

Mechanisms of Type 2 Diabetes Susceptibility

A thesis submitted to the University of Oxford



in candidature for the degree of Doctor of Philosophy

by

Mary E. Travers

St. Hilda's College, University of Oxford

and

Oxford Centre for Diabetes, Endocrinology and Metabolism
Nuffield Department of Clinical Medicine, University of Oxford

Trinity 2012

Memorandum

The work in this thesis is the original work of the author. Experiments were carried out 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 Medical Research Council.

Investigation of the *KCNQ1* T2D susceptibility locus (chapter 3) has been published in *Diabetes*. The manuscript (Travers et al., 2012) is provided as an appendix. Meta-analysis of genotyping data obtained in experiments described in this thesis with data from the Early Growth Genetics (EGG) consortium was performed by Dr. Rachel Freathy.

Expression profiling (chapter 4) was performed as part of the Meta-analysis of Glucose and Insulin-related Traits Consortium (MAGIC). Genome-wide association analysis was performed by consortium members and led by Dr. Rona J. Strawbridge. The results detailed in this thesis form part of the published manuscript Strawbridge et al., 2011, provided as an appendix

CNV analysis (chapter 5) was conducted as part of the Wellcome Trust Case Control Consortium (WTCCC). Genome-wide association analysis was performed by consortium members and lead by Dr. Nick Craddock. The published manuscript Craddock et al., 2010 is provided as an appendix.

Rare and low frequency variant genotyping (chapter 6) was performed on the basis of genome-wide association analysis conducted by Dr. Andrew Morris.

I hereby state that no part of this thesis has been submitted for any other degree at this or any other university.

This thesis is approximately 48,000 words.

Mechanisms of Type 2 Diabetes Susceptibility

Mary Travers, St Hilda's College

Submitted in candidature for the degree of Doctor of Philosophy, Trinity term 2012

Abstract

Type 2 diabetes (T2D) has a genetic component which is only partially understood. The majority of genetic variance in disease susceptibility is unaccounted for, whilst the precise transcripts and molecular mechanisms through which most risk variants exert their effect is unclear. A complete understanding of T2D susceptibility mechanisms could have benefits in risk prediction, and in drug discovery through the identification of novel therapeutic targets. Work presented in this thesis aims to define relevant transcripts and disease mechanisms at known susceptibility loci, and to identify disease association with classes of genetic variation other than common single nucleotide polymorphisms (SNPs).

KCNQ1 contains intronic variants associated with T2D susceptibility and β -cell dysfunction, but only maternally-inherited alleles confer increased disease risk. It maps within an imprinted domain with an established role in congenital and islet-specific growth phenotypes. Using human adult islet and foetal pancreas samples, I refined the transcripts and developmental stage at which T2D susceptibility must be conferred by demonstrating developmentally plastic monoallelic and biallelic expression. I identified a potential risk mechanism through the effect of T2D risk alleles upon DNA methylation.

The disease-associated regions identified through genome-wide association (GWA) studies often contain multiple transcripts. I performed mRNA expression profiling of genes within loci associated with raised proinsulin/insulin ratios in human islets and metabolically relevant tissues. Some genes (notably *CT62*) were not expressed and therefore excluded from consideration for a risk effect, whilst others (for example *C2CD4A*) were highlighted as good regional candidates due to specific expression in relevant tissues.

GWA studies for T2D risk have focused predominantly upon common single nucleotide polymorphisms. As part of a consortium conducting GWA analysis for copy number variation (CNV) and T2D risk, I optimised and compared alternative methods of CNV genotyping, before using this information to validate two signals of disease association. I genotyped three rare single nucleotide variants emerging from an association study with T2D risk based on imputed data, providing an indication of imputation accuracy and more powerful disease association analysis.

These data underscore the challenge of translating association signals to causal mechanisms, and of identifying alternative forms of genomic variation which contribute to T2D risk. My work highlights candidates for functional analysis around proinsulin-associated loci, and makes significant progress towards uncovering risk mechanisms at the *KCNQ1* locus.

Acknowledgements

Whilst I know that this is the first (and possibly only) section of my thesis that many people will read, it is the section which I have put off writing for as long as possible. This is not because I have no-one to thank, but because it is so difficult to pull out individuals and events from such an amazing and eventful four years.

Heartfelt thanks go to my supervisors, Prof Mark McCarthy and Dr Anna Gloyn, for being such supportive and encouraging guides. I have learned so much from you, and am grateful that you took a punt on the girl who had never held a pipette! In return, I hope that I have instilled a life-long love of tractors in Freya. Thanks also go to Leif Groop and Neil Hanley for making me feel so welcome during my time in Malmö and Manchester, and to Andrew Morris and Deborah Mackay for fielding my many statistical and pyrosequencing questions.

I am incredibly lucky to have spent my journeys to work over the past four years looking forward to cheery hellos and coffee-time banter. I count the OCDEM crew as real friends as well as colleagues. Thank you to my lab-dad, -mum and -auntie (Chris, Mandy and Amy) for tea, much-needed laboratory wisdom and occasional hugs. I can imagine no more all-round awesome office than the Fellows' room, and I thank its inhabitants past (Olly, Nick) and present (Lily, Martijn, Hannah, Anne, Jana) for the constant laughs, therapeutic moans and sticker collecting. Thank you also for indulging my unhealthy tea habit. Thank you to the upstairs crew (Aparna, Gaya, Peta, Matt, Saima and Katharine) for your brilliant laughs and inappropriate senses of humour - I didn't always visit you just to steal biscuits, I promise!

Laura, Nic, Coops, Neelam, Matt and Lisa – thank you for the nutella, the bananagrams, the microwave cake, the vodka and all-round fun times that we have shared in houses together. Nic and Laura - thank you for being my sounding boards and sources of wisdom through the writing process. Coops and Neelam - I hope I can do the same for you.

Sam, John & Ben, Natalie, Sarah F and Jacquie – you may not fit neatly into any of the categories above, but my time in Oxford would not have been the same without you. Sarah G, Jenny, Roberta, Sarah C and Heather – thank you for not letting me forget the 'Stowe. And Felixstowe library – thank you for bending the rules to provide me with an office.

Finally, thank you to my parents for welcoming me back home for the price of only one Jamie Oliver dinner per week, and to Catherine for allowing me to convert her bedroom into my chaotic office. I may well have lost my sanity without the company, support and balanced diet which you have provided over the past five months.

Thank you.

List of common abbreviations

λs	Sibling relative risk
58BC	1958 birth cohort
ADP	Adenosine diphosphate
ATP	Adenosine triphosphate
β-cell	Beta cell
BAC	Bacterial artificial chromosome
BBC	Berlin birth cohort
BMI	Body mass index
bp	Base pairs
cDNA	Complementary deoxyribonucleic acid
ChIP-seq	Chromatin Immunoprecipitation sequencing
Chr	Chromosome
CNV	Copy number variant
ddNTPs	Dideoxyribonucleotide
dH₂O	Distilled water
DIAGRAM+	Diabetes genetics replication and meta-analysis consortium
DMR	Differentially methylated region
DNA	Deoxyribonucleic acid
dNTPs	Deoxyribonucleotide
EFSD	European foundation for the study of diabetes
EGG	Early growth genetics (consortium)
ENU	N-ethyl-nitrosourea
eQTL	Expression quantitative trait loci
fpg	Fasting plasma glucose
gDNA	genomic DNA
GSIS	Glucose-stimulated insulin secretion
GWA(S)	Genome-wide association (study)
HKG	Housekeeping gene
HOMA-B	Homeostatic model assessment of beta-cell function
HOMA-IR	Homeostatic model assessment of insulin resistance
INDEL	Insertion/deletion
Kb	Kilobase
LD	Linkage disequilibrium
LOD	Logarithm [base10] of odds
MAF	Minor allele frequency

MAGIC	Meta-analysis of glucose and insulin related traits consortium
MAQ	Multiplex amplicon quantitation
Mb	Megabase
miRNA	Micro ribonucleic acid
MODY	Maturity onset diabetes of the young
mQTL	Methylation quantitative trait loci
mRNA	Messenger ribonucleic acid
NAHR	Non-allelic homologous recombination
NCBI	National Centre for Biotechnology Information
NEFAs	Non-esterified fatty acids
NTC	Non-template control
OGTT	Oral glucose tolerance test
OR	Odds ratio
OXCIT	Oxford centre for islet transplantation
PBS	Phosphate buffered saline
PCR	Polymerase chain reaction
PI/I	Proinsulin / insulin ratio
PNDM	Permanent neonatal diabetes
QC	Quality control
QPCR	Quantitative polymerase chain reaction
RAF	Risk allele frequency
RIN	RNA integrity number
RNA	Ribonucleic acid
RT	Reverse transcriptase
SD	Standard deviation
SNP	Single nucleotide polymorphism
T1D	Type 1 diabetes
T2D	Type 2 diabetes
TBE	Tris/Borate/EDTA
TNDM	Transient neonatal diabetes mellitus
UCSC	University of California Santa Cruz
UKBS	United Kingdom Blood Service
UTR	Untranslated region
WHO	World health organisation
WTCCC	Wellcome Trust Case Control Consortium
WT	Wildtype

Contents

CHAPTER 1: General introduction

1.1	Type 2 diabetes and global health.....	2
1.2	Pathophysiology of T2D	3
1.2.1	Diagnosis and clinical presentation	3
1.2.2	Causation: insulin resistance and insulin secretion	4
1.2.2 (a)	Insulin resistance	4
1.2.2 (b)	Insulin secretion.....	6
1.3	T2D and genetics: the story so far	10
1.3.1	Genetics plays a role in T2D predisposition	10
1.3.2	Lessons from monogenic diabetes.....	10
1.3.3	A brief history of susceptibility variant discovery	11
1.3.3 (a)	Linkage analysis.....	11
1.3.3 (b)	Candidate gene association studies.....	12
1.3.3 (c)	Genome-wide association studies.....	12
1.3.4	Insight from genetic analysis so far	16
1.3.5	Where next? Challenges remaining.....	18
1.4	Finding the missing heritability.....	19
1.4.1	Copy number variation	20
1.4.1 (a)	Copy number variation as a source of heritable variation.....	20
1.4.1 (b)	Copy number variation in the human genome	21
1.4.1 (c)	Copy number variation and human disease.....	23
1.4.2	Rare variants.....	25
1.4.2 (a)	Rare variation in the human genome	25
1.4.2 (b)	Rare variation and human disease:.....	25
1.5	Understanding causal mechanisms	29
1.5.1	Fine and trans-ethnic mapping to determine causal variants.....	29
1.5.2	Functional variant and cellular analysis.....	30
1.5.3	Phenotypic refinement	31
1.5.4	Expression profiling and eQTL analysis	32
1.5.5	Methylation profiling and epigenetics.....	34
1.5.6	Animal models.....	37
1.6	Thesis Aims.....	40

CHAPTER 2: General Methods

2.1	DNA amplification	42
2.1.1	Primer design	42
2.1.2	PCR reaction components and cycling conditions	42
2.1.3	PCR product validation using agarose gel electrophoresis	45
2.2	Sequencing.....	46
2.2.1	Clean-up of PCR products	46
2.2.2	Sequencing reactions.....	46
2.2.3	Sequencing product clean-up.....	47
2.2.4	Capillary electrophoresis and analysis	47
2.3	RNA extraction.....	48
2.3.1	Human islet processing and storage for RNA extraction.....	48
2.3.2	RNA extraction from human islets.....	48
2.3.3	RNA extraction from human foetal pancreas	50
2.3.4	RNA extraction from human adult pancreas & adipose tissue	50
2.3.5	RNA quality determination	50
2.4	DNA extraction	52
2.4.1	DNA extraction from human islets and human pancreas.....	52
2.4.3	DNA extraction from human spleen and whole blood.....	53
2.5	Genotyping.....	54
2.5.1	Assay design	54
2.5.2	TaqMan PCR.....	54
2.5.3	Genotype determination	55
2.6	cDNA synthesis	56
2.7	mRNA quantification	57
2.7.1	TaqMan probe selection.....	57
2.7.2	Housekeeping gene selection.....	58
2.7.3	Standard curve generation.....	59
2.7.4	Quantitative real-time PCR	59
2.7.5	Analysis	60
2.7.6	Quantitative PCR as a method to determine human islet purity.....	62
2.8	Multiplex Amplicon Quantitation	64
2.8.1	Assay design	64
2.8.2	Amplification.....	64

2.8.3	Analysis	65
2.9	Fragment analysis	66
2.10	Quantification of DNA methylation.....	67
2.10.1	Bisulphite treatment of DNA	68
2.10.2	Primer design and PCR	68
2.10.3	Pyrosequencing reactions	69
2.10.4	Analysis	70
2.11	Statistical methods.....	71

CHAPTER 3: Functional interrogation of T2D association at the *KCNQ1* locus

3.1	Introduction	73
3.1.1	The need for functional interrogation of T2D-association loci.....	73
3.1.2	<i>KCNQ1</i> locus variants are associated with T2D & β -cell phenotypes.....	74
3.1.3	Molecular mechanisms for T2D risk at the <i>KCNQ1</i> locus are not clear ...	77
3.1.4	<i>KCNQ1</i> lies in the 11p15.5 imprinted gene cluster	78
3.1.5	Regional genes are of relevance to β -cell proliferation & foetal growth.	81
3.1.6	The <i>KCNQ1</i> locus may be relevant to birth weight.....	82
3.1.7	Experimental aims.....	84
3.2	Samples and Methods	85
3.2.1	Samples used.....	85
3.2.2	RNA and DNA extraction and quality assessment.....	87
3.2.3	SNP selection and Genotyping.....	87
3.2.4	cDNA synthesis.....	88
3.2.5	Fragment analysis.....	89
3.2.6	Sequencing.....	89
3.2.7	Total mRNA expression quantification	89
3.2.8	Allele-specific mRNA expression quantification	92
3.2.9	Islet purity analysis	92
3.2.10	Bisulphite-treatment and Pyrosequencing	93
3.2.11	Statistical procedures.....	96
3.3	Results	97
3.3.1	Genotyping of islet and foetal pancreas samples.....	97
3.3.2	Gene expression at the 11p15.5 cluster.....	99

3.3.3	Temporal differences in imprinting status.....	100
3.3.3 (a)	<i>CDKN1C</i> imprinting status.....	100
3.3.3 (b)	<i>KCNQ1</i> & <i>KCNQ1OT1</i> imprinting status	102
3.3.3 (c)	<i>PHLDA2</i> , <i>SLC22A18</i> & <i>SLC22A18AS</i> imprinting status.....	104
3.3.4	T2D risk genotype and methylation.....	106
3.3.5	T2D risk genotype and gene expression	110
3.3.6	<i>KCNQ1</i> T2D risk variants and birth weight	116
3.4	Discussion	122

CHAPTER 4: Gene expression profiling for proinsulin analysis

4.1	Introduction	130
4.1.1	Limitations of genome-wide association studies.....	130
4.1.2	Gene expression profiling as an interpretive tool	131
4.1.3	Proinsulin as a relevant trait to T2D	132
4.1.4	The MAGIC proinsulin genome-wide association study.....	134
4.1.5	Experimental aims.....	137
4.2	Samples and Methods	139
4.2.1	Samples used.....	139
4.2.2	RNA extraction	141
4.2.3	cDNA synthesis.....	141
4.2.4	mRNA expression quantification.....	141
4.3	Results.....	144
4.3.1	mRNA gene expression profiling.....	144
4.3.2	Comparisons of genes within each locus.....	152
4.4	Discussion	156
4.4.1	Comparison of data between pancreas, islets and β -cells.....	156
4.4.2	Consideration of individual loci.....	160
4.4.2 (a)	<i>LARP6</i> locus	160
4.4.2 (b)	<i>MADD</i> locus.....	161
4.4.2 (c)	<i>SGSM2</i> locus	163
4.4.2 (d)	<i>VPS13C</i> locus.....	164
4.4.2 (e)	<i>ARAP1</i> locus	166
4.4.3	Study Limitations	168

CHAPTER 5: Common structural variation and T2D risk

5.1	Introduction	170
5.1.1	The Wellcome Trust Case Control Consortium (WTCCC) CNV study....	170
5.1.2	The necessity and challenge of genotyping copy number variants.....	171
5.1.3	Experimental aims.....	173
5.1.4	Test CNVs.....	174
5.2	Samples and Methods	176
5.2.1	Samples Used	176
5.2.2	SNP genotyping.....	177
5.2.3	Breakpoint assays	177
5.2.4	Quantitative PCR for copy number analysis	179
5.2.5	Multiplex Amplicon Quantification for copy number analysis.....	181
5.3	Results	184
5.3.1	Assay optimisation and validation.....	184
5.3.1 (a)	Breakpoint assays	184
5.3.1 (b)	Quantitative PCR	185
5.3.1 (c)	Multiplex Amplicon Quantitation	188
5.3.2	Method comparison	190
5.3.2 (a)	<i>NEGR1</i>	190
5.3.2 (b)	<i>FCGR3</i>	194
5.3.3	Follow-up of T2D-associated CNVs arising from WTCCC CNV scan	197
5.3.3 (a)	CNV1943.1	199
5.3.3 (b)	CNV5583.1	203
5.4	Discussion	208
5.4.1	Genotyping method comparison	208
5.4.2	Linkage Disequilibrium between CNVs and SNPs	210
5.4.3	Copy number variation and T2D	212

CHAPTER 6: Rare single nucleotide variation and T2D risk

6.1	Introduction	215
6.1.1	Rare variation and common disease susceptibility analysis.....	215
6.1.2	A genome-wide association study for rare variants and T2D risk	217
6.1.3	Experimental aims.....	220
6.2	Samples and Methods	222
6.2.1	Samples used.....	222
6.2.2	Genotyping.....	222
6.2.3	Positive control sequencing.....	224
6.3	Results	225
6.3.1	Genotyping results.....	225
6.3.2	Imputation accuracy	229
6.3.3	T2D association analysis	234
6.4	Discussion	236

CHAPTER 7: General Discussion.....239

APPENDICES

References.....	248
Published and submitted Manuscripts.....	262

Chapter 1

General Introduction

1.1 Type 2 diabetes and global health

The prevalence of type 2 diabetes (T2D) is a growing challenge in global health. There were 346 million sufferers in 2011 (WHO, 2011a; figure 1.1), and, if current trends continue, one in ten of the global adult population will be affected by 2030 (IDF, 2012). T2D is not only a first-world problem: it is the most common non-communicable disease (NCD) in 75% of countries, and more than 80% of cases occur in low or middle income nations (WHO, 2011a).

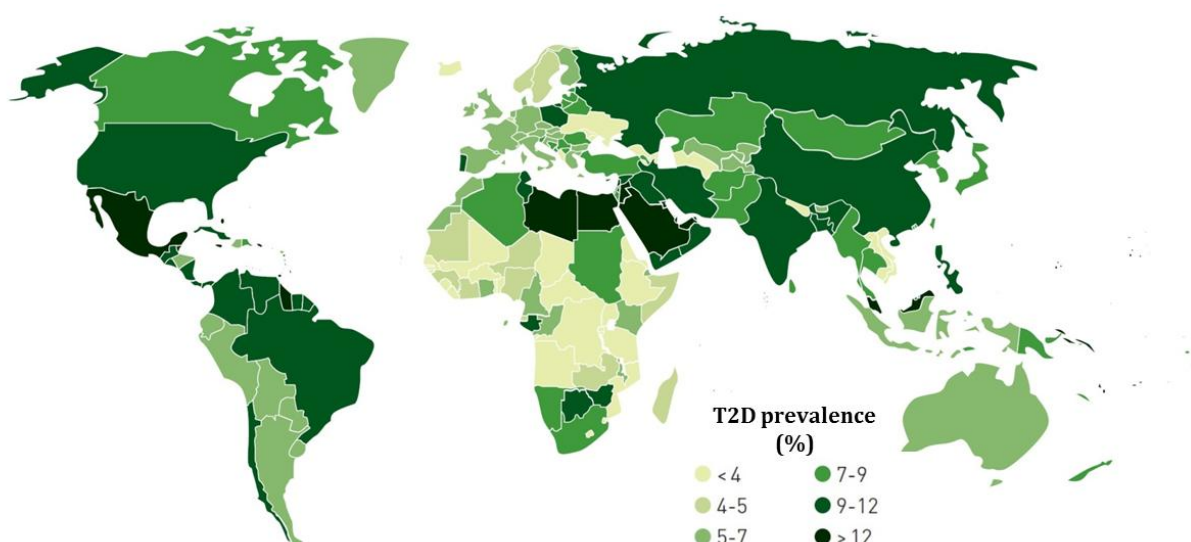


Figure 1.1: Worldwide T2D prevalence, 2011. Prevalence (%) of T2D in the adult population (20-79 years). Figure taken from the International Diabetes Federation Diabetes Atlas (<http://www.idf.org/diabetesatlas/5e/diabetes>).

T2D has substantial implications for mortality. It was responsible for approximately 4.6 million deaths in 2011; more than AIDS, malaria and tuberculosis combined and constituting 8.2% of all global mortality (WHO, 2011a). Whilst mortality rates from many other NCDs are falling, those from T2D increased by 13.3% from 2010-2011 (IDF, 2012). Diabetes and its complications are a heavy burden for healthcare systems, accounting for an estimated 11% (\$465 billion) of all worldwide healthcare expenditure in 2011 (WHO, 2011a).

1.2 Pathophysiology of T2D

1.2.1 Diagnosis and clinical presentation

T2D is a metabolic disorder, characterised by raised blood glucose in the context of insulin resistance and relative impairment of insulin secretion. Clinical diagnosis is typically made through a combination of raised fasting glucose and impaired response to glucose ingestion in the absence of the autoantibodies - predominantly glutamate decarboxylase (GAD) and islet antigen-2 (IA-2) - which destroy pancreatic β -cells in type 1 diabetes (T1D) (Sabbah et al., 2000). World Health Organisation (WHO) guidelines suggest diagnostic criteria of ≥ 7 mmol/L for fasting glucose and ≥ 11.1 mmol/L for blood glucose two hours following a glucose challenge (WHO, 2011a). They also suggest that fasting glucose $\geq 6.1 < 7.0$ mmol/L and 2-hr glucose $\geq 7.8 < 11.1$ mmol/L reflect “impaired” glycaemia and a pre-diabetic population at high risk of developing T2D. Recently, glycated haemoglobin (HbA_{1c}) has been recommended as an additional diagnostic tool, with levels $\geq 6.5\%$ indicating T2D reflecting abnormal glycaemic control over the preceding 2-3 months (ADA, 2010; WHO, 2011b).

The classic symptoms of T2D include frequent and excessive urination (polyuria), increased thirst (polydipsia), increased hunger (polyphagia), fatigue and weight change. Patients may also present with blurred vision, skin irritation, peripheral neuropathy, and, very rarely, nonketotic hyperosmolar coma. However, T2D may remain undiagnosed for many years, with mild symptoms unnoticed.

Although the long-term complications of hyperglycaemia develop gradually, they can eventually be disabling or even life-threatening – reducing life expectancy amongst

sufferers by an average of ten years (UK, 2010). Over 50% of T2D patients die from cardiovascular disease (WHO, 2011a), with a 2-3 times increased risk of heart disease and stroke (Stamler et al., 1993). Hyperglycaemic damage to blood vessels in the retina (diabetic retinopathy) and nephrons (diabetic nephropathy) make T2D a leading non-traumatic cause of blindness and kidney failure (Ripsin et al., 2009). Nerve damage (diabetic neuropathy) can cause limb ulcers, numbness and pain, and, in conjunction with peripheral vascular disease, puts T2D patients at a significantly increased risk of lower limb amputation (Gordois et al., 2003).

1.2.2 Causation: insulin resistance and insulin secretion

Despite its prevalence and severe consequences, the precise pathophysiology of T2D remains unclear. It is typically characterised by an inadequate β -cell response to increased peripheral insulin resistance (Stumvoll et al., 2008), but the mechanisms of β -cell dysfunction and impaired insulin sensitivity are not fully understood. The relative importance of insulin secretion and resistance in the onset of T2D is unclear, and likely to differ between individuals (ADA, 2010).

1.2.2 (a) Insulin resistance

Insulin is central to the regulation of carbohydrate and fat metabolism. It regulates postprandial storage and fasting release of glucose, preserving constant glucose availability. In healthy individuals, circulating insulin is detected by cell surface insulin receptors, particularly on adipose, muscle and liver cells. It causes glucose to be taken up from the bloodstream and stored as glycogen, lipolysis to be inhibited and hepatic

glucose production to be reduced. Insulin resistant cells fail to respond appropriately to circulating insulin: glucose uptake is reduced, hepatic gluconeogenesis is insufficiently suppressed, and lipolysis is increased (Dinneen et al., 1992).

The onset of insulin resistance is complex, but there is strong evidence to suggest that it is significantly related to weight gain (Zimmet et al., 2001). Visceral adipose cells produce lipolysis-inducing inflammatory cytokines such as tumour necrosis factor- α (TNF- α) (Hotamisligil et al., 1995), whilst enlarged adipocytes are less responsive to the lipolysis-inhibiting effects of insulin (Boden and Shulman, 2002).

The importance of insulin sensitivity is demonstrated by patients with severe monogenic insulin resistance disorders, most often caused by mutation of the insulin receptor gene *INSR*. Donohue syndrome is caused by recessive frameshift or nonsense mutations, and usually fatal in the first year of life (Psiachou et al., 1993). Less severe mutations of *INSR* can cause the recessive disorder Rabson-Mendenhall syndrome, characterised by developmental delay and severe hirsutism, as well as potential diabetic ketoacidosis (Longo et al., 1994). Mutations in the receptor's intracellular domain often result in milder insulin resistance (Musso et al., 2004), whilst mutation of other genes in the insulin signalling pathway – such as *AKT2* – can also cause insulin resistance syndromes (George et al., 2004).

Insulin sensitising T2D therapies combat insulin resistance by increasing the responsiveness of peripheral tissues to circulating insulin. Thiazolidinediones bind to the nuclear regulatory protein PPAR γ , upregulating the transcription of insulin-dependent enzymes to enable more efficient use of glucose by peripheral cells (O'Moore-Sullivan and Prins, 2002) – although concerns have been raised over an

accompanying increase in cardiovascular disease risk (Nissen and Wolski, 2007). The mechanism of action of biguanides (notably metformin) is not fully understood, but they are thought to activate AMPK (AMP-activated protein kinase), reducing hepatic gluconeogenesis increasing peripheral glucose uptake (Kim et al., 2008).

1.2.2 (b) Insulin secretion

Insulin resistance does not lead directly to T2D, because pancreatic β -cells can usually compensate for some insensitivity by increasing their insulin output (Stumvoll et al., 2003). Hyperglycaemia and progression to T2D occur when β -cells can no longer meet the increased insulin demand, and often show decreased secretory function.

Figure 1.2 shows the normal process of glucose-stimulated insulin secretion (GSIS). Following a postprandial increase in blood glucose in healthy individuals, glucose enters the β -cell by facilitated diffusion through glucose transporters. Intracellular glucose is phosphorylated by glucokinase (*GCK*) to form glucose-6-phosphate (Matschinsky, 2002), which is further metabolised to generate ATP. The subsequent increase in the ATP:ADP ratio causes ATP-sensitive potassium (K_{ATP}) channels on the cell membrane to close, thereby depolarising the β -cell from its resting potential of -70mV (Ashcroft et al., 1984). Membrane depolarisation opens voltage-gated calcium channels, and the subsequent calcium influx triggers insulin exocytosis from secretory vesicles.

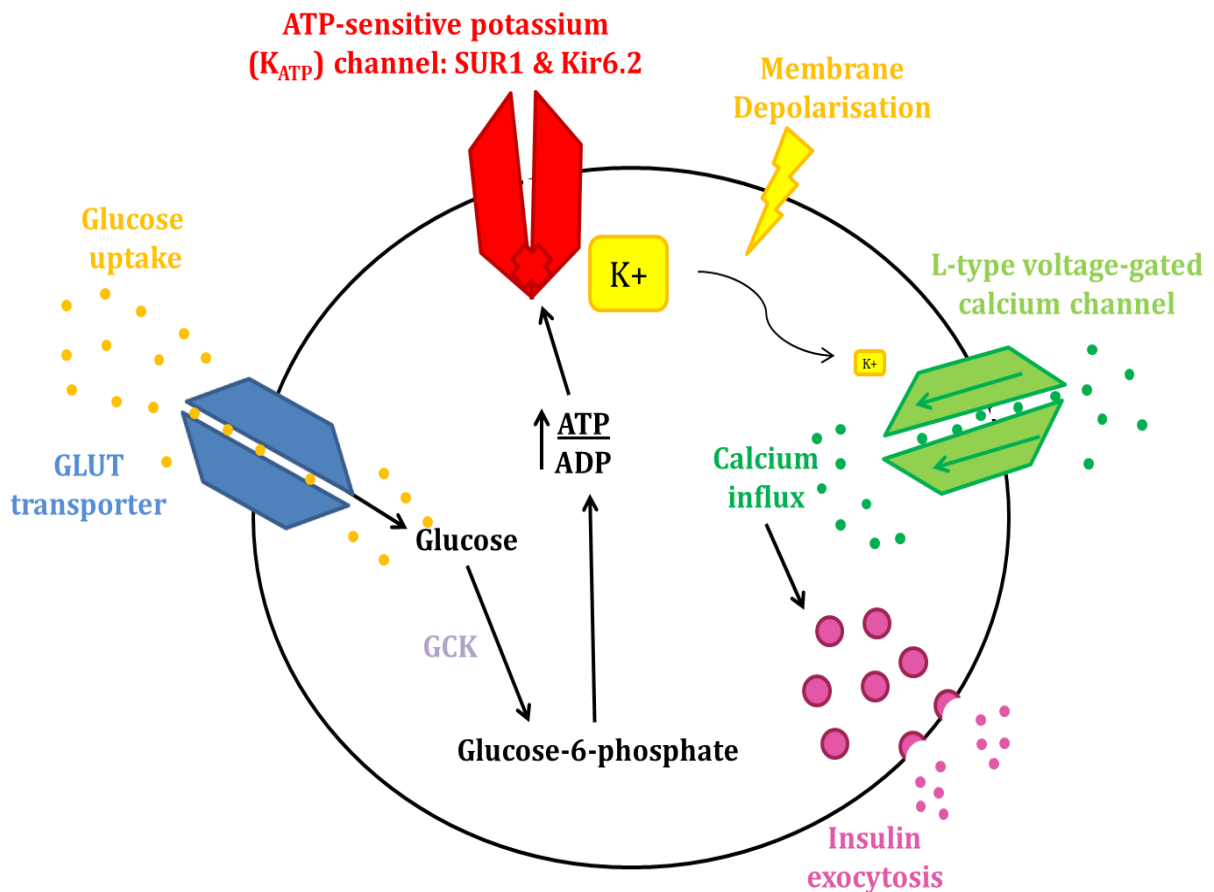


Figure 1.2: Insulin secretion from the pancreatic β -cell. Glucose enters cell via glucose transporters. Once inside, glucose is metabolised by glucokinase (*GCK*), the first enzyme in glycolysis. The product of this reaction, glucose-6-phosphate, is metabolised further, leading to an increase in the cellular ATP:ADP ratio. This leads to closure of the ATP sensitive potassium (K_{ATP}) channel, 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.

Insulin secretion could be impaired at any stage in this process, or outside it. Pancreatic islets contain α , δ , PP and ϵ cells as well as β -cells, secreting glucagon, somatostatin, pancreatic polypeptide and ghrelin respectively. Glucagon in particular is crucial to the maintenance of glucose homeostasis, raising blood glucose in response to hypoglycaemia by stimulating glycogenolysis and gluconeogenesis. The incretin hormones glucagon-like peptide-1 (GLP-1) and gastric inhibitory polypeptide (GIP) are also key components of a normal response to glucose intake. Secreted by

gastrointestinal K-cells and L-cells respectively, GLP-1 and GIP induce insulin and suppress glucagon secretion, accounting for the greater insulin response to oral than intravenously-administered glucose (McIntyre et al., 1964).

Monogenic disorders of β -cell function account for only 1-2% of all non-insulin dependent diabetes in Europe (Thanabalasingham and Owen, 2011), but their molecular mechanisms offer insight into how β -cell processes can be disrupted to impair insulin secretion. *KCNJ11* and *ABCC8* encode two subunits of the K_{ATP} channel. Heterozygous activating mutations within them are the most common cause of permanent neonatal diabetes (PNDM; Babenko et al., 2006; Gloyn et al., 2004), and rarer causes of transient neonatal diabetes (TNDM; Flanagan et al., 2007; Gloyn et al., 2005). These mutations reduce the channel's sensitivity to changing ATP:ADP ratios, inhibiting closure and membrane depolarisation (Babenko, 2008; Gloyn et al., 2004). Conversely, inactivating mutations are the most common cause of the opposite phenotype hyperinsulinaemic hypoglycaemia (HH; Kapoor et al., 2011). Here, loss of function causes the K_{ATP} channel to close and stimulate inappropriate insulin secretion despite hypoglycaemia (Thomas et al., 1996; Thomas et al., 1995).

Similar opposite phenotypes result from mutation of the glucokinase gene, *GCK*. Inactivating mutations are one of the commonest causes of Maturity-onset diabetes of the young (MODY), reducing the affinity of GCK for glucose and setting the threshold for GSIS at a slightly increased blood glucose level (Byrne et al., 1996; Froguel et al., 1992). Meanwhile, heterozygous activating *GCK* mutations are a rare cause of HH, increasing the affinity of GCK for glucose such that the threshold for GSIS is lowered, stimulating insulin secretion at inappropriately low blood glucose levels (Glaser et al., 1998).

Secretagogue T2D therapies target the pancreatic β -cell, aiming to increase insulin secretion. Sulphonylureas (such as glibenclamide) bind to and close the K_{ATP} channel, causing the β -cell to depolarise and stimulating insulin secretion. Endogenous GLP-1 is rapidly broken down by dipeptidyl peptidase-4 (DPP-4); peptide analogue secretagogue therapies therefore increase insulin secretion through the administration of DPP-4 inhibitors, and GLP-1 analogues with an increased half-life.

1.3 T2D and genetics: the story so far

1.3.1 Genetics plays a role in T2D predisposition

The current increase in T2D prevalence must be due primarily to environmental change; the timescales involved are far too short for shifts in susceptibility variant frequency. But not everyone exposed to the increasingly pervasive “obesogenic” environment seems at equivalent risk. Most estimates place the heritability of T2D between 25% and 50% (Almgren et al., 2011; Poulsen et al., 1999), and there has been little evidence of an attenuation in heritability in more recent studies (Wardle et al., 2008). An individual’s risk of developing T2D therefore reflects the intersection of inherited variation and exposure to modern environmental stressors, particularly increased energy intake and decreased physical activity (Stumvoll et al., 2008).

1.3.2 Lessons from monogenic diabetes

Thanks to their high penetrance, the alleles responsible for rare, monogenic forms of non-autoimmune diabetes were relatively easily identified through linkage analysis and candidate gene studies (reviewed in Steck and Winter, 2011). These discoveries have led to molecular classifications of disease with demonstrable prognostic and therapeutic relevance (Thanabalasingham and Owen, 2011). For example, individuals with MODY due to mutations in the transcription factor *HNF1A* respond particularly well to treatment with low dose sulphonylureas, whilst those with mutations in glucokinase (*GCK*) can often come off medication entirely given their relatively benign prognosis (Pearson et al., 2003; Stride et al., 2002).

As a consequence of such advances in genetic understanding and classification, molecular diagnostics and personalised therapy are now standard components of clinical care for patients with monogenic diabetes. An improved understanding of the genetic basis of common T2D aims to inspire similar insight into disease biology, and to underpin the development of novel therapeutic agents and biomarkers.

1.3.3 A brief history of susceptibility variant discovery

1.3.3 (a) Linkage analysis

Family-based linkage approaches, so successful in identifying the mutations responsible for monogenic and syndromic subtypes of diabetes, proved poorly suited to revealing lower penetrance variants in late-onset disease. Numerous scans identified apparently linked loci (for example Lindgren et al., 2002; Reynisdottir et al., 2003; reviewed in McCarthy, 2003), but few were replicated. The most convincing linkage signal emerged at chromosome 1q, identified initially in Pima Indians (Hanson et al., 1998) and subsequently replicated in European populations (Wiltshire et al., 2001). The linked region encompasses plausible candidate genes – such as the potassium channel gene *KCNJ10* – but fine-mapping has failed to identify any T2D-associated common variants (Prokopenko et al., 2009), and, to date, resequencing has not revealed any rarer coding variants which could account for the linkage signal (Wang et al., 2009).

1.3.3 (b) Candidate gene association studies

Attention turned from linkage analysis to association approaches in larger, unrelated samples sets (Merikangas and Risch, 2003). Association analyses rely upon typing the causal variant or a closely-correlated proxy, and so initial efforts were constrained by practical limitations of genotyping cost and capacity to the evaluation of variants within candidate genes selected on the basis of known biology. Nonetheless, this approach heralded the first wave of robust T2D associations. Non-synonymous variants in genes encoding the targets of two drugs widely used in T2D treatment (P12A in *PPARG*; Altshuler et al., 2000 for thiazolidinediones and E23K in *KCNJ11*; Gloyn et al., 2003 for sulphonylureas) showed consistent, though modest (per-allele odds ratios of ~ 1.2), evidence of association with disease risk. Subsequent analysis has suggested that the A1369 variant in *ABCC8*, in complete LD with E23K, may be a stronger candidate for mediating the T2D risk effect (Florez et al., 2004; Hamming et al., 2009).

1.3.3 (c) Genome-wide association studies

The candidate gene approach is restricted by its intrinsic reliance upon prior knowledge. When understanding of disease pathogenesis is imperfect, there is a manifest need to extend the search for susceptibility variants across the entire genome in an unbiased, hypothesis-free manner. The first gene to be implicated in T2D susceptibility without prior biological candidacy was *TCF7L2*, discovered following systematic association analysis across a region of previously-identified linkage (Grant et al., 2006). The most strongly associated variants at this locus have the greatest effect on T2D-susceptibility of any common variant so far identified.

The advent of genome-wide association studies (GWAS) proved transformative. The first wave of GWAS in 2007 confirmed the known loci at *PPARG*, *KCNJ11* and *TCF7L2*, but added a further six novel loci including signals near *CDKAL1*, *HHEX*, *SLC30A8*, *IGF2BP2* and *CDKN2A* (Saxena et al., 2007; Scott et al., 2007; Sladek et al., 2007; Steinthorsdottir et al., 2007; Zeggini et al., 2007). At the sixth locus, near *FTO* (Frayling et al., 2007), association with T2D was predicated entirely on case-control differences in adiposity, serendipitously revealing *FTO* as the first common variant signal for body mass index (BMI) and risk of obesity.

Successive rounds of GWA meta-analysis (Saxena et al., 2012; Voight et al., 2010; Zeggini et al., 2008) have brought the count of confirmed common variant signals for T2D to more than sixty (table 1.1). In European populations, the Diabetes Genetics Replication and Meta-analysis (DIAGRAM) consortium first combined data from three published GWAS to reveal six novel loci (Zeggini et al., 2008), and subsequently aggregated data from an additional five GWAS to capture a further twelve signals (Voight et al., 2010). The improvement in power derived from increasing sample size has been particularly beneficial in exposing variants of smaller effect size and more extreme risk allele frequency.

Genome-wide association analysis and meta-analysis in South and East Asian populations (Cho et al., 2012; Kooner et al., 2011; Unoki et al., 2008; Yamauchi et al., 2010; Yasuda et al., 2008) revealed a further seventeen associated loci, most of which have been replicated in European populations. A recent meta-analysis of 50,000 SNPs across candidate gene regions in Asian, Hispanic, European and African-American populations has revealed a novel susceptibility locus, as well as independently associated variants within known loci (Saxena et al., 2012).

