility, it clearly demonstrates a link between T2D data and individual non-coding RNAs. These data therefore provide strong evidence for the involvement of non-coding RNAs in mediating common T2D association signal.

Whole-genome sequencing data from the GoT2D study provided an excellent first opportunity to study the potential impact of rarer variation in islet lncRNAs and miRNAs on T2D susceptibility. The highly conserved nature of miRNA transcripts meant that the number of observed variants per transcript was too low to effectively assess whether or not rare variation is likely to impact on T2D risk. However, since some of these miRNA transcripts were captured as part of the whole-exome sequencing project in 13,000 individuals with and without T2D, the number of observed variants could be increased by about 2.5-fold for 46 miRNA transcripts. This expansion enabled one potentially interesting miRNA *(mir-152)* to be identified, showing a consistent (and nominally significant) excess of rare variants in both datasets.

Neither miRNA targets nor lncRNAs showed strong evidence for a role of rare variants in mediating T2D risk, which, given the study size, suggests that large effect sizes in those annotations are unlikely. However, the possibility that rare variation in these types of

annotation can a role in T2D susceptibility cannot be excluded. Unlike protein-coding genes, the impact of variation on lncRNAs function cannot be easily established, so each test will include many neutral variants detrimental to the power of the test. Equally, predictions of miRNA targets are far from perfect, and as such each set will be large and contain many false-positive targets.

Finally, the refined T2D association loci from chapter five were combined with many more of the high quality islet annotations to provide a proof-of-principle that these annotations can be used to home in on causal genes or regulatory elements underlying T2D association loci. Since the first T2D GWAS in 2007, causal gene underlying association signals has been robustly identified in a minority of the cases (e.g. *KLF14*[109,110]), and potential non-coding regulatory mechanisms have been uncovered in a few others, including *TCF7L2*[70,71] and the primary signal at *KCNQ1*[72,87]. Results from the analyses in this chapter provide putative genes, as well as regulatory mechanisms, at a further four loci where islet eQTLs, T2D association signal and islet active elements coincide. Several recent studies have also identified enrichment of common variation across this type of chromatin with enhancer signatures in islets[143,268]. Although functional follow-up is clearly needed to validate the findings, they could represent a substantial increase in the number of loci at which the causal mechanism can be determined.

Furthermore, at the *SLC30A8* locus, strong allelic imbalance of one of the most associated T2D variants (rs3802177)[57,80] suggest that alternative splicing or changes in mRNA stability could be a causal mechanism. Such a mechanism would require a shift away from the prevailing assumption that the single amino acid substitution introduced by the missense variant in strong LD with rs3802177 is causal. There is support for rs3802177 from a genetic point-of-view – it being the reported lead variant in several T2D meta-analyses[57,80]. Again, the result required validation on an alternative platform, since technical bias resulting in a consistent allele-specific bias cannot be excluded. However, rs3802177 is not in close proximity to any known insertion or deletion, has not been blacklisted due to difficulties in

mapping by the ENCODE project, and shows no significant mapping bias when simulating reads mapping to this region (N. Panousis, personal communication). Based on the GENCODE annotations, the non-risk genotype would lead to a shorter expressed isoform. – consistent with the findings that protein-truncating variants in *SLC30A8* protect against T2D in humans[310]. However, the possibility that the ASE at rs3802177 is secondary to changes in RNA degradation/splicing caused by the coding variant cannot be excluded.

In conclusion, high quality islet (and other tissues involved in T2D pathogenesis) annotations will be essential for assessing the functional impact of T2D association signals. Larger sample sizes, and improvements in methodologies to aggregate functional annotations currently being developed, will improve our understanding even further.

Chapter 7

General Discussion

The work presented in this thesis had three main aims; to improve regulatory annotations of human islets, to assess whether statistical fine-mapping could be a viable option for common variant association loci, and, finally, to combine the fine-mapping and annotations to establish whether our understanding of the causal mechanisms underlying T2D loci could be improved.

Chapters three and four established, respectively, the miRNA and mRNA expression profile of human pancreatic islets by next generation sequencing approaches. Chapter three provides the first published catalogue of the miRNA expression profile in human islets and beta-cell samples enriched through fluorescence-activated cell sorting (FACS) using next generation sequencing. The study identified many novel highly-expressed miRNAs in islets, and confirms high human islet expression for miRNAs previously identified in rodent islets such as miR-375[173]. Since several beta-cell samples were also sequenced, the profile of this specific endocrine cell type could be compared against that of the whole islet. There was a strong correlation between the two profiles, with a clear (although not statistically significant) enrichment of the aforementioned miR-375 in the beta-cell population. Comparing the islet expression profile against those of sixteen non-islet tissues identified many other miRNAs with enriched expression in islets. After these data had been published, several further studies were published which corroborate my findings[196,197]. Several of the miRNAs identified as islet-enriched have now also been linked to islet function and T2D: it was demonstrated that miR-184 plays an important role in regulating insulin secretion and beta-cell proliferation in rodent and human islets[204], whereas the beta-cell specific miRNA cluster on chromosome 14q32.2 (containing miR-136 and miR-127) was shown to be downregulated in islets from T2D donors[196].

The mRNA expression profile of ~130 human islet preparations established in chapter four was used to study the impact of genetic variation on gene expression levels (through eQTL analysis) and allelic balance in transcribed genes (through ASE analysis). Around nine percent of the 18,519 protein-coding and long non-coding RNA (lncRNAs) transcripts had

significant *cis*-eQTLs, whereas ~67% of these genes contained a variant which displayed significant allelic imbalance in at least one individual. Using publicly available *cis*-eQTL data, a substantial number of the islet *cis*-eQTL signals were shown to be shared between tissues. When intersecting the *cis*-eQTL data with chromatin state definitions from human islets, significant *cis*-eQTLs detected in the islet data showed a clear depletion of overlapping repressive elements. Furthermore, when including the most significant variant for each gene, there was an enrichment of overlap between these variants and active islet chromatin. These results therefore indicate that the detected *cis*-eQTLs are indeed of islet origin and many more, true, islet *cis*-eQTLs remain to be identified amongst the non-significant findings.

Chapter five then set out to determine whether statistical fine-mapping could provide a viable method for refining common variant association regions to have the best chance of identifying causal mechanisms when intersecting with annotation data. Through simulations, the work in this chapter showed that fine-mapping methods could indeed reduce associated regions to smaller genomic interval or set of variants. These "credible" sets and intervals were well calibrated in terms of their inclusion of the causal variant. On average, the association signal could be refined to fewer than fifty variants and less than fifty kilobases, but not all regions were amenable to fine-mapping. The strength of the association signal was the most influential factor in determining the fine-mapping potential of a region (i.e. little association signal equalled low fine-mapping potential). Combining two methods to disentangled local disequilibrium (in this case Bayesian fine-mapping combined with multi-ethnic study design) further improved the ability of fine-mapping to accurately identify the causal variant. These simulated results were validated in an empirical dataset with good power to detect association. When applied to known common variant T2D associated loci imputed into a high-density panel derived from around 2,700 whole-genome sequenced individuals from the GoT2D study, Bayesian fine-mapping could refine twenty one of these loci to < 10% of the original genomic interval as well as < 10% of the original number of variants.

The final chapter (six) investigated whether improved islet annotations – in particular the islet regulatory RNA annotations (lncRNAs, miRNAs) I had been working on extensively – could provide new insights in T2D genetics. Common variant T2D genome-wide meta-analysis data showed significant enrichment of association signal across two islet regulatory RNA annotations (islet lncRNAs and target gene sets for islet expressed miRNAs). Whole-genome sequencing, performed as part of the GoT2D study, allowed me to explore most of the allele frequency spectrum across the regulatory RNA annotations. These analyses identified at least one miRNA precursor transcript (*mir-152*) with a consistent nominally significant excess of rare variation, but did not identify candidate annotations significant after multiple testing correction. They also uncovered one islet lncRNA overlapping a common association signal with a significance which makes it a good candidate for replication in a larger cohort. Finally, as proof-of-principle for the utility of integrating islet annotations to uncover novel potential biology underlying T2D association loci, a variety of annotations were applied to the twenty one fine-mapped common variant T2D association loci from chapter five. This analysis identified potential causal mechanisms or causal genes at six T2D association loci. For some, such as the *SLC30A8* locus, the results hinted at an unexplored causal mechanism, whereas for others it implicated genes not currently thought to be involved in the mechanistic underpinnings of the association signal (e.g. *PUS10*, *SLC22A18AS*).

The clearest next step from here would involve validating the findings at T2D association loci. Whilst the combination of nominally significant eQTLs coinciding with T2D association signals in regions with histone modifications suggestive for islet enhancer activity at the *ADCY5*, *ZMIZ1*, *BCL11A* and *KCNQ1* (rs163184) loci is very encouraging, all but the T2D association will need to be confirmed. It should be reasonably straight-forward to establish whether these regions indeed act as enhancer elements in, for example, a luciferase reporter system. However, as exemplified by the *TCF7L2* locus[70,71], even when it is possible to show activity (and allelic differences) at an islet enhancer, this does not definitively link it to a gene. Ideally the enhancer would harbour a study-wide significant eQTL, but unfortunately

this was not possible to detect in the islet RNA-seq dataset. Chromosome conformation capture studies (e.g. 4C) could instead be used to determine whether the enhancer region interacts with the gene of interest and to provide functional support for the nominally significant eQTL signal.

The islet dataset in chapter four consists of around one hundred and thirty human islet preparations. Whilst this size enabled eQTL analysis and is larger than any of the currently published studies on islets (which are mainly based on array technology)[214], it is much smaller than similar studies from more easily accessible tissues[205-207,243]. The mixed cell-types within the islet, as well as the differences in purity leading to contamination with exocrine cells, further reduce power for eQTL detection. Given that most T2D association signals seem to also associate with changes in beta-cell function and provided the loci harbour eQTLs, one might expect such eQTLs to act in a tissue-specific manner. This makes the reduction of power due to the mixed cell-types and small sample size an even more pressing problem, since tissue-specific eQTLs tend to have smaller effect sizes than those shared between tissues[205,206].

The resolution of these studies can be increased by isolating the specific cell-types normally combined in the islet. In analyses such as the miRNA profiles generated in chapter three, contributions of the different cell types can be untangled. When RNA-seq was applied to purified beta-cells (and the leftover "non-beta" component), it uncovered many genes which were differentially expressed in the different islet cell-types[108]. Observations such as this will be instrumental in efforts to improve generation of human beta-cell lines from, for example, induced pluripotent stem-cells. They provide a blueprint of beta-cell gene expression, and can help identify novel genes potentially important for cellular identity (e.g. islet lncRNAs upregulated during islet development). Stable protocols to generate pancreatic endocrine cells from easily accessible tissues from living donors would take away many of the difficulties associated with the islet isolation procedure. Standardisation of culture times and cellular viability, removal of selection bias towards older donors with high body mass

indices, and absence (or even equal levels) of exocrine contamination would be just a few of the major obstacles in islet work that would be resolved.

However, until such protocol is developed, the main way forward for improving eQTL detection is increasing sample size. This is very likely to require collaborative efforts between groups, since a single group is unlikely to obtain the sample sizes needed to detect smaller effect, tissue-specific, *cis*-eQTL signals. For transcriptome-wide studies into transcript usage or *trans*-eQTLs, the noise introduced during islet isolation proves more problematic. Whilst the properties of isoform usage per gene across samples in chapter four follow previously published characteristics[106,207], even after correction these data are too noisy for accurate identification of variation affecting isoform usage. Correcting for bias in *trans*-eQTL analysis is possible but requires very careful selection of covariates, since the correlation between different *trans* genes can be easily lost when some are subject to, for example, purity or viability factors whereas others are not.

Future work will undoubtedly provide information on *trans* associations and isoform usage, which is likely to provide even deeper insights into T2D associated loci. *Trans*-eQTLs in subcutaneous adipose tissue have already been implicated at the *KLF14* locus[110], and have been found to coincide with association signals at many other trait associated loci[205,243]. However, none of these analyses will be able to detect molecular mechanisms which rely on temporal expression in developing tissues, nor will they account for expression changes affecting only a subset of cells. For example, a recent publication by Weedon *et al*. found mutations in a distal enhancer element for *PTF1A* active in pancreatic progenitor cells to be a frequent cause of isolated pancreatic agenesis[313]. This finding clearly demonstrates the potential for such regulatory elements to be involved in disease pathogenesis, but also highlights how easily these effects are missed when not looking in the right tissue at the right time. The specific enhancer element – identified through chromatin profiling of pancreatic endoderm cells – in which the mutations resided did not exhibit any markers of active chromatin in any of the 68 cell types analysed as part of the Epigenome Roadmap project

and in 125 cell types from ENCODE (including adult exocrine pancreas). Detailed studies of methylation patterns around the *KCNQ1* locus associated with T2D equally identify temporal changes in methylation[72], which would be missed by profiling only adult tissue.

Maturing single-cell transcriptome sequencing technologies will soon provide opportunities to study samples previously inaccessible due to low quantities of available material (e.g. embryonic tissue)[314-317]. Although currently not many such studies have been published, those that have been provide evidence for substantial heterogeneity in gene expression between different cells of identical cell-type and developmental stage[314,315,317]. This suggests that to find molecular mechanisms at all association loci, we may well need to zoom in even further –profiling not just the tissue of interest, nor the specific cell-type of interest, but all the way down to the single cell level.

Particularly when interrogating rare variation, the results in chapter six highlight the need for improvements in computational tools assigning functionality to non-coding regulatory elements. The ability to profile novel non-coding RNAs, including miRNAs and lncRNAs, outpaces methods of assigning function to such elements. miRNA target predictions show only average overlap between different prediction algorithms (and poor overlap when looking at the specific genomic sequence predicted to be targeted). For lncRNAs, the only characteristic that can be derived from their name is the length of the transcript. Whilst many of the lncRNAs are thought to be functional, their function can currently be derived only from experimental studies. Integrated analysis of all the different (epi)genomic annotations currently being generated – especially on commonly studied cell-types such as LCLs – should facilitate detection of common patterns and lead to improvements in prediction algorithms.

Finally, the Bayesian fine-mapping approach validated in chapter five, and applied to T2D association data in chapter six, provides a flexible framework which should allow the inclusion of priors to further improve fine-mapping efforts. Post-hoc analysis to determine overlap between fine-mapped regions and functional elements is a useful starting point, but

ideally such information should be included in the model to allow formal testing of the increased likelihood of causality for that variant over other variants present in the fine-mapped interval. With correct prior selection, and the variety of functional annotations now available, such analysis would also provide more evidence for the proposed pivotal role of islets in mediating T2D association signal.

The data presented in this thesis have addressed two key challenges in understanding mechanisms of T2D susceptibility: improving islet genomic annotations (chapters three and four) and fine-mapping common association loci (chapter five). Together, these data provided a powerful means of uncovering molecular mechanisms underlying T2D associated loci (chapter six), and demonstrate the power of utilising a broad, multidisciplinary approach to investigate the genetic susceptibility mechanisms of T2D. Continued investigation is of paramount importance in developing more effective therapeutic approaches to combat the global diabetes epidemic.

Appendices

Additional tables, references and manuscripts

Tables

Appendix Table 1

Normalized average read counts for all 385 miRNAs present in human islets and beta-cells. Each row denotes an individual miRNA present above background levels (>100 observed reads). The first column has the miRNA identifiers according to miRBase 18. The entries in the two columns are the normalised average read counts for, respectively, human islets and beta-cells. Zero denotes the expression of the miRNA was below background levels (<100 total reads observed).

miRNA	Normalized Islet Average	Normalized Beta-cell Average
hsa-miR-375	2,046,518	5,804,575
hsa-miR-143-3p	1,233,087	284,207
hsa-let-7f-5p	544,989	966,619
hsa-let-7a-5p	343,033	668,000
hsa-miR-21-5p	284,834	60,031
hsa-miR-7-5p	259,457	670,521
hsa-miR-27b-3p	226,272	698,396
hsa-miR-192-5p	190,726	383,622
hsa-miR-148a-3p	176,002	389,911
hsa-let-7b-5p	153,729	386,363
hsa-miR-182-5p	152,455	327,398
hsa-miR-30d-5p	149,301	111,933
hsa-miR-26a-5p	132,411	389,655
hsa-miR-127-5p	126,729	427,356
hsa-miR-30a-5p	117,912	21,116
hsa-miR-22-3p	87,376	250,404
hsa-miR-200c-3p	72,125	60,837
hsa-miR-24-3p	69,287	34,032
hsa-miR-99b-5p	65,508	31,322
hsa-miR-101-3p	64,622	45,868
hsa-miR-184	58,167	151,568
hsa-miR-151a-3p	53,886	57,370
hsa-miR-378a-3p	53,505	11,167
hsa-miR-183-5p	52,753	83,327
hsa-let-7e-5p	48,850	81,004
hsa-miR-103a-3p	41,454	34,228
hsa-let-7i-5p	40,949	66,351
hsa-miR-10a-5p	37,020	14,299
hsa-miR-141-3p	32,860	93,408
hsa-let-7g-5p	32,464	58,038
hsa-miR-181a-5p	31,491	56,942
hsa-miR-30e-5p	29,462	11,512
hsa-miR-92a-3p	22,367	51,442
hsa-miR-129-5p	22,186	10,962
hsa-miR-379-5p	22,099	15,652
hsa-let-7c	20,411	23,638
hsa-miR-191-5p	20,388	61,566
hsa-miR-148b-3p	17,018	22,964
hsa-miR-744-5p	15,742	26,865
hsa-miR-200b-3p	14,969	22,134
hsa-miR-381-3p	14,546	15,696
hsa-miR-26b-5p	13,910	21,614
hsa-miR-29a-3p	13,379	15,815
hsa-miR-23b-3p	13,276	9,987
hsa-miR-200a-3p	12,885	12,698

hsa-miR-27a-3p	12,470	5,378
hsa-miR-411-5p	12,065	16,426
hsa-miR-30a-3p	11,855	3,071
hsa-miR-186-5p	11,459	31,188
hsa-miR-140-5p	11,017	2,453
hsa-miR-10b-5p	10,614	6,933
hsa-miR-16-5p	10,389	18,254
hsa-miR-889	10,356	15,668
hsa-miR-598	9,999	3,425
hsa-miR-125b-5p	9,513	16,076
hsa-miR-30e-3p	9,486	6,622
hsa-miR-136-3p	9,268	18,006
hsa-miR-92b-3p	8,705	40,263
hsa-miR-204-5p	8,346	32,817
hsa-miR-410	8,043	37,884
hsa-miR-409-5p	7,730	29,566
hsa-miR-98-5p	7,662	20,770
hsa-miR-134	7,626	10,575
hsa-miR-429	7,411	10,225
hsa-miR-28-5p	7,262	16,560
hsa-miR-194-5p	7,054	10,955
hsa-miR-30c-5p	6,760	3,355
hsa-miR-320a	6,516	11,772
hsa-miR-126-5p	6,384	2,735
hsa-miR-29c-3p	6,063	10,082
hsa-miR-222-3p	5,874	6,233
hsa-miR-221-3p	5,870	10,681
hsa-let-7d-5p	5,859	10,790
hsa-miR-181c-5p	5,719	12,695
hsa-miR-432-5p	5,680	18,508
hsa-miR-126-3p	5,637	888
hsa-miR-132-3p	5,531	7,607
hsa-miR-423-5p	5,337	11,840
hsa-miR-30b-5p	5,137	6,043
hsa-miR-25-3p	5,015	9,954
hsa-miR-23a-3p	4,993	1,533
hsa-miR-199a-5p	4,342	804
hsa-miR-3184-5p	4,276	9,264
hsa-miR-181b-5p	4,243	3,681
hsa-miR-374a-3p	4,041	1,102
hsa-miR-654-5p	3,994	11,054
hsa-miR-217	3,723	2,066
hsa-miR-382-5p	3,544	6,052
hsa-miR-340-5p	3,488	7,701
hsa-miR-487b	3,157	8,264
hsa-miR-95	2,830	3,121
hsa-miR-21-3p	2,576	1,222
hsa-miR-370	2,379	2,188
hsa-miR-128	2,352	2,847
hsa-miR-136-5p	2,274	3,381
hsa-miR-199b-5p	2,146	395
hsa-miR-9-5p	2,028	687
hsa-miR-93-5p	2,018	914
hsa-miR-15a-5p	1,982	2,914
hsa-miR-301a-3p	1,918	4,742
hsa-miR-100-5p	1,866	447
hsa-miR-153	1,755	3,025
hsa-miR-376a-5p	1,710	905

hsa-miR-152	1,659	1,656
hsa-miR-542-5p	1,601	244
hsa-miR-3065-5p	1,557	902
hsa-miR-29b-3p	1,500	1,212
hsa-miR-338-5p	1,495	2,911
hsa-miR-1299	1,402	321
hsa-miR-486-5p	1,394	4,312
hsa-miR-151b	1,370	3,563
hsa-miR-1468	1,336	6,172
hsa-miR-19b-3p	1,321	2,081
hsa-miR-216a-5p	1,275	980
hsa-miR-215	1,259	424
hsa-miR-335-5p	1,243	2,146
hsa-miR-503-5p	1,107	251
hsa-miR-369-5p	1,101	1,340
hsa-miR-210	1,056	581
hsa-miR-107	1,043	2,267
hsa-miR-1307-3p	1,039	1,070
hsa-miR-345-5p	970	1,228
hsa-miR-181d	929	954
hsa-miR-874	928	1,583
hsa-miR-660-5p	914	1,497
hsa-miR-146a-5p	903	825
hsa-miR-139-5p	865	757
hsa-miR-323a-3p	862	1,178
hsa-miR-99a-5p	854	666
hsa-miR-339-5p	851	1,415
hsa-miR-148a-5p	843	4,133
hsa-miR-652-3p	838	954
hsa-miR-145-5p	804	49
hsa-miR-130a-3p	746	401
hsa-miR-376c-3p	742	1,009
hsa-miR-27b-5p	732	1,549
hsa-miR-484	723	1,050
hsa-miR-424-5p	722	95
hsa-miR-493-5p	718	1,584
hsa-miR-200a-5p	716	760
hsa-miR-421	712	1,897
hsa-miR-493-3p	697	696
hsa-miR-425-5p	675	757
hsa-miR-135a-5p	657	286
hsa-miR-335-3p	653	751
hsa-miR-1246	653	0
hsa-miR-106b-5p	651	348
hsa-miR-221-5p	651	579
hsa-miR-320b	647	1,711
hsa-miR-361-5p	643	1,201
hsa-miR-129-1-3p	639	664
hsa-miR-1285-3p	637	386
hsa-miR-200b-5p	623	623
hsa-miR-331-5p	620	357
hsa-miR-99b-3p	610	473
hsa-miR-582-5p	607	801
hsa-miR-433	603	1,453
hsa-miR-873-5p	596	1,775
hsa-miR-20a-5p	588	77
hsa-miR-342-5p	586	1,017
hsa-miR-203a	583	243

hsa-miR-574-5p	563	829
hsa-miR-125b-2-3p	553	543
hsa-miR-4488	530	139
hsa-miR-185-5p	527	407
hsa-miR-941	524	1,207
hsa-miR-212-3p	520	1,382
hsa-miR-195-5p	516	213
hsa-miR-34a-5p	503	190
hsa-miR-96-5p	499	514
hsa-miR-1179	499	679
hsa-miR-181a-2-3p	492	432
hsa-miR-216b	490	1,086
hsa-miR-141-5p	487	692
hsa-miR-328	485	974
hsa-miR-4792	477	79
hsa-miR-30c-2-3p	473	340
hsa-miR-130b-3p	468	2,114
hsa-miR-106b-3p	459	750
hsa-miR-3934-5p	442	99
hsa-miR-17-5p	437	114
hsa-miR-1260b	436	1,326
hsa-miR-1260a	429	1,315
hsa-miR-181c-3p	393	514
hsa-miR-500a-3p	390	977
hsa-miR-1180	387	592
hsa-miR-323b-5p	386	1,115
hsa-miR-125b-1-3p	384	1,507
hsa-miR-197-3p	383	401
hsa-miR-330-5p	382	607
hsa-let-7d-3p	382	746
hsa-miR-132-5p	373	899
hsa-miR-1	369	187
hsa-miR-664a-3p	366	527
hsa-miR-374b-5p	357	557
hsa-miR-30d-3p	356	633
hsa-miR-454-3p	342	464
hsa-miR-378c	337	62
hsa-miR-214-3p	332	45
hsa-miR-15b-5p	330	442
hsa-miR-187-3p	325	80
hsa-let-7b-3p	312	151
hsa-miR-181a-3p	310	286
hsa-miR-4532	302	0
hsa-miR-31-5p	302	0
hsa-miR-155-5p	301	39
hsa-miR-32-5p	288	174
hsa-miR-374a-5p	281	560
hsa-miR-551b-3p	276	381
hsa-miR-29b-2-5p	272	125
hsa-miR-27a-5p	271	177
hsa-miR-340-3p	261	299
hsa-miR-219-5p	260	117
hsa-miR-668	258	858
hsa-miR-455-5p	258	291
hsa-miR-664a-5p	251	453
hsa-miR-145-3p	249	33
hsa-miR-3960	246	49
hsa-miR-589-5p	246	371

hsa-miR-3615	245	517
hsa-miR-4446-5p	241	388
hsa-miR-337-5p	241	123
hsa-miR-4634	227	0
hsa-miR-19a-3p	212	267
hsa-miR-452-5p	205	0
hsa-miR-494	197	147
hsa-miR-3200-5p	192	106
hsa-miR-497-5p	187	87
hsa-miR-4286	185	813
hsa-miR-33b-5p	181	183
hsa-miR-92b-5p	181	248
hsa-miR-376a-3p	179	183
hsa-miR-708-5p	177	245
hsa-miR-1301	176	286
hsa-miR-4448	176	160
hsa-miR-1275	174	780
hsa-miR-1296	172	345
hsa-miR-450a-5p	172	81
hsa-miR-149-5p	172	308
hsa-miR-324-5p	171	286
hsa-miR-299-5p	170	245
hsa-miR-154-5p	170	214
hsa-miR-885-5p	170	217
hsa-miR-125a-5p	165	204
hsa-miR-137	160	80
hsa-miR-758-3p	159	383
hsa-let-7a-3p	157	204
hsa-miR-671-5p	157	732
hsa-miR-539-5p	156	163
hsa-miR-532-5p	153	100
hsa-miR-320c	152	266
hsa-miR-377-5p	151	400
hsa-miR-3928	147	63
hsa-miR-22-5p	144	75
hsa-miR-708-3p	143	629
hsa-miR-431-3p	137	289
hsa-miR-29c-5p	137	79
hsa-miR-363-3p	135	51
hsa-miR-365a-3p	135	121
hsa-miR-495-3p	134	138
hsa-miR-190b	134	432
hsa-let-7f-2-3p	131	280
hsa-miR-329	130	62
hsa-miR-411-3p	129	97
hsa-miR-502-5p	129	226
hsa-miR-2110	128	122
hsa-miR-891a	126	0
hsa-miR-577	121	396
hsa-miR-124-3p	120	41
hsa-miR-485-5p	116	137
hsa-miR-377-3p	114	289
hsa-let-7i-3p	113	58
hsa-miR-489	105	94
hsa-miR-1292-5p	104	136
hsa-miR-296-5p	103	149
hsa-miR-431-5p	103	211
hsa-miR-146b-5p	103	91

hsa-miR-148b-5p	100	95
hsa-miR-17-3p	100	110
hsa-miR-301b	99	482
hsa-miR-320d	96	166
hsa-miR-138-5p	93	0
hsa-miR-33a-5p	92	0
hsa-let-7e-3p	92	124
hsa-miR-379-3p	92	95
hsa-miR-147b	91	106
hsa-miR-541-3p	91	232
hsa-miR-1197	90	163
hsa-miR-193b-3p	86	198
hsa-miR-1185-5p	84	142
hsa-miR-133a	83	153
hsa-miR-655	81	66
hsa-miR-487a	79	179
hsa-miR-501-5p	79	109
hsa-miR-3605-5p	78	211
hsa-miR-543	78	135
hsa-miR-130b-5p	74	519
hsa-miR-590-5p	74	79
hsa-miR-424-3p	73	40
hsa-miR-4454	73	64
hsa-miR-4508	72	0
hsa-miR-30c-1-3p	71	156
hsa-miR-194-3p	71	132
hsa-miR-656	71	193
hsa-miR-3656	68	0
hsa-miR-190a	67	0
hsa-miR-3529-5p	65	88
hsa-miR-143-5p	65	0
hsa-miR-7-1-3p	63	82
hsa-miR-1268b	59	0
hsa-miR-1270	59	108
hsa-miR-224-5p	58	0
hsa-miR-760	58	225
hsa-miR-378d	57	0
hsa-miR-548k	55	137
hsa-miR-628-5p	54	84
hsa-miR-4492	54	0
hsa-miR-641	54	80
hsa-miR-142-5p	53	0
hsa-miR-629-5p	53	55
hsa-miR-218-5p	52	0
hsa-miR-29a-5p	51	52
hsa-miR-3158-5p	50	45
hsa-miR-1287	49	52
hsa-miR-3909	49	99
hsa-miR-1268a	49	0
hsa-miR-15b-3p	46	0
hsa-miR-496	46	77
hsa-miR-380-3p	46	98
hsa-miR-625-5p	46	92
hsa-miR-642a-5p	46	73
hsa-miR-627	44	0
hsa-miR-99a-3p	43	0
hsa-miR-4662b	43	76
hsa-miR-26b-3p	43	190

hsa-miR-3177-5p	43	91
hsa-miR-451a	42	0
hsa-miR-378f	42	0
hsa-miR-1262	41	80
hsa-miR-642b-3p	41	62
hsa-miR-1291	41	0
hsa-miR-205-5p	40	0
hsa-miR-185-3p	40	49
hsa-miR-425-3p	40	63
hsa-miR-3074-5p	39	48
hsa-miR-454-5p	39	86
hsa-miR-150-5p	39	0
hsa-miR-665	39	129
hsa-miR-1224-5p	39	0
hsa-miR-183-3p	38	105
hsa-miR-135b-5p	38	0
hsa-miR-326	37	0
hsa-miR-4781-5p	37	0
hsa-miR-105-5p	37	34
hsa-miR-3912	36	51
hsa-miR-2355-5p	36	0
hsa-miR-3120-5p	36	0
hsa-miR-18a-5p	36	0
hsa-miR-548aa	35	0
hsa-miR-9-3p	35	0
hsa-let-7f-1-3p	35	0
hsa-miR-1276	35	0
hsa-miR-374b-3p	35	48
hsa-miR-1303	34	37
hsa-miR-548e	34	38
hsa-miR-4520a-5p	34	58
hsa-miR-23c	0	244
hsa-miR-802	0	199
hsa-miR-592	0	133
hsa-miR-541-5p	0	114
hsa-miR-191-3p	0	93
hsa-miR-1193	0	67
hsa-miR-4677-5p	0	66
hsa-miR-3929	0	62
hsa-miR-211-5p	0	60
hsa-miR-770-5p	0	49
hsa-miR-744-3p	0	48
hsa-miR-550a-5p	0	46
hsa-miR-92a-1-5p	0	44
hsa-miR-30b-3p	0	42
hsa-miR-7-2-3p	0	42
hsa-miR-192-3p	0	38
hsa-miR-2682-5p	0	38
hsa-miR-93-3p	0	36

Appendix Table 2

miRNA level comparison between tissues. Each row represents a different miRNA observed for at least 1000 reads across all tissues. The first column has the miRNA identifiers according to miRBase 18. Each other column represents a distinct (set of) tissues used for the comparison. The entries in the columns denote the tissue specificity score as described in the section 3.2.6.

miRNA	Enrich-ment score islets	Enrich-ment score fat	Enrich-ment score pigment cells	Enrich-ment score liver	Enrich-ment score heart/ muscle	Enrich-ment score orbital gyrus	Enrich-ment score pancreas	Enrich-ment score lung/ spleen/ kidney	Enrich-ment score bone marrow/ CD24+ progenitor cells	Enrich-ment score B-cells/thy mus	Enrich-ment score skin
hsa-miR-184	0.97	0.00	0.00	0.00	0.00	0.01	0.00	0.00	0.00	0.00	0.00
hsa-miR-375	0.96	0.00	0.00	0.00	0.00	0.00	0.03	0.00	0.00	0.00	0.00
hsa-miR-409-5p	0.93	0.01	0.00	0.01	0.00	0.02	0.00	0.01	0.01	0.00	0.02
hsa-miR-182-5p	0.92	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.07
hsa-miR-1468	0.91	0.01	0.01	0.01	0.00	0.01	0.02	0.01	0.01	0.01	0.00
hsa-miR-183-5p	0.89	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.10
hsa-miR-136-3p	0.89	0.01	0.00	0.01	0.00	0.04	0.01	0.01	0.00	0.00	0.03
hsa-miR-127-5p	0.86	0.00	0.00	0.01	0.00	0.07	0.01	0.01	0.01	0.00	0.03
hsa-miR-370	0.83	0.03	0.00	0.00	0.00	0.07	0.00	0.00	0.01	0.00	0.04
hsa-miR-153	0.83	0.04	0.00	0.00	0.00	0.09	0.01	0.00	0.00	0.00	0.01
hsa-miR-200a-5p	0.82	0.00	0.00	0.00	0.00	0.00	0.01	0.00	0.01	0.01	0.14
hsa-miR-7-5p	0.82	0.00	0.01	0.00	0.00	0.03	0.02	0.00	0.03	0.08	0.00
hsa-miR-410	0.80	0.00	0.00	0.00	0.00	0.11	0.00	0.00	0.05	0.00	0.03
hsa-miR-889	0.79	0.00	0.00	0.00	0.00	0.14	0.02	0.02	0.01	0.00	0.01
hsa-miR-668	0.79	0.00	0.00	0.01	0.00	0.16	0.01	0.02	0.01	0.00	0.00
hsa-miR-654-5p	0.78	0.01	0.00	0.02	0.01	0.01	0.01	0.01	0.01	0.00	0.14
hsa-miR-216a	0.78	0.02	0.00	0.00	0.00	0.00	0.18	0.00	0.00	0.00	0.01
hsa-miR-376a-5p	0.77	0.03	0.00	0.01	0.00	0.03	0.00	0.00	0.00	0.00	0.15
hsa-miR-204-5p	0.77	0.01	0.00	0.00	0.00	0.10	0.01	0.10	0.00	0.00	0.01
hsa-miR-381	0.76	0.02	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.22
hsa-miR-411-5p	0.74	0.02	0.00	0.03	0.00	0.05	0.01	0.01	0.00	0.00	0.14
hsa-miR-1179	0.73	0.13	0.00	0.00	0.01	0.05	0.04	0.01	0.01	0.01	0.01
hsa-miR-493-5p	0.71	0.01	0.01	0.02	0.00	0.00	0.01	0.02	0.05	0.00	0.17

hsa-miR-212-3p	0.69	0.02	0.04	0.01	0.00	0.11	0.00	0.04	0.02	0.03	0.02
hsa-miR-132-3p	0.66	0.06	0.02	0.01	0.01	0.07	0.01	0.03	0.01	0.04	0.07
hsa-miR-141-3p	0.66	0.00	0.00	0.00	0.00	0.00	0.01	0.00	0.00	0.00	0.33
hsa-miR-493-3p	0.63	0.02	0.01	0.03	0.01	0.01	0.02	0.05	0.02	0.00	0.20
hsa-miR-1180	0.62	0.04	0.00	0.01	0.01	0.17	0.01	0.07	0.01	0.01	0.06
hsa-miR-429	0.60	0.00	0.00	0.01	0.00	0.00	0.05	0.02	0.00	0.00	0.32
hsa-miR-27b-3p	0.60	0.05	0.01	0.00	0.01	0.00	0.00	0.01	0.01	0.00	0.31
hsa-miR-148a-5p	0.60	0.06	0.01	0.02	0.00	0.00	0.02	0.00	0.00	0.02	0.27
hsa-miR-125b-1-3p	0.60	0.05	0.04	0.02	0.01	0.15	0.02	0.02	0.00	0.01	0.09
hsa-miR-582-5p	0.59	0.02	0.24	0.01	0.00	0.01	0.00	0.00	0.04	0.01	0.06
hsa-miR-129-5p	0.58	0.00	0.06	0.00	0.00	0.33	0.02	0.00	0.00	0.00	0.00
hsa-miR-671-5p	0.56	0.04	0.11	0.02	0.00	0.00	0.00	0.01	0.12	0.10	0.04
hsa-miR-200a-3p	0.55	0.00	0.00	0.03	0.00	0.00	0.10	0.03	0.00	0.01	0.29
hsa-miR-148a-3p	0.53	0.04	0.01	0.07	0.00	0.00	0.06	0.01	0.00	0.04	0.24
hsa-miR-181c-5p	0.52	0.04	0.02	0.01	0.01	0.13	0.01	0.01	0.13	0.05	0.08
hsa-miR-323b-5p	0.52	0.00	0.01	0.01	0.01	0.33	0.06	0.01	0.05	0.00	0.00
hsa-miR-141-5p	0.51	0.00	0.00	0.00	0.00	0.00	0.02	0.00	0.00	0.00	0.46
hsa-miR-369-5p	0.47	0.04	0.01	0.07	0.01	0.18	0.05	0.03	0.00	0.00	0.14
hsa-miR-190b	0.46	0.07	0.00	0.04	0.03	0.00	0.24	0.07	0.02	0.04	0.03
hsa-miR-134	0.45	0.01	0.00	0.01	0.01	0.31	0.04	0.08	0.04	0.00	0.04
hsa-miR-194-5p	0.44	0.03	0.00	0.38	0.00	0.00	0.02	0.03	0.01	0.02	0.06
hsa-miR-3065-5p	0.44	0.05	0.01	0.01	0.00	0.30	0.02	0.02	0.01	0.00	0.13
hsa-miR-340-3p	0.44	0.45	0.01	0.01	0.00	0.00	0.00	0.01	0.01	0.01	0.08
hsa-miR-132-5p	0.44	0.07	0.03	0.02	0.01	0.31	0.03	0.03	0.01	0.05	0.02
hsa-miR-148b-3p	0.43	0.11	0.06	0.03	0.00	0.02	0.01	0.02	0.04	0.07	0.23
hsa-miR-744-5p	0.42	0.03	0.10	0.01	0.01	0.13	0.01	0.02	0.10	0.12	0.04
hsa-miR-720	0.42	0.08	0.01	0.02	0.00	0.00	0.00	0.00	0.01	0.09	0.37
hsa-miR-200b-3p	0.40	0.00	0.00	0.03	0.00	0.00	0.08	0.02	0.00	0.00	0.47
hsa-miR-577	0.39	0.01	0.02	0.00	0.00	0.27	0.11	0.03	0.01	0.08	0.07
hsa-mir-151	0.38	0.06	0.15	0.08	0.02	0.05	0.04	0.03	0.02	0.06	0.10
hsa-miR-301a-3p	0.37	0.19	0.05	0.02	0.01	0.01	0.01	0.01	0.25	0.08	0.03
hsa-miR-95	0.37	0.12	0.00	0.03	0.08	0.08	0.03	0.04	0.00	0.05	0.21
hsa-miR-30d-3p	0.37	0.18	0.08	0.03	0.04	0.01	0.03	0.02	0.02	0.07	0.14
hsa-miR-1260a	0.36	0.08	0.01	0.01	0.01	0.01	0.01	0.02	0.01	0.02	0.46
hsa-miR-1260b	0.35	0.06	0.01	0.01	0.01	0.01	0.02	0.01	0.01	0.05	0.45
hsa-miR-130b-5p	0.35	0.03	0.14	0.04	0.01	0.01	0.03	0.00	0.10	0.24	0.06

hsa-miR-1299	0.33	0.03	0.03	0.01	0.11	0.25	0.07	0.12	0.01	0.03	0.01
hsa-miR-432-5p	0.33	0.00	0.00	0.02	0.01	0.52	0.04	0.06	0.01	0.00	0.00
hsa-miR-4446-5p	0.33	0.01	0.00	0.01	0.01	0.16	0.00	0.02	0.21	0.26	0.01
hsa-miR-92b-3p	0.32	0.00	0.13	0.01	0.01	0.30	0.03	0.10	0.06	0.02	0.02
hsa-miR-98	0.32	0.08	0.05	0.02	0.05	0.10	0.05	0.07	0.06	0.12	0.08
hsa-miR-99b-5p	0.32	0.09	0.17	0.02	0.02	0.10	0.03	0.05	0.01	0.00	0.18
hsa-miR-4286	0.31	0.11	0.01	0.06	0.01	0.02	0.03	0.03	0.04	0.04	0.33
hsa-miR-484	0.31	0.11	0.06	0.05	0.02	0.02	0.01	0.06	0.08	0.13	0.15
hsa-miR-377-5p	0.30	0.02	0.02	0.24	0.00	0.25	0.05	0.04	0.03	0.00	0.05
hsa-miR-135a-5p	0.30	0.01	0.00	0.01	0.01	0.39	0.14	0.10	0.01	0.01	0.01
hsa-miR-338-5p	0.30	0.14	0.00	0.02	0.01	0.24	0.02	0.07	0.06	0.00	0.13
hsa-miR-30d-5p	0.28	0.06	0.07	0.11	0.08	0.03	0.03	0.07	0.02	0.06	0.19
hsa-miR-500a-3p	0.28	0.06	0.23	0.04	0.02	0.02	0.02	0.04	0.05	0.07	0.16
hsa-miR-487b	0.28	0.01	0.00	0.03	0.00	0.54	0.03	0.05	0.05	0.00	0.01
hsa-miR-1275	0.28	0.03	0.01	0.03	0.01	0.06	0.11	0.05	0.06	0.26	0.10
hsa-miR-382-5p	0.28	0.01	0.00	0.02	0.01	0.51	0.05	0.05	0.04	0.00	0.03
hsa-miR-136-5p	0.27	0.23	0.00	0.02	0.02	0.27	0.06	0.06	0.01	0.00	0.06
hsa-miR-200b-5p	0.26	0.00	0.00	0.07	0.00	0.00	0.17	0.03	0.00	0.01	0.45
hsa-miR-708-3p	0.25	0.23	0.01	0.02	0.01	0.23	0.01	0.01	0.01	0.01	0.21
hsa-miR-589-5p	0.25	0.05	0.15	0.01	0.01	0.08	0.04	0.05	0.08	0.23	0.05
hsa-miR-379-5p	0.24	0.09	0.00	0.03	0.01	0.15	0.01	0.02	0.01	0.00	0.44
hsa-miR-96-5p	0.24	0.08	0.00	0.00	0.00	0.00	0.05	0.03	0.01	0.02	0.56
hsa-miR-421	0.24	0.04	0.11	0.03	0.03	0.09	0.01	0.03	0.26	0.13	0.03
hsa-miR-361-5p	0.24	0.05	0.07	0.03	0.01	0.03	0.01	0.04	0.08	0.35	0.08
hsa-miR-423-5p	0.24	0.10	0.19	0.03	0.01	0.02	0.01	0.02	0.12	0.16	0.11
hsa-miR-26a-5p	0.23	0.08	0.03	0.05	0.03	0.04	0.03	0.07	0.10	0.13	0.22
hsa-miR-186-5p	0.23	0.16	0.03	0.07	0.02	0.03	0.04	0.05	0.05	0.13	0.19
hsa-miR-3184-5p	0.22	0.09	0.21	0.03	0.01	0.02	0.01	0.02	0.14	0.15	0.10
hsa-miR-542-5p	0.22	0.25	0.10	0.03	0.04	0.01	0.01	0.05	0.11	0.04	0.15
hsa-miR-28-5p	0.22	0.05	0.10	0.17	0.05	0.02	0.03	0.04	0.04	0.17	0.11
hsa-miR-27b-5p	0.22	0.10	0.04	0.17	0.16	0.03	0.03	0.03	0.01	0.01	0.20
hsa-mir-323	0.21	0.00	0.00	0.02	0.01	0.70	0.02	0.02	0.01	0.00	0.01
hsa-miR-22-3p	0.21	0.26	0.07	0.11	0.11	0.04	0.04	0.07	0.01	0.01	0.07
hsa-miR-873-5p	0.20	0.00	0.07	0.00	0.00	0.68	0.02	0.02	0.01	0.00	0.01
hsa-let-7d-3p	0.19	0.05	0.08	0.07	0.06	0.18	0.06	0.10	0.02	0.02	0.17
hsa-miR-154-5p	0.18	0.01	0.00	0.06	0.04	0.42	0.08	0.13	0.01	0.00	0.07

hsa-miR-874	0.18	0.02	0.00	0.01	0.02	0.65	0.01	0.05	0.01	0.02	0.03
hsa-miR-551b-3p	0.18	0.06	0.00	0.03	0.03	0.19	0.14	0.05	0.29	0.01	0.02
hsa-miR-151b	0.17	0.11	0.23	0.09	0.02	0.07	0.02	0.04	0.03	0.13	0.09
hsa-miR-30b-5p	0.17	0.38	0.01	0.02	0.01	0.01	0.01	0.02	0.01	0.02	0.34
hsa-miR-339-5p	0.17	0.14	0.07	0.03	0.04	0.11	0.08	0.09	0.10	0.08	0.08
hsa-miR-376c	0.17	0.06	0.00	0.15	0.01	0.21	0.05	0.10	0.07	0.00	0.18
hsa-miR-200c-3p	0.17	0.00	0.00	0.00	0.00	0.00	0.07	0.01	0.00	0.00	0.75
hsa-miR-335-3p	0.16	0.68	0.00	0.01	0.00	0.01	0.01	0.01	0.00	0.01	0.11
hsa-miR-3615	0.15	0.04	0.02	0.01	0.01	0.01	0.01	0.02	0.40	0.29	0.03
hsa-miR-125b-2-3p	0.15	0.19	0.00	0.02	0.01	0.03	0.00	0.00	0.00	0.00	0.58
hsa-miR-494	0.15	0.12	0.01	0.11	0.02	0.13	0.03	0.11	0.09	0.00	0.23
hsa-miR-197-3p	0.15	0.11	0.06	0.01	0.02	0.04	0.01	0.04	0.08	0.10	0.37
hsa-miR-29b-2-5p	0.14	0.33	0.00	0.03	0.02	0.17	0.03	0.06	0.06	0.05	0.11
hsa-miR-454-3p	0.14	0.30	0.05	0.01	0.01	0.00	0.01	0.02	0.08	0.15	0.24
hsa-miR-598	0.14	0.01	0.01	0.01	0.01	0.66	0.03	0.07	0.03	0.01	0.03
hsa-miR-328	0.14	0.03	0.02	0.02	0.05	0.54	0.04	0.07	0.01	0.01	0.07
hsa-miR-664-3p	0.13	0.02	0.04	0.04	0.04	0.07	0.04	0.08	0.13	0.19	0.21
hsa-miR-106b-3p	0.13	0.03	0.16	0.10	0.02	0.03	0.01	0.03	0.20	0.22	0.07
hsa-miR-574-5p	0.13	0.05	0.08	0.15	0.02	0.03	0.06	0.09	0.01	0.00	0.38
hsa-miR-30e-3p	0.13	0.08	0.22	0.04	0.02	0.02	0.01	0.02	0.07	0.23	0.17
hsa-miR-181c-3p	0.13	0.02	0.02	0.00	0.01	0.40	0.04	0.02	0.21	0.10	0.05
hsa-miR-99b-3p	0.12	0.02	0.18	0.03	0.07	0.30	0.06	0.14	0.05	0.00	0.02
hsa-miR-21-3p	0.12	0.00	0.39	0.25	0.00	0.00	0.01	0.03	0.08	0.08	0.01
hsa-miR-374a-3p	0.12	0.23	0.04	0.15	0.04	0.07	0.05	0.06	0.01	0.10	0.13
hsa-miR-143-3p	0.11	0.54	0.00	0.01	0.01	0.00	0.01	0.01	0.00	0.00	0.31
hsa-miR-340-5p	0.11	0.03	0.27	0.02	0.03	0.17	0.03	0.05	0.21	0.06	0.01
hsa-miR-30e-5p	0.11	0.19	0.02	0.06	0.04	0.02	0.02	0.03	0.04	0.08	0.40
hsa-miR-149-5p	0.11	0.06	0.00	0.00	0.00	0.03	0.00	0.00	0.00	0.00	0.79
hsa-miR-30a-5p	0.11	0.16	0.07	0.10	0.08	0.04	0.06	0.18	0.00	0.00	0.20
hsa-miR-27a-5p	0.11	0.04	0.08	0.01	0.00	0.00	0.00	0.00	0.15	0.02	0.59
hsa-miR-652-3p	0.11	0.45	0.02	0.01	0.04	0.04	0.02	0.04	0.05	0.04	0.16
hsa-miR-324-5p	0.10	0.17	0.04	0.05	0.03	0.22	0.02	0.11	0.08	0.07	0.10
hsa-miR-4488	0.10	0.05	0.15	0.11	0.07	0.02	0.11	0.20	0.04	0.04	0.09
hsa-miR-1285-3p	0.10	0.05	0.19	0.11	0.02	0.01	0.03	0.03	0.15	0.17	0.14
hsa-miR-23b-3p	0.10	0.04	0.03	0.18	0.08	0.04	0.03	0.07	0.02	0.01	0.40
hsa-miR-3934	0.10	0.02	0.10	0.01	0.02	0.00	0.02	0.06	0.09	0.51	0.07

hsa-miR-532-5p	0.10	0.05	0.09	0.10	0.03	0.04	0.03	0.13	0.09	0.07	0.26
hsa-miR-3960	0.10	0.03	0.02	0.03	0.12	0.03	0.10	0.47	0.03	0.01	0.07
hsa-miR-33b-5p	0.10	0.25	0.04	0.07	0.01	0.26	0.05	0.03	0.12	0.04	0.03
hsa-miR-187-3p	0.09	0.00	0.00	0.04	0.02	0.24	0.06	0.39	0.05	0.01	0.11
hsa-let-7a-3p	0.09	0.07	0.03	0.08	0.04	0.12	0.08	0.07	0.03	0.09	0.29
hsa-miR-125b-5p	0.09	0.18	0.01	0.07	0.03	0.15	0.02	0.03	0.00	0.00	0.43
hsa-let-7b-3p	0.09	0.06	0.04	0.07	0.01	0.08	0.07	0.07	0.03	0.03	0.46
hsa-miR-30a-3p	0.09	0.19	0.26	0.04	0.05	0.05	0.05	0.15	0.00	0.00	0.13
hsa-miR-30c-1-3p	0.08	0.02	0.12	0.16	0.08	0.10	0.05	0.03	0.05	0.25	0.07
hsa-miR-146b-5p	0.08	0.02	0.00	0.31	0.01	0.05	0.07	0.12	0.21	0.10	0.04
hsa-miR-192-5p	0.08	0.00	0.00	0.81	0.00	0.00	0.06	0.02	0.01	0.02	0.00
hsa-miR-129-1-3p	0.08	0.00	0.02	0.00	0.00	0.85	0.04	0.00	0.00	0.00	0.00
hsa-miR-590-5p	0.08	0.45	0.05	0.05	0.03	0.02	0.02	0.01	0.02	0.12	0.14
hsa-miR-941	0.08	0.01	0.15	0.04	0.03	0.04	0.02	0.04	0.31	0.24	0.02
hsa-miR-628-5p	0.08	0.13	0.01	0.04	0.14	0.41	0.04	0.08	0.03	0.02	0.03
hsa-let-7e-5p	0.08	0.02	0.47	0.05	0.06	0.14	0.06	0.08	0.02	0.01	0.01
hsa-miR-450a-5p	0.08	0.28	0.02	0.08	0.08	0.01	0.01	0.06	0.15	0.09	0.14
hsa-miR-23c	0.08	0.02	0.10	0.34	0.04	0.07	0.14	0.11	0.07	0.02	0.03
hsa-miR-1185-5p	0.08	0.06	0.01	0.10	0.01	0.62	0.04	0.03	0.02	0.00	0.04
hsa-miR-376a-3p	0.08	0.06	0.01	0.07	0.02	0.19	0.07	0.14	0.04	0.00	0.33
hsa-miR-660-5p	0.07	0.07	0.12	0.07	0.03	0.03	0.04	0.22	0.08	0.13	0.15
hsa-miR-21-5p	0.07	0.08	0.09	0.09	0.01	0.01	0.06	0.16	0.12	0.13	0.18
hsa-miR-625-5p	0.07	0.05	0.20	0.05	0.02	0.04	0.04	0.02	0.08	0.37	0.06
hsa-miR-92a-3p	0.07	0.02	0.29	0.02	0.01	0.01	0.01	0.03	0.15	0.32	0.09
hsa-miR-433	0.07	0.00	0.00	0.02	0.01	0.85	0.02	0.01	0.01	0.00	0.00
hsa-miR-1301	0.07	0.01	0.27	0.01	0.04	0.28	0.01	0.09	0.14	0.05	0.01
hsa-miR-32-5p	0.07	0.37	0.10	0.03	0.01	0.01	0.01	0.03	0.10	0.05	0.22
hsa-miR-1270	0.07	0.04	0.02	0.01	0.04	0.04	0.04	0.04	0.28	0.36	0.07
hsa-miR-10a-5p	0.07	0.12	0.09	0.04	0.02	0.00	0.07	0.25	0.04	0.00	0.32
hsa-miR-30c-2-3p	0.07	0.16	0.15	0.17	0.05	0.09	0.10	0.14	0.00	0.00	0.08
hsa-miR-30c-5p	0.06	0.36	0.01	0.02	0.01	0.01	0.00	0.02	0.02	0.04	0.44
hsa-miR-191-5p	0.06	0.03	0.12	0.11	0.01	0.01	0.01	0.02	0.29	0.27	0.06
hsa-miR-210	0.06	0.05	0.06	0.13	0.03	0.03	0.01	0.06	0.05	0.15	0.37
hsa-miR-641	0.06	0.01	0.10	0.18	0.01	0.03	0.06	0.02	0.13	0.41	0.00
hsa-miR-101-3p	0.06	0.06	0.05	0.34	0.06	0.09	0.04	0.07	0.06	0.06	0.11
hsa-miR-345-5p	0.06	0.05	0.20	0.14	0.05	0.07	0.02	0.06	0.26	0.05	0.04

hsa-miR-7-1-3p	0.06	0.02	0.04	0.06	0.02	0.17	0.04	0.03	0.09	0.38	0.09
hsa-let-7i-3p	0.06	0.10	0.19	0.02	0.03	0.06	0.02	0.08	0.06	0.06	0.31
hsa-miR-455-5p	0.06	0.03	0.01	0.57	0.01	0.01	0.02	0.07	0.00	0.01	0.22
hsa-miR-34a-5p	0.06	0.17	0.03	0.02	0.02	0.08	0.04	0.21	0.01	0.00	0.35
hsa-miR-16-5p	0.06	0.06	0.04	0.03	0.01	0.01	0.01	0.04	0.39	0.26	0.12
hsa-miR-1307-3p	0.06	0.03	0.13	0.08	0.06	0.20	0.03	0.15	0.10	0.13	0.03
hsa-miR-329	0.06	0.01	0.01	0.03	0.00	0.83	0.01	0.02	0.01	0.00	0.04
hsa-miR-502-5p	0.05	0.09	0.29	0.08	0.04	0.03	0.04	0.09	0.11	0.08	0.10
hsa-miR-374a-5p	0.05	0.25	0.04	0.17	0.03	0.05	0.05	0.08	0.04	0.15	0.09
hsa-miR-29a-5p	0.05	0.20	0.03	0.09	0.07	0.21	0.07	0.12	0.02	0.02	0.12
hsa-miR-19b-3p	0.05	0.38	0.02	0.03	0.02	0.02	0.05	0.09	0.13	0.11	0.10
hsa-miR-181a-3p	0.05	0.02	0.04	0.01	0.01	0.15	0.01	0.03	0.08	0.57	0.03
hsa-miR-629-5p	0.05	0.30	0.06	0.01	0.00	0.00	0.01	0.02	0.11	0.28	0.15
hsa-miR-181a-5p	0.05	0.03	0.10	0.01	0.01	0.15	0.01	0.03	0.10	0.45	0.05
hsa-miR-181a-2-3p	0.05	0.02	0.36	0.03	0.06	0.07	0.03	0.03	0.04	0.17	0.15
hsa-miR-543	0.05	0.00	0.02	0.08	0.02	0.59	0.05	0.06	0.10	0.00	0.02
hsa-miR-26b-5p	0.04	0.05	0.02	0.10	0.04	0.03	0.06	0.15	0.17	0.23	0.11
hsa-miR-15a-5p	0.04	0.11	0.02	0.02	0.01	0.02	0.01	0.05	0.32	0.29	0.12
hsa-miR-3150b-5p	0.04	0.02	0.06	0.04	0.00	0.00	0.00	0.05	0.38	0.39	0.02
hsa-miR-4508	0.04	0.03	0.02	0.10	0.05	0.02	0.18	0.30	0.08	0.03	0.16
hsa-let-7a-5p	0.04	0.07	0.18	0.06	0.07	0.09	0.11	0.09	0.05	0.13	0.11
hsa-miR-29c-3p	0.04	0.05	0.01	0.16	0.10	0.15	0.10	0.14	0.10	0.12	0.03
hsa-miR-486-5p	0.04	0.13	0.00	0.02	0.27	0.02	0.06	0.29	0.05	0.01	0.10
hsa-miR-425-5p	0.04	0.26	0.02	0.03	0.01	0.04	0.01	0.03	0.18	0.12	0.26
hsa-miR-27a-3p	0.04	0.16	0.02	0.01	0.01	0.00	0.00	0.01	0.02	0.01	0.72
hsa-let-7f-5p	0.04	0.02	0.20	0.03	0.08	0.08	0.07	0.10	0.12	0.22	0.03
hsa-miR-103a-3p	0.04	0.02	0.23	0.10	0.02	0.11	0.02	0.05	0.13	0.23	0.05
hsa-miR-374b-5p	0.04	0.16	0.10	0.10	0.04	0.07	0.04	0.09	0.09	0.16	0.12
hsa-miR-10b-5p	0.04	0.23	0.04	0.00	0.00	0.00	0.00	0.02	0.00	0.00	0.67
hsa-miR-2110	0.04	0.02	0.05	0.07	0.10	0.09	0.05	0.14	0.23	0.18	0.03
hsa-miR-24-3p	0.04	0.20	0.03	0.01	0.03	0.01	0.01	0.04	0.04	0.01	0.58
hsa-miR-222-3p	0.04	0.02	0.29	0.05	0.03	0.13	0.01	0.05	0.23	0.10	0.05
hsa-miR-19a-3p	0.03	0.80	0.01	0.01	0.00	0.00	0.01	0.01	0.02	0.04	0.07
hsa-miR-185-3p	0.03	0.02	0.53	0.06	0.05	0.12	0.01	0.03	0.07	0.06	0.02
hsa-miR-217	0.03	0.00	0.00	0.00	0.00	0.00	0.96	0.00	0.00	0.00	0.00
hsa-miR-548k	0.03	0.02	0.24	0.09	0.04	0.02	0.01	0.03	0.20	0.28	0.01

hsa-miR-331-5p	0.03	0.01	0.07	0.04	0.09	0.35	0.02	0.14	0.03	0.19	0.03
hsa-let-7i-5p	0.03	0.02	0.42	0.01	0.02	0.03	0.02	0.05	0.18	0.19	0.03
hsa-miR-625-3p	0.03	0.02	0.24	0.06	0.02	0.05	0.05	0.03	0.09	0.34	0.06
hsa-miR-130b-3p	0.03	0.01	0.13	0.04	0.00	0.01	0.05	0.01	0.45	0.26	0.02
hsa-miR-3928	0.03	0.03	0.22	0.06	0.03	0.13	0.04	0.04	0.25	0.14	0.03
hsa-miR-495	0.03	0.00	0.00	0.10	0.01	0.66	0.04	0.09	0.04	0.00	0.02
hsa-miR-92b-5p	0.03	0.00	0.52	0.01	0.01	0.20	0.04	0.08	0.09	0.01	0.00
hsa-miR-195-5p	0.03	0.13	0.01	0.11	0.01	0.02	0.02	0.08	0.03	0.02	0.54
hsa-miR-425-3p	0.02	0.03	0.05	0.03	0.03	0.20	0.01	0.05	0.46	0.06	0.04
hsa-miR-320a	0.02	0.01	0.45	0.09	0.04	0.03	0.04	0.03	0.16	0.11	0.02
hsa-miR-4792	0.02	0.07	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.89
hsa-let-7b-5p	0.02	0.06	0.09	0.06	0.10	0.13	0.26	0.17	0.04	0.02	0.05
hsa-miR-29b-3p	0.02	0.12	0.25	0.04	0.06	0.22	0.03	0.14	0.04	0.03	0.06
hsa-miR-29c-5p	0.02	0.09	0.00	0.07	0.15	0.33	0.08	0.16	0.02	0.04	0.04
hsa-miR-221-3p	0.02	0.01	0.29	0.02	0.07	0.18	0.03	0.06	0.16	0.12	0.04
hsa-miR-503	0.02	0.01	0.27	0.11	0.07	0.01	0.01	0.06	0.40	0.04	0.01
hsa-miR-93-5p	0.02	0.09	0.04	0.04	0.01	0.03	0.02	0.07	0.25	0.26	0.16
hsa-miR-335-5p	0.02	0.85	0.00	0.03	0.00	0.03	0.02	0.02	0.01	0.01	0.01
hsa-miR-378a-3p	0.02	0.19	0.21	0.12	0.11	0.00	0.01	0.02	0.09	0.05	0.17
hsa-miR-766-3p	0.02	0.06	0.05	0.02	0.01	0.10	0.01	0.03	0.19	0.42	0.09
hsa-miR-216b	0.02	0.00	0.00	0.00	0.00	0.00	0.98	0.00	0.00	0.00	0.00
hsa-miR-181d	0.02	0.00	0.02	0.01	0.02	0.58	0.05	0.04	0.21	0.05	0.01
hsa-miR-760	0.02	0.01	0.09	0.00	0.03	0.81	0.01	0.02	0.01	0.01	0.00
hsa-miR-34b-5p	0.02	0.08	0.05	0.02	0.03	0.22	0.04	0.38	0.04	0.06	0.07
hsa-miR-126-5p	0.02	0.72	0.00	0.04	0.02	0.01	0.01	0.03	0.00	0.00	0.16
hsa-miR-15b-3p	0.02	0.05	0.06	0.03	0.04	0.04	0.01	0.10	0.11	0.28	0.27
hsa-miR-4492	0.02	0.01	0.01	0.09	0.03	0.02	0.21	0.41	0.05	0.01	0.13
hsa-miR-887	0.02	0.08	0.07	0.02	0.12	0.41	0.04	0.16	0.00	0.01	0.07
hsa-miR-505-3p	0.02	0.20	0.03	0.14	0.02	0.02	0.01	0.06	0.19	0.09	0.21
hsa-miR-99a-3p	0.02	0.09	0.01	0.17	0.11	0.23	0.07	0.05	0.01	0.00	0.24
hsa-miR-708-5p	0.02	0.38	0.00	0.01	0.01	0.23	0.02	0.03	0.01	0.02	0.29
hsa-miR-25-3p	0.02	0.00	0.17	0.04	0.03	0.03	0.06	0.07	0.15	0.42	0.01
hsa-miR-145-3p	0.02	0.64	0.00	0.03	0.03	0.02	0.04	0.05	0.00	0.00	0.16
hsa-miR-664-5p	0.02	0.02	0.05	0.13	0.11	0.16	0.18	0.12	0.05	0.15	0.02
hsa-miR-23a-3p	0.02	0.08	0.09	0.06	0.04	0.01	0.02	0.06	0.11	0.01	0.50
hsa-miR-485-5p	0.02	0.00	0.00	0.04	0.01	0.83	0.03	0.03	0.02	0.00	0.01

hsa-miR-29a-3p	0.02	0.02	0.31	0.14	0.04	0.14	0.03	0.08	0.08	0.08	0.05
hsa-miR-152	0.01	0.11	0.32	0.07	0.08	0.02	0.04	0.06	0.06	0.03	0.21
hsa-miR-320d	0.01	0.01	0.25	0.12	0.06	0.08	0.08	0.07	0.23	0.06	0.02
hsa-miR-342-5p	0.01	0.04	0.01	0.05	0.01	0.10	0.01	0.09	0.21	0.44	0.03
hsa-miR-424-5p	0.01	0.16	0.03	0.27	0.02	0.01	0.01	0.12	0.30	0.02	0.04
hsa-let-7d-5p	0.01	0.01	0.37	0.14	0.04	0.04	0.03	0.03	0.16	0.14	0.02
hsa-miR-221-5p	0.01	0.00	0.11	0.03	0.04	0.11	0.03	0.05	0.36	0.24	0.02
hsa-miR-576-5p	0.01	0.05	0.06	0.10	0.04	0.01	0.04	0.03	0.10	0.45	0.11
hsa-miR-15b-5p	0.01	0.02	0.04	0.02	0.00	0.00	0.00	0.03	0.44	0.34	0.09
hsa-miR-181b-5p	0.01	0.01	0.13	0.01	0.01	0.25	0.02	0.06	0.16	0.32	0.02
hsa-miR-30b-3p	0.01	0.02	0.08	0.32	0.13	0.06	0.08	0.06	0.04	0.18	0.02
hsa-miR-4420	0.01	0.00	0.00	0.02	0.00	0.00	0.00	0.02	0.59	0.34	0.01
hsa-miR-215	0.01	0.04	0.00	0.74	0.01	0.01	0.11	0.05	0.02	0.02	0.00
hsa-miR-365a-5p	0.01	0.02	0.66	0.14	0.02	0.01	0.03	0.02	0.05	0.02	0.02
hsa-miR-490-5p	0.01	0.01	0.02	0.41	0.28	0.21	0.04	0.01	0.01	0.00	0.01
hsa-miR-146a-5p	0.01	0.15	0.40	0.02	0.00	0.00	0.00	0.02	0.02	0.17	0.19
hsa-miR-214-3p	0.01	0.07	0.00	0.47	0.02	0.00	0.07	0.14	0.00	0.00	0.21
hsa-let-7g-5p	0.01	0.02	0.04	0.01	0.02	0.03	0.02	0.02	0.39	0.41	0.02
hsa-miR-330-5p	0.01	0.00	0.23	0.01	0.01	0.54	0.02	0.02	0.09	0.08	0.00
hsa-miR-548e	0.01	0.01	0.07	0.08	0.02	0.01	0.02	0.02	0.35	0.40	0.01
hsa-miR-22-5p	0.01	0.13	0.12	0.22	0.14	0.08	0.04	0.13	0.06	0.03	0.05
hsa-miR-1306-3p	0.01	0.01	0.08	0.05	0.20	0.20	0.04	0.05	0.24	0.12	0.00
hsa-miR-4448	0.01	0.03	0.07	0.06	0.04	0.12	0.19	0.14	0.17	0.02	0.14
hsa-miR-106b-5p	0.01	0.06	0.06	0.01	0.01	0.01	0.01	0.03	0.37	0.24	0.21
hsa-miR-424-3p	0.01	0.00	0.18	0.32	0.05	0.01	0.01	0.07	0.31	0.04	0.01
hsa-miR-3656	0.01	0.05	0.00	0.03	0.08	0.02	0.11	0.14	0.01	0.00	0.55
hsa-miR-1246	0.01	0.00	0.36	0.02	0.06	0.00	0.06	0.10	0.15	0.09	0.14
hsa-miR-218-5p	0.01	0.44	0.00	0.00	0.00	0.08	0.01	0.05	0.00	0.00	0.40
hsa-let-7c	0.01	0.07	0.02	0.16	0.10	0.28	0.13	0.10	0.03	0.03	0.07
hsa-miR-1277-3p	0.01	0.07	0.09	0.09	0.04	0.11	0.05	0.07	0.28	0.17	0.02
hsa-miR-18a-5p	0.01	0.29	0.03	0.02	0.00	0.01	0.00	0.01	0.21	0.20	0.22
hsa-miR-365a-3p	0.01	0.06	0.14	0.05	0.13	0.06	0.06	0.05	0.00	0.05	0.39
hsa-miR-4521	0.01	0.01	0.71	0.01	0.00	0.00	0.00	0.00	0.10	0.15	0.01
hsa-miR-1255a	0.01	0.01	0.21	0.01	0.00	0.00	0.01	0.00	0.39	0.32	0.04
hsa-miR-940	0.01	0.01	0.09	0.03	0.02	0.48	0.03	0.15	0.11	0.05	0.01
hsa-miR-31-5p	0.01	0.00	0.63	0.01	0.00	0.17	0.08	0.05	0.03	0.01	0.01

hsa-miR-130a-3p	0.01	0.03	0.13	0.08	0.01	0.02	0.03	0.07	0.49	0.03	0.10
hsa-miR-4532	0.01	0.13	0.00	0.00	0.02	0.00	0.05	0.09	0.00	0.00	0.70
hsa-miR-1291	0.01	0.00	0.01	0.17	0.16	0.14	0.11	0.17	0.16	0.04	0.04
hsa-miR-9-5p	0.01	0.00	0.00	0.00	0.00	0.96	0.00	0.00	0.00	0.00	0.01
hsa-miR-128	0.01	0.00	0.04	0.01	0.02	0.76	0.00	0.01	0.02	0.13	0.00
hsa-miR-100-5p	0.01	0.12	0.02	0.18	0.00	0.03	0.01	0.01	0.00	0.00	0.61
hsa-miR-1271-5p	0.01	0.02	0.06	0.02	0.12	0.13	0.02	0.03	0.20	0.36	0.04
hsa-miR-193b-3p	0.01	0.21	0.05	0.07	0.12	0.03	0.07	0.03	0.00	0.02	0.38
hsa-miR-383	0.01	0.01	0.00	0.15	0.01	0.77	0.02	0.01	0.00	0.00	0.04
hsa-miR-378f	0.01	0.07	0.25	0.22	0.14	0.01	0.03	0.01	0.13	0.10	0.03
hsa-miR-190a	0.01	0.93	0.00	0.00	0.00	0.01	0.00	0.00	0.00	0.00	0.05
hsa-miR-224-5p	0.01	0.46	0.00	0.05	0.03	0.00	0.05	0.02	0.01	0.00	0.36
hsa-miR-139-5p	0.01	0.10	0.00	0.07	0.05	0.53	0.02	0.17	0.00	0.00	0.05
hsa-miR-497-5p	0.00	0.19	0.00	0.11	0.04	0.10	0.07	0.15	0.01	0.00	0.34
hsa-miR-92a-1-5p	0.00	0.00	0.53	0.05	0.01	0.00	0.01	0.01	0.14	0.23	0.00
hsa-miR-378a-5p	0.00	0.45	0.03	0.07	0.27	0.01	0.01	0.03	0.02	0.03	0.08
hsa-miR-20a-5p	0.00	0.14	0.01	0.01	0.00	0.00	0.00	0.01	0.15	0.13	0.54
hsa-miR-17-5p	0.00	0.08	0.02	0.03	0.00	0.00	0.00	0.02	0.41	0.21	0.21
hsa-miR-126-3p	0.00	0.66	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.33
hsa-miR-155-5p	0.00	0.05	0.01	0.01	0.00	0.00	0.01	0.04	0.12	0.69	0.06
hsa-miR-504	0.00	0.05	0.01	0.08	0.04	0.65	0.03	0.09	0.01	0.01	0.03
hsa-miR-99a-5p	0.00	0.07	0.00	0.60	0.03	0.10	0.03	0.02	0.01	0.00	0.13
hsa-miR-1255b-5p	0.00	0.01	0.11	0.12	0.07	0.02	0.05	0.07	0.16	0.37	0.01
hsa-miR-449c-5p	0.00	0.01	0.06	0.02	0.01	0.13	0.03	0.58	0.09	0.06	0.00
hsa-miR-1278	0.00	0.01	0.21	0.09	0.04	0.03	0.03	0.04	0.14	0.40	0.01
hsa-miR-33a-5p	0.00	0.05	0.09	0.06	0.02	0.24	0.04	0.04	0.32	0.10	0.02
hsa-miR-150-5p	0.00	0.28	0.00	0.02	0.00	0.00	0.00	0.02	0.08	0.35	0.24
hsa-miR-208b	0.00	0.00	0.00	0.01	0.96	0.01	0.01	0.00	0.00	0.00	0.00
hsa-miR-1261	0.00	0.01	0.32	0.25	0.00	0.01	0.00	0.01	0.10	0.08	0.20
hsa-miR-548j	0.00	0.03	0.10	0.21	0.04	0.03	0.19	0.05	0.17	0.16	0.02
hsa-miR-1827	0.00	0.01	0.10	0.03	0.06	0.06	0.06	0.05	0.17	0.45	0.00
hsa-miR-320b	0.00	0.00	0.54	0.11	0.05	0.04	0.04	0.03	0.14	0.04	0.00
hsa-miR-219-5p	0.00	0.00	0.00	0.00	0.00	0.97	0.01	0.00	0.00	0.00	0.00
hsa-miR-514a-3p	0.00	0.01	0.41	0.00	0.00	0.01	0.00	0.10	0.00	0.01	0.45
hsa-miR-4634	0.00	0.09	0.00	0.00	0.00	0.00	0.01	0.02	0.00	0.00	0.86
hsa-miR-935	0.00	0.01	0.00	0.00	0.01	0.95	0.00	0.00	0.00	0.00	0.01

hsa-miR-17-3p	0.00	0.01	0.06	0.03	0.01	0.01	0.02	0.03	0.67	0.15	0.02
hsa-miR-145-5p	0.00	0.18	0.00	0.06	0.09	0.03	0.09	0.23	0.01	0.00	0.30
hsa-miR-193a-5p	0.00	0.12	0.12	0.22	0.03	0.02	0.10	0.18	0.07	0.03	0.11
hsa-miR-138-5p	0.00	0.00	0.05	0.00	0.00	0.84	0.00	0.06	0.00	0.03	0.01
hsa-miR-199a-5p	0.00	0.03	0.01	0.40	0.05	0.00	0.13	0.14	0.09	0.00	0.14
hsa-miR-199b-5p	0.00	0.03	0.01	0.40	0.05	0.00	0.13	0.15	0.09	0.00	0.14
hsa-miR-140-5p	0.00	0.01	0.10	0.06	0.04	0.11	0.04	0.11	0.18	0.31	0.04
hsa-miR-4510	0.00	0.04	0.04	0.19	0.05	0.21	0.11	0.07	0.09	0.18	0.01
hsa-miR-23b-5p	0.00	0.01	0.19	0.22	0.18	0.09	0.09	0.13	0.04	0.03	0.01
hsa-miR-107	0.00	0.00	0.28	0.12	0.02	0.13	0.03	0.06	0.15	0.20	0.00
hsa-miR-944	0.00	0.01	0.00	0.04	0.11	0.00	0.00	0.01	0.00	0.03	0.79
hsa-miR-885-5p	0.00	0.00	0.00	0.43	0.02	0.54	0.00	0.00	0.00	0.00	0.00
hsa-miR-150-3p	0.00	0.04	0.00	0.00	0.01	0.01	0.00	0.01	0.23	0.68	0.01
hsa-miR-4429	0.00	0.01	0.50	0.10	0.02	0.02	0.02	0.01	0.23	0.09	0.00
hsa-miR-584-5p	0.00	0.02	0.36	0.00	0.03	0.50	0.01	0.03	0.04	0.00	0.01
hsa-miR-106a-5p	0.00	0.10	0.02	0.01	0.00	0.01	0.00	0.02	0.50	0.26	0.08
hsa-miR-133a	0.00	0.00	0.00	0.00	0.98	0.00	0.00	0.01	0.00	0.00	0.00
hsa-miR-378e	0.00	0.02	0.26	0.28	0.22	0.01	0.04	0.03	0.09	0.04	0.01
hsa-miR-505-5p	0.00	0.01	0.17	0.51	0.02	0.01	0.01	0.04	0.09	0.12	0.01
hsa-miR-320c	0.00	0.00	0.53	0.11	0.05	0.05	0.04	0.04	0.16	0.02	0.00
hsa-miR-363-3p	0.00	0.05	0.06	0.00	0.00	0.01	0.01	0.01	0.36	0.47	0.03
hsa-miR-143-5p	0.00	0.42	0.00	0.05	0.08	0.04	0.09	0.12	0.01	0.00	0.17
hsa-miR-185-5p	0.00	0.04	0.29	0.10	0.07	0.12	0.03	0.11	0.14	0.08	0.01
hsa-miR-193b-5p	0.00	0.04	0.18	0.20	0.24	0.07	0.16	0.02	0.00	0.07	0.02
hsa-miR-20b-5p	0.00	0.15	0.03	0.00	0.01	0.01	0.01	0.02	0.30	0.32	0.14
hsa-miR-9-3p	0.00	0.00	0.01	0.00	0.00	0.97	0.00	0.00	0.00	0.00	0.00
hsa-miR-211-5p	0.00	0.00	0.98	0.00	0.00	0.01	0.00	0.00	0.00	0.00	0.02
hsa-miR-452-5p	0.00	0.58	0.00	0.01	0.04	0.00	0.04	0.04	0.00	0.00	0.29
hsa-miR-223-5p	0.00	0.01	0.00	0.01	0.00	0.00	0.00	0.02	0.86	0.09	0.00
hsa-miR-509-3-5p	0.00	0.00	0.92	0.00	0.00	0.00	0.00	0.05	0.00	0.00	0.01
hsa-miR-378b	0.00	0.01	0.52	0.22	0.05	0.01	0.01	0.01	0.10	0.07	0.01
hsa-miR-203	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00
hsa-miR-509-5p	0.00	0.00	0.93	0.00	0.00	0.00	0.00	0.04	0.00	0.00	0.01
hsa-miR-378c	0.00	0.02	0.36	0.18	0.20	0.01	0.02	0.03	0.12	0.05	0.01
hsa-miR-25-5p	0.00	0.00	0.44	0.02	0.00	0.01	0.01	0.01	0.25	0.27	0.00
hsa-miR-142-5p	0.00	0.49	0.00	0.01	0.00	0.00	0.00	0.03	0.12	0.16	0.19

hsa-miR-144-5p	0.00	0.49	0.00	0.01	0.01	0.00	0.01	0.03	0.00	0.00	0.45
hsa-miR-144-3p	0.00	0.83	0.00	0.00	0.01	0.00	0.01	0.05	0.03	0.00	0.05
hsa-miR-2964a-5p	0.00	0.00	0.00	0.00	0.00	0.98	0.00	0.00	0.00	0.00	0.00
hsa-miR-205-5p	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00
hsa-miR-499a-5p	0.00	0.00	0.00	0.01	0.96	0.01	0.00	0.01	0.00	0.00	0.00
hsa-miR-223-3p	0.00	0.06	0.00	0.01	0.00	0.01	0.01	0.06	0.77	0.04	0.04
hsa-miR-196a-5p	0.00	0.29	0.02	0.00	0.01	0.00	0.00	0.03	0.00	0.00	0.64
hsa-miR-133b	0.00	0.00	0.00	0.00	0.98	0.00	0.00	0.01	0.00	0.00	0.00
hsa-miR-196b-5p	0.00	0.36	0.01	0.00	0.00	0.00	0.00	0.04	0.09	0.02	0.48
hsa-miR-508-5p	0.00	0.00	0.91	0.00	0.00	0.00	0.00	0.04	0.00	0.00	0.04
hsa-miR-124-3p	0.00	0.00	0.00	0.00	0.00	0.99	0.00	0.01	0.00	0.00	0.00
hsa-miR-378d	0.00	0.00	0.41	0.15	0.26	0.01	0.02	0.03	0.08	0.05	0.00
hsa-miR-451a	0.00	0.36	0.00	0.02	0.02	0.00	0.02	0.14	0.01	0.00	0.43
hsa-miR-206	0.00	0.00	0.00	0.00	0.96	0.00	0.01	0.03	0.00	0.00	0.00
hsa-miR-1	0.00	0.00	0.00	0.00	0.99	0.00	0.00	0.00	0.00	0.00	0.00
hsa-miR-3591-5p	0.00	0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
hsa-miR-122-5p	0.00	0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
hsa-miR-20b-3p	0.00	0.01	0.23	0.01	0.00	0.03	0.01	0.02	0.31	0.38	0.00
hsa-miR-363-5p	0.00	0.01	0.29	0.01	0.00	0.02	0.01	0.05	0.18	0.44	0.00
hsa-miR-122-3p	0.00	0.00	0.00	0.98	0.02	0.00	0.00	0.00	0.00	0.00	0.00

Appendix Table 3

Results for the significantly enriched predicted target gene sets for islet-expressed miRNAs from the MAGENTA analysis with the 75th percentile *P*-value cut-off used to determine significance.

Prediction type	miRNA target-gene set	Effective genes in set	Significant genes expected	Significant genes observed	Enrichment value	Enrichment *P*-value	Enrichment *q*-value
MiRanda	hsa-miR-424-5p	1160	290	344	1.19	9.90E-06	2.82E-03
MiRanda	hsa-miR-23a-3p	1425	356	413	1.16	1.00E-05	2.82E-03
MiRanda	hsa-miR-23b-3p	1436	359	410	1.14	9.00E-05	2.82E-03
MiRanda	hsa-miR-224-5p	1329	332	381	1.15	1.40E-04	2.82E-03
MiRanda	hsa-miR-376c-3p	1351	338	384	1.14	1.40E-04	2.82E-03
MiRanda	hsa-miR-10b-5p	791	198	237	1.20	3.30E-04	2.82E-03
MiRanda	hsa-miR-143-3p	1204	301	343	1.14	4.40E-04	2.82E-03
MiRanda	hsa-miR-214-3p	1204	301	344	1.14	4.60E-04	2.82E-03
MiRanda	hsa-miR-149-5p	1003	251	291	1.16	5.20E-04	2.82E-03
MiRanda	hsa-miR-181a-5p	1474	369	411	1.11	6.00E-04	2.82E-03
MiRanda	hsa-miR-22-3p	764	191	227	1.19	6.20E-04	2.82E-03
MiRanda	hsa-miR-377-3p	1495	374	416	1.11	7.30E-04	2.82E-03
MiRanda	hsa-miR-185-5p	1391	348	389	1.12	8.80E-04	2.82E-03
MiRanda	hsa-miR-212-3p	1279	320	359	1.12	9.20E-04	2.82E-03
MiRanda	hsa-miR-329	1021	255	293	1.15	9.20E-04	2.82E-03
MiRanda	hsa-miR-15b-5p	1159	290	328	1.13	1.04E-03	2.82E-03
MiRanda	hsa-miR-199b-5p	1054	264	300	1.14	1.27E-03	2.82E-03
MiRanda	hsa-miR-141-3p	1473	368	407	1.11	1.37E-03	2.82E-03
MiRanda	hsa-miR-330-5p	961	240	276	1.15	1.49E-03	2.82E-03
MiRanda	hsa-miR-10a-5p	842	211	244	1.16	1.55E-03	2.82E-03
MiRanda	hsa-miR-199a-5p	994	249	283	1.14	1.68E-03	2.82E-03
MiRanda	hsa-miR-98-5p	691	173	203	1.17	2.05E-03	2.82E-03
MiRanda	hsa-miR-132-3p	1360	340	377	1.11	2.22E-03	2.82E-03
MiRanda	hsa-let-7b-5p	744	186	217	1.17	2.36E-03	2.82E-03
MiRanda	hsa-miR-200a-3p	1416	354	390	1.10	2.76E-03	2.82E-03
MiRanda	hsa-miR-134	910	228	260	1.14	2.80E-03	2.82E-03
MiRanda	hsa-miR-181b-5p	1613	403	439	1.09	2.83E-03	2.82E-03
MiRanda	hsa-miR-217	1202	301	335	1.11	2.98E-03	2.82E-03

MiRanda	hsa-let-7c	714	179	208	1.16	3.03E-03	2.82E-03
MiRanda	hsa-miR-320c	1363	341	375	1.10	3.71E-03	2.82E-03
MiRanda	hsa-miR-16-5p	1152	288	321	1.11	3.88E-03	2.82E-03
MiRanda	hsa-miR-92a-3p	856	214	244	1.14	3.88E-03	2.82E-03
MiRanda	hsa-miR-320d	1318	330	363	1.10	3.88E-03	2.82E-03
MiRanda	hsa-let-7e-5p	794	199	228	1.15	4.08E-03	2.82E-03
MiRanda	hsa-let-7a-5p	707	177	205	1.16	4.16E-03	2.82E-03
MiRanda	hsa-miR-18a-5p	989	247	278	1.13	4.49E-03	2.82E-03
MiRanda	hsa-let-7d-5p	699	175	202	1.15	4.56E-03	2.82E-03
MiRanda	hsa-miR-320b	1414	354	387	1.09	4.72E-03	2.82E-03
MiRanda	hsa-miR-24-3p	857	214	244	1.14	4.78E-03	2.82E-03
MiRanda	hsa-miR-320a	1414	354	386	1.09	5.38E-03	2.82E-03
MiRanda	hsa-miR-150-5p	1003	251	281	1.12	5.55E-03	2.82E-03
MiRanda	hsa-miR-15a-5p	1148	287	318	1.11	5.94E-03	2.82E-03
MiRanda	hsa-miR-485-5p	1254	314	345	1.10	6.05E-03	2.82E-03
MiRanda	hsa-miR-195-5p	1142	286	316	1.10	6.34E-03	2.82E-03
MiRanda	hsa-miR-181c-5p	1379	345	376	1.09	6.78E-03	2.82E-03
MiRanda	hsa-miR-551b-3p	153	38	52	1.37	6.79E-03	2.82E-03
MiRanda	hsa-miR-154-5p	994	249	277	1.11	7.42E-03	2.86E-03
MiRanda	hsa-miR-92b-3p	796	199	226	1.14	7.65E-03	2.86E-03
MiRanda	hsa-miR-136-5p	1489	372	403	1.08	7.90E-03	2.86E-03
MiRanda	hsa-miR-497-5p	1051	263	291	1.11	9.02E-03	3.01E-03
MiRanda	hsa-miR-181d	1729	432	462	1.07	9.34E-03	3.02E-03
TargetScan	hsa-miR-26b-5p	252	63	89	1.41	1.40E-04	2.34E-02
TargetScan	hsa-miR-26a-5p	254	64	88	1.38	2.80E-04	2.34E-02
TargetScan	hsa-miR-24-3p	281	70	92	1.31	1.63E-03	2.34E-02
TargetScan	hsa-miR-454-3p	184	46	63	1.37	2.86E-03	2.34E-02
TargetScan	hsa-miR-874	145	36	51	1.42	3.58E-03	2.34E-02
TargetScan	hsa-miR-455-5p	79	20	30	1.50	6.95E-03	2.34E-02
TargetScan	hsa-miR-194-5p	141	35	48	1.37	9.09E-03	2.34E-02
TargetScan	hsa-miR-134	94	24	34	1.42	9.68E-03	2.34E-02
TargetScan	hsa-miR-130a-3p	187	47	61	1.30	9.83E-03	2.34E-02
MiRanda+TargetScan	hsa-miR-93-5p	873	218	257	1.18	3.70E-04	9.19E-03
MiRanda+TargetScan	hsa-miR-200b-3p	780	195	232	1.19	4.80E-04	9.19E-03
MiRanda+TargetScan	hsa-miR-200c-3p	784	196	232	1.18	5.10E-04	9.19E-03
MiRanda+TargetScan	hsa-miR-20a-5p	864	216	253	1.17	6.40E-04	9.19E-03
MiRanda+TargetScan	hsa-miR-429	783	196	232	1.18	7.00E-04	9.19E-03
MiRanda+TargetScan	hsa-miR-101-3p	533	133	164	1.23	7.00E-04	9.19E-03

MiRanda+TargetScan	hsa-miR-106b-5p	865	216	251	1.16	1.29E-03	9.19E-03
MiRanda+TargetScan	hsa-miR-27a-3p	723	181	213	1.18	1.44E-03	9.19E-03
MiRanda+TargetScan	hsa-miR-17-5p	867	217	250	1.15	1.93E-03	9.19E-03
MiRanda+TargetScan	hsa-miR-27b-3p	723	181	211	1.17	2.35E-03	9.19E-03
MiRanda+TargetScan	hsa-miR-539-5p	423	106	130	1.23	2.83E-03	9.19E-03
MiRanda+TargetScan	hsa-miR-98-5p	515	129	154	1.19	3.62E-03	9.19E-03
MiRanda+TargetScan	hsa-let-7c	522	131	156	1.19	3.65E-03	9.19E-03
MiRanda+TargetScan	hsa-miR-301b	651	163	190	1.17	3.95E-03	9.19E-03
MiRanda+TargetScan	hsa-miR-543	519	130	155	1.19	4.03E-03	9.19E-03
MiRanda+TargetScan	hsa-miR-340-5p	957	239	270	1.13	4.43E-03	9.19E-03
MiRanda+TargetScan	hsa-miR-590-5p	209	52	69	1.33	5.07E-03	9.19E-03
MiRanda+TargetScan	hsa-miR-219-5p	213	53	70	1.32	5.28E-03	9.19E-03
MiRanda+TargetScan	hsa-miR-130b-3p	639	160	186	1.16	5.36E-03	9.19E-03
MiRanda+TargetScan	hsa-let-7b-5p	523	131	155	1.18	5.39E-03	9.19E-03
MiRanda+TargetScan	hsa-miR-197-3p	111	28	40	1.43	5.64E-03	9.19E-03
MiRanda+TargetScan	hsa-miR-23a-3p	794	199	226	1.14	6.16E-03	9.19E-03
MiRanda+TargetScan	hsa-miR-301a-3p	649	162	188	1.16	6.60E-03	9.19E-03
MiRanda+TargetScan	hsa-miR-130a-3p	639	160	185	1.16	7.00E-03	9.19E-03
MiRanda+TargetScan	hsa-let-7a-5p	519	130	153	1.18	7.21E-03	9.19E-03
MiRanda+TargetScan	hsa-miR-34a-5p	372	93	113	1.22	8.21E-03	9.19E-03
MiRanda+TargetScan	hsa-miR-421	289	72	90	1.25	8.51E-03	9.19E-03
MiRanda+TargetScan	hsa-let-7d-5p	518	130	152	1.17	8.85E-03	9.19E-03
MiRanda+TargetScan	hsa-miR-454-3p	643	161	185	1.15	9.36E-03	9.19E-03
MiRanda+TargetScan	hsa-miR-23b-3p	792	198	224	1.13	9.88E-03	9.19E-03
MiRanda+miRDB	hsa-miR-374a-5p	749	187	216	1.16	4.09E-03	5.73E-02
MiRanda+miRDB	hsa-miR-130b-3p	378	95	115	1.21	6.70E-03	5.73E-02
MiRanda+miRDB	hsa-miR-155-5p	332	83	102	1.23	8.31E-03	5.73E-02
MiRanda+miRDB	hsa-miR-129-5p	652	163	187	1.15	9.55E-03	5.73E-02
MiRanda+miRDB+TargetScan	hsa-miR-17-5p	476	119	143	1.20	4.43E-03	8.55E-02
MiRanda+miRDB+TargetScan	hsa-miR-93-5p	465	116	140	1.21	4.54E-03	8.55E-02
MiRanda+miRDB+TargetScan	hsa-miR-20a-5p	476	119	141	1.18	8.65E-03	8.55E-02

References

1. Danaei, G. *et al.* National, regional, and global trends in fasting plasma glucose and diabetes prevalence since 1980: systematic analysis of health examination surveys and epidemiological studies with 370 country-years and 2.7 million participants. *Lancet* **378**, 31-40 (2011).
2. International Diabetes Federation. IDF Diabetes Atlas, 5th edn. (International Diabetes Federation, Brussels, Belgium, 2011).
3. World Health Organisation. Global status report on noncommunicable diseases 2010. (World Health Organisation, Geneva, 2011).
4. Hex, N., Bartlett, C., Wright, D., Taylor, M. & Varley, D. Estimating the current and future costs of Type 1 and Type 2 diabetes in the UK, including direct health costs and indirect societal and productivity costs. *Diabet Med* **29**, 855-62 (2012).
5. American Diabetes Association. Diagnosis and classification of diabetes mellitus. *Diabetes Care* **36 Suppl 1**, S67-74 (2013).
6. World Health Organisation. Use of glycated haemoglobin (HbA1c) in the diagnosis of diabetes mellitus. *Diabetes Research and Clinical Practice* **93**, 299-309 (2011).
7. World Health Organisation. Definition and diagnosis of diabetes mellitus and intermediate hyperglycaemia. (World Health Organisation, Geneva, 2006).
8. Stamler, J., Vaccaro, O., Neaton, J.D. & Wentworth, D. Diabetes, other risk factors, and 12-yr cardiovascular mortality for men screened in the Multiple Risk Factor Intervention Trial. *Diabetes Care* **16**, 434-44 (1993).
9. Holman, R.R., Paul, S.K., Bethel, M.A., Matthews, D.R. & Neil, H.A. 10-year follow-up of intensive glucose control in type 2 diabetes. *N Engl J Med* **359**, 1577-89 (2008).
10. Turner, R.C. *et al.* Risk factors for coronary artery disease in non-insulin dependent diabetes mellitus: United Kingdom Prospective Diabetes Study (UKPDS: 23). *BMJ* **316**, 823-8 (1998).
11. Reaven, G.M. Banting lecture 1988. Role of insulin resistance in human disease. *Diabetes* **37**, 1595-607 (1988).
12. Dinneen, S., Gerich, J. & Rizza, R. Carbohydrate metabolism in non-insulin-dependent diabetes mellitus. *N Engl J Med* **327**, 707-13 (1992).
13. Yoshimasa, Y. *et al.* Insulin-resistant diabetes due to a point mutation that prevents insulin proreceptor processing. *Science* **240**, 784-7 (1988).
14. Kobayashi, M. *et al.* Insulin resistance by unprocessed insulin proreceptors point mutation at the cleavage site. *Biochem Biophys Res Commun* **153**, 657-63 (1988).
15. Moller, D.E. & Flier, J.S. Detection of an alteration in the insulin-receptor gene in a patient with insulin resistance, acanthosis nigricans, and the polycystic ovary syndrome (type A insulin resistance). *N Engl J Med* **319**, 1526-9 (1988).
16. George, S. *et al.* A family with severe insulin resistance and diabetes due to a mutation in AKT2. *Science* **304**, 1325-8 (2004).
17. Agarwal, A.K. & Garg, A. A novel heterozygous mutation in peroxisome proliferator-activated receptor-gamma gene in a patient with familial partial lipodystrophy. *J Clin Endocrinol Metab* **87**, 408-11 (2002).
18. Cao, H. & Hegele, R.A. Nuclear lamin A/C R482Q mutation in canadian kindreds with Dunnigan-type familial partial lipodystrophy. *Hum Mol Genet* **9**, 109-12 (2000).
19. Shackleton, S. *et al.* LMNA, encoding lamin A/C, is mutated in partial lipodystrophy. *Nat Genet* **24**, 153-6 (2000).
20. Reaven, G.M. Pathophysiology of insulin resistance in human disease. *Physiol Rev* **75**, 473-86 (1995).
21. Zimmet, P., Alberti, K.G. & Shaw, J. Global and societal implications of the diabetes epidemic. *Nature* **414**, 782-7 (2001).