Chrom-osome	Locus	Index SNP	RAF	OR	Population	Reference
3	<i>PPARG</i>	rs1801282	0.92	1.14	European	Altshuler (2000)
11	<i>KCNJ11</i>	rs5219	0.50	1.14	European	Gloyn (2003)
10	<i>TCF7L2</i>	rs7903146	0.25	1.37	European	Grant (2006)
10	<i>HHEX-IDE</i>	rs1111875	0.56	1.15	European	Sladek (2007)
17	<i>HNF1B</i>	rs757210	0.43	1.10	European	Winkler (2007)
16	<i>FTO</i>	rs8050136	0.45	1.17	European	Frayling (2007)
3	<i>IGF2BP2</i>	rs4402960	0.29	1.14	European	Zeggini (2007)
4	<i>WFS1</i>	rs1001013	0.32	1.12	European	Sandhu (2007)
6	<i>CDKAL1</i>	rs7754840	0.31	1.14	European	Zeggini (2007)
8	<i>SLC30A8</i>	rs13266634	0.75	1.15	European	Sladek (2007)
9	<i>CDKN2A/B</i>	rs10811661	0.79	1.20	European	Zeggini (2007)
1	<i>NOTCH2</i>	rs10923931	0.11	1.13	European	Zeggini (2008)
2	<i>THADA</i>	rs7578597	0.92	1.15	European	Zeggini (2008)
3	<i>ADAMT59</i>	rs4607103	0.81	1.09	European	Zeggini (2008)
7	<i>JAZF1</i>	rs864745	0.52	1.10	European	Zeggini (2008)
10	<i>CDC123</i>	rs12779790	0.23	1.11	European	Zeggini (2008)
12	<i>TSPAN8</i>	rs7961581	0.23	1.09	European	Zeggini (2008)
11	<i>KCNQ1</i>	rs2237892	0.61	1.40	E. Asian	Yasuda/Unoki (2008)
11	<i>MTNR1B</i>	rs10830963	0.30	1.09	European	Prokopenko (2009)
2	<i>IRS1</i>	rs2943641	0.61	1.19	European	Rung (2009)
11	<i>DUSP8</i>	rs2334499	0.45	1.08	European	Kong (2009)
2	<i>GCKR</i>	rs780094	0.62	1.06	European	Dupuis (2010)
3	<i>ADCY5</i>	rs11708067	0.78	1.12	European	Dupuis (2010)
7	<i>DGKB</i>	rs2191349	0.47	1.06	European	Dupuis (2010)
7	<i>GCK</i>	rs4607517	0.2	1.07	European	Dupuis (2010)
9	<i>PTPRD</i>	rs17584499	0.11	1.57	E. Asian	Tsai (2010)
17	<i>SRR</i>	rs391300	0.76	1.28	E. Asian	Tsai (2010)
1	<i>PROX1</i>	rs340874	0.5	1.07	European	Voight (2010)
2	<i>BCL11A</i>	rs243021	0.46	1.08	European	Voight (2010)
5	<i>ZBED3</i>	rs4457053	0.26	1.08	European	Voight (2010)
7	<i>KLF14</i>	rs972283	0.55	1.07	European	Voight (2010)

8	<i>TP53INP1</i>	rs896854	0.48	1.06	European	Voight (2010)
9	<i>CHCHD9</i>	rs13292136	0.93	1.11	European	Voight (2010)
11	<i>CENTD2</i>	rs1552224	0.88	1.14	European	Voight (2010)
11	<i>KCNQ1</i>	rs231362	0.52	1.08	European	Voight (2010)
12	<i>HMGA2</i>	rs1531343	0.1	1.10	European	Voight (2010)
12	<i>HNF1A</i>	rs7957197	0.85	1.07	European	Voight (2010)
15	<i>PRC1</i>	rs8042680	0.22	1.07	European	Voight (2010)
15	<i>ZFAND6</i>	rs11634397	0.56	1.06	European	Voight (2010)
X	<i>DUSP9</i>	rs5945326	0.12	1.27	European	Voight (2010)
2	<i>RBMS1</i>	rs7593730	0.83	1.10	European	Qi (2010)
3	<i>UBE2E2</i>	rs6780569,	0.85	1.18	E. Asian	Yamauchi (2010)
15	<i>C2CD4A/B</i>	rs7172432	0.58	1.13	E. Asian	Yamauchi (2010)
2	<i>GRB14</i>	rs3923113	0.74	1.06	S. Asian	Kooner (2011)
3	<i>ST6GAL1</i>	rs16861329	0.75	1.06	S. Asian	Kooner (2011)
10	<i>VPS26A</i>	rs1802295	0.26	1.05	S. Asian	Kooner (2011)
15	<i>AP3S2</i>	rs2028299	0.31	1.07	S. Asian	Kooner (2011)
15	<i>HMG20A</i>	rs7178572	0.52	1.06	S. Asian	Kooner (2011)
20	<i>HNF4A</i>	rs4812829	0.29	1.06	S. Asian	Kooner (2011)
3	<i>PSMD6</i>	rs831571	0.68	1.09	E. Asian	Cho (2011)
4	<i>MAEA</i>	rs6815464	0.58	1.13	E. Asian	Cho (2011)
6	<i>KCNK16</i>	rs1535500	0.42	1.08	E. Asian	Cho (2011)
6	<i>ZFAND3</i>	rs9470794	0.32	1.12	E. Asian	Cho (2011)
9	<i>GLIS3</i>	rs7041847	0.51	1.10	E. Asian	Cho (2011)
19	<i>PEPD</i>	rs3786897	0.56	1.10	E. Asian	Cho (2011)
20	<i>R3HDML-HNF4A</i>	rs6017317	0.47	1.09	E. Asian	Cho (2011)
20	<i>GCC1-HNF4A</i>	rs6467136	0.79	1.11	E. Asian	Cho (2011)
18	<i>BCL2</i>	rs12454712	0.59*	1.09	Multi-ethnic	Saxena (2012)
19	<i>GATAD2A</i>	rs19471596	0.08	1.12	European	Saxena (2012)

Table 1.1: Established common variant T2D susceptibility loci arranged by order of discovery (then chromosome). Gene name represents closest or best regional candidate gene and labelling SNP represents strongest association; neither are necessarily causal. RAF = risk allele frequency; frequency of the allele which increases T2D risk, based on HapMap information for the listed population. OR = odds ratio; per-allele increase in T2D risk. Population represents discovery group – most signals have been replicated across multiple population. *RAF averaged across multiple populations used in trans-ethnic meta-analysis.

1.3.4 Insight from genetic analysis so far

Despite its reliance on linkage disequilibrium and limitation to common variants, the unbiased genome-wide association approach has offered considerable insight into the broad pathophysiological processes of T2D. Analyses in normoglycaemic individuals have shown that most associated loci exert their primary effects on disease risk through reduced insulin secretion rather than increased insulin resistance (Perry and Frayling, 2008; Voight et al., 2010), suggesting that the β -cells' ability to compensate for increased insulin demand is central to T2D pathogenesis. Amongst loci where T2D-risk alleles are associated with reduced insulin sensitivity, only the signal at *FTO* seems to be driven by an effect on BMI.

At a subset of T2D-susceptibility loci, including *GCKR* and *SLC30A8*, there is substantial statistical and biological evidence to support specific coding variants as causal. The signal on chromosome 2p23 (Saxena et al., 2007) can be attributed to the P446L variant in *GCKR* - encoding the glucokinase regulatory protein (GKRP) - one of 17 genes mapping to the original 420kb interval of association (Orho-Melander et al., 2008). Functional characterisation has shown that the T2D risk allele alters fructose-6-phosphate-mediated regulation of GKRP, with consequences for glycolytic flux which explain the variant's effects on glucose and lipid metabolism (Beer et al., 2009; Rees et al., 2012). The R325W variant in *SLC30A8* (Sladek et al. 2007) represents a further functionally active missense polymorphism which appears causal for its local association. *SLC30A8* encodes a zinc transporter, ZnT8, known to be expressed in the pancreatic islet and implicated in the proper function of β -cell insulin granules (Nicolson et al., 2009). In mice, some studies have shown that β -cell specific knockouts of *Znt8* are glucose intolerant and display defects in insulin production, crystallization,

packaging and secretion (Wijesekara et al., 2010), although the necessity of ZnT8 for normal mouse glucose homeostasis is not certain (Lemaire et al., 2009).

Limited progress has also been made at loci for which no obvious causal coding variant can be identified. At *TCF7L2*, fine-mapping studies have converged upon the intronic SNP rs7903146 as the most compelling candidate variant (Helgason et al., 2007). FAIRE-seq (formaldehyde-assisted isolation of regulatory elements sequencing) studies have shown that rs7903146 maps within a region of islet-specific open chromatin, and that its two alleles differ in their capacity to achieve or maintain this state (Gaulton et al., 2010). *TCF7L2* mRNA levels in human pancreatic islets increase with number of risk alleles, and are five-fold higher in human islets isolated from T2D patients than those isolated from controls (Lyssenko et al., 2007).

Linkage and association studies together have demonstrated that there is significant overlap between monogenic and common diabetes genes. Rare mutations in *KCNJ11*, *GCK*, *HNF1A*, *PPARG* and *WFS1* cause monogenic and syndromic diabetes, whilst common variants in the same genes contribute to T2D susceptibility. In a review of 891 GWA studies, at least 268 of the genes implicated in complex disease susceptibility also carry mutations causing rare Mendelian disorders (Lupski et al., 2011) - suggesting a wide allelic spectrum in which common, low penetrance and very rare, high penetrance variants at the same locus can independently influence disease susceptibility. The possibility of an intermediate set of rarer variants with intermediate penetrance will be addressed in section 1.4.

1.3.5 Where next? Challenges remaining

Candidate association, genome-wide association and meta-analysis have together identified more than 60 loci conferring susceptibility to T2D. However, effect sizes at all of these loci are modest, and, taken together, they explain a small proportion (~10%) of the familial aggregation of T2D. The 37 loci confirmed by 2010 produce a sibling relative T2D risk (λ_s) of 1.16 (Voight et al. 2010) - well below epidemiological λ_s estimates of 2-3 (Köbberling, 1982). Furthermore, disease mechanisms at the majority of established susceptibility loci remain unclear. A list of associated SNPs is of little value unless they can be used to improve models of disease biology, by identifying causal variants and unravelling the mechanisms by which those variants exert their effects on disease predisposition.

The challenges to understanding molecular mechanisms of genetic T2D susceptibility are twofold: to identify the missing majority of genetic variance, and to determine the causal mechanisms at disease-associated loci. Work presented in this thesis addresses both of these challenges, and possible strategies are outlined in sections 1.4 and 1.5 respectively.

1.4 Finding the missing heritability

It is possible that epidemiological analyses have overestimated the heritability of T2D – for example by failing to account for a more similar prenatal environment amongst monozygotic than dizygotic twins - or that unexplored epistatic interactions between known loci will account for some of the heritability currently thought to be missing (Zuk et al., 2012). Nevertheless, there is clearly a significant amount of genetic variation conferring T2D susceptibility yet to be identified. The scale and power of recent GWA meta-analyses mean that, as far as common variants are concerned, it is likely that future discoveries in European populations will be limited to alleles of small effect size. Identifying the variants underlying further genetic variance is likely to require the extension of genome-wide surveys across a wider diversity of ethnic groups and relevant traits, and analysis of a more complete range of variation types across the full allele frequency spectrum (Manolio et al., 2009).

Most early GWA studies involved individuals of European descent, but a growing number of discoveries are being made in other ethnic groups (Cho et al., 2012; Kooner et al., 2011; Saxena et al., 2012; Unoki et al., 2008; Yamauchi et al., 2010; Yasuda et al., 2008). Differences in allele frequency and effect size influence the likelihood of initial discoveries being made in one or other ethnic group (Myles et al., 2008), but almost all common variant signals so far examined have a consistent effect on T2D risk across multiple ethnic groups (Waters et al., 2010). Common variant GWAS in diverse populations therefore represent a promising avenue to uncovering susceptibility variants relevant to all populations – although the small effect sizes of common variants are likely to account for a limited proportion of overall genetic variance.

Analysis of T2D-relevant traits in the normal population has already revealed novel disease loci (Dupuis et al., 2010), and a better understanding of the interaction between genetic and environmental risk factors may help to reveal how common variant risk effects vary across populations (Helgadóttir et al., 2006). However, it is unlikely that common variation will account for all genetic susceptibility to T2D. Two alternative sources of variation, copy number and rare variants, are investigated in chapters five and six and discussed in more detail below.

1.4.1 Copy number variation

1.4.1 (a) Copy number variation as a source of heritable variation

A copy number variant (CNV) is a deletion or duplication of a genomic region. The human default copy number is two; regions of zero or one copy therefore represent homozygous or heterozygous deletions respectively, and regions of more than two copies reflect duplication. The size definition of copy number variation is not straightforward. Small genomic deletions are conventionally termed 'indels' (insertion/deletions) and small repeated sequence motifs micro-and mini-satellites, whilst very large deletions or duplications are considered structural or chromosomal aberrations. Copy number variation occupies the intermediate size bracket, defined by the database of genomic variants (DGV) as larger than 1kb but smaller than 1Mb.

There are two main ways in which deletion or duplication can create a CNV. Non-allelic homologous recombination (NAHR) may occur between regions with a high degree of sequence homology - notably segmental duplications - but which are not homologous alleles. This results in the gain or loss of regions adjacent to the recombination event

(Chen et al., 1997), and is thought to be one of the major mechanisms of CNV-formation (figure 1.3). CNVs are often flanked by segmental duplications or other repeated sequence motifs (Sharp et al., 2005), and, in detailed sequencing of the breakpoints of 350 CNVs, NAHR was the likely mechanism of formation for around 25% of variants (Conrad et al., 2010).

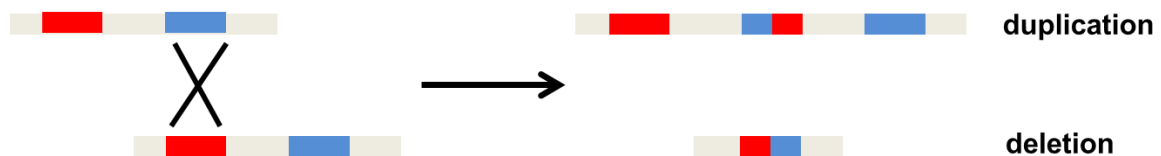


Figure 1.3: Non allelic homologous recombination. Schematic representation of NAHR, generating duplications and reciprocal deletions by unequal alignment of homologous chromosomes.

Copy number variation can also be formed where there is minimal homology. When the exposed single DNA strand formed during replication contains short lengths of sequence identity, replication slippage can delete or duplicate sequence between the identical regions (Hastings et al., 2009). Similarly, the remaining single-stranded template at a collapsed replication fork in a stressed cell will anneal to any nearby single stranded end with which it shares micro-homology (Arlt et al., 2009), deleting the intervening sequence.

1.4.1 (b) Copy number variation in the human genome

Many discovery studies have suggested that copy number polymorphism is a significant source of variation in the human genome. Perhaps the largest estimate of its significance came from an analysis of 270 individuals, identifying 1,447 copy number variable regions encompassing 360Mb – 12% of the entire genome and more nucleotides than are involved in SNP variation (Redon et al., 2006). Other

contemporary studies identified 221 copy number variants in 20 individuals (Sebat et al., 2004), 586 copy number variable regions in 189 HapMap families (Conrad et al., 2006), and 112 copy number differences between the human reference sequence and a single alternative genome (Tuzun et al., 2005), encompassing 25, 21 and 15 megabases of sequence respectively.

Most of these discovery studies used bacterial artificial chromosome (BAC) clones, into which large fragments (80–200kb) of genomic DNA were inserted. They therefore lacked precision in identifying variant breakpoints, so are likely to have overestimated the total size of copy number variable sequence. A later study used intensity data from SNP genotypes to map copy number variation at 2kb resolution in the same 270 HapMap samples used by Rendon et. al. (McCarroll et al., 2008b). This more refined analysis confirmed copy number variation at the majority (82%) of regions identified by Rendon, but the variant sequences were 5-15 times smaller than those determined by BAC analysis – affecting a reduced (2-4%), but still significant, proportion of the genome.

McCarroll's analysis (McCarroll et al., 2008b) showed that, like single nucleotide variation, variation due to copy number arises from a combination of rare and common alleles. The majority of CNVs are rare, but 85% of copy number differences between any two individuals arose from a limited set of common ($MAF > 0.05$) variants, and only 8% from rare or private CNVs ($MAF < 0.01$). The majority of copy number variation is therefore likely to be derived from inheritance rather than *de novo* mutation, and a limited set of common CNVs will capture most copy number differences between individuals.

1.4.1 (c) Copy number variation and human disease

Even at conservative estimates, any two individuals of European descent can be expected to differ in copy number at 175 sites, spanning 5.8Mb of the genome and overlapping the transcribed regions of around 100 genes (McCarroll et al., 2008b). CNVs therefore have significant potential for effects on disease susceptibility.

As for single nucleotide variation, the first evidence of a role for copy number variation in human disease came from monogenic disorders. For example, large deletions on Chr17.2 which disrupt the transcription factor *HNF1B* are a major cause of MODY (Bellanne-Chantelot et al., 2005), whilst paternally derived duplications of 6q24 are responsible for transient neonatal diabetes (TMDM; Gardner et al., 1998).

The potential role of CNVs in altering gene dosage and regulatory landscapes is clear. Large variants, in contrast with SNPs or small indels, may affect entire protein-coding genes and substantial regulatory regions. They may influence mRNA expression levels, deleting entire transcripts or increasing gene dosage through duplication. Parts of transcripts may be affected, producing truncated and non-functional proteins. Large variants may duplicate or delete transcriptional regulatory elements, or bring a gene into a new regulatory environment by deleting intervening sequence (Cooper et al., 2007). Studies in mice have shown both that the expression of genes within CNVs correlates with copy number, and that the presence of copy number variation can alter expression patterns across megabases of the genome (Henrichsen et al., 2009).

There have already been some successes in relating CNVs to complex disease susceptibility – particularly for neuropsychiatric disorders. For example, deletions at chromosome 15q13.3 are enriched in patients with autism and schizophrenia

(Pagnamenta et al., 2009; Stefansson et al., 2008), whilst there is a seven-fold and eight-fold genome-wide enrichment for *de novo* CNVs in sporadic autism and schizophrenia cases respectively (Sebat et al., 2007; Xu et al., 2008). A genome-wide analysis concluded that 15-20% of schizophrenia cases can be attributed to the greater burden of CNVs amongst patients (Walsh et al., 2008).

Rare copy number variation has also been implicated in susceptibility to other complex diseases. Obese individuals carry a heavier burden of large (>1Mb) CNVs – particularly deletions – than their normal weight counterparts (Wang et al., 2010a), whilst rare deletions close to *GPRC5B*, *IRGM* and *UGT2B17* have been implicated in BMI variation, Crohn's disease and osteoporosis respectively (McCarroll et al., 2008a; Speliotes et al., 2010; Yang et al., 2008).

Common CNV associations have been less forthcoming, but deletions close to *NEGR1* (MAF 0.36) and including *LCE3B* (MAF 0.4) have been implicated in BMI variation and psoriasis susceptibility (de Cid et al., 2009; Willer et al., 2009). An analysis in Korean populations identified a 200kb deletion in the leptin receptor gene *LEPR* associated with T2D risk and raised fasting glucose (Jeon et al., 2010), although the association has not been replicated. Chapter five will present a genome-wide association analysis to reveal the influence of common copy number polymorphism on complex diseases including T2D. In 2008, when the study was planned, common copy number polymorphism remained a largely unexplored potential source of disease susceptibility.

1.4.2 Rare variants

1.4.2 (a) Rare variation in the human genome

Rare ($MAF < 0.005$) and low frequency ($MAF < 0.05 \geq 0.005$) variants are the major source of single-nucleotide variation in the human genome. Rare variants already make up 38% of all single nucleotide variation documented in the HapMap database (Gorlov et al., 2011), and exhaustive sequencing projects with the capacity to discover a reasonable proportion of rare and low frequency variation are in their infancy.

The initial stages of the 1000 Genomes Project demonstrated the extent of undocumented rare variation in the human genome (Consortium, 2010). In its pilot phase, virtually all SNPs identified with $MAF > 0.05$ were already listed in dbSNP but almost all variants with a $MAF < 0.005$ were novel, with a consistently negative relationship between MAF and proportion of novel variants. The mean MAF of previously discovered variants was 0.26, whilst the mean MAF of novel variants was 0.02. The completed 1000 Genomes project is expected to identify almost all low frequency variants, and a significant proportion of rare variants (Consortium, 2010). However, it is likely that many sites harbouring private ($MAF < 0.00001$) or lineage-specific variants will remain undiscovered - even in this large cohort.

1.4.2 (b) Rare variation and human disease:

The common disease-common variant (CD-CV) hypothesis (Reich and Lander, 2001) holds that genetic susceptibility to common disease is likely to be primarily influenced by common alleles. Late onset diseases are likely to have minimal effects on reproductive fitness, and mildly deleterious alleles can rise to moderate frequency in

populations which have undergone recent expansion. The frequency of alleles which predispose to T2D may also have risen as a consequence of advantages which they conferred in the evolutionary past – the ‘thrifty genotype’ hypothesis (Neel, 1962). These models have been partially vindicated by the emergence of disease-associated common variants from GWA studies, but there is a growing body of evidence to suggest that rare variants are also an important component of the genetic architecture of common human disease (Eyre-Walker, 2010; Gorlov et al., 2011; Manolio et al., 2009).

Evolutionary arguments also suggest that variants which strongly promote disease should be subject to purifying selection, and therefore restricted to low allele frequencies. However, this does not mean that rare variants are unimportant for overall disease susceptibility. In a model of complex disease architecture in which variants influencing disease susceptibility also affect fitness, variants with larger effect sizes become rarer, but also contribute more overall susceptibility variance than common variants of small effect size (Eyre-Walker, 2010). The relevance of evolutionary models to a late-onset disease may be questionable, but there is empirical evidence to suggest that rare SNPs are more likely to be functional and confer a larger effect size than common ones (Gorlov et al., 2011). Across all variants in dbSNP, there is a negative correlation between MAF and proportion of SNPs predicted to be functional, and between MAF and proportion of non-synonymous SNPs predicted to be protein damaging (Gorlov et al., 2011).

Whilst the consistency of T2D-associated common variants across populations (Tan et al., 2010; Waters et al., 2010) suggests that rare variants are unlikely to underlie common variant signals by producing synthetic associations (Dickson et al., 2010), the

potential of these rare alleles with large (but not Mendelian) effect sizes for playing a significant role in complex disease susceptibility should not be underestimated. The human population explosion means that most genomic variation is rare and of recent origin, and has been subject to relaxed purifying selection. The total mutational burden in an individual, primarily comprising recently arisen rare mutations, may therefore have a greater influence on disease susceptibility than ancient and common variation with effect sizes estimated from population risk (Lupski et al., 2011).

It is therefore possible that a large proportion of the unidentified genetic variance will be attributable to rare and low frequency variants with intermediate effect sizes, which have been largely invisible to linkage and existing genome-wide association studies. Whilst fully penetrant rare variants causing Mendelian disorders are likely to have been revealed by linkage analysis, the penetrance of intermediate effect size variants is too low for Mendelian segregation. GWAS are powered to detect small effect sizes amongst common variants, but do not adequately cover variants with a MAF <0.05 . Whereas over 800 common (MAF 0.05-0.5) variants with allelic odds ratios between 1.05 and 1.10 would be needed to explain the heritability of T2D (as estimated from epidemiology), the inclusion of just 30 rare (MAF 0.01) variants of intermediate effect size (odds ratio 3) could account for it all (Janssens et al., 2006; Pawitan et al., 2009).

Large-scale whole genome sequencing is only just underway, but exome sequencing of candidate genes has already revealed rare and low frequency variant associations with common disease. Resequencing of three genes which cause monogenic low high-density lipoprotein-cholesterol (HDL-C) identified an excess of rare variants in complex low HDL-C phenotypes, whilst genotyping of candidate variants in the MODY

gene *HNF4A* revealed a low frequency (MAF 0.03) coding variant which confers (OR 1.2) increased T2D risk (Cohen et al., 2004; Jafar-Mohammadi et al., 2011). Resequencing of *MTNR1B*, in which common variants contribute to T2D susceptibility, identified 40 rare variants (average MAF <0.001) with a mean T2D-risk odds ratio of 3.31 (Bonnetfond et al., 2012). Four of these variants resulted in complete loss of melatonin binding and signalling capabilities, with an average odds ratio of 5.67. Exome resequencing of GWAS candidates has also revealed rare variants with a role in asthma and inflammatory bowel disease susceptibility (Rivas et al., 2011; Torgerson et al., 2012), and whole exome sequencing identified four rare variants in the vitamin D-activating gene *CYP27B1* associated with multiple sclerosis risk (Ramagopalan et al., 2011).

These studies provide evidence that common and rare variation in the same gene can contribute to complex disease risk, and that rare variants undetected by existing GWA studies can have large (but not monogenic) effect sizes. Rare variation represents a significant potential source of T2D susceptibility, and chapter six will present an investigation into its influence.

1.5 Understanding causal mechanisms

For most of the common variant loci revealed by GWA studies, the transition from association signal to causal mechanism has proved far from straightforward. It has been challenging, because of the extensive linkage disequilibrium in most human populations, to tie association signals to a single causal variant. With so many signals mapping to non-coding sequence, it has been difficult to define which of many regional transcripts is likely to mediate associations. Both issues represent obstacles to the translation of genetic discoveries into an improved understanding of disease biology. The approaches outlined below aim to identify causal variants, and connect them to the genes and disease pathways which they modulate.

1.5.1 Fine and trans-ethnic mapping to determine causal variants

Genome-wide association studies make use of linkage disequilibrium (LD) to limit the number of variants which must be typed to capture the majority of common variation across the human genome. However, LD also means that associated regions may be large and encompass many candidate genes. Detailed genotyping of additional variants within associated loci can help to refine the region of association, and, ideally, define the functional variant. In a notable success story, genotyping 104 additional variants within the T2D association signal on chromosome 2p23 in European populations narrowed the field from a 417-kb region, encompassing 17 genes and represented by an intronic SNP, to a single missense variant (rs1260326; P446L) within the glucokinase regulatory protein gene *GCKR* (Orho-Melander et al., 2008).

The differing allele frequency and linkage disequilibrium patterns between populations mean that trans-ethnic fine mapping, particularly in genetically diverse African populations, offers new opportunities to close in on causal variants. For example, analysis in a West African population determined which of the three markers in *TCF7L2* associated with T2D in Europeans was most likely to be the causal variant – or its closest correlate. In European populations, the markers fell in a 64-kb block of strong LD and were genetically indistinguishable, but weaker LD in West Africans meant that disease association was replicated only at rs7903146 (Helgason et al., 2007). More recently, a GWAS for coronary heart disease in African-Americans has identified five loci for which the predominant association signal is at SNPs strongly correlated with the index SNP in HapMap CEPH populations, but only weakly so in HapMap YRI individuals (Lettre et al., 2011).

1.5.2 Functional variant and cellular analysis

Where fine-mapping is able to determine a causal coding variant, in-vitro comparison of wild-type and variant proteins can reveal the mechanism through which genetic variation impairs protein function. The number of T2D associations which have been attributable to coding variants and therefore amenable to functional analysis is limited, but the approach has been successful in determining disease mechanisms for the P446L, R325W and E23K/S1369A variants in *GCKR*, *SLC30A8* and *KCNJ11/ABCC8* respectively (Beer et al., 2009; Hamming et al., 2009; Nicolson et al., 2009; Rees et al., 2012).

Functional characterisation of P446L-variant protein (associated with reduced T2D risk) has revealed an attenuated response to fructose-6-phosphate resulting in increased glucokinase activity and reduced blood glucose (Beer et al., 2009), and altered cellular localisation (Rees et al., 2012). The R325W-variant protein (associated with increased T2D risk) shows reduced zinc transport activity (Nicolson et al., 2009), suggesting a T2D risk mechanism in defective insulin crystallisation and secretion, whilst recombinant human K_{ATP} channels carrying the T2D-associated E23K/S1369A variants display reduced ATP inhibition (Hamming et al., 2009).

Whereas recombinant protein analysis is restricted to coding variants, in-vitro analysis of cellular phenotypes may be illuminating for associated loci both coding and non-coding regions, and particularly where precise causal variants have not been determined. For example, a detailed functional characterisation of human islets revealed how T2D-associated variants alter islet function. Variants at *TCF7L2* and *ADRA2A* were also associated with reduced glucose-stimulated insulin secretion, whereas susceptibility variants near *KCNJ11* and *KCNQ1* were associated with reduced exocytosis in response to depolarisation (Rosengren et al., 2012).

1.5.3 Phenotypic refinement

Functional analysis interrogates variant effects at a highly detailed level, but the broader physiological effects of T2D-associated variants at a population level can also be instructive. Genome-wide analysis of normal variation in insulin and fasting glucose levels demonstrates that the mechanisms influencing physiological and pathophysiological variation in glucose homeostasis are only partially overlapping

(Dupuis et al., 2010), but the effect of T2D-associated variants on insulin resistance or insulin secretion can help establish their biological relevance.

In a meta-analysis of European GWA studies (Voight et al., 2010), the majority of T2D-associated variants were also associated with reduced fasting insulin and β -cell dysfunction, suggesting a mechanism of action linked to insulin production, processing or secretion. Risk alleles at only four loci (*PPARG*, *FTO*, *IRS1* and *KLF14*) were associated with raised fasting insulin and a primary effect on insulin sensitivity. The attenuation of *FTO*'s T2D association upon correction for BMI remains unique, suggesting that it is the only one of the known T2D signals driven by a causal association with obesity. Even in the absence of established variants or transcripts therefore, physiological analysis of T2D-related traits can prove invaluable in determining which of the broad pathogenic pathways any particular associated locus may be working through.

1.5.4 Expression profiling and eQTL analysis

The lead SNPs at most established T2D-associated loci are non-coding (intronic or intergenic); many lie kilobases away from the nearest coding sequence, and the loci which they represent may encompass many plausible candidate genes, or no genes. Their effects are likely to be through regulation of gene expression, but the relevant gene and coding exons could lie hundreds of kilobases away and outside the region of maximal association (Birney et al., 2007). Expression profiling and expression quantitative trait locus (eQTL) analysis are invaluable tools in defining the range of potentially causal transcripts and in determining gene expression effects.

On the premise that only genes expressed in tissues relevant to a disease phenotype can mediate a susceptibility effect, expression profiling of candidate genes within or close to a region of disease association is valuable. It can eliminate a gene from consideration by demonstrating that it is not expressed in relevant tissues (though concerns remain about temporal specificity of expression), or highlight a gene as particularly interesting if it is expressed strongly or specifically in those same tissues. Chapter four will use expression profiling as a tool to investigate causal genes at loci associated with a T2D-relevant trait.

By tying together genotype and expression data, eQTL analysis has the potential to reveal a disease-associated SNP's effect on expression of a specific gene – ideally, transcript levels of which are associated with the disease phenotype. This may be the case for the intronic T2D SNP rs7903146, which has also been shown to associate with expression of *TCF7L2* (Lyssenko et al., 2007). *TCF7L2* mRNA levels are five-fold higher in human islets isolated from T2D patients than those isolated from controls, and over-expression of *TCF7L2* leads to reduced glucose-stimulated insulin secretion (Lyssenko et al., 2007). A pathogenic increase of *TCF7L2* expression is therefore a possible T2D susceptibility mechanism for rs7903146 or a closely linked causal variant.

Expression QTL analysis was conducted for rs7903146 and *TCF7L2* because rs7903146 lies within an intron of the gene. Where associated variants have no such obvious candidate genes for an expression effect, the situation is more challenging. Genome wide expression data in conjunction with genotype information for human islets or β -cells would be invaluable, but the challenges of sample collection mean that this resource is not yet available. Genome-wide eQTL analysis for T2D-associated variants has instead been conducted using data from blood and adipose tissue (Small

et al., 2011; Voight et al., 2010). Interesting effects have emerged - notably the association of the T2D risk variant rs972283 with *KLF14* expression in adipose tissue – but, given the emerging importance of the β -cell in T2D pathogenesis, genome-wide expression data in islets or purified β -cells are likely to provide the most instructive eQTL analysis for the majority of T2D-associated variants.

1.5.5 Methylation profiling and epigenetics

The relevance of epigenetics to diabetes was first highlighted from a genetic perspective by the monogenic disorder transient neonatal diabetes (TNDM). The majority of TNDM cases are caused by imprinting abnormalities on chromosome 6q24. Like most imprinted regions, 6q24 contains a small (~1kb) differentially methylated region (DMR), which overlaps a promoter shared between the imprinted genes *PLAGL1* and *HYMA1*. In healthy individuals, the DMR is unmethylated on the active paternal chromosome and methylated on the silenced maternal allele (Gardner et al., 2000). In TNDM patients however, paternal uniparental disomy of chromosome 6, paternal duplication of 6q24 or maternal hypomethylation of the DMR lead to overexpression of *PLAGL1* and *HYMA1* (Mackay and Temple, 2010). It is not clear which of the two genes is responsible for TNDM pathogenesis, but the zinc finger protein *PLAGL1* is a particularly good candidate. It has an alternative, unimprinted promoter, and a switch between the two could account for the postnatal remission characteristic of TNDM. Recent parent-of-origin specific association analysis has highlighted the potential relevance of imprinting to T2D (Kong et al., 2009). Variants at two loci within *KCNQ1* confer their risk effect only when maternally inherited, whilst

the same variant within *DUSP8* is associated with both increased and decreased T2D, when inherited paternally and maternally respectively.

Methylation can also influence the expression of non-imprinted genes. Approximately 70% of all CpG dinucleotides in the human genome are methylated, and most of the unmethylated sites are specific regions of CpG-rich sequence (CpG islands) close to coding sequence. CpG islands act as promoters for around half of all mammalian genes (Simmons, 2007), and their methylation level influences transcriptional activity. In general, higher methylation represses transcription - both directly by inhibiting transcription-factor binding, and indirectly by modifying histone and chromatin conformation.

Epigenetic influences on gene expression could be stably inherited or environmentally determined. Genome-wide epigenetic reprogramming occurs during gametogenesis, implicating early foetal development as a period susceptible to environmental influence. The Barker-Hales (or 'Foetal origins') hypothesis (Hales et al., 1991) suggests that intrauterine programming in response to maternal malnutrition might be responsible for the link between low birth weight and increased lifetime risk of T2D. Reduced insulin-mediated foetal growth as a result of alleles which predispose to T2D in adulthood (the foetal insulin hypothesis; Hattersley and Tooke, 1999) might also account for this relationship, but there is substantial evidence that early environment can adversely influence complex disease outcomes in later life - and that epigenetic regulation of gene expression may mediate the effects (Heijmans et al., 2008). DNA methylation of the insulin-like growth factor gene *IGF2* is altered in individuals born during acute famine (Heijmans et al., 2008), whilst poor nutrition in pregnant rats

leads to epigenetic silencing and subsequently reduced expression of the established MODY and T2D susceptibility gene *Hnf4a* (Sandovici et al., 2011).

The interaction of genotype, epigenotype and environment is poorly understood, but epigenetic change may provide a mechanism through which environmental exposure is translated into pathogenic phenotype. However, this type of environmentally determined epigenetic influence is unlikely to underlie the disease associations identified through GWA studies: a stable effect of genotype on epigenetic markers is more likely to be responsible. For example, genomic variation within regulatory regions could alter their susceptibility to methylation, thus influencing gene expression. The promoter of *PPARGC1A*, a regulator of mitochondrial gene expression with a role in ATP production, is twice as highly methylated in islets isolated from T2D patients than normoglycaemic donors, and *PPARGC1A* expression is reduced in T2D islets (Ling et al., 2008). There is an association between haplotype and promoter methylation, although none of the SNPs on the high methylation haplotype has reached genome-wide significance for association with T2D susceptibility.

This type of genotype-epigenotype interaction is amenable to QTL analysis in the same way as expression. An integration of SNP and DNA methylation data (from whole blood) at established T2D loci identified increased methylation on the obesity and T2D susceptibility haplotype at *FTO*, driven by the co-ordinated phase of CpG-creating SNPs across the risk haplotype (Bell et al., 2010). However, like expression analysis, genome-wide methylation QTL (mQTL) analysis is restricted by sample availability. The first comprehensive methylation profiling in pancreatic islets from T2D and non-diabetic donors was published in January 2012, and 276 CpG sites in the promoters of 254 genes with differential methylation diabetic and non-diabetes islets were

identified (Volkmar et al., 2012). These differences were not present in blood cells and not induced in non-diabetic islets by experimental exposure to high glucose, suggesting that they may be a cause rather than a consequence of hyperglycaemia. Differentially methylated sites were overrepresented within and close to genes involved in β -cell function and survival, but did not overlap with established T2D susceptibility loci.

Epigenetic changes are largely unexplored as potential causal mechanisms for T2D susceptibility. Chapter three will investigate the potential epigenetic effects of disease-associated SNPs at the imprinted *KCNQ1* locus.

1.5.6 Animal models

Genetic manipulation in animals allows the biological role of genes implicated in human T2D susceptibility to be investigated in far more detail than would be possible in human samples. There are many established rodent models of T2D, including monogenic insulin resistance (for example the ob/ob mouse; Ingalls et al., 1950), monogenic impaired insulin secretion (for example the Akita mouse; Yoshioka et al., 1997) and polygenic hyperinsulinaemia and insulin resistance (for example the Kuo Kondo mouse; Nakamura and Yamada, 1967).

In addition to rodents' ease of breeding and environmental control, the rodent genome can be manipulated to reveal the biological functions and relevant tissues for disease genes of interest. Whole genes can be knocked out – either universally, or in a tissue specific manner using the site-specific Cre-recombinase system (Akagi et al., 1997).

The International Knockout Mouse Consortium (IKMC, 2012) aims to derive mouse strains with universal knockouts of every protein coding gene, providing a public resource for understanding the biology of any gene (with a mouse homolog) identified through GWAS, CNV or rare variant analysis.

Conditional knockdown in rodent models can reveal gene function in specific tissues. For example, GSIS is impaired in β -cell specific insulin receptor knockout (β IRKO) mice, suggesting that insulin has an autocrine function which stimulates its own secretion via β -cell insulin receptors (Kulkarni et al., 1999). Similarly, impaired glycogen synthesis and GSIS in liver-specific *Gck* knockout mice suggests that hepatic glucokinase activity is relevant to T2D risk (Postic et al., 1999). Conditional knockout mice were invaluable in tying the T2D susceptibility effect of *Slc30a8* to the β -cell, by demonstrating in some studies that β -cell-specific inactivation of *Slc30a8* results in glucose intolerance (Wijesekara et al., 2010).

Rodent models also allow phenotypic assessment of point mutations within a gene of interest. Archives of N-ethyl-N-nitrosourea (ENU)-mutagenised spermatogonia (Justice et al., 1999) can be used to breed mice with relevant point mutations, whose insulin secretion, glucose tolerance and other T2D-relevant phenotypes can be quantified. Analysis of point mutations in the obesity and T2D-associated *Fto* gene demonstrated that reduced gene function leads to reduced fat mass (Church et al., 2009), whilst overexpression increases food intake and results in obesity (Church et al., 2010).

Care must be taken in extrapolating rodent to human physiology. For example, there is currently no evidence of a relationship between risk genotype and *FTO* expression in human adipose tissue (Grunnet et al., 2009), and variants at *FTO* are equally prevalent

in lean and obese humans (Meyre et al., 2010). Particular caution should be exercised for islet phenotypes, given evidence that human and rodent islets differ considerably in their β -cell proportion and arrangement – with β -cells being dispersed throughout a human islet in close proximity to blood vessels, rather than clustered in a central core (Bosco et al., 2010; Cabrera et al., 2006). There is also evidence that human and rodent islets differ in their predominant glucose transporters and ion channel composition (Braun et al., 2008; McCulloch et al., 2011). However, animal models offer an effective approach for establishing the biological relevance of genes implicated in T2D susceptibility.

1.6 Thesis Aims

This introduction has outlined the importance of determining the molecular mechanisms of T2D, the progress made to date and the challenges remaining. Section 1.4 outlined approaches to identifying the outstanding heritable component of disease risk, whilst section 1.5 outlined methods used to determine causal mechanisms at known loci. The heterogeneity of techniques demonstrates the necessity of a multidisciplinary approach to investigating the genetic component of T2D and its related complications.

Work presented in this thesis uses a combination of approaches to address the challenge of determining the molecular mechanisms of T2D susceptibility. Chapters three and four aim to determine causal mechanisms at established susceptibility loci, whilst chapters five and six look for new variants conferring disease susceptibility in under-investigated forms of genetic variation. More specifically:

- chapter three aims to determine T2D susceptibility mechanisms at the imprinted *KCNQ1* locus by interrogating effects of risk genotype upon gene expression and methylation;
- chapter four aims to highlight or eliminate genes from candidacy for an effect on proinsulin levels, using gene expression profiling in relevant tissues;
- chapter five aims to identify heritable components of T2D susceptibility by replicating results from an analysis of copy number variation;
- chapter six aims to identify rare variants associated with T2D susceptibility through follow-up genotyping of a rare variant imputation analysis.

Chapter 2

General Methods

2.1 DNA amplification

DNA was amplified to perform sequencing, pyrosequencing and fragment analysis of islet gDNA and cDNA (chapter three), to determine copy number status via breakpoint assays (chapter five), and to confirm rare variant heterozygosity via sequencing (chapter six).

2.1.1 Primer design

Forward and reverse primers were designed using Primer3 (v0.3.0 and v0.4.0) web-based software (<http://frodo.wi.mit.edu/>), and gene sequences obtained from the University of California Santa Cruz Genome Bioinformatics website (<http://genome.ucsc.edu/>). Primer pairs were selected on the basis of their length (18-27 bp), size of amplicon generated (200-600bp), melting temperature (57°C-63°C), guanine and cytosine (GC) content (40-60%), and location outside polymorphic sites. For complementary DNA (cDNA), wherever possible primers were designed such that the resulting amplicon would cross an exon-exon boundary, preventing amplification of any contaminating genomic DNA (gDNA). Universal M13 tags (F-TGTAACGACGGCCAGT, R-CAGGAAACAGCTATGACC) were added for sequencing ease where appropriate. Specific primer sequences are detailed in relevant experimental chapters. Primers were synthesised by Eurofins MWG Operon (Ebersberg, Germany).

2.1.2 PCR reaction components and cycling conditions

Polymerase Chain reaction (PCR) was used to exponentially amplify gDNA and cDNA, producing quantities sufficient for gel electrophoresis, fragment analysis and dye-

terminator sequencing. Unless otherwise stated, standard reaction mixes were used as detailed in tables 2.1, 2.2 and 2.3. Unless otherwise specified, reactions contained 20ng DNA. Reaction volumes quoted allow for the addition of 1µl DNA (plus 11.5µl of Applied Biosystems [Warrington, UK] reagent mix or 9µl of Qiagen [Crawley, UK] and PFU [Promega, Southampton, UK] reagent mix): where DNA was available at higher or lower concentrations, the volume of distilled water was adjusted to compensate.

Reaction component	Volume in 10µl reaction (µl)	Final reaction concentration
Buffer	1	n/a
Q-solution	2	n/a
dNTPs (2.5mM each)	0.8	0.2mM (each)
Forward primer (5mM)	0.5	0.25mM
Reverse primer (5mM)	0.5	0.25mM
Hotstar Taq	0.2	1U
dH ₂ O	4	n/a

Table 2.1: PCR reaction components – Qiagen reagents. Buffer, Q-solution and Hotstar Taq all supplied by Qiagen. Volumes provided for n=1 standard 10µl reaction. Reaction volumes scaled up linearly.

Reaction component	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 ₂ O	7.95	n/a

Table 2.2: PCR reaction components - Applied Biosystems reagents. Buffer, Magnesium Chloride, dNTPs and Amplitaq gold all supplied by Applied Biosystems. Volumes provided for n=1 standard 12.5µl reaction. Reaction volumes scaled up linearly.

Reaction component	Volume in 10 μ l reaction (μ l)	Final reaction concentration
PFU Buffer	1	n/a
dNTPs (2.5mM each)	1.6	0.4 mM (each)
Forward primer (5mM)	0.5	0.25mM
Reverse primer (5mM)	0.5	0.25mM
PFU polymerase	0.1	1U
dH ₂ O	5.3	n/a

Table 2.3: PCR reaction components – PFU reagents. Buffer and polymerase supplied by Promega. Volumes provided for n=1 standard 10 μ l reaction. Reaction volumes scaled up linearly.

After brief centrifugation, samples were amplified on a PTC-225 Peltier Thermal-Cycler (MJ Research, Waltham, Massachusetts, USA) according to thermal cycling conditions as detailed in table 2.4 unless otherwise specified.

Step	Action	Temp (°C)	Time	Temp (°C)	Time	Temp (°C)	Time
1	Polymerase activation	95	12min	95	15min	95	2min
2	DNA melting	94	30s	94	1min	95	1min
3	Primer annealing	58	30s	60	1min	60	30s
4	Synthesis & Extension	72	1min	72	1min	72	10min
5	Steps 2-4 repeat	35 times		35 times		35 times	
6	Final extension	72	10min	72	10min	72	10min
Reagent set		Applied Biosystems		Qiagen		PFU (Promega)	

Table 2.4: Standard PCR thermal cycling conditions

2.1.3 PCR product validation using agarose gel electrophoresis

The success and specificity of all PCR reactions was confirmed via agarose gel electrophoresis of the PCR product. For products less than 1kb in length, a 2% gel was generated by dissolving agarose (Sigma Aldrich, Gillingham, UK) in 1X Tris/Borate/EDTA (TBE) (NBS Biologicals, Huntingdon, UK), with the aid of an 800W microwave for 2min. After cooling for 15min but prior to gel solidification, the intercalating agent Ethidium Bromide (Sigma Aldrich, Gillingham, UK) was added, to produce a final 0.002% solution. Volumes depended upon the number of samples to be run and therefore the size of gel required; a standard small gel contained 3g agarose, 150mls TBE and 3 μ l ethidium bromide. Reagents were multiplied up accordingly for larger gels. For samples greater than 1kb in length, a 1% gel was produced by halving the mass of agarose used. Gels were set in an electrophoresis tank of appropriate size containing combs of appropriate spacing (both Appleton Woods, Birmingham, UK). Once set, gels were covered with 1xTBE.

5 μ l PCR product plus 5 μ l Orange G loading buffer (4g sucrose in 10ml dH₂O and 20mg Orange G [Sigma Aldrich, Gillingham, UK]) was added to a sample well, alongside 4 μ l (0.2 μ g) 100bp or 1kb DNA ladder (New England BioLabs, Hitchin, UK) as appropriate. For a standard small gel with anticipated PCR product sizes of 300-600bp, 120V was passed through the gel for 40minutes. Voltage and time were increased for larger gels and larger amplicon sizes. Gels were visualised using Ultra Violet (UV) technology on a Gel Doc 2000 illuminator system and Quantity One v4.3.1 software (both BioRad, Hemel Hempstead, UK).

2.2 Sequencing

Sequencing was performed to establish genetic imprinting status in islets and pancreas (chapter three), and to confirm rare variant heterozygosity (chapter six). The sequence of PCR-amplified products was obtained via a modified version of the chain-termination (Sanger) sequencing method using the Applied Biosystems BigDye v3.1 Cycle Sequencing Kit. The kit includes fluorescently-labelled dideoxynucleotide triphosphates (ddNTPs), which lack a 3'-hydroxyl group and are therefore unable to form a phosphodiester bond with subsequent nucleotides in an extending DNA sequence. Each ddNTP has a different fluorescence emission spectrum, allowing premature chain termination at each nucleotide to be determined via capillary electrophoresis (Rosenblum et al., 1997; Sanger et al., 1977; Smith et al., 1986).

2.2.1 Clean-up of PCR products

Exonuclease I (EXO, New England Biolabs UK Ltd) and shrimp alkaline phosphatase (SAP, Promega UKLtd) were combined in a 1:1 volume ratio, equivalent to 10 and 0.5 units of each enzyme respectively per microliter. One microlitre of EXO-SAP mix was added to 1µl of PCR product and centrifuged briefly. Samples were incubated at 37°C for 30min to allow the enzymes to remove residual PCR contaminants, before heating at 80°C for 15min to inactivate the enzymes.

2.2.2 Sequencing reactions

Each sequencing reaction contained 2µl PCR product-EXOSAP mix, together with 8µl mastermix as detailed in table 2.5. Sequencing primers (F-tgtaaaacgacggccagt, R-caggaaacagctatgacc) were complementary to the M13 tails added to PCR primers. After

mixing, the sequencing reaction comprised 25 cycles of: 96°C for 10s, 50°C for 5s and 60°C for 4min. Sequencing products were centrifuged at 1000xg for 2min and 5µl deionised water was added to each sample.

Reagent	Volume in reaction (µl)
Big Dye Terminator Sequencing Mix v1.1	0.25
Big Dye Dilution Buffer	1.88
M13 sequencing primer (2.5µM)	2.00
Distilled water	3.88

Table 2.5: DNA sequencing reagents for n=1 reaction. Sequencing Mix and Dilution Buffer from Applied Biosystems. Reaction mix was scaled up linearly.

2.2.3 Sequencing product clean-up

To remove unused sequencing reagents, samples were passed through a 96well Performa DTR V3 short clean up plate (VH Bio, Gateshead, UK) according to manufacturer's instructions. Briefly, sequencing product plus water (total 15µl) was passed through the column well plate by centrifugation (850xg for 5min). Eluent was collected and heated for 90min at 70°C to remove water.

2.2.4 Capillary electrophoresis and analysis

Samples were sequenced via capillary electrophoresis (ABI3730xl DNA Analyser, standard run module, 15s injection, v1.1 dye set and analysis module) at the University's core sequencing facility based at the Department of Zoology. Fluorescence emitted by ddNTPs was converted to .ab1 output files, which were visualised and compared to appropriate reference sequences using Mutation Surveyor v3.4 (Soft Genetics, Cambridge, UK).

2.3 RNA extraction

RNA was extracted from human adult islets, adult pancreas and foetal pancreas to quantify mRNA expression and determine genetic imprinting at the *KCNQ1* locus (chapter three), and from adult adipose tissue (omental and subcutaneous) to quantify mRNA expression of proinsulin-associated loci (chapter four). Details of sample origin, donor characteristics and ethical considerations are provided in the relevant experimental chapters.

2.3.1 Human islet processing and storage for RNA extraction

Islets cells in suspension were transferred to a 15ml centrifuge tube, before pelleting by centrifugation at 100xg for 3min (4°C). Suspension media was removed and sufficient 1X phosphate buffered saline (PBS, Sigma Aldrich, Gillingham, UK) added to cover the pellet. Islets were again pelleted by centrifugation (100xg, 3min, 4°C) and the PBS removed. An additional two PBS washes were performed. Following the final wash, islets were re-suspended in 5 volumes of *RNAlater* solution (Ambion, Paisley, UK) to prevent RNA degradation. Samples were stored at 4°C for 24 hours to allow *RNAlater* to fully penetrate the cells, before transferral to -80°C for longer term storage.

2.3.2 RNA extraction from human islets

RNA was extracted from human islets using the guanidinium-thiocyanate-phenol-chloroform method (Chomczynski and Sacchi, 1987). Islets suspended in *RNAlater* were thawed on ice, before pelleting by centrifugation at 200xg for 3min (4°C). If a pellet failed to form due to the viscosity of *RNAlater*, an equal volume of 1X PBS was

added to dilute the solution. After removal of *RNAlater*, 1 ml Trizol (Ambion, Warrington, UK) was added to the cells. To ensure rapid inhibition of RNase activity, cells were lysed immediately by gently passing the suspension ten times through a 20 gauge needle (Becton Dickinson, Oxford, UK). Following incubation at room temperature for 10min, lysates were transferred to clean 1.5ml RNase-free centrifuge tubes (Applied Biosystems) with 200µl chloroform (FisherScientific, Loughborough, UK). Tubes were shaken vigorously by hand for 15s to begin organic and aqueous phase separation, before incubation at room temperature for 5min. Phase separation was completed by centrifugation at 12,000xg for 15min (4°C), producing a pink organic phenol phase at the base of the tube, a white interphase containing DNA and a clear top aqueous phase containing RNA. The aqueous phase was transferred by pipette to a clean 1.5ml RNase-free tube, and 500µl isopropanol (Fisher Scientific, Loughborough, UK) was added to precipitate the RNA. Remaining organic and DNA phases were stored at -20°C for subsequent DNA extraction (see section 2.4). After incubation for 5min at room temperature with repeated inversion, the solution was stored overnight at -20°C. The following day, samples were centrifuged at 12,000xg for 50min (4°C) to pellet RNA. Isopropanol was removed carefully, ensuring the pellet was not dislodged, and 1ml 75% ethanol (Sigma Aldrich, Gillingham, UK) was added to wash the pellet. After centrifugation at 12,000xg for 30min, the same procedure was repeated with 1ml fresh 75% ethanol. The final ethanol wash was removed and the RNA pellet allowed to air-dry for 10min, before resuspension in at least 20µl RNase-free water (more as necessary for complete resuspension). After determining sample quality (see section 2.3.5), samples were stored at -80°C for future use.

2.3.3 RNA extraction from human foetal pancreas

Samples were thawed on ice, placed in a 2ml tube (Eppendorf, Cambridge, UK) with 1ml Trizol and disrupted using a tissue lyser (Qiagen). Lysate was transferred to a clean 1.5ml RNase-free tube and centrifuged at 12,000xg for 20min (4°C) to pellet any insoluble fragments. Supernatant was transferred to a fresh 1.5ml RNase-free tube with 200µl chloroform, for extraction according to the protocol detailed in section 2.3.2.

2.3.4 RNA extraction from human adult pancreas, omental adipose and subcutaneous adipose tissue

Samples were rapidly dissected on dry ice before transferral to 2ml tubes (Eppendorf) containing a ball bearing (Retsch, Leeds, UK) and 1ml Trizol. Large samples were dissected into pieces <100mg and processed in sections. Tissues were homogenised for 5min on the highest frequency of a Retsch free standing homogeniser. Lysate was transferred to a clean 1.5ml RNase-free tube and centrifuged at 12,000xg for 20min (4°C) to pellet any insoluble fragments and separate a lipid layer. Supernatant was removed by pipette through the lipid layer and transferred to a clean 1.5ml RNase-free tube with 200µl chloroform, for extraction according to the protocol detailed in section 2.3.2.

2.3.5 RNA quality determination

The quality of RNA was assessed via a combination of spectrophotometry and microfluidics. A NanoDrop 1000 (ThermoScientific, Wilmington, USA) spectrophotometer was used to assess RNA purity. After blanking with RNase-free

water, 1µl of sample was used to assess absorbance at wavelengths of 230, 260 and 280nm. RNA with a 260/280 ratio in the range 1.7-2.1 and a 230/260 ratio in the range 1.5-2.2 was deemed to be of acceptable purity for further analysis.

RNA integrity was determined using an Agilent 2100 Bioanalyser and total eukaryotic RNA 6000 nano chip (both Agilent, Wokingham, UK). The chip priming station and chip were prepared according to manufacturer's instructions, using 1µl sample. Samples are separated on the basis of size, and quantified as they migrate through the capillary and pass a laser. Size peaks are used to determine an RNA Integrity Number (RIN score) according to a proprietary algorithm (figure 2.1), where 0 indicates completely degraded and 10 indicates perfect integrity. Samples with a RIN score >6 were deemed acceptable for experimental use.

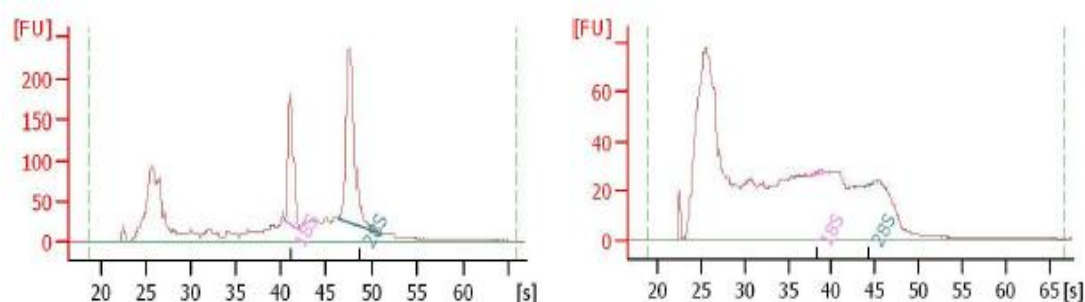


Figure 2.1: RNA integrity testing as performed by the Agilent Bioanalyser. Left hand trace has identifiable miRNA peak which passed through the laser at around 25seconds, followed by clear 18S and 28S ribosomal RNA peaks at 41 and 48 seconds respectively (RIN score = 8.1, good quality). Right hand trace shows considerable degradation, with most RNAs broken down to smaller sizes and ribosomal peaks barely evident (RIN score = 2.3, RNA considered unusable).

2.4 DNA extraction

DNA was extracted from human adult islets, adult pancreas and foetal pancreas to determine genetic imprinting at the *KCNQ1* locus, and from spleen and whole blood to determine T2D risk genotype status (all chapter three).

2.4.1 DNA extraction from human islets and human pancreas

DNA was extracted from human islets and pancreas using the fraction remaining after phenol-chloroform extraction of RNA (see section 2.3), to maximise the use of scarce resources and minimise sample variability by obtaining RNA and DNA from the same tissue. After thawing on ice, 300µl pure (100%) ethanol was added to precipitate DNA and mixed by inversion. The mixture was incubated at room temperature for 15min with frequent inversion, before pelleting by centrifugation at 4000xg for 5min (4°C). The supernatant was removed, and the remaining pellet covered with 1ml DNA wash solution (0.1M trisodium citrate [Sigma Aldrich] in 10% ethanol). The pellet was left in wash solution at room temperature for 30min, with occasional inversion. After centrifugation (4000xg, 5min, 4°C), wash solution was removed and a fresh ml was added. After a further 30min at room temperature and a further centrifugation as above, wash solution was removed and the pellet covered in 75% ethanol. Following a further 30min at room temperature with occasional inversion, the sample was centrifuged as above and ethanol removed. Pellets were air-dried for 10min to remove residual ethanol, and resuspended in 8mM NaOH (Sigma Aldrich) (as little volume as required for resuspension, minimum 100µl). If required, resuspension was aided by mixing at 400rpm for 1hr on a plate mixer heated to 50°C (Eppendorf). Successful DNA extraction was confirmed by PCR with a validated primer set and an agarose gel check

as described in section 2.1.3. Stock DNA was stored at -20°C; working dilutions were stored at 4°C.

2.4.3 DNA extraction from human spleen and whole blood

Where possible, the OXCIT team obtained a spleen sample from human islet donors. This allowed genotypes, which are not tissue-specific, to be obtained without depleting DNA obtained from islets. Prior to extraction, spleen tissue was stored at 4°C and whole blood was stored at -80°C. DNA was extracted from spleen and whole blood using the automated Maxwell 16 system with a Tissue DNA Purification Kit or Whole blood Purification Kit (both Promega, Madison, USA) according to manufacturer's instructions. Briefly, tissue was lysed by shearing and genomic DNA captured on proprietary MagneSil Paramagnetic Particles. Particles were carried by magnetism through wash buffers, before releasing DNA into 300µl elution buffer. Eluted stock DNA was stored at -20°C. Successful DNA extraction was confirmed by PCR with a validated primer set and an agarose gel check as in section 2.1.3. Stock DNA was stored at -20°C; working dilutions were stored at 4°C.

2.5 Genotyping

All genotyping was performed using TaqMan technology (Applied Biosystems), amplifying the relevant genomic region in the presence of differentially labelled fluorescent probes (VIC/FAM) specific to the two possible alleles.

2.5.1 Assay design

Where available, inventoried genotyping assays were obtained from Applied Biosystems. Assay details are provided in the relevant chapters. Where no inventoried assays were available, custom assays were designed using genomic sequence obtained from the National Centre for Biotechnology Information (NCBI) and Primer Express v3.0 (Applied Biosystems). This software selects primers and probes on the basis of amplicon length (50-150bp), probe length (13-25bp), GC content (30-80%), probe annealing temperature (65-67°C) and primer annealing temperature (58-60°C). Designed primer/probe mixes were synthesised by Applied Biosystems.

2.5.2 TaqMan PCR

Two-step genotyping PCRs were conducted using 10ng DNA in a maximum 2µl volume, dried in to a 384-well PCR plate (4titude, Wotten, UK) by incubating at 80°C for 10min. Reactions were conducted in a 2µl volume, comprising 1µl genotyping mastermix (containing polymerase and dNTPs, KBiosciences, Hoddesdon, UK), 0.98µl distilled water and 0.02µl assay. After brief centrifugation, samples were heated to 95°C for 10min (polymerase activation) and underwent 40 or 50 cycles (detailed in relevant chapters) comprising: 92°C for 15s (DNA melting) and 60°C for 1min (combined annealing and extension step).

2.5.3 Genotype determination

After centrifugation, genotype was determined through post-read allelic discrimination on an Applied Biosystems 7900HT machine, using SDSv2.3 software and standard analysis settings. Genotypes were assigned according to proprietary algorithms, but visually checked. Some genotype re-assignments were made where myself and a postdoctoral colleague (Dr. Amanda Bennett, OCDEM, University of Oxford) independently agreed, but results were discarded if more than 5% of samples were deemed to require re-assignment. At least five non-template control (NTC) wells (containing dH₂O) and five sample duplicates were included on each genotyping plate, and results were discarded if NTCs showed amplification or if duplicate well results were not concordant.

2.6 cDNA synthesis

cDNA was synthesised to perform mRNA expression quantification of genes at the *KCNQ1* locus (chapter three) and genes around proinsulin-associated loci (chapter four), and for sequencing to determine genetic imprinting (chapter three).

Single stranded cDNA was generated from RNA using the QuantiTect reverse transcription kit (Qiagen, Crawley, UK), containing Omniscript and Sensiscript reverse transcriptases and random primers. RNA (1µg) in 12µl distilled water was mixed with 2µl gDNA wipeout buffer and incubated at 42°C for 2min to eliminate any residual gDNA contamination. Reverse transcription (RT) reactions were performed for each sample in the presence (RT+ve) and absence (RT-ve) of reverse transcriptase enzyme, allowing further expression experiments to test for genomic DNA contamination. Mastermix (6µl) as detailed in table 2.6 was added to each sample and mixed by pipetting.

Reaction component	Volume RT +ve (µl)	Volume RT -ve (µl)
5x RT Buffer	4	4
RT random primer mix	1	1
Reverse transcriptase enzyme	1	-
dH ₂ O	-	1

Table 2.6: Reaction components for cDNA synthesis. All reagents included in QuantiTect reverse transcription kit (Qiagen).

After centrifugation, the reverse transcription reaction took place during incubation at 42°C for 30min, followed by 95°C for 3min to inactivate the enzyme. Stock cDNA samples were stored at 4°C.

2.7 mRNA quantification

mRNA expression was quantified for genes at the *KCNQ1* locus (chapter three) and genes around proinsulin-associated loci (chapter four).

mRNA expression was quantified by quantitative (real-time) PCR (QPCR), using TaqMan chemistry (Applied Biosystems). TaqMan probes have a fluorophore and quencher attached to their 5' and 3' end respectively, in sufficiently close proximity to prevent fluorescence emission. During PCR, the 5'-3' exonuclease activity of Taq polymerase releases the fluorophore from its quenched state (Holland *et al.*, 1991), such that the fluorescent level is directly proportional to the quantity of cDNA generated. During the exponential phase of a PCR, the quantity of cDNA at a given cycle number will be proportional to the input quantity. The cycle number at which fluorescence crosses a critical threshold will therefore be proportional to the initial cDNA (and originally mRNA) concentration.

2.7.1 TaqMan probe selection

Where possible, inventoried expression assays were obtained from Applied Biosystems. Assays were selected to amplify as many alternative transcripts of the relevant gene as possible. To prevent amplification of potentially contaminating gDNA, wherever possible probes were selected which crossed an exon-exon boundary (_m1), rather than lying within a single exon (_s1). When _m1 probes were unavailable (for example in single exon gene), RT-ve reactions were included in expression experiments to check for gDNA contamination. Where no inventoried assays were available, custom assays were designed using Primer Express software as detailed in

section 2.5.1. All assays for genes of interest were FAM-labelled. Specific assay details are provided in the relevant chapters.

2.7.2 Housekeeping gene selection

Accurate quantitation via QPCR requires normalisation to endogenous control ('housekeeping') genes, which are uniformly expressed across samples and thus correct for any differences in input cDNA concentration. It is not clear which are the most consistent housekeeping genes for specific tissues. Previous investigation in our laboratory (data not shown) has suggested that hypoxanthine-guanine phosphoribosyltransferase 1 (*HPRT1*) and peptidylprolyl isomerase A (*PPIA*) are particularly appropriate for use with islet cDNA. The islet experiments in chapter three therefore use these genes and ribonuclease P RNA component H1 (*RNaseP*) as endogenous controls. Selecting appropriate housekeepers for the tissue panel expression experiments reported in chapter four was more challenging. The above three genes, as well as the established endogenous controls glyceraldehydes-3-phosphate dehydrogenase (*GAPDH*) and beta-2-microglobulin (*B2M*) were all tested across the tissue panel. *HPRT1*, *PPIA*, *RNaseP* and *B2M* showed consistent results and were used for further analysis.

All experiments used the geometric mean (Vandesompele *et al.*, 2002) of three endogenous control genes for normalisation. Endogenous control genes were VIC-labelled to enable multiplexing with FAM-labelled gene of interest assays. Details of the housekeeping genes used are provided in table 2.7.

Gene	Endogenous control assay ID
<i>HPRT1</i>	4326321
<i>PPIA</i>	4326316
<i>GAPDH</i>	4326317
<i>RPPH1 (RNAseP)</i>	4316844
<i>B2M</i>	4326319

Table 2.7: Assay details for total expression quantification.
Assay IDs are VIC-labelled TaqMan endogenous control assays (Applied Biosystems).

2.7.3 Standard curve generation

Standard curves were generated from serial dilution of pooled cDNA and used to calculate amplification efficiency. A concentrated pool was created using 1µl neat cDNA from each sample to be used in the experiment, and diluted as appropriate in 0.01M Tris HCl. Six concentrations were generated, encompassing the dilution at which the full experiment was conducted. For example, for a QPCR experiment in which samples were run at a 1:100 dilution, standards were generated at 1:50, 1:100, 1:200, 1:400, 1:800 and 1:1600 dilutions. For samples run at 1:10, standard dilutions started at 1:5. Fresh standards were generated for each experiment.

2.7.4 Quantitative real-time PCR

Prior to main experiments, each assay was tested on a sub-set of samples (usually those with the greatest cDNA availability) at dilutions (in 0.01M Tris HCl) of 1:5, 1:10, 1:50 and 1:100. Results were assessed to select the optimum experimental dilution, based on achieving quantifiable amplification within the optimum experimental range

(cycle 18-30), whilst minimising cDNA use. An assay was deemed unquantifiable if no amplification could be detected after 40 amplification cycles at a 1:5 dilution. Dilutions used for individual assays are provided in relevant chapters.

QPCR was performed in a reaction volume of 10 μ l, comprising 4 μ l diluted cDNA, 5 μ l gene expression mastermix (Applied Biosystems), 0.5 μ l assay and 0.5 μ l dH₂O. Where assays were multiplexed, 0.5 μ l of each assay and no water was used. Samples and standards were amplified in triplicate, with standards for a given assay always on the same plate as samples for that assay. 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.

2.7.5 Analysis

Fluorescence readings were taken after every amplification cycle on an Applied Biosystems 7900HT machine and analysed using SDSv2.3 software (Applied Biosystems). Output values were comparative threshold (C_T) scores: the cycle number at increasing which fluorescence level crossed a critical threshold (figure 2.2). Threshold levels were set to be in the exponential phase of amplification. In general, gene expression was quantified by the $\Delta\Delta C_T$ method (Pfaffl, 2001). C_T values for target genes were compared to a control sample or value (ΔC_T), before normalising to the endogenous control C_T value ($\Delta\Delta C_T$). In general, the control value was taken as the C_T value for the standard sample pool of the same dilution as experimental samples. For example, if samples were run at a 1:100 dilution, the C_T value for the pooled standard

at 1:100 was used as the control value. The resulting value was linearised, using the formula:

$$(1 + \text{assay efficiency})^{-\Delta\Delta CT}$$

In order for assays to be comparable, any difference in their amplification efficiency must be taken in to account. Assay efficiency was calculated from the standard curve samples, using the formula :

$$\left(10^{\frac{-1}{\text{slope}}}\right) - 1$$

where slope refers to the gradient when C_T (x-axis) was plotted against log-dilution (y-axis). Further analysis details are provided within relevant chapters.

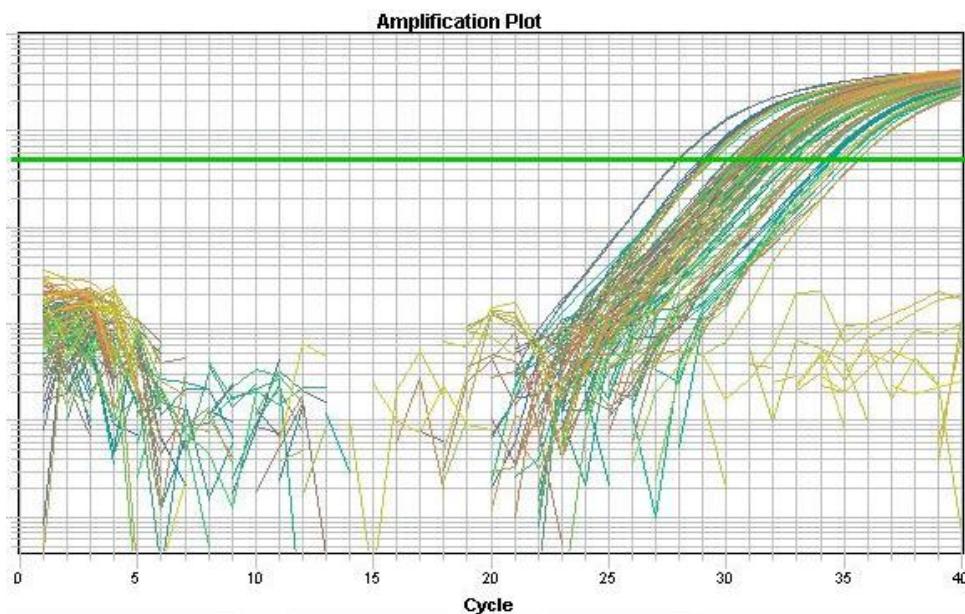


Figure 2.2: Example amplification plot for mRNA quantification (B2M). PCR cycle number (x-axis) is plotted against normalised fluorescence intensity (y-axis). Each line represents amplification in a single sample well. Fluorescent intensity increases as amplification progress. Green line represents the critical threshold, as set by SDSv2.3 software during the exponential phase of the reaction. C_T score is the x-axis value where the amplification curve for each sample crosses the critical threshold – in this case, between 28 and 36 cycles.

2.7.6 Quantitative PCR as a method to determine human islet purity

The islet preparations analysed were of varying purity. Since the gene expression profiles of the endocrine and exocrine pancreas are different, it was important to quantify the purity of every sample. This was achieved by comparing expression levels of three endocrine-specific and three exocrine-specific markers, detailed in table 2.8.

	Gene Details	Assay ID
Endocrine Markers	Insulin (<i>INS</i>)	Hs00355773_m1
	Glucagon (<i>GCG</i>)	Hs00174967_m1
	Somatostatin (<i>SST</i>)	Hs00356144_m1
Exocrine Markers	Chymotrypsin (<i>CTRB</i>)	Hs00200713_m1
	Pancreatic Lipase (<i>PNLIP</i>)	Hs00609591_m1
	Amylase, alpha-2A (<i>AMY2A</i>)	Hs00420710_g1

Table 2.8: Genes and expression assays for islet purity analysis

Purity marker assays were amplified (all at 1:100 dilutions) and analysed as detailed in section 2.7.5. Since this analysis depends upon comparing expression levels between genes, threshold fluorescence was fixed at 0.1 normalised units of fluorescence (ΔR_n) for all assays. Purity was determined by calculating the proportion of overall expression which came from the endocrine markers, using the formula:

$$\frac{(\text{mean of endocrine assay expression})}{(\text{mean of endocrine assay expression} + \text{mean of exocrine assay expression})}$$

A value of 0 therefore represents entirely exocrine material, whilst a value of 1 represents pure endocrine islets. This assessment correlated well ($r^2 = 0.70$) with

purity estimates provided by the OXCIT team and Swedish laboratories on the basis of dithazone staining (figure 2.3).

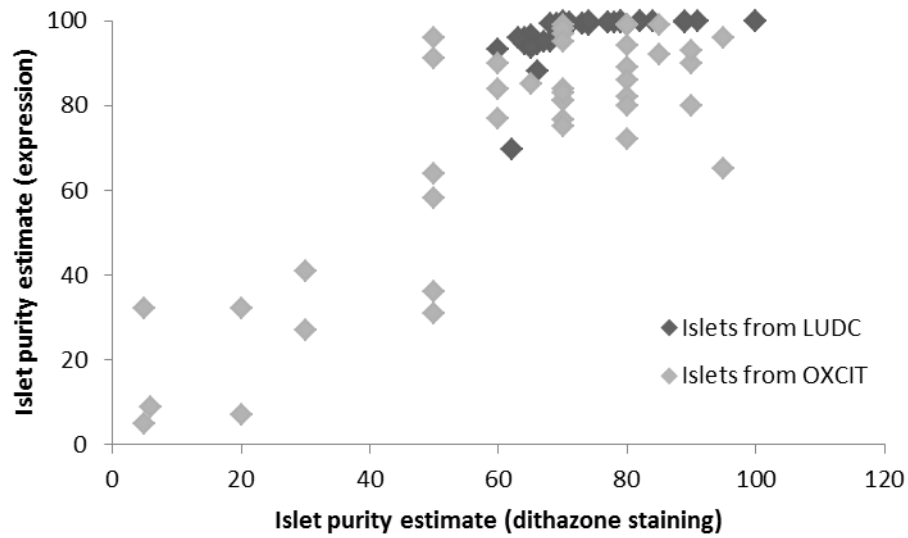


Figure 2.3: Comparison of islet purity measurements based on dithazone staining in islet isolation laboratories (x-axis) and endocrine-vs-exocrine gene expression analysis (y-axis). LUDC = Lund University Diabetes Centre, OXCIT = Oxford Centre for Islet Transplantation. Correlation coefficient (r^2) = 0.70. Note that only samples >80% purity (according to expression analysis) were used for the work described in chapters three and four.

2.8 Multiplex Amplicon Quantitation

The copy number of repeated or deleted gDNA regions (chapter four) was determined using Multiplex Amplicon Quantitation (MAQ), a method which amplifies fluorescently labelled test and control regions in a multiplexed PCR reaction and compares the resulting fluorescent intensity.

2.8.1 Assay design

MAQ plexes were designed and manufactured by Multiplicom (Antwerp, Belgium), on the basis of DNA sequence obtained from the University of California Santa Cruz Genome Bioinformatics website (<http://genome.ucsc.edu/>). Each plex contained 5 or 6 FAM-labelled primer sets for amplicons within the copy number variable region, and 6 or 7 FAM-labelled primer sets for amplicons within control non copy-number variable regions. Primer sets were selected on the basis of uniform melting temperatures and non-complementarity, to allow multiplexing. Further assay details are provided in chapter five.

2.8.2 Amplification

PCR was performed with 50ng DNA in 10µl dH₂O, 5µl MAQ buffer (containing primer sets, dNTPs and 1.5mM MgCl₂) and 0.075µl MAQ polymerase (Multiplicom). After mixing and centrifugation, samples were subjected to thermal cycling (table 2.9) and gel-checked as described in section 2.1.3. PCR products were stored at 4°C (for a maximum of 48hr) pending fragment analysis.

Step	Action	Temperature (°C)	Time
1	Polymerase activation	98	10 min
2	DNA melting	95	45 s
3	Primer annealing	60	45 s
4	Synthesis and Extension	68	2 min
5	Steps 2-4 repeat	23 times	
6	Final extension	72	10 min

Table 2.9: MAQ PCR thermal cycling conditions

2.8.3 Analysis

Fragment analysis was performed on an Applied Biosystems 3130xl machine (Gene Scan 500 size standard, dye set D) at the University's core facility, using 2µl PCR product. The resulting .fsa files were analysed using MAQs software v1.0 (Multiplicom), which quantifies fragment peak areas and produces a dosage quotient according to the following formula:

$$\frac{(\text{Control sample target amplicon})}{(\text{Control sample reference amplicon})} \div \frac{(\text{Test sample target amplicon})}{(\text{Test sample reference amplicon})}$$

Further analysis details are provided in chapter five.

2.9 Fragment analysis

Fragment analysis was used to determine genotype of a 12-bp insertion/deletion (indel) in chapter three by distinguishing between differently sized products.

Primers were designed across the indel location as described in section 2.1.1, with the addition of a 6-FAM label on the 5' end of forward primers. Primer details are provided in chapter 3. Using 20ng DNA or 1µl of a 1/10 cDNA dilution (diluted in 0.01M TRIS), samples were PCR-amplified using Qiagen reagents as outlined in section 2.1.2 and gel-checked for integrity as in section 2.1.3. Fragment analysis was performed on an Applied Biosystems 3130xl machine at the University's core facility, using 2µl PCR product (Gene Scan 500 standards, Dye set D). The resulting .fsa files were visualised using Peak Scanner v1.0 software (Applied Biosystems). Indel status was called manually, by comparing the size of FAM-fluorescence peaks to standard peak sizes. Further details are provided in chapter 3.

2.10 Quantification of DNA methylation

Methylation of human islet and pancreas DNA was quantified as part of investigations into T2D risk at the *KCNQ1* locus (chapter three), via pyrosequencing of bisulphite-treated DNA.

Bisulphite treatment converts unmethylated cytosine bases to uracil, leaving methylated cytosines unaffected. PCR amplification converts uracil to thymine, creating a T/C polymorphism in which T is representative of unmethylated cytosines and C is representative of methylated cytosines in the original DNA (figure 2.4). Quantification of this polymorphism via pyrosequencing is therefore equivalent to quantification of methylation levels in the original DNA sequence.

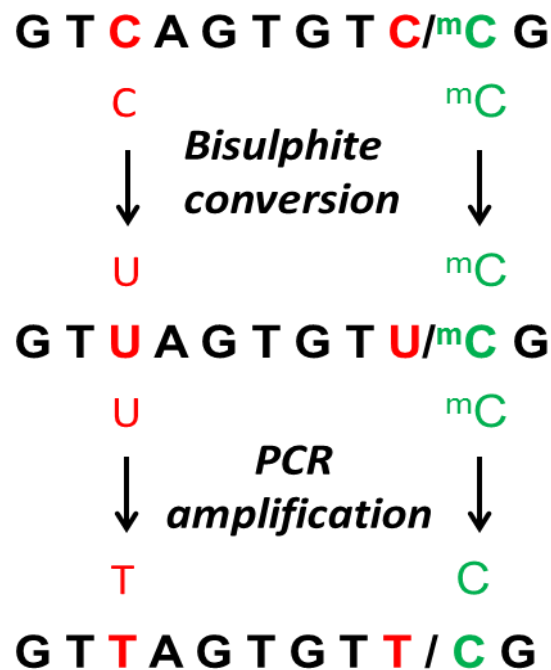


Figure 2.4: Bisulphite treatment of DNA for pyrosequencing. Bisulphite treatment converts unmethylated cytosine bases (C) to uracil (U). Methylated cytosines (mC) are left unchanged. Uracil bases are converted to thymine (T) during PCR amplification, producing a C/T polymorphism according to methylation status in the template DNA

The pyrosequencing procedure releases free nucleotides successively into the reaction mix. As they are incorporated into the growing nucleotide strand, pyrophosphate is released and fuels the generation of light via conversion of luciferin to oxyluciferin. Solutions of A, C, G and T bases are added and removed sequentially; the sequence of solutions which produce light therefore determine the sequence of template, and quantifying the intensity of light produced upon addition of C or T is proportional equivalent to quantifying that polymorphism.