22. Brethauer, S.A. *et al.* Can diabetes be surgically cured? Long-term metabolic effects of bariatric surgery in obese patients with type 2 diabetes mellitus. *Ann Surg* **258**, 628-36; discussion 636-7 (2013).
23. Kitabchi, A.E. *et al.* Role of insulin secretion and sensitivity in the evolution of type 2 diabetes in the diabetes prevention program: effects of lifestyle intervention and metformin. *Diabetes* **54**, 2404-14 (2005).
24. Pal, A. *et al.* PTEN mutations as a cause of constitutive insulin sensitivity and obesity. *N Engl J Med* **367**, 1002-11 (2012).
25. Yki-Jarvinen, H. Thiazolidinediones. *N Engl J Med* **351**, 1106-18 (2004).
26. Nissen, S.E. & Wolski, K. Effect of rosiglitazone on the risk of myocardial infarction and death from cardiovascular causes. *N Engl J Med* **356**, 2457-71 (2007).
27. Effect of intensive blood-glucose control with metformin on complications in overweight patients with type 2 diabetes (UKPDS 34). UK Prospective Diabetes Study (UKPDS) Group. *Lancet* **352**, 854-65 (1998).
28. Kim, Y.D. *et al.* Metformin inhibits hepatic gluconeogenesis through AMP-activated protein kinase-dependent regulation of the orphan nuclear receptor SHP. *Diabetes* **57**, 306-14 (2008).
29. Stumvoll, M., Tataranni, P.A., Stefan, N., Vozarova, B. & Bogardus, C. Glucose allostasis. *Diabetes* **52**, 903-9 (2003).
30. McCulloch, L.J. *et al.* GLUT2 (SLC2A2) is not the principal glucose transporter in human pancreatic beta cells: implications for understanding genetic association signals at this locus. *Mol Genet Metab* **104**, 648-53 (2011).
31. Matschinsky, F.M. Regulation of pancreatic beta-cell glucokinase: from basics to therapeutics. *Diabetes* **51 Suppl 3**, S394-404 (2002).
32. Ashcroft, F.M., Harrison, D.E. & Ashcroft, S.J. Glucose induces closure of single potassium channels in isolated rat pancreatic beta-cells. *Nature* **312**, 446-8 (1984).
33. Gloyn, A.L. *et al.* Activating mutations in the gene encoding the ATP-sensitive potassium-channel subunit Kir6.2 and permanent neonatal diabetes. *N Engl J Med* **350**, 1838-49 (2004).
34. Babenko, A.P. *et al.* Activating mutations in the ABCC8 gene in neonatal diabetes mellitus. *N Engl J Med* **355**, 456-66 (2006).
35. Kapoor, R.R. *et al.* Hyperinsulinaemic hypoglycaemia and diabetes mellitus due to dominant ABCC8/KCNJ11 mutations. *Diabetologia* **54**, 2575-83 (2011).
36. Froguel, P. *et al.* Close linkage of glucokinase locus on chromosome 7p to early-onset non-insulin-dependent diabetes mellitus. *Nature* **356**, 162-4 (1992).
37. Byrne, M.M. *et al.* Altered insulin secretory responses to glucose in diabetic and nondiabetic subjects with mutations in the diabetes susceptibility gene MODY3 on chromosome 12. *Diabetes* **45**, 1503-10 (1996).
38. Njolstad, P.R. *et al.* Neonatal diabetes mellitus due to complete glucokinase deficiency. *N Engl J Med* **344**, 1588-92 (2001).
39. Glaser, B. *et al.* Familial hyperinsulinism caused by an activating glucokinase mutation. *N Engl J Med* **338**, 226-30 (1998).
40. Kaprio, J. *et al.* Concordance for type 1 (insulin-dependent) and type 2 (non-insulin-dependent) diabetes mellitus in a population-based cohort of twins in Finland. *Diabetologia* **35**, 1060-7 (1992).
41. Newman, B. *et al.* Concordance for type 2 (non-insulin-dependent) diabetes mellitus in male twins. *Diabetologia* **30**, 763-8 (1987).
42. Köbberling, J. & Tillil, H. Empirical risk figures for first degree relatives of non-insulin dependent diabetics. in *The genetics of diabetes mellitus* (eds. Köbberling, J. & Tattersall, R.) 201–209 (Academic Press, London, 1982).
43. Almgren, P. *et al.* Heritability and familiality of type 2 diabetes and related quantitative traits in the Botnia Study. *Diabetologia* **54**, 2811-9 (2011).
44. Poulsen, P., Kyvik, K.O., Vaag, A. & Beck-Nielsen, H. Heritability of type II (non-insulin-dependent) diabetes mellitus and abnormal glucose tolerance--a population-based twin study. *Diabetologia* **42**, 139-45 (1999).

45. Altshuler, D. *et al.* The common PPARgamma Pro12Ala polymorphism is associated with decreased risk of type 2 diabetes. *Nat Genet* **26**, 76-80 (2000).
46. Gloyn, A.L. *et al.* Large-scale association studies of variants in genes encoding the pancreatic beta-cell KATP channel subunits Kir6.2 (KCNJ11) and SUR1 (ABCC8) confirm that the KCNJ11 E23K variant is associated with type 2 diabetes. *Diabetes* **52**, 568-72 (2003).
47. Hattersley, A.T. *et al.* Linkage of type 2 diabetes to the glucokinase gene. *Lancet* **339**, 1307-10 (1992).
48. Bell, G.I. *et al.* Gene for non-insulin-dependent diabetes mellitus (maturity-onset diabetes of the young subtype) is linked to DNA polymorphism on human chromosome 20q. *Proc Natl Acad Sci U S A* **88**, 1484-8 (1991).
49. Vaxillaire, M. *et al.* A gene for maturity onset diabetes of the young (MODY) maps to chromosome 12q. *Nat Genet* **9**, 418-23 (1995).
50. Yamagata, K. *et al.* Mutations in the hepatocyte nuclear factor-4alpha gene in maturity-onset diabetes of the young (MODY1). *Nature* **384**, 458-60 (1996).
51. Yamagata, K. *et al.* Mutations in the hepatocyte nuclear factor-1alpha gene in maturity-onset diabetes of the young (MODY3). *Nature* **384**, 455-8 (1996).
52. Horikawa, Y. *et al.* Mutation in hepatocyte nuclear factor-1 beta gene (TCF2) associated with MODY. *Nat Genet* **17**, 384-5 (1997).
53. Stoffers, D.A., Ferrer, J., Clarke, W.L. & Habener, J.F. Early-onset type-II diabetes mellitus (MODY4) linked to IPF1. *Nat Genet* **17**, 138-9 (1997).
54. The International HapMap Project. *Nature* **426**, 789-96 (2003).
55. Dupuis, J. *et al.* New genetic loci implicated in fasting glucose homeostasis and their impact on type 2 diabetes risk. *Nat Genet* **42**, 105-16 (2010).
56. Morris, A.P. *et al.* Large-scale association analysis provides insights into the genetic architecture and pathophysiology of type 2 diabetes. *Nat Genet* **44**, 981-90 (2012).
57. Voight, B.F. *et al.* Twelve type 2 diabetes susceptibility loci identified through large-scale association analysis. *Nat Genet* **42**, 579-89 (2010).
58. Cho, Y.S. *et al.* Meta-analysis of genome-wide association studies identifies eight new loci for type 2 diabetes in east Asians. *Nat Genet* **44**, 67-72 (2012).
59. Yamauchi, T. *et al.* A genome-wide association study in the Japanese population identifies susceptibility loci for type 2 diabetes at UBE2E2 and C2CD4A-C2CD4B. *Nat Genet* **42**, 864-8 (2010).
60. Zeggini, E. *et al.* Replication of genome-wide association signals in UK samples reveals risk loci for type 2 diabetes. *Science* **316**, 1336-41 (2007).
61. Saxena, R. *et al.* Genome-wide association analysis identifies loci for type 2 diabetes and triglyceride levels. *Science* **316**, 1331-6 (2007).
62. Sladek, R. *et al.* A genome-wide association study identifies novel risk loci for type 2 diabetes. *Nature* **445**, 881-5 (2007).
63. Kooner, J.S. *et al.* Genome-wide association study in individuals of South Asian ancestry identifies six new type 2 diabetes susceptibility loci. *Nat Genet* **43**, 984-9 (2011).
64. Grant, S.F. *et al.* Variant of transcription factor 7-like 2 (TCF7L2) gene confers risk of type 2 diabetes. *Nat Genet* **38**, 320-3 (2006).
65. Schaub, M.A., Boyle, A.P., Kundaje, A., Batzoglou, S. & Snyder, M. Linking disease associations with regulatory information in the human genome. *Genome Res* **22**, 1748-59 (2012).
66. Teslovich, T.M. *et al.* Biological, clinical and population relevance of 95 loci for blood lipids. *Nature* **466**, 707-13 (2010).
67. Samani, N.J. *et al.* Genomewide association analysis of coronary artery disease. *N Engl J Med* **357**, 443-53 (2007).
68. Kathiresan, S. *et al.* Genome-wide association of early-onset myocardial infarction with single nucleotide polymorphisms and copy number variants. *Nat Genet* **41**, 334-41 (2009).

69. Musunuru, K. *et al.* From noncoding variant to phenotype via SORT1 at the 1p13 cholesterol locus. *Nature* **466**, 714-9 (2010).
70. Gaulton, K.J. *et al.* A map of open chromatin in human pancreatic islets. *Nat Genet* **42**, 255-9 (2010).
71. Stitzel, M.L. *et al.* Global epigenomic analysis of primary human pancreatic islets provides insights into type 2 diabetes susceptibility loci. *Cell Metab* **12**, 443-55 (2010).
72. Travers, M.E. *et al.* Insights into the molecular mechanism for type 2 diabetes susceptibility at the KCNQ1 locus from temporal changes in imprinting status in human islets. *Diabetes* **62**, 987-92 (2013).
73. Baerenwald, D.A. *et al.* Multiple functional polymorphisms in the G6PC2 gene contribute to the association with higher fasting plasma glucose levels. *Diabetologia* **56**, 1306-16 (2013).
74. McCarthy, M.I. Genomics, type 2 diabetes, and obesity. *N Engl J Med* **363**, 2339-50 (2010).
75. Dimas, A.S. *et al.* Impact of type 2 diabetes susceptibility variants on quantitative glycemic traits reveals mechanistic heterogeneity. *Diabetes* **63**, 2158-71 (2014).
76. Sakura, H. *et al.* Sequence variations in the human Kir6.2 gene, a subunit of the beta-cell ATP-sensitive K-channel: no association with NIDDM in while Caucasian subjects or evidence of abnormal function when expressed in vitro. *Diabetologia* **39**, 1233-6 (1996).
77. Schwanstecher, C., Meyer, U. & Schwanstecher, M. K(IR)6.2 polymorphism predisposes to type 2 diabetes by inducing overactivity of pancreatic beta-cell ATP-sensitive K(+) channels. *Diabetes* **51**, 875-9 (2002).
78. Riedel, M.J., Boora, P., Steckley, D., de Vries, G. & Light, P.E. Kir6.2 polymorphisms sensitize beta-cell ATP-sensitive potassium channels to activation by acyl CoAs: a possible cellular mechanism for increased susceptibility to type 2 diabetes? *Diabetes* **52**, 2630-5 (2003).
79. Villareal, D.T. *et al.* Kir6.2 variant E23K increases ATP-sensitive K+ channel activity and is associated with impaired insulin release and enhanced insulin sensitivity in adults with normal glucose tolerance. *Diabetes* **58**, 1869-78 (2009).
80. Perry, J.R. *et al.* 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 Genet* **8**, e1002741 (2012).
81. Sankaran, V.G. *et al.* Cyclin D3 coordinates the cell cycle during differentiation to regulate erythrocyte size and number. *Genes Dev* **26**, 2075-87 (2012).
82. van der Harst, P. *et al.* Seventy-five genetic loci influencing the human red blood cell. *Nature* **492**, 369-75 (2012).
83. Beer, N.L. *et al.* The P446L variant in GCKR associated with fasting plasma glucose and triglyceride levels exerts its effect through increased glucokinase activity in liver. *Hum Mol Genet* **18**, 4081-8 (2009).
84. Rees, M.G. *et al.* Cellular characterisation of the GCKR P446L variant associated with type 2 diabetes risk. *Diabetologia* **55**, 114-22 (2012).
85. Bort, R., Martinez-Barbera, J.P., Beddington, R.S. & Zaret, K.S. Hex homeobox gene-dependent tissue positioning is required for organogenesis of the ventral pancreas. *Development* **131**, 797-806 (2004).
86. Steneberg, P. *et al.* The type 2 diabetes-associated gene ide is required for insulin secretion and suppression of alpha-synuclein levels in beta-cells. *Diabetes* **62**, 2004-14 (2013).
87. Dayeh, T.A. *et al.* Identification of CpG-SNPs associated with type 2 diabetes and differential DNA methylation in human pancreatic islets. *Diabetologia* **56**, 1036-46 (2013).
88. Gutierrez-Arcelus, M. *et al.* Passive and active DNA methylation and the interplay with genetic variation in gene regulation. *Elife* **2**, e00523 (2013).

89. Shu, L. *et al.* Transcription factor 7-like 2 regulates beta-cell survival and function in human pancreatic islets. *Diabetes* **57**, 645-53 (2008).
90. da Silva Xavier, G. *et al.* TCF7L2 regulates late events in insulin secretion from pancreatic islet beta-cells. *Diabetes* **58**, 894-905 (2009).
91. da Silva Xavier, G. *et al.* Abnormal glucose tolerance and insulin secretion in pancreas-specific Tcf7l2-null mice. *Diabetologia* **55**, 2667-76 (2012).
92. Boj, S.F. *et al.* Diabetes risk gene and Wnt effector Tcf7l2/TCF4 controls hepatic response to perinatal and adult metabolic demand. *Cell* **151**, 1595-607 (2012).
93. Lyssenko, V. *et al.* Mechanisms by which common variants in the TCF7L2 gene increase risk of type 2 diabetes. *J Clin Invest* **117**, 2155-63 (2007).
94. Le Bacquer, O. *et al.* TCF7L2 rs7903146 impairs islet function and morphology in non-diabetic individuals. *Diabetologia* **55**, 2677-81 (2012).
95. Rosengren, A.H. *et al.* Reduced insulin exocytosis in human pancreatic beta-cells with gene variants linked to type 2 diabetes. *Diabetes* **61**, 1726-33 (2012).
96. Altshuler, D.M. *et al.* Integrating common and rare genetic variation in diverse human populations. *Nature* **467**, 52-8 (2010).
97. Abecasis, G.R. *et al.* An integrated map of genetic variation from 1,092 human genomes. *Nature* **491**, 56-65 (2012).
98. Manolio, T.A. *et al.* Finding the missing heritability of complex diseases. *Nature* **461**, 747-53 (2009).
99. Reich, D.E. & Lander, E.S. On the allelic spectrum of human disease. *Trends Genet* **17**, 502-10 (2001).
100. Hunt, K.A. *et al.* Negligible impact of rare autoimmune-locus coding-region variants on missing heritability. *Nature* **498**, 232-5 (2013).
101. Merikangas, K.R. & Risch, N. Genomic priorities and public health. *Science* **302**, 599-601 (2003).
102. Bonnefond, A. *et al.* Rare MTNR1B variants impairing melatonin receptor 1B function contribute to type 2 diabetes. *Nat Genet* **44**, 297-301 (2012).
103. Wu, Y. *et al.* Trans-ethnic fine-mapping of lipid loci identifies population-specific signals and allelic heterogeneity that increases the trait variance explained. *PLoS Genet* **9**, e1003379 (2013).
104. Wang, Z., Gerstein, M. & Snyder, M. RNA-Seq: a revolutionary tool for transcriptomics. *Nat Rev Genet* **10**, 57-63 (2009).
105. Ozsolak, F. & Milos, P.M. RNA sequencing: advances, challenges and opportunities. *Nat Rev Genet* **12**, 87-98 (2011).
106. Djebali, S. *et al.* Landscape of transcription in human cells. *Nature* **489**, 101-8 (2012).
107. Kapranov, P. *et al.* RNA maps reveal new RNA classes and a possible function for pervasive transcription. *Science* **316**, 1484-8 (2007).
108. Nica, A.C. *et al.* Cell-type, allelic, and genetic signatures in the human pancreatic beta cell transcriptome. *Genome Res* (2013).
109. Kong, A. *et al.* Parental origin of sequence variants associated with complex diseases. *Nature* **462**, 868-74 (2009).
110. Small, K.S. *et al.* Identification of an imprinted master trans regulator at the KLF14 locus related to multiple metabolic phenotypes. *Nat Genet* **43**, 561-4 (2011).
111. Bernstein, B.E. *et al.* An integrated encyclopedia of DNA elements in the human genome. *Nature* **489**, 57-74 (2012).
112. Thurman, R.E. *et al.* The accessible chromatin landscape of the human genome. *Nature* **489**, 75-82 (2012).
113. van Arensbergen, J. *et al.* Derepression of Polycomb targets during pancreatic organogenesis allows insulin-producing beta-cells to adopt a neural gene activity program. *Genome Res* **20**, 722-32 (2010).
114. van Arensbergen, J. *et al.* Ring1b bookmarks genes in pancreatic embryonic progenitors for repression in adult beta cells. *Genes Dev* **27**, 52-63 (2013).

115. Xie, R. *et al.* Dynamic chromatin remodeling mediated by polycomb proteins orchestrates pancreatic differentiation of human embryonic stem cells. *Cell Stem Cell* **12**, 224-37 (2013).
116. Chomczynski, P. & Sacchi, N. Single-step method of RNA isolation by acid guanidinium thiocyanate-phenol-chloroform extraction. *Anal Biochem* **162**, 156-9 (1987).
117. Parnaud, G. *et al.* Proliferation of sorted human and rat beta cells. *Diabetologia* **51**, 91-100 (2008).
118. Scott, L.J. *et al.* A genome-wide association study of type 2 diabetes in Finns detects multiple susceptibility variants. *Science* **316**, 1341-5 (2007).
119. Wiltshire, S. *et al.* A genomewide scan for loci predisposing to type 2 diabetes in a U.K. population (the Diabetes UK Warren 2 Repository): analysis of 573 pedigrees provides independent replication of a susceptibility locus on chromosome 1q. *Am J Hum Genet* **69**, 553-69 (2001).
120. Spector, T.D. & MacGregor, A.J. The St. Thomas' UK Adult Twin Registry. *Twin Res* **5**, 440-3 (2002).
121. Wichmann, H.E., Gieger, C. & Illig, T. KORA-gen--resource for population genetics, controls and a broad spectrum of disease phenotypes. *Gesundheitswesen* **67 Suppl 1**, S26-30 (2005).
122. Groop, L. *et al.* Metabolic consequences of a family history of NIDDM (the Botnia study): evidence for sex-specific parental effects. *Diabetes* **45**, 1585-93 (1996).
123. Bog-Hansen, E. *et al.* Risk factor clustering in patients with hypertension and non-insulin-dependent diabetes mellitus. The Skaraborg Hypertension Project. *J Intern Med* **243**, 223-32 (1998).
124. Taylor, H.A., Jr. *et al.* Toward resolution of cardiovascular health disparities in African Americans: design and methods of the Jackson Heart Study. *Ethn Dis* **15**, S6-4-17 (2005).
125. Yu, H., Bowden, D.W., Spray, B.J., Rich, S.S. & Freedman, B.I. Linkage analysis between loci in the renin-angiotensin axis and end-stage renal disease in African Americans. *J Am Soc Nephrol* **7**, 2559-64 (1996).
126. Chahal, N.S. *et al.* Ethnicity-related differences in left ventricular function, structure and geometry: a population study of UK Indian Asian and European white subjects. *Heart* **96**, 466-71 (2010).
127. Chahal, N.S. *et al.* Does subclinical atherosclerosis burden identify the increased risk of cardiovascular disease mortality among United Kingdom Indian Asians? A population study. *Am Heart J* **162**, 460-6 (2011).
128. Lavanya, R. *et al.* Methodology of the Singapore Indian Chinese Cohort (SICC) eye study: quantifying ethnic variations in the epidemiology of eye diseases in Asians. *Ophthalmic Epidemiol* **16**, 325-36 (2009).
129. Cho, Y.S. *et al.* A large-scale genome-wide association study of Asian populations uncovers genetic factors influencing eight quantitative traits. *Nat Genet* **41**, 527-34 (2009).
130. Hughes, K., Aw, T.C., Kuperan, P. & Choo, M. Central obesity, insulin resistance, syndrome X, lipoprotein(a), and cardiovascular risk in Indians, Malays, and Chinese in Singapore. *J Epidemiol Community Health* **51**, 394-9 (1997).
131. Hughes, K. *et al.* Cardiovascular diseases in Chinese, Malays, and Indians in Singapore. II. Differences in risk factor levels. *J Epidemiol Community Health* **44**, 29-35 (1990).
132. Tan, C.E., Emmanuel, S.C., Tan, B.Y. & Jacob, E. Prevalence of diabetes and ethnic differences in cardiovascular risk factors. The 1992 Singapore National Health Survey. *Diabetes Care* **22**, 241-7 (1999).
133. Cutter, J., Tan, B.Y. & Chew, S.K. Levels of cardiovascular disease risk factors in Singapore following a national intervention programme. *Bull World Health Organ* **79**, 908-15 (2001).

134. Mitchell, B.D. *et al.* Genetic and environmental contributions to cardiovascular risk factors in Mexican Americans. The San Antonio Family Heart Study. *Circulation* **94**, 2159-70 (1996).
135. Hunt, K.J. *et al.* Genome-wide linkage analyses of type 2 diabetes in Mexican Americans: the San Antonio Family Diabetes/Gallbladder Study. *Diabetes* **54**, 2655-62 (2005).
136. Coletta, D.K. *et al.* Genome-wide linkage scan for genes influencing plasma triglyceride levels in the Veterans Administration Genetic Epidemiology Study. *Diabetes* **58**, 279-84 (2009).
137. Knowler, W.C. *et al.* The Family Investigation of Nephropathy and Diabetes (FIND): design and methods. *J Diabetes Complications* **19**, 1-9 (2005).
138. Hanis, C.L. *et al.* Diabetes among Mexican Americans in Starr County, Texas. *Am J Epidemiol* **118**, 659-72 (1983).
139. Stancakova, A. *et al.* Changes in insulin sensitivity and insulin release in relation to glycemia and glucose tolerance in 6,414 Finnish men. *Diabetes* **58**, 1212-21 (2009).
140. Atzmon, G. *et al.* Abraham's children in the genome era: major Jewish diaspora populations comprise distinct genetic clusters with shared Middle Eastern Ancestry. *Am J Hum Genet* **86**, 850-9 (2010).
141. Quinlan, A.R. & Hall, I.M. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**, 841-2 (2010).
142. Garcia, D.M. *et al.* Weak seed-pairing stability and high target-site abundance decrease the proficiency of lsy-6 and other microRNAs. *Nat Struct Mol Biol* **18**, 1139-46 (2011).
143. Pasquali, L. *et al.* Pancreatic islet enhancer clusters enriched in type 2 diabetes risk-associated variants. *Nat Genet* (2014).
144. Mikkelsen, T.S. *et al.* Comparative epigenomic analysis of murine and human adipogenesis. *Cell* **143**, 156-69 (2010).
145. McLaren, W. *et al.* Deriving the consequences of genomic variants with the Ensembl API and SNP Effect Predictor. *Bioinformatics* **26**, 2069-70 (2010).
146. Flicek, P. *et al.* Ensembl 2013. *Nucleic Acids Res* **41**, D48-55 (2013).
147. R Development Core Team. R: A Language and Environment for Statistical Computing. (R Foundation for Statistical Computing, Vienna, Austria, 2008).
148. Lee, R.C., Feinbaum, R.L. & Ambros, V. The C. elegans heterochronic gene lin-4 encodes small RNAs with antisense complementarity to lin-14. *Cell* **75**, 843-54 (1993).
149. Krol, J., Loedige, I. & Filipowicz, W. The widespread regulation of microRNA biogenesis, function and decay. *Nat Rev Genet* **11**, 597-610 (2010).
150. Kozomara, A. & Griffiths-Jones, S. miRBase: integrating microRNA annotation and deep-sequencing data. *Nucleic Acids Res* **39**, D152-7 (2011).
151. Huntzinger, E. & Izaurralde, E. Gene silencing by microRNAs: contributions of translational repression and mRNA decay. *Nat Rev Genet* **12**, 99-110 (2011).
152. Hafner, M. *et al.* PAR-CliP--a method to identify transcriptome-wide the binding sites of RNA binding proteins. *J Vis Exp* (2010).
153. Chi, S.W., Zang, J.B., Mele, A. & Darnell, R.B. Argonaute HITS-CLIP decodes microRNA-mRNA interaction maps. *Nature* **460**, 479-86 (2009).
154. Giraldez, A.J. *et al.* MicroRNAs regulate brain morphogenesis in zebrafish. *Science* **308**, 833-8 (2005).
155. Wienholds, E. *et al.* MicroRNA expression in zebrafish embryonic development. *Science* **309**, 310-1 (2005).
156. Wang, Y., Medvid, R., Melton, C., Jaenisch, R. & Blelloch, R. DGCR8 is essential for microRNA biogenesis and silencing of embryonic stem cell self-renewal. *Nat Genet* **39**, 380-5 (2007).
157. Huang, T.C. *et al.* Regulation of lipid metabolism by Dicer revealed through SILAC mice. *J Proteome Res* **11**, 2193-205 (2012).

158. Lu, J. *et al.* MicroRNA expression profiles classify human cancers. *Nature* **435**, 834-8 (2005).
159. Kumar, M.S., Lu, J., Mercer, K.L., Golub, T.R. & Jacks, T. Impaired microRNA processing enhances cellular transformation and tumorigenesis. *Nat Genet* **39**, 673-7 (2007).
160. Esquela-Kerscher, A. & Slack, F.J. Oncomirs - microRNAs with a role in cancer. *Nat Rev Cancer* **6**, 259-69 (2006).
161. O'Donnell, K.A., Wentzel, E.A., Zeller, K.I., Dang, C.V. & Mendell, J.T. c-Myc-regulated microRNAs modulate E2F1 expression. *Nature* **435**, 839-43 (2005).
162. He, L. *et al.* A microRNA polycistron as a potential human oncogene. *Nature* **435**, 828-33 (2005).
163. van Rooij, E. *et al.* A signature pattern of stress-responsive microRNAs that can evoke cardiac hypertrophy and heart failure. *Proc Natl Acad Sci U S A* **103**, 18255-60 (2006).
164. Williams, A.H. *et al.* MicroRNA-206 delays ALS progression and promotes regeneration of neuromuscular synapses in mice. *Science* **326**, 1549-54 (2009).
165. Wu, H. *et al.* Genome-wide analysis reveals methyl-CpG-binding protein 2-dependent regulation of microRNAs in a mouse model of Rett syndrome. *Proc Natl Acad Sci U S A* **107**, 18161-6 (2010).
166. Clop, A. *et al.* A mutation creating a potential illegitimate microRNA target site in the myostatin gene affects muscularity in sheep. *Nat Genet* **38**, 813-8 (2006).
167. Sethupathy, P. *et al.* Human microRNA-155 on chromosome 21 differentially interacts with its polymorphic target in the AGTR1 3' untranslated region: a mechanism for functional single-nucleotide polymorphisms related to phenotypes. *Am J Hum Genet* **81**, 405-13 (2007).
168. Brest, P. *et al.* A synonymous variant in IRGM alters a binding site for miR-196 and causes deregulation of IRGM-dependent xenophagy in Crohn's disease. *Nat Genet* **43**, 242-5 (2011).
169. Abelson, J.F. *et al.* Sequence variants in SLITRK1 are associated with Tourette's syndrome. *Science* **310**, 317-20 (2005).
170. Jostins, L. *et al.* Host-microbe interactions have shaped the genetic architecture of inflammatory bowel disease. *Nature* **491**, 119-24 (2012).
171. McCarroll, S.A. *et al.* Deletion polymorphism upstream of IRGM associated with altered IRGM expression and Crohn's disease. *Nat Genet* **40**, 1107-12 (2008).
172. Singh, S.B., Davis, A.S., Taylor, G.A. & Deretic, V. Human IRGM induces autophagy to eliminate intracellular mycobacteria. *Science* **313**, 1438-41 (2006).
173. Poy, M.N. *et al.* A pancreatic islet-specific microRNA regulates insulin secretion. *Nature* **432**, 226-30 (2004).
174. Lovis, P., Gattesco, S. & Regazzi, R. Regulation of the expression of components of the exocytotic machinery of insulin-secreting cells by microRNAs. *Biol Chem* **389**, 305-12 (2008).
175. Tang, X., Muniappan, L., Tang, G. & Ozcan, S. Identification of glucose-regulated miRNAs from pancreatic {beta} cells reveals a role for miR-30d in insulin transcription. *RNA* **15**, 287-93 (2009).
176. Wijesekara, N. *et al.* miR-33a modulates ABCA1 expression, cholesterol accumulation, and insulin secretion in pancreatic islets. *Diabetes* **61**, 653-8 (2012).
177. Jordan, S.D. *et al.* Obesity-induced overexpression of miRNA-143 inhibits insulin-stimulated AKT activation and impairs glucose metabolism. *Nat Cell Biol* **13**, 434-46 (2011).
178. Trajkovski, M. *et al.* MicroRNAs 103 and 107 regulate insulin sensitivity. *Nature* **474**, 649-53 (2011).
179. Zhou, B. *et al.* Downregulation of miR-181a upregulates sirtuin-1 (SIRT1) and improves hepatic insulin sensitivity. *Diabetologia* **55**, 2032-43 (2012).
180. Kornfeld, J.W. *et al.* Obesity-induced overexpression of miR-802 impairs glucose metabolism through silencing of Hnf1b. *Nature* **494**, 111-5 (2013).

181. van de Bunt, M. & Gloyn, A.L. A tale of two glucose transporters: how GLUT2 re-emerged as a contender for glucose transport into the human beta cell. *Diabetologia* **55**, 2312-5 (2012).
182. Cabrera, O. *et al.* The unique cytoarchitecture of human pancreatic islets has implications for islet cell function. *Proc Natl Acad Sci U S A* **103**, 2334-9 (2006).
183. Ravassard, P. *et al.* A genetically engineered human pancreatic beta cell line exhibiting glucose-inducible insulin secretion. *J Clin Invest* **121**, 3589-97 (2011).
184. Flicek, P. *et al.* Ensembl 2011. *Nucleic Acids Res* **39**, D800-6 (2011).
185. Anders, S. & Huber, W. Differential expression analysis for sequence count data. *Genome Biol* **11**, R106 (2010).
186. Kuchen, S. *et al.* Regulation of microRNA expression and abundance during lymphopoiesis. *Immunity* **32**, 828-39 (2010).
187. Hou, J. *et al.* Identification of miRNomes in human liver and hepatocellular carcinoma reveals miR-199a/b-3p as therapeutic target for hepatocellular carcinoma. *Cancer Cell* **19**, 232-43 (2011).
188. Stark, M.S. *et al.* Characterization of the Melanoma miRNAome by Deep Sequencing. *PLoS One* **5**, e9685 (2010).
189. Schotte, D. *et al.* Discovery of new microRNAs by small RNAome deep sequencing in childhood acute lymphoblastic leukemia. *Leukemia* **25**, 1389-99 (2011).
190. Joyce, C.E. *et al.* Deep sequencing of small RNAs from human skin reveals major alterations in the psoriasis miRNAome. *Hum Mol Genet* **20**, 4025-40 (2011).
191. Faghihi, M.A. *et al.* Evidence for natural antisense transcript-mediated inhibition of microRNA function. *Genome Biol* **11**, R56 (2010).
192. Parts, L. *et al.* Extent, causes, and consequences of small RNA expression variation in human adipose tissue. *PLoS Genet* **8**, e1002704 (2012).
193. Friedlander, M.R., Mackowiak, S.D., Li, N., Chen, W. & Rajewsky, N. miRDeep2 accurately identifies known and hundreds of novel microRNA genes in seven animal clades. *Nucleic Acids Res* **40**, 37-52 (2012).
194. Bolmeson, C. *et al.* Differences in islet-enriched miRNAs in healthy and glucose intolerant human subjects. *Biochem Biophys Res Commun* **404**, 16-22 (2011).
195. Melkman-Zehavi, T. *et al.* miRNAs control insulin content in pancreatic beta-cells via downregulation of transcriptional repressors. *EMBO J* **30**, 835-45 (2011).
196. Kameswaran, V. *et al.* Epigenetic Regulation of the DLK1-MEG3 MicroRNA Cluster in Human Type 2 Diabetic Islets. *Cell Metab* (2013).
197. Klein, D. *et al.* MicroRNA expression in alpha and beta cells of human pancreatic islets. *PLoS One* **8**, e55064 (2013).
198. Bravo-Egana, V. *et al.* Quantitative differential expression analysis reveals miR-7 as major islet microRNA. *Biochem Biophys Res Commun* **366**, 922-6 (2008).
199. Poy, M.N. *et al.* miR-375 maintains normal pancreatic alpha- and beta-cell mass. *Proc Natl Acad Sci U S A* **106**, 5813-8 (2009).
200. Nieto, M. *et al.* Antisense miR-7 impairs insulin expression in developing pancreas and in cultured pancreatic buds. *Cell Transplant* **21**, 1761-74 (2012).
201. Saini, H.K., Griffiths-Jones, S. & Enright, A.J. Genomic analysis of human microRNA transcripts. *Proc Natl Acad Sci U S A* **104**, 17719-24 (2007).
202. Zhu, H. *et al.* The Lin28/let-7 axis regulates glucose metabolism. *Cell* **147**, 81-94 (2011).
203. Frost, R.J. & Olson, E.N. Control of glucose homeostasis and insulin sensitivity by the Let-7 family of microRNAs. *Proc Natl Acad Sci U S A* **108**, 21075-80 (2011).
204. Tattikota, S.G. *et al.* Argonaute2 Mediates Compensatory Expansion of the Pancreatic beta Cell. *Cell Metab* (2013).
205. Fairfax, B.P. *et al.* Genetics of gene expression in primary immune cells identifies cell type-specific master regulators and roles of HLA alleles. *Nat Genet* **44**, 502-10 (2012).
206. Grundberg, E. *et al.* Mapping cis- and trans-regulatory effects across multiple tissues in twins. *Nat Genet* **44**, 1084-9 (2012).

207. Lappalainen, T. *et al.* Transcriptome and genome sequencing uncovers functional variation in humans. *Nature* **501**, 506-11 (2013).
208. Hamming, K.S. *et al.* Coexpression of the type 2 diabetes susceptibility gene variants KCNJ11 E23K and ABCC8 S1369A alter the ATP and sulfonylurea sensitivities of the ATP-sensitive K(+) channel. *Diabetes* **58**, 2419-24 (2009).
209. Moran, I. *et al.* Human beta cell transcriptome analysis uncovers lncRNAs that are tissue-specific, dynamically regulated, and abnormally expressed in type 2 diabetes. *Cell Metab* **16**, 435-48 (2012).
210. Bramswig, N.C. *et al.* Epigenomic plasticity enables human pancreatic alpha to beta cell reprogramming. *J Clin Invest* **123**, 1275-84 (2013).
211. Bugliani, M. *et al.* Microarray analysis of isolated human islet transcriptome in type 2 diabetes and the role of the ubiquitin-proteasome system in pancreatic beta cell dysfunction. *Mol Cell Endocrinol* **367**, 1-10 (2013).
212. Eizirik, D.L. *et al.* The human pancreatic islet transcriptome: expression of candidate genes for type 1 diabetes and the impact of pro-inflammatory cytokines. *PLoS Genet* **8**, e1002552 (2012).
213. Mahdi, T. *et al.* Secreted frizzled-related protein 4 reduces insulin secretion and is overexpressed in type 2 diabetes. *Cell Metab* **16**, 625-33 (2012).
214. Taneera, J. *et al.* A systems genetics approach identifies genes and pathways for type 2 diabetes in human islets. *Cell Metab* **16**, 122-34 (2012).
215. Negi, S. *et al.* Analysis of beta-cell gene expression reveals inflammatory signaling and evidence of dedifferentiation following human islet isolation and culture. *PLoS One* **7**, e30415 (2012).
216. Dorrell, C. *et al.* Transcriptomes of the major human pancreatic cell types. *Diabetologia* **54**, 2832-44 (2011).
217. Martens, G.A. *et al.* Clusters of conserved beta cell marker genes for assessment of beta cell phenotype. *PLoS One* **6**, e24134 (2011).
218. Dominguez, V. *et al.* Class II phosphoinositide 3-kinase regulates exocytosis of insulin granules in pancreatic beta cells. *J Biol Chem* **286**, 4216-25 (2011).
219. Marselli, L. *et al.* Gene expression profiles of Beta-cell enriched tissue obtained by laser capture microdissection from subjects with type 2 diabetes. *PLoS One* **5**, e11499 (2010).
220. Kutlu, B. *et al.* Detailed transcriptome atlas of the pancreatic beta cell. *BMC Med Genomics* **2**, 3 (2009).
221. Marselli, L. *et al.* Gene expression of purified beta-cell tissue obtained from human pancreas with laser capture microdissection. *J Clin Endocrinol Metab* **93**, 1046-53 (2008).
222. Orom, U.A. *et al.* Long noncoding RNAs with enhancer-like function in human cells. *Cell* **143**, 46-58 (2010).
223. Guttman, M. *et al.* lincRNAs act in the circuitry controlling pluripotency and differentiation. *Nature* **477**, 295-300 (2011).
224. Pasmant, E., Sabbagh, A., Vidaud, M. & Bieche, I. ANRIL, a long, noncoding RNA, is an unexpected major hotspot in GWAS. *FASEB J* **25**, 444-8 (2011).
225. Barrett, J.C. *et al.* Genome-wide association study and meta-analysis find that over 40 loci affect risk of type 1 diabetes. *Nat Genet* **41**, 703-7 (2009).
226. Yi, F., Brubaker, P.L. & Jin, T. TCF-4 mediates cell type-specific regulation of proglucagon gene expression by beta-catenin and glycogen synthase kinase-3beta. *J Biol Chem* **280**, 1457-64 (2005).
227. Delaneau, O., Marchini, J. & Zagury, J.F. A linear complexity phasing method for thousands of genomes. *Nat Methods* **9**, 179-81 (2012).
228. Howie, B., Fuchsberger, C., Stephens, M., Marchini, J. & Abecasis, G.R. Fast and accurate genotype imputation in genome-wide association studies through pre-phasing. *Nat Genet* **44**, 955-9 (2012).
229. Purcell, S. *et al.* PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am J Hum Genet* **81**, 559-75 (2007).

230. Price, A.L. *et al.* Principal components analysis corrects for stratification in genome-wide association studies. *Nat Genet* **38**, 904-9 (2006).
231. Lamble, S. *et al.* Improved workflows for high throughput library preparation using the transposome-based nextera system. *BMC Biotechnol* **13**, 104 (2013).
232. Kim, D. *et al.* TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. *Genome Biol* **14**, R36 (2013).
233. Harrow, J. *et al.* GENCODE: the reference human genome annotation for The ENCODE Project. *Genome Res* **22**, 1760-74 (2012).
234. Montgomery, S.B. *et al.* Transcriptome genetics using second generation sequencing in a Caucasian population. *Nature* **464**, 773-7 (2010).
235. Stegle, O., Parts, L., Piipari, M., Winn, J. & Durbin, R. Using probabilistic estimation of expression residuals (PEER) to obtain increased power and interpretability of gene expression analyses. *Nat Protoc* **7**, 500-7 (2012).
236. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature* **447**, 661-78 (2007).
237. Huang da, W., Sherman, B.T. & Lempicki, R.A. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nat Protoc* **4**, 44-57 (2009).
238. Benjamini, Y. & Hochberg, Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J R Stat Soc Series B* **57**, 289–300 (1995).
239. Stranger, B.E. *et al.* Patterns of cis regulatory variation in diverse human populations. *PLoS Genet* **8**, e1002639 (2012).
240. Dimas, A.S. *et al.* Common regulatory variation impacts gene expression in a cell type-dependent manner. *Science* **325**, 1246-50 (2009).
241. Senee, V. *et al.* Mutations in GLIS3 are responsible for a rare syndrome with neonatal diabetes mellitus and congenital hypothyroidism. *Nat Genet* **38**, 682-7 (2006).
242. Yang, Y. *et al.* The Kruppel-like zinc finger protein GLIS3 transactivates neurogenin 3 for proper fetal pancreatic islet differentiation in mice. *Diabetologia* **54**, 2595-605 (2011).
243. Westra, H.J. *et al.* Systematic identification of trans eQTLs as putative drivers of known disease associations. *Nat Genet* **45**, 1238-43 (2013).
244. Storey, J.D. & Tibshirani, R. Statistical significance for genomewide studies. *Proc Natl Acad Sci U S A* **100**, 9440-5 (2003).
245. Irmler, M. *et al.* Inhibition of death receptor signals by cellular FLIP. *Nature* **388**, 190-5 (1997).
246. Kataoka, T. *et al.* The caspase-8 inhibitor FLIP promotes activation of NF-kappaB and Erk signaling pathways. *Curr Biol* **10**, 640-8 (2000).
247. Maedler, K. *et al.* FLIP switches Fas-mediated glucose signaling in human pancreatic beta cells from apoptosis to cell replication. *Proc Natl Acad Sci U S A* **99**, 8236-41 (2002).
248. Maedler, K. *et al.* Low concentration of interleukin-1beta induces FLICE-inhibitory protein-mediated beta-cell proliferation in human pancreatic islets. *Diabetes* **55**, 2713-22 (2006).
249. Richiusa, P. *et al.* Up-regulation of c-FLIPshort and reduction of activation-induced cell death in T-cells from patients with Type 1 diabetes. *Diabetes Nutr Metab* **17**, 6-11 (2004).
250. Waters, K.M. *et al.* Consistent association of type 2 diabetes risk variants found in europeans in diverse racial and ethnic groups. *PLoS Genet* **6**(2010).
251. Dickson, S.P., Wang, K., Krantz, I., Hakonarson, H. & Goldstein, D.B. Rare variants create synthetic genome-wide associations. *PLoS Biol* **8**, e1000294 (2010).
252. Beecham, A.H. *et al.* Analysis of immune-related loci identifies 48 new susceptibility variants for multiple sclerosis. *Nat Genet* (2013).
253. Cortes, A. *et al.* Identification of multiple risk variants for ankylosing spondylitis through high-density genotyping of immune-related loci. *Nat Genet* **45**, 730-8 (2013).

254. Maller, J.B. *et al.* Bayesian refinement of association signals for 14 loci in 3 common diseases. *Nat Genet* **44**, 1294-301 (2012).
255. Hemani, G. *et al.* Inference of the genetic architecture underlying BMI and height with the use of 20,240 sibling pairs. *Am J Hum Genet* **93**, 865-75 (2013).
256. Stahl, E.A. *et al.* Bayesian inference analyses of the polygenic architecture of rheumatoid arthritis. *Nat Genet* **44**, 483-9 (2012).
257. Todd, J.A. *et al.* Identification of susceptibility loci for insulin-dependent diabetes mellitus by trans-racial gene mapping. *Nature* **338**, 587-9 (1989).
258. Gong, J. *et al.* Fine Mapping and Identification of BMI Loci in African Americans. *Am J Hum Genet* **93**, 661-671 (2013).
259. Lango Allen, H. *et al.* Hundreds of variants clustered in genomic loci and biological pathways affect human height. *Nature* **467**, 832-8 (2010).
260. Wakefield, J. A Bayesian measure of the probability of false discovery in genetic epidemiology studies. *Am J Hum Genet* **81**, 208-27 (2007).
261. Su, Z., Marchini, J. & Donnelly, P. HAPGEN2: simulation of multiple disease SNPs. *Bioinformatics* **27**, 2304-5 (2011).
262. Howie, B.N., Donnelly, P. & Marchini, J. A flexible and accurate genotype imputation method for the next generation of genome-wide association studies. *PLoS Genet* **5**, e1000529 (2009).
263. Magi, R. & Morris, A.P. GWAMA: software for genome-wide association meta-analysis. *BMC Bioinformatics* **11**, 288 (2010).
264. Stephens, M. & Balding, D.J. Bayesian statistical methods for genetic association studies. *Nat Rev Genet* **10**, 681-90 (2009).
265. Fuchsberger, C. *et al.* The value of population-specific reference panels for genotype imputation in the age of whole-genome sequencing. in *American Society for Human Genetics Annual Meeting* (San Francisco, 2012).
266. Bernstein, B.E. *et al.* The NIH Roadmap Epigenomics Mapping Consortium. *Nat Biotechnol* **28**, 1045-8 (2010).
267. The Genotype-Tissue Expression (GTEx) project. *Nat Genet* **45**, 580-5 (2013).
268. Parker, S.C. *et al.* Chromatin stretch enhancer states drive cell-specific gene regulation and harbor human disease risk variants. *Proc Natl Acad Sci U S A* **110**, 17921-6 (2013).
269. van de Bunt, M. *et al.* The miRNA profile of human pancreatic islets and beta-cells and relationship to type 2 diabetes pathogenesis. *PLoS One* **8**, e55272 (2013).
270. Segre, A.V., Groop, L., Mootha, V.K., Daly, M.J. & Altshuler, D. Common inherited variation in mitochondrial genes is not enriched for associations with type 2 diabetes or related glycemic traits. *PLoS Genet* **6**(2010).
271. Betel, D., Wilson, M., Gabow, A., Marks, D.S. & Sander, C. The microRNA.org resource: targets and expression. *Nucleic Acids Res* **36**, D149-53 (2008).
272. Wang, X. miRDB: a microRNA target prediction and functional annotation database with a wiki interface. *RNA* **14**, 1012-7 (2008).
273. Speliotes, E.K. *et al.* Association analyses of 249,796 individuals reveal 18 new loci associated with body mass index. *Nat Genet* **42**, 937-48 (2010).
274. Heid, I.M. *et al.* Meta-analysis identifies 13 new loci associated with waist-hip ratio and reveals sexual dimorphism in the genetic basis of fat distribution. *Nat Genet* **42**, 949-60 (2010).
275. Lee, S., Wu, M.C. & Lin, X. Optimal tests for rare variant effects in sequencing association studies. *Biostatistics* **13**, 762-75 (2012).
276. Lango Allen, H. *et al.* GATA6 haploinsufficiency causes pancreatic agenesis in humans. *Nat Genet* **44**, 20-2 (2012).
277. Xu, Z. *et al.* Antisense expression increases gene expression variability and locus interdependency. *Mol Syst Biol* **7**, 468 (2011).
278. Katayama, S. *et al.* Antisense transcription in the mammalian transcriptome. *Science* **309**, 1564-6 (2005).

279. Hongay, C.F., Grisafi, P.L., Galitski, T. & Fink, G.R. Antisense transcription controls cell fate in Saccharomyces cerevisiae. *Cell* **127**, 735-45 (2006).
280. Rinn, J.L. *et al.* Functional demarcation of active and silent chromatin domains in human HOX loci by noncoding RNAs. *Cell* **129**, 1311-23 (2007).
281. Clemson, C.M., McNeil, J.A., Willard, H.F. & Lawrence, J.B. XIST RNA paints the inactive X chromosome at interphase: evidence for a novel RNA involved in nuclear/chromosome structure. *J Cell Biol* **132**, 259-75 (1996).
282. Zhang, L.F., Huynh, K.D. & Lee, J.T. Perinucleolar targeting of the inactive X during S phase: evidence for a role in the maintenance of silencing. *Cell* **129**, 693-706 (2007).
283. Tanzer, A. & Stadler, P.F. Molecular evolution of a microRNA cluster. *J Mol Biol* **339**, 327-35 (2004).
284. Houbaviy, H.B., Murray, M.F. & Sharp, P.A. Embryonic stem cell-specific MicroRNAs. *Dev Cell* **5**, 351-8 (2003).
285. Ventura, A. *et al.* Targeted deletion reveals essential and overlapping functions of the miR-17 through 92 family of miRNA clusters. *Cell* **132**, 875-86 (2008).
286. Berezikov, E. *et al.* Phylogenetic shadowing and computational identification of human microRNA genes. *Cell* **120**, 21-4 (2005).
287. Altuvia, Y. *et al.* Clustering and conservation patterns of human microRNAs. *Nucleic Acids Res* **33**, 2697-706 (2005).
288. Chen, K. & Rajewsky, N. Natural selection on human microRNA binding sites inferred from SNP data. *Nat Genet* **38**, 1452-6 (2006).
289. Moraes, C.T., Diaz, F. & Barrientos, A. Defects in the biosynthesis of mitochondrial heme c and heme a in yeast and mammals. *Biochim Biophys Acta* **1659**, 153-9 (2004).
290. Valnot, I. *et al.* A mutation in the human heme A:farnesyltransferase gene (COX10) causes cytochrome c oxidase deficiency. *Hum Mol Genet* **9**, 1245-9 (2000).
291. Antonicka, H. *et al.* Mutations in COX10 result in a defect in mitochondrial heme A biosynthesis and account for multiple, early-onset clinical phenotypes associated with isolated COX deficiency. *Hum Mol Genet* **12**, 2693-702 (2003).
292. Coenen, M.J. *et al.* Cytochrome c oxidase biogenesis in a patient with a mutation in COX10 gene. *Ann Neurol* **56**, 560-4 (2004).
293. Pitceathly, R.D. *et al.* COX10 Mutations Resulting in Complex Multisystem Mitochondrial Disease That Remains Stable Into Adulthood. *JAMA Neurol* (2013).
294. Snogdal, L.S. *et al.* Common variation in oxidative phosphorylation genes is not a major cause of insulin resistance or type 2 diabetes. *Diabetologia* **55**, 340-8 (2012).
295. Tsuruta, T. *et al.* miR-152 is a tumor suppressor microRNA that is silenced by DNA hypermethylation in endometrial cancer. *Cancer Res* **71**, 6450-62 (2011).
296. Kitano, K. *et al.* CpG island methylation of microRNAs is associated with tumor size and recurrence of non-small-cell lung cancer. *Cancer Sci* **102**, 2126-31 (2011).
297. Nielsen, L.B. *et al.* Circulating levels of microRNA from children with newly diagnosed type 1 diabetes and healthy controls: evidence that miR-25 associates to residual beta-cell function and glycaemic control during disease progression. *Exp Diabetes Res* **2012**, 896362 (2012).
298. Weksberg, R., Shuman, C. & Beckwith, J.B. Beckwith-Wiedemann syndrome. *Eur J Hum Genet* **18**, 8-14 (2010).
299. Manning, A.K. *et al.* A genome-wide approach accounting for body mass index identifies genetic variants influencing fasting glycemic traits and insulin resistance. *Nat Genet* **44**, 659-69 (2012).
300. Gromada, J. *et al.* Glucagon-like peptide 1 (7-36) amide stimulates exocytosis in human pancreatic beta-cells by both proximal and distal regulatory steps in stimulus-secretion coupling. *Diabetes* **47**, 57-65 (1998).
301. Roger, B. *et al.* Adenylyl cyclase 8 is central to glucagon-like peptide 1 signalling and effects of chronically elevated glucose in rat and human pancreatic beta cells. *Diabetologia* **54**, 390-402 (2011).

302. Chung, C.D. *et al.* Specific inhibition of Stat3 signal transduction by PIAS3. *Science* **278**, 1803-5 (1997).
303. Sharma, M. *et al.* hZimp10 is an androgen receptor co-activator and forms a complex with SUMO-1 at replication foci. *EMBO J* **22**, 6101-14 (2003).
304. Dai, X.Q. *et al.* SUMOylation regulates insulin exocytosis downstream of secretory granule docking in rodents and humans. *Diabetes* **60**, 838-47 (2011).
305. McCleverty, C.J., Hornsby, M., Spraggon, G. & Kreusch, A. Crystal structure of human Pus10, a novel pseudouridine synthase. *J Mol Biol* **373**, 1243-54 (2007).
306. Aza-Blanc, P. *et al.* Identification of modulators of TRAIL-induced apoptosis via RNAi-based phenotypic screening. *Mol Cell* **12**, 627-37 (2003).
307. Nicolson, T.J. *et al.* Insulin storage and glucose homeostasis in mice null for the granule zinc transporter ZnT8 and studies of the type 2 diabetes-associated variants. *Diabetes* **58**, 2070-83 (2009).
308. Wijesekara, N. *et al.* Beta cell-specific Znt8 deletion in mice causes marked defects in insulin processing, crystallisation and secretion. *Diabetologia* **53**, 1656-68 (2010).
309. Pound, L.D. *et al.* The physiological effects of deleting the mouse SLC30A8 gene encoding zinc transporter-8 are influenced by gender and genetic background. *PLoS One* **7**, e40972 (2012).
310. Flannick, J. *et al.* Loss of function mutations in SLC30A8 protect against type 2 diabetes. in *American Society for Human Genetics Annual Meeting* (Boston, 2013).
311. Inoue, H. *et al.* A gene encoding a transmembrane protein is mutated in patients with diabetes mellitus and optic atrophy (Wolfram syndrome). *Nat Genet* **20**, 143-8 (1998).
312. Strom, T.M. *et al.* Diabetes insipidus, diabetes mellitus, optic atrophy and deafness (DIDMOAD) caused by mutations in a novel gene (wolframin) coding for a predicted transmembrane protein. *Hum Mol Genet* **7**, 2021-8 (1998).
313. Weedon, M.N. *et al.* Recessive mutations in a distal PTF1A enhancer cause isolated pancreatic agenesis. *Nat Genet* (2013).
314. Islam, S. *et al.* Characterization of the single-cell transcriptional landscape by highly multiplex RNA-seq. *Genome Res* **21**, 1160-7 (2011).
315. Sasagawa, Y. *et al.* Quartz-Seq: a highly reproducible and sensitive single-cell RNA sequencing method, reveals non-genetic gene-expression heterogeneity. *Genome Biol* **14**, R31 (2013).
316. Ramskold, D. *et al.* Full-length mRNA-Seq from single-cell levels of RNA and individual circulating tumor cells. *Nat Biotechnol* **30**, 777-82 (2012).
317. Shalek, A.K. *et al.* Single-cell transcriptomics reveals bimodality in expression and splicing in immune cells. *Nature* **498**, 236-40 (2013).

OPEN ACCESS Freely available online

PLOS ONE

The miRNA Profile of Human Pancreatic Islets and Beta-Cells and Relationship to Type 2 Diabetes Pathogenesis

Martijn van de Bunt[1,2], Kyle J. Gaulton[2], Leopold Parts[3,4], Ignasi Moran[5], Paul R. Johnson[1,6,7], Cecilia M. Lindgren[2], Jorge Ferrer[5], Anna L. Gloyn[1,7], Mark I. McCarthy[1,2,7]*

1 Oxford Centre for Diabetes, Endocrinology & Metabolism, University of Oxford, Oxford, United Kingdom, **2** Wellcome Trust Centre for Human Genetics, University of Oxford, Oxford, United Kingdom, **3** Wellcome Trust Sanger Institute, Hinxton, United Kingdom, **4** Donnelly Centre for Cellular and Biomolecular Research, Toronto, Canada, **5** Genomic Programming of Beta-cells Laboratory, Institut d'Investigacions August Pi I Sunyer (IDIBAPS), Barcelona, Spain, **6** Nuffield Department of Surgery, University of Oxford, Oxford, United Kingdom, **7** Oxford NIHR Biomedical Research Centre, Churchill Hospital, Oxford, United Kingdom

Abstract

Recent advances in the understanding of the genetics of type 2 diabetes (T2D) susceptibility have focused attention on the regulation of transcriptional activity within the pancreatic beta-cell. MicroRNAs (miRNAs) represent an important component of regulatory control, and have proven roles in the development of human disease and control of glucose homeostasis. We set out to establish the miRNA profile of human pancreatic islets and of enriched beta-cell populations, and to explore their potential involvement in T2D susceptibility. We used Illumina small RNA sequencing to profile the miRNA fraction in three preparations each of primary human islets and of enriched beta-cells generated by fluorescence-activated cell sorting. In total, 366 miRNAs were found to be expressed (i.e. >100 cumulative reads) in islets and 346 in beta-cells; of the total of 384 unique miRNAs, 328 were shared. A comparison of the islet-cell miRNA profile with those of 15 other human tissues identified 40 miRNAs predominantly expressed (i.e. >50% of all reads seen across the tissues) in islets. Several highly-expressed islet miRNAs, such as miR-375, have established roles in the regulation of islet function, but others (e.g. miR-27b-3p, miR-192-5p) have not previously been described in the context of islet biology. As a first step towards exploring the role of islet-expressed miRNAs and their predicted mRNA targets in T2D pathogenesis, we looked at published T2D association signals across these sites. We found evidence that predicted mRNA targets of islet-expressed miRNAs were globally enriched for signals of T2D association (p-values <0.01, q-values <0.1). At six loci with genome-wide evidence for T2D association (*AP3S2*, *KCNK16*, *NOTCH2*, *SCL30A8*, *VPS26A*, and *WFS1*) predicted mRNA target sites for islet-expressed miRNAs overlapped potentially causal variants. In conclusion, we have described the miRNA profile of human islets and beta-cells and provide evidence linking islet miRNAs to T2D pathogenesis.

Citation: van de Bunt M, Gaulton KJ, Parts L, Moran I, Johnson PR, et al. (2013) The miRNA Profile of Human Pancreatic Islets and Beta-Cells and Relationship to Type 2 Diabetes Pathogenesis. PLoS ONE 8(1): e55272. doi:10.1371/journal.pone.0055272

Editor: Carol J. Wilusz, Colorado State University, United States of America

Received August 25, 2012; **Accepted** December 22, 2012; **Published** January 25, 2013

Copyright: © 2013 van de Bunt et al. This is an open-access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

Funding: MvdB is a Nuffield Department of Medicine Prize Student. LP is a Canadian Institute for Advanced Research Junior Fellow. CML is a Wellcome Trust Research Career Development Fellow (086596/Z/08/Z). ALG is a Wellcome Trust Senior Research Fellow in Basic & Biomedical Science (095101/2/10/Z). The work was funded by the Oxford NIHR Biomedical Research Centre, NIDDK (R01 DK073490), and the Wellcome Trust (090532/Z/09/Z). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The authors have declared that no competing interests exist.

* E-mail: Mark.McCarthy@drl.ox.ac.uk

Introduction

The overwhelming majority of the disease-associated variation identified by genome-wide association studies (GWAS) maps to the non-protein-coding genome. Efforts to unlock the functional impact of these variants therefore rely on an understanding of the processes involved in the regulation of transcription, most particularly those which are active in the cells and tissues implicated in disease pathogenesis [1,2].

miRNAs, short (~22 nucleotides) non-coding RNAs, are thought to play a key role in the regulation of cellular function through effects on mRNA destabilisation and/or translational repression [3,4]. Altered miRNA function has been implicated in the pathogenesis of a growing number of diseases, including Tourette's syndrome and a variety of cancers [5,6]. There is also substantial evidence linking miRNAs to the regulation of glucose homeostasis. For example, miR-375 has been reproducibly shown to be involved in the regulation of glucose-stimulated insulin secretion in the murine insulin-secreting cell-line MIN6 [7,8] and other miRNAs (let-7, miR-103 and -107) influence insulin sensitivity in rodents [9,10].

Many of the variants robustly associated with type 2 diabetes (T2D) in GWAS exert their diabetogenic effect via a primary reduction in insulin secretion, placing the pancreatic islet, and the insulin-secreting beta-cell in particular, center-stage in terms of T2D pathogenesis [11]. Given significant differences in islet physiology between rodents and humans [12], and with few suitable human beta-cell lines available [13], the genomic characterisation of primary human islet preparations provides an important opportunity to develop the functional annotations that can support biological inference at T2D association signals.

In this study, we set out to define the miRNA profile of primary human islets and enriched human beta-cell preparations, and to relate these findings to patterns of T2D predisposition.

Results

The miRNA profile of human pancreatic islets and beta-cells

Human islets were acquired from six donors with appropriate ethical consent. For three of these samples the islets were subjected to fluorescence-activated cell sorting (FACS) to select for beta-cells. The purity of the beta-cells following FACS ranged between 90–95%, and we refer to these as "enriched beta-cell" preparations. All six samples (3 islet and 3 enriched beta-cells; see Table S1 for clinical characteristics) were sequenced using 50 base pair reads on either Illumina GAIIx or Hiseq2000 platforms. After quality control, a total of 83 million reads mapping to small RNA species were obtained across the three islet samples (648,000, 14 million, 68 million per sample), and 68 million across the enriched beta-cell samples (3 million, 38 million, 26 million per sample). The differences in read-depths reflect improvements in sequencing technology over the period of data collection.

Almost all reads (92%) mapped to mature miRNA sequences annotated in miRBase v18 [14], the remainder being snoRNAs (2.4%) and other non-coding species in Ensembl v63 [15]. As the number of annotated human miRNAs is still expanding, we used the miRDeep2 package – a probabilistic method for discovering miRNAs from small RNA sequencing data using the predicted secondary structure of potential miRNA precursors [16] – to uncover novel, unannotated miRNAs in our sequencing data, but there were no instances of high confidence predictions of novel miRNAs.

In islets, 366 miRNAs were expressed above background levels (i.e. more than 100 combined reads across the three samples), and 346 in enriched beta-cells (Figure 1; Table S2). In total, 384 unique miRNAs were identified, of which 85% (n = 328) were shared between islets and beta-cells (Table S2). The expression profile was of medium complexity with on average 38 miRNAs responsible for 90% of the total miRNA aligned reads in islets, and 21 in enriched beta-cells. There was good reproducibility across the three samples of each type (Spearman rho range 0.61–0.96), with all but one expressed miRNA present in at least two samples of each type.

Several of the most highly expressed miRNAs in human islets (including miR-375, miR-7-5p, miR-148a-3p, miR-26a-5p and miR-127-3p) have been previously described in rodent or human islets [7,17–20]. The expression in islets of others (e.g. miR-27b-3p, miR-192-5p) is novel. Other islet-expressed miRNAs such as miR-143-3p and let-7 family members have been implicated in glucose homeostasis, but primarily in the context of insulin action rather than insulin secretion [10,21,22].

Comparison of the miRNA expression profiles of islets and enriched beta-cells revealed a strong correlation between the two (Figure 1), but highlighted some interesting differences. For example, amongst the most highly expressed miRNAs, miR-375 shows substantially higher abundance in enriched beta-cells as compared to islets (42% vs 27%) whilst the opposite is true for miR-143-3p (2% vs 16%). On the basis that an average islet contains ~50% beta-cells [23], and, given an estimated 90% purity of the FACS-enriched beta-cell preparations, the ~1.5 fold enrichment of miR-375 in beta-cells compared to islets indicates that, in human as in rodents [7], this miRNA is predominantly expressed in beta-cells. On the same grounds, these data suggest that miR-143-3p is mostly expressed outside pancreatic beta-cells.

Tissue specificity of miRNA expression

We compared the miRNA expression profiles in islets against equivalent, publicly available, next-generation sequencing-derived miRNA expression data from other human tissues, including B-cells [24], liver [25], pigment cells [26], pooled thymocytes, bone marrow, CD34+ progenitor cells [27], skin [28], lung, kidney, skeletal muscle, heart, whole pancreas, frontal orbital gyrus, spleen, liver tissue [29], and adipose tissue [30]. Raw reads were downloaded, realigned and normalized in parallel with our islet and beta-cell samples (see Methods). From these data, we generated a tissue specificity score (see Methods), defined as the expression of a miRNA in the reference tissue divided by the sum of its expression in all tissues, and considered a value >0.5 to indicate tissue-specificity.

By this definition (which indicated that half or more of all observed reads for that miRNA were derived from islets) 40 of the 366 islet miRNAs were islet-specific (Table S3). Equivalent specificity analyses using other tissues as reference found a mean of only 10 enriched miRNAs per tissue (ranging from 1 [in spleen, kidney and lung] to 31 [orbital gyrus]; Table S3). Inevitably, these comparisons are influenced by the panel of tissues for which data were available, but they suggest that the miRNA profile of human islets is relatively distinct (Figure 2A). By permuting the data, we estimated that the false-discovery rate associated with an assertion of tissue specificity for a given miRNA was approximately 8%.

Few of the 10 most islet-specific miRNAs (Figure 2B; all with specificity scores >0.8) have previously been implicated in islet function. For miR-184, miR-182-5p and miR-127-3p, there is published evidence for a role in insulin biosynthesis and secretion, though for miR-184 and miR-127-3p this is restricted to a correlation between islet expression levels and glucose-stimulated insulin secretion [17,18]. For other miRNA transcripts, such as miR-409-5p and miR-183-5p, the high degree of islet-specificity may point to novel roles in the development and maintenance of islet cellular phenotype.

Primary miRNA sequences tend to cluster together in genomic intervals spanning <10 kb, and it is thought that members of these clusters are transcribed together [30,31]. We found examples of such clustering amongst the miRNAs showing the greatest absolute abundance in islets, as well as those that were most islet-specific. For example, mir-182 and mir-183 (both amongst the most islet-specific transcripts) originate from the same cluster on chromosome 7q32.2, whilst mir-136 and mir-127 map together on chromosome 14q32.2. Across samples, expression levels of miRNAs within these clusters are highly correlated both in islets and enriched beta-cells ($r^2>0.6$).

The role of islet-expressed miRNAs in T2D predisposition

To explore the role in T2D predisposition played by genetic variants that influence miRNA function and/or expression, we analysed genome-wide association data from 8,130 T2D cases and 38,987 controls available through the Diabetes Genetics Replication and Meta-analysis (DIAGRAM) consortium [11].

We first looked for evidence that T2D-associated variants mapped to sequences encoding miRNAs themselves. Of the ~2.5 million variants for which directly-typed or HapMap-imputed T2D-association p-values were available, seven overlapped precursor miRNA transcripts for islet-expressed mature miRNAs (there were a total of 364 of these in miRBase v18 corresponding to the 366 mature miRNAs we document in islets).

Next, we asked whether the predicted targets of islet-expressed miRNAs were enriched for evidence of association with T2D. To determine whether results were robust to different algorithms for predicting these target sites, we used three different prediction algorithms (TargetScan, miRanda and miRDB) [32–34]. As two of these algorithms (TargetScan and miRanda) provide genomic coordinates for the target sites they predict, we were able, for

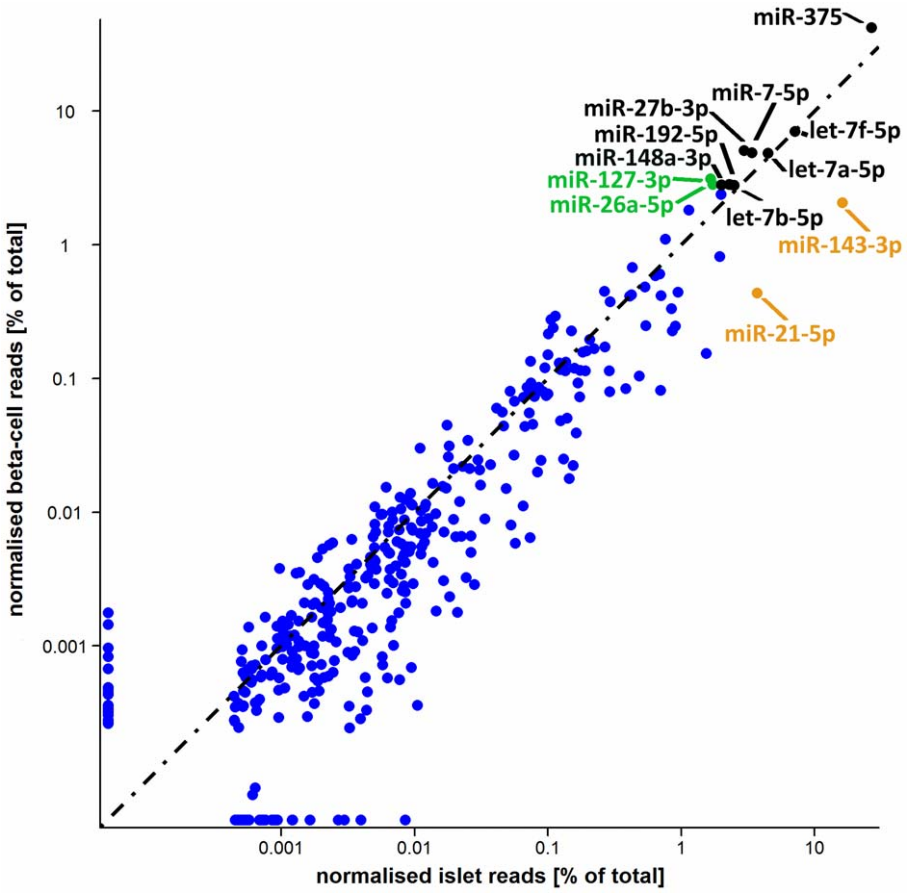


Figure 1. Islet and beta-cell miRNA profiles. The top 10 most abundant miRNAs in islets and beta-cells are annotated, with colors denoting those shared between the two top 10s (black), in islet top 10 only (orange), and in beta-cell top 10 only (green). There was good correlation between the two profiles ($r^2 = 0.78$). A black dashed line represents equality.
doi:10.1371/journal.pone.0055272.g001

these, to consider both the target genes and these specific target sites. Looking at the predicted target sites of those 366 islet miRNAs (a total of 39,103 and 749,873 such target sites were predicted by TargetScan and miRanda respectively), we identified 6,496 variants overlapping the association data. The relatively low density of common variants observed in both islet miRNA and miRNA target sequence, is consistent with the constraints imposed by negative selection [19,35,36]. Of the 6,503 variants overlapping miRNA transcripts and predicted miRNA target sites, one variant, rs3802177 at the *SLC30A8* locus (p-value = 1.45×10^{-8}) showed genome-wide significant association for T2D in the DIAGRAM meta-analysis data.

We also sought overlap with miRNA sequence and/or predicted targets within the set of 58 loci for which there was consistent evidence for genome-wide significant associations with T2D (as of July 2012 [11,37,38]). As the causal variant at most T2D-association loci is not known, we defined a broad set of 1,403 variants comprising the lead SNPs at the 58 loci plus all good proxies (LD; $r^2 > 0.8$ in CEU 1000 Genomes pilot data) which was likely to include the alleles causal for these associations. We found overlap of target sites (all predicted by miRanda) at ten of these variants, mapping to six independent loci (*AP3S2*, *KCNK16*, *NOTCH2*, *SLC30A8*, *VPS26A*, and *WFS1*).

Finally, we implemented an approach where we sought evidence for enrichment of T2D association signals across sets of islet-expressed genes predicted to be targeted by an individual islet-expressed miRNA. As we used three different target prediction algorithms, with often differing predicted target genes for individual miRNAs, we looked at the degree of enrichment of signal in the target genes predicted by each method alone, or by the overlap (Figure 3). To test for enrichment within the predicted transcriptional targets of each miRNA, we employed MAGENTA [39], a tool that uses meta-analysis summary statistics to determine enrichment for trait association across a given gene set while correcting for differences in transcript size, number of variants, and patterns of LD. We found significant enrichment (p-values <0.01, q-value <0.1) across islet miRNA targeted gene sets with nearly all of the (combinations of) prediction algorithms used, with significant enrichment (miR-17-5p, p-value = 0.004, q-value = 0.09; miR-93-5p, p-value = 0.005, q-value = 0.09; miR-20a-5p, p-value = 0.009, q-value = 0.09) even in the set of targets predicted by all three different algorithms (Figure 3, Table S4).

Discussion

Using next-generation sequencing, we have established the first catalog of miRNAs in human pancreatic islets and beta-cells, and explored the overlap between these miRNAs and T2D genetic susceptibility. Our catalog not only serves as a valuable resource for those interested in the roles of specific miRNAs in normal islet physiology and beta-cell function, it also provides a reference for the study of miRNA mediated abnormalities in islets from type 2 diabetic donors.

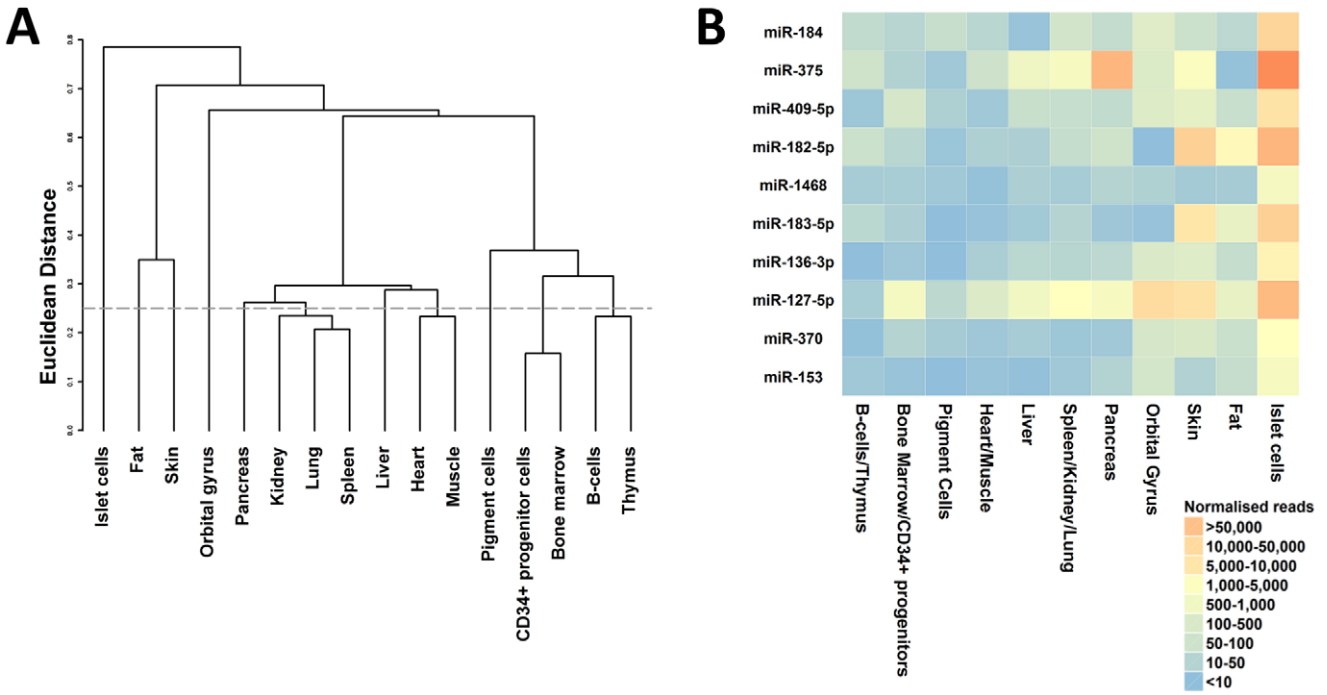


Figure 2. Comparison of miRNA profiles across tissues. The left panel (A) shows the single-linkage hierarchical clustering of inter-tissue profile correlations. In the right panel (B) the top 10 most tissue specific islet miRNAs are displayed in descending order. The colors indicate the normalized expression levels of these miRNAs across the different profiles used in the analysis.
doi:10.1371/journal.pone.0055272.g002

The abundance of miR-375 in the miRNA profile provides valuable support for a critical role in human pancreatic beta-cells, mirroring the well-established role in rodent islet biology. miR-375 null mice are hyperglycaemic and exhibit reduced beta-cell mass [40]. In a clonal rodent beta-cell line (MIN6), knockdown or over-expression of this miRNA influences glucose-stimulated insulin secretion [7]. Furthermore, knockdown of miR-375 in obese ob/ob mice results in a more profound effect on glycaemia leading to a severe diabetic phenotype in these mice [40]. Our study establishes that miR-375 is also abundantly expressed in human islets and warrants further studies to define the contribution of miR-375 to the pathogenesis of T2D.

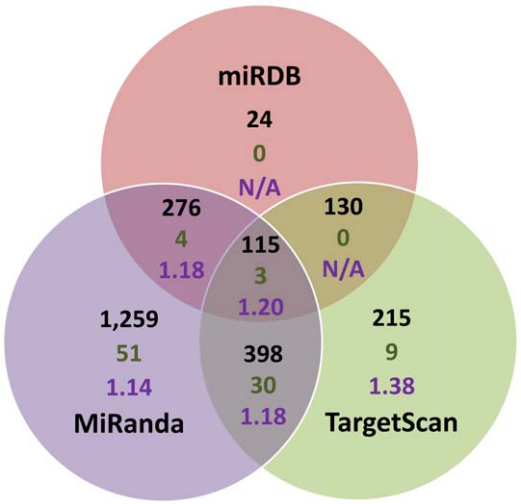


Figure 3. Results for the T2D association signal enrichment analysis in miRNA target gene sets. The Venn diagram represents the sets of miRNA target genes for each miRNA as predicted by miRDB, miRanda and TargetScan alone, or overlap of these methods. Annotated are the median number of genes in each set (black text), the number of significantly (p-value <0.001, q-value <0.1) enriched gene sets (green text), and the median enrichment of the significantly enriched gene sets (purple text).
doi:10.1371/journal.pone.0055272.g003

Most of the other abundant miRNAs show comparable levels of expression between islets and beta-cells, and the islet-cell profiles are relatively distinct from most of the other tissues studied. Our data adds support for an islet-enriched expression pattern of six miRNAs previously highlighted by microarray studies (miR-184, miR-183-5p, miR-7-5p, miR-127-3p, miR-375, and miR-493-5p) [17,20]. However, for many of the miRNAs that are most abundant and/or enriched in human islets, there is little, if any, existing data concerning their effects on islet development or function. For example, these data point to the need to further functional investigation of the let-7 family of miRNAs. Two recent studies have demonstrated that manipulation of let-7 levels in mice, by overexpression or knockdown, alters insulin sensitivity [10,22]. One of these studies however also showed reduced insulin secretion during an intraperitoneal glucose tolerance test in the let-7 overexpressing mice, hinting at a role for let-7 in glucose-stimulated insulin secretion, but did not characterize the effects of altered let-7 expression on pancreatic islet physiology [22].

When we investigated the overlap between the islet miRNAs and published T2D genome-wide association data, we found evidence suggestive of a role for miRNAs in diabetes pathogenesis. We identified both global enrichment of T2D association signals within islet-expressed miRNA target genes, and overlap of potentially causal T2D-associated variants with predicted target sites for islet-expressed miRNAs.

The strongest signal of overlap we identified was for a variant (rs3802177) within the 3′ UTR of the *SLC30A8* gene, which maps to miRanda predicted target sites for six islet-expressed miRNAs (miR-363-3p, miR-25-3p, miR-32-5p, miR-92a-3p, miR-33a-5p, and miR-33b-5p) and reaches genome wide significance in T2D-association studies [11]. Whilst, in the DIAGRAM meta-analysis, rs3802177 is the variant with the strongest association with T2D [11], it is also in complete linkage disequilibrium with a non-synonymous coding variant in *SLC30A8* (rs13266634), which is generally considered to be the causal allele [41,42]. Further functional studies will be required to establish whether rs3802177 could be contributing to perturbation of SLC30A8 expression and/or function.

These data are based on the common variant associations hitherto accessible to interrogation by GWAS. Extension of risk variant discovery efforts through next-generation sequencing to include low frequency and rare alleles will allow a more complete evaluation of the role of miRNAs and their targets with respect to pancreatic islet development and/or function and the pathogenesis of T2D.

Materials and Methods

Samples

Human islets were obtained (with research consent) from the Oxford DRWF Human Islet Isolation Facility (n = 4) and through existing collaborations with Barcelona (n = 2) from deceased donors of European descent (clinical characteristics of the donors can be found in Table S1). High islet purity was established both by dithizone labeling, as well as through comparative quantification by qRT-PCR of endocrine (insulin [*INS*], glucagon [*GCG*], somatostatin [*SST*]) and exocrine (pancreatic lipase [*PNLIP*], pancreatic amylase [*AMY2A*] and chymotrypsin C [*CTRC*]) markers.

The beta-cell enriched preparation was collected through fluorescence-activated cell sorting of human islet preparations using a previously described method [43,44]. RNA was extracted from all samples using TRI reagent (Applied Biosystems, Warrington, UK).

Library preparation and sequencing

Only samples with a total RNA yield in excess of 1ug and a quality score (RIN) >7 were selected for library preparation. Libraries were prepared at the High-Throughput Genomics group (Wellcome Trust Centre for Human Genetics, University of Oxford, Oxford, UK) using Illumina v1.5 (2 islets and 1 beta-cell sample) or TruSeq v1 (2 beta-cell and 1 islet sample) small library preparation protocols, and sequenced using 50 base reads on Illumina GAIIx and HiSeq2000 platforms respectively.

Raw sequence processing

Raw data was obtained in Fastq format, and pre-processed to remove residual 3′-adaptor sequences using the fastx_clipper function from the FASTX-toolkit (http://hannonlab.cshl.edu/fastx_toolkit/index.html). Sequences less than 16 bases after adaptor stripping, and reads containing primarily N bases, were removed. The remaining read sequences should correspond to short RNAs. Length histograms showed enrichment of a peak around 22 bases, which corresponds to the expected size of miRNAs.

Sequence mapping and quantification

As ambiguities in read mapping, and frequent contamination by large numbers of adaptor dimers pose significant challenges to the accurate quantification of small RNAs, we used a three step alignment approach. Reads were aligned using novoalign v. 2.07.11 (http://www.novocraft.com, parameters -h 60 60 -t30 -s -m -l 16 -R 0 -r A 30), which is aware of sequence ambiguities, to three consecutive references:

1. Contaminating sequences of adaptors, linkers, adaptor-linker, and adaptor-tag combinations.
2. Full-length miRBase v17 hairpin sequences, combined with other known and predicted human small RNA sequences present in Ensembl v63 [14,15].
3. Human reference genome (NCBI build 37).

Reads mapping to contaminating sequences were excluded. The remaining reads were prioritised in order miRNA > ncRNA > genome, and alignments with the smallest edit distance were selected. miRNA hairpin aligning reads were further split out into mature miRNAs, star miRNAs and those reads mapping only to the hairpin by using the respective coordinates on the hairpin from miRBase v17 extended by 3 bases either side of the mature and star sequences.

Finally, expression levels of each ncRNA were quantified by counting the number of reads aligning to it. In case of multiple mapping, where a read mapped between k alternative sequences in one reference, $1/k$ was added to the count of each. Those RNA sequences observed less than 100 times across all samples were excluded. This cut-off was chosen as the reproducibility across samples dropped when going lower, and means that per sample >95% of all mapped sequences are included in the analysis.

Normalization

Data was normalized to be able to compare read counts between samples, which all differ in read depth. An inflation factor i was calculated for each library l using a method proposed by Anders and Huber [45]. This was determined as the median inflation factor across all genes g using the following formula: $il = \text{median}_g(n_{gl}/GM_l(n_{gl}))$, where n_{gl} is the read count each gene g in that library l, and GM is the geometric mean. Next, the data was log2-transformed ($\log_2(n_{gl}/il)$) to to account for the heteroskedastic distribution of the data. These values were used in all subsequent analyses.

Tissue specific analysis

Publicly available human small RNA sequencing data using Illumina technology from B-cells [24], liver [25], pigment cells [26], pooled thymocytes, bone marrow, CD34+ progenitor cells [27], skin [28], lung, kidney, skeletal muscle, heart, pancreas, frontal orbital gyrus, spleen, and liver tissue [29], as well as to data from adipose tissue available through collaboration with the MuTHER consortium [30], were downloaded, realigned and normalized together with the islet and beta-cell samples. To minimalize background, only those miRNAs observed at least 1000 times across the normalized values (n = 367) were taken forward for tissue specific analysis.