2.10.1 Bisulphite treatment of DNA

Islet DNA was Bisulphite treated using the EZ methylation gold kit (Zymo Research, Irvine, US). 1µg DNA in 20µl dH₂O was mixed with 130µl CT conversion reagent and incubated at 98°C for 10min for enzyme activation, followed by 64°C overnight (12hr) for conversion. The mixture was treated with 200µl desulphonation buffer and 100µl wash buffer as per manufacturer's instructions, before being eluted in 25µl 8mM NaOH. Eluted DNA was stored at 4°C pending pyrosequencing.

2.10.2 Primer design and PCR

PCR and sequencing primers were designed using PyroMark v1.0 software (Qiagen), against in-silico bisulphite treated DNA. Sequencing primers were placed such that the analysed sequence would contain as many informative CpG dinucleotides as possible, as well as a cytosine base not in a CpG dinucleotide to act as a Bisulphite conversion control. Reverse PCR primers were biotinylated to enable products to be captured by sepharose beads during pyrosequencing. Primers were synthesised by Eurofins; specific assay details are provided in chapter 3. Bisulphite-treated DNA (2µl) was added to 48µl PCR mix (detailed in table 2.10), mixed and subjected to thermal cycling

as detailed in table 2.11. Products were gel-checked (section 2.1.3) before pyrosequencing.

Reaction component	Volume in 50µl reaction (µl)	Final reaction concentration
Titanium Buffer	5	n/a
dNTPs (2.5mM each)	4	0.2mM each
Forward primer (10mM)	1.5	0.3mM
Reverse primer (10mM)	1.5	0.3mM
Titanium Taq	0.2	1U
Distilled water	35.8	n/a

Table 2.10: PCR for pyrosequencing reaction components. Titanium buffer and Taq supplied by Clontech (Saint-Germain-en-Laye, France). Volumes provided for n=1 standard 50µl reaction. Reaction volumes scaled up linearly.

Step	Action	Temperature (°C)	Time
1	Polymerase activation	94	1 min
2	DNA melting	95	20 s
3	Primer annealing	57	10 s
4	Synthesis and Extension	72	10 s
5	Steps 2-4 repeat	44 times	
6	Final extension	72	2 min

Table 2.11: Pyrosequencing PCR thermal cycling conditions

20µl of PCR product was added to 60µl of binding mix, comprising 3µl streptavidin sepharose beads (GE Healthcare, Amersham, UK), 37µl pyrosequencing binding buffer (Qiagen) and 20µl dH₂O. Samples were agitated on a plate shaker at 1400rpm for 15min at room temperature. Meanwhile, 45µl of sequencing mix, comprising 40µl annealing buffer (Qiagen) and 5µl (at 10mM) of the appropriate sequencing primer

was added to each well of a PSQ-plate low (Qiagen). After agitation, beads were captured using a PyroMark Q96 vacuum prep workstation and successively washed in 70% ethanol, PyroMark denaturation solution and PyroMark wash buffer (both Qiagen) before being released into the sequencing mix. DNA was denatured at 80°C for 2 min and left to cool at room temperature for 10min. Pyrosequencing was performed on a PSQ96 ID machine, using enzyme, substrate and dNTP from the PyroMark Gold Q96 kit (both Qiagen). Dispensation order was determined by PyroMark software v2.0 (Qiagen), according to the sequence to be analysed. Enzyme, substrate and dNTPs were added to the pyrosequencing cartridge as per requirements for the sequence to be analysed (i.e. dictated by sequence length and base composition) and pyrosequencing reaction was performed according to standard settings.

2.10.4 Analysis

Each potentially methylated cytosine base was designated as a T/C polymorphism. Additionally, in some assays, one cytosine base outside a CpG dinucleotide was also designated a T/C polymorphism, to act as a check for complete bisulphite conversion. Polymorphisms were analysed using the absolute quantification setting of PyroMark v2.0 software (Qiagen). The output %T and %C was taken to be equivalent to % unmethylated and % methylated respectively in the original DNA sequence.

2.11 Statistical methods

Linear regression, Kolmogorov-Smirnov, Kruskal-Wallis and cluster analyses were conducted in SPSSv18.0 (PASW statistics 18). Statistical significance was defined as $p < 0.05$ for all tests.

Assessment of genotype effects on methylation and expression (chapter three) was conducted by linear regression, with donor age, donor gender, sample purity and centre of origin (for adult islets) or days post-conception (for foetal pancreas) included as covariates. Birth weight association (chapter three) was tested via linear regression of birth weight on genotype, with sex and gestational age as covariates. Since foetal and maternal genotypes are not independent, analysis was also conducted with an adjustment (through inclusion as a covariate) of foetal genotypes for maternal genotype, and of maternal genotypes for foetal genotype. Parental origin-specific birth weight effects were tested via linear regression of birth weight on risk allele parent-of-origin, corrected for sex and gestational age. This is equivalent to a t-test to compare birth weights in paternal and maternal origin groups.

K-means cluster analysis of DNA copy number variation (chapter five) was conducted with cluster number set at three, and a maximum of 50 iterations.

Chapter 3

Functional interrogation of T2D association at the *KCNQ1* locus

3.1 Introduction

3.1.1 The need for functional interrogation of T2D-association loci

Five years following publication of the first genome-wide association study (GWAS) for susceptibility to type 2 diabetes (T2D), the list of confirmed loci stands at above 60 (Altshuler et al., 2000; Cho et al., 2012; Dupuis et al., 2010; Frayling et al., 2007; Gloyn et al., 2003; Grant et al., 2006; Kooner et al., 2011; Saxena et al., 2012; Saxena et al., 2007; Scott et al., 2007; Sladek et al., 2007; Steinthorsdottir et al., 2007; Tsai et al., 2010; Unoki et al., 2008; Voight et al., 2010; Winckler et al., 2007; WTCCC, 2007; Yamauchi et al., 2010; Yasuda et al., 2008; Zeggini et al., 2008; Zeggini et al., 2007). However, progress made in using these data to understand T2D pathology is so far limited. Functional assays have made significant headway where associated regions contain a candidate coding mutation (Beer et al., 2009; Hamming et al., 2009; Nicolson et al., 2009; Rees et al., 2012), but most signals fall outside exonic regions. With the notable exception of chromatin effects at *TCF7L2* (Gaulton et al., 2010), determining the physiological effects of non-coding loci has proved challenging. Whilst it is known that T2D loci predominantly work through impaired insulin secretion rather than increased insulin resistance (Voight et al., 2010), biological insight demands a precise characterisation of the mechanisms through which risk alleles exert their T2D risk effect by disrupting β -cell development or function.

3.1.2 Variants at the *KCNQ1* locus are associated with T2D and β -cell phenotypes

The potassium channel gene *KCNQ1* (voltage-gated channel, KQT-like subfamily, member 1) harbours at least two independent and genome-wide significant regions of association with T2D risk. A signal in intron 15 was initially identified (Unoki et al., 2008; Yasuda et al., 2008) and replicated (Liu et al., 2009; Tsai et al., 2010) by genome-wide association scans in samples of East Asian origin. However, the SNPs with strongest T2D association in these studies differ markedly in allele frequency between East Asian and European populations. The most associated risk allele in these initial studies, rs2237892, has a frequency of 0.41 and 0.33 in Japanese and Chinese populations respectively, but only 0.08 in Europeans (table 3.1). The intron 15 region of association was therefore more readily identifiable amongst East Asians.

However, a large meta-analysis of T2D genome-wide association scans in European populations conducted by the DIAGRAM consortium confirmed the presence of this signal (Dupuis et al., 2010) in Europeans. It also revealed a second, independent signal within intron 10 of *KCNQ1* (figure 3.1 and table 3.1). Figure 3.1 shows association data from the DIAGRAM meta-analysis, with the two association signals (separated by recombination peaks) clearly visible.

The intron 10 signal was more readily detected in Europeans, with the most strongly associated SNP having a lower MAF amongst East Asian populations (table 3.1). Both association signals have now been replicated in both populations at the level of genome-wide significance (Tsai et al., 2010; Unoki et al., 2008).

	SNP	T2D risk allele	Non risk allele	MAF (Europeans)	MAF (East Asians)	T2D association p-value (European)	T2D odds ratio (European)	T2D association p-value (East Asian)	T2D odds ratio (East Asian)
INTRON 10	rs231362	G	A	0.44 ^a	0.13 ^c /0.16 ^d	^g 2.8x10 ⁻¹³	1.08 ^g	^h 3.43x10 ⁻⁸	1.33 ^h
	rs163184	G	T	0.44 ^a	0.39 ^c /0.47 ^d	^g 6.8x10 ⁻⁵	1.08 ^g	-	-
INTRON 15	rs2237892	C	T	0.08 ^a	0.41 ^c /0.33 ^d	^g 2.7x10 ⁻³	1.12 ^g	^f 6.7x10 ⁻¹³	1.49 ^f
	rs2237895	C	A	0.46 ^b	0.41 ^e	^e 3.7x10 ⁻¹¹	1.24 ^e	^e 7.3x10 ⁻⁹	1.32 ^e

Table 3.1: T2D-associated SNPs at *KCNQ1* ^aMAF according to HapMap CEPH data ^bMAF according to experiments presented in section 3.3.1 (SNP not in HapMap) ^cMAF according to HapMap Japanese data ^dMAF according to HapMap Han Chinese data ^eUnoki et. al., 2008 ^fYasuda et. al., 2008 ^gVoight et. al., 2010 ^hTsai et. al., 2010 – association p-value and odds ratio reported for rs231361, in tight LD with rs231362

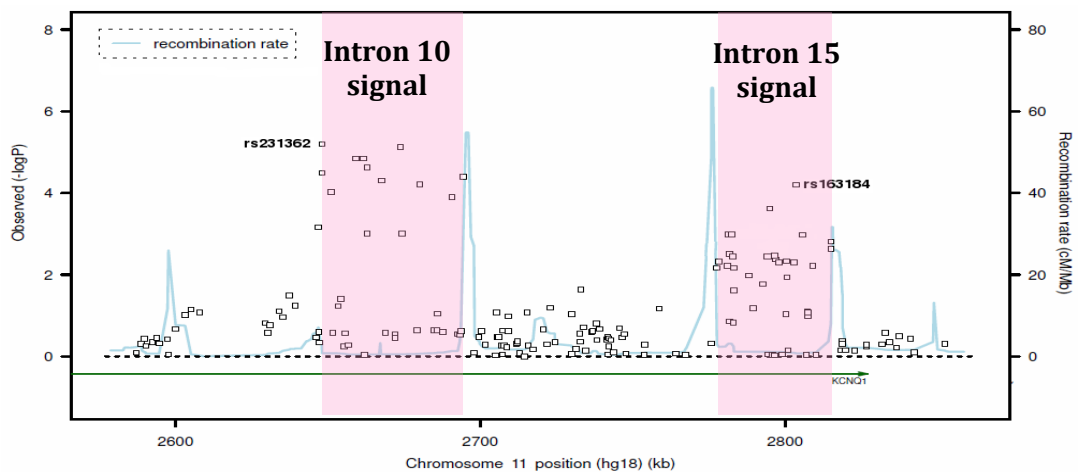


Figure 3.1: T2D association at the *KCNQ1* locus (European population). Chromosome location (x-axis) is plotted against strength of association (y-axis), each point representing a genotyped SNP. Two association peaks (in pink) are evident, clearly separated by recombination hotspots. Data taken from DIAGRAM meta-analysis of European populations, plotted via SNAP (<http://www.broadinstitute.org/mpg/snap/>).

Unlike most large disease and control cohorts, the deCODE group are able to infer the parental origin of SNP alleles on the basis of known genealogy and long-range phasing. Their analysis showed that standard GWA studies mask a parent-of-origin effect in transmission of disease risk at *KCNQ1*: associated SNPs at both loci confer their T2D risk only when maternally inherited (Kong et al., 2009). In the deCODE cohort, rs2237892 (intron 15) and rs231362 (intron 10) reached p-values for association with T2D of 0.0084 and 6.2×10^{-5} respectively when maternally inherited, but 0.71 and 0.73 when inherited paternally.

Further studies have shown that T2D risk at this locus is mediated through impaired islet function. Risk alleles are associated with raised fasting plasma glucose (Tan et al., 2009), a reduction in insulin response during oral glucose tolerance tests and hyperglycaemic glucose clamps (Holmkvist et al., 2009; van Vliet-Ostaptchouk et al., 2012) reduced HOMA-B (Tan et al., 2009) and a reduction in the disposition index

measure of β -cell function (insulin secretion adjusted for insulin sensitivity; Jonsson et al., 2009), with no effect upon HOMA-IR measurements of insulin resistance.

3.1.3 Molecular mechanisms for T2D risk at the *KCNQ1* locus are not clear

KCNQ1 encodes the Kv7.1 voltage-gated potassium channel subunit which, together with *KCNE1*, is responsible for repolarisation of the cardiac action potential following myocardial contraction (Jespersen et al., 2005). It is expressed in human β -cells (MacDonald and Wheeler, 2003), but its potential role in the insulin secretion pathway is not fully understood. It has been suggested that it may play a role in repolarisation of the β -cell following action-potential-induced insulin secretion, closing after depolarisation and opening to restore outward flux of K^+ (MacDonald and Wheeler, 2003). It might therefore be possible to infer a risk mechanism through direct effects upon *KCNQ1* protein function, particularly given the established role of the ATP-sensitive potassium channel *KCNJ11* in insulin secretion and T2D risk (Gloyn et al., 2003). If this was the case, the associated intronic variants might be expected to effect a gain-of-function in the *KCNQ1* potassium channel unit, increasing the rate of β -cell repolarisation and therefore prematurely terminating insulin secretion. It is also possible that a loss-of-function effect might increase T2D susceptibility, through Ca^{2+} overload and eventual β -cell exhaustion or Ca^{2+} -dependent apoptosis.

However, the reality may be more complex. Mutations within *KCNQ1* are a well-documented cause of cardiac arrhythmia: loss-of function mutations are the commonest cause of Long QT syndrome (Hedley et al., 2009), a disorder of delayed repolarisation following myocardial contraction, whilst gain-of functional mutations are responsible for the opposite condition Short QT syndrome (Bellocq et al., 2004;

Morita et al., 2008), as well as atrial fibrillation (Chen et al., 2003). Yet no patient with a *KCNQ1* mutation, whether gain or loss of function, has been reported to suffer from a glycaemic disorder – and neither do *Kcnq1*-null mice exhibit glycaemic phenotypes (Pan et al., 2010; Rivas and Francis, 2005). Some studies have reported that blocking the K_v7.1 channel can influence insulin secretion (Ullrich et al., 2005), but experiments within the department have cast doubt upon the specificity of the blocker (chromanol 293B) used in these experiments (Matthias Braun, personal communication). Other voltage-gated potassium channels, notably K_v2.1 and K_v2.2, are important in β -cell electrophysiology (Braun et al., 2008), and it is not clear to what extent results based on chromanol 293B measure inhibition of these channels as well as K_v7.1. A recent study using a more specific blocker, JNJ303, reported a 5% decrease in β -cell K⁺ current and no effect upon exocytosis or insulin secretion. In the same study, siRNA knockdown of *KCNQ1* in human islets did alter exocytosis, but not insulin secretion (Rosengren et al., 2012).

3.1.4 *KCNQ1* lies in the 11p15.5 imprinted gene cluster

A parent of origin-specific T2D risk effect is consistent with the genomic location of *KCNQ1*. It is situated at 11p15.5 (homologous to mouse chromosome 7F5), an imprinted region which has been extensively studied in mice and in relation to human disease (Maher and Reik, 2000; Zollino et al., 2009). In both humans and mice, 11p15.5 contains two imprinted domains, regulated by functionally independent imprinting centres (Casparly et al., 1998; Horike et al., 2000). The (human) telomeric imprinted cluster contains *H19*, *IGF2* and *INS*, but *KCNQ1* lies within the centromeric imprinted

domain, which also contains *KCNQ1DN*, *CDKN1C*, *PHLDA2*, *SLC22A18*, *SLC22A18AS* and *TRPM5*.

The standard view holds that imprinting regulation is via a differentially methylated region (DMR, also known as K_v DMR1) at the promoter of *KCNQ1* overlapping transcript 1 (*KCNQ1OT1*, previously known as *LIT1*; Smilnich et al., 1999), a non-translated antisense RNA whose transcription regulates expression of downstream genes (Thakur et al., 2004). The DMR is unmethylated on a paternally inherited chromosome, allowing transcription of *KCNQ1OT1* and inhibiting expression of downstream genes. Conversely, the DMR is highly methylated on a maternally inherited chromosome, blocking transcription of *KCNQ1OT1* and allowing maternal-specific expression of downstream genes (figure 3.2). In mice, paternal inheritance of a

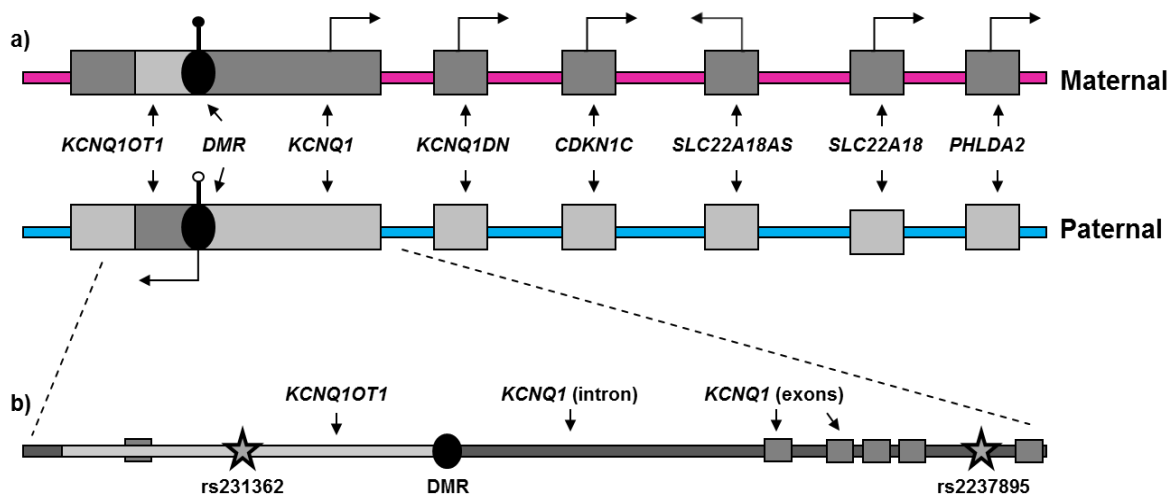


Figure 3.2: Schematic representation of imprinting control (the standard view) and T2D-associated SNPs at 11p15.5. a) Chromosome 11:2,450,000-2,960,000. Closed circle at differentially methylated region (DMR) represents high level of methylation, open circle represents low level of methylation. Transcribed sequence is shown in dark grey, untranscribed in light. Arrows indicate direction of transcription. **b)** Chromosome 11:2,650,000-2,880,000. Smaller region distinguishing exonic (boxes) and intronic regions of *KCNQ1* (dark grey), region of *KCNQ1OT1* transcription (light grey) and relative positions of top T2D-associated SNPs (rs231362, chr11:2,691,471 and rs2237985, chr11:2,857,194). All genomic coordinates are b37/hg19 (figures not to scale).

DMR deletion results in de-repression of downstream genes, whilst maternal inheritance has no effect on gene expression (Fitzpatrick et al., 2002). Removal of *KCNQ1OT1* causes a loss of silencing activity, demonstrating that the antisense transcript is required for imprinting control (Thakur et al., 2004). *KCNQ1OT1* itself is likely to regulate the expression of downstream genes by recruiting DNA methyltransferase 1 (*DNMT1*) to their promoters (Mohammad et al., 2010), or by interacting with chromatin to recruit repressive histone modifications (Nagano and Fraser, 2009; Redrup et al., 2009).

The DMR itself has methylation-sensitive chromatin insulator properties. Studies in mice have shown that it inhibits expression when methylated and inserted between gene enhancers and promoters, yet when unmethylated it has no such inhibitory effect (Kanduri et al., 2002; Mancini-DiNardo et al., 2003). Further studies have demonstrated the existence of key regulatory sites within the DMR. At a putative *PLAGL1*-binding site, *PLAGL1* protein binds preferentially to unmethylated paternal alleles and functions as transcriptional activator of *KCNQ1OT1* (Arima et al., 2005). In mice, there are two CTCF binding sites which lie in the DMR but just outside the *KCNQ1OT1* promoter region. They are occupied only on the unmethylated paternal allele and show repressive activities in enhance-blocking assays (Fitzpatrick et al., 2007). This may be an independent regulatory mechanism to *KCNQ1OT1*, in which bound CTCF creates a repressive expression environment for downstream genes on the paternal chromosome (Fitzpatrick et al., 2007).

3.1.5 Genes in the 11p15.5 imprinted cluster are of relevance to β -cell proliferation and foetal growth

11p15.5 has a well-established role in phenotypes relevant to T2D. Disruption of regional genomic architecture can cause the congenital growth and hyperglycaemic disorder Beckwith-Wiedemann syndrome (BWS), characterised by pre and post-natal overgrowth and macroglossia as well as hypoglycaemia. BWS has a complex aetiology. It can arise from mosaic partial paternal uniparental disomy of chromosome 11, paternal duplication of 11p15.5 (Zollino et al., 2009), maternal-specific deletion of *KCNQ10T1* (Niemitz et al., 2004) and a loss of imprinting of *IGF2*. But by far the most common cause (50-60% of cases) is loss of maternal methylation at the DMR, leading to abnormal biallelic expression of *KCNQ10T1* and inhibiting expression of downstream genes from the maternal as well as normally repressed paternal chromosome (Lee et al., 1999).

Of particular interest amongst these downstream genes is the cyclin-dependent kinase inhibitor *CDKN1C* (encoding p57^{kip2}). *CDKN1C* is a negative regulator of cell proliferation, binding to a variety of cyclin-dependent kinase complexes and causing arrest of cell cycling in the G1 growth phase. It may be one of the major causal factors in BWS, since germline point mutations in *CDKN1C* are sufficient to cause BWS in rare cases (Hatada et al., 1996). Fibroblasts from BWS patients with maternal hypomethylation at the DMR show diminished expression of *CDKN1C* (Diaz-Meyer et al., 2003).

CDKN1C also has established roles in islet-specific growth phenotypes. Within a normal pancreas, p57^{kip2} is expressed exclusively in endocrine tissue and almost exclusively in β -cells. It is expressed by 30-40% of β -cells at any one time in healthy

individuals, but never concurrently with Ki67 (a marker of cell proliferation), suggesting that the expression of p57^{kip2} inhibits proliferation (Kassem et al., 2001).

The role of *CDKN1C* in β -cell proliferation is further evidenced by its expression patterns in patients suffering from focal Hyperinsulinism of Infancy (focal HI). Focal HI is characterised by pancreatic lesions, formed as a result of dramatically increased β -cell proliferation rates (Sempoux et al., 1998). Within the cells of these hyperproliferative lesions, *CDKN1C* expression is completely abolished (Giurgea et al., 2006; Kassem et al., 2001), but is retained at normal levels in unaffected tissue of the same individuals. It appears, therefore, that *CDKN1C* plays a key role in regulating normal β -cell proliferation.

3.1.6 The *KCNQ1* locus may be relevant to birth weight

Common variants in *KCNQ1* are associated with T2D risk and insulin secretion. They are therefore also good candidates for association with birth weight, under the hypothesis that foetal T2D risk alleles may inhibit pre-natal growth by reducing foetal insulin in utero (Andersson et al., 2010; Freathy et al., 2009), whilst maternal T2D risk alleles may associate with higher offspring birth weight through raised maternal glucose levels (Freathy et al., 2007). The relevance of 11p15.5 to the pre-natal growth syndrome BWS also makes it a particularly good candidate for a birth weight effect. Although the sample size was small, one study has suggested that a 12-bp coding indel in *CDKN1C* (del171APVA) may be associated with increased birth weight (Nielsen et al., 2005).

Data produced by the Early Growth Genetics (EGG) consortium, in collaboration with Dr. Rachel Freathy and my supervisor, supports this hypothesis. Figure 3.3 shows the association of SNPs representing 37 T2D-associated loci with birth weight. Six are significantly ($p < 0.05$) associated, demonstrated by deviation from the expected shaded area (Horikoshi et al, under review at Nature Genetics). Birth weight effects at *ADCY5* and *CDKAL1* are well-established (Freathy et al., 2009; Freathy et al., 2010), but this plot suggests that the *MTNR1B*, *GCK*, *HHEX-IDE* and *KCNQ1* loci may also be associated with birth weight as well as T2D.

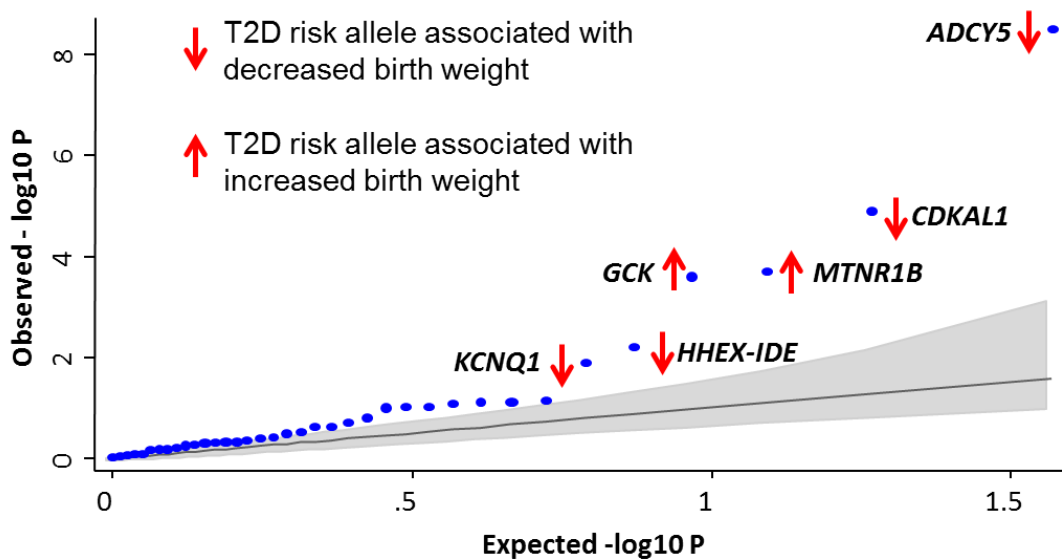


Figure 3.3: Significance of association with birth weight for 37 T2D-associated SNPs. Expected $-\log_{10}$ p-value (x-axis) is plotted against observed $-\log_{10}$ p-value (y-axis), with the black line representing expected relationship and the grey area denoting 95% confidence interval for that relationship. Red arrows indicate the direction of birth weight effect for T2D risk alleles. Confirmed birth weight SNPs are clear, but *KCNQ1* also lies outside the 95% region of expectation.

3.1.7 Experimental aims

The *KCNQ1* region contains imprinted genes with a role in generalised and islet-specific growth phenotypes, harbours SNPs which contribute to T2D risk in a parent-of-origin specific manner, and is a good candidate for birth weight effects. I therefore hypothesised that T2D risk may be mediated through disruption of methylation and imprinted gene expression - leading to impaired islet proliferation, development or function, and that T2D-associated SNPs may also be associated with altered birth weight. To test these hypotheses, my study aimed to:

- determine the imprinting status of *KCNQ1*, *KCNQ1OT1*, *KCNQ1DN*, *CDKN1C*, *PHLDA2*, *SLC22A18* and *SLC22A18AS* in human adult islet and foetal pancreas;
- assess the effect of T2D risk genotype status (lead SNP at both independently associated loci) upon DNA methylation at relevant regulatory regions;
- assess the effect of T2D risk genotype status upon total expression levels of imprinted genes in the 11p15.5 centromeric cluster (*KCNQ1*, *KCNQ1OT1*, *KCNQ1DN*, *CDKN1C*, *PHLDA2*, *SLC22A18*, *SLC22A18AS* and *TRPM5*);
- assess the effect of T2D risk genotype status upon allele-specific expression of the same imprinted genes;
- assess the effect of T2D risk genotype status upon birth weight, distinguishing between foetal and maternal risk alleles and between maternally and paternally transmitted risk alleles.

Since methylation and gene expression are frequently tissue-specific, experiments were conducted in samples of relevance to a β -cell phenotype; namely human islets. Given the established relevance of 11p15.5 to prenatal growth, methylation and expression were also explored as described above in human foetal pancreas samples.

3.2 Samples and Methods

3.2.1 Samples used

Human islets were obtained with full research consent (Oxfordshire Regional Ethics Committee B) from the Oxford Centre for Islet Transplantation (OXCIT, n=30) and the Human Tissue Laboratory at Lund University Diabetes Centre (LUDC, n=42). Donor characteristics are provided in table 3.2. Islets were not hand-picked and preparations therefore differed in purity (table 3.2). Islet samples from Oxford were collected over two years and DNA/RNA was stored as an islet biobank resource. All donors were of European descent and had no known diabetic phenotype. Islets were cultured at 37°C (5% CO₂) in CMRL 1066 (ICN Biomedicals, USA) supplemented with 10 mmol/l HEPES, 2 mmol/l L-glutamine, 50 µg/ml gentamicin, 0.25 µg/ml Fungizone (GIBCO, BRL, USA), 20 µg/ml ciprofloxacin (Bayer Healthcare, Germany), and 10 mmol/l nicotinamide. Prior to DNA and RNA extraction, islets were suspended in RNA later (Ambion, Warrington, UK) and stored at -80°C.

		LUDC islets, n=42		OXCIT islets, n=30
		Mean (range)		Mean (range)
Adult islets	Age (years)	57	(24 – 75)	46 (22 - 62)
	BMI (kg/m ²)	26.1	(17.6 – 36.6)	29.5 (21.7 – 36.7)
	Islet purity	96%	(89% - 99%)	89% (83% - 99%)
	Gender	48% male 52% female		60% male 40% female

Table 3.2: Donor and sample details for human adult islets and foetal pancreas

Foetal pancreas samples (n=18) were obtained in collaboration with Professor Neil Hanley at the University of Manchester with informed consent and ethical approval (North West Regional Ethics committee). Samples were obtained at mean 79 days (range 42-133 days) post conception.

Adult whole pancreas samples were obtained in collaboration with the Oxford MolSURG project, collected as part of the MolPAGE (Molecular Phenotyping to Accelerate Genomic Epidemiology) consortium. Tissue and blood samples were collected with full research consent (Oxfordshire Regional Ethics Committee B) from participants undergoing elective surgery, snap frozen and stored as whole tissue in liquid nitrogen prior to extraction.

Birth weight association analysis was conducted in the Berlin Birth Cohort (1292 mother-child pairs). Cohort details are provided in table 3.3. Whole blood (for genotyping as described in section 2.5) and phenotype information was obtained in collaboration with Dr. Berthold Hofer at the University of Potsdam, Germany. Meta-analysis of these results and results from other cohorts in the EGG consortium (detailed in table 3.4) was conducted by Dr. Rachel Freathy at the University of Exeter.

Berlin Birth cohort (n=1292 mother-child pairs)	Mean (range)
Birth weight (g)	3468 (1815 – 5405)
Gestational age (weeks)	39 (37 – 49)
Parity	Median 1 (1 – 4)
Child's sex	52% male, 48% female
Maternal age (years)	30 (14 – 45)

Table 3.3: Berlin Birth Cohort phenotypes

Study	Cohort Type	N
Exeter Family Study of Childhood (EFSOCH)	Parent-child trios	700 trios
Avon Longitudinal Study of Parents and Children (ALSPAC)	Mother-child pairs	4000 pairs
Helsinki Birth Cohort (HBCS)	Children only	4500 children

Table 3.4: Early Growth Genetics (EGG) cohorts used in birth weight meta-analysis

3.2.2 RNA and DNA extraction and quality assessment

Phenol-chloroform extraction of DNA and RNA from human adult islets, foetal pancreas and adult pancreas was performed using Trizol as described in sections 2.3 and 2.4. DNA was extracted from human adult spleen and whole blood using the Maxwell 16 system, as described in section 2.4. DNA was suspended in 1xTE (minimum volume 100µl). RNA was suspended in RNase free water (minimum volume 20µl) and checked for integrity using a Bioanalyser 2100 (section 2.3.5). Samples with RNA Integrity Numbers (RIN scores) <7 were excluded from further analysis.

3.2.3 SNP selection and Genotyping

T2D-associated SNPs were selected for genotyping on the basis of their disease association strength (table 3.1). In the largest European meta-analysis to date, rs231362 is the lead SNP in the intron 10 signal (Voight et al., 2010). rs2237892 was initially reported as the intron 15 lead SNP amongst East Asian populations, whilst rs163184 was the intron 15 lead SNP in European populations. rs2237895 is not in HapMap and was therefore not included in the original GWA studies, but subsequently emerged through fine-mapping as the lead intron 15 SNP in both populations (Holmkvist et al., 2009). Assay details are provided in table 3.5.

	SNP	Location	Assay ID	PCR cycles
Intron 10 signal	rs231362	Chr11:2648047	C__3075844	40
	rs2237892	Chr11:2796327	C_16171025	40
Intron 15 signal	rs2237895	Chr11:2813770	C_16171034	50
	rs163184	Chr11:2803645	C__3075708	40

Table 3.5: Genotyping assay details - T2D-associated SNPs. Locations are b36/hg18. assay IDs refer to Applied Biosystems TaqMan genotyping assays.

Two reporter coding SNPs each in *KCNQ*, *KCNQ1OT1*, *PHLDA2*, *SLC22A18* and *SLC22A18AS* were genotyped for imprinting and allele-specific expression analysis. SNPs with maximum possible MAFs were selected, in order to maximise the number of heterozygous samples able to differentiate mRNA products from homologous chromosomes. Assay details are provided in table 3.6.

Gene	SNP	Location	MAF	Assay ID	PCR cycles
<i>KCNQ1</i>	rs1057128	Chr11: 2753813	0.17	C__11364860	40
	rs8234	Chr11: 2870108	0.38	C__11368280	40
<i>KCNQ1OT1</i>	rs231362	Chr11: 2691471	0.48	C__3075844	40
	rs760419	Chr11: 2683357	0.43	C__3075849	50
<i>PHLDA2</i>	rs13390	Chr11: 2950558	0.20	C__8793507	40
	rs1056819	Chr11: 2949861	0.16	C__8793508	40
<i>SLC22A18</i>	rs1129782	Chr11: 2943671	0.19	C__8793517	40
	rs1048047	Chr11: 2924610	0.01	C__2600571	50
<i>SLC22A18-AS</i>	rs367035	Chr11: 2923826	0.44	C__2278305	40
	rs10741735	Chr11: 2920283	0.50	C__2600562	40

Table 3.6: Genotyping assay details - coding SNPs for allele-specific expression analysis. Locations are b36/hg18, MAF = minor allele frequency (HapMap CEU samples). Assay IDs refer to Applied Biosystems TaqMan genotyping assays.

All genotyping was performed using TaqMan technology, as described in section 2.5. Two-step PCR was performed on an Applied Biosystems 7900HT, before post-read allelic discrimination using SDSv2.3 software (Applied Biosystems).

3.2.4 cDNA synthesis

cDNA was generated via random primed first strand synthesis using the QuantiTect reverse transcription kit (Qiagen, Crawley, UK) from 1µg RNA as detailed in section 2.6.

3.2.5 Fragment analysis

Indel genotype at del171APVA was determined by fragment analysis of a fluorescently tagged PCR, as described in section 2.9. FAM-labelled primers (F- TGGACCGAAGTGGACAGCGA, R- GGGGCCAGGACCGCGACC, Eurofins) were designed as described in section 2.1.1 and PCR was performed as described in section 2.1 using HotStar Taq and reagents (Qiagen).

3.2.6 Sequencing

Primers for PCR amplification were designed as described in section 2.1.1. Primer sequences and accompanying reagents are detailed in table 3.7. Bidirectional sequencing was performed using universal M13 primers as detailed in section 2.2 and analysed as outlined in section 2.2.4.

3.2.7 Total mRNA expression quantification

Total gene expression was quantified using TaqMan gene expression assays (Applied Biosystems) as described in section 2.7. Genes were selected for analysis on the basis of inclusion in the GeneImprint database (www.geneimprint.com) and expression regulation by *KCNQ10T1*'s differentially methylated promoter (*KCNQ1*, *KCNQ10T1*, *CDKN1C* (2 assays for comparison), *KCNQ1DN*, *PHLDA2*, *SLC22A18*, *SLC22A18AS* and *TRPM5*. *IFG2*, *H19* and *INSR* were also analysed as nearby genes (in the adjacent but independently regulated imprinted domain) with strong biological candidacy. Assay details are provided in table 3.8.

Gene	Primer type		Primer Sequence	Reagents
KCNQ1	cDNA	FWD	AAGGGGAGACTCTGCTGACA	Applied Biosystems
		REV	AAGGAGAGCAGCTGGTGAAG	
	gDNA	FWD	TGCCGGTGAGTAGACAGGAA	Applied Biosystems
		REV	CGGGGAACAAGGTAGAGCTG	
KCNQ10T1	gDNA	FWD	CTACCAAGGATGTGGGCTGT	Applied Biosystems
		REV	TACAGGCGGAAACCTAGCAC	
	cDNA		None – intronic transcript	
	PHLDA2	gDNA	FWD	GCTCGGCACGACATGAAA
REV			CTTGAGGATGGAGTGAAGC	
cDNA		None – insufficient intronic sequence		
SLC22A18		cDNA	FWD	GGGCTCTTCATGGTCATGTT
	REV		CGGACTTTGTCCTTCCTCTG	
	gDNA	FWD	GGGGAGGGACAGGTAAGACC	Applied Biosystems
		REV	CAGCTGGGCTAGGTGGTAGG	
SLC22A18-AS	cDNA	FWD	TACTCACTGGCACGATGGAG	Qiagen
		REV	CATTCCCTCAGAACCAGGAA	
	gDNA	FWD	CTCTGCCTTACCTCCCACAG	Qiagen
		REV	GGTCCCTGTTGCAAGTCATC	

Table 3.7: Primer sequences and PCR reagents for gDNA and cDNA sequencing

Gene	Assay ID	Experimental dilution
<i>KCNQ1</i>	Hs00923522_m1	1:10
<i>KCNQ10T1</i>	Hs03665990_s1	1:10
<i>CDKN1C</i>	HS00175938_m1	1:10
<i>CDKN1C_2</i>	Hs00908986_g1	1:10
<i>KCNQ1DN</i>	Hs00218884_m1	1:5
<i>PHLDA2</i>	Hs00169368_m1	1:10
<i>SLC22A18</i>	Hs00180039_m1	1:10
<i>SLC22A18AS</i>	Hs00757934_m1	1:10
<i>TRPM5</i>	Hs00175822_m1	1:5
<i>IGF2</i>	Hs01005963_m1	1:100
<i>INSR</i>	Hs02741908_m1	1:100
<i>H19</i>	Hs00262142_g1	1:100

Table 3.8: Assay details for total expression quantification. Assay IDs are TaqMan inventoried gene expression assays (Applied Biosystems).

cDNA samples were diluted in 0.01M Tris and standard curves generated by serially diluting a pool of all cDNA samples as detailed in section 2.7.3.

C_T scores for the two *CDKN1C* assays showed excellent correlation ($r^2 = 0.98$, figure 3.4). Since one assay was designed across exon boundaries (_m1) whilst the other was within a single exon (_g1), the highly comparable results indicate that there is no significant gDNA contamination in the sample set. The results also provide reassuring evidence of highly consistent experimental conditions across plates.

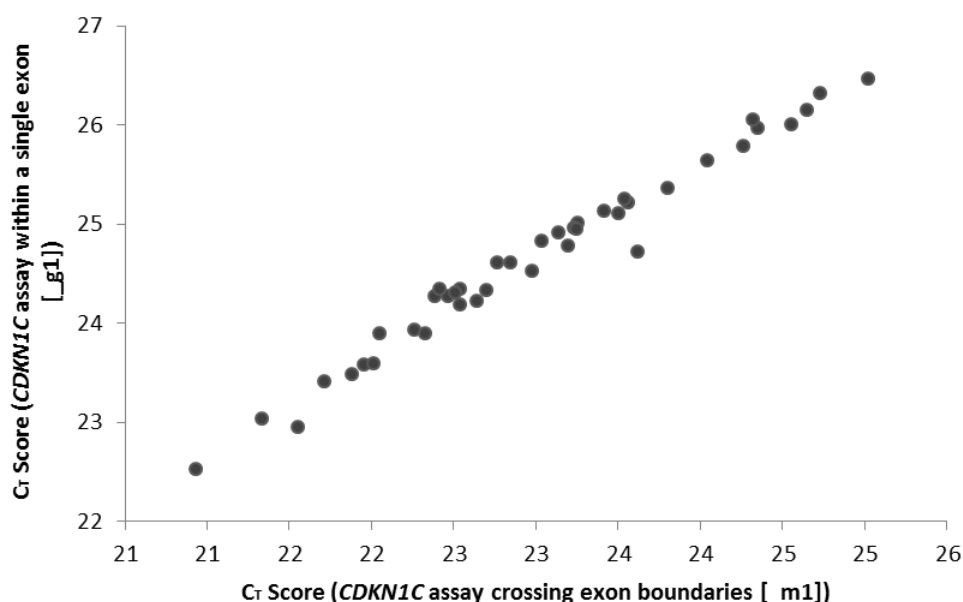


Figure 3.4: Comparison of two alternative total expression assays for *CDKN1C*. _m1 assay (x-axis), designed across exon boundaries and therefore unable to amplify gDNA against _g1 assay (y-axis), designed within a single exon and therefore able to amplify gDNA. Strong correlation ($r^2=0.98$) indicates experimental consistency and lack of contaminating gDNA.

Each reaction multiplexed a FAM-labelled test assay and a VIC-labelled endogenous control assay (one of *HPRT*, *PPIA* or *RNaseP*), after standard curve validation showed that multiplexing had no impact upon amplification efficiency. QPCR was performed on

an Applied Biosystems 7900HT machine in accordance with manufacturer's guidelines and as detailed in section 2.7.4. All samples were run in triplicate, and in the presence of a reverse-transcriptase negative control. Any samples with amplification in the negative control wells (1 sample) or with a coefficient of variance between replicates >1 (4 samples) were excluded from further analysis. Analysis was performed using the $\Delta\Delta C_t$ method as described in section 2.7.5, with respect to the pooled sample standard dilution as appropriate and normalised to the geometric mean of endogenous controls.

3.2.8 Allele-specific mRNA expression quantification

Allele-specific expression was quantified using the genotyping assays detailed in table 3.6. These assays contain two differentially labelled (VIC/FAM) probes, annealing over the two alleles of a coding SNP. Quantitative PCR was conducted as for total expression quantification, but in the absence of an endogenous control assay. Analysis was conducted by dividing the expression level of the most highly expressed allele by the expression level of the more lowly expressed allele, to provide an indication of the balance between the two. A ratio of one would reflect equal expression from the two alleles, whilst ratios deviating further from one would represent increasingly skewed allelic expression. For *CDKN1C*, the same analysis was conducted, using fragment analysis peak heights from the two alleles.

3.2.9 Islet purity analysis

All islet samples were assessed for purity by comparing expression levels of three endocrine markers as described in section 2.7.6. Samples of less than 80% purity were

excluded from further analysis, leaving a sample set of high endocrine purity (table 3.2). Islet purity was used as a covariate in statistical analysis. Although foetal samples were obtained at too early a developmental stage to isolate islets (mean 79 days post conception, range 42-133 days), all samples (with the single exception of the 42-day sample) showed expression of insulin, somatostatin and glucagon.

3.2.10 Bisulphite-treatment and Pyrosequencing

500ng DNA was bisulphite-treated using the EZ methylation-gold kit (Zymo research, Orange, CA, UK) as detailed in section 2.10.1 and eluted in 25µl 8mN NaOH. Methylation levels were quantified at five regulatory regions of interest, comprising:

- Sequence within *KCNQ1OT1*'s differentially methylated promoter region used for diagnosis of BWS (Mackay et al., 2008). Subsequently termed "DMR assay"
- CTCF binding-site sequence homologous to the identified methylation-sensitive CTCF binding site at mouse chromosome 7F4 (Fitzpatrick et al., 2007)
- Putative PLAGL1 binding-site sequence (Arima et al., 2005)
- Candidate assay_1: designed on the "shore" of the differentially methylated CpG island, chosen on the basis of proximity to numerous CpG dinucleotides and the strongly T2D-associated SNP rs231354
- Candidate assay_2: designed on the "shore" of the differentially methylated CpG island, chosen on the basis of proximity to numerous CpG dinucleotides and the strongly T2D-associated SNP rs2283202

These assays interrogate the three known regulatory regions, and two candidate regions representing strongly T2D-associated SNPs in close proximity to CpG dinucleotides.

Primers were designed using Pyromark software (sequences detailed in table 3.9) and biotinylated on the reverse strand (Operon Biotechnologies, Ebersberg, Germany). Where two alternative sequencing primers are detailed for a single PCR primer set, a double volume of PCR amplification product was produced and the mean of results from the two sequencing primers was used for analysis.

Assay	Primer type		Primer Sequence
DMR	PCR	FWD	TAATTAGTAGGTGGGGGG
		REV	CTAAAAAACTCCCTAAAAATC
	Sequencing		GGGGGTAGTCGGAG
CTCF binding-site	PCR	FWD	ATTTGAGTAGTTGGGGTTTTAGTT
		REV	TCAAAACCTCTACCCAAAAACAAA
	Sequencing_1		TAGTTGGGGTTTTAGTTAGT
	Sequencing_2		GTTGAGATTTAAGGGAGG
PLAGL1 binding-site	PCR	FWD	TGTTTAGGTTAGGTTGTATTGTTG
		REV	CTCCCCATCTCTCTAAAAAATTT
	Sequencing_1		GTATATAGTTTATTTTAGTA
	Sequencing_2		GTTTGTGATTTGGGA
Candidate assay_1	PCR	FWD	TTTTTGTGGGTTGGTTTGTTAAG
		REV	TCTACCAATCACCCAATATTCAAA
	Sequencing_1		AGATTTTAGATTTTTTATTA
	Sequencing_2		GTTATAGTAGGTTTAAAGTT
Candidate assay_2	PCR	FWD	ATGGGAGGATTGAAGTAGTGTTTT
		REV	TATCTTCTTACCATCCCTATCTCA
	Sequencing		TTTATATTTGATTAGTTTTG

Table 3.9: Primer sequences for pyrosequencing assays

One µl of treated DNA was amplified and pyrosequencing was performed on a PSQ96A machine (Qiagen) as described in section 2.10.3. Sequences for analysis and base dispensation orders are provided in table 3.10. Analysis was performed using Pyromark software v2.0 (Qiagen), including cytosine bases outside CpG dinucleotides to confirm efficiency of bisulphite conversion. Each assessed sequence contained two CpG dinucleotides (not including bisulphite controls), and results presented represent the mean methylation of CpG sites within each assay. Each pyrosequencing reaction was conducted in duplicate, using product from each of two independent PCR amplifications. Results are therefore the mean of four reactions for each sample. Three samples with high variability between replicates (coefficient of variance >1) were excluded from further analysis.

Assay	Sequence for analysis	Dispensation order
DMR	T/CGTTGTT/CGT/CAGTT	ATCAGTGTCAGTCGA
CTCF binding site_1	TT/CGGGTTTGGAGTTT/CGT	ATCAGTGAGTCTG
CTCF binding site_2	TTT/CGAGTTGGT/CGATTGGA	ATCTGAGTGTCTG
PLAGL1 binding site_1	AT/CGTTAGTGATTATTT/CGTTT/C GT/CGT	GATCAGTAGTGATATCAGTC AGTCTG
PLAGL1 binding site_2	T/CGGTT/CGT/CGGGTATATAGTTT ATTTT	ATCAGTCAGTCAGTA
Candidate assay_1_1	T/CGGAAGGTAGTT/CGTG	ATCTGAGTAGTCTG
Candidate assay_1_2	TT/CGGAGATTTTAGATTTTTTATTAT /CGGAA	TCTAGCAGATAGATTATATC TG
Candidate assay_2	ATT/CGTTAGATTTTTT/CGTG	GATCGAGTAGATCTG

Table 3.10: Sequences and dispensation orders for pyrosequencing analysis. Potentially methylated cytosines are highlighted in grey. Bisulphite controls are bordered.

3.2.11 Statistical procedures

Islet samples were obtained from Lund (n=42) and Oxford (n=30). Results from the two centres for all methylation and expression assays were not significantly different ($p \geq 0.05$, Kolmogorov-Smirnov independent samples test). All islet samples were therefore combined for further analysis, but centre of origin was included as a covariate where appropriate. Expression and methylation was compared between tissue types via a Kolmogorov-Smirnov independent samples test with a threshold significance value of 0.05. Analysis of genotype groups was conducted by linear regression, with donor age, donor gender, sample purity and centre of origin (for adult islets) or days post-conception (for foetal pancreas) included as covariates. Analysis was also performed using a Kruskal-Wallis analysis of ranks, which retained significance in all cases. All statistical tests were performed in SPSSv18.0.

3.3 Results

3.3.1 Genotyping of islet and foetal pancreas samples

A total of 14 SNPs were genotyped via TaqMan assays (figure 3.4), across both association signals (4 assays, table 2.5) and in the coding regions of genes of interest for allele specific expression analysis (10 assays, table 2.6).

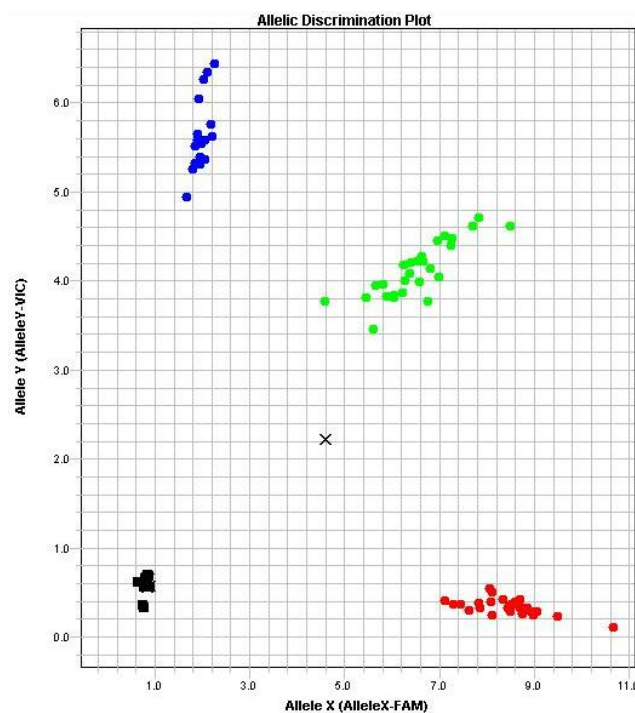


Figure 3.5: Example genotyping cluster plot (rs10747135) as produced by SDSv2.3 software. FAM (G allele) fluorescence (x-axis) is plotted against VIC (A allele) fluorescence (y-axis). **Red points** represent GG homozygous samples, **blue points** represent AA homozygotes and **green points** represent AG heterozygotes. Black cross (X) denotes a samples of undetermined genotype, whilst black squares represent non-template controls (dH₂O) which should not amplify.

All SNP assays resulted in successful genotype calls for >90% of samples (genotyping pass rates >90%), were present in accordance with expected (HapMap CEPH) minor allele frequencies and did not depart from Hardy-Weinberg equilibrium ($p > 0.30$ in all cases) (tables 3.11 and 3.12).

SNP	Genotyping pass rate	HapMap MAF	Experimental MAF	Hardy-Weinberg p-value
rs231362	99%	0.48	0.50	0.46
rs2237892	97%	0.08	0.08	0.80
rs2237895	95%	-	0.46	0.82
rs163184	98%	0.44	0.38	1.00

Table 3.11: Genotyping results - T2D-associated SNPs. MAF = minor allele frequency. HapMap allele frequency data in CEPH (European) population. rs2237895 is not in HapMap and does not therefore have a CEPH reference allele frequency. Previous studies have placed its MAF at around 0.48.

SNP	Genotyping pass rate	HapMap MAF	Experimental MAF	Hardy-Weinberg p-value
rs1057128	98%	0.17	0.17	0.83
rs8234	98%	0.38	0.33	1.00
rs231362	95%	0.48	0.50	0.46
rs760419	96%	0.43	0.49	0.83
rs13390	99%	0.20	0.12	0.95
rs1056819	99%	0.16	0.10	0.41
rs1129782	93%	0.19	0.33	0.31
rs1048047	97%	0.42	0.39	0.67
rs367035	99%	0.44	0.69	0.37
rs10741735	98%	0.50	0.56	0.36

Table 3.12: Genotyping results - SNPs for use in allele-specific expression analysis. MAF = minor allele frequency. HapMap allele frequency data in CEPH (European) population

3.3.2 Gene expression at the 11p15.5 cluster

KCNQ1, *KCNQ1OT1*, *CDKN1C*, *PHLDA2*, *SLC22A18* and *SLC22A18AS* were all quantifiably expressed in both human adult islets and foetal pancreas. *KCNQ1DN* and *TRPM5* showed no expression in either tissue (probes were successfully amplified in other tissues) and were therefore not carried forward for further analysis (figure 3.6).

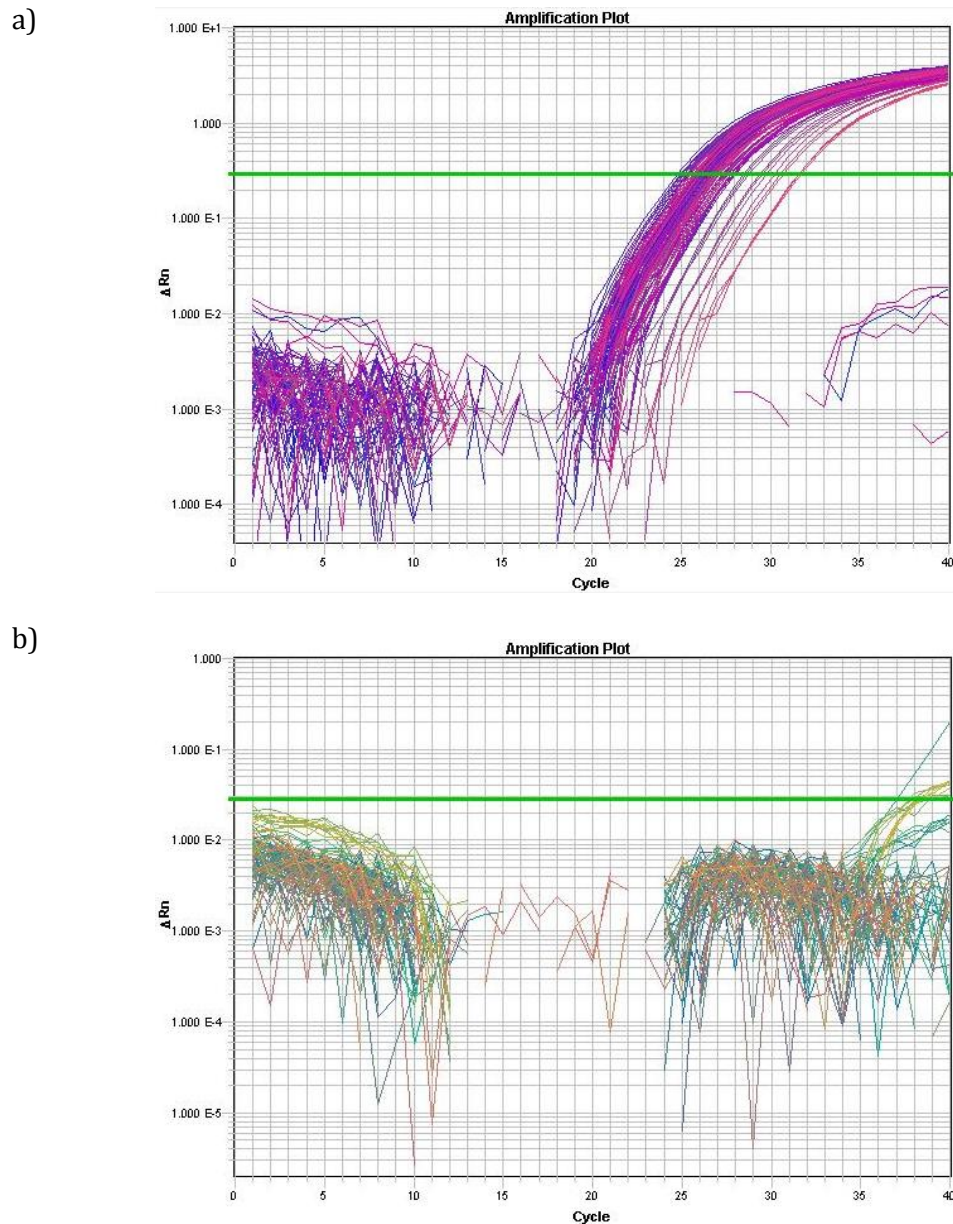


Figure 3.6: Total expression of genes at 11p15.5: example amplification plots. PCR cycle number (x-axis) is plotted against normalised fluorescence intensity (y-axis). Green line represents critical threshold (as set by SDSv2.3 software). **a)** Example gene (*PHLDA2*) with quantifiable expression. Non-amplifying wells on the right hand site represent non-template controls. **b)** Example gene (*KCNQ1DN*) without quantifiable expression.

3.3.3 Genes at the 11p15.5 cluster show temporal differences in imprinting status

3.3.3 (a) *CDKN1C* is imprinted in both adult islets and foetal pancreas

Common coding SNPs were used to determine imprinting status of *KCNQ1*, *KCNQ10T1*, *PHLDA2*, *SLC22A18* and *SLC22A18AS* (table 2.6). *CDKN1C* does not contain any coding SNPs with a MAF over 0.1, and was therefore unlikely to yield a useful number of heterozygous samples. It does, however, harbour a 12-bp coding insertion deletion (indel; del171APVA) with a MAF of approximately 0.26 (Nielsen et al., 2005).

Indels can be more expensive and labour-intensive to type than SNPs. We therefore attempted to identify an indel-tagging SNP for ease of future experiments. del171APVA was typed via fragment analysis in 90 HapMap CEPH samples, and the resulting indel genotypes combined with publicly available SNP genotype data for those samples (<http://hapmap.ncbi.nlm.nih.gov/>) using Haploview software (v4.1). However, no effective tagging SNP for the indel was identified (figure 3.7). Its strongest relationship



Figure 3.7: Linkage disequilibrium (LD) plot of del171APVA and regional SNPs. Plot numbered by r^2 (0-100), coloured by D' . del171APVA is represented by marker 41 (marked with arrow). It does not lie within a region of strong LD, and its strongest relationship is with marker 54 (rs2237901), with an r^2 of 0.59.

was with rs2237901, but an r^2 of 0.59 is not sufficient to reliably impute indel status from SNP genotype. Imputation accuracy could not be improved by genotyping multiple SNPs. Further allele-specific expression work on *CDKN1C* was therefore conducted via fragment analysis of fluorescently-tagged PCR product to directly type the indel.

In samples heterozygous for del171APA, the resulting 12-bp fragment size difference was used to distinguish between mRNA products from homologous chromosomes. Imprinting status was determined by the presence of either one size peak (monoallelic expression) or two size peaks (biallelic expression) in cDNA from samples shown to be heterozygous in (gDNA) from the same individual.

Consistent with expectation for a gene in an imprinted region, *CDKN1C* was monoallelically expressed in both adult and foetal samples (figure 3.8). In all cases,

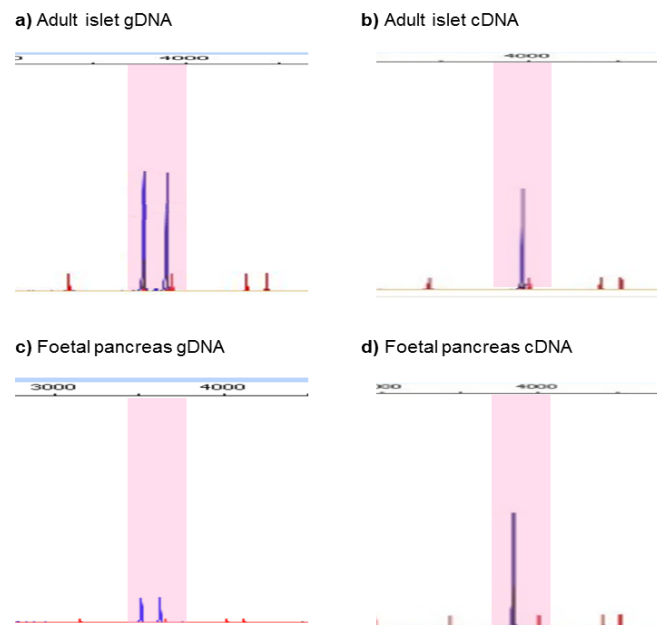


Figure 3.8: Fragment analysis traces demonstrating monoallelic (imprinted) expression of *CDKN1C* in adult islet (top panels) and foetal pancreas (bottom panels) samples. Left hand panels show genomic DNA, with the two size peaks characteristic of a heterozygote for del171APVA highlighted. Right hand panels show complementary DNA from the same heterozygous samples. In both cases, only one size peak is evident, indicating the presence of mRNA from only one chromosome.

samples heterozygous for del171APVA at the gDNA level (n=23 adult islet heterozygotes, n=5 foetal pancreas heterozygotes) appeared homozygous at the cDNA level, indicating gene transcription from only one chromosome. Parental origin could not be determined in the absence of parental DNA, but, on the basis of previous studies (Smilnich et al., 1999; Thakur et al., 2004), is likely to be maternal.

3.3.3 (b) *KCNQ1* and *KCNQ1OT1* are imprinted in foetal pancreas but biallelically expressed in adult islets

Imprinting status of *KCNQ1*, *KCNQ1OT1*, *PHLDA2*, *SLC22A18* and *SLC22A18AS* was determined using common coding SNPs (table 2.6) to distinguish between mRNA products from homologous chromosomes. Gene expression analysis compared the proportion of each allele present in cDNA from human islets and foetal pancreas (section 3.2.8). cDNA from heterozygous individuals is informative. In a biallelically expressed gene, cDNA would remain heterozygous, with equal proportions of both alleles present. In an imprinted gene, however, cDNA would appear homozygous, indicating transcription from only one chromosome.

For *KCNQ1* and *KCNQ1OT1*, temporal changes in imprinting status were observed. Figure 3.9 shows the ratio of expression between alleles in homozygous samples (left and right hand clusters of each plot) and heterozygous samples (centre cluster of each plot). Heterozygous adult islet samples (n=16 and n=30 in *KCNQ1* and *KCNQ1OT1* respectively) and heterozygous foetal pancreas samples (n=3 and n=9 respectively) show clearly different patterns. In adult islets, the allelic expression of heterozygotes is around one, indicating equal expression of the two alleles and falling in the centre of the two homozygous sample clusters. In foetal pancreas samples however,

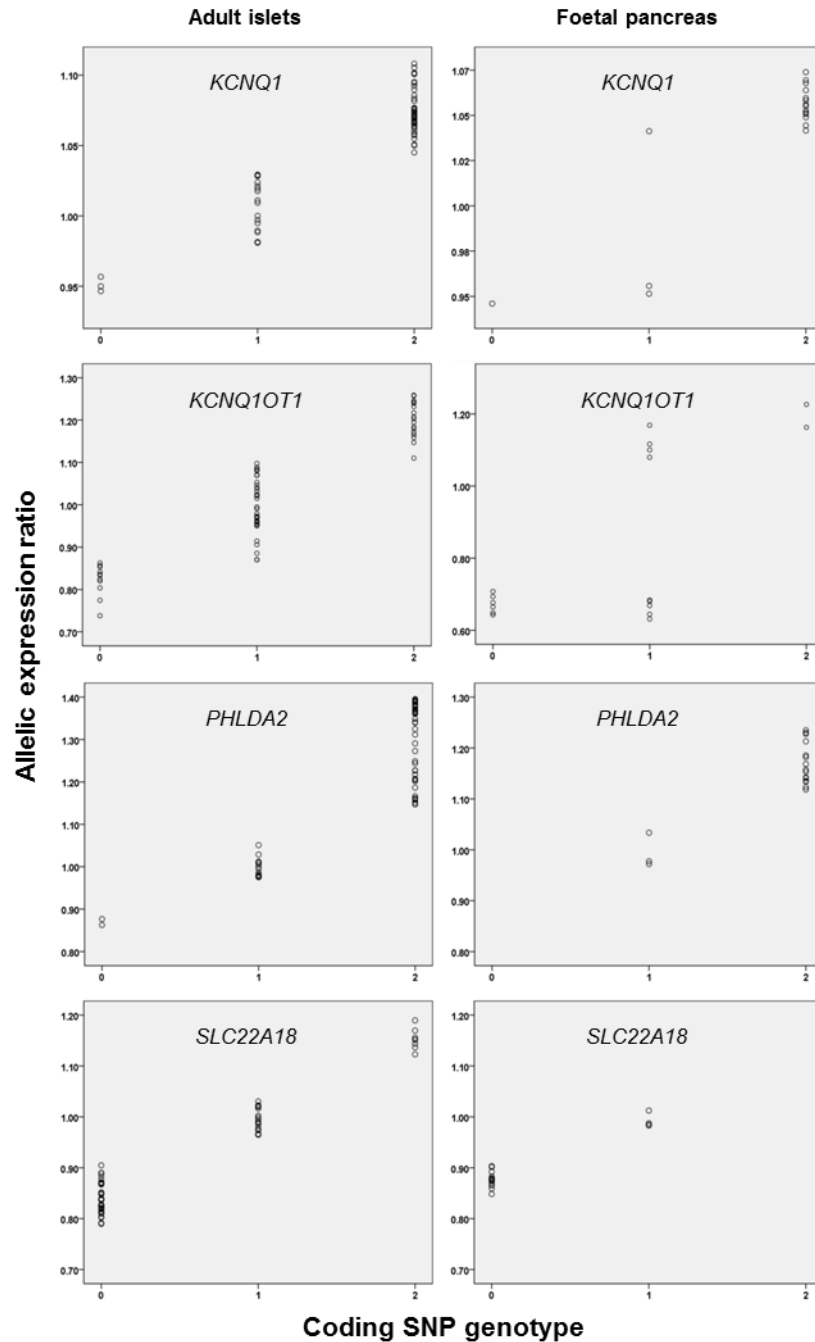


Figure 3.9: Flexibility of imprinting at the 11p15.5 cluster. X-axis groups represent genotypes of reporter coding SNPs within the relevant gene (numbering arbitrary): left and right hand clusters contain the two homozygote groups, whilst the centre cluster contains heterozygous samples. Y-axis values are a ratio of mRNA expression level of a fluorescent probe specific to one allele of the coding SNP, against expression level of a differently labelled probe specific to the alternative allele. The two homozygous groups are therefore expected to have relatively high and low ratios, representing substantially more amplification of the probe specific for one allele. Under biallelic expression (all adult islet samples, *PHLDA2* and *SLC22A18* foetal pancreas samples), heterozygous samples form a central cluster with a ratio around 1, indicating equal expression from homologous chromosomes. Under monoallelic (imprinted) expression (*KCNQ1* and *KCNQ1OT1* foetal pancreas samples), heterozygote samples separate into two distinct groups, corresponding to the two homozygous clusters and indicating expression from only one chromosome.

heterozygous samples split into two expression ratio clusters, corresponding to each of the homozygous sample clusters. This indicates imprinted expression of *KCNQ1* and *KCNQ10T1* in foetal pancreas, but biallelic expression all in adult islets. These results were confirmed by sequencing cDNA from all heterozygous samples (figure 3.10).

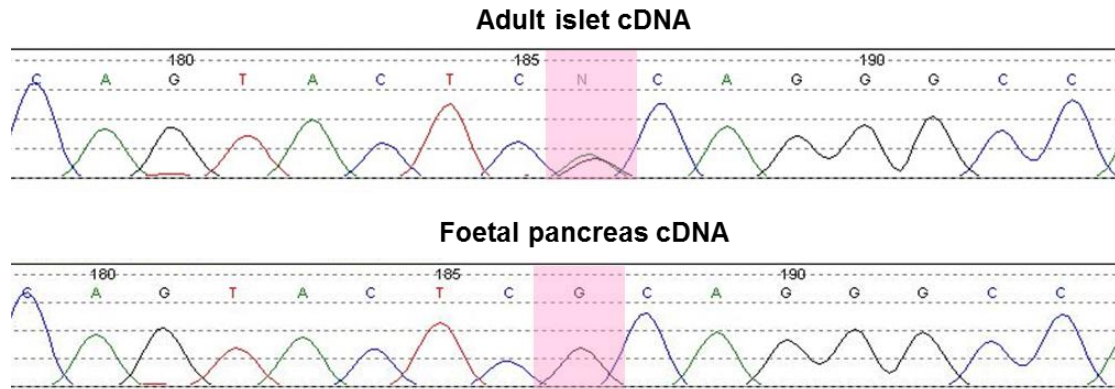


Figure 3.10: Electropherograms demonstrating temporal changes in imprinting status at *KCNQ1*. Electropherogram traces from capillary sequencing of cDNA across a heterozygous coding SNP within *KCNQ1* (highlighted in pink). The adult islet sample retains heterozygosity, indicating transcription from both chromosomes, whereas the foetal pancreas sample appears monoallelic – revealing imprinted expression from only one chromosome. The same pattern was seen for all samples at *KCNQ1* and *KCNQ10T1*.

3.3.3 (c) *PHLDA2*, *SLC22A18* and *SLC22A18AS* are biallelically expressed in both adult islets and foetal pancreas

The allelic ratio expression pattern for both adult islet and foetal pancreas samples at *PHLDA2* and *SLC22A18* is the same as that for adult samples at *KCNQ1* and *KCNQ10T1* (figure 3.9). Heterozygous samples at both developmental stages (n=13 and n=14 for adults islets in *PHLDA2* and *SLC22A18* respectively, n=3 and n=4 for foetal pancreas) form a cluster around one and between the two homozygous clusters, indicating biallelic expression. Allele specific assays for *SLC22A18AS* were not successful, but capillary sequencing revealed the same pattern (figure 3.11). Heterozygous samples

from *PHLDA2* and *SLC22A18* were also sequenced across their coding SNPs and confirmed biallelic expression in both adult and foetal samples.

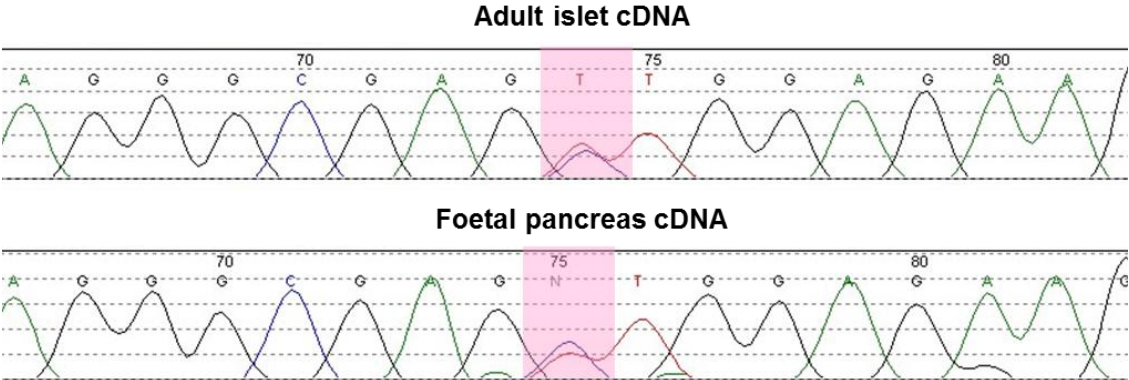


Figure 3.11: Electropherograms demonstrating biallelic expression at *SLC22A18AS*. Electropherogram traces from capillary sequencing of cDNA across a heterozygous coding SNP within *SLC22A18AS* (highlighted in pink). Both samples retain heterozygosity, indicating transcription from both chromosomes

The data above show that imprinting of genes at this locus is not universal. Not only does it vary between genes mapping to the region, but it also varies across developmental stages (summarised in figure 3.12).

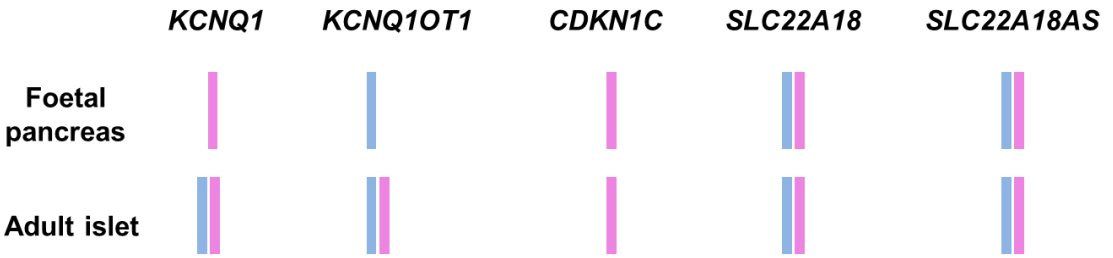


Figure 3.12: Imprinting status of genes at 11p15.5 in foetal pancreas and adult islet samples. Two chromosomes represents biallelic expression; one chromosome represents monoallelic (imprinted) expression. Note that maternal (pink) or paternal-specific (blue) expression has not been demonstrated in these experiments, but is assumed on the basis of data outlined in section 3.1.4.

3.3.4 T2D risk genotype status influences regional methylation in a developmentally flexible manner

Gene expression at the 11p15.5 region is thought to be governed by DNA methylation at the *KCNQ10T1* differentially methylated promoter region (outlined in section 3.1.4). We hypothesised that any influence of T2D risk SNPs on neighbouring gene expression would be mediated through an effect on DNA methylation in this region. Quantitative measurements of methylation were obtained via pyrosequencing (figure 3.13) of bisulphite-induced C/T polymorphisms in five distinct regions of sequence (section 3.1.10).

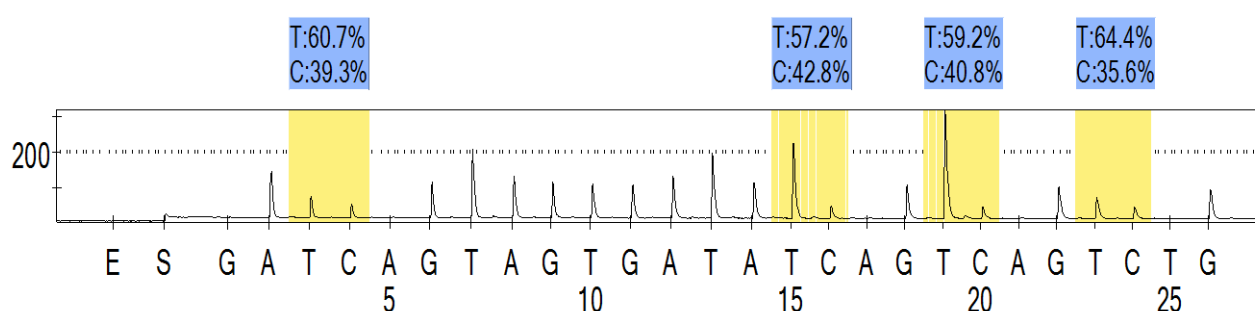


Figure 3.13: Pyrosequencing trace for methylation quantification. Relative peak heights at a Bisulphite-induced C/T polymorphism (highlighted in yellow) are used to quantify methylation at that cytosine residue. Note that “T” peak heights at polymorphisms 2 and 3 are increased by >1 thymine base in the preceding sequence.

There were no significant differences in methylation between adult and foetal samples for 4 of the 5 regions tested (figure 3.14). Mean methylation for adult and foetal samples was 44% and 43% respectively for the diagnostic DMR assay, 42% and 43% for the *CTCF* binding-site, 38% and 40% for the *PLAGL1* binding-site and 82% and 83% for the rs231354 candidate region ($p>0.2$). Methylation levels around 50% are consistent with the location of the transcription binding site assays within the hemimethylated DMR, whilst the rs231354 candidate region lies on the border of the DMR. However, methylation did appear to be significantly higher in adult islet samples

at the rs2283202 candidate CpG region, with a mean methylation of 71% in adult islets versus 50% in foetal samples ($p=2.2 \times 10^{-13}$).

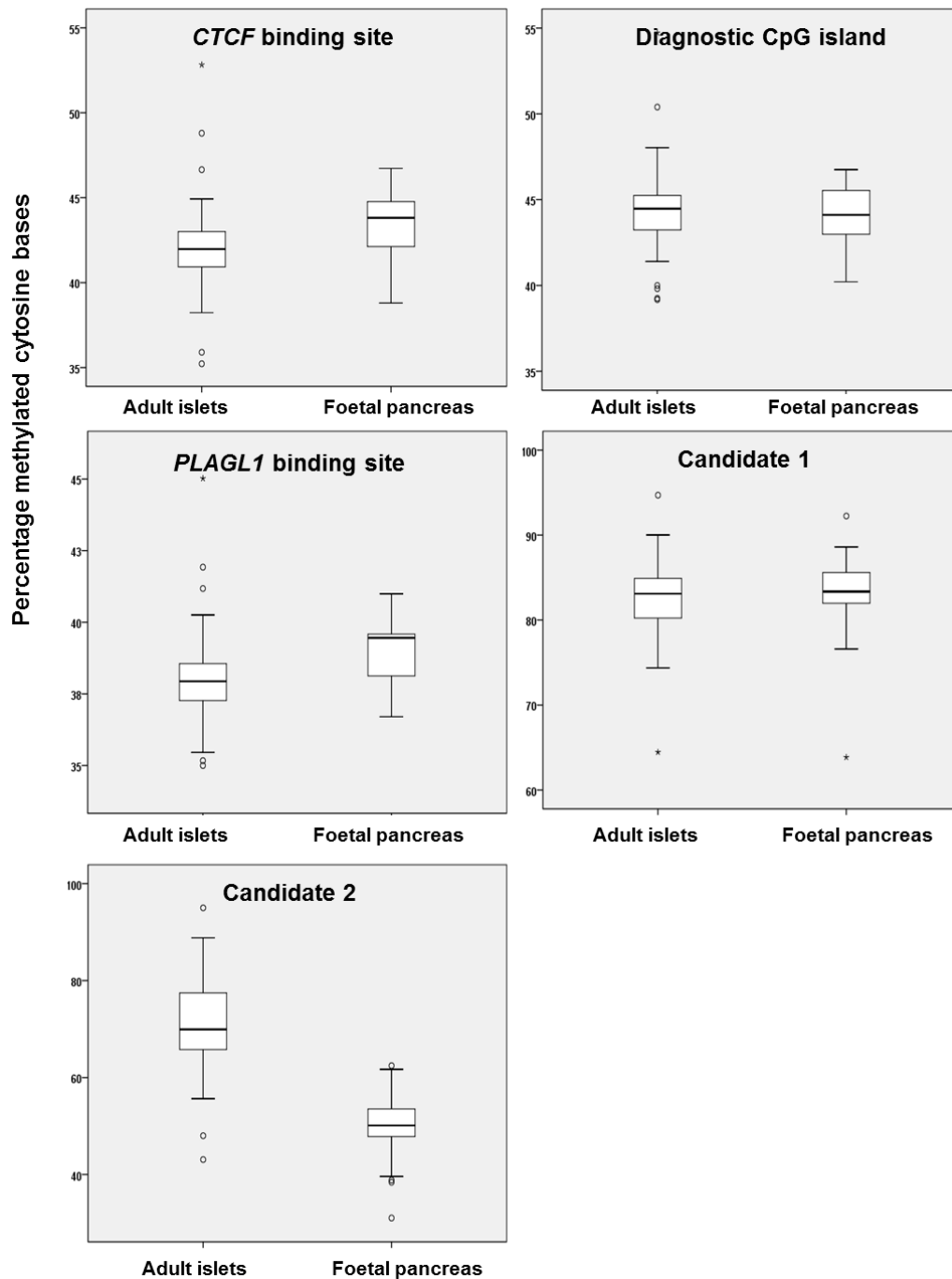


Figure 3.14: Methylation levels in adult islet and foetal pancreas samples. Boxes represent quartiles; whiskers encompass values within 1.5x interquartile range. Methylation is significantly higher in adult islets than in foetal pancreas at the candidate 2 region. All other differences are non-significant.