Subsequently we calculated a tissue specificity score, defined as the expression of a miRNA in a tissue divided by the sum of its expression in all tissues, for each miRNA in each tissue. The resulting fractional expression is an indicator of what proportion of the expression of that miRNA is attributable to a given tissue. As this approach would be limited in a set of tissues containing very similar profiles, we combined the results from the islets and the beta-cells and calculated a correlation matrix between the different profiles and subjected this to single-linkage hierarchical clustering in R version 2.14.0 using *hclust* [46]. To create profiles as unrelated

as possible, all tissues with a Euclidian distance <0.25 were averaged along their branch. For each of the remaining 11 profiles, tissue specific scores were calculated by dividing the normalized reads for a miRNA in a tissue by all reads in all tissue for that miRNA. The resulting score thus represented the fraction of miRNA reads explained by that tissue. Tissue-specificity was defined by a score >0.5, which, using a permutation-based approach comprising 1 million permutations of the observed data, amounts to a false-discovery rate of 8%.

Involvement of miRNAs in T2D association

For all reported T2D loci, variants in strong LD ($R^2>0.8$) in 1000 Genomes pilot CEU data were combined with the reported T2D most associated variants using custom scripts. Genomic locations of miRNAs were downloaded from miRBase v18 (http://www.mirbase.org) [14], Targetscan 6.2 [32] conserved predictions from http://www.targetscan.org, miRanda predicted targets with good mirSVR scores (release August 2010) [33] from http://www.microrna.org/microrna/getDownloads.do and miRDB (version 4) [34] predicted target genes from http://mirdb.org/miRDB/download.html.

For the overlap with T2D associated variants, genomic coordinates were, where appropriate, converted to hg18 using the UCSC liftOver tool with *-minMatch = 1*. For the enrichment analysis we determined the intersection between predicted islet-expressed target genes of all three algorithms for each miRNA using custom Perl scripts. MAGENTA [39] was run using pre-defined settings using 100000 permutations for the p-value calculation, and the 75th percentile cut-off p-values for establishing enrichment. To correct for multiple testing q-values, which represent the expected proportion of false positives incurred when calling that gene set significant, were calculated using the *qvalue* command from the package "qvalue" in R version 2.14.0 [46] with the setting "robust = TRUE". Those gene sets with p-values <0.01 and q-values <0.1 were deemed significant.

Accession numbers

Datasets have been deposited in Gene Expression Omnibus GEO (In process).

Supporting Information

Table S1 Clinical characteristics of human islet donors. N/A denotes information not available.
(DOCX)

Table S2 Normalized average read counts for all 385 miRNAs present in human islets and beta-cells. Each row denotes an individual miRNA present above background levels (>100 observed reads). The first column has the miRNA identifiers according to miRBase 18. The entries in the two columns are the normalised average read counts for, respectively, human islets and beta-cells. 0 denotes the expression of the miRNA was below background levels (<100 total reads observed).
(DOCX)

Table S3 miRNA level comparison between tissues. Each row represents a different miRNA observed for at least 1000 reads across all tissues. The first column has the miRNA identifiers according to miRBase 18. Each other column represents a distinct (set of) tissues used for the comparison. The entries in the columns denote the tissue specificity score as described in the Methods.
(DOCX)

Table S4 Results for the significantly enriched predicted target gene sets for islet-expressed miRNAs from the MAGENTA analysis with the 75 percentile cut-off used to determine significance.
(DOCX)

Acknowledgments

We like to thank Drew Worth for assistance on the fluorescence-assisted cell sorting of human islets, the Oxford DRWF human islet isolation facility for providing human islets, the Multiple Tissue Human Expression Resource (MuTHER) and Diabetes Genetics Replication And Meta-analysis (DIAGRAM) consortia for access to their data, and the High-Throughput Genomics Group at the Wellcome Trust Centre for Human Genetics for the generation of the Sequencing data.

Author Contributions

Critically reviewed and edited the paper: MvdB KJG LP IM CML JF ALG MIM. Conceived and designed the experiments: MvdB ALG MIM. Performed the experiments: MvdB. Analyzed the data: MvdB LP IM. Contributed reagents/materials/analysis tools: KJG PRJ CML JF ALG MIM. Wrote the paper: MvdB ALG MIM.

References

1. Hindorff LA, Sethupathy P, Junkins HA, Ramos EM, Mehta JP, et al. (2009) Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. Proc Natl Acad Sci U S A 106: 9362–9367.
2. Dunham I, Kundaje A, Aldred SF, Collins PJ, Davis CA, et al. (2012) An integrated encyclopedia of DNA elements in the human genome. Nature 489: 57–74.
3. Krol J, Loedige I, Filipowicz W (2010) The widespread regulation of microRNA biogenesis, function and decay. Nat Rev Genet 11: 597–610.
4. Huntzinger E, Izaurralde E (2011) Gene silencing by microRNAs: contributions of translational repression and mRNA decay. Nat Rev Genet 12: 99–110.
5. Abelson JF, Kwan KY, O'Roak BJ, Baek DY, Stillman AA, et al. (2005) Sequence variants in SLITRK1 are associated with Tourette's syndrome. Science 310: 317–320.
6. Esteller M (2011) Non-coding RNAs in human disease. Nat Rev Genet 12: 861–874.
7. Poy MN, Eliasson L, Krutzfeldt J, Kuwajima S, Ma X, et al. (2004) A pancreatic islet-specific microRNA regulates insulin secretion. Nature 432: 226–230.
8. El Ouaamari A, Baroukh N, Martens GA, Lebrun P, Pipeleers D, et al. (2008) miR-375 targets 3′-phosphoinositide-dependent protein kinase-1 and regulates glucose-induced biological responses in pancreatic beta-cells. Diabetes 57: 2708–2717.
9. Trajkovski M, Hausser J, Soutschek J, Bhat B, Akin A, et al. (2011) MicroRNAs 103 and 107 regulate insulin sensitivity. Nature 474: 649–653.
10. Zhu H, Shyh-Chang N, Segre AV, Shinoda G, Shah SP, et al. (2011) The Lin28/let-7 axis regulates glucose metabolism. Cell 147: 81–94.
11. Voight BF, Scott LJ, Steinthorsdottir V, Morris AP, Dina C, et al. (2010) Twelve type 2 diabetes susceptibility loci identified through large-scale association analysis. Nat Genet 42: 579–589.
12. van de Bunt M, Gloyn AL (2012) A tale of two glucose transporters: how GLUT2 re-emerged as a contender for glucose transport into the human beta cell. Diabetologia.
13. Ravassard P, Hazhouz Y, Pechberty S, Bricout-Neveu E, Armanet M, et al. (2011) A genetically engineered human pancreatic beta cell line exhibiting glucose-inducible insulin secretion. J Clin Invest 121: 3589–3597.
14. Kozomara A, Griffiths-Jones S (2011) miRBase: integrating microRNA annotation and deep-sequencing data. Nucleic Acids Res 39: D152–157.
15. Flicek P, Amode MR, Barrell D, Beal K, Brent S, et al. (2011) Ensembl 2011. Nucleic Acids Res 39: D800–806.
16. Friedlander MR, Mackowiak SD, Li N, Chen W, Rajewsky N (2012) miRDeep2 accurately identifies known and hundreds of novel microRNA genes in seven animal clades. Nucleic Acids Res 40: 37–52.
17. Bolmeson C, Esguerra JL, Salehi A, Speidel D, Eliasson L, et al. (2011) Differences in islet-enriched miRNAs in healthy and glucose intolerant human subjects. Biochem Biophys Res Commun 404: 16–22.
18. Melkman-Zehavi T, Oren R, Kredo-Russo S, Shapira T, Mandelbaum AD, et al. (2011) miRNAs control insulin content in pancreatic beta-cells via downregulation of transcriptional repressors. EMBO J 30: 835–845.

19. Nieto M, Hevia P, Garcia E, Klein D, Alvarez-Cubela S, et al. (2011) Anti sense miR-7 impairs insulin expression in developing pancreas and in cultured pancreatic buds. Cell Transplant.
20. Bravo-Egana V, Rosero S, Molano RD, Pileggi A, Ricordi C, et al. (2008) Quantitative differential expression analysis reveals miR-7 as major islet microRNA. Biochem Biophys Res Commun 366: 922–926.
21. Jordan SD, Kruger M, Willmes DM, Redemann N, Wunderlich FT, et al. (2011) Obesity-induced overexpression of miRNA-143 inhibits insulin-stimulated AKT activation and impairs glucose metabolism. Nat Cell Biol 13: 434–446.
22. Frost RJ, Olson EN (2011) Control of glucose homeostasis and insulin sensitivity by the Let-7 family of microRNAs. Proc Natl Acad Sci U S A 108: 21075–21080.
23. Cabrera O, Berman DM, Kenyon NS, Ricordi C, Berggren PO, et al. (2006) The unique cytoarchitecture of human pancreatic islets has implications for islet cell function. Proc Natl Acad Sci U S A 103: 2334–2339.
24. Kuchen S, Resch W, Yamane A, Kuo N, Li Z, et al. (2010) Regulation of microRNA expression and abundance during lymphopoiesis. Immunity 32: 828–839.
25. Hou J, Lin L, Zhou W, Wang Z, Ding G, et al. (2011) Identification of miRNomes in human liver and hepatocellular carcinoma reveals miR-199a/b-3p as therapeutic target for hepatocellular carcinoma. Cancer Cell 19: 232–243.
26. Stark MS, Tyagi S, Nancarrow DJ, Boyle GM, Cook AL, et al. (2010) Characterization of the Melanoma miRNAome by Deep Sequencing. PLoS One 5: e9685.
27. Schotte D, Moqadam FA, Lange-Turenhout EA, Chen C, van Ijcken WF, et al. (2011) Discovery of new microRNAs by small RNAome deep sequencing in childhood acute lymphoblastic leukemia. Leukemia 25: 1389–1399.
28. Joyce CE, Zhou X, Xia J, Ryan C, Thrash B, et al. (2011) Deep sequencing of small RNAs from human skin reveals major alterations in the psoriasis miRNAome. Hum Mol Genet 20: 4025–4040.
29. Faghihi MA, Zhang M, Huang J, Modarresi F, Van der Brug MP, et al. (2010) Evidence for natural antisense transcript-mediated inhibition of microRNA function. Genome Biol 11: R56.
30. Parts L, Hedman AK, Keildson S, Knights AJ, Abreu-Goodger C, et al. (2012) Extent, causes, and consequences of small RNA expression variation in human adipose tissue. PLoS Genet 8: e1002704.
31. Saini HK, Griffiths-Jones S, Enright AJ (2007) Genomic analysis of human microRNA transcripts. Proc Natl Acad Sci U S A 104: 17719–17724.
32. Garcia DM, Baek D, Shin C, Bell GW, Grimson A, et al. (2011) Weak seed-pairing stability and high target-site abundance decrease the proficiency of lsy-6 and other microRNAs. Nat Struct Mol Biol 18: 1139–1146.
33. Betel D, Wilson M, Gabow A, Marks DS, Sander C (2008) The microRNA.org resource: targets and expression. Nucleic Acids Res 36: D149–153.
34. Wang X (2008) miRDB: a microRNA target prediction and functional annotation database with a wiki interface. RNA 14: 1012–1017.
35. Chen K, Rajewsky N (2006) Natural selection on human microRNA binding sites inferred from SNP data. Nat Genet 38: 1452–1456.
36. Quach H, Barreiro LB, Laval G, Zidane N, Patin E, et al. (2009) Signatures of purifying and local positive selection in human miRNAs. Am J Hum Genet 84: 316–327.
37. Kooner JS, Saleheen D, Sim X, Sehmi J, Zhang W, et al. (2011) Genome-wide association study in individuals of South Asian ancestry identifies six new type 2 diabetes susceptibility loci. Nat Genet 43: 984–989.
38. Cho YS, Chen CH, Hu C, Long J, Ong RT, et al. (2012) Meta-analysis of genome-wide association studies identifies eight new loci for type 2 diabetes in east Asians. Nat Genet 44: 67–72.
39. Segre AV, Groop L, Mootha VK, Daly MJ, Altshuler D (2010) Common inherited variation in mitochondrial genes is not enriched for associations with type 2 diabetes or related glycemic traits. PLoS Genet 6.
40. Poy MN, Hausser J, Trajkovski M, Braun M, Collins S, et al. (2009) miR-375 maintains normal pancreatic alpha- and beta-cell mass. Proc Natl Acad Sci U S A 106: 5813–5818.
41. Stancakova A, Kuulasmaa T, Paananen J, Jackson AU, Bonnycastle LL, et al. (2009) Association of 18 confirmed susceptibility loci for type 2 diabetes with indices of insulin release, proinsulin conversion, and insulin sensitivity in 5,327 nondiabetic Finnish men. Diabetes 58: 2129–2136.
42. van de Bunt M, Gloyn AL (2010) From genetic association to molecular mechanism. Curr Diab Rep 10: 452–466.
43. Parnaud G, Bosco D, Berney T, Pattou F, Kerr-Conte J, et al. (2008) Proliferation of sorted human and rat beta cells. Diabetologia 51: 91–100.
44. McCulloch LJ, van de Bunt M, Braun M, Frayn KN, Clark A, et al. (2011) GLUT2 (SLC2A2) is not the principal glucose transporter in human pancreatic beta cells: implications for understanding genetic association signals at this locus. Mol Genet Metab 104: 648–653.
45. Anders S, Huber W (2010) Differential expression analysis for sequence count data. Genome Biol 11: R106.
46. R Development Core Team (2011) R: A Language and Environment for Statistical Computing. Vienna, Austria: R Foundation for Statistical Computing.

nature genetics

Pancreatic islet enhancer clusters enriched in type 2 diabetes risk-associated variants

Lorenzo Pasquali[1,2,14], Kyle J Gaulton[3–5,14], Santiago A Rodríguez-Seguí[1,13,14], Loris Mularoni[1,2], Irene Miguel-Escalada[6], İldem Akerman[1,2], Juan J Tena[7], Ignasi Morán[8], Carlos Gómez-Marín[7], Martijn van de Bunt[3–5], Joan Ponsa-Cobas[8], Natalia Castro[1,2], Takao Nammo[1,13], Inês Cebola[8], Javier García-Hurtado[1,2], Miguel Angel Maestro[1,2], François Pattou[9], Lorenzo Piemonti[10], Thierry Berney[11], Anna L Gloyn[4,5], Philippe Ravassard[12], José Luis Gómez-Skarmeta[7], Ferenc Müller[6], Mark I McCarthy[3–5] & Jorge Ferrer[1,2,8]

Type 2 diabetes affects over 300 million people, causing severe complications and premature death, yet the underlying molecular mechanisms are largely unknown. Pancreatic islet dysfunction is central in type 2 diabetes pathogenesis, and understanding islet genome regulation could therefore provide valuable mechanistic insights. We have now mapped and examined the function of human islet *cis*-regulatory networks. We identify genomic sequences that are targeted by islet transcription factors to drive islet-specific gene activity and show that most such sequences reside in clusters of enhancers that form physical three-dimensional chromatin domains. We find that sequence variants associated with type 2 diabetes and fasting glycemia are enriched in these clustered islet enhancers and identify trait-associated variants that disrupt DNA binding and islet enhancer activity. Our studies illustrate how islet transcription factors interact functionally with the epigenome and provide systematic evidence that the dysregulation of islet enhancers is relevant to the mechanisms underlying type 2 diabetes.

© 2014 Nature America, Inc. All rights reserved.

Despite recent progress in mapping transcriptional regulatory elements and chromatin states[1–7], there is still limited understanding of how transcription factor networks interact with chromatin to control genome function. Furthermore, little is known about how the dysregulation of such networks contributes to disease. As pancreatic islet cells are pivotal in diabetes pathogenesis[8–10] and genomic *cis*-regulatory maps in islets are still markedly underdeveloped compared to those for other cell types[5,7,11–16], we have now mapped chromatin states, binding sites of key transcription factors and transcripts in human pancreatic islets (**Fig. 1a**). Integrative analysis and genetic perturbations were combined to provide reference charts of the *cis*-regulatory elements functioning in human pancreatic islets. Our studies identified enhancer domains that are central determinants of islet-specific gene activity and linked sequence variation in these regions to susceptibility for type 2 diabetes (T2D) and to variation in fasting glycemia levels.

RESULTS

Topology of an essential islet transcription factor network

To characterize the transcription factor networks that control gene activity in human islet cells, we focused on five key β cell transcription factors—FOXA2, MAFB and NKX2.2, which are expressed in insulin-producing β cells and other major islet cell types; NKX6.1, which is specific to β cells; and PDX1, which is present in β cells and more scarce somatostatin-producing islet cells[17–21] (**Supplementary Fig. 1a–c** and **Supplementary Table 1**). Mouse genetic experiments have shown that these five transcription factors are essential for the differentiation or function of β cells[22–24], yet little is known about how these factors function at the level of regulatory networks or how they interact with human islet chromatin to create cell type–specific transcriptional activity. To determine the genomic binding sites of these islet transcription factors, we used chromatin immunoprecipitation and sequencing (ChIP-seq) in duplicate human islet samples

[1]Genomic Programming of Beta-Cells Laboratory, Institut d'Investigacions Biomèdiques August Pi i Sunyer (IDIBAPS), Barcelona, Spain. [2]Centro de Investigación Biomédica en Red de Diabetes y Enfermedades Metabólicas Asociadas (CIBERDEM), Barcelona, Spain. [3]Wellcome Trust Centre for Human Genetics, University of Oxford, Oxford, UK. [4]Oxford Centre for Diabetes, Endocrinology and Metabolism, Churchill Hospital, Oxford, UK. [5]Oxford National Institute for Health Research (NIHR) Biomedical Research Centre, Churchill Hospital, Oxford, UK. [6]School of Clinical and Experimental Medicine, College of Medical and Dental Sciences, University of Birmingham, Birmingham, UK. [7]Centro Andaluz de Biología del Desarrollo (CABD), Consejo Superior de Investigaciones Científicas-Universidad Pablo de Olavide–Junta de Andalucía, Seville, Spain. [8]Department of Medicine, Imperial College London, London, UK. [9]University of Lille 2, INSERM U859 Biotherapies of Diabetes, Lille, France. [10]Diabetes Research Institute, San Raffaele Scientific Institute, Milan, Italy. [11]Cell Isolation and Transplantation Center, Department of Surgery, Geneva University Hospitals and University of Geneva, Geneva, Switzerland. [12]Centre de Recherche de l'Institut du Cerveau et de la Moelle, Biotechnology and Biotherapy Team, CNRS UMR 7225, INSERM U975, University Pierre et Marie Curie, Paris, France. [13]Present addresses: Instituto de Fisiología, Biología Molecular y Neurociencias (IFIBYNE-Consejo Nacional de Investigaciones Científicas y Técnicas), Facultad de Ciencias Exactas y Naturales, Universidad de Buenos Aires, Buenos Aires, Argentina (S.A.R.-S.) and Department of Metabolic Disorder, Diabetes Research Center, Research Institute, National Center for Global Health and Medicine, Shinjuku-ku, Tokyo, Japan (T.N.). [14]These authors contributed equally to this work. Correspondence should be addressed to J.F. (j.ferrer@imperial.ac.uk).

Received 27 March 2013; accepted 12 December 2013; published online 12 January 2014; doi:10.1038/ng.2870

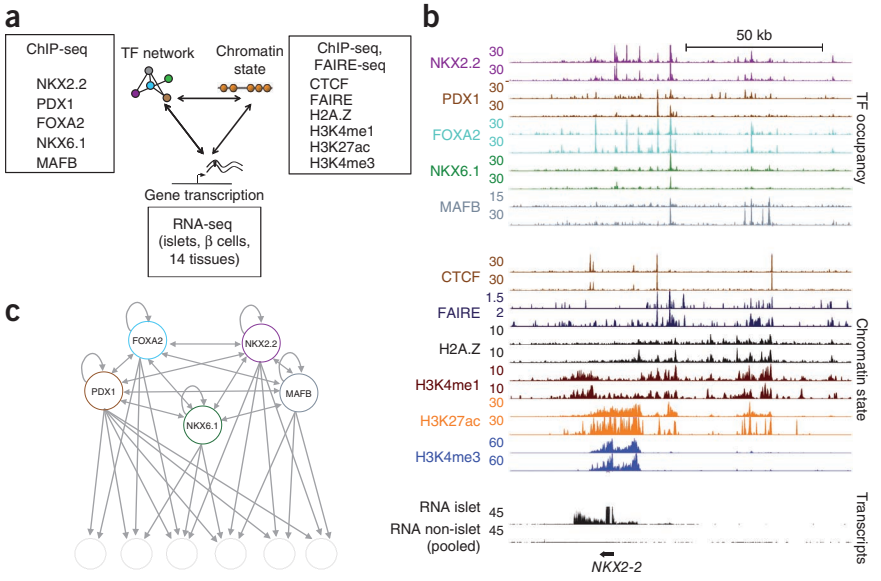


Figure 1 Integrative regulatory maps of human pancreatic islet cells. (**a**) Transcription factor (TF) binding, active chromatin and gene transcription maps in human islet cells. (**b**) Integrative map of the *NKX2-2* locus. Transcription factor binding and chromatin state profiles are shown for duplicate human islet samples. RNA-seq tracks correspond to human islets or pooled data from 14 non-pancreatic tissues are shown highlighting islet-specific transcripts. (**c**) Network topology diagram showing that all five islet-specific transcription factors have direct auto- and cross-regulatory interactions and frequently target adjacent genomic sites (**Supplementary Fig. 2**).

and identified 3,911–32,747 high-confidence sites per transcription factor (**Supplementary Figs. 1d**,**e** and **2a**,**b**). We found that all five transcription factors invariably bound to multiple sites in the vicinity of their own and each other's genes, a pattern indicating auto- and cross-regulatory interactions (**Fig. 1b**,**c** and **Supplementary Fig. 2c**–**f**). Furthermore, all five transcription factors frequently bound to overlapping genomic sites, and, consequently, their genomic binding signals were highly correlated (r = 0.36–0.68 between pairs of islet transcription factors in contrast with r = 0.03–0.11 between islet transcription factors and a control non-islet transcription factor) (**Fig. 1b**,**c** and **Supplementary Fig. 2c**–**h**). This systematic analysis shows that, in agreement with earlier descriptions of transcription factor networks in diverse species and tissues[25,26], human islet-specific transcription factors form a remarkably interconnected network (**Fig. 1c**). These genome-scale maps of binding sites for key islet transcription factors provide a framework for understanding how transcription factors control islet-specific gene transcription.

Islet transcription factors bind to distinct chromatin states

We next sought to understand the relationship between the binding sites of islet-specific transcription factors and underlying chromatin states in human islets. We first mapped all accessible, or open, chromatin sites in pancreatic islets. We combined two assays that capture overlapping, although not identical, sets of accessible chromatin sites: FAIRE-seq, which uses formaldehyde-assisted isolation of regulatory elements to identify open (nucleosome-depleted) sites[5], and ChIP-seq for H2A.Z, a histone variant that is enriched in accessible chromatin sites[27,28]. We defined ~95,000 sites that were enriched in either of these 2 accessible chromatin types in 2 or more human islet samples. Next, we mapped key histone modifications (monomethylation of histone H3 at lysine 4 (H3K4me1), trimethylation of histone H3 at lysine 4 (H3K4me3) and acetylation of histone H3 at lysine 27 (H3K27ac)) and CTCF-binding sites in human islets and used k-median clustering to divide all accessible chromatin sites into five classes (C1–C5) that displayed clearly recognizable enrichment patterns for histone modifications and CTCF binding[1–3], namely, promoters (C1), poised or inactive enhancers (C2), active enhancers (C3), CTCF-bound sites (C4) and additional accessible sites lacking active histone modifications and CTCF occupancy (C5) (**Fig. 2a** and **Supplementary Fig. 3a**–**c**). Consistent with our expectation that transcription factors would bind to accessible DNA, 92% of the genomic sites bound by any of the five islet-specific transcription factors mapped to sites that showed enrichment in FAIRE-seq or H2A.Z ChIP-seq in at least one sample, and islet transcription factors bound to all major classes of accessible chromatin (**Fig. 2b** and **Supplementary Fig. 4**). Integrating histone marks and transcription factor binding maps thus resolved discrete classes of transcription factor–bound accessible chromatin sites in human islets.

Cis-regulatory clusters drive islet-specific transcription

Next, we examined the functional relationship between these different classes of transcription factor–bound, accessible chromatin sites and islet-specific gene transcription. To study islet-specific gene transcription, we analyzed RNA sequencing (RNA-seq) profiles from 3 human islet samples[29], 2 fluorescence-activated cell sorting (FACS)-purified β cell samples[29] and 14 non-pancreatic tissues and used these profiles to define 3 classes of genes: (i) islet-specific genes, which included the 1,000 genes that showed the strongest transcript enrichment in human islet cells, (ii) ubiquitously transcribed genes, which included 1,000 genes that were similarly active in islet cells and other tissues, and (iii) inactive genes, which included the 1,000 genes that showed the lowest expression levels in islet cells. Contrary to our expectation that cell type–specific transcription factors should primarily bind to cell type–specific genes, islet-specific transcription factors bound to promoter (C1) accessible chromatin to a similar extent at both islet-specific and ubiquitously transcribed genes (**Fig. 2c**,**d** and **Supplementary Fig. 5a**,**b**). Similarly, binding events of islet-specific transcription factors at C2, C4 and C5 accessible chromatin sites were not enriched at loci harboring genes specifically expressed in islet cells (**Fig. 2c** and **Supplementary Fig. 5c**). In sharp contrast, gene loci with islet-specific expression had a higher density of islet transcription factor–bound active enhancer (C3) accessible chromatin sites (**Fig. 2c**,**e**). These genes with islet-specific expression had a median of 3 transcription factor–bound C3 sites (interquartile range (IQR) = 1–5) within 50 kb of their transcriptional start site (TSS), whereas ubiquitously transcribed genes had 1 (IQR = 0–3) and islet-inactive genes had a median of 0 (IQR = 0–1) transcription factor–bound C3 sites (Kruskall-Wallis $P < 1 \times 10^{-16}$). Consistent with studies showing that combinatorial transcription factor interactions are critical for enhancer activation[30,31], 72.8% of genomic sites bound by three or more islet-specific transcription factors were associated with C3 accessible chromatin (**Fig. 2f**). Although some promoter (C1) accessible chromatin sites were also bound by multiple islet transcription factors, these sites were not associated with islet-specific transcription, suggesting that the link between transcription

© 2014 Nature America, Inc. All rights reserved.

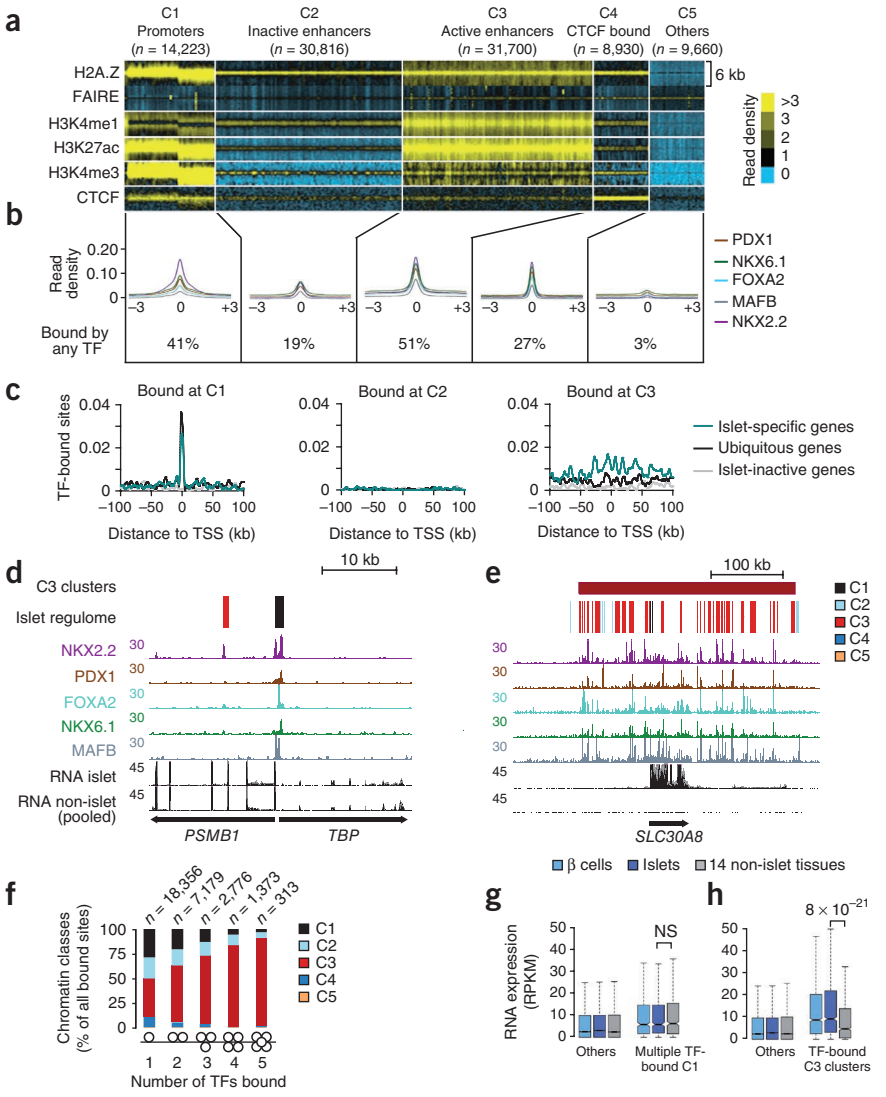


Figure 2 Transcription factor networks establish distinct types of interactions with the epigenome. (**a**) *k*-median clustering of 95,329 islet accessible chromatin sites (defined by FAIRE and/or H2A.Z enrichment in 2 samples) showing 5 classes of accessible chromatin (C1–C5) that we refer to as promoters, inactive enhancers, active enhancers, CTCF-bound sites and other, on the basis of previously defined patterns of histone modification and CTCF binding[1–3] (**Supplementary Fig. 3a**). Histone- and CTCF-bound reads were divided into 100-bp bins across 6-kb windows centered on merged FAIRE- or H2A.Z-enriched sites. (**b**) Average transcription factor binding signal distribution relative to the center of each accessible chromatin class and the percentage of sites in each class bound by at least one transcription factor. Non-quantile-normalized reads were processed as described for **a**. (**c**) Density of C1, C2 and C3 accessible chromatin sites bound by 2 or more transcription factors in the regions surrounding the TSSs of the 1,000 genes with the most islet-specific expression, 1,000 ubiquitously active genes and 1,000 islet-inactive genes. (**d**,**e**) Examples of islet-specific transcription factor binding to the 5′ ends of the ubiquitously active genes *TBP* and *PSMB1* (**d**) or to the islet-specific T2D susceptibility gene *SLC30A8* (**e**). The regulome track depicts colored chromatin states for classes C1–C5. (**f**) Binding by multiple transcription factors is more common at C3 (active enhancer) chromatin. The numbers of transcription factor–bound sites consistent in two samples for each category are shown above. (**g**) Binding by three or more transcription factors at C1 (promoter) chromatin is not associated with islet-specific activity of the adjacent gene. NS, not significant. (**h**) Genes located <25 kbfrom clusters of C3 sites that are highly bound by transcription factors show enriched expression in islets and β cells relative to 14 non-islet tissues. Boxes show IQR, notches indicate 95% confidence intervals of the median, and whiskers extend to either 1.5 times the IQR or to extreme values. Gene expression data from non-pancreatic tissues in **g**,**h** are presented by individual tissue in **Supplementary Figures 5d** and **7b**. *P* values were calculated with the Wilcoxon rank-sum test.

© 2014 Nature America, Inc. All rights reserved.

factor occupancy and cell type–specific transcription is largely confined to active enhancers (**Fig. 2g** and **Supplementary Fig. 5d**). Taken together, these findings showed remarkable differences between transcription factor binding at sites in the different accessible chromatin classes. These findings suggest that certain binding events of islet-specific transcription factors, such as those occurring at promoters, are not generally associated with cell type–specific gene transcription, which is instead tightly linked to clusters of active enhancer sites bound by multiple transcription factors.

We scanned the genome to identify all clusters of three or more enhancer sites active in islets (Online Methods and **Supplementary Fig. 6a**,**b**). We identified 3,677 such clusters, which mapped near genes with strong islet-enriched expression (median expression in islets of 8.3 reads per kilobase per million mapped reads (RPKM) (IQR = 2.9–19.6) and in non-islet tissues of 4.8 RPKM (IQR = 0.7–14.2); Wilcoxon $P < 1 \times 10^{-30}$), in contrast to more modest islet enrichment in genes near non-clustered (orphan) enhancers (median expression in islets of 6.0 RPKM (IQR = 1.5–14.3) and in non-islet tissues of 5.0 RPKM (IQR = 0.8–13.9); Wilcoxon $P < 1 \times 10^{-5}$) (**Supplementary Fig. 6c**). Islet-enriched transcription was most pronounced near the 1,813 enhancer clusters that showed higher than median occupancy by islet transcription factors (**Fig. 2h** and **Supplementary Fig. 7a–c**). Remarkably, most genes that are currently known to be important for islet cell identity, function or disease were associated with an islet enhancer cluster (92% of a manually annotated list of 65 such genes, 90% of which belonged to the subset with high transcription factor occupancy; **Supplementary Fig. 7d** and **Supplementary Table 2**). These findings reinforce the view that the genomic program that underlies islet cell identity is tightly associated with clusters of enhancers bound by islet transcription factors.

To investigate the function of transcription factor–bound enhancer clusters, we first verified that individual clustered C3 sites truly acted as cell type–specific enhancers. Consistent with their chromatin signature, 8 of 12 transcription factor–bound active enhancer C3 sites (but not transcription factor–bound C2 or C5 sites) displayed β cell–selective enhancer activity in episomal reporter assays in mouse cell lines (**Fig. 3a**). Five conserved transcription factor–bound C3 sites were tested in transgenic zebrafish assays, and all exhibited

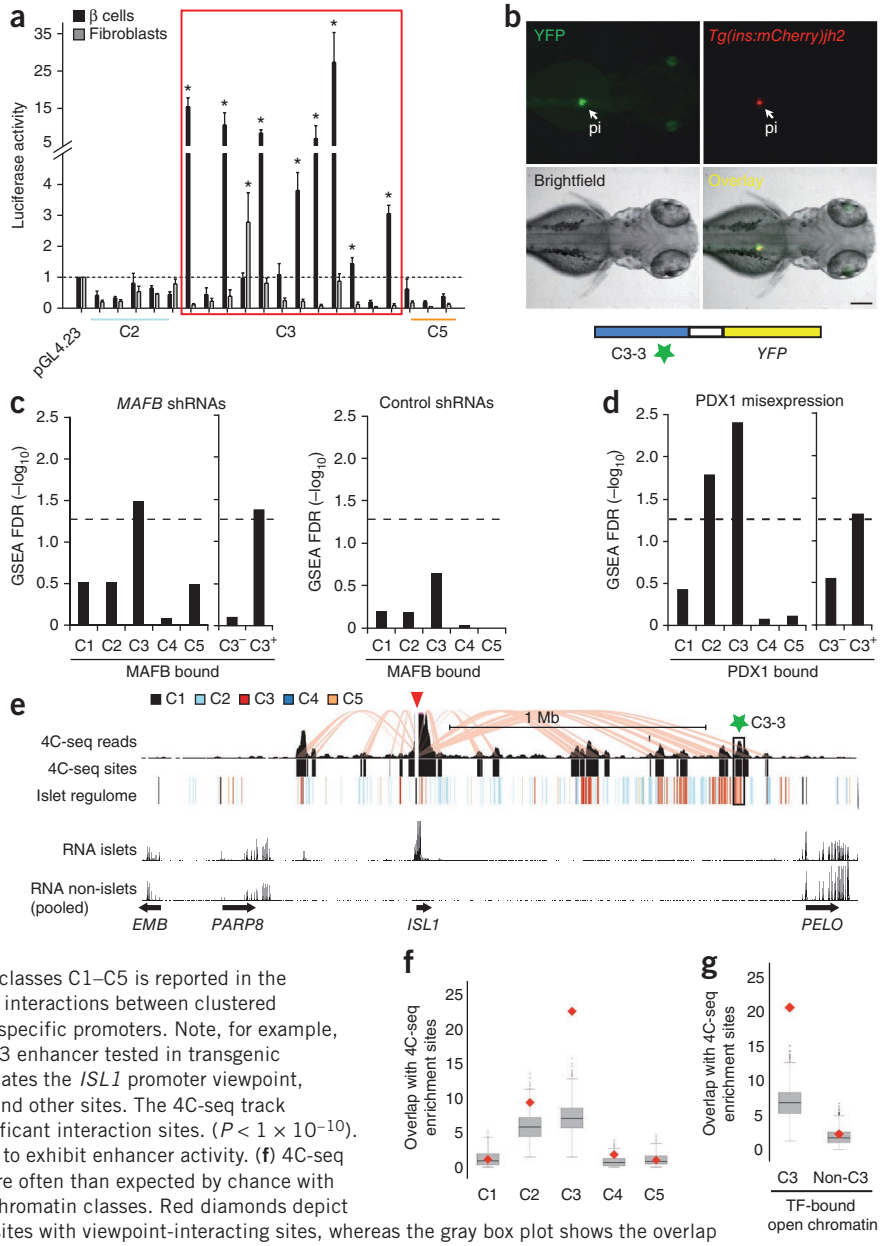


Figure 3 Enhancer clusters form functional three-dimensional chromatin domains. (**a**) Luciferase assays in mouse MIN6 β cells and NIH3T3 fibroblasts showing that 8 of 12 transcription factor–bound C3 sites (red box) but not transcription factor–bound C2 or C5 sites confer cell type–specific expression to a minimal promoter (*$P < 0.005$, *t* test for comparison with empty vector; $n = 3$ transfections per condition). Data are presented as mean ± s.d. (**b**) C3-3, a C3 element in a cluster >1 Mb from *ISL1*, shows selective enhancer activity in pancreatic islet from zebrafish embryos 70 h post-fertilization. The enhancer transgene (*YFP*) was injected into a transgenic line that exhibits fluorescence (mCherry) in insulin-positive cells. Scale bar, 0.1 mm. (**c**) *MAFB* knockdown in human EndoC-β H1 β cells causes the downregulation of genes bound by MAFB at clustered ($C3^+$) rather than orphan ($C3^-$) MAFB-bound enhancers. Bar plots show GSEA FDRs for different MAFB-bound gene sets for the genes that are downregulated after infecting cells with viruses expressing two short hairpin RNAs (shRNAs) for *MAFB* versus four non-targeting shRNAs. As a control, we repeated the analysis using the same number of arrays but compared sets of different control shRNAs. Horizontal dashed lines signify FDR = 0.05 as a reference. (See also **Supplementary Fig. 9a–e**.) (**d**) Misexpression of PDX1 with MAFA and NGN3 in HEK293T cells preferentially activates genes associated with PDX1-bound clustered ($C3^+$) enhancers but not genes bound by PDX1 at promoter accessible chromatin sites or those associated with orphan ($C3^-$) PDX1-bound enhancers. The number of transcription factor–bound genes for classes C1–C5 is reported in the Online Methods. (**e**) 4C-seq analysis shows selective interactions between clustered transcription factor–bound enhancers and cell type–specific promoters. Note, for example, interactions between the *ISL1* promoter and the C3-3 enhancer tested in transgenic zebrafish located >1 Mb away. The red triangle indicates the *ISL1* promoter viewpoint, and pink lines highlight interactions between *ISL1* and other sites. The 4C-seq track shows aligned sequences, and 4C-seq sites are significant interaction sites. ($P < 1 \times 10^{-10}$). A green star marks region C3-3, which is shown in **b** to exhibit enhancer activity. (**f**) 4C-seq analysis of nine loci shows that C3 sites interact more often than expected by chance with nearby promoters in contrast with other accessible chromatin classes. Red diamonds depict the mean overlap of different accessible chromatin sites with viewpoint-interacting sites, whereas the gray box plot shows the overlap of randomized accessible chromatin sites (classes C1–C5) after 1,000 iterations at the same loci. (**g**) Transcription factor–bound C3 but not other transcription factor–bound sites establish frequent interactions with nearby promoters. Boxes show IQR, and whiskers extend to 1.5 times the IQR.

enhancer activity (three specifically in the pancreatic islet), whereas transcription factor–bound C2 and C5 sites showed no activity (**Fig. 3b** and **Supplementary Fig. 8a–c**). These experiments show that clustered sites have islet enhancer activity.

We next designed experiments to directly test whether transcription factor binding sites that were present in clusters of enhancers, but not necessarily other transcription factor binding sites, were functionally important for the activation of islet-selective genes. First, we transduced human β cells (EndoC-β H1) with two independent interfering hairpin RNAs for *MAFB* or with four control hairpins (**Supplementary Fig. 9a**). Knockdown of *MAFB* caused downregulation of genes linked to MAFB-bound enhancer clusters (Gene Set Enrichment Analysis (GSEA) normalized enrichment score (NES) = 1.98, false discovery rate (FDR) $q < 1.5 \times 10^{-2}$) (**Fig. 3c**), such as *ROBO2* and *G6PC2* (**Supplementary Fig. 9b**,**c**). However, the expression of genes bound by MAFB only at their promoters or at other classes of accessible chromatin was not significantly affected (**Fig. 3c** and **Supplementary Fig. 9d**,**e**). Next, we transfected HEK293T embryonic kidney cells with expression vectors encoding PDX1 and assessed which PDX1-bound genes were activated. We coexpressed PDX1 with MAFA and NGN3 because this transcription factor combination, which has previously been employed to activate β cell genes[32], was more efficient than PDX1 alone in gene activation in heterologous cells. We observed transcriptional enhancement of genes near PDX1-bound enhancer clusters (GSEA NES = 2.2, FDR $q < 1 \times 10^{-3}$), whereas the expression of genes that were exclusively bound by PDX1 at promoter or other accessible chromatin classes was unaffected (GSEA NES = −1.3, FDR $q = 0.37$) (**Fig. 3d**). These studies confirm that chromatin profiling can identify subsets of transcription

© 2014 Nature America, Inc. All rights reserved.

factor binding sites that are important for cell type–specific gene activity.

Having shown that active enhancer clusters are pervasive at genes with islet-specific expression and, furthermore, that these clusters are critical for islet-specific gene activity, we examined the relationship between enhancer multiplicity and the formation of broad three-dimensional chromatin structures. We selected nine loci, eight of which contained active enhancer clusters near a gene with islet-enriched expression, and used human islet chromatin to perform circular chromosome conformation capture coupled with high-throughput sequencing (4C-seq) (**Fig. 3e**)[33]. Transcription factor–bound C3 sites from these clusters displayed frequent, strong interactions with the promoters of genes at the same loci with islet-enriched expression, including C3 sites located >1 Mb from their target promoter (**Fig. 3e–g**; additional loci are shown in **Supplementary Fig. 10a–e**). In contrast, transcription factor–bound non-C3 sites at the same loci did not show greater interactions than expected by chance ($P < 1 \times 10^{-3}$ for transcription factor–bound C3 sites versus randomized sites and $P = 0.41$ for transcription factor–bound non-C3 sites versus randomized sites) (**Fig. 3f,g**). These observations suggest that islet enhancer clusters are three-dimensional structural units in which multiple transcription factor–bound active enhancers (rather than other transcription factor–bound sites) interact with target promoters. Collectively, the results indicate that human islet-specific gene transcription is largely driven by combinatorial transcription factor binding at clusters of enhancers that form three-dimensional chromatin structures.

Islet *cis*-regulatory sequence code

The identification of the genomic sites that underlie islet gene activity allowed us to explore the sequence code that is recognized by islet transcription factor networks to drive cell type–specific transcription. We identified 46 highly enriched sequence motifs in clustered enhancers, including recognition sequences for key β cell transcription factors (HNF1A, RFX, FOXA, NEUROD1, NKX6.1, PDX1, MAFA and MAFB), as well as previously unrecognized motifs (**Fig. 4** and **Supplementary Table 3**). We further identified combinations of these 46 motifs that were most enriched in islet enhancers relative to enhancers from 9 non-pancreatic cell types[34] (**Supplementary Fig. 11a–c**). Motifs for RFX[35] or the pioneer FOXA factors[36] were present in all of the most islet-enriched combinations (**Supplementary Fig. 11b**), consistent with a major role for these factors in the activation of islet enhancers. To independently validate this analysis, we mapped all instances of the ten most enriched motif combinations to the mouse genome and found that they were also located near genes that show islet-enriched expression in mice (median expression of nearby genes in mouse islets versus non-islet tissues of 1.0 (IQR = 0.0–5.2) RPKM versus 0.6 (IQR = 0.0–3.7) RPKM; Wilcoxon $P = 9.6 \times 10^{-31}$) (**Supplementary Fig. 11d**). *In silico* analysis thus expanded the range of candidate combinatorial islet transcription factor interactions and provides a resource to discover new transcription factors and upstream signaling pathways that control β cell transcriptional programs.