The T2D risk SNP rs2237895 was significantly associated with temporal-specific changes in methylation status at three of the sites tested (figure 3.15). At the diagnostic DMR region and *CTCF* binding site, methylation was higher (7.3% and 5.6% respectively) in foetal samples homozygous for the T2D risk allele than in foetal non-risk homozygotes ($p=0.02$ and 0.08 under a linear regression model). However, this effect was not replicated in adult islets; risk genotype had no effect on methylation levels at these same sites ($p=0.32$ and 0.28 respectively). Conversely, an effect was seen only in adult tissues at the *PLAGL1* binding site. Here, methylation was 1.6% higher in T2D risk genotype homozygotes than non-risk homozygotes ($p=0.006$ under a linear regression model) in adult islets, but no effect was seen in foetal pancreas ($p=0.36$).

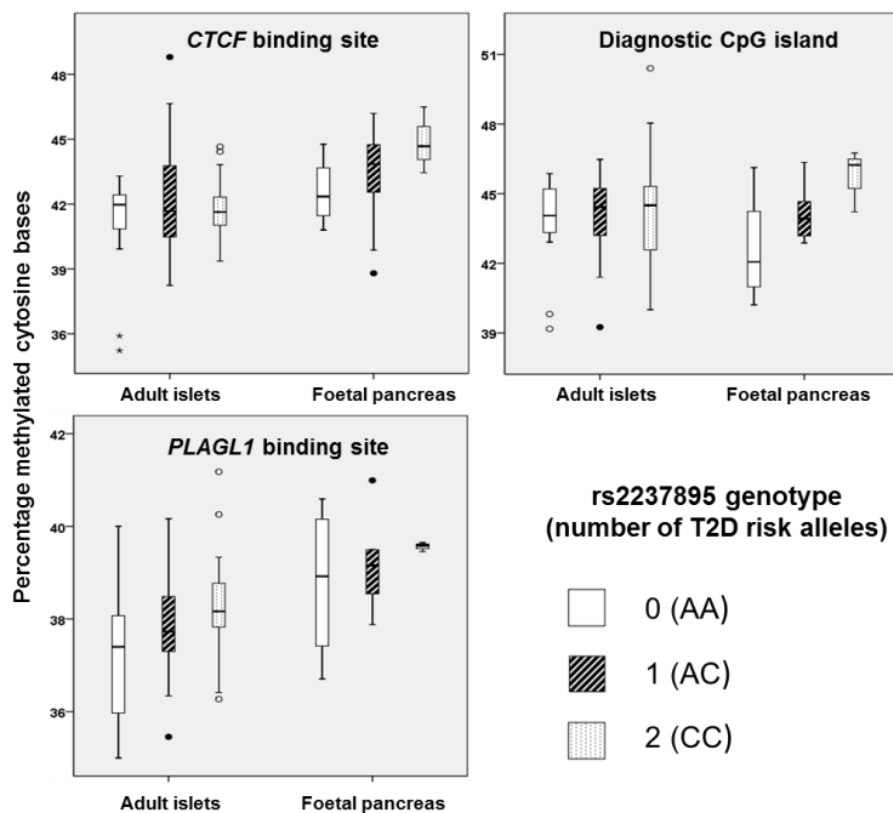


Figure 3.15: rs2237895 and methylation. Boxplots showing effect of rs2237895 T2D risk allele number (x-axis) upon methylation level (y-axis), separated by tissue type. Boxes represent quartiles; whiskers encompass values within 1.5x interquartile range. T2D risk genotype is significantly associated with increased methylation levels at the *CTCF* binding site and Beckwith-Weidman diagnostic CpG island in foetal samples only, and at the *PLAGL1* binding site in adult islets only.

There were no significant associations between rs231362 and methylation level at any of the tested sites ($p > 0.1$, figure 3.16).

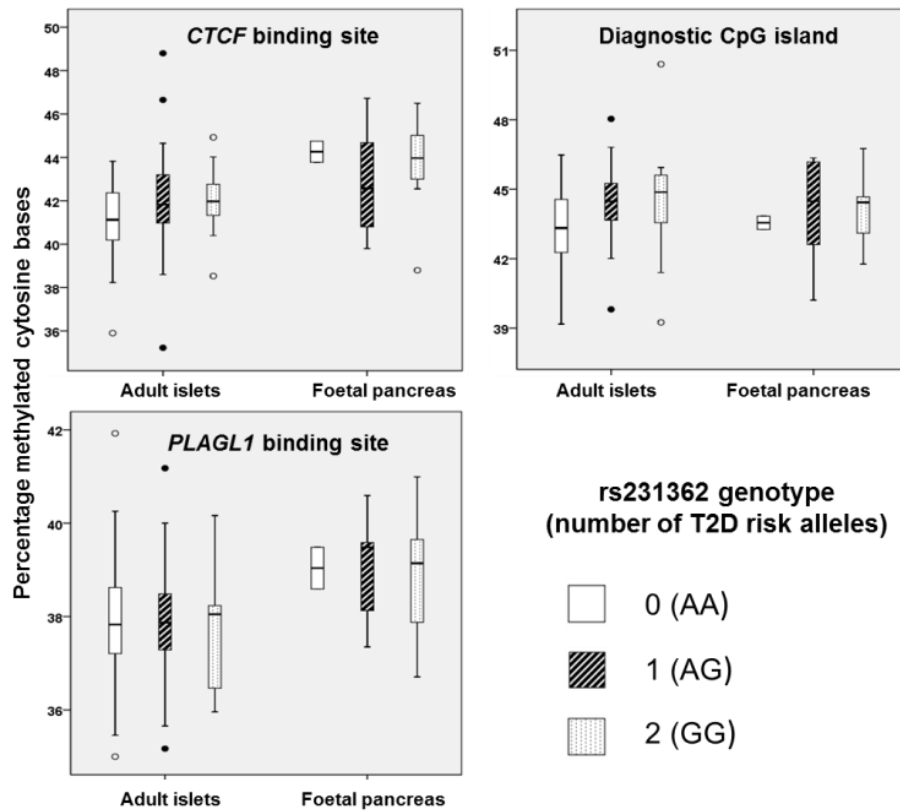


Figure 3.16: rs231362 and methylation. Boxplots showing effect of rs231362 T2D risk allele number (x-axis) upon methylation level (y-axis), separated by tissue type. Boxes represent quartiles; whiskers encompass values within 1.5x interquartile range. No significant associations were identified.

3.3.5 No evidence for an effect of T2D risk genotype on total or allele-specific gene expression

To determine whether the identified changes in methylation dictate regional gene expression effects, we assessed mRNA levels of *KCNQ1*, *KCNQ1OT1*, *PHLDA2*, *SLC22A18* and *SLC22A18AS* via QPCR.

As expected, there were significant differences in total expression between adult islets and whole foetal pancreas (figure 3.17). Expression of *KCNQ1*, *PHLDA2*, *SLC22A18* and *SLC22A18AS* was significantly higher in adult islets than in foetal pancreas (2.9, 26.4, 3.5 and 2.8 times respectively; $p=4.1 \times 10^{-9}$, $p=1.6 \times 10^{-25}$, $p=2.6 \times 10^{-9}$ and $p=4.1 \times 10^{-3}$), whilst *KCNQ1OT1* and *CDKN1C* were 8.8 and 3.6 times more highly expressed in foetal pancreas than in adult islets ($p=2.5 \times 10^{-5}$ and $p=1.1 \times 10^{-10}$).

To assess whether these differences were due to tissue type or developmental stage, we also quantified total mRNA expression in whole adult pancreas ($n=5$, figure 3.17). For *PHLDA2*, *SLC22A18* and *SLC22A18AS*, expression levels in adult pancreas were comparable to those in foetal pancreas, suggesting that the observed differences were related to tissue type and not developmental stage. In contrast, adult pancreas expression of *KCNQ1*, *KCNQ1OT1* and *CDKN1C* was similar to adult islet expression and markedly different to foetal pancreas, consistent with temporal differences in gene expression dependent on developmental stage.

Despite the observed differences in methylation levels according to T2D risk genotype, we were unable to demonstrate any relationship between either risk variant and total

gene expression level for any of the genes tested ($p > 0.05$ in all cases, figures 3.18 and 3.19).

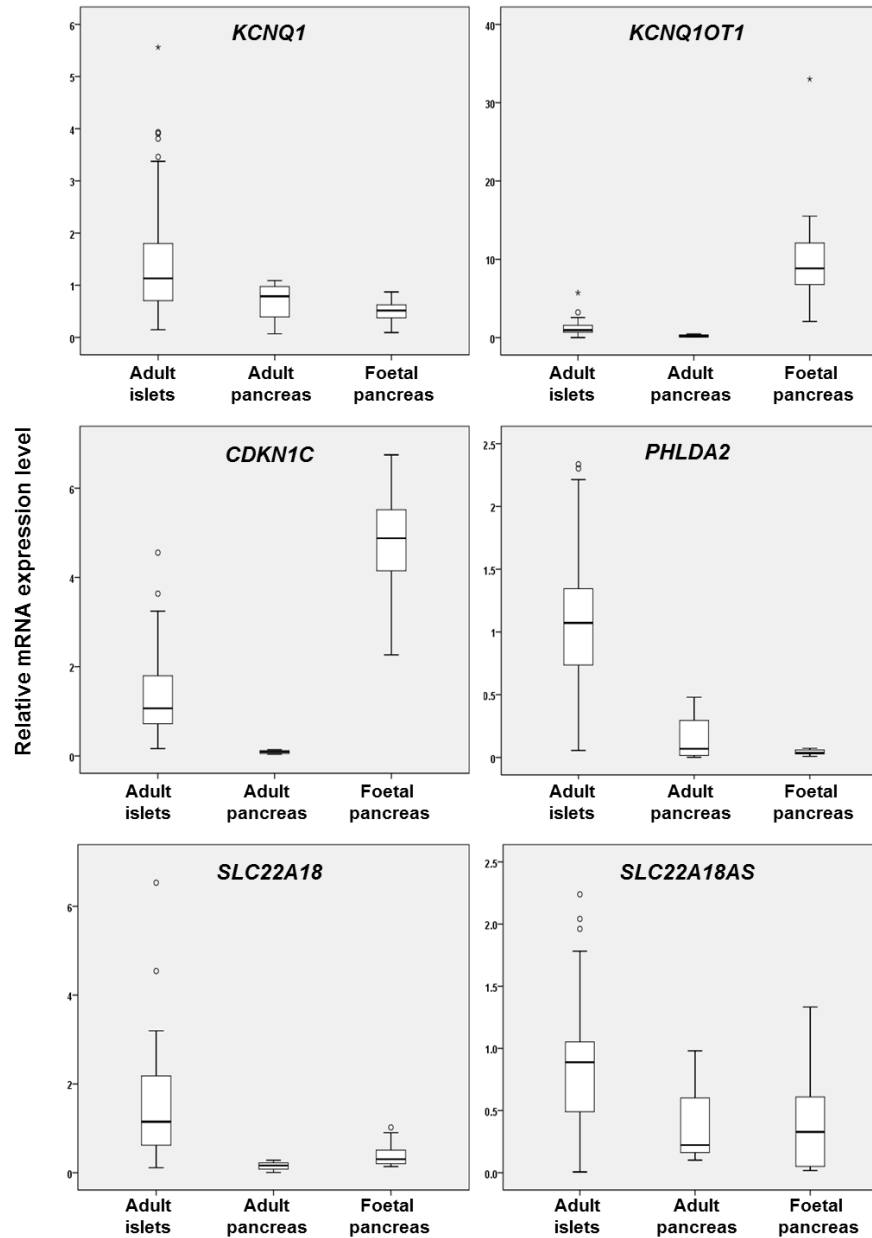


Figure 3.17: Expression levels in adult islets, adult pancreas and foetal pancreas samples. Expression is significantly higher in adult islets than foetal pancreas for *KCNQ1*, *PHLDA2*, *SLC22A18* and *SLC22A18AS*, but significantly higher in foetal pancreas for *KCNQ10T1* and *CDKN1C*. Adult pancreas expression is more in line with adult islet expression for *KCNQ1*, *KCNQ10T1* and *CDKN1C*, but more in line with foetal pancreas expression for *PHLDA2*, *SLC22A18* and *SLC22A18AS*.

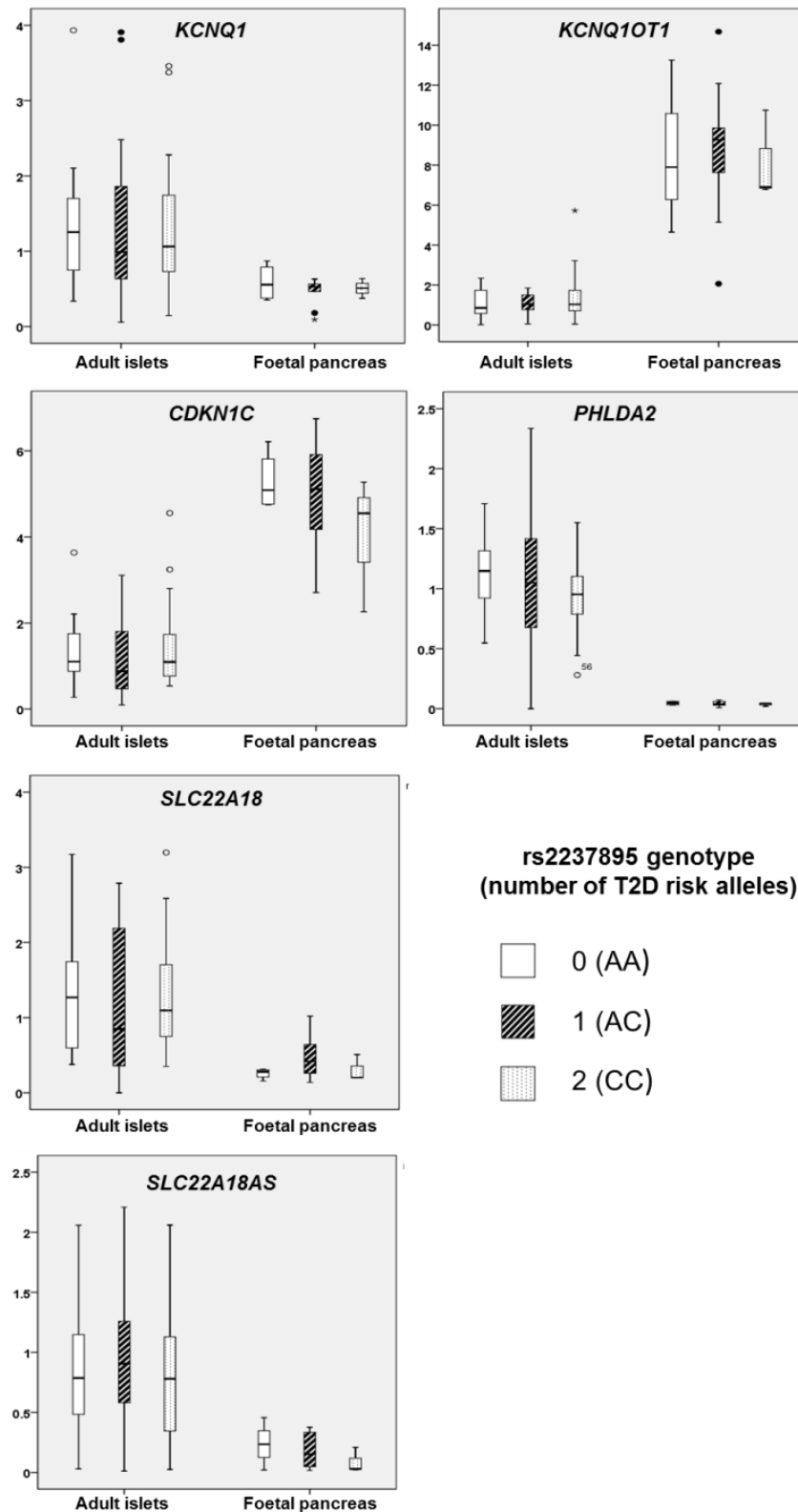


Figure 3.18: rs223795 and expression. Boxplots showing the effect of rs223795 T2D risk allele number (x-axis) upon total mRNA expression level (y-axis), separated by tissue type. Boxes represent quartiles; whiskers encompass values within 1.5x interquartile range. There was no evidence for an effect of risk allele number on expression levels of any of the tested genes in either tissue type ($p > 0.05$ in all cases).

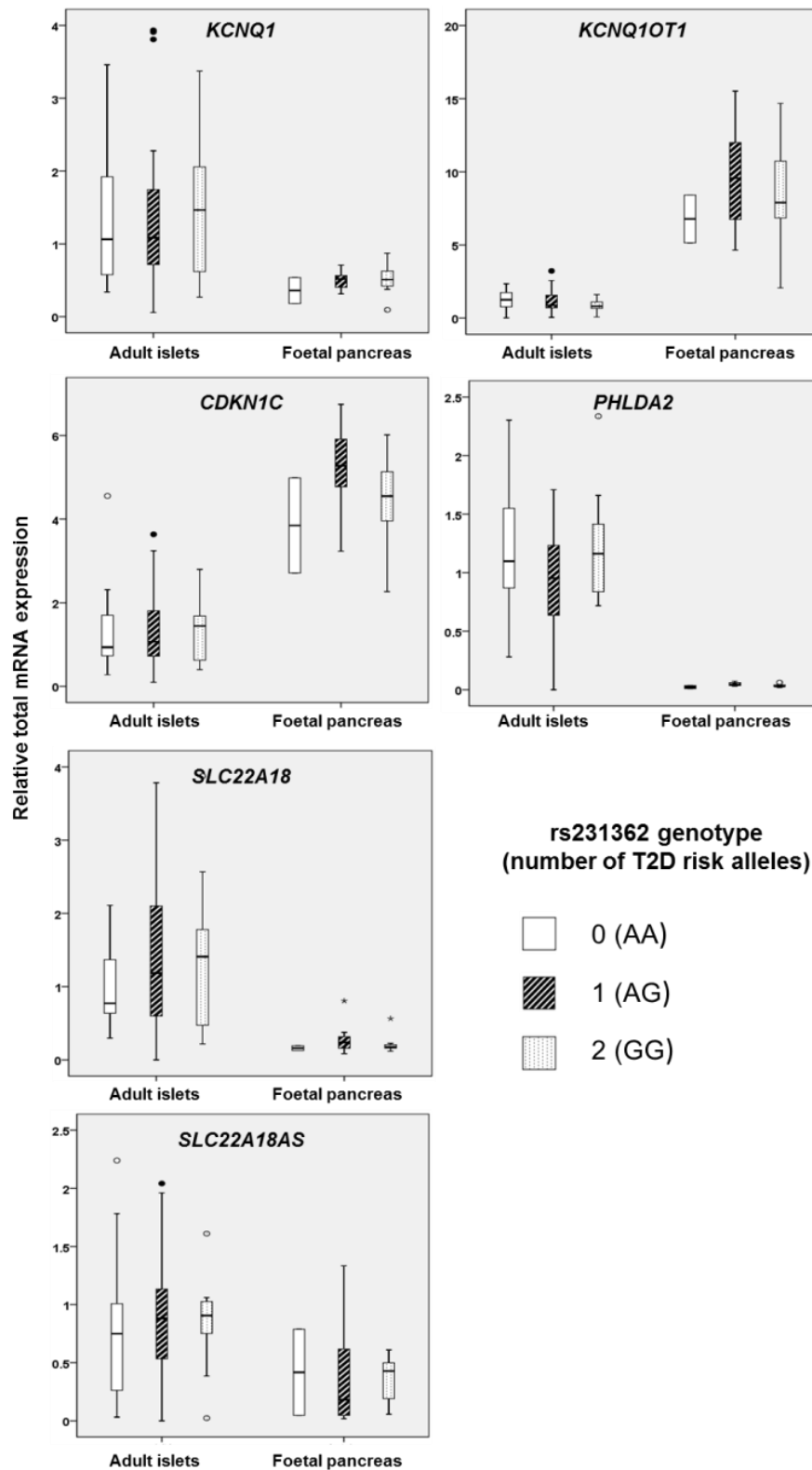


Figure 3.19: rs231362 and expression. Boxplots showing the effect of rs231362 T2D risk allele number (x-axis) upon total mRNA expression level (y-axis), separated by tissue type. Boxes represent quartiles; whiskers encompass values within 1.5x interquartile range. There was no evidence for an effect of risk allele number on expression levels of any of the tested genes in either tissue type ($p > 0.05$ in all cases).

Expression of *IGF2*, *INS* and *H19*, which lie in an adjacent but independently regulated imprinted locus at 11p15 (Maher and Reik, 2000) was also assessed, with no effect of T2D risk genotype status. On the basis of observed variance in expression, these experiments had limited power to detect expression effects. Despite representing the largest islet sample set reported to date, 72 adult islet and 18 foetal pancreas samples provide 40-60% and 10-20% (depending on variance for individual assays) power to detect a 50% change in expression per risk allele.

To explore allele specific effects of reduced DNA methylation on gene expression, we also examined the effect of T2D risk genotype on allele-specific expression ratios. For this analysis, expression data were used from samples heterozygous for coding SNPs as detailed in section 3.2.8. The expression level of the most highly expressed allele was divided by the expression level of the more lowly expressed allele, to provide an indication of the balance between the two (figure 3.20). For *CDKN1C*, the same analysis was conducted, using fragment analysis peak heights from the two alleles. Neither of the *KCNQ1* lead SNPs (data for rs231362 not shown) had any detectable impact upon the balance of expression between the chromosomes for any tested gene ($p>0.2$).

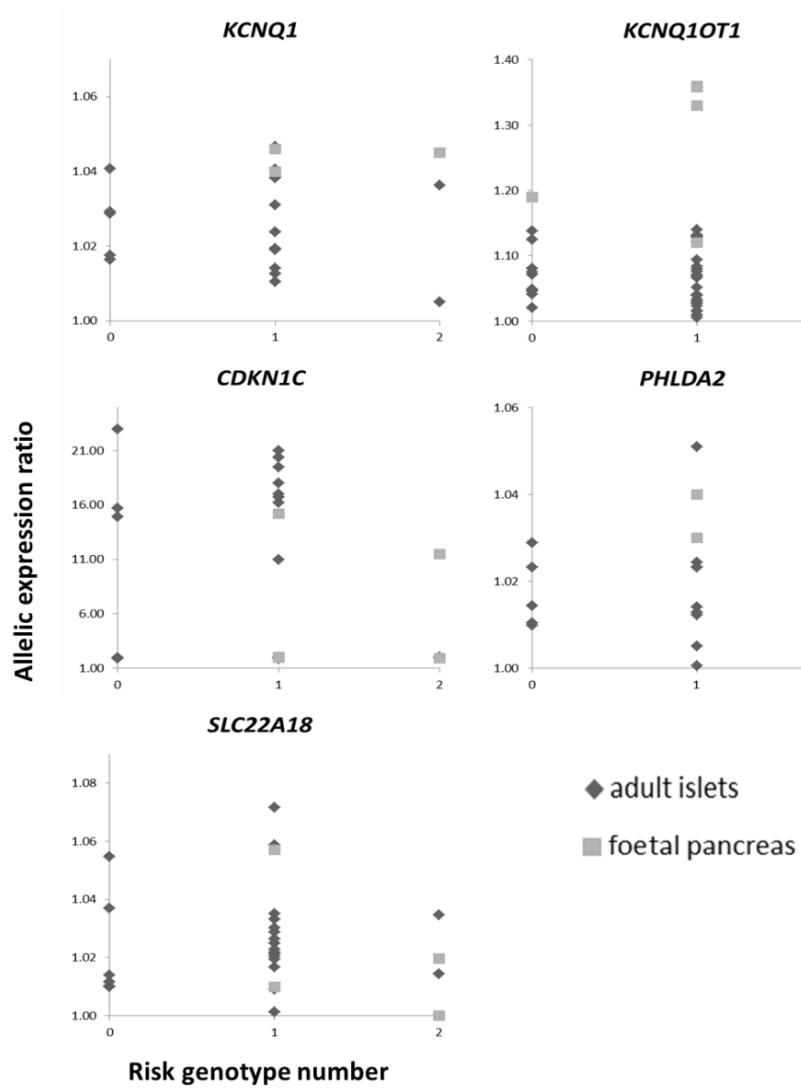


Figure 3.20: Allelic expression ratios. Scatter plots show rs2237895 T2D risk allele number (x-axis), plotted against allelic expression ratios for heterozygote samples based on coding SNPs (for *KCNQ1*, *KCNQ1OT1*, *PHLDA2* and *SLC22A18*) or a coding indel (for *CDKN1C*)(y-axis). No significant correlations were observed ($p > 0.2$ in all cases).

3.3.6 *KCNQ1* T2D risk variants and birth weight

T2D risk variants at the *KCNQ1* locus are good candidates for effects on birth weight. They influence insulin secretion, and regional variation is responsible for generalised foetal and early growth phenotypes such as Beckwith-Wiedemann syndrome. Preliminary data from the EGG (Early Growth Genetics) consortium suggests that *KCNQ1* variants may have some association with birth weight, as does a study into the *CDKN1C* indel del171APVA (Nielsen et al., 2005; detailed in section 3.1.6).

I genotyped the lead risk SNPs in both association signals (rs231362 in intron 10 and rs2237895 in intron 15) and del171APVA in 1292 mother-child pairs from the Berlin Birth Cohort (BBC). All participants included in the analysis were of European descent and singleton births of gestational length ≥ 37 weeks.

Both SNPs assays resulted in successful genotype calls for >90% of samples (genotyping pass rates >90%), but the more challenging indel assay had a lower pass rate of 84%. All variants were present in accordance with expected minor allele frequencies and did not depart from Hardy-Weinberg equilibrium ($p > 0.25$ in all cases) (table 3.13).

Variant	Genotyping pass rate	Reference MAF	Experimental MAF	Hardy-Weinberg p-value
rs231362	95%	0.48	0.47	0.27
rs2237895	90%	0.46	0.44	0.98
del171APVA	84%	0.26	0.26	0.85

Table 3.13: BBC Genotyping results. MAF = minor allele frequency. rs231362 reference MAF from HapMap allele frequency data in CEPH (European) population, rs2237895 reference MAF from experimental data outlined in previous sections, del171APVA reference MAF from Nielsen et al., 2005.

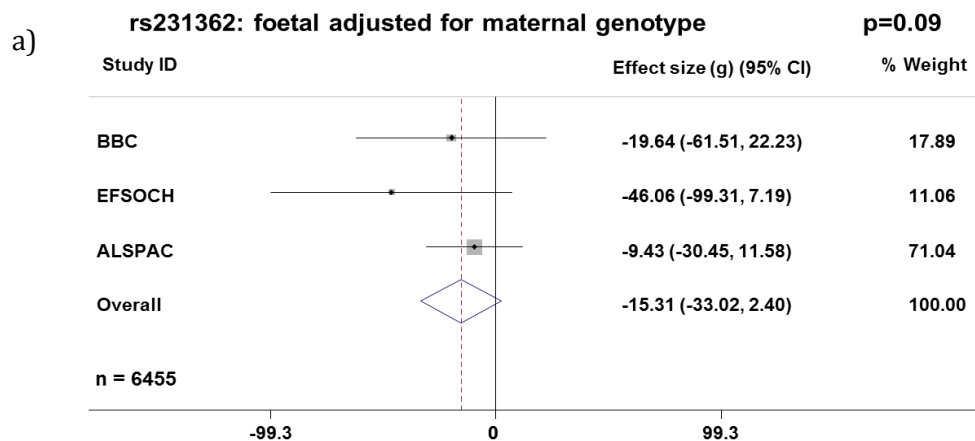
Association was tested via a linear regression of birth weight on genotype, with sex and gestational age as covariates (analysis conducted in SPSSv18). Since foetal and maternal genotypes are not independent, analysis was also conducted with an adjustment (through inclusion as a covariate) of foetal genotypes for maternal genotypes, and of maternal genotypes for foetal genotypes. This may help to determine whether any birth weight effect originates from T2D risk alleles carried by the mother or by the foetus.

Variant	Effect size in grams (95% CI)	p-value for association with birth weight
rs231362_foetal genotypes	-23 (-59, 13)	0.21
rs231362_maternal genotypes	-8 (-44, 28)	0.66
rs231362 foetal adjusted for maternal genotypes	-20 (-62, 22)	0.37
rs231362 maternal adjusted for foetal genotypes	3 (-39, 46)	0.92
rs2237895_foetal genotypes	-56 (-103, -9)	0.02
rs2237895_maternal genotypes	-65 (-112, -18)	0.01
rs2237895 foetal adjusted for maternal genotypes	-38 (-93, 17)	0.18
rs2237895 maternal adjusted for foetal genotypes	-55 (-110, 0)	0.05
del171APVA_foetal genotypes	-10 (-45, 25)	0.59
del171APVA_maternal genotypes	3 (-33, 40)	0.85
del171APVA foetal adjusted for maternal genotypes	-13, (-51, 26)	0.51
del171APVA maternal adjusted for foetal genotypes	9 (-31, 49)	0.67

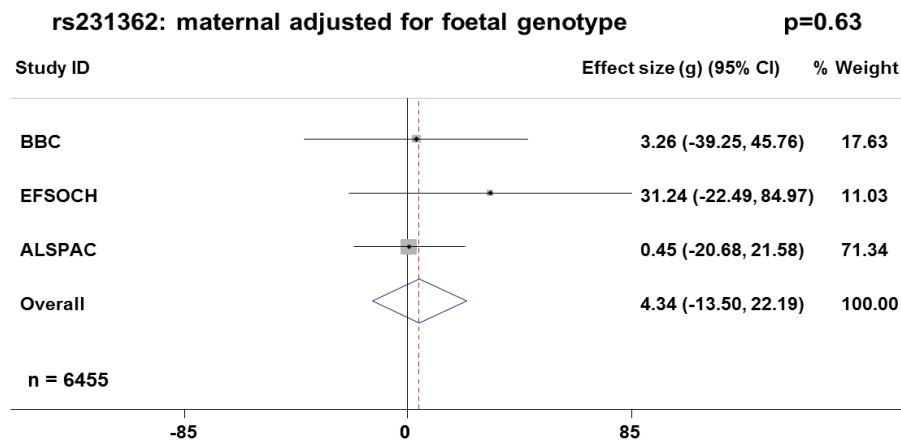
Table 3.14: Birth weight effects at the *KCNQ1* locus: Berlin Birth Cohort. Analysis conducted via a linear regression of birth weight on genotype, with sex and gestational age as covariates. Maternal and foetal genotypes were added as covariates for adjustment as indicated. Effect size (in grams) = unstandardized coefficient (B) (and accompanying 95% confidence intervals) from linear regression analysis,

No significant associations between birth weight and either the intron 10 SNP (rs231362) or del171APVA were identified in the BBC samples (table 3.14). Significant associations were identified between foetal genotypes, maternal genotypes and maternal genotypes adjusted for foetal genotypes at the intron 15 SNP (rs2237895). The negative effect sizes indicate that increased T2D risk allele number is associated with decreased birth weight. The attenuation of significance in foetal genotypes when adjusted for maternal genotype may suggest that the risk effect is maternal in origin.

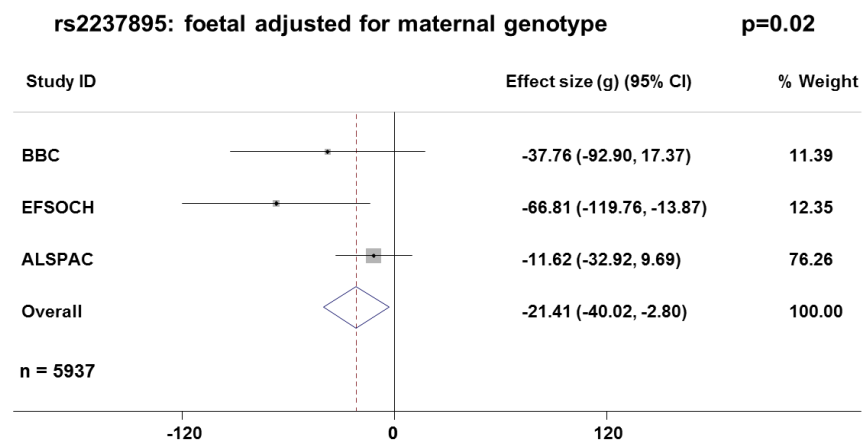
The BBC sample set is relatively small. To increase power to detect birth weight effects, and to reduce the chance of spurious results, BBC data was combined with data from three EGG cohorts (section 3.2.1) via inverse variance meta-analysis (figure 3.21:a-d). Genotyping and analysis was conducted by Dr. Rachel Freathy in collaboration with the EGG consortium. Indel genotypes were not available in these cohorts.



b)



c)



d)

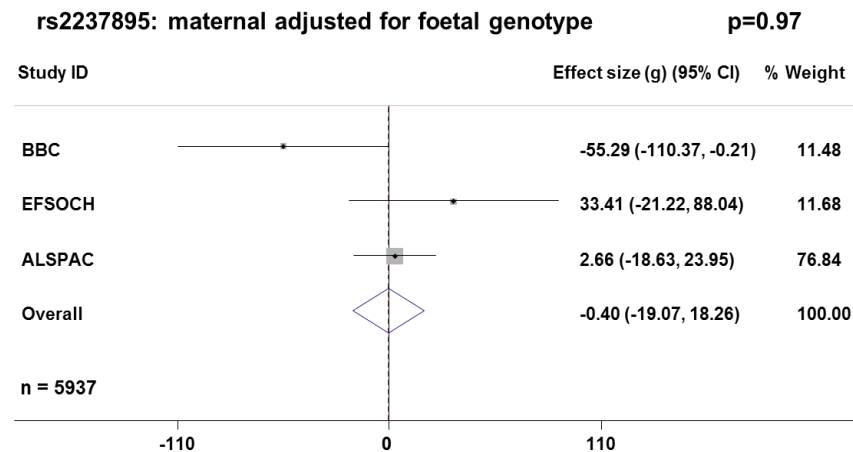


Figure 3.21a-d: Birth weight effects at the *KCNQ1* locus: Meta-analysis. Forest plots detailing association with birth weight in the BBC, EFSOCH, ALSPAC and HBCS (e only) sample sets, and overall inverse variance meta-analysis result. X-axis indicates effect size (grams), calculated via linear regression of birth weight on genotype, with sex, gestational age and maternal or foetal genotype (as appropriate) as covariates. Red hashed line and diamond indicates effect size and confidence intervals for overall meta-analysis.

As in the BBC data alone, no significant association between birth weight and rs231362 risk genotype was observed. There was, however, a trend ($p=0.09$) towards decreased birth weight (15g per risk allele) with increased number of foetal (adjusted for maternal) T2D risk alleles (figure 3.21a). This direction of effect is as predicted by the foetal insulin hypothesis. The trend for rs231362 maternal (adjusted for foetal) risk alleles was also in accordance with the maternal glucose hypothesis; in all studies, an increased number of maternal risk alleles tended to increase birth weight (figure 3.21b), although these effects were non-significant.

Foetal (adjusted for maternal) rs2237895 risk alleles were associated with a reduction in birth weight of 21grams per allele (figure 3.21c). Although this trend did not reach statistical significance in data from the BBC alone, in combination with data from EFSOCH and ALSPAC a significant p-value of 0.02 was reached. The significant effect of maternal (adjusted for foetal) risk genotype on reduced birth weight observed in the BBC (figure 3.21d) is surprising given the maternal glucose hypothesis, and contradictory to data in the EFSOC and ALSPAC sample sets.

Given the parent of origin-specific nature of T2D risk effects at the *KCNQ1* locus, we also tested whether effects on birth weight could be dependent upon the parental origin of risk alleles. The BBC consists of mother-child pairs, so risk allele parent of origin could only be inferred in a subset of samples. For heterozygote babies only, the risk allele was assumed to be paternally inherited when the mother was a non-risk homozygote, and assumed to be maternally inherited when the mother was a risk allele homozygote. Parental origin of risk alleles could not be inferred when the baby was homozygous, or when the mother was a heterozygote.

Parental origin effects were tested via linear regression of birth weight on parental origin of risk allele, corrected for sex and gestational age (table 3.15). This is equivalent to a t-test to compare birth weights in paternal and maternal origin groups. The reduced sample size provided only 11% power to detect an effect size of the magnitude of that observed for all samples.

Variant	n (of known parental origin)	Effect size in grams (95% CI)	p-value for association with birth weight
rs231362	304	-23 (-129, 84)	0.68
rs2237895	265	-100 (-236, 36)	0.15

Table 3.15: Parent of origin-specific birth weight effects at the *KCNQ1* locus: Berlin Birth Cohort. Analysis conducted via a linear regression of birth weight on parental origin of risk allele, with sex and gestational age as covariates. Effect size (in grams) = unstandardized coefficient (B) (and accompanying 95% confidence intervals) from linear regression analysis. Positive effect size indicates lower birth weight in paternal origin group; negative effect size indicates lower birth weight in maternal origin group.

Data was coded such that group 0 represented babies with a paternally inherited risk allele, whilst group 1 contained babies with a maternally inherited risk allele. Negative effect sizes therefore indicate that babies inheriting risk alleles from their mother had a lower birth weight than babies inheriting the same risk allele from their father. Although they are not statistically significant, the fact that the effect sizes identified in this analysis are similar to those in the main association analysis, particularly for rs231362 (both -23 grams for rs231362, and -55 versus -100grams for the main and parent of origin analysis respectively for rs2237895), suggest that birth weight effects may indeed originate from parental-specific inheritance of T2D risk alleles.

3.4 Discussion

The finding that imprinting at the 11p15.5 locus is not universal or static has two key implications; firstly for the attribution of T2D risk at this locus, and secondly for broader understanding and study of the region.