Sequence variation in islet enhancers is associated with T2D

Genome-wide association studies (GWAS) have identified dozens of loci associated with T2D and glycemic traits[37,38], but the molecular mechanisms linking specific alleles to cellular functions remain poorly described. The catalog of islet *cis*-regulatory elements allowed us to explore to what extent genetic variants that contribute to T2D susceptibility and to variation in fasting glycemia in non-diabetic individuals act through islet regulatory mechanisms.

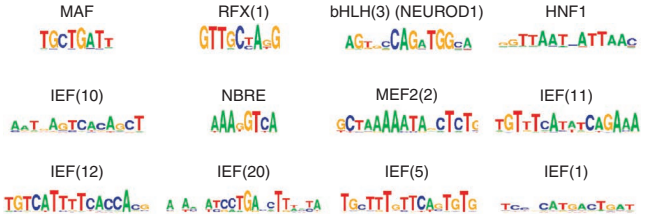


Figure 4 Known and new transcription factor motifs are enriched in clustered islet enhancers. Examples are shown of sequence motifs that are enriched in clustered enhancers at $P < 1 \times 10^{-60}$ (HOMER)[50]. Several motifs match known islet transcription factor recognition sequences, whereas others are candidate binding sites for new regulators. A complete list of motifs that showed enrichment at $P < 1 \times 10^{-60}$ (HOMER) is given in **Supplementary Table 3**. The name of each motif corresponds to the annotations shown in **Supplementary Table 3**. IEF, islet-enriched factor.

First, we examined loci with genome-wide significant association with T2D or fasting glycemia in European populations, considering all variants in high linkage disequilibrium (LD; 1000 Genomes Project CEU (Utah residents of Northern and Western European ancestry) $r^2 > 0.8$) with a lead SNP reported in the National Human Genome Research Institute (NHGRI) GWAS catalog[39]. These associated variants were enriched in C3 sites and occurred more prominently in clustered C3 rather than in orphan C3 sites (T2D: C3 $P = 0.004$; clustered C3 $P = 0.012$, orphan C3 $P = 0.20$; fasting glycemia: C3 $P < 1 \times 10^{-5}$; clustered C3 $P = 3 \times 10^{-5}$, orphan C3 $P = 0.12$) (**Fig. 5a**). There was no significant enrichment in variants associated with T2D or fasting glycemia for other accessible chromatin classes, and variants associated with other complex traits were not enriched at clustered islet C3 sites (all traits $P > 0.05$). Moreover, little enrichment was seen in the 200-bp sites immediately flanking C3 sites (**Fig. 5a**; T2D $P = 0.44$, fasting glycemia $P = 0.043$), suggesting that C3 enrichment is not driven solely by the genomic context of these sites. These findings indicate that a subset of loci with genome-wide association with T2D and fasting glycemia are likely to harbor functional *cis*-regulatory variants in islet active enhancers.

To further examine the contribution of islet *cis*-regulatory variation to T2D and fasting glycemia, we considered association results for the full set of ~2.5 million variants represented in the largest available genome-wide association meta-analyses for these traits rather than restricting analysis to signals reaching stringent genome-wide significance thresholds[37,38]. For each islet accessible chromatin class, we identified the overlapping variants that exceeded a series of association significance thresholds and calculated fold enrichment values using a null distribution of permuted variant counts (Online Methods). Variants in C3 sites were collectively enriched for T2D and fasting glycemia genome-wide association values of $P < 0.001$ (T2D, C3 $P = 2.1 \times 10^{-11}$; fasting glycemia, C3 $P < 1 \times 10^{-16}$). This enrichment was predominantly a feature of C3 clusters rather than of orphan C3 sites (T2D: clustered C3 $P = 5.1 \times 10^{-11}$, orphan C3 $P = 0.006$; fasting glycemia: clustered C3 $P < 1 \times 10^{-11}$, orphan C3 $P = 0.23$) (**Fig. 5b** and **Supplementary Fig. 12**). Fold enrichment patterns were unchanged when the genome-wide association data were aggressively pruned ($r^2 > 0.2$) to retain only the most associated signal in each LD cluster and remained significant for variants associated with T2D and fasting glycemia in clustered C3 sites (T2D: C3 $P = 0.2$, clustered C3 $P = 0.031$; fasting glycemia: C3 $P = 3.6 \times 10^{-6}$, clustered C3 $P = 5.0 \times 10^{-5}$). Enrichment was also retained when established (genome-wide significant)

© 2014 Nature America, Inc. All rights reserved.

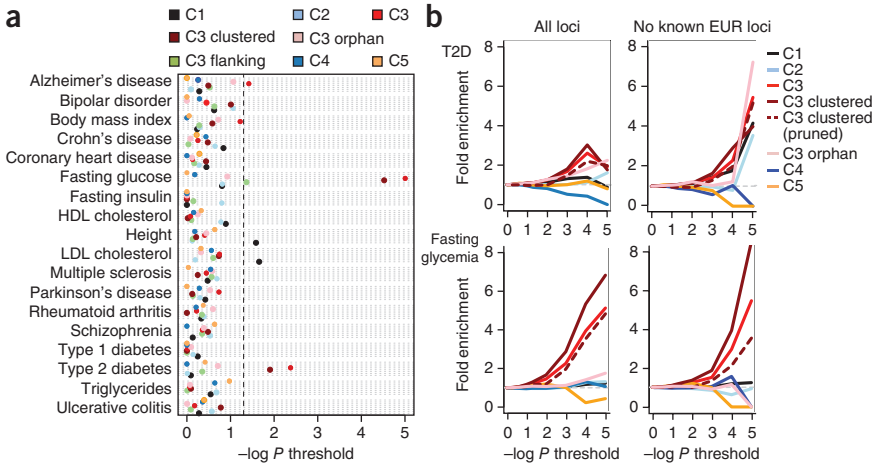


Figure 5 Islet enhancers are enriched in loci associated with T2D and fasting glycemia. (**a**) Loci associated with complex traits were tested for enrichment of overlap with classes of islet sites (C1–C5, clustered C3 and orphan C3) compared to matched background loci. Loci associated with T2D and fasting glycemia were enriched for islet active enhancer sites (C3) but not for directly flanking sites (C3 flanking) or for other islet classes. The enrichment of C3 sites for both T2D and fasting glycemia was stronger when considering clustered rather than orphan (non-clustered) enhancers. The most significant enrichments for other complex diseases were found for C1 sites with low-density lipoprotein (LDL) cholesterol ($P = 0.022$) and height ($P = 0.026$) and for C3 sites with Alzheimer's disease ($P = 0.038$). No significant enrichment of variants associated with other complex traits was seen in clustered C3 enhancers. HDL, high-density lipoprotein. The horizontal dashed line shows P value = 0.05 as a reference. (**b**) Enrichment of islet regulome sites in T2D[38] and fasting glycemia[37] genome-wide association data. We determined the number of variants overlapping sites from each islet accessible chromatin class that surpassed a series of association significance thresholds for T2D and fasting glycemia. We then calculated fold enrichment values at each threshold by comparison to the number of matched background variants significant at that threshold. C3 sites and clustered C3 sites were more enriched for T2D and fasting glycemia association at increasing P-value thresholds (left), even after removing known European (EUR) T2D and fasting glycemia loci (right). Patterns in C3 and clustered C3 sites were maintained when pruning variants to retain a single variant in each LD block ($r^2 > 0.2$). In all such analyses, clustered C3 sites displayed significant enrichment relative to the null distribution of permuted counts for T2D and fasting glycemia association P values of <0.001 (P values for these analyses are given in the text).

T2D and fasting glycemia loci were removed (**Fig. 5b**; T2D: C3 $P = 7.3 \times 10^{-5}$, clustered C3 $P = 1.1 \times 10^{-5}$; fasting glycemia: C3 $P = 1.1 \times 10^{-5}$, clustered C3 $P = 4.1 \times 10^{-9}$), suggesting that the association of variants at C3 sites (and, in the T2D analysis, also of variants at C1 sites) extends to new loci that have yet to be confirmed to show genome-wide significant association with these traits.

To further examine this point, we sought examples of variants overlapping islet accessible chromatin sites that mapped within such moderately associated loci. We found that rs72695654 disrupted a sequence motif and abolished binding of a β cell protein complex to a C3 site that is located in an enhancer cluster within the *ACSL1* gene, which encodes a long-chain fatty acyl-CoA synthase that is highly expressed in islets (**Supplementary Fig. 13a–c**). Long-chain fatty acyl-CoA synthesis has been implicated in β cell stimulus-secretion coupling, survival and lipotoxicity[40–42]. The rs72695654 variant was in high LD (1000 Genomes Project CEU $r^2 = 0.94$) with rs735949, which is strongly associated with T2D ($P = 3.7 \times 10^{-6}$) and, independently, with

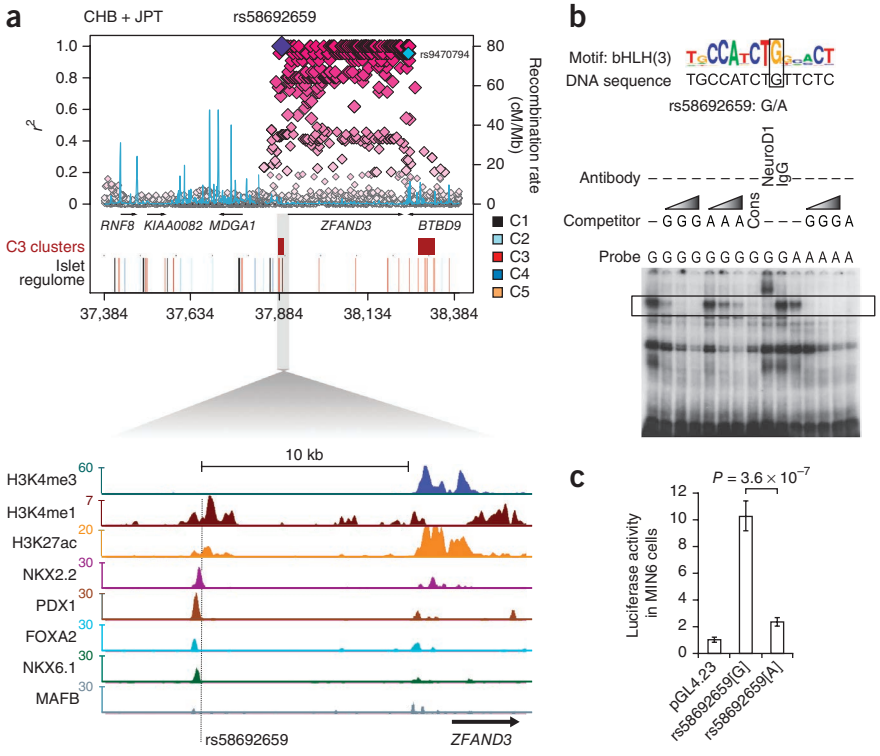


Figure 6 A T2D risk variant at *ZFAND3* disrupts islet enhancer activity. (**a**) Regional plot of 1000 Genomes Project variants and islet accessible chromatin elements at the *ZFAND3* locus associated with T2D in East Asian individuals (r^2 values based on 1000 Genomes Project CHB and JPT LD with rs58692659). rs9470794 (cyan) at this locus shows the strongest T2D association from Cho *et al.*[44]. The rs58692659 SNP maps to a C3 site >10 kb upstream of *ZFAND3* bound by PDX1, FOXA2, NKX2.2 and NKX6.1. (**b**) The rs58692659 SNP disrupts an islet-enriched bHLH-like motif that matches the recognition site for NEUROD1 (ref. 51), a known and important islet transcription factor[22]. The electrophoretic mobility shift assay (EMSA) shows that the minor A allele at this variant abolishes binding of a protein complex that is supershifted by antibody to NEUROD1. The black box highlights the sequence-specific retardation complex. Competition gradients, represented by gray triangles, correspond to 5-fold, 50-fold and 100-fold excess of cold probe. (**c**) Luciferase assays showing reduced enhancer activity of the A allele compared to the G allele of rs58692659 in MIN6 β cells. Data are presented as mean ± s.d. Three independent experiments were performed in triplicate. P values were calculated by two-sided t test.

© 2014 Nature America, Inc. All rights reserved.

fasting glycemia levels in non-diabetic individuals in large genome-wide association meta-analyses ($P = 1.6 \times 10^{-5}$)[37,38] (**Supplementary Fig. 13a**). This finding suggests that *cis*-regulatory maps can be used to prioritize loci on the basis of which genetic evidence is strong, although not genome-wide significant, to guide hypothesis-driven experiments that define susceptibility variants.

We next attempted to define functional enhancer variants in established genome-wide associated loci that might be causal in T2D susceptibility. We first catalogued all variants that resided within C3 sites and were in high LD with established genome-wide significant T2D and fasting glycemia association signals (**Supplementary Tables 4–6**)[37,38]. We confirmed that the *TCF7L2* intronic SNP[43] rs7903146, previously shown to be located in an islet FAIRE-enriched site and to affect chromatin state and enhancer function[5], mapped to a C3 site that is bound by NKX2.2, FOXA2 and MAFB, yet did not show active chromatin marks in an extensive panel of non-islet human cell types (**Supplementary Fig. 14a**,**b**). Similar observations were made at multiple established T2D and fasting glycemia regions, including at *SLC2A2*, *CDKN2A*, *C2CD4A*, *C2CD4B*, *SLC30A8*, *DGKB* and *PCSK1*, where we identified associated variants that mapped to transcription factor–bound C3 sites, many of which disrupted islet-enriched sequence motifs (see **Supplementary Fig. 15a**,**b**, which describes a public browser that facilitates the visualization of T2D and fasting glycemia genome-wide associated variants and the islet regulome, and **Supplementary Fig. 16**). For example, at the *ZFAND3* T2D locus discovered in East Asians[44], SNP rs58692659 formed part of an array of variants that was highly correlated with the reported lead SNP (rs9470794) (1000 Genomes Project CHB (Han Chinese in Beijing, China) and JPT (Japanese in Tokyo, Japan) $r^2 = 0.96$) and mapped to a C3 element bound by multiple islet transcription factors within an enhancer cluster (**Fig. 6a**). This variant altered sequence-specific DNA binding of a key islet-enriched transcription factor, NEUROD1, and abolished enhancer activity in β cells (**Fig. 6b**,**c**). We note that this SNP, along with many others in the region, is monomorphic in European samples (1000 Genomes Project CEU) and that there is no T2D association at *ZFAND3* in Europeans (rs9470794 DIAbetes Genetics Replication And Meta-analysis (DIAGRAM) $P = 0.8$)[44], consistent with rs58692659 having a causal role at this locus. These results show how islet accessible chromatin maps can pinpoint functional *cis*-regulatory variants that are strong candidates for having a causal role in driving T2D association signals.

DISCUSSION

Our work provides reference *cis*-regulatory maps for ongoing efforts to dissect the transcriptional program of pancreatic β cells[45] and the molecular mechanisms of human T2D[10]. We show that islet-specific transcription factors establish widespread binding to accessible chromatin sites where they are apparently not engaged in regulation in *cis* of cell type–specific gene transcription. We further demonstrate that binding events that do drive the transcription of genes underlying islet cell identity reside in clusters of transcriptional enhancers. These clusters mirror previously reported clusters of accessible chromatin in human islets[5], and, although they constitute groups of discrete enhancers, they share many features and may thus represent the same phenomenon as recently reported for large enhancers (superenhancers)[46,47]. Our 4C-seq analysis allows us to speculate that enhancer multiplicity serves an architectural role by creating active chromatin structures in genomic domains that are transcriptionally silent in most non-islet cell types. These findings have therefore uncovered central *cis*-regulatory determinants of islet cell gene transcription.

Our systematic analysis implicates sequence variation at islet enhancer clusters in susceptibility to T2D and in variation in fasting glycemia levels. Recently, SNPs associated with common polygenic diseases were shown to be enriched in noncoding genomic elements defined by DNase I hypersensitivity, which marks accessible chromatin regions such as active promoters, insulators, repressors, and poised and active enhancers[48], or by H3K4me3, which is enriched at promoters and, more weakly, at enhancers[49]. We have now defined for the first time, to our knowledge, functionally distinct transcription factor–bound genomic sites in human islets and have thereby disclosed that T2D susceptibility is specifically associated with allelic variation in pancreatic islet distant enhancers. Our results therefore link islet *cis*-regulatory networks to the mechanisms underlying T2D susceptibility and variation in glycemic traits. The availability of integrated *cis*-regulatory maps in pancreatic islets will facilitate hypothesis-driven experiments to establish the exact manner in which common and lower frequency genetic variants affect islet cells in human diabetes.

The islet regulome, including information on transcription factor occupancy, chromatin states, motifs, enhancer clusters and genome-wide significant *P* values for association with T2D and fasting glycemia, can be accessed and visualized online (see URLs and **Supplementary Fig. 15** for a description of this browser).

URLs. Islet Regulome Browser, http://www.isletregulome.org/.

METHODS

Methods and any associated references are available in the online version of the paper.

Accession codes. Raw data are available at ArrayExpress under accession E-MTAB-1919. Illumina human BodyMap2 data are available from ArrayExpress under accession E-MTAB-513.

Note: Any Supplementary Information and Source Data files are available in the online version of the paper.

ACKNOWLEDGMENTS

We thank J. Rios (IDIBAPS) for expert statistical advice, M. Parsons (Johns Hopkins University) for *Tg*(*ins*:*mCherry*)*jh2* transgenics and R. Stein (Vanderbilt University), J. Habener (Harvard University) and G. Gradwohl (Institute of Genetics and Molecular and Cellular Biology) for MAFA, IDX1, and NGN3 DNA constructs. We thank the DIAGRAM and MAGIC consortia, the Singapore Prospective Study Program, the Singapore Consortium of Cohort Studies, the Singapore Indian Eye Study, the Singapore Malay Eye Study and Y.Y. Teo, E.S. Tai, T.Y. Wong, W.Y. Lim and X. Wang (National University of Singapore; funded by the National Medical Research Council of Singapore, Singapore Translational Researcher Award schemes, the Biomedical Research Council of Singapore and the National Research Foundation (NRF) Fellowship scheme). This work was carried out in part at the Centre Esther Koplowitz. This work was funded by grants from a European Foundation for the Study of Diabetes Lilly fellowship (L. Pasquali), the Ministerio de Economía y Competitividad (SAF2011-27086 to J.F., BFU2010-14839 and CSD2007-00008 to J.L.G.S.), the Innovative Medicines Initiative (DIRECT to M.I.M. and J.F.), the Andalusian Government (CVI-3488 to J.L.G.S.), the Biology of Liver and Pancreatic Development and Disease Marie Curie Initial Training Network (F.M. and J.F.), the Wellcome Trust (090532, 098381 and 090367 to M.I.M., 095101 to A.L.G., 101033 to J.F.), Juvenile Diabetes Research Foundation (31-2012-783 to T.B., F.P. and L. Piemonti) and Framework Programme 7 (HEALTH-F4-2007-201413 to M.I.M.).

AUTHOR CONTRIBUTIONS

L. Pasquali, K.J.G., J.L.G.S., I.M.-E., F.M., M.I.M. and J.F. analyzed integrated data and wrote and edited the manuscript, which all authors have approved. İ.A., S.A.R.-S., J.P.-C., J.G.-H., T.N., I.M.-E., C.G.-M., I.C., N.C., M.A.M. and J.J.T. performed and analyzed experiments. L. Pasquali, K.J.G., L.M., J.P.-C. and I.M. performed

© 2014 Nature America, Inc. All rights reserved.

computational analysis. A.L.G., M.v.d.B., F.P., L. Piemonti, T.B. and P.R. provided materials and reagents. L. Pasquali and J.F. conceived and coordinated the project.

COMPETING FINANCIAL INTERESTS

The authors declare competing financial interests: details are available in the online version of the paper.

Reprints and permissions information is available online at http://www.nature.com/reprints/index.html.

1. Rada-Iglesias, A. *et al.* A unique chromatin signature uncovers early developmental enhancers in humans. *Nature* **470**, 279–283 (2011).
2. Heintzman, N.D. *et al.* Histone modifications at human enhancers reflect global cell-type-specific gene expression. *Nature* **459**, 108–112 (2009).
3. Creyghton, M.P. *et al.* Histone H3K27ac separates active from poised enhancers and predicts developmental state. *Proc. Natl. Acad. Sci. USA* **107**, 21931–21936 (2010).
4. Bonn, S. *et al.* Tissue-specific analysis of chromatin state identifies temporal signatures of enhancer activity during embryonic development. *Nat. Genet.* **44**, 148–156 (2012).
5. Gaulton, K.J. *et al.* A map of open chromatin in human pancreatic islets. *Nat. Genet.* **42**, 255–259 (2010).
6. Shen, Y. *et al.* A map of the *cis*-regulatory sequences in the mouse genome. *Nature* **488**, 116–120 (2012).
7. ENCODE Project Consortium. An integrated encyclopedia of DNA elements in the human genome. *Nature* **489**, 57–74 (2012).
8. Kahn, S.E. Clinical review 135: the importance of β-cell failure in the development and progression of type 2 diabetes. *J. Clin. Endocrinol. Metab.* **86**, 4047–4058 (2001).
9. Lyssenko, V. *et al.* Clinical risk factors, DNA variants, and the development of type 2 diabetes. *N. Engl. J. Med.* **359**, 2220–2232 (2008).
10. Ashcroft, F.M. & Rorsman, P. Diabetes mellitus and the β cell: the last ten years. *Cell* **148**, 1160–1171 (2012).
11. Bhandare, R. *et al.* Genome-wide analysis of histone modifications in human pancreatic islets. *Genome Res.* **20**, 428–433 (2010).
12. Stitzel, M.L. *et al.* Global epigenomic analysis of primary human pancreatic islets provides insights into type 2 diabetes susceptibility loci. *Cell Metab.* **12**, 443–455 (2010).
13. Khoo, C. *et al.* Research resource: the Pdx1 cistrome of pancreatic islets. *Mol. Endocrinol.* **26**, 521–533 (2012).
14. Bernstein, B.E. *et al.* The NIH Roadmap Epigenomics Mapping Consortium. *Nat. Biotechnol.* **28**, 1045–1048 (2010).
15. Zhu, J. *et al.* Genome-wide chromatin state transitions associated with developmental and environmental cues. *Cell* **152**, 642–654 (2013).
16. Tennant, B.R. *et al.* Identification and analysis of murine pancreatic islet enhancers. *Diabetologia* **56**, 542–552 (2013).
17. Artner, I. *et al.* MafB is required for islet β cell maturation. *Proc. Natl. Acad. Sci. USA* **104**, 3853–3858 (2007).
18. Ahlgren, U., Jonsson, J., Jonsson, L., Simu, K. & Edlund, H. β-cell-specific inactivation of the mouse *Ipf1/Pdx1* gene results in loss of the β-cell phenotype and maturity onset diabetes. *Genes Dev.* **12**, 1763–1768 (1998).
19. Sander, M. *et al.* Homeobox gene *Nkx6.1* lies downstream of *Nkx2.2* in the major pathway of β-cell formation in the pancreas. *Development* **127**, 5533–5540 (2000).
20. Sund, N.J. *et al.* Tissue-specific deletion of *Foxa2* in pancreatic β cells results in hyperinsulinemic hypoglycemia. *Genes Dev.* **15**, 1706–1715 (2001).
21. Sussel, L. *et al.* Mice lacking the homeodomain transcription factor Nkx2.2 have diabetes due to arrested differentiation of pancreatic β cells. *Development* **125**, 2213–2221 (1998).
22. Servitja, J.M. & Ferrer, J. Transcriptional networks controlling pancreatic development and β cell function. *Diabetologia* **47**, 597–613 (2004).
23. Wilson, M.E., Scheel, D. & German, M.S. Gene expression cascades in pancreatic development. *Mech. Dev.* **120**, 65–80 (2003).
24. Oliver-Krasinski, J.M. & Stoffers, D.A. On the origin of the β cell. *Genes Dev.* **22**, 1998–2021 (2008).
25. Gerstein, M.B. *et al.* Integrative analysis of the *Caenorhabditis elegans* genome by the modENCODE project. *Science* **330**, 1775–1787 (2010).
26. Kyrmizi, I. *et al.* Plasticity and expanding complexity of the hepatic transcription factor network during liver development. *Genes Dev.* **20**, 2293–2305 (2006).
27. Jin, C. *et al.* H3.3/H2A.Z double variant–containing nucleosomes mark 'nucleosome-free regions' of active promoters and other regulatory regions. *Nat. Genet.* **41**, 941–945 (2009).
28. Ong, C.T. & Corces, V.G. Enhancer function: new insights into the regulation of tissue-specific gene expression. *Nat. Rev. Genet.* **12**, 283–293 (2011).
29. Morán, I. *et al.* Human β cell transcriptome analysis uncovers lncRNAs that are tissue-specific, dynamically regulated, and abnormally expressed in type 2 diabetes. *Cell Metab.* **16**, 435–448 (2012).
30. Stefflova, K. *et al.* Cooperativity and rapid evolution of cobound transcription factors in closely related mammals. *Cell* **154**, 530–540 (2013).
31. Zinzen, R.P., Girardot, C., Gagneur, J., Braun, M. & Furlong, E.E. Combinatorial binding predicts spatio-temporal *cis*-regulatory activity. *Nature* **462**, 65–70 (2009).
32. Zhou, Q., Brown, J., Kanarek, A., Rajagopal, J. & Melton, D.A. *In vivo* reprogramming of adult pancreatic exocrine cells to β-cells. *Nature* **455**, 627–632 (2008).
33. Noordermeer, D. *et al.* The dynamic architecture of *Hox* gene clusters. *Science* **334**, 222–225 (2011).
34. Ernst, J. *et al.* Mapping and analysis of chromatin state dynamics in nine human cell types. *Nature* **473**, 43–49 (2011).
35. Smith, S.B. *et al.* Rfx6 directs islet formation and insulin production in mice and humans. *Nature* **463**, 775–780 (2010).
36. Zaret, K.S. *et al.* Pioneer factors, genetic competence, and inductive signaling: programming liver and pancreas progenitors from the endoderm. *Cold Spring Harb. Symp. Quant. Biol.* **73**, 119–126 (2008).
37. Scott, R.A. *et al.* Large-scale association analyses identify new loci influencing glycemic traits and provide insight into the underlying biological pathways. *Nat. Genet.* **44**, 991–1005 (2012).
38. Morris, A.P. *et al.* Large-scale association analysis provides insights into the genetic architecture and pathophysiology of type 2 diabetes. *Nat. Genet.* **44**, 981–990 (2012).
39. Hindorff, L.A. *et al. A Catalog of Published Genome-Wide Association Studies* http://www.genome.gov/26525384.
40. El-Assaad, W., *et al.* Saturated fatty acids synergize with elevated glucose to cause pancreatic β-cell death. *Endocrinology* **144**, 4154–4163 (2003).
41. Shimabukuro, M., Zhou, Y.T., Levi, M. & Unger, R.H. Fatty acid–induced β cell apoptosis: a link between obesity and diabetes. *Proc. Natl. Acad. Sci. USA* **95**, 2498–2502 (1998).
42. Hu, L. *et al.* Regulation of lipolytic activity by long-chain acyl-coenzyme A in islets and adipocytes. *Am. J. Physiol. Endocrinol. Metab.* **289**, E1085–E1092 (2005).
43. Helgason, A. *et al.* Refining the impact of *TCF7L2* gene variants on type 2 diabetes and adaptive evolution. *Nat. Genet.* **39**, 218–225 (2007).
44. Cho, Y.S. *et al.* Meta-analysis of genome-wide association studies identifies eight new loci for type 2 diabetes in east Asians. *Nat. Genet.* **44**, 67–72 (2012).
45. Zaret, K.S. & Grompe, M. Generation and regeneration of cells of the liver and pancreas. *Science* **322**, 1490–1494 (2008).
46. Whyte, W.A. *et al.* Master transcription factors and mediator establish super-enhancers at key cell identity genes. *Cell* **153**, 307–319 (2013).
47. Lovén, J. *et al.* Selective inhibition of tumor oncogenes by disruption of super-enhancers. *Cell* **153**, 320–334 (2013).
48. Maurano, M.T. *et al.* Systematic localization of common disease-associated variation in regulatory DNA. *Science* **337**, 1190–1195 (2012).
49. Trynka, G. *et al.* Chromatin marks identify critical cell types for fine mapping complex trait variants. *Nat. Genet.* **45**, 124–130 (2013).
50. Heinz, S. *et al.* Simple combinations of lineage-determining transcription factors prime *cis*-regulatory elements required for macrophage and B cell identities. *Mol. Cell* **38**, 576–589 (2010).
51. Seo, S., Lim, J.W., Yellajoshyula, D., Chang, L.W. & Kroll, K.L. Neurogenin and NeuroD direct transcriptional targets and their regulatory enhancers. *EMBO J.* **26**, 5093–5108 (2007).

© 2014 Nature America, Inc. All rights reserved.

ONLINE METHODS

Human islets. Islets were isolated from donors without a history of glucose intolerance[52], shipped in culture medium and then cultured for 3 d before analysis. β cells were purified by FACS as described[53]. Islet samples were selected from a set of 120 islet samples with ascertainment of minimal exocrine contamination, using dithizone staining and quantitative RT-PCR of lineage-specific markers (**Supplementary Table 7**). High purity was also ascertained by strong H3K4me3 and H3K27ac enrichment at key β cell–specific genes, including *INS*, which shows no enrichment of histone marks in publically available human islet regulatory maps (**Supplementary Fig. 2i**)[11,12,14].

Ethics. Human pancreata were harvested from brain-dead organ donors after obtaining informed consent from family members. Islet isolation centers had permission to use islets for scientific research if they were insufficient for clinical islet transplantation, in accordance with national laws and institutional ethical requirements. Ethical approval for the project was given by the Clinical Research Ethics Committee of Hospital Clinic de Barcelona.

FAIRE and ChIP. Human islets processed for FAIRE and ChIP as described[5,54,55] using 1–1.5 μg of antibodies against H3K4me1 (ab-8895, Abcam), H3K4me3 (Upstate, 05-745), H3K27ac (ab4729, Abcam), CTCF (07-729, Millipore), H2A.Z (ab4174, Abcam; 09-862, Millipore), Pdx1 (AB2027, BCBC), FOXA2 (sc-6554, Santa Cruz Biotechnology), Nkx2.2 (HPA003468, Sigma), MafB (HPA005653, Sigma) and Nkx6.1 (HPA036774, Sigma). Expected cell-specific and subcellular patterns of formalin-fixed epitopes were ascertained by protein blotting on human islet extracts and dual immunofluorescence of human pancreas (**Supplementary Fig. 1b**,**c**).

RNA sequencing. PolyA$^+$ RNA-seq data from three highly purified human islet samples and two FACS-purified human β cells have been described[29]. Non-pancreatic RNA-seq reads were from the Illumina human BodyMap2 project. We processed ~2.5 billion 50-bp paired-end reads from 14 human tissues or cell types (adipose, adrenal gland, breast, colon, heart, kidney, liver, lung, lymphatic node, muscle, ovary, prostate, thyroid gland and white blood cell) identically as for the islet data.

We defined (i) 1,000 islet-specific, non-redundant RefSeq transcripts with the highest median expression ratio in islets versus non-islet samples that were expressed in islets at >10 RPKM, (ii) 1,000 ubiquitously expressed transcripts with median islet versus non-islet expression ratios closest to 1 that had expression in islets of >10 RPKM and (iii)1,000 inactive transcripts with the lowest median RPKM expression in islets.

ChIP-seq analysis. Sequencing was performed on an Illumina HiSeq or Genome Analyzer IIx platform. Quality-filtered 36- to 51-bp single-end reads were aligned to the human genome (NCBI36/hC18) using Bowtie v0.11.3 (ref. 56), allowing only one mismatch and unique mapping. Clonal reads were removed. To calculate genome coverage densities, reads were extended *in silico* to a final length equal to the fragment size estimation based on MACS[57] and were averaged by the number of aligned reads in each ChIP-seq experiment. Sites enriched for transcription factor binding and H2A.Z were detected with MACS[57] using default parameters. Transcription factor peaks were called at $P < 1 \times 10^{-10}$ and H2A.Z peaks were called at $P < 1 \times 10^{-5}$ over a background model derived by sequencing input DNA (**Supplementary Table 8**). We only retained enrichment sites found in replicate experiments. To calculate FDR for replicated peaks, we balanced the number of input and sample reads with MACS and ascertained that >94% of peaks called at these thresholds showed FDR of <0.01. For FAIRE, we tuned the local background noise to $\lambda_{local} = \max \lambda_{5000}, \lambda_{125000}$ and used a cutoff of $P < 1 \times 10^{-3}$, retaining only enrichment sites found in replicate experiments. We considered peaks to be overlapping if they shared a minimum of one base.

Accessible chromatin classes. Accessible chromatin was defined by merging all FAIRE- and H2A.Z-enriched sites that were found in at least two islet samples. These two established measures of accessible chromatin[5,28] proved to be complementary, as FAIRE enrichment was weak in promoters yet readily identified central regions of enhancers and CTCF-bound sites, whereas H2A.Z enrichment was strongest in promoters and also marked enhancers and CTCF-bound sites (**Fig. 2a** and **Supplementary Fig. 3a**). We generated 6-kb windows centered on each accessible chromatin site and computed the read coverage for each feature (H2A.Z, FAIRE, H3K4me3, H3K4me1, H3K27ac and CTCF binding) over 100-bp bins. Cluster3 (ref. 58) was used for *k*-median clustering with HI32, a single human islet sample for which data were available for all marks. For C1, two clusters that showed nearly identical histone modification enrichment patterns were artificially merged into a single class. Enrichment patterns at clusters (or accessible chromatin classes) were concordant in replicate samples (**Supplementary Fig. 3b**).

Genome scan of enhancer clusters. To define enhancer clusters, we first created 1,000 iterations of randomized C3 sites in the mappable genome of individual chromosomes. We then calculated for each chromosome the 25th percentile of the between-site distances of randomized C3 sites (**Supplementary Fig. 6a**). Next, we defined clusters of islet C3 sites as any group of ≥3 C3 sites in which all adjacent C3 sites were separated by less than the above-mentioned 25th-percentile distance for randomized sites in the same chromosome. The distribution of islet C3 clusters differed from those of clusters generated with randomized C3 sites (**Supplementary Fig. 6b**).

We next created a score to measure the average transcription factor occupancy of islet enhancer (C3) clusters. For each cluster, we summed the number of binding sites for each transcription factor at C3 sites and divided this value by the number of C3 sites in the cluster. Clusters with higher average transcription factor occupancy were associated with genes with enriched transcription in islets and β cells (**Supplementary Fig. 7a**). We defined high-occupancy clusters as those showing scores in the top two quartiles (**Fig. 2h** and **Supplementary Fig. 7a**,**b**).

For comparisons of RNA expression in islets and non-islet samples, we identified genes containing a transcription factor–bound enhancer cluster within 25 kb of their annotated TSS (**Fig. 2h** and **Supplementary Fig. 7b**).

For sequence motif analysis, we examined all 19,624 C3 sites that formed clusters, including C3 sites not bound by the transcription factors we had profiled, as we had only profiled a subset of all islet factors. This set of clustered enhancers was also associated with islet-enriched transcription (**Supplementary Fig. 6c**), and only 46% overlapped enhancer sites from nine non-pancreatic cells[34].

We used the Genomic Regions Enrichment of Annotations Tool (GREAT)[59] with default parameters to study enriched functional categories among genes linked to transcription factor–bound enhancer clusters (**Supplementary Fig. 7c**). We also used GREAT to determine which clusters were associated with a list of 65 genes with established importance for islet-specialized function and identity determined on the basis of a systematic literature search (**Supplementary Fig. 7d** and **Supplementary Table 2**).

Transcription factor occupancy, chromatin state and islet-specific transcription. To compare transcription factor occupancy at different chromatin states in islet-specific and ubiquitous versus islet-inactive genes, we identified sites bound by >1 transcription factor in each chromatin class and calculated the density of such sites ±100 kb from the TSSs of the three groups of genes. This level of transcription factor occupancy was chosen to highlight that even high transcription factor occupancy at promoters is not selective for islet-specific gene expression, although conclusions remained unaltered using different numbers of transcription factors.

To directly examine genes bound by islet-specific transcription factors at different chromatin states, we compared quantile-normalized RNA expression in human islets, β cells and 14 non-pancreatic tissues. Data for non-pancreatic tissues were pooled in **Figure 2g**,**h** and are given for individual tissues in **Supplementary Figures 5b**,**d** and **7b**. We compared the mRNA expression of genes bound by transcription factors at C1 sites but lacking transcription factor–bound C3 chromatin within 25 kb of the TSS versus all other genes or genes containing transcription factor–bound C3 clusters within 25 kb of the TSS versus all other genes.

Enhancer function. Human sequences were cloned upstream of the *hsp70* zebrafish core promoter[60] or the *gata2* promoter[61] linked to a Venus reporter and used to inject >200 eggs from wild-type (AB*) and *Tg*(*ins*:*mCherry*)*jh2* zebrafish in three independent experiments. Reporter expression patterns

© 2014 Nature America, Inc. All rights reserved.

were documented using NIS-Elements software (Nikon). Expression was quantified by counting the number of embryos with Venus-positive cells in different tissues (**Supplementary Table 9**). Islet-specific expression patterns from transient transgenic zebrafish were confirmed in stable transgenic lines (**Supplementary Fig. 8**). Transcription factor–bound C3 sites were also cloned in Gateway-adapted pGL4.23 and cotransfected in triplicate wells with pRL in MIN6 and NIH3T3 cells, and luciferase activity was measured after 48 h. Results were expressed as luciferase/*Renilla* ratios with vectors carrying putative enhancers relative to the ratio with empty pGL4.23 vector (**Fig. 3a**). See **Supplementary Table 10** for oligonucleotides.

Transcription factor functional studies. Human EndoC-β H1 β cells[62] were independently transduced with lentiviral vectors expressing two independent shRNAs that target *MAFB* mRNA or four negative-control, nontargeting shRNAs. Each shRNA was transduced in duplicate. See **Supplementary Table 10** for oligonucleotides. *MAFB* shRNAs led to 64% and 55% inhibition of *MAFB* mRNA. We note that *MAFB* was chosen as a transcription factor target on the basis of our ability to design two efficient inhibitory shRNAs.

Mycoplasma-negative HEK293T cells (ATCC CRL-11268) were transfected with pcDNA3-PDX1, pcDNA-NGN3 or pcDNA3.1-MAFA vector or with a control empty vector using Lipofectamine2000 (Invitrogen). After 72 h, RNA was extracted (TRIzol, Invitrogen) and hybridized to GeneChip Human Genome U133 Plus2.0 arrays. To assess the enrichment of predefined gene sets in transcription factor perturbations, robust multi-array average (RMA)-normalized data were analyzed with GSEA[63] using default parameters over 1,000 permutations to calculate FDR. The same analysis was carried out in parallel for an identical number of arrays that compared two non-targeting control and three non-targeting control shRNAs. Gene sets for GSEA were created with genes whose TSSs were closest to C1–C5 sites bound by either MAFB or PDX1. We thus created the following five non–mutually exclusive sets of genes for each transcription factor: MAFB-bound C1 ($n = 363$), C2 ($n = 511$), C3 ($n = 1{,}891$), C4 ($n = 155$) and C5 ($n = 16$); PDX1-bound C1 ($n = 830$), C2 ($n = 878$), C3 ($n = 2{,}874$), C4 ($n = 247$) and C5 ($n = 65$). Similar results were observed using mutually exclusive gene sets (i.e., only sites bound at C3 showed NES > 1 and FDR $q < 0.05$). We also compared the behavior of two mutually exclusive sets of genes, one associated with orphan C3 sites only and another that was only associated with clusters of C3. Gene set sizes were as follows: MAFB-bound orphan C3 ($n = 577$) and clustered C3 ($n = 657$); PDX1-bound orphan C3 ($n = 1{,}233$) and clustered C3 ($n = 1{,}331$).

4C-seq. Chromosome conformation capture assays were performed as described[64] and were adapted for 4C-seq as described[33,65]. Human islets (~5×10^6 cells) were treated with lysis buffer (10 mM Tris-HCl, pH 8, 10 mM NaCl, 0.3% IGEPAL-CA-630 (Sigma-Aldrich) and 1× protease inhibitor (Complete, Roche)). Nuclear DNA was digested with DpnII (New England Biolabs) and ligated with T4 DNA ligase (Promega). Csp6I endonuclease (Fermentas) was used in a second round of digestion, and DNA was religated. Locus-specific primers containing Illumina adaptors (**Supplementary Table 10**) were designed with Primer3 v. 0.4.0 (ref. 66) using the following gene promoters as viewpoints: *ISL1*, *GNAS*, *C2CD4A*, *C2CD4B*, *TM4SF4*, *TM4SF1*, *PDX1*, *MAFB* and *G6PC2*. Eight loci, with the exception of *TM4SF1*, contained enhancer clusters linked to genes with islet-enriched expression. PCR was performed with the Expand Long-Template PCR System (Roche) for each viewpoint, and reactions were pooled for sequencing. 4C-seq data were analyzed by adapting a previously described procedure[33]. Briefly, 4C-seq reads were sorted, aligned and translated to restriction fragments. A moving average of 30 fragments per window was used to smoothen reads. Next, we calculated for each fragment the Poisson probability of it containing a given number of smoothened reads. To this end, all aligned fragments were randomized 1,000 times in a 2-Mb window centered on the viewpoint and smoothened in the same way. We then defined significant interactions in the 4C-seq experiment as those with a Poisson probability of $< 1 \times 10^{-10}$. We tested two C3 promoter interactions by 3C capture, both of which were confirmed. To assess overlaps with chromatin sites, we computed for all nine loci the overlap of different classes of accessible chromatin with 4C-seq interaction sites located in a 2-Mb window centered on the viewpoint, excluding 40 kb on each side of the viewpoint. To contrast expected and observed overlap (**Fig. 3d**,**e**), we randomized different chromatin sites or transcription factor–bound sites 1,000 times in mappable genomic sequences of the same 2-Mb window and calculated the overlap of 4C-seq enrichment sites with these positions.

Electrophoretic mobility shift assays. EMSAs with mouse MIN6 β-cell nuclear extracts were performed as described[67]. We used 1 µl of antibody to NEUROD1 (sc-1084, Santa Cruz Biotchnology) for supershifts. Findings were confirmed with binding experiments carried out on a separate day. See **Supplementary Table 10** for oligonucleotides.

Motif analysis. *De novo* motif discovery was performed with HOMER[50], using a 500-bp window centered on the FAIRE or H2A.Z peaks of clustered C3 sites. We searched for motifs of 6–20 bp in length and retained 46 non-redundant matrices with $P < 1 \times 10^{-60}$. Motifs were annotated using HOMER[50], TOMTOM[68] and manual comparisons (**Supplementary Table 3**). All possible combinations of 3 motifs from the 46 enriched motifs were computed in clustered islet C3 versus analogous genomic sites (H3K4me1 and H3K27ac enriched not H3K4me3 enriched, referred to as strong enhancers by Ernst *et al.*[34]) in 9 non-pancreatic cell lines (GM12878, HepG2, HSMM, HUVEC, ESC, HMEC, K562, NHEK and NHLF). We limited the motif search window to ±250 bp from the center of the genomic site and computed the $\log_2$ value of ratios between motif combination frequencies in islet versus non-islet sites as a metric of islet enrichment. We focused on combinations enriched at χ^2 $P < 1 \times 10^{-3}$. To ensure that results were not affected by differences in the data types used to define enhancers, we also analyzed enhancers from four non-pancreatic cell lines (GM12878, HepG2, HUVEC and K562) for which data on the same chromatin marks were available, including for FAIRE and H2A.Z enrichment[69]. We identified non-islet strong enhancer sites that overlapped FAIRE or H2A.Z sites and calculated the density of the 100 most islet-enriched motif combinations in enhancers from islet and non-islet cells in a 2-kb window centered on merged FAIRE and/or H2A.Z enrichment sites (**Supplementary Fig. 11c**).

To compute the ability of motif combinations to predict islet gene activity in mice, we used HOMER to scan the mouse genome (mm9) for all instances of the ten most islet-enriched motif combinations. We then identified all possible three-motif combinations spanning <500 bp, extended sequences on both sides of motif combinations to create 500-bp segments and filtered for non-redundant combinations. We linked these motif combinations to nearby genes using GREAT with default parameters[59]. Then, RNA-seq RPKM values for mouse islets[29] and nine non-islet tissues (bone marrow, cerebellum, heart, kidney, liver, lung, embryonic fibroblast, embryonic stem cell and spleen; obtained from the Encyclopedia of DNA Elements (ENCODE)/Ludwig Institute for Cancer Research[70]) were aligned and processed as described for human RNA-seq, and expression values were quantile normalized across all tissues[71] (**Supplementary Fig. 11d**).

We also performed *de novo* motif discovery in all accessible chromatin classes (C1–C5) as described above for genome-wide association variant analyses.

Regulome browser. To facilitate the exploitation of integrated data sets, we created a browser (**Supplementary Fig. 15a**,**b**) that enables data downloads and visualization at desired levels of resolution for islet transcription factor binding, chromatin states, motifs and Meta-Analyses of Glucose and Insulin-related traits Consortium (MAGIC)[37] or DIAGRAM *P* values[38] (see URLs).

Genome-wide association analysis. We identified SNPs with genome-wide association ($P < 5 \times 10^{-8}$) to any trait in European individuals present in the GWAS catalog. Index SNPs were pruned (CEU $r^2 > 0.2$) so that each independent locus was only represented by one index SNP. Each index SNP was then used to identify variants in 1000 Genomes Project pilot 1 data with CEU $r^2 > 0.8$. Thus, an associated locus consisted of an index SNP and the set of 1000 Genomes Project SNPs in high LD with it in the CEU population. We then created a background set containing all qualifying loci binned on the basis of the number of total variants (index SNP + high-LD SNPs) in the locus.

For each C3 site, 200-bp regions directly flanking the left and right ends were obtained. For each trait listed in **Figure 5a**, we calculated the number

© 2014 Nature America, Inc. All rights reserved.

doi:10.1038/ng.2870

of loci in the background set associated with that trait containing a variant overlapping a site for each class (C1–C5) as well as clustered C3, orphan C3 and C3 flanking sites. We then permuted the set of loci by drawing from matching bins in the background set and recalculated the number of loci with a variant overlapping a site for that class. The significance of each overlap was calculated as the number of permuted sets with greater overlap than the observed set divided by the number of permutations. We ran 10,000 permutations except for fasting glycemia, for which we ran 100,000 permutations.

We then identified a comprehensive set of T2D and fasting glycemia index SNPs identified in any population. For T2D, we used all genome-wide significant lead SNPs reported in Morris *et al.*[38] and further identified lead SNPs for loci not genome-wide significant in that study but reported as such by previous studies. For fasting glycemia, we used all genome-wide significant lead SNPs reported in Scott *et al.*[37] for both fasting glycemia levels and fasting glycemia levels adjusted for body mass index loci (the former was retained when different lead SNPs at the same locus were reported).

For all lead SNPs, we identified variants in high LD ($r^2 > 0.8$) in 1000 Genomes Project phase I data for CEU samples (European loci), CHB and JPT samples (Asian loci) or YRI samples (African loci; Yoruba from Ibadan, Nigeria).

HapMap enrichment analysis. We obtained *P* values for all HapMap variants from DIAGRAM[38] and MAGIC[37] studies for T2D and fasting glycemia, respectively. Variants were pruned using LD from HapMap CEU samples as follows: first, all variants with $r^2 > 0.2$ with a lead SNP for a trait in Europeans were removed; second, remaining variants were sorted by *P* value, and variants were included if not having $r^2 > 0.2$ with a more significant variant. Both sets of variants (all HapMap and LD-pruned HapMap variants) were binned on the basis of CEU minor allele frequency and distance to the closest GENCODEv12 TSS. For analyses excluding known associated loci, variants in the 500-kb region surrounding each European lead SNP were removed.

A series of trait association *P*-value thresholds were set, and, at each threshold, the number of variants overlapping an islet chromatin class (C1–C5, clustered C3 and orphan C3) was counted. For each variant, a matching variant was then selected at random from the same bin, and the number of matched variants with a *P* value below the same threshold was counted. Fold enrichment was calculated for each threshold by dividing the observed counts by the matched counts averaged over 1,000 permutations.

To evaluate the significance of these enrichments, we focused on variants overlapping each chromatin class that attained $P < 0.001$ in T2D or fasting glycemia meta-analyses. We obtained counts of matched variants significant at the same trait *P*-value threshold across all permutations. As the matched variant counts for these replicates were normally distributed, we calculated a *z* score for the observed islet class counts using the mean and s.d. from the distribution of permuted counts. Reported *P* values were then obtained from *z* scores using a one-sided test.

52. Bucher, P. *et al.* Assessment of a novel two-component enzyme preparation for human islet isolation and transplantation. *Transplantation* **79**, 91–97 (2005).
53. McCulloch, L.J. *et al. GLUT2* (*SLC2A2*) is not the principal glucose transporter in human pancreatic β cells: implications for understanding genetic association signals at this locus. *Mol. Genet. Metab.* **104**, 648–653 (2011).
54. Servitja, J.M. *et al.* Hnf1α (MODY3) controls tissue-specific transcriptional programs and exerts opposed effects on cell growth in pancreatic islets and liver. *Mol. Cell. Biol.* **29**, 2945–2959 (2009).
55. van Arensbergen, J. *et al.* Derepression of Polycomb targets during pancreatic organogenesis allows insulin-producing β-cells to adopt a neural gene activity program. *Genome Res.* **20**, 722–732 (2010).
56. Langmead, B., Trapnell, C., Pop, M. & Salzberg, S.L. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol.* **10**, R25 (2009).
57. Zhang, Y. *et al.* Model-based analysis of ChIP-Seq (MACS). *Genome Biol.* **9**, R137 (2008).
58. de Hoon, M.J., Imoto, S., Nolan, J. & Miyano, S. Open source clustering software. *Bioinformatics* **20**, 1453–1454 (2004).
59. McLean, C.Y. *et al.* GREAT improves functional interpretation of *cis*-regulatory regions. *Nat. Biotechnol.* **28**, 495–501 (2010).
60. Blechinger, S.R. *et al.* The heat-inducible zebrafish *hsp70* gene is expressed during normal lens development under non-stress conditions. *Mech. Dev.* **112**, 213–215 (2002).
61. Meng, A., Tang, H., Ong, B.A., Farrell, M.J. & Lin, S. Promoter analysis in living zebrafish embryos identifies a *cis*-acting motif required for neuronal expression of GATA-2. *Proc. Natl. Acad. Sci. USA* **94**, 6267–6272 (1997).
62. Ravassard, P. *et al.* A genetically engineered human pancreatic β cell line exhibiting glucose-inducible insulin secretion. *J. Clin. Invest.* **121**, 3589–3597 (2011).
63. Subramanian, A. *et al.* Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl. Acad. Sci. USA* **102**, 15545–15550 (2005).
64. Tena, J.J. *et al.* An evolutionarily conserved three-dimensional structure in the vertebrate Irx clusters facilitates enhancer sharing and coregulation. *Nat. Commun.* **2**, 310 (2011).
65. Splinter, E., de Wit, E., van de Werken, H.J., Klous, P. & de Laat, W. Determining long-range chromatin interactions for selected genomic sites using 4C-seq technology: from fixation to computation. *Methods* **58**, 221–230 (2012).
66. Rozen, S. & Skaletsky, H. Primer3 on the WWW for general users and for biologist programmers. *Methods Mol. Biol.* **132**, 365–386 (2000).
67. Boj, S.F., Parrizas, M., Maestro, M.A. & Ferrer, J. A transcription factor regulatory circuit in differentiated pancreatic cells. *Proc. Natl. Acad. Sci. USA* **98**, 14481–14486 (2001).
68. Gupta, S., Stamatoyannopoulos, J.A., Bailey, T.L. & Noble, W.S. Quantifying similarity between motifs. *Genome Biol.* **8**, R24 (2007).
69. Rosenbloom, K.R. *et al.* ENCODE data in the UCSC Genome Browser: year 5 update. *Nucleic Acids Res.* **41**, D56–D63 (2013).
70. ENCODE Project Consortium. A user's guide to the encyclopedia of DNA elements (ENCODE). *PLoS Biol.* **9**, e1001046 (2011).
71. Hansen, K.D., Irizarry, R.A. & Wu, Z. Removing technical variability in RNA-seq data using conditional quantile normalization. *Biostatistics* **13**, 204–216 (2012).

© 2014 Nature America, Inc. All rights reserved.

Cell Metabolism
Article

Human β Cell Transcriptome Analysis Uncovers lncRNAs That Are Tissue-Specific, Dynamically Regulated, and Abnormally Expressed in Type 2 Diabetes

Ignasi Morán,[1,19] İldem Akerman,[1,19] Martijn van de Bunt,[3,4] Ruiyu Xie,[5] Marion Benazra,[6] Takao Nammo,[1,2,7] Luis Arnes,[8] Nikolina Nakić,[1,2] Javier García-Hurtado,[1,2] Santiago Rodríguez-Seguí,[1,2] Lorenzo Pasquali,[1,2] Claire Sauty-Colace,[6] Anthony Beucher,[1,2] Raphael Scharfmann,[9] Joris van Arensbergen,[1,2] Paul R. Johnson,[4,10,16] Andrew Berry,[11] Clarence Lee,[12] Timothy Harkins,[12] Valery Gmyr,[13] François Pattou,[13] Julie Kerr-Conte,[13] Lorenzo Piemonti,[14] Thierry Berney,[15] Neil Hanley,[11] Anna L. Gloyn,[4,16] Lori Sussel,[8] Linda Langman,[17] Kenneth L. Brayman,[17] Maike Sander,[5] Mark I. McCarthy,[3,4,14] Philippe Ravassard,[6] and Jorge Ferrer[1,2,18,*]

[1]Genomic Programming of Beta Cells Laboratory, Institut d'Investigacions Biomèdiques August Pi I Sunyer (IDIBAPS), 08036 Barcelona, Spain
[2]CIBER de Diabetes y Enfermedades Metabólicas Asociadas (CIBERDEM), 08036 Barcelona, Spain
[3]The Wellcome Trust Centre for Human Genetics, University of Oxford, OX1 2JD Oxford, UK
[4]Oxford Centre for Diabetes, Endocrinology and Metabolism, Churchill Hospital, OX3 7LJ Oxford, UK
[5]Department of Cellular and Molecular Medicine, University of California at San Diego, La Jolla, CA 92093, USA
[6]Centre de recherche de l'institut du cerveau et de la moelle, Biotechnology & Biotherapy team, CNRS UMR7225; INSERM U975; University Pierre et Marie Curie, 75005 Paris, France
[7]Department of Metabolic Disorders, Diabetes Research Center, Research Institute, National Center for Global Health and Medicine, Shinjuku-ku, 162-8655 Tokyo, Japan
[8]Department of Genetics and Development, Russ Berrie Medical Pavilion, Columbia University, New York, NY 10032, USA
[9]Institut National de la Santé et de la Recherche Médicale (INSERM) U845, Research Center Growth and Signalling, Paris Descartes University, Sorbonne Paris Cité, Necker Hospital, 75001 Paris, France
[10]Oxford Islet Transplant Programme, Nuffield Department of Surgical Sciences, John Radcliffe Hospital, OX3 9DU Oxford, UK
[11]Developmental Biomedicine Research Group, School of Biomedicine, Manchester Academic Health Sciences Centre, University of Manchester and Central Manchester University Hospitals NHS Foundation Trust, M13 9WL Manchester, UK
[12]Genome Sequencing Collaborations Group, Life Technologies, Beverly, MA 01915, USA
[13]University of Lille Nord de France, INSERM U859 Biotherapies of Diabete, Lille, France
[14]Diabetes research institute (HSR-DRI), San Raffaele Scientific Institute, 20132 Milano, Italy
[15]Cell Isolation and Transplantation Center, 1121 Geneva 4, Switzerland
[16]Oxford NIHR Biomedical Research Centre, Churchill Hospital, OX3 7LJ Oxford, UK
[17]Division of Transplantation, Department of Surgery, Center for Cellular Therapy and Biotherapeutics, University of Virginia, Charlottesville, VA 22904, USA
[18]Department of Endocrinology and Nutrition, Hospital Clínic de Barcelona, 170 08036 Barcelona, Spain
[19]These authors contributed equally to this work
*Correspondence: jferrer@clinic.ub.es
http://dx.doi.org/10.1016/j.cmet.2012.08.010

SUMMARY

A significant portion of the genome is transcribed as long noncoding RNAs (lncRNAs), several of which are known to control gene expression. The repertoire and regulation of lncRNAs in disease-relevant tissues, however, has not been systematically explored. We report a comprehensive strand-specific transcriptome map of human pancreatic islets and β cells, and uncover >1100 intergenic and antisense islet-cell lncRNA genes. We find islet lncRNAs that are dynamically regulated and show that they are an integral component of the β cell differentiation and maturation program. We sequenced the mouse islet transcriptome and identify lncRNA orthologs that are regulated like their human counterparts. Depletion of *HI-LNC25,* a β cell-specific lncRNA, downregulated *GLIS3* mRNA, thus exemplifying a gene regulatory function of islet lncRNAs. Finally, selected islet lncRNAs were dysregulated in type 2 diabetes or mapped to genetic loci underlying diabetes susceptibility. These findings reveal a new class of islet-cell genes relevant to β cell programming and diabetes pathophysiology.

INTRODUCTION

During recent years, it has become apparent that the genomes of species as diverse as zebrafish, mice, and humans transcribe thousands of RNAs that do not encode for proteins (Bertone et al., 2004; Birney et al., 2007; Carninci et al., 2005; Guttman et al., 2009; Ulitsky et al., 2011). A subset of noncoding transcripts are larger than 200 nucleotides and are known as long noncoding RNAs (lncRNAs) (Mattick and Makunin, 2006). The function of most lncRNAs remains unknown. However, several dozen lncRNAs are known to exert nonredundant roles in processes such as X inactivation, imprinting, splicing, transcriptional regulation, pluripotency, cancer, cell cycle, or survival

(Gupta et al., 2010; Guttman et al., 2011; Hu et al., 2011; Penny et al., 1996; Rinn et al., 2007; Sleutels et al., 2002). In one example, a lncRNA has been shown to promote reprogramming of pluripotent cells from somatic cells (Loewer et al., 2010). Available evidence thus indicates that lncRNAs represent a still poorly understood layer of gene regulation.

Numerous mammalian lncRNAs are expressed in a cell-type specific manner (Cabili et al., 2011; Mercer et al., 2008). Together with knowledge that several such transcripts are functional, this raises the intriguing possibility that lncRNAs could be previously unsuspected mediators of lineage-specific differentiation or specialized cellular functions. Defects in lncRNAs could thus underlie human disease, and cell-specific regulatory lncRNAs might provide therapeutic targets. This warrants the need to explore the repertoires of lncRNAs of disease-relevant cell types and tissues.

Pancreatic islets of Langerhans are an excellent model of a specialized tissue that is closely linked to human disease. Islets comprise insulin-secreting β cells and other polypeptide hormone-producing cells, including glucagon-secreting α-cells. Islet-cell dysfunction is central to the pathophysiology of type 2 diabetes (T2D), the most prevalent form of diabetes (Bell and Polonsky, 2001). Recent genome-wide association studies for T2D and related traits have revealed >50 susceptibility loci, most of which are not known to carry variants that alter protein-coding sequences (McCarthy, 2010). A common hypothesis is that such variants impact regulatory elements of protein-coding genes, although they could equally affect other non-protein-coding elements such as lncRNAs.

In type 1 diabetes, β cells are destroyed by autoimmune mechanisms, and consequently several experimental approaches are being developed to replace destroyed cells (Halban et al., 2001). One approach is based on the recent discovery of β cell transcription factors, some of which have been misexpressed in somatic cells to create insulin-expressing cells (Collombat et al., 2009; Zhou et al., 2008). Another approach is to derive β cells from pluripotent cells (Kroon et al., 2008). However, existing strategies have not yet succeeded in generating fully functional therapeutic β cells in vitro. Clearly, the identification of novel regulators of β cell differentiation and maturation remains a major challenge, and islet-specific regulatory transcripts are logical targets for this effort. Despite the potential implications for human diabetes, information on islet-cell lncRNAs is lacking.

In this study, we integrated sequence-based transcriptome and chromatin maps of human islets and β cells to define 1128 islet lncRNA genes. We show that these lncRNAs are an integral component of the dynamic β cell-specific differentiation program, suggesting a role as biomarkers and potential regulators for programming efforts. We extend existing knowledge on lncRNAs by disclosing orthologous mouse transcripts that are regulated in an evolutionarily conserved manner. We focused on one islet-specific lncRNA and show that it impacts the expression of a regulatory target. Finally, we show dysregulation of islet lncRNAs in T2D and map selected lncRNAs to human diabetes genetic susceptibility loci. Collectively, these studies describe a new class of islet genes and provide diverse lines of evidence to suggest that islet lncRNAs may impact diabetes pathophysiology and efforts to program therapeutic β cells.

RESULTS

High-Resolution Maps of Active Human Pancreatic Islet Cell Genes

To identify human β cell lncRNAs we integrated transcriptional and chromatin maps of purified human islet cells (Figure 1A). We generated directional and nondirectional cDNA libraries from PolyA+ and Ribo− RNA fractions of six human islets and two FACS-purified β cell samples, and we collected >460 million uniquely mapped paired-end sequence reads (Table S1). To increase our ability to define the structure of novel genes, we mapped epigenetic gene landmarks in three human islet samples, including H3K4me3, which is enriched in active promoter regions (Table S1). These data sets were then integrated to build strand-specific models of active human islet-cell genes with their predominant splicing patterns (Figure 1A).

As a measure of quality, we first examined annotated genes and recovered ~70% as expressed above a threshold level of 0.5 reads per million mapped reads (RPKM) in islet cells. Despite the fact that the samples originated from diverse human islet donors (Table S2), both mRNA and H3K4me3 levels showed a high correlation between the different human islet samples ($r^2 > 0.94$; Figures S1A and S1B). Consistent with the high islet purity, the most abundant H3K4me3-positive transcripts included known islet-cell and βcell genes (Figure 1B, Table S3). Remarkably, ~20% of exonic reads in islets and ~45% in β cells originated from the Proinsulin mRNA (Figure 1B). Moreover, known islet-specific genes displayed coherent chromatin enrichment patterns (see, for example, Figures 1C and S1C).

Genomic inventories of human islets are unavoidably confounded by signals from contaminant exocrine cells. Our approach, however, allowed us to define lncRNAs that are truly expressed in β cells and other islet cell types. Transcripts from few contaminating acinar cells were efficiently excluded because they lacked H3K4me3 enrichment in islet preparations (Figures S1D and S1H, Table S3). Accordingly, <1% of genes that we defined as active in islet cells were contaminant acinar transcripts, on the basis of a relative enrichment in acinar versus islet-cell RNA. This initial analysis thus validated our strategy to map bona-fide active genes in human islets and β cells.

Human Islet Cells Transcribe Intergenic and Antisense lncRNAs

Over 19% of the transcribed genome in human islets mapped outside of annotated protein-coding genes (Figure S1E). To define discrete islet-cell lncRNA genes, we selected 1128 transcripts with the following properties: (1) length >200 bp; (2) H3K4me3 enrichment in a coherent location relative to the transcribed strand; (3) expression >0.5 RPKM in all five islet and β cell PolyA+ samples; (4) no splicing or overlap with any coding gene present in RefSeq, UCSC, or Ensembl annotations; and (5) low protein-coding potential. Of these, 761 were "antisense" (AS) lncRNAs, located <1 kb from an annotated gene but in a divergent orientation (Figure 1D). Thirty-two were <1 kb from an annotated gene but in a convergent orientation. Another set of 335 were "intergenic" (IG) lncRNAs, located >1 kb from any coding gene. Finally, we identified 55 annotated lncRNAs that were located within the boundaries of coding genes and termed these "overlapping antisense" lncRNAs (Figure 1D).

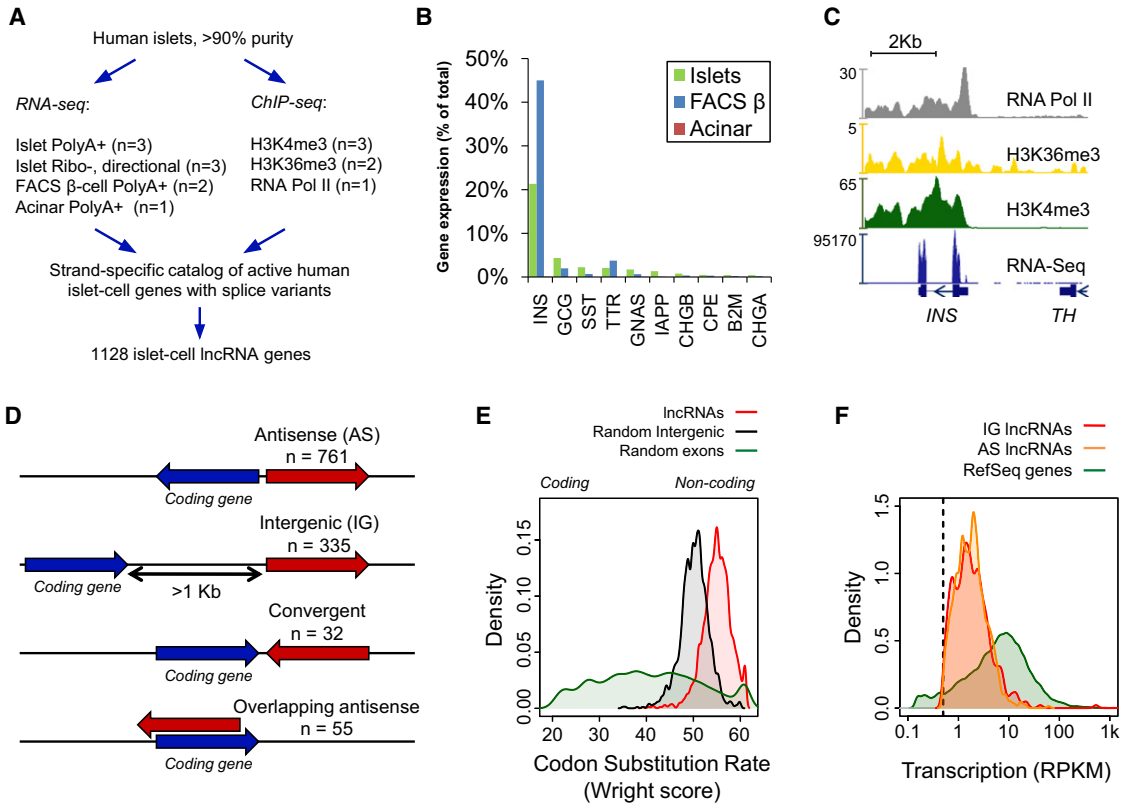


Figure 1. Systematic Identification of Human Islet and β cell lncRNAs

(A) Outline of the analysis pipeline.

(B) RNA-seq expression levels in the three pancreatic cell fractions for the ten most-expressed H3K4me3-positive genes in human islets.

(C) Chromatin and RNA-seq landscape of the *INS* locus.

(D) Definitions and counts of different lncRNA classes.

(E) Kernel density plot of the codon substitution rate of islet lncRNAs (red), intergenic control regions (black) and randomly selected exons of protein-coding genes (green).

(F) Kernel density plot of the transcriptional abundance of all RefSeq genes (green), intergenic (red), and antisense (orange) lncRNAs. The dashed line marks the 0.5 RPKM expression threshold used for defining lncRNAs. See also Figure S1 and Tables S1–S5.

For subsequent analysis we merged convergent and intergenic lncRNAs (a total of 367) as neither was closely associated with a promoter of a protein-coding gene. Comparison with acinar RNA-seq showed that only one lncRNA showed >3-fold acinar enrichment. Furthermore, qPCR analysis confirmed expression of 31/31 lncRNAs in human islets, and 26/31 in the human β cell-line EndoC-β H1 (Ravassard et al., 2011) (data not shown). These transcripts are therefore bona-fide islet-cell and in most cases β cell lncRNAs. Their genomic location and exon annotations are provided in Table S4.

Further analysis of the genomic properties of the 1128 islet lncRNAs showed that their overall protein-coding parameters resembled those of randomly chosen intergenic regions, confirming that most are likely to be truly noncoding transcripts (Figures 1E and S1F). Antisense and intergenic lncRNAs had a similar average length as RefSeq genes, yet showed ~10-fold lower expression (Figure 1F). Despite the low abundance of lncRNAs, the top quartile were expressed at levels comparable to mRNAs encoding transcriptional regulators linked to human diabetes (*HNF1A*, *HNF4A*, and *TCF7L2*; 6.4, 7.1 and 5.7 RPKM in human islets, respectively). Compared with RefSeq transcripts, islet lncRNA genes displayed similar RNA Pol II enrichment but were less often H3K36me3 enriched, in keeping with less frequent splicing (36% for intergenic lncRNAs, versus 93% in RefSeq genes) (Table S5). Finally, lncRNAs overlapped interspersed repeat regions more often than coding exons, yet less than random intergenic control regions (Figure S1G). In conclusion, these studies uncover a high-confidence set of 1128 human islet-cell genes that have several expected properties for lncRNAs.

Islet lncRNAs Are Highly Tissue Specific

Because many islet lncRNAs have escaped annotation, we reasoned that this could be in part because this class of transcripts is often cell-type specific. We realigned 16 human non-pancreatic RNA-seq data sets (Table S1), and found that 9.4% of RefSeq annotated genes were islet specific (Figure 2A, Table S6; see Experimental Procedures for a definition of the islet-specificity score). By contrast, 55% of intergenic lncRNAs, and 40% of antisense islet lncRNAs were islet specific, a significant enrichment over protein-coding genes ($p < 3 \cdot 10^{-16}$ for both comparisons) (Figure 2B). This set of lncRNAs was similarly

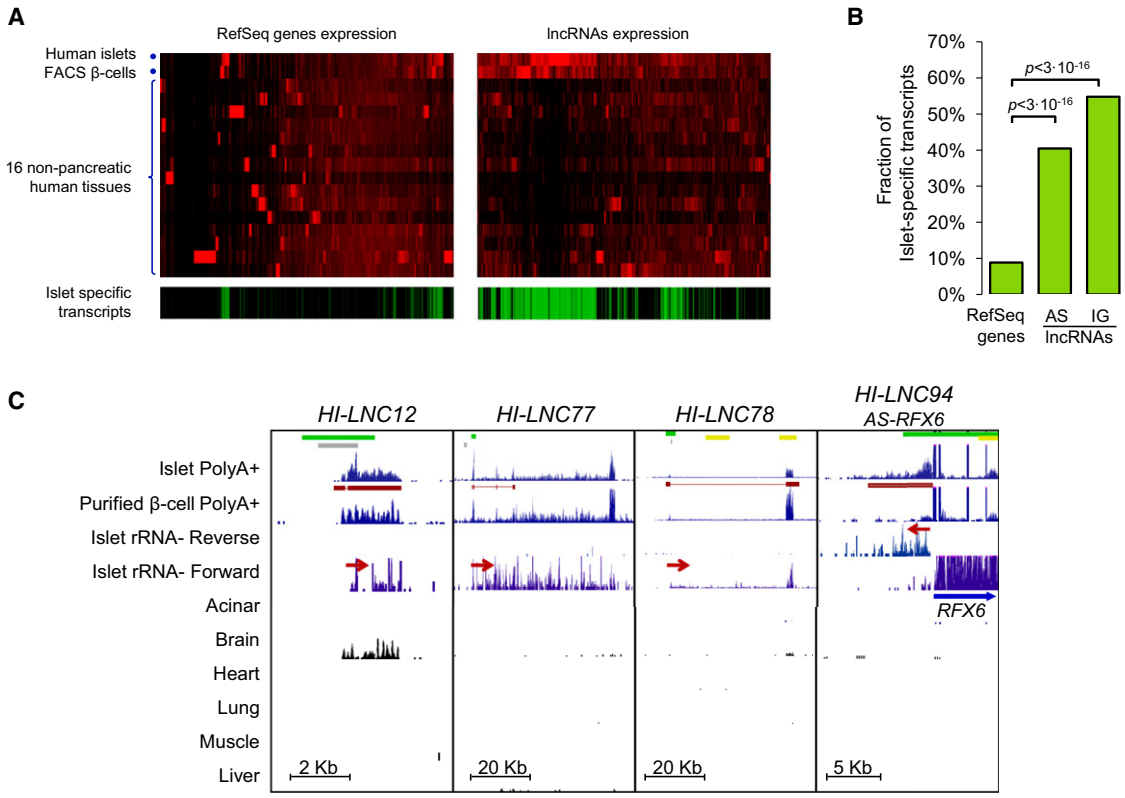


Figure 2. Islet lncRNAs Are Highly Tissue Specific

(A) Hierarchical clustering of expression levels of all RefSeq genes (left) and lncRNAs (right) across human islets, FACS-purified β cells and 16 nonpancreatic human tissues. Bottom panels highlight islet-specific transcripts in green.

(B) Islet specificity of antisense and intergenic lncRNAs compared to all RefSeq genes. *P* values were calculated by Chi-Square test.

(C) Representative examples of islet-enriched lncRNAs. Gene models are depicted in red and transcript orientation with a red arrow. All rows depict RNA-seq except the top row, which shows consistent peaks for H3K4me3 (green), H3K36 (yellow), and RNA Polymerase II (gray). The vertical axis is scaled at the same expression level for all tissues. Only 5/16 representative nonpancreatic tissues are shown for simplicity. See also Figure S2 and Table S6.

enriched in purified β cells (Figure 2A and not shown). Many examples of lncRNAs were thus found to be unique or highly specific to β cells within the entire panel of 18 tissues (Figure 2C). The islet selectivity was confirmed by qPCR in 12/12 lncRNAs using an independent panel of nine tissues (Figure S2). Thus, human islet-cell lncRNAs are frequently transcribed in a highly cell-type specific manner.

Islet lncRNAs Often Map near Islet-Specific Chromatin Domains and Coding Genes

To explore the biological significance of human islet lncRNAs, we next examined their relationship with nearby protein-coding genes. The abundance of each lncRNA transcript in human islets was not related to that of the neighboring protein-coding mRNA (Figure S3A). However, the expression of lncRNAs and their most proximal protein-coding gene significantly correlated across different tissues ($p < 3 \cdot 10^{-14}$) (Figure 3A). Accordingly, the closest protein-coding genes to islet intergenic and antisense lncRNAs were more often islet-specific than the entire collection of annotated genes ($p < 10^{-10}$), up to >3-fold when considering only islet-specific lncRNAs ($p < 2 \cdot 10^{-11}$) (Figures 3B and 3C). Notably, islet lncRNAs were preferentially located in genomic regions where the closest protein-coding gene has been linked to β cell function, development, and transcription (Figure 3C). Examples of known islet protein-coding genes paired with intergenic lncRNAs include *MAFB*, *FOXA2, PCSK1*, and *ISL1* (Figure S3B and data not shown). We compiled a list of 20 genes encoding known islet-enriched transcription factors, and found that remarkably 13 (65%), including *HNF1A*, *PDX1, PAX6*, *ISL1*, *INSM1*, *NEUROD1, GATA6, NKX2-2,* and *RFX6,* were associated with antisense lncRNAs (Figure S3B and data not shown). Thus, islet lncRNAs often map to the same genomic region as islet-enriched protein-coding genes.

It was also apparent that lncRNAs were often located in gene-poor areas. We examined the genomic intergenic spaces where lncRNAs reside and found that they are on average nearly 3-fold larger than the entire set of intergenic intervals in the human genome ($p < 3 \cdot 10^{-12}$) (Figure 3D). The fact that islet lncRNAs were often located in large gene-poor spaces, yet associated to genes with similar cell-type specific expression, suggested the existence of broad cell-specific regulatory domains. Human islets harbor tissue-specific clusters of open chromatin regulatory elements (COREs), often linked to a single islet-enriched protein-coding gene (Gaulton et al., 2010). We thus examined the relationship of islet lncRNAs to islet-specific COREs (see Experimental Procedures). Both intergenic and antisense

lncRNAs, as well as their nearby annotated genes, were more often located near islet-selective COREs than random intergenic regions ($p < 3 \cdot 10^{-9}$) or expressed annotated genes ($p < 0.013$) (Figure 3E). Thus, islet lncRNAs are often associated to broad cell-specific regulatory domains, many of which appear to be shared by protein-coding genes.

HI-LNC25 illustrates many of the above-mentioned features. It is a multiexonic transcript located in a broad ~1.6 Mb space that lacks any protein-coding gene but contains clusters of islet-specific active chromatin (Figure 3F). The most proximal protein-coding gene is *MAFB*, an essential regulator of islet-cell maturation that is abundantly expressed in human β cells (Artner et al., 2007; Dorrell et al., 2011). Interestingly, whereas *MAFB* is expressed in several tissues, *HI-LNC25* shows a much more restricted tissue distribution (Figure 3F). Taken together, these results indicate that islet-selective lncRNAs are frequently associated with cell-specific higher-order chromatin domains that potentially underlie coregulation of noncoding and coding transcript pairs.

Stage-Specific Activation of lncRNAs during β Cell Maturation

To understand the developmental regulation of islet lncRNAs, we first used qPCR to test the expression of β cell selective lncRNAs (ten intergenic, three antisense) in dissected Carnegie stage 17–19 human embryonic pancreas, a progenitor stage that shows scarce signs of cytodifferentiation (Piper et al., 2004). Of the 13 lncRNAs that were examined, all except one were silent or expressed at low levels in pancreatic progenitors and were subsequently active in adult islets (Figure 4A). Thus, islet lncRNAs are not only often islet specific, but they are also linked to the pancreatic endocrine differentiation program.

We next examined the dynamics of islet lncRNAs in a human embryonic stem-cell (hES) differentiation model. Current methods to differentiate β cells from stem cells are limited by difficulties in completing the maturation of β cells in vitro. We thus used a hES cell protocol that involves multiple in vitro differentiation steps, followed by encapsulation of differentiated pancreatic endoderm, which after implantation into mice produces mature endocrine cells over a 140 day in vivo incubation period (Kroon et al., 2008). We profiled the same lncRNAs at each stage of the protocol and discovered that all islet lncRNAs were markedly induced during the in vivo maturation step (Figure 4B). Notably, six lncRNAs were expressed at very low or undetectable levels throughout all in vitro differentiation steps and were only activated during the in vivo maturation step (Figure 4B). Thus, islet-specific lncRNA gene activation is linked to pancreatic endocrine differentiation during embryogenesis and during in vivo differentiation of hES cell-derived pancreatic progenitors. Islet lncRNAs are therefore candidate markers and/or regulators of β cell differentiation and maturation.

Glucose-Dependent Regulation of Islet lncRNAs

Mature β cells exhibit a transcriptional response to increased demand (Bensellam et al., 2009; Schuit et al., 2002). Selected islet protein-coding transcripts are thus moderately induced in response to high glucose under appropriate experimental conditions. To assess if islet lncRNAs are also dynamically regulated, we exposed human islets to 4 mM or 11 mM glucose for 72 hrs. This revealed glucose-dependent changes in protein-coding mRNAs related to insulin secretion (*INS*, *IAPP*, *PCSK1*), whereas the stress-responsive marker *DDIT3* (CHOP) was not induced (Figure 4C). We then tested the set of 13 islet lncRNAs and discovered that *HI-LNC78* and *HI-LNC80,* two intergenic multiexonic lncRNAs that are not located near islet-enriched protein coding genes, were consistently upregulated in a glucose-dependent manner in five individual donors (Figure 4C). Thus, selected islet lncRNAs are dynamically regulated in settings that are relevant for mature islet-cell physiology.

Identification of Conserved β Cell-Specific lncRNAs in Mice

We next assessed the evolutionary conservation of islet lncRNAs. Mammalian sequence conservation scores of islet lncRNAs were markedly lower than protein-coding exons, yet higher than random intergenic regions, in keeping with previous results (Pauli et al., 2012; Ulitsky et al., 2011) (Figure 5A). For nearly 70% of human lncRNAs, we identified orthologous mouse genomic regions of >200 bp, compared to 58% of random human intergenic regions ($p < 2 \cdot 10^{-9}$) (Figure 5B).

The degree of sequence identity that is required for functional conservation of noncoding transcripts is uncertain. To improve our understanding of the evolutionary conservation of islet lncRNAs, we examined if orthologous lncRNAs were transcribed in mouse islets. To this end, we generated 78 million paired-end reads from purified mouse islet PolyA+ RNA directional libraries (Table S7). We found that 47% of the mouse orthologous regions were transcribed in mouse islets, as opposed to <3% random intergenic regions of the same size ($p < 3 \cdot 10^{-16}$) (Figure 5C). This is likely an underestimate of orthologous mouse lncRNAs since qPCR analysis confirmed the expression of 7/7 orthologous mouse lncRNAs, and it also showed that 5/7 orthologous mouse genomic regions that did not surpass our detection threshold in the RNA-seq analysis were nevertheless detected at low levels in mouse islet and/or β cell-line RNA, whereas control intergenic regions were undetectable (Figures 5D and 5E).

Next, we assessed if mouse orthologous lncRNA transcripts are regulated in a similar manner as their human counterparts. Like human lncRNAs, mouse orthologous transcripts were >4-fold more frequently islet-cell-specific than protein-coding genes ($p < 3 \cdot 10^{-16}$) (Figures 5F, 5G, and S4A). Moreover, the analysis of E13.5 embryonic pancreas showed that five out eight lncRNAs were inactive prior to endocrine differentiation, indicating that they are typically regulated in a similar stage-specific fashion as their human orthologs (Figures 5H and S4B). Furthermore, a comparison of islets from neonate and adult mice showed that most lncRNAs were further upregulated during postnatal islet maturation (Figure S4C). Likewise, five out of eight tested mouse islet lncRNAs were induced upon exposure to high-glucose concentrations, including *Mi-Linc80*, whose ortholog is glucose-responsive in human islets (Figure 5I). Several islet lncRNAs were also regulated in vivo in glucose intolerant Leptin-deficient mice (Figure S4D). Thus, genomic mouse regions that are orthologous to human islet lncRNA genes are frequently also transcribed in islets, and like their human counterparts they exhibit a highly dynamic and cell-specific regulation.

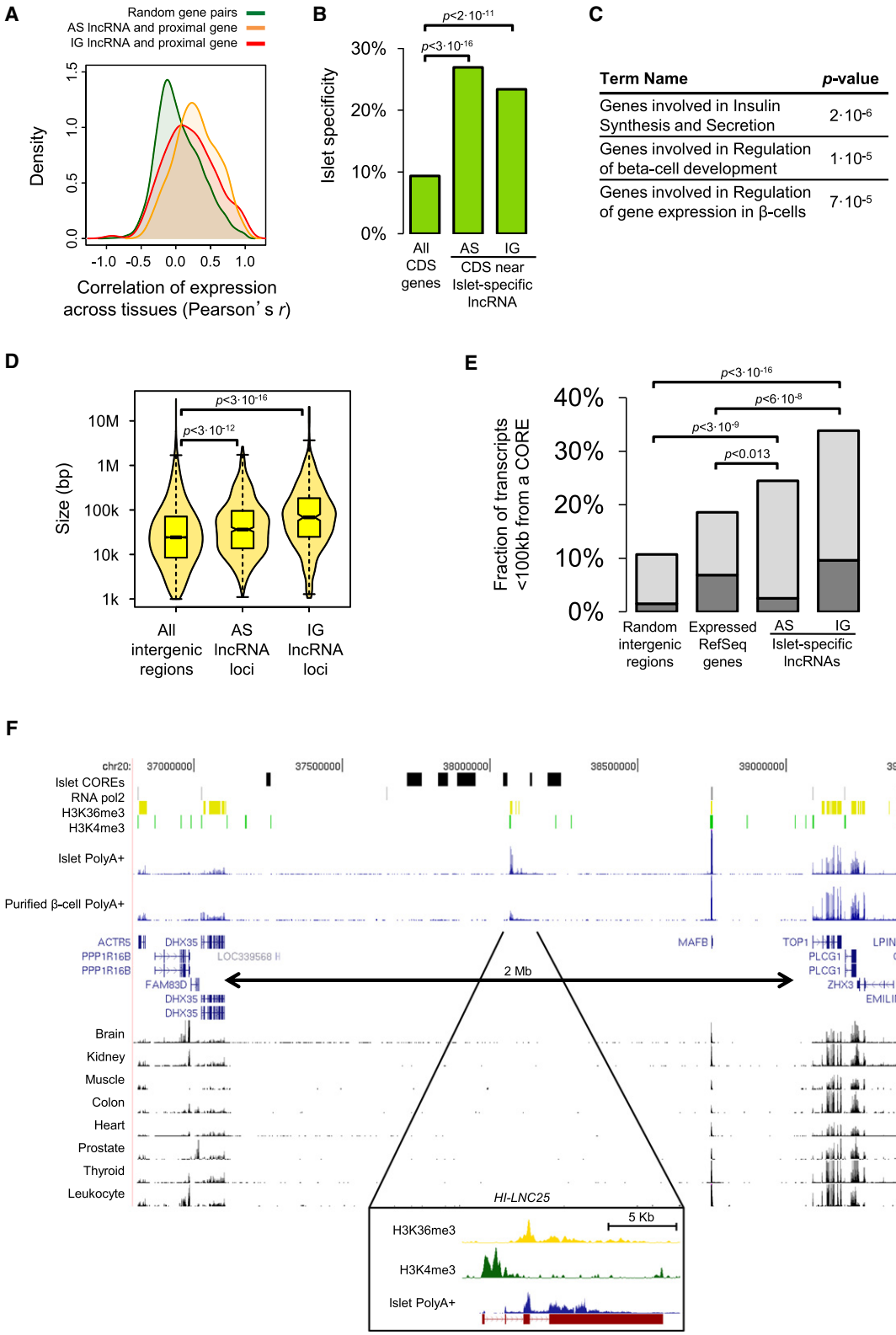


Figure 3. Many Islet lncRNAs Map Near Islet-Specific Chromatin Domains and Regulatory Genes

(A) LncRNA transcript levels correlate with those of their nearest coding gene across tissues. Kernel density plots of correlation of RNA levels for random gene pairs (green), intergenic (IG) lncRNAs paired with their nearest coding genes (red) ($p < 3 \cdot 10^{-14}$ compared to random genes pairs, Mann-Whitney *U* test), and antisense (AS) lncRNAs with their proximal coding genes (orange) ($p < 3 \cdot 10^{-16}$).

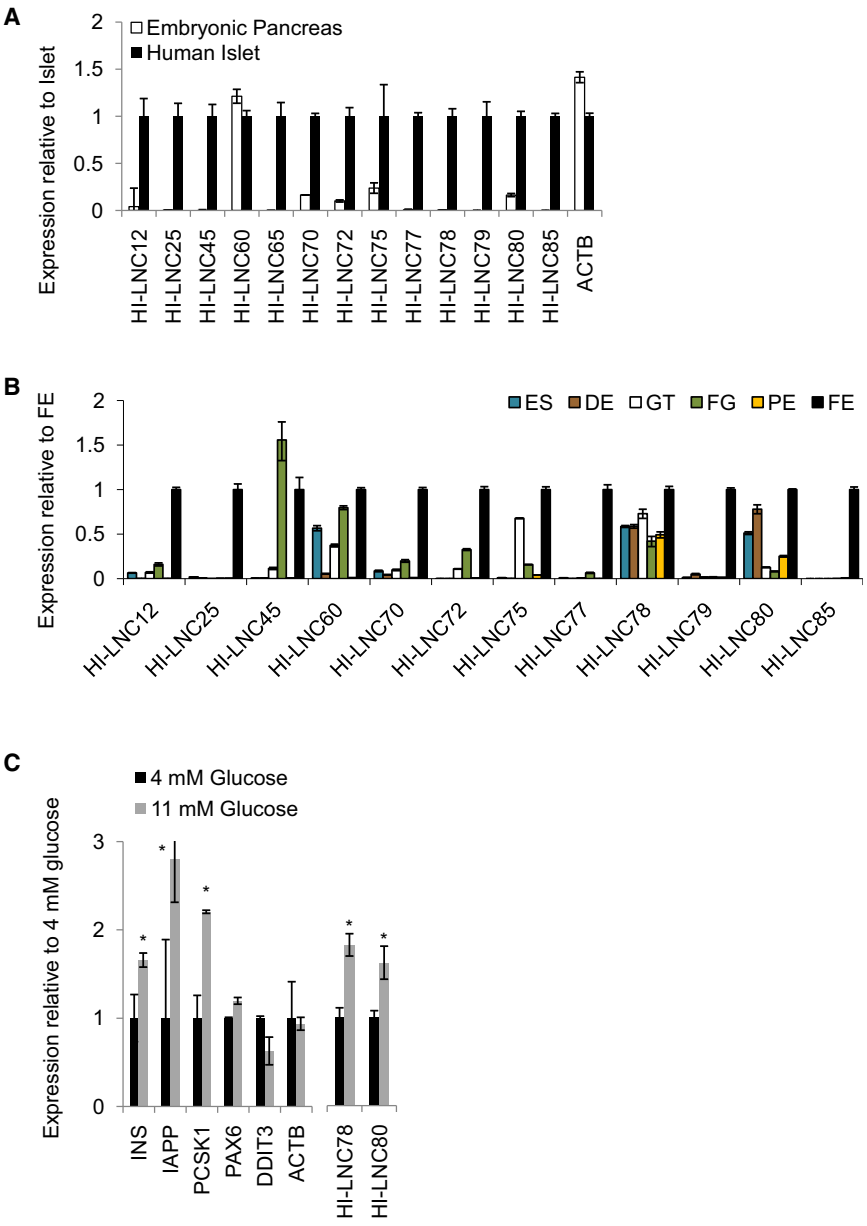


Figure 4. Human lncRNAs Are Dynamically Regulated

(A) Expression of human islet-specific lncRNAs (10 intergenic, 3 antisense) was assayed in mature islets and pooled embryonic pancreas (Carnegie stage 17–19) by qPCR. The results were normalized to *TBP* mRNA and expressed as a fraction of expression in islets. All but one lncRNAs were activated after the pancreatic progenitor stage.