The parent of origin-specific nature of T2D risk at the *KCNQ1* locus (Kong et al., 2009) provides compelling evidence that diabetes risk is mediated through a gene with imprinted expression. Given the association of risk variants with insulin secretion and β -cell function (Holmkvist et al., 2009; Jonsson et al., 2009; Tan et al., 2009), diabetes risk is also likely to be mediated through a gene expressed in β -cells, and therefore also in islets. By demonstrating that *PHLDA2*, *SLC22A18* and *SLC22A18AS* are biallelically expressed in both adult and foetal pancreas and islets, the data in section 3.3.3 show that they are unlikely to be involved in a proximal molecular mechanism for diabetes risk. It is theoretically possible that a parent of origin-specific effect might operate through a biallelically expressed transcript, if expression from one allele is at a fixed level whilst expression from the other allele can be altered in cis. However, the existence of regional monoallelically expressed transcripts and extensive literature detailing imprinted expression suggests that this is not the case. Furthermore, this kind of effect would result in an altered allelic balance, which we did not detect - although numbers were limited for this analysis.

Similarly, the flexibly imprinted status of *KCNQ1* and *KCNQ1OT1* (imprinted in foetal pancreas, biallelically expressed in adult islets) suggests that any T2D susceptibility mechanism working through these genes is likely to be relatively early in islet development. It is not possible to determine the exact developmental stage at which

imprinting of these genes is relaxed, since they exhibited monoallelic expression in all our foetal samples, ranging between 42 and 133 days post conception. However, *IGF2* - located at the adjacent but independently regulated imprinted cluster on chromosome 11 and with multiple promoters - may provide a useful model. *IGF2* expression is imprinted from conception to six months after birth, relaxed between 8 and 10 months and fully biallelic by two years of age (Davies, 1994).

The variable and flexible nature of imprinting at 11p15.5 also has implications for the way in which future studies in this region should be conducted. It suggests that methylation and expression results should not be extrapolated from mouse to human, from one human tissue to another, or even from adult to early developmental time-points. Previous studies have demonstrated monoallelic expression of *KCNQ1* and *KCNQ10T1* in adult peripheral blood and adipose tissues, biallelic expression of the same genes in term placenta, and predominantly monoallelic expression of *PHLDA2* and *SLC22A18* in whole foetus samples (Kong et al., 2009; Monk et al., 2006). Existing data also demonstrate differences at *CDKN1C* between humans and mice. In mice, *Cdkn1c* has its own differentially methylated promoter region (Yatsuki et al., 2002), whereas the *CDKN1C* human promoter is unmethylated on both chromosomes (Chung et al., 1996). Results presented here highlight the fact that even extensive and well-documented analysis of imprinting status and mechanisms in other animals should not be assumed to apply to humans, or equally to all human tissues at all developmental stages. Empirical studies must be conducted not only in the relevant species and tissue, but also at the appropriate developmental time point.

On the basis of imprinting status data presented here, T2D risk at 11p15.5 could be mediated through *KCNQ1* or *KCNQ10T1* at an early developmental stage (since this is

when they show imprinted expression), or through *CDKN1C* in either foetal development or adulthood (since *CDKN1C* was imprinted in both foetal pancreas and adult islets). It cannot be definitively conclude whether *KCNQ1* is responsible for the risk effect, as has generally been assumed in the literature. I have also contributed data to a collaborative study (Rosengren et al., 2012) which has shed some light upon, but not fully resolved, this issue. However, a *decrease* in insulin secretion through this mechanism would require *gain of function* in $K_v7.1$, causing premature β -cell repolarisation and termination of insulin secretion – something which has not been experimentally demonstrated, or observed in humans with gain of function *KCNQ1* mutations causing short QT syndrome (Bellocq et al., 2004; Morita et al., 2008) . The study also demonstrated that risk genotype (intron 15 SNP rs2237895 only) was associated with reduced exocytosis (although not insulin secretion). However, repolarisation was experimentally controlled in their analysis. An altered rate of repolarisation through $K_v7.1$ cannot therefore be responsible for this phenotype.

Given its established role in glycaemic phenotypes and β -cell proliferation, the cell cycle regulator *CDKN1C* ($p57^{KIP2}$) is a particularly strong alternative regional candidate. Genes involved in cell-cycle regulation are generally overrepresented in T2D-associated loci (Voight et al., 2010), contributing to the debate over whether it is decreased β -cell mass or impaired β -cell function which is responsible for the impairment of insulin secretion seen for many T2D-associated variants (Clark et al., 2001). Recent evidence that the adult human β -cell population is largely static (Cnop et al., 2010; Perl et al., 2010) suggests that cell cycle regulation in *adult* β -cells is unlikely to contribute significantly to insulin secretory potential.

However, an effect mediated early in islet development may not necessitate a choice between β -cell mass and function. A T2D risk mechanism mediated through prenatal β -cell proliferation, or proliferation in infancy, could influence adult β -cell mass without any assumption of β -cell replication or turnover in adulthood. An altered adult 'starting' point for β -cell mass could dictate the extent to which β -cells are able to respond to an increased demand for insulin secretion, and therefore the propensity to develop T2D.

My work has shown that T2D risk genotype status is related to levels of DNA methylation, and that the developmentally-specific nature of this effect is consistent with a molecular mechanism for diabetes which is mediated at a specific developmental stage. Although the relationship between T2D risk genotype and DNA methylation levels is statistically stronger in adult than foetal samples ($p=0.006$ vs. $p=0.021$ and 0.081 respectively), the magnitude of effect appears larger in foetal samples (7.3% and 5.6% higher methylation in T2D risk homozygotes at the two associated sites in foetal pancreas, 1.6% in adult islets). The greater significance in adult islets may be attributable to improved power from a larger sample size (72 adult islet samples compared to 18 foetal pancreases). Importantly, the true effect size is likely to be underestimated, since data were obtained by pyrosequencing of PCR products amplified from both chromosomes. If the effect is indeed allele-specific, any change in the affected allele will therefore be partially disguised by static methylation on the unaffected allele. That said, effect sizes reported here are similar to those identified in the neighbouring *IGF2* DMR of Dutch Hunger Winter survivors, which had significant consequences for risk of heart disease and diabetes (Heijmans et al., 2008).

The direction of methylation effect identified - higher methylation in individuals carrying more T2D risk alleles - is consistent with a molecular mechanism for diabetes mediated through a cell-cycle repressor. An increase in methylation could reduce *KCNQ10T1* transcription, diminishing the repressive histone modifications which it induces through chromatin interaction (Nagano and Fraser, 2009; Redrup et al., 2009) and ultimately leading to increased expression of a growth inhibitor. It was disappointing not to identify changes in expression (particularly of *CDKN1C*) in this sample set, but it is possible that expression effects were not detectable due to greater individual variability and therefore reduced power in gene expression than DNA methylation. *CDKN1C* is expressed in two isoforms, produced through alternative splicing. The shortened form lacks eleven amino acids at the 5' end of the coding sequence, but was not distinguished by the expression assay used in this study. Altered isoform balance, transcript half-life or protein level are all mechanisms through which T2D-associated SNPs could influence *CDKN1C* gene function which were not investigated here.

Although my work identified effects of the intron 15 variant (rs2237895) on DNA methylation, no effects were seen with the top SNP in the intron 10 signal (rs231362). This is surprising, given that intron 10 actually contains the DMR for *KCNQ10T1* and much of *KCNQ10T1* itself. It is possible that effects at rs231362 were not identified due to a less favourable division of samples between the risk genotype groups, and therefore that a greater sample size with improved power may detect effects at both association signals. On the basis of observed methylation variants, effect sizes and sample sizes, the significant differences in adult and foetal samples were detected with 23% and 84% power respectively - so it is possible that effects of a similar magnitude from rs231362 may have been missed due to chance. Alternatively, there may be a

highly complex mechanism whereby single nucleotide variation can influence methylation at a distance of many kilobases. It is interesting that electrophysiological analysis (Rosengren et al., 2012) also identified effects on exocytosis associated with rs2237895 but not rs231362 genotype. Further work is required to interrogate these possible mechanisms, as well as to demonstrate that methylation effects have consequences for local gene expression and to determine the precise developmental stage at which imprinted expression of *KCNQ1* and *KCNQ1OT1* is lost in human islets.

The birth weight analysis reported in section 3.3.6 provides some evidence that T2D-associated variants at *KCNQ1* are also associated with altered foetal development. As predicted by the foetal insulin hypothesis, T2D risk alleles correlate with reduced birth weight, lending further weight to the hypothesis that the biological consequences of T2D-associated variants may manifest at an early developmental stage. The small number of samples in which parent-of-origin could be inferred limit power to detect parent-specific risk effects, and the differences in birth weight between maternally and paternally-inherited risk allele groups were not significant. For example, reduced numbers in the parent-of-origin analysis provide only 11% power to detect an effect of the same magnitude as that reported in the main analysis (at a 5% significance level). However, the similarity of effect sizes between the parent-of-origin and main analyses do suggest that the overall birth weight effect could be coming from one parentally inherited allele only. Furthermore, the direction of effect - lower birth weight for maternally inherited risk alleles - is consistent with a T2D risk mechanism mediated only from maternally inherited chromosomes.

Like methylation effects, significant birth weight associations were identified only with rs2237895. However, meta-analysis did reveal a clear trend (of nominal significance)

in the anticipated direction for rs231362. Whilst care should be taken in interpreting non-significant results, this may suggest that methylation effects could be uncovered for rs231362, given more samples to provide greater power. A birth weight effect for del171APVA of the magnitude reported in Neilsen et. al. (2005) can be confidently excluded, since the larger numbers reported here provide 98% power to detect a similar effect size.

Using the largest sample set of human islets currently available and material from a human developmental biology research programme, this chapter reports the first assessment of imprinting status at 11p15.5 in human adult islets and foetal pancreas. It demonstrates temporal changes in imprinted gene expression at 11p15.5, and developmentally-specific effects of T2D risk genotype on DNA methylation. The data provide insights into the complexity of imprinting control at 11p15.5, and have implications for the molecular mechanism by which associated variants in this region exert their effect on diabetes risk.

Chapter 4

Gene expression
profiling as a tool
for interpreting
genome-wide
association signals:

Proinsulin analysis

4.1 Introduction

4.1.1 Limitations of genome-wide association studies

Genome-wide association studies have been spectacularly successful in identifying common variation which contributes to T2D risk. In some instances, fine-mapping has identified a causative coding variant and functional studies have revealed its mechanism, such as P446L in glucokinase regulatory protein (*GCKR*) (Beer et al., 2009; Rees et al., 2012) and R325W in *SLC30A8* (Nicolson et al., 2009).

However, the majority of diabetes-associated loci have proved less tractable. Some comprise a linkage disequilibrium block which encompasses multiple plausible genes (such as the chromosome 10q T2D-associated locus which contain *HHEX*, *KIF11* and *IDE*), whilst others lie in a gene desert far away from any obvious candidate genes. The region of T2D association on chromosome 9p, for example, is over 200kb away from *CDKN2A/B*, the genes through which it is thought most likely to operate (Krishnamurthy et al., 2006; Moritani et al., 2005). Whilst association signals are generally labelled by a gene name, in most cases this label is not proven. In the absence of functional evidence, a variant lying close to or even within a gene (see chapter 3) does not prove that gene's causality. Furthermore, SNPs may regulate gene expression at a long distance. In many cases, it is not therefore possible to deduce the relevant gene simply from the location of associated SNPs.

4.1.2 Gene expression profiling as a tool to aid interpretation of genome-wide association studies

Genome-wide association studies require complementary analysis and investigation, to refine genetic signals and to determine true causal genes and mechanisms. Replication of association studies in diverse populations with differing LD structures may help to narrow the region of association (Hassanein et al., 2010), whilst expression quantitative trait loci (eQTL) mapping aims to determine whether disease-associated variants also influence the expression levels of nearby genes – although the scarcity of human islet samples means that islet eQTL data are not yet available. Physiological characterisation can help to determine which biological processes are disrupted, distinguishing between T2D-associated variants which work through insulin resistance or impaired insulin secretion (Ingelsson et al., 2010).

Gene expression profiling of candidate genes is one such biological tool which can aid the interpretation of genome-wide association studies. On the premise that only genes expressed in tissues relevant to the phenotype can mediate an effect, demonstrating that any particular gene is not expressed in relevant tissues weakens that gene as a causal candidate. Similarly, any gene particularly highly or specifically expressed in relevant tissues becomes a stronger candidate.

For T2D, it is pertinent to screen liver, muscle and adipose tissue to account for a mechanism mediated through altered insulin sensitivity, as well as pancreas, islets and pancreatic β -cells for variants acting through impaired insulin processing or secretion. Wherever possible, expression profiling should be conducted in human tissues. This is particularly pertinent for islets, given evidence that human and rodent islets differ considerably in their β -cell proportion and arrangement, predominant glucose

transporters and ion channel composition. Expression profile analysis should be seen as complementary to physiological data, which can also distinguish between secretion and sensitivity phenotypes.

4.1.3 Proinsulin as a relevant trait to T2D

I performed expression profiling as part of the follow-up from a GWA study of proinsulin levels (association analysis and expression profiling published in Strawbridge et al., 2011). Proinsulin is the precursor of mature (immunoreactive) insulin. It is processed in secretory granules of the pancreatic β -cell through a series of enzymatic conversions (figure 4.1), and secreted once split into mature insulin and C-peptide. High levels of circulating insulin are often indicative of insulin resistance and T2D, but there is considerable evidence that levels of proinsulin itself are also of interest (Fritsche et al., 2002; Ingelsson et al., 2010; Roder et al., 1998; Saad et al., 1990; Wareham et al., 1999). Conversion of proinsulin to insulin is an important aspect of β -cell function. The relative proportion of proinsulin (PI) and insulin (the PI/I ratio) therefore represents an estimate of proinsulin processing efficiency. A decrease in the PI/I ratio would indicate an increase in the rate of proinsulin processing, whilst an increase in the ratio would indicate a decrease in the rate of processing (Fritsche et al., 2002). High levels of circulating proinsulin relative to mature insulin can therefore indicate β -cell stress, as a result of insulin resistance, impaired β -cell function, or insulin processing and secretion abnormalities.

A 2-3 times increase in PI/I ratio has indeed been observed in T2D (Gorden et al., 1974; Saad et al., 1990; Yoshioka et al., 1988), and the extent of this disproportional

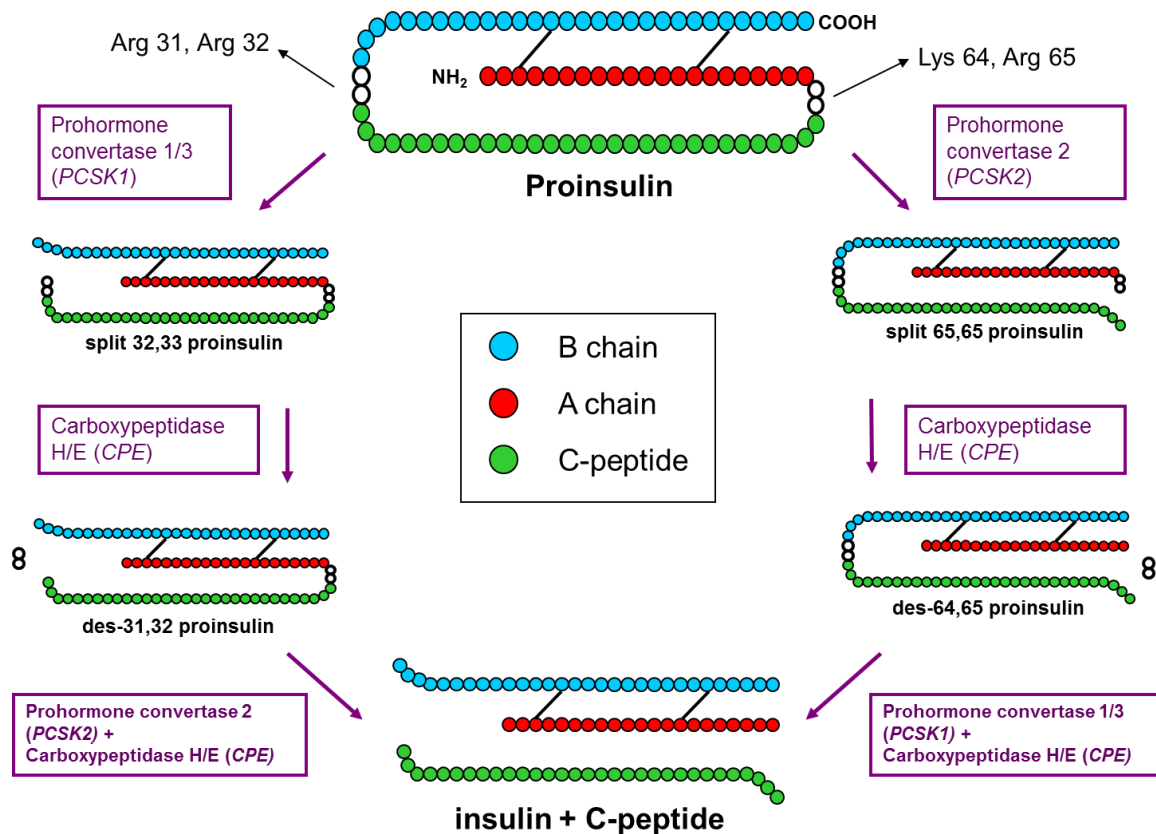


Figure 4.1: The proinsulin processing pathway. Insulin biosynthesis via proinsulin processing occurs in secretory granules of the pancreatic beta cell. Proinsulin is composed of two chains, A and B, joined by a connecting peptide (C-peptide). Protein (encoding gene) shown. Prohormone convertase 1/3 (PCSK1) cleaves the carboxyl sites Arginine 31, Arginine 32 at the B/C chain junction and prohormone convertase 2 (PCSK2) cleaves the carboxyl sites Lysine 64, Arginine 65 at the A/C chain junction to produce split proinsulin. The prohormone convertase 1/3 pathway is predominant in the physiological state. Split fragments are further modified by carboxypeptidases to remove exposed basic residues, producing des-proinsulin. Finally, proinsulin is cleaved at the remaining carboxyl site to separate mature insulin and C-peptide molecules, which are secreted in equimolar amounts. Figure adapted from Assmann et al., 2009.

elevation is inversely related to maximal β -cell secretory capacity (Roder et al., 1998).

High levels of circulating proinsulin predict future T2D, even after taking fasting glucose into account (Kahn et al., 1995; Wareham et al., 1999). Furthermore, some loci identified through GWAS for fasting glucose (*MADD*) or T2D risk (*TCF7L2*, *SLC30A8*, and *CDKAL1*) have, on further analysis, been shown to also associate with raised circulating proinsulin (Ingelsson et al., 2010; Kirchhoff et al., 2008; Loos et al., 2007; Stolerma et al., 2009), providing evidence that the genetic determinants of proinsulin can also be relevant to T2D causation.

4.1.4 Meta-Analysis of Glucose-and Insulin-related traits Consortium (MAGIC) proinsulin genome-wide association study

The MAGIC consortium, of which my supervisor's group is a part, was formed on the premise that variants influencing continuous insulin and glucose-related traits in healthy individuals may highlight not only physiological regulatory mechanisms, but also genes and pathways implicated in pathological variance. Five loci (*ADCY5*, *GCK*, *GCKR*, *PROX1* and *DGKB-TMEM195*) identified through association analysis of fasting glucose, fasting insulin, β -cell function (HOMA-B) and insulin resistance (HOMA-IR) have emerged as novel T2D loci, providing confirmation of the link for some loci between physiologically and pathophysiologically relevant genetic variants (Dupuis et al., 2010; Ingelsson et al., 2010).

Following this success, the MAGIC investigators turned their attention to other insulin-related traits, including proinsulin. Disproportionately (compared to mature insulin) raised circulating proinsulin levels are associated with impaired β -cell function, raised glucose levels, insulin resistance and T2D. By performing a GWAS for proinsulin levels, the MAGIC consortium aimed to capture an increase in proinsulin relative to the non-specific activation of the insulin processing pathway induced by generalised insulin resistance, identifying loci influencing the β -cell's ability to process proinsulin adequately and thus revealing additional loci for T2D susceptibility.

Consortium statisticians led by Dr. Rona Strawbridge performed a meta-analysis of four studies, incorporating 2.5million genotyped or imputed SNPs and 10,701 non-diabetic (fasting glucose <7mmol/L) adults of European ancestry. In order to identify variants associated with high proinsulin relative to fasting insulin (the PI/I ratio), proinsulin levels were adjusted for mature insulin – as well as age and sex. Twenty-one

loci were taken forward to replication in 11 cohorts including 16,378 healthy individuals. After replication, nine SNPs at eight loci were associated with proinsulin levels at genome-wide significance ($p < 5 \times 10^{-8}$, table 4.1, figure 4.2).

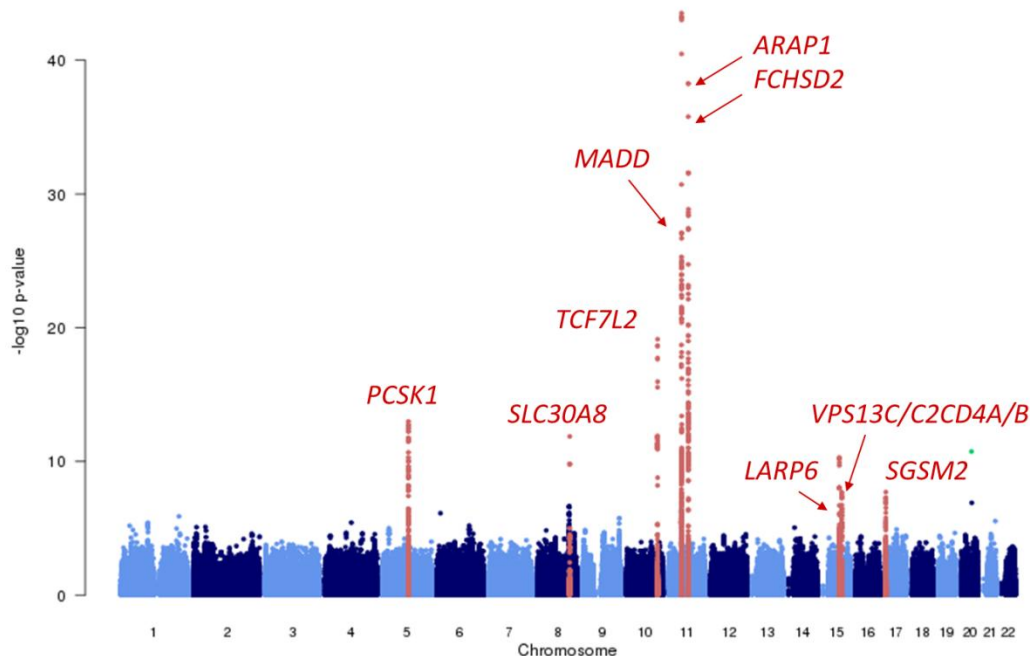


Figure 4.2: Manhattan plot of fasting proinsulin loci. Genotyped SNPs are plotted according to their genomic position (x-axis) and $-\log_{10}$ p-value (y-axis) for association with fasting proinsulin adjusted for fasting insulin. SNPs which achieved genome-wide significance ($p < 5 \times 10^{-8}$) upon follow-up are shown in red. Plot taken from (Strawbridge et al., 2011)

PCSK1 encodes prohormone convertase 1/3, the first enzyme in the proinsulin processing pathway (figure 4.1). The *MADD* locus, which has been previously associated with fasting glucose (Ingelsson et al., 2010), contained two independent ($r^2=0.07$) signals 19kb apart. *TCF7L2*, *SLC30A8*, *VPS13C* and *ARAP1* (formerly *CENTD2*) are already implicated in T2D risk, whilst *LARP6* and *SGSM2* are novel loci not previously associated with metabolic traits. Together, these loci explained 8.1% of overall proinsulin/insulin ratio variance in the replication cohorts.

SNP	Locus labelling gene	Position	Effect allele	Effect Allele frequency	β	p-value
rs11603334	<i>ARAP1</i>	Chr11:72110633	A	0.15	0.0928	3.2×10^{-102}
rs10501320	<i>MADD</i>	Chr11:47250375	G	0.72	0.0805	1.1×10^{-88}
rs6235	<i>PCSK1</i>	Chr5:95754654	G	0.28	0.0394	9.8×10^{-27}
rs7903146	<i>TCF7L2</i>	Chr10:114748339	T	0.30	0.0321	2.3×10^{-20}
rs4502156	<i>VPS13C</i>	Chr15:60170447	T	0.58	0.0294	3.5×10^{-20}
rs11558471	<i>SLC30A8</i>	Chr8:118254914	A	0.69	0.0280	3.1×10^{-18}
rs10838687	<i>MADD*</i>	Chr11:47269468	T	0.80	0.0253	6.9×10^{-12}
rs1549318	<i>LARP6</i>	Chr15:68896201	T	0.61	0.0192	2.4×10^{-10}
rs4790333	<i>SGSM2</i>	Chr17:2209453	T	0.45	0.0154	3.0×10^{-9}

Table 4.1: Loci associated with fasting proinsulin levels at a genome-wide level of statistical significance. Effect size (β) represents per allele change in PI/I ratio, after adjustment for sex and age. β and p-values for combined analysis including replication. *Second, independent signal at the *MADD* locus. Association was not attenuated by adjustment for fasting glucose. Data taken from (Strawbridge et al., 2011).

4.1.5 Experimental aims

As part of the MAGIC consortium, I performed gene expression profiling for loci identified in the proinsulin GWAS. These experiments aimed to:

- exclude genes as candidates for a proinsulin effect by demonstrating that they are not expressed in relevant tissues;
- highlight genes as particularly interesting candidates for a proinsulin effect, by demonstrating that they are particularly strongly or specifically expressed in relevant tissues;
- determine expression in tissues relevant to insulin secretion or insulin sensitivity, to provide information about the possible mechanism of action for associated variants.

Three loci were not included in expression analysis because the genes and biological mechanisms through which they operate were considered to be clear. The lead SNP at the *PCSK1* locus (rs6235) is a missense polymorphism (S690T), likely to disrupt the function of the enzyme which cleaves proinsulin to 32,33-split proinsulin. The lead SNP at the *TCF7L2* locus (rs7903146) has previously been shown to associate with diabetes risk. The same allele is associated with raised proinsulin levels but decreased insulin secretion (Voight et al., 2010), which suggests a defect in proinsulin processing. Similarly, the lead proinsulin-raising allele at *SLC30A8* (rs11558471) is a coding variant in perfect LD with R325W, already known to associate with fasting glucose and T2D (Dupuis et al., 2010). On-going functional and expression analyses of these genes are consistent with an effect on insulin processing, through granule fusion and insulin crystallisation respectively (da Silva Xavier et al., 2009; Nicolson et al., 2009).

For the remaining five loci, the expression of a subset of genes transcribed from a 1Mb region around each association signal was assessed, selected on the basis of biological plausibility. Genes assessed are detailed in table 4.2. As is the case in T2D, tissues involved in both insulin secretion and insulin resistance can be considered relevant to the PI/I ratio. Islet and β -cells are pertinent to the proinsulin processing pathway itself, including enzymatic cleavage, vesicle trafficking and exocytosis. The tissues of insulin action (adipose, muscle and liver) may play a role through increased insulin resistance. If the β -cell proinsulin processing pathways cannot keep up with increased insulin demand, a greater number of vesicles may reach exocytosis whilst still containing immature insulin, thus increasing the PI/I ratio even if there is no intrinsic abnormality in the β -cell.

Gene closest to lead SNP	Other regional and biologically plausible genes
<i>ARAP1</i>	<i>INPPL1, STARD10, RAB6A</i>
<i>MADD</i>	<i>CELFI, LRP4, MTCH2, SLC38A13</i>
<i>VPS13C</i>	<i>C2CD4A, C2CD4B, RORA</i>
<i>LARP6</i>	<i>CT62</i>
<i>SGSM2</i>	<i>MYOC1, RIPIGAP2, SKIP, YWHAE</i>

Table 4.2: Candidate genes at proinsulin-associated loci. Biologically plausible genes within a 1Mb region around proinsulin association signal

Expression was assessed across a commercial human tissue panel (including the metabolically relevant tissues muscle and liver), supplemented by human omental adipose, subcutaneous adipose, pancreas and islet samples gathered in Oxford.

As part of the MAGIC consortium investigations, expression analysis was also conducted in enriched β -cell preparations. This work was performed by collaborators in Lille, France and is therefore presented as part of the discussion rather than results.

4.2 Samples and Methods

4.2.1 Samples used

Total RNA samples (bone marrow, spleen, thymus, adult liver, foetal liver, adrenal gland, salivary gland, thyroid gland, heart, skeletal muscle, placenta, prostate, testis, uterus, adult whole brain, foetal whole brain, cerebellum, spinal cord, colon, kidney, lung, trachea, small intestine, stomach) were purchased as part of a commercially available tissue panel (Clontech, Oxford, UK), with each sample comprising a pool of between one and 64 donors (table 4.3).

Omental adipose, subcutaneous adipose and pancreas tissue (pool of 5 donors each) were obtained through Oxford University's MolSURG project, collected as part of the MolPAGE (Molecular Phenotyping to Accelerate Genomic Epidemiology) consortium (MolPAGE, 2012). Tissue and blood samples were collected with full research consent (Oxfordshire Regional Ethics Committee B) from participants undergoing elective surgery, snap frozen and stored as whole tissue in liquid nitrogen prior to extraction.

Adult human islets (3 samples, not pooled) were obtained with full research consent (Oxfordshire Regional Ethics Committee B) from the Oxford Centre for Islet Transplantation (OXCIT). All donors were of European descent and with no known diabetic phenotype. Samples were stored in RNA later at -80°C prior to RNA extraction. Islet samples were >80% pure, determined by endocrine-vs-exocrine gene expression analysis as described in section 2.7.6.

Tissue	N donors	Donor age range (years)
bone marrow	16	22-68
spleen	15	22-69
thymus	3	18-57
adult liver	3	29-51
foetal liver	63	22-40 (weeks)
adrenal gland	62	15-61
salivary gland	24	16-60
thyroid gland	64	15-61
heart	3	30-39
skeletal muscle	3	43-46
placenta	15	19-33
prostate	12	20-58
testis	39	14-64
uterus	8	23-63
adult whole brain	1	18
foetal whole brain	59	20-33 (weeks)
cerebellum	10	22-68
spinal cord	4	29-64
colon	1	23
kidney	3	30-55
lung	3	32-61
trachea	22	18-54
small intestine	5	20-61
stomach	4	23-61

Table 4.3: Human tissue RNA panel donor details

4.2.2 RNA extraction

Phenol-chloroform extraction of RNA from human islets, omental adipose, subcutaneous adipose and pancreas tissue was performed using Trizol (Ambion, Warrington, UK) as described in section 2.3. RNA was suspended in RNase free water (minimum volume 20µl) and checked for integrity checked using an Agilent Bioanalyser 2100 as described in section 2.3.5. All samples used had RIN scores >7.

4.2.3 cDNA synthesis

cDNA was generated via random primed first strand synthesis using the QuantiTect reverse transcription kit (Qiagen, Crawley, UK) from 1ug RNA, as detailed in section 2.6. cDNA synthesis was also performed in the absence of the reverse transcriptase enzyme to generate negative control samples.

4.2.4 mRNA expression quantification

Total gene expression was quantified using TaqMan gene expression assays (Applied Biosystems) as described in section 2.7. A subset of the genes transcribed from within 1Mb of each associated locus were analysed, selected by a MAGIC project committee on the basis of biological candidacy. Inventoried assays were selected to cover the maximum number of transcripts and, where possible, to span an exon junction (_m1 suffix) to prevent amplification of any residual contaminating DNA. Assay details are provided in table 4.4. All cDNA samples were diluted one to 50 parts (1:50) in 0.01M Tris, and standard curves generated by serially diluting a pool of all cDNA samples as described in section 2.7.3.

Gene ID	Full gene name	Assay ID
<i>INPPL1</i>	inositol polyphosphate phosphatase-like 1	Hs00155533_m1
<i>ARAP1 (CENTD2)</i>	ArfGAP with RhoGAP domain, ankyrin repeat and PH domain 1	Hs00373707_m1
<i>STARD10</i>	StAR-related lipid transfer (START) domain containing 10	Hs00246405_m1
<i>RAB6A</i>	RAB6A, member RAS oncogene family	Hs03044017_m1
<i>LRP4</i>	low density lipoprotein receptor-related protein 4	Hs00391006_m1
<i>MADD</i>	MAP-kinase activating death domain	Hs00366249_m1
<i>SLC39A13</i>	solute carrier family 39 (zinc transporter), member 13	Hs00378317_m1
<i>CELF1 (CUGBP1)</i>	CUGBP, Elav-like family member 1	Hs01549074_m1
<i>MTCH2</i>	mitochondrial carrier 2	Hs01013175_m1
<i>RORA</i>	RAR-related orphan receptor A	Hs00536545_m1
<i>VPS13C</i>	vacuolar protein sorting 13 homolog C	Hs01569419_m1
<i>C2CD4A</i>	C2 calcium-dependent domain containing 4A	Hs01387999_g1
<i>C2CD4B</i>	C2 calcium-dependent domain containing 4B	Hs02379187_s1
<i>LARP6</i>	La ribonucleoprotein domain family, member 6	Hs00217969_m1
<i>CT62</i>	cancer/testis antigen 62	Hs03405497_m1
<i>YWHAE</i>	tyrosine 3-monooxygenase/tryptophan 5-monooxygenase activation protein, epsilon polypeptide	Hs00356749_g1
<i>MYO1C</i>	myosin IC	Hs00300761_m1
<i>SKIP (INPP5K)</i>	inositol polyphosphate-5-phosphatase K	Hs00213017_m1
<i>SGSM2 (RUTBC1)</i>	small G protein signaling modulator 2	Hs00390054_m1
<i>RAP1GAP2</i>	RAP1 GTPase activating protein 2	Hs00970363_m1

Table 4.4: Assay details for total expression quantification. Assay IDs are TaqMan inventoried gene expression assays (Applied Biosystems).

Each reaction multiplexed a FAM-labelled test assay and a VIC-labelled endogenous control assay (one of *B2M*, *RNaseP*, *HPRT* or *PPIA*), after standard curve validation showed that multiplexing had no impact upon amplification efficiency. Real-time PCR was carried out on an Applied Biosystems 7900HT machine as described in section 2.7.4. All samples were run in triplicate and in the presence of a reverse-transcriptase negative control. No samples showed amplification in negative control wells.

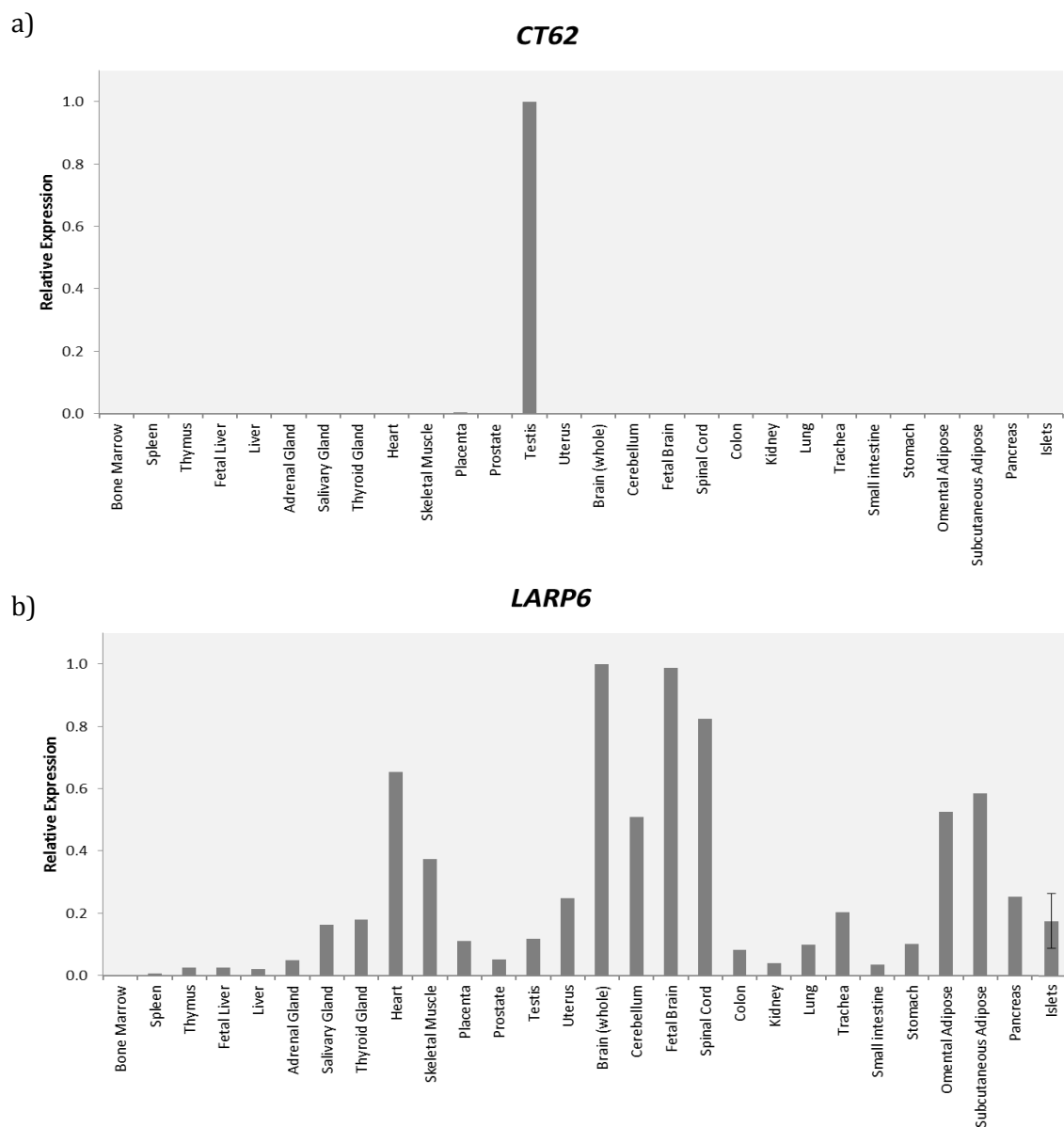
For expression quantification within one gene across many tissues (section 4.3.1), analysis was performed using the $\Delta\Delta C_t$ method as described in section 2.7.5. Expression was determined with respect to the 1:50 pooled sample standard and normalised to the geometric mean of three endogenous controls (*B2M*, *HPRT* and *PPIA*; *RNaseP* proved the least consistent of the four endogenous control genes tested). For ease of presentation, the maximum expression level in each gene is set to 1, and expression in other tissues reported as a fraction of this.