(B) Human islet lncRNAs are induced upon in vivo differentiation of embryonic stem cells. Human embryonic stem cells (ES) were subjected to a differentiation protocol that undergoes stage-specific differentiation to definitive endoderm (DE), primitive gut tube (GT), foregut endoderm (FG), and pancreatic endoderm (PE), respectively. The cells are then transplanted into mice to allow for in vivo maturation of functional endocrine cells (FE). ES, DE, GT, FG, PE, and FE represent days 0, 2, 5, 7, 10, and 150 of the differentiation protocol. *HI-LNC65* was not detected in any of the maturation steps and is not shown.

(C) Expression of a subset of human lncRNAs is regulated by glucose. Human islets were either cultured under low (4 mM) or high (11 mM) glucose concentrations for 72 hr. Panels from left to right illustrate expected glucose responsiveness of *INS*, *IAPP*, and *PCSK1*, and the glucose responsiveness of *HI-LNC78* and *HI-LNC80* in five independent human islet samples. Values are normalized to *TBP* mRNA and expressed as a fraction of the condition denoted in the vertical axis label. Bars represent means ± SEM and asterisks denote $p < 0.001$ (Student's t test).

HI-LNC25 Regulates *GLIS3* mRNA

The evolutionary conserved regulation of islet lncRNAs points to a functional role. LncRNAs have been linked to diverse types of functions, frequently involving direct or indirect regulation of expression of protein-coding genes (Wang and Chang, 2011). To begin to understand islet lncRNA function we focused on *HI-LNC25,* which fulfills several typical features of human islet lncRNAs, as shown in Figure 3F. As an experimental model we used a human β cell line (EndoC-β H1) that exhibits glucose-induced insulin secretion (Ravassard et al., 2011). We transduced human β cells with lentiviral vectors expressing two independent shRNA hairpins that suppressed *HI-LNC25* RNA and examined the expression of a panel of 24 islet mRNAs. This screen identified *GLIS3* mRNA as a potential regulatory target of *HI-LNC25*. *GLIS3* encodes an islet transcription factor, it is mutated in a form of monogenic diabetes, and contains T2D risk variants (Cho et al., 2012; Senée et al., 2006). We therefore

(B) Islet specificity of all coding RefSeq genes (CDS) compared with the nearest coding RefSeq gene to islet-specific antisense and intergenic lncRNAs. *P* values were calculated by Chi-square test.

(C) Top-enriched gene ontology terms for genes closest to islet-specific lncRNAs.

(D) LncRNAs are preferentially located in gene-poor regions. Violin plots of the length distribution of all intergenic spaces in the genome >1 kb, compared with the intergenic spaces where antisense and intergenic lncRNAs reside.

(E) Percentage of islet-specific lncRNAs located near islet-specific clusters of regulatory elements (COREs), compared with random intergenic regions and RefSeq genes. Dark gray areas of the bars represent lncRNAs directly overlapping a CORE element, while light gray areas represent non-overlapping lncRNAs located <100 kb away from COREs.

(F) Birds-eye view of a ~2 Mb region surrounding the islet regulator *MAFB*, centered on *HI-LNC25*. All rows depict RNA-seq (vertical axis fixed at 50 RPKMs), except for the first four rows, which show the islet CORE regions (black), enriched regions for RNA polymerase II (gray), H3K36 (yellow), and H3K4me3 (green). The bottom inset shows a zoomed-in view of the *HI-LNC25* gene. Only 8/16 representative human tissues are shown for simplicity. See also Figure S3.

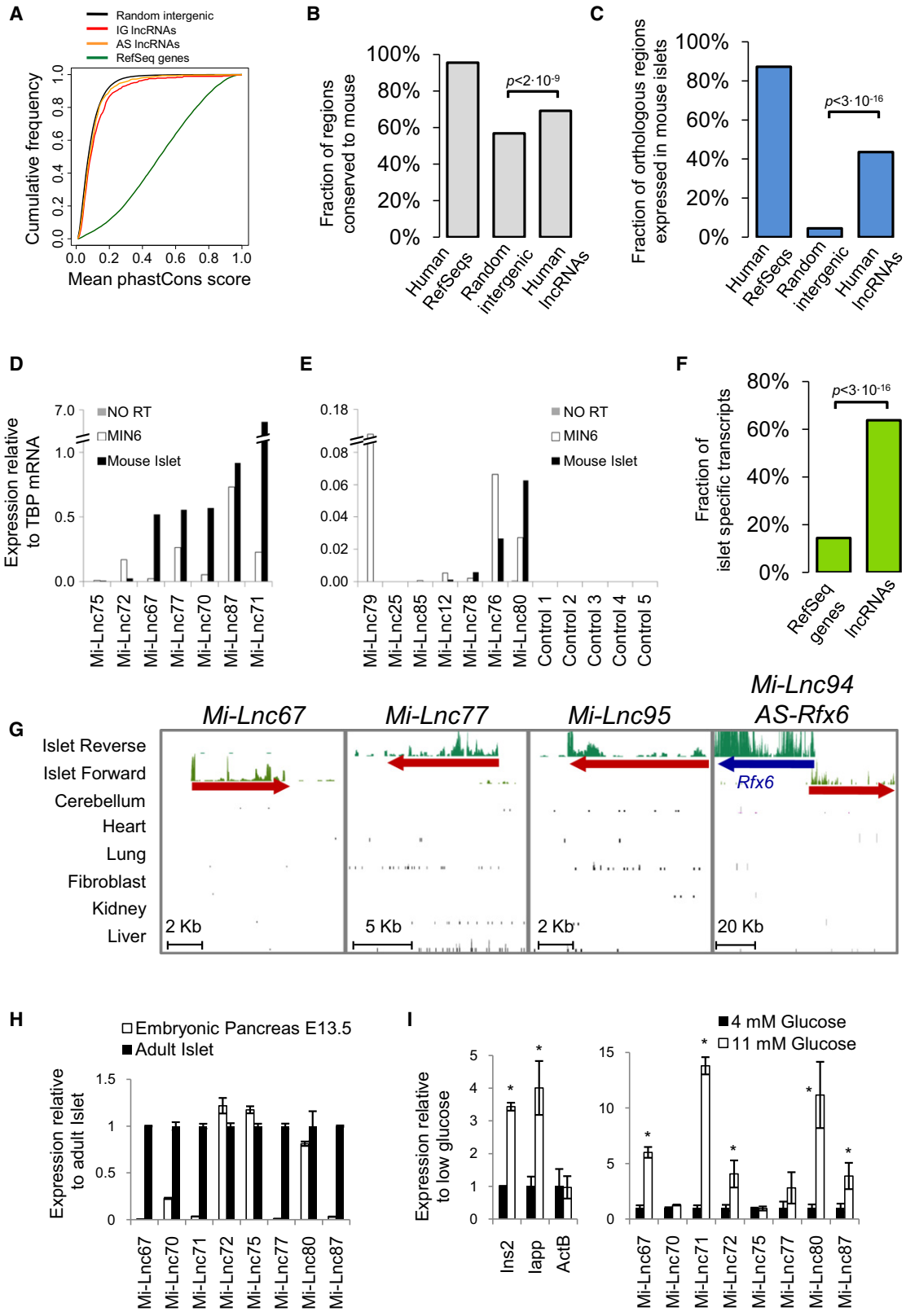


Figure 5. Mouse Islet lncRNA Orthologs Exhibit Conserved Regulation

(A) Mean phastCons scores for intergenic lncRNAs (red), antisense lncRNAs (orange), random intergenic fragments (black), and coding genes (green).

(B) Fraction of human RefSeq exons, random intergenic regions, and lncRNAs with an orthologous mouse genomic region of at least 200 bp.

performed four independent *HI-LNC25* knockdown experiments with two separate shRNA hairpins. Although this depletion of *HI-LNC25* did not cause significant changes in glucose-stimulated insulin secretion (not shown), we observed a consistent reduction of *GLIS3* mRNA in comparison to five control shRNA hairpins, whereas other control genes remained unaltered (Figure 6A). This effect was stably maintained over several days in culture (Figure 6B). Thus, *HI-LNC25* positively regulates *GLIS3* mRNA, supporting the notion that our collection of islet lncRNAs contains regulatory transcripts.

Dysregulation of Islet lncRNAs in Human T2D

To determine the potential role of islet lncRNAs in the pathogenesis of diabetes, we first examined whether lncRNAs are abnormally expressed in human islets from donors with T2D.

We examined the panel of 13 lncRNAs and added *KCNQ1OT1,* a lncRNA that was detected as an overlapping antisense lncRNA in our analysis, and has been previously genetically associated with T2D (Voight et al., 2010). We compared these lncRNAs in islets from 19 nondiabetic and 16 T2D donors. We found that two lncRNAs, namely *KCNQ1OT1* and *HI-LNC45*, were significantly increased or decreased in T2D islets, respectively ($p < 0.02$) (Figure 7A). To ensure that this result was not due to differences in islet purity between groups of samples, we normalized the expression levels of HI-LNCs relative to the islet transcription factor *PAX6*, and found similar results (data not shown). Thus, selected lncRNAs are dysregulated in T2D islets.

A Subset of lncRNAs Map to Genomic Loci Underlying Human Diabetes

To further examine the potential role of islet lncRNAs in the genetic susceptibility for diabetes, we first examined loci implicated in monogenic syndromes of β cell dysfunction disorders. Islet antisense lncRNAs were found within loci underlying neonatal diabetes (*ABCC8/KCNJ11*), pancreas agenesis (*GATA6*), and monogenic diabetes (*HNF1A*) (Allen et al., 2012; Bell and Polonsky, 2001; Gloyn et al., 2004). Because several antisense lncRNAs have been implicated in *cis* regulatory effects (Wilusz et al., 2009; Xu et al., 2011), these islet lncRNAs provide additional candidate regulatory elements that may contain pathogenic variants.

Next, we examined genetic association data from multifactorial T2D and related continuous glycemic measures, including fasting glucose in nondiabetic individuals (Dupuis et al., 2010; Voight et al., 2010). Currently, >50 loci are known to show association to T2D and related traits, and at most loci noncoding variants are likely to be causal (McCarthy, 2010). Earlier studies showed that two annotated lncRNAs, *ANRIL/CDKN2BAS* and *KCNQ1OT1*, map within established T2D susceptibility loci (Voight et al., 2010; Zeggini et al., 2007). To evaluate the potential overlap between newly identified islet lncRNAs with T2D loci, we used MAGENTA (Segrè et al., 2010). This tool tests whether a prespecified set of genes is enriched for trait associations in genome-wide data, and was applied here to available data sets for T2D and related continuous glycemic measures, as well as five nonislet related control phenotypes.

We found that islet lncRNA transcripts were, as a group, enriched for association with both T2D ($p = 0.02$) and fasting glucose ($p = 0.01$), with no enrichment for the control phenotypes apart from height ($p = 0.02$) (Lango Allen et al., 2010) and waist-hip ratio ($p = 0.001$) (Heid et al., 2010). When lncRNAs were examined by type (intergenic versus antisense), enrichment of association signal for intergenic lncRNAs was only observed for T2D and fasting glucose (T2D $p = 0.03$; fasting glucose $p = 0.003$) (Table S8). Of 55 T2D susceptibility loci, 9 contained islet lncRNAs within 150 kb of the reported lead SNP, 6 of which have been linked directly to β cell dysfunction (Figure 7A) (Cho et al., 2012; Dupuis et al., 2010; Kooner et al., 2011; Strawbridge et al., 2011; van de Bunt and Gloyn, 2010; Voight et al., 2010). Examples include a lncRNA in the vicinity of *PROX1*, which overlaps the region of strongest association, very likely to contain the casual SNP (Figure 7B), and the most significant lncRNA in the MAGENTA analysis, near *WFS1* (Figure 7C). These studies therefore offer a new class of genomic elements that can be interrogated to dissect the functional etiology of T2D susceptibility.

DISCUSSION

Increasing evidence points to a regulatory function of many lncRNAs, which suggests a pivotal role in physiology and disease (Guttman and Rinn, 2012; Wang and Chang, 2011). We have integrated transcriptional and chromatin maps to systematically annotate lncRNA genes in pancreatic islet cells, a key tissue for human diabetes. Strand-specific analysis uncovered hundreds of intergenic islet lncRNAs and many others that were antisense to protein-coding genes and hence not discernable using conventional cDNA libraries. Human islet lncRNAs were found to be highly cell-type specific, developmentally regulated, and tightly linked to the differentiation of ES cell-derived islet cells. We identified mouse orthologous transcripts, and discovered that they are dynamically regulated in a similar manner as human islet lncRNAs. We focused on a prototypical islet lncRNA and demonstrate that it acts as a positive regulator of an islet mRNA. Finally, we show examples of lncRNAs

(C) Fraction of orthologous regions to human RefSeq exons, random intergenic regions, and lncRNAs that are transcribed higher than 0.5 RPKM in mouse islet RNA-seq.
(D) qPCR confirmation in mouse islets and MIN6 mouse β cell line for 7/7 orthologous mouse islet lncRNAs.
(E) qPCR reveals islet expression of 4/7 orthologous regions where islet RNA-seq did not reveal transcription, but not in random intergenic regions. No RT denotes control reactions lacking reverse transcriptase.
(F) Orthologous mouse lncRNAs are more frequently islet specific than RefSeq genes.
(G) Examples of conserved mouse islet-specific lncRNAs. Red arrows depict lncRNA orientation. The y axis for all tissues is adjusted to the same value as islets (RPKM).
(H) qPCR shows that 4/7 lncRNAs were silent in E13.5 mouse embryonic pancreas and were induced in adult islets.
(I) Expression of several orthologous lncRNAs is regulated by glucose. Mouse islets were cultured either under low (4 mM) or high (11 mM) glucose concentrations for 72 hr. Expression is shown as a fraction of expression under low-glucose conditions (n = 4). All qPCR values are normalized to *Tbp* mRNA. Bars represent average ± SEM and asterisks denote $p < 0.001$ (Student's t test). See also Figure S4 and Table S7.

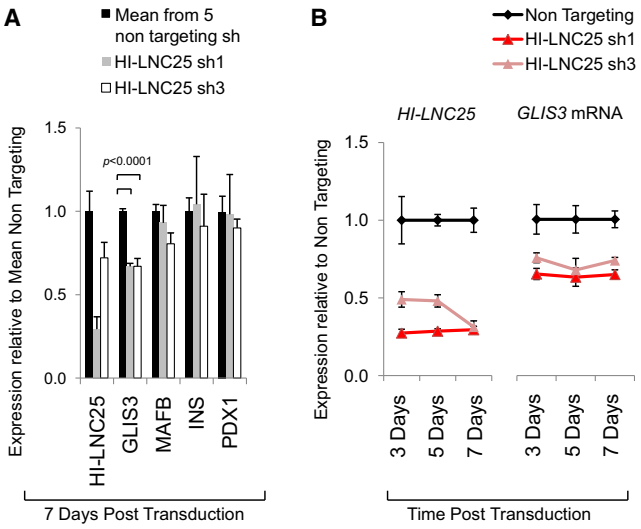


Figure 6. Knockdown of *HI-LNC25* in β Cells Causes Downregulation of *GLIS3*

(A) Human β cells (EndoC-β H1) were transduced with lentiviral vectors expressing two independent RNA hairpins that target *HI-LNC25* RNA or five independently transduced negative control nontargeting shRNA sequences. Cells were harvested at 7 days and assayed for *GLIS3* mRNA and several control genes.

(B) Effects on *GLIS3* mRNA were stable 3–7 days post transduction. Expression levels were normalized to *TBP* mRNA and shown as a fraction of nontargeting vector cells. The results reflect four independent experiments. Bars are average ± SEM, *p* values were obtained with Student's t test.

that are dysregulated in T2D islets and others that map to T2D susceptibility loci. These studies open new avenues to study the role of functional lncRNAs in islet-cell disease and therapeutic programming.

LncRNAs Are Dynamically Regulated during Islet Cell Maturation

Islet lncRNAs show a striking cell-type specific expression pattern. This finding is in line with a recent analysis of intergenic lncRNAs across multiple tissues (Pauli et al., 2012), and with earlier descriptions of lncRNAs that have distinct cell-type specific expression patterns in brain (Mercer et al., 2008). Our studies additionally demonstrate that islet-specific lncRNA genes are inactive in embryonic pancreatic progenitors and subsequently become activated in islet cells. Likewise, numerous lncRNAs were activated during the final endocrine differentiation step of a pluripotent cell differentiation protocol. This indicates that islet lncRNAs are an integral component of the endocrine differentiation program.

Islet-specific lncRNA transcription was linked to clusters of open chromatin. Such clusters were previously shown to be associated with islet-specific transcription (Gaulton et al., 2010). Interestingly, we also observed concordant cell-specific transcription of lncRNAs with their nearest protein-coding genes, in accordance with other genomic analysis of lncRNAs (Ponjavic et al., 2009). These data suggest that concordant tissue-specific expression of lncRNA and protein-coding pairs is linked to shared chromosomal regulatory domains.

The islet specificity of lncRNAs is reminiscent of the recent discovery of transcription factors that show a similar stage- and cell-specific activation pattern (reviewed in Servitja and Ferrer, 2004). Several such islet-enriched factors have now been shown to control islet development and have consequently been exploited to program insulin-expressing cells from somatic cells (Collombat et al., 2009; Zhou et al., 2008). By analogy, islet lncRNAs that exert gene-regulatory functions could also be employed in efforts to program functional β cells. This prospect is supported by recent studies showing that a lncRNA can promote reprogramming pluripotent cells from somatic cells (Loewer et al., 2010).

Several islet lncRNAs were specifically activated during the in vivo maturation step of the ES cell differentiation protocol. This result is significant because the major roadblock for efforts to derive fully functional β cells from pluripotent cells lies in the inability to complete the differentiation process in vitro (Van Hoof et al., 2009). Our findings suggest that lncRNAs that are induced at this late stage can be used as biomarkers of mature β cells and potentially be exploited as effectors to promote β cell programming.

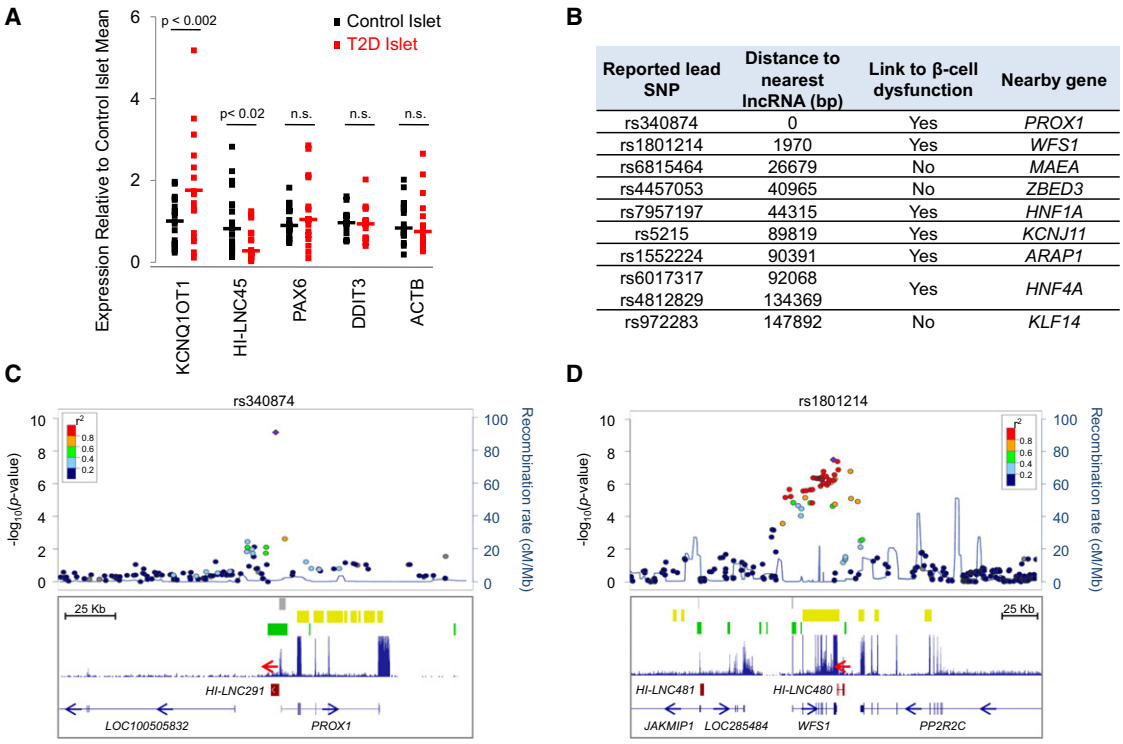
Evolutionary Conserved Regulation of Islet lncRNAs

For a major fraction of lncRNAs we identified orthologous mouse genomic regions, although the overall sequence identity was clearly not subject to a similar evolutionary constraint as protein-coding genes, in keeping with earlier studies of lncRNAs (Pauli et al., 2012; Ulitsky et al., 2011). Several mechanisms have been previously proposed to explain this comparatively low sequence conservation, including a hypothetically less stringent requirement for primary sequence conservation to maintain functional secondary structures, or the existence of very short stretches of functional sequences (Ulitsky et al., 2011). It is also conceivable that evolutionary changes in lncRNA sequences are functionally coupled to other parallel genomic changes, such as the acquisition of species-specific repetitive elements.

Importantly, in the current study we discovered that a significant fraction of orthologous mouse sequences were transcribed in mouse islet cells and β cell lines. Orthologous lncRNA pairs display similar cell-type specific expression and stage-specific regulation during embryonic development. This finding points to an evolutionary conserved functional property of lncRNAs that extends beyond primary sequence and further supports that lncRNAs are an integral component of the mammalian islet-cell differentiated phenotype.

Islet lncRNAs and Diabetes

The identification of lncRNAs opens a new framework to study human pathophysiology. Our study revealed lncRNA genes at six neonatal diabetes loci and in *HNF1A*, the most common monogenic diabetes locus (Bell and Polonsky, 2001). Most such lncRNAs run antisense from the coding gene. Antisense transcripts have been shown to control the transcription of protein-coding genes in *cis* (Wilusz et al., 2009; Xu et al., 2011). Antisense lncRNAs identified in this study are thus candidate regulators of protein-coding genes implicated in human β cell monogenic disorders and can be potentially affected by pathogenic mutations.



Reported lead SNP	Distance to nearest lncRNA (bp)	Link to β-cell dysfunction	Nearby gene
rs340874	0	Yes	*PROX1*
rs1801214	1970	Yes	*WFS1*
rs6815464	26679	No	*MAEA*
rs4457053	40965	No	*ZBED3*
rs7957197	44315	Yes	*HNF1A*
rs5215	89819	Yes	*KCNJ11*
rs1552224	90391	Yes	*ARAP1*
rs6017317	92068	Yes	*HNF4A*
rs4812829	134369		
rs972283	147892	No	*KLF14*

Figure 7. Several Human lncRNAs Map to T2D-Associated Loci

(A) *KCNQ1OT1* and *HI-LNC45* are dysregulated in islets from T2D patients. Islet LncRNAs and control mRNAs were assessed by qPCR in islets isolated from 19 control and 16 T2D donors. Values are normalized to TBP mRNA and expressed as a fraction of controls. *P* values represent Mann-Whitney significance test.

(B) Nine T2D-associated loci contain islet lncRNAs mapping within 150 kb of the reported lead SNPs. Six of these loci have been linked to β cell dysfunction (Dupuis et al., 2010; van de Bunt and Gloyn, 2010; Voight et al., 2010).

(C and D) Islet lncRNAs map to established T2D association signals located near *PROX1* and *WFS1*. The top panel shows T2D association *p* values for all analyzed SNPs in the published DIAGRAM genome-wide association meta analysis (Voight et al., 2010) and the combined *p* value for the "lead" SNP after further follow up (rs340874, $p = 7.2 \cdot 10^{-10}$) (rs1801214, $p = 3.16 \cdot 10^{-8}$), which are denoted as purple diamonds. SNP colors indicate LD relationships. RNA-seq and ChIP-seq is presented in the bottom panel as described in the legend for Figure 2. See also Table S8.

These results are also relevant to the molecular etiology of T2D, a disease caused by abnormal β cell function or growth (Bell and Polonsky, 2001; McCarthy, 2010). We identified examples of islet lncRNAs that are dysregulated in T2D. Our findings also revealed islet-cell lncRNAs as candidates to dissect the underpinnings of noncoding variants underlying T2D risk. More generally, these results set the stage for future studies to dissect the potential role of inherited and acquired defects in lncRNA genes in human β cell physiology and disease. Several experimental model systems, including mouse genetics coupled with perturbation studies in human β cells, are now available to address this challenge.

EXPERIMENTAL PROCEDURES

Pancreatic Islets and Tissues

Human islets were isolated at the University of Virginia, University of Geneva, University of Lille, and San Raffaele Scientific Institute islet centers (Bucher et al., 2005). Human islets used for RNA-seq and ChIP-seq were cultured with CMRL 1066 medium containing 10% fetal calf serum (FCS) before shipment, after which they were cultured for 3 days with RPMI 1640 medium containing 11 mM glucose, supplemented with 10% FCS. For glucose regulation experiments, human or mouse islets were cultured for 72 hr in RPMI 1640 medium containing 4 or 11 mM glucose. Donor information for nondiabetic and diabetic donors is provided in Table S2. Samples were selected based on islet cell purity, as assessed by dithizone staining, immunofluorescence analysis, and qPCR analysis of cell-specific mRNAs.

RNA Analysis

RNA was isolated with Trizol (Invitrogen) or RNeasy (QIAGEN). DNase I treatment and a control lacking reverse transcriptase were performed on all RNA samples. Quantitative PCR was performed as described based on SYBR green detection (van Arensbergen et al., 2010). See Table S9 for oligonucleotide sequences.

miR-Based Knockdown Assays

Lentiviral vectors carrying two miRNA-based shRNAs targeting *HI-LNC25* and five nontargeting control sequences were transduced into the EndoC-β H1 human β cell line (Castaing et al., 2005; Ravassard et al., 2011; Scharfmann et al., 2008). Oligonucleotide sequences are shown in Table S9. Nontransduced cells were assayed in parallel. Cells were assayed 3 or 7 days post transduction.

ES Cell Differentiation

Human ES cells were differentiated to pancreatic insulin-producing cells as described (Kroon et al., 2008).

RNA-Seq and ChIP-Seq

Chromatin immunoprecipitations (ChIP) were performed as described (Boj et al., 2009), with modifications described in Supplemental Experimental

Procedures. ChIP libraries were prepared according to Illumina protocols and sequencing of single-end reads was performed on a GAIIx system. RNA-seq was performed from unidirectional and nondirectional cDNA libraries prepared from PolyA+ or rRNA-depleted pancreatic islet RNA, and sequenced using GAIIx, HiSeq2000, or SOLiD 4 systems. Nonpancreatic RNA-seq reads were retrieved from the Illumina human BodyMap2 data set (http://www.ebi.ac.uk/arrayexpress/browse.html?keywords=E-MTAB-513) and the ENCODE/LICR Project (Birney et al., 2007). Reads were aligned to the NCBI36/hg18 and NCBI37/mm9 genomes using Bowtie v0.11.3 for ChIP-seq, allowing one mismatch per read, no multimapping, and no clonal reads. Nondirectional RNA-seq libraries were mapped with TopHat v1.2.0, and unidirectional libraries with Bowtie, using parameters described in the Supplemental Experimental Procedures. The number of uniquely aligned reads, read length, and sample information are shown in Tables S1 and S7.

LncRNA Characterization and Analysis

Transcriptional units were defined from the RNA-seq alignments and joined by splice junctions and paired-end mapping to create putative gene models. Transcriptional units that were expressed at >0.5 RPKMs, with H3K4me3 enrichment in the 5′ region in at least two samples, and either did not overlap with annotated protein-coding genes or did overlap but were transcribed in the opposite strand and were assessed for protein-coding potential and processed through our lncRNA discovery pipeline. To assess evolutionary sequence conservation human islet lncRNA exon locations or the same fragments randomized 1000 times in the alignable portion of intergenic space were analyzed. Orthologous mouse genomic sequences were identified with the LiftOver tool from the UCSC browser. Mouse islet RNA-seq was used to detect transcripts expressed at >0.5 RPKM in orthologus or randomized control regions. To evaluate islet specificity of lncRNAs, Cluster 3.0 v1.5a was used for hierarchical clustering of transcript expression in islet cells and 16 nonpancreatic tissues. Furthermore, a score was created that measures the difference between the average expression in 3 human islet PolyA+ samples and the average +2 standard deviations in the remaining 16 tissues. MAGENTA (Segrè et al., 2010) was used to evaluate whether lncRNAs were enriched for association to T2D and related glycemic traits, with adaptions described in the Supplemental Experimental Procedures.

Statistical Analysis

Chi-square, Pearson's *r* coefficient, and independent 2-group Mann-Whitney *U* tests were implemented with the R statistical package (http://www.r-project.org). Student's t test was performed with Microsoft Excel for quantitative PCR experiments. Error bars in Figures 4–6 represent SEM.

ACCESSION NUMBERS

Data sets have been deposited in the ArrayExpress Archive under accession number E-MTAB-1294. Islet lncRNAs coordinates are available at http://www.betacellregulation.net.

SUPPLEMENTAL INFORMATION

Supplemental Information includes nine figures, four tables, Supplemental Experimental Procedures, and Supplemental References and can be found with this article online at http://dx.doi.org/10.1016/j.cmet.2012.08.010.

ACKNOWLEDGMENTS

This work was funded by grants from the NIH-BCBC (2U01 DK072473-06 to J.F., P.R., and L.S.; U01-DK089567 to M.S. and J.F.), the Juvenile Diabetes Research Foundation (26-2008-633 to J.F. and 31-2008-416 to T.B., F.P., and L.P. to support human islet isolation), Ministerio de Economía y Competitividad (SAF2008-03116 to J.F. and a Ph.D. fellowship to Ignasi Morán). A.L.G. is a Wellcome Trust Senior Fellow in Basic Biomedical Research (WT 09510/Z/10/Z). We thank Rudolph Leibel (University of Columbia) for providing B6.V-Lepob/J mice, Bing Ren (Ludwig Institute for Cancer Research) for generating the ENCODE mouse tissue RNA-seq data, Kelli Bramlett (Genome Sequencing Collaborations Group) for her help in generation of SOLiD data and Viacyte, Inc. for performing implantations of hES cell derivatives. We are also grateful to Ayellet Segrè (Broad Institute) for helpful discussion on MAGENTA, and to Chris Stoeckert and Elisabetta Manduchi (University of Pennsylvania) for help with the online submission.

Received: March 21, 2012
Revised: July 30, 2012
Accepted: August 31, 2012
Published online: October 2, 2012

REFERENCES

Allen, H.L., Flanagan, S.E., Shaw-Smith, C., De Franco, E., Akerman, I., Caswell, R., Ferrer, J., Hattersley, A.T., and Ellard, S. (2012). GATA6 haploinsufficiency causes pancreatic agenesis in humans. Nat. Genet. *44*, 20–22.

Artner, I., Blanchi, B., Raum, J.C., Guo, M., Kaneko, T., Cordes, S., Sieweke, M., and Stein, R. (2007). MafB is required for islet beta cell maturation. Proc. Natl. Acad. Sci. USA *104*, 3853–3858.

Bell, G.I., and Polonsky, K.S. (2001). Diabetes mellitus and genetically programmed defects in beta-cell function. Nature *414*, 788–791.

Bensellam, M., Van Lommel, L., Overbergh, L., Schuit, F.C., and Jonas, J.C. (2009). Cluster analysis of rat pancreatic islet gene mRNA levels after culture in low-, intermediate- and high-glucose concentrations. Diabetologia *52*, 463–476.

Bertone, P., Stolc, V., Royce, T.E., Rozowsky, J.S., Urban, A.E., Zhu, X., Rinn, J.L., Tongprasit, W., Samanta, M., Weissman, S., et al. (2004). Global identification of human transcribed sequences with genome tiling arrays. Science *306*, 2242–2246.

Birney, E., Stamatoyannopoulos, J.A., Dutta, A., Guigó, R., Gingeras, T.R., Margulies, E.H., Weng, Z., Snyder, M., Dermitzakis, E.T., Thurman, R.E., et al; ENCODE Project Consortium; NISC Comparative Sequencing Program; Baylor College of Medicine Human Genome Sequencing Center; Washington University Genome Sequencing Center; Broad Institute; Children's Hospital Oakland Research Institute. (2007). Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. Nature *447*, 799–816.

Boj, S.F., Servitja, J.M., Martin, D., Rios, M., Talianidis, I., Guigo, R., and Ferrer, J. (2009). Functional targets of the monogenic diabetes transcription factors HNF-1alpha and HNF-4alpha are highly conserved between mice and humans. Diabetes *58*, 1245–1253.

Bucher, P., Mathe, Z., Morel, P., Bosco, D., Andres, A., Kurfuest, M., Friedrich, O., Raemsch-Guenther, N., Buhler, L.H., and Berney, T. (2005). Assessment of a novel two-component enzyme preparation for human islet isolation and transplantation. Transplantation *79*, 91–97.

Cabili, M.N., Trapnell, C., Goff, L., Koziol, M., Tazon-Vega, B., Regev, A., and Rinn, J.L. (2011). Integrative annotation of human large intergenic noncoding RNAs reveals global properties and specific subclasses. Genes Dev. *25*, 1915–1927.

Carninci, P., Kasukawa, T., Katayama, S., Gough, J., Frith, M.C., Maeda, N., Oyama, R., Ravasi, T., Lenhard, B., Wells, C., et al; FANTOM Consortium; RIKEN Genome Exploration Research Group and Genome Science Group (Genome Network Project Core Group). (2005). The transcriptional landscape of the mammalian genome. Science *309*, 1559–1563.

Castaing, M., Duvillié, B., Quemeneur, E., Basmaciogullari, A., and Scharfmann, R. (2005). Ex vivo analysis of acinar and endocrine cell development in the human embryonic pancreas. Dev. Dyn. *234*, 339–345.

Cho, Y.S., Chen, C.H., Hu, C., Long, J., Ong, R.T., Sim, X., Takeuchi, F., Wu, Y., Go, M.J., Yamauchi, T., et al; DIAGRAM Consortium; MuTHER Consortium. (2012). Meta-analysis of genome-wide association studies identifies eight new loci for type 2 diabetes in east Asians. Nat. Genet. *44*, 67–72.

Collombat, P., Xu, X., Ravassard, P., Sosa-Pineda, B., Dussaud, S., Billestrup, N., Madsen, O.D., Serup, P., Heimberg, H., and Mansouri, A. (2009). The ectopic expression of Pax4 in the mouse pancreas converts progenitor cells into alpha and subsequently beta cells. Cell *138*, 449–462.

Dorrell, C., Schug, J., Lin, C.F., Canaday, P.S., Fox, A.J., Smirnova, O., Bonnah, R., Streeter, P.R., Stoeckert, C.J., Jr., Kaestner, K.H., and

Grompe, M. (2011). Transcriptomes of the major human pancreatic cell types. Diabetologia *54*, 2832–2844.

Dupuis, J., Langenberg, C., Prokopenko, I., Saxena, R., Soranzo, N., Jackson, A.U., Wheeler, E., Glazer, N.L., Bouatia-Naji, N., Gloyn, A.L., et al; DIAGRAM Consortium; GIANT Consortium; Global BPgen Consortium; Anders Hamsten on behalf of Procardis Consortium; MAGIC investigators. (2010). New genetic loci implicated in fasting glucose homeostasis and their impact on type 2 diabetes risk. Nat. Genet. *42*, 105–116.

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

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., et al. (2004). Activating mutations in the gene encoding the ATP-sensitive potassium-channel subunit Kir6.2 and permanent neonatal diabetes. N. Engl. J. Med. *350*, 1838–1849.

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., et al. (2010). Long non-coding RNA HOTAIR reprograms chromatin state to promote cancer metastasis. Nature *464*, 1071–1076.

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

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

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

Halban, P.A., Kahn, S.E., Lernmark, A., and Rhodes, C.J. (2001). Gene and cell-replacement therapy in the treatment of type 1 diabetes: how high must the standards be set? Diabetes *50*, 2181–2191.

Heid, I.M., Jackson, A.U., Randall, J.C., Winkler, T.W., Qi, L., Steinthorsdottir, V., Thorleifsson, G., Zillikens, M.C., Speliotes, E.K., Mägi, R., et al; MAGIC. (2010). Meta-analysis identifies 13 new loci associated with waist-hip ratio and reveals sexual dimorphism in the genetic basis of fat distribution. Nat. Genet. *42*, 949–960.

Hu, W., Yuan, B., Flygare, J., and Lodish, H.F. (2011). Long noncoding RNA-mediated anti-apoptotic activity in murine erythroid terminal differentiation. Genes Dev. *25*, 2573–2578.

Kooner, J.S., Saleheen, D., Sim, X., Sehmi, J., Zhang, W., Frossard, P., Been, L.F., Chia, K.S., Dimas, A.S., Hassanali, N., et al; DIAGRAM; MuTHER. (2011). Genome-wide association study in individuals of South Asian ancestry identifies six new type 2 diabetes susceptibility loci. Nat. Genet. *43*, 984–989.

Kroon, E., Martinson, L.A., Kadoya, K., Bang, A.G., Kelly, O.G., Eliazer, S., Young, H., Richardson, M., Smart, N.G., Cunningham, J., et al. (2008). Pancreatic endoderm derived from human embryonic stem cells generates glucose-responsive insulin-secreting cells in vivo. Nat. Biotechnol. *26*, 443–452.

Lango Allen, H., Estrada, K., Lettre, G., Berndt, S.I., Weedon, M.N., Rivadeneira, F., Willer, C.J., Jackson, A.U., Vedantam, S., Raychaudhuri, S., et al. (2010). Hundreds of variants clustered in genomic loci and biological pathways affect human height. Nature *467*, 832–838.

Loewer, S., Cabili, M.N., Guttman, M., Loh, Y.-H., Thomas, K., Park, I.H., Garber, M., Curran, M., Onder, T., Agarwal, S., et al. (2010). Large intergenic non-coding RNA-RoR modulates reprogramming of human induced pluripotent stem cells. Nat. Genet. *42*, 1113–1117.

Mattick, J.S., and Makunin, I.V. (2006). Non-coding RNA. Hum. Mol. Genet. *15* (Spec No 1), R17–R29.

McCarthy, M.I. (2010). Genomics, type 2 diabetes, and obesity. N. Engl. J. Med. *363*, 2339–2350.

Mercer, T.R., Dinger, M.E., Sunkin, S.M., Mehler, M.F., and Mattick, J.S. (2008). Specific expression of long noncoding RNAs in the mouse brain. Proc. Natl. Acad. Sci. USA *105*, 716–721.

Pauli, A., Valen, E., Lin, M.F., Garber, M., Vastenhouw, N.L., Levin, J.Z., Fan, L., Sandelin, A., Rinn, J.L., Regev, A., and Schier, A.F. (2012). Systematic identification of long noncoding RNAs expressed during zebrafish embryogenesis. Genome Res. *22*, 577–591.

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

Piper, K., Brickwood, S., Turnpenny, L.W., Cameron, I.T., Ball, S.G., Wilson, D.I., and Hanley, N.A. (2004). Beta cell differentiation during early human pancreas development. J. Endocrinol. *181*, 11–23.

Ponjavic, J., Oliver, P.L., Lunter, G., and Ponting, C.P. (2009). Genomic and transcriptional co-localization of protein-coding and long non-coding RNA pairs in the developing brain. PLoS Genet. *5*, e1000617.

Ravassard, P., Hazhouz, Y., Pechberty, S., Bricout-Neveu, E., Armanet, M., Czernichow, P., and Scharfmann, R. (2011). A genetically engineered human pancreatic β cell line exhibiting glucose-inducible insulin secretion. J. Clin. Invest. *121*, 3589–3597.

Rinn, J.L., Kertesz, M., Wang, J.K., Squazzo, S.L., Xu, X., Brugmann, S.A., Goodnough, L.H., Helms, J.A., Farnham, P.J., Segal, E., and Chang, H.Y. (2007). Functional demarcation of active and silent chromatin domains in human HOX loci by noncoding RNAs. Cell *129*, 1311–1323.

Scharfmann, R., Xiao, X., Heimberg, H., Mallet, J., and Ravassard, P. (2008). Beta cells within single human islets originate from multiple progenitors. PLoS ONE *3*, e3559.

Schuit, F., Flamez, D., De Vos, A., and Pipeleers, D. (2002). Glucose-regulated gene expression maintaining the glucose-responsive state of beta-cells. Diabetes *51* (*Suppl 3*), S326–S332.

Segrè, A.V., Groop, L., Mootha, V.K., Daly, M.J., and Altshuler, D.; DIAGRAM Consortium; MAGIC investigators. (2010). Common inherited variation in mitochondrial genes is not enriched for associations with type 2 diabetes or related glycemic traits. PLoS Genet. *6*, 6.

Senée, V., Chelala, C., Duchatelet, S., Feng, D., Blanc, H., Cossec, J.C., Charon, C., Nicolino, M., Boileau, P., Cavener, D.R., et al. (2006). Mutations in GLIS3 are responsible for a rare syndrome with neonatal diabetes mellitus and congenital hypothyroidism. Nat. Genet. *38*, 682–687.

Servitja, J.M., and Ferrer, J. (2004). Transcriptional networks controlling pancreatic development and beta cell function. Diabetologia *47*, 597–613.

Sleutels, F., Zwart, R., and Barlow, D.P. (2002). The non-coding Air RNA is required for silencing autosomal imprinted genes. Nature *415*, 810–813.

Strawbridge, R.J., Dupuis, J., Prokopenko, I., Barker, A., Ahlqvist, E., Rybin, D., Petrie, J.R., Travers, M.E., Bouatia-Naji, N., Dimas, A.S., et al; DIAGRAM Consortium; GIANT Consortium; MuTHER Consortium; CARDIoGRAM Consortium; C4D Consortium. (2011). Genome-wide association identifies nine common variants associated with fasting proinsulin levels and provides new insights into the pathophysiology of type 2 diabetes. Diabetes *60*, 2624–2634.

Ulitsky, I., Shkumatava, A., Jan, C.H., Sive, H., and Bartel, D.P. (2011). Conserved function of lincRNAs in vertebrate embryonic development despite rapid sequence evolution. Cell *147*, 1537–1550.

van Arensbergen, J., García-Hurtado, J., Moran, I., Maestro, M.A., Xu, X., Van de Casteele, M., Skoudy, A.L., Palassini, M., Heimberg, H., and Ferrer, J. (2010). Derepression of Polycomb targets during pancreatic organogenesis allows insulin-producing beta-cells to adopt a neural gene activity program. Genome Res. *20*, 722–732.

van de Bunt, M., and Gloyn, A.L. (2010). From genetic association to molecular mechanism. Curr. Diab. Rep. *10*, 452–466.

Van Hoof, D., D'Amour, K.A., and German, M.S. (2009). Derivation of insulin-producing cells from human embryonic stem cells. Stem Cell Res. (Amst.) *3*, 73–87.

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., et al; MAGIC investigators; GIANT Consortium. (2010). Twelve type 2 diabetes susceptibility loci identified through large-scale association analysis. Nat. Genet. *42*, 579–589.

Wang, K.C., and Chang, H.Y. (2011). Molecular mechanisms of long noncoding RNAs. Mol. Cell *43*, 904–914.

Wilusz, J.E., Sunwoo, H., and Spector, D.L. (2009). Long noncoding RNAs: functional surprises from the RNA world. Genes Dev. *23*, 1494–1504.

Xu, Z., Wei, W., Gagneur, J., Clauder-Münster, S., Smolik, M., Huber, W., and Steinmetz, L.M. (2011). Antisense expression increases gene expression variability and locus interdependency. Mol. Syst. Biol. *7*, 468.

Zeggini, E., Weedon, M.N., Lindgren, C.M., Frayling, T.M., Elliott, K.S., Lango, H., Timpson, N.J., Perry, J.R., Rayner, N.W., Freathy, R.M., et al; Wellcome Trust Case Control Consortium (WTCCC). (2007). Replication of genome-wide association signals in UK samples reveals risk loci for type 2 diabetes. Science *316*, 1336–1341.

Zhou, Q., Brown, J., Kanarek, A., Rajagopal, J., and Melton, D.A. (2008). In vivo reprogramming of adult pancreatic exocrine cells to beta-cells. Nature *455*, 627–632.