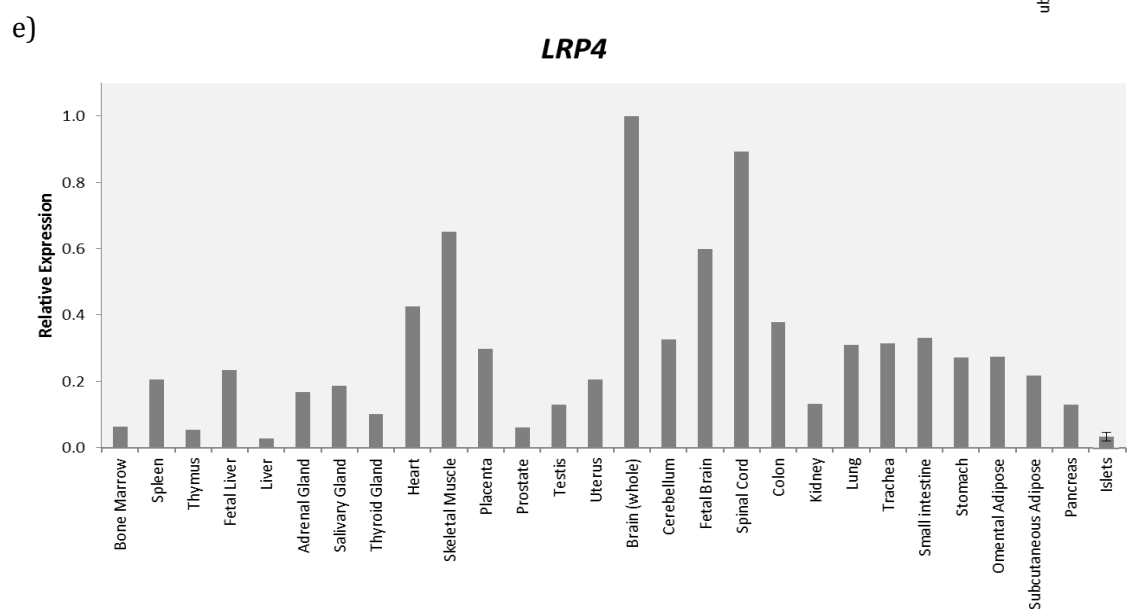
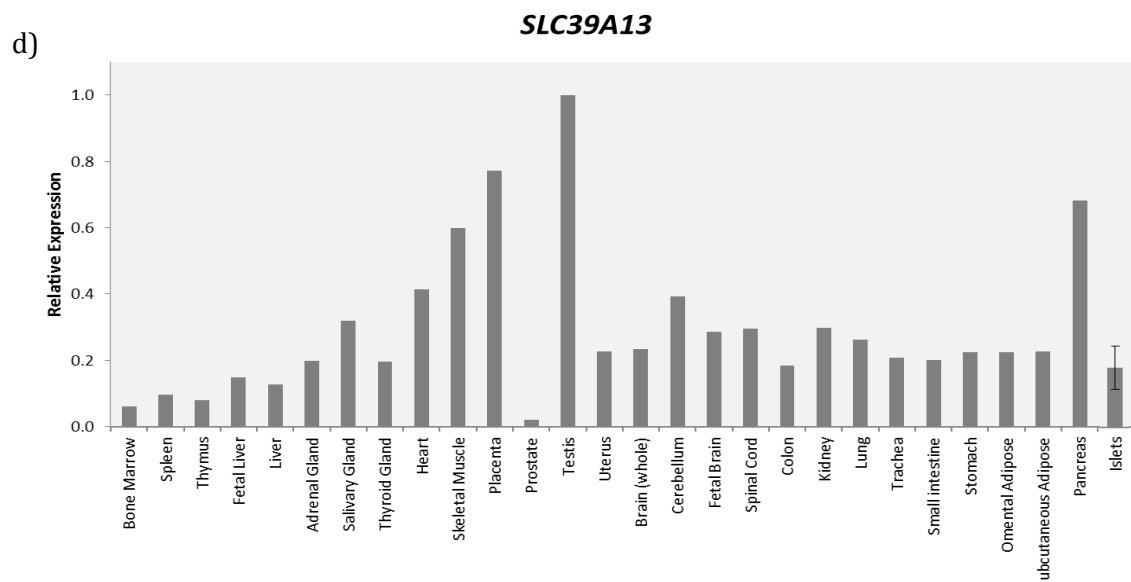
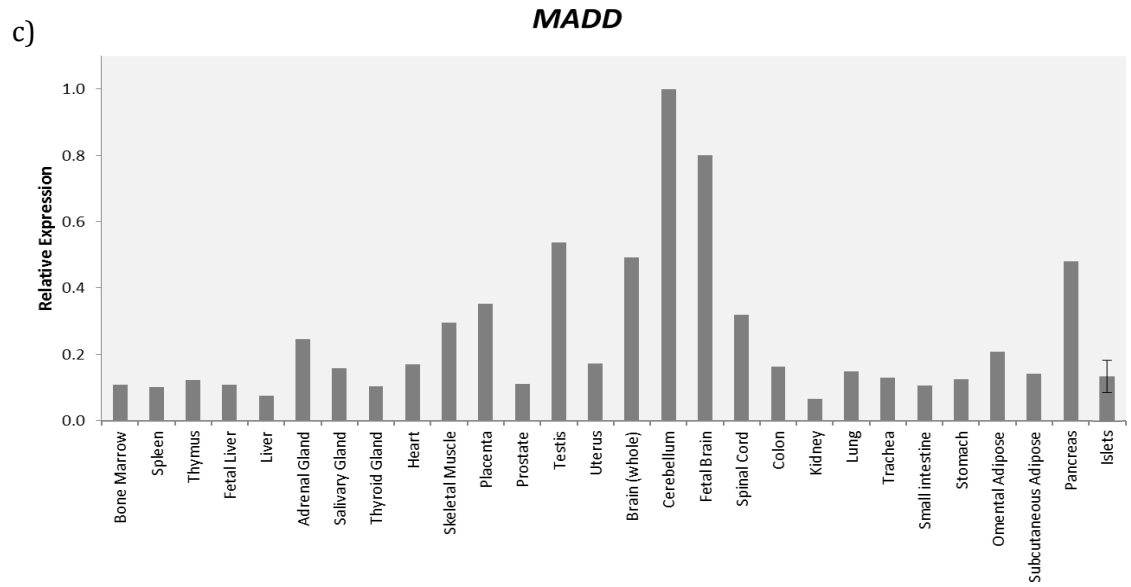
For expression quantification of many genes in human islets and metabolically relevant tissues, analysis was performed using the ΔC_t method. To make comparisons uniform, the critical threshold was set to the same value for each assay (test and endogenous control). Relative expression was then calculated through normalisation to the geometric mean of housekeeping genes *B2M*, *HPRT* and *PPIA*.

4.3 Results

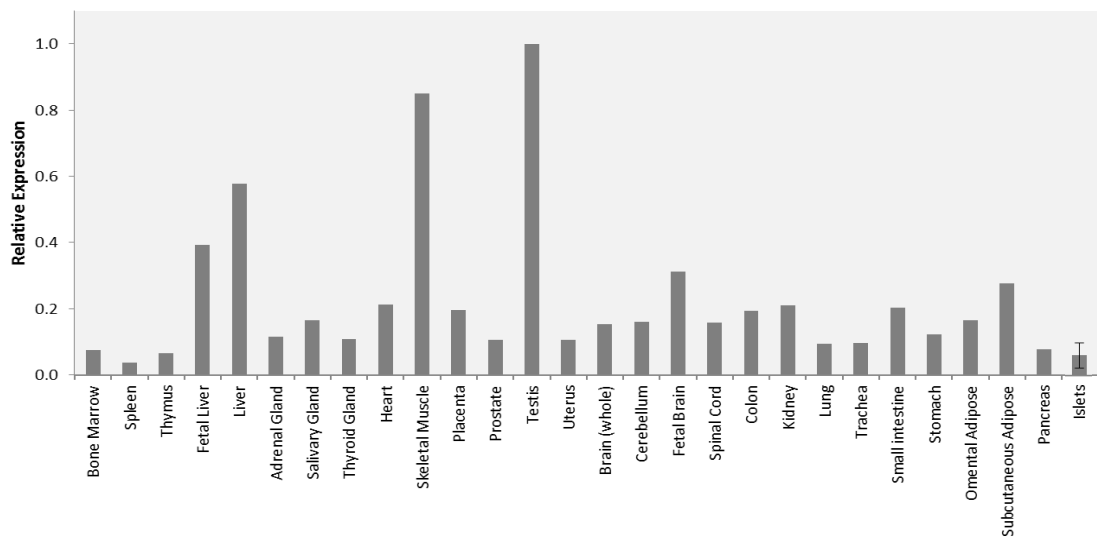
4.3.1 mRNA expression profiling of genes close to novel proinsulin loci across a human tissue panel

Figure 4.3(a-t) shows relative expression levels of genes transcribed from a 1Mb region around proinsulin associated variants, in a range of tissues from a commercial human tissue panel, supplemented with omental adipose, subcutaneous adipose, pancreas and islets. Expression is reported as a fraction of the gene tissue with maximum expression for each gene. Genes are arranged according to loci

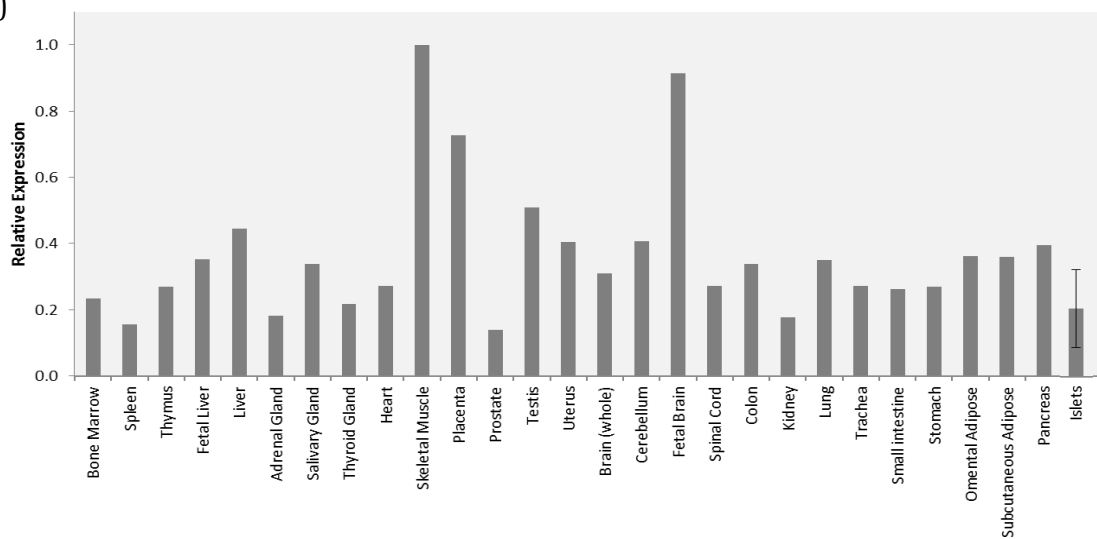




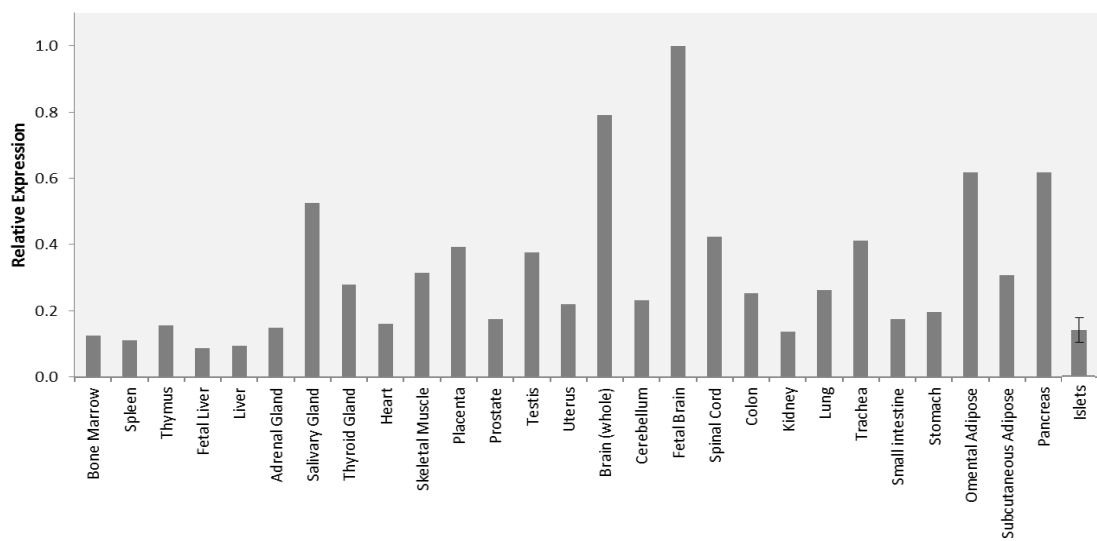
f)

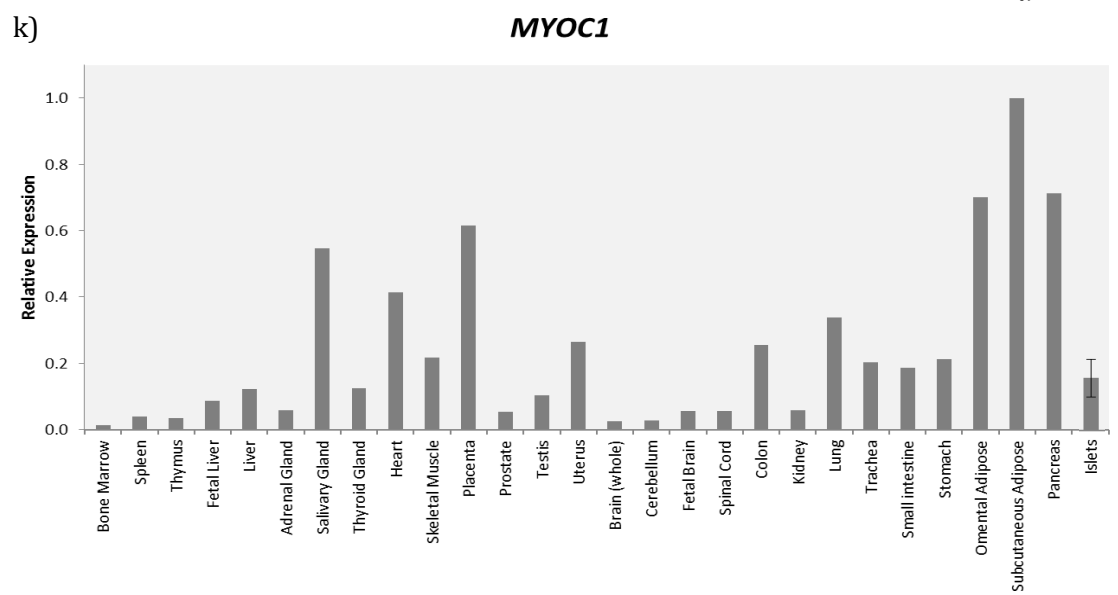
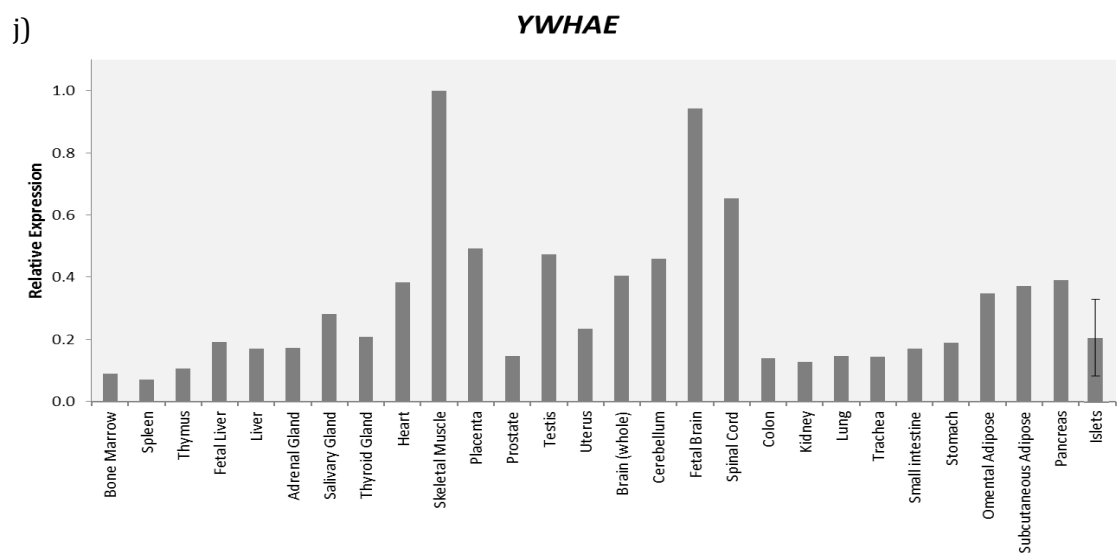
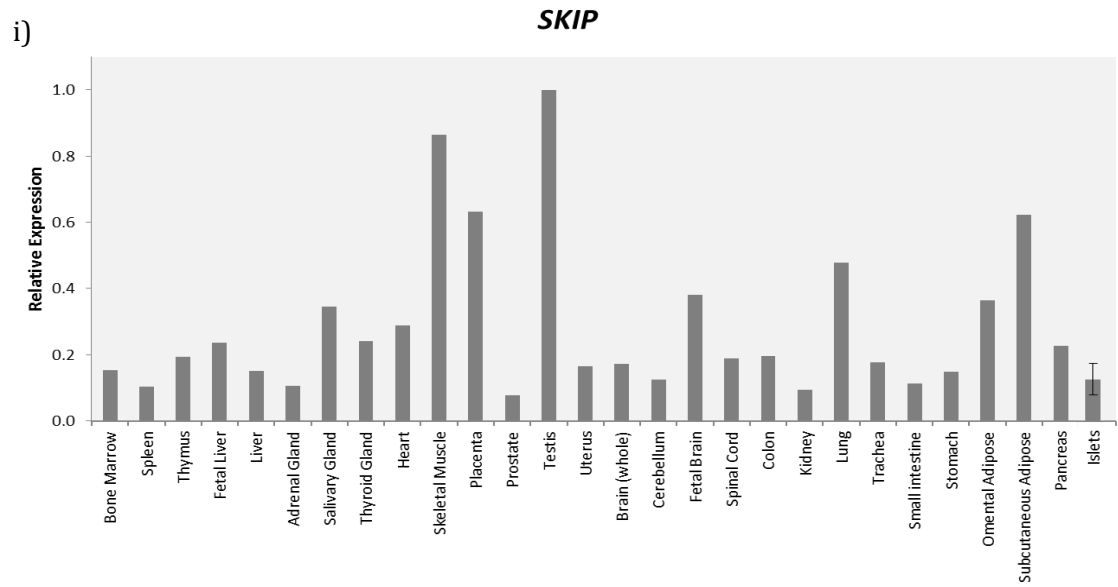
MTCH2

g)

CELF1

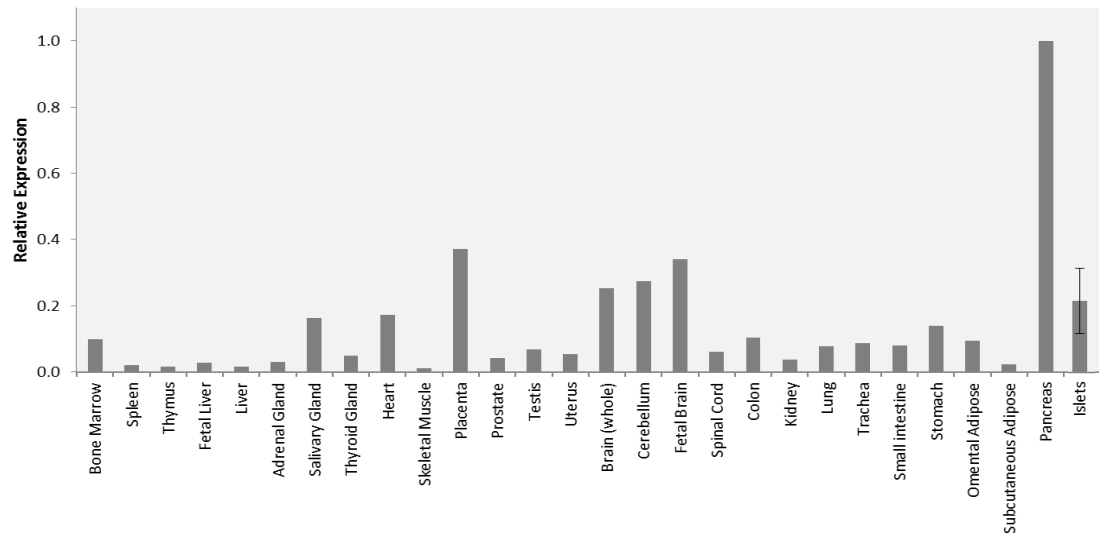
h)

SGSM2



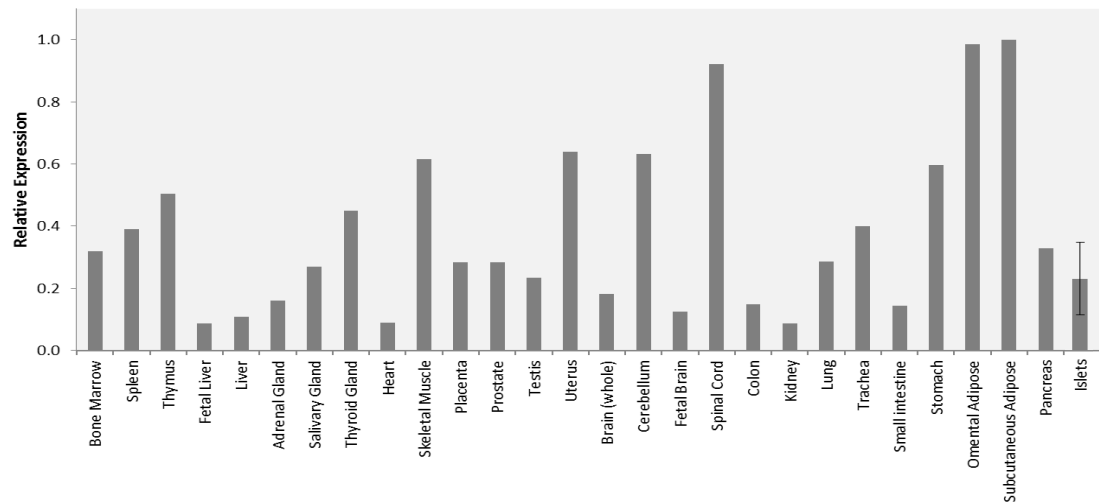
l)

RIP1GAP2



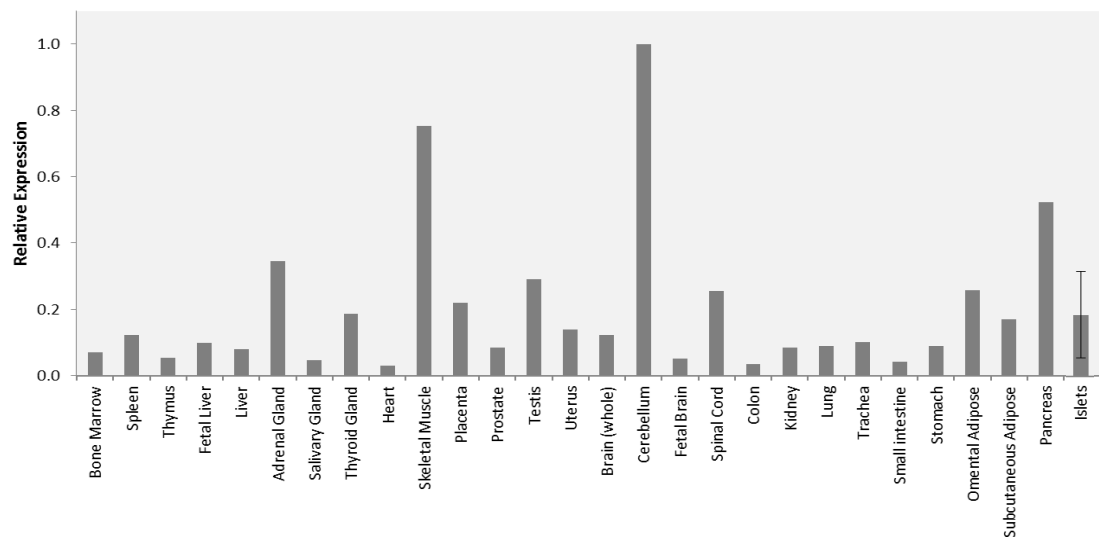
m)

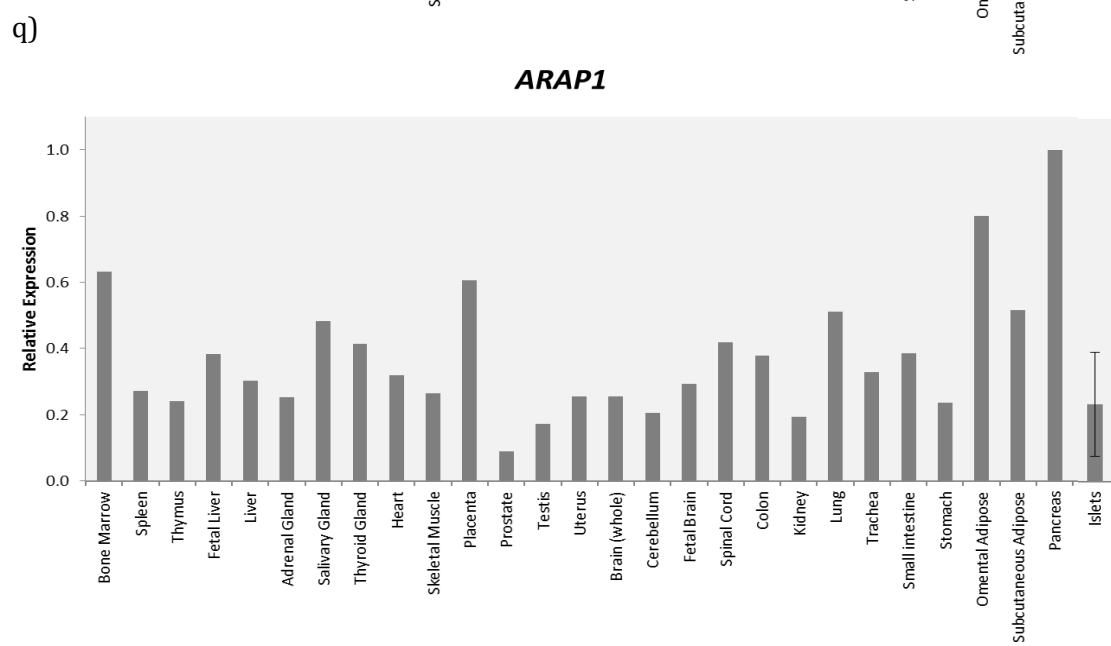
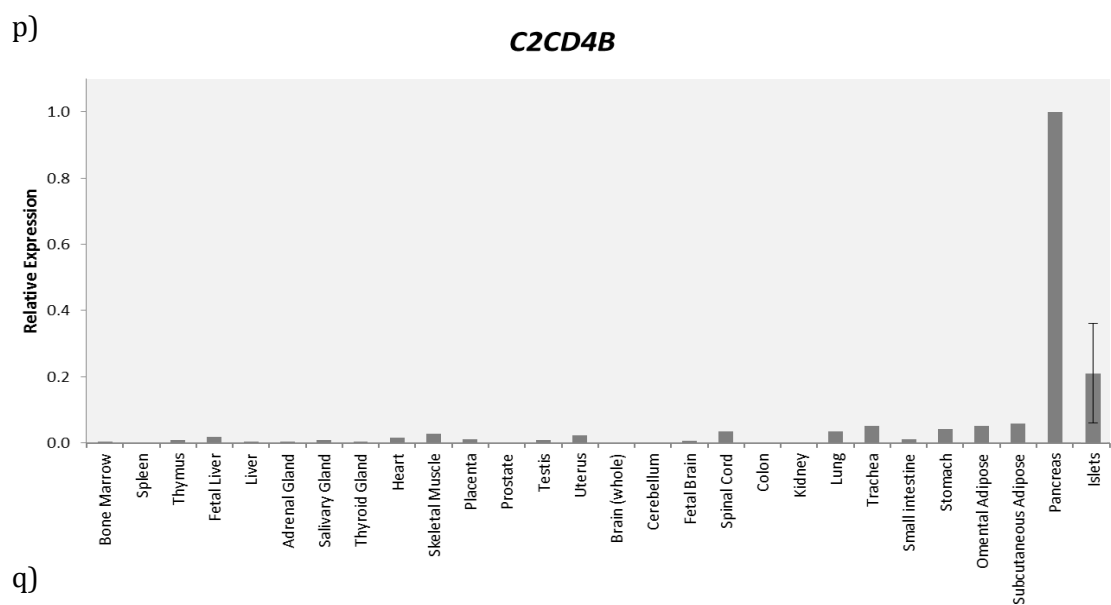
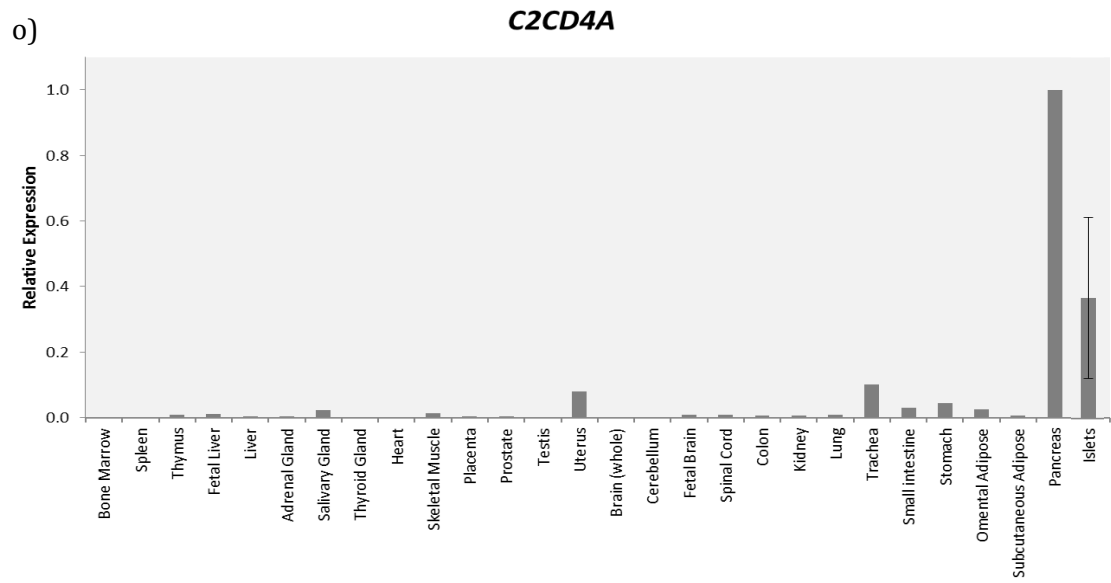
VPS13C



n)

RORA





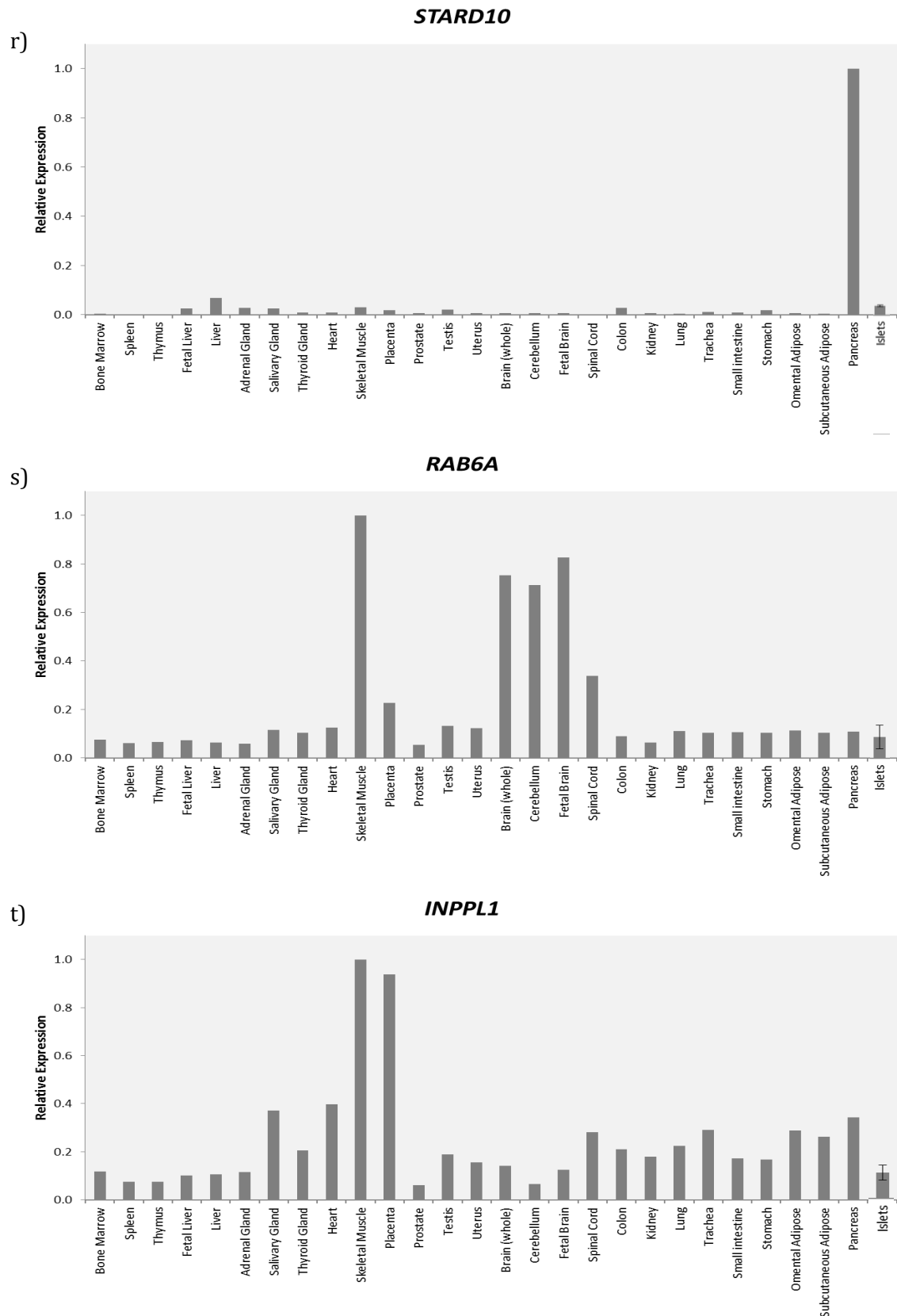


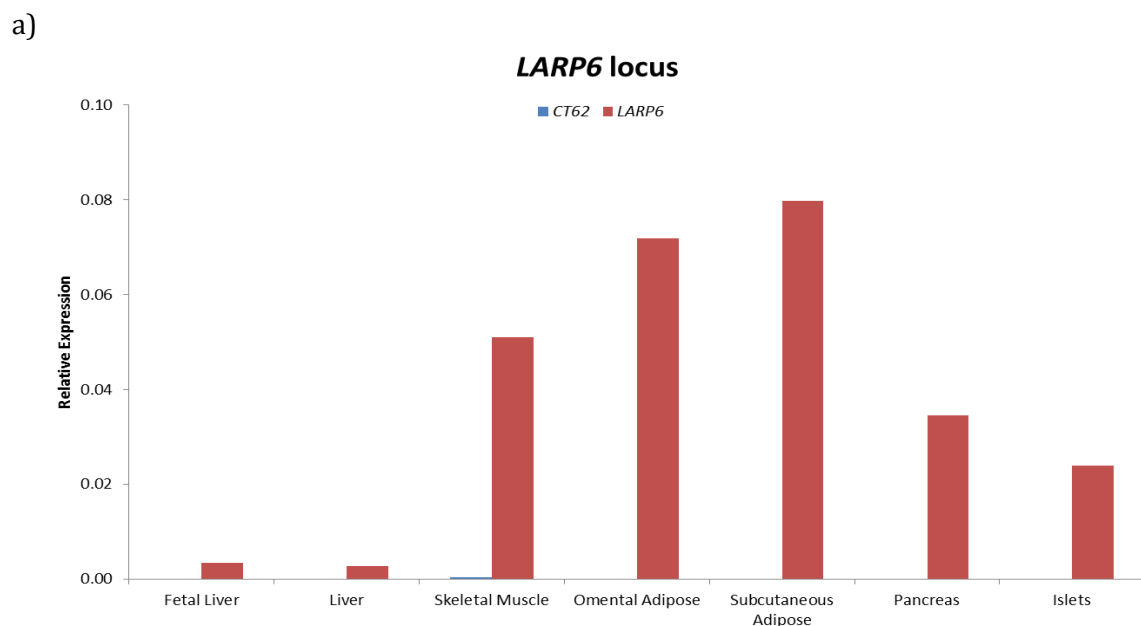
Figure 4.3(a-t): Relative expression levels of genes transcribed from a 1Mb region around pro-insulin associated loci. Expression shown in a range of human tissues, including liver, muscle, omental adipose, subcutaneous adipose, pancreas and islets. The maximum expression for each gene is set to one, and other expression levels reported as a fraction of this. Islet expression is presented as mean $\pm$ SD of three donors.

Many of the tested genes were widely expressed. *INPPL1* was ubiquitously expressed, with highest expression seen in skeletal muscle and placenta. *LARP6* and *LRP4* were widely expressed, with highest expression levels in nervous system tissues (brain and spinal cord). *MADD* was universally expressed, particularly highly in cerebellum and foetal brain, as well as pancreas tissue. *SGSM2* and *SLC39A13* were ubiquitously expressed, at highest levels in brain and testis. *VPS13C* and *MYOC1* were also widely expressed with highest levels in adipose tissues. *CELF1* and *YWHAE* showed widespread expression, particularly strongly in skeletal muscle. *MTCH2*, *RAB6A*, *RORA*, and *SKIP* were also universally expressed with particularly high levels in skeletal muscle, and also with strong expression in liver and testis, nervous system tissues, cerebellum and testis respectively. *ARAP1* showed widespread expression, with highest expression in pancreas.

Other genes were not widely expressed. *CT62* was expressed exclusively in testis tissue, whilst *C2CD4A* and *C2CD4B* showed expression in pancreases and islets almost to the exclusion of any other tissues tested. *STARD10* was expressed at dramatically higher levels in pancreas than any other tissue, including islets. *RIP1GAP2* was also expressed most strongly in pancreas, but with quantifiable expression across almost all other tissues.

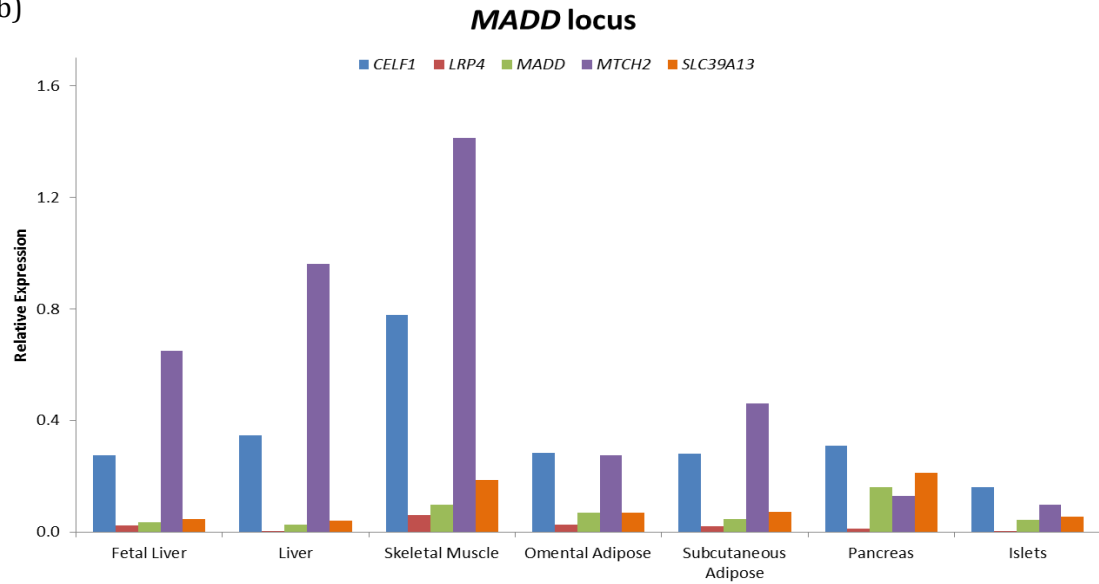
4.3.2 Comparisons of genes within each locus

It is not only the expression profiles of individual genes which is of potential interest in understanding disease associated loci. Comparative expression levels in relevant tissues of genes within each associated locus may be helpful in determining which gene is most likely to play an important role in those tissues. Figure 4.4 (a-e) presents the relative expression levels of genes within each of the loci associated with proinsulin levels, in metabolically relevant tissues. Expression levels are presented relative to the expression of endogenous control genes (see section 4.2.4).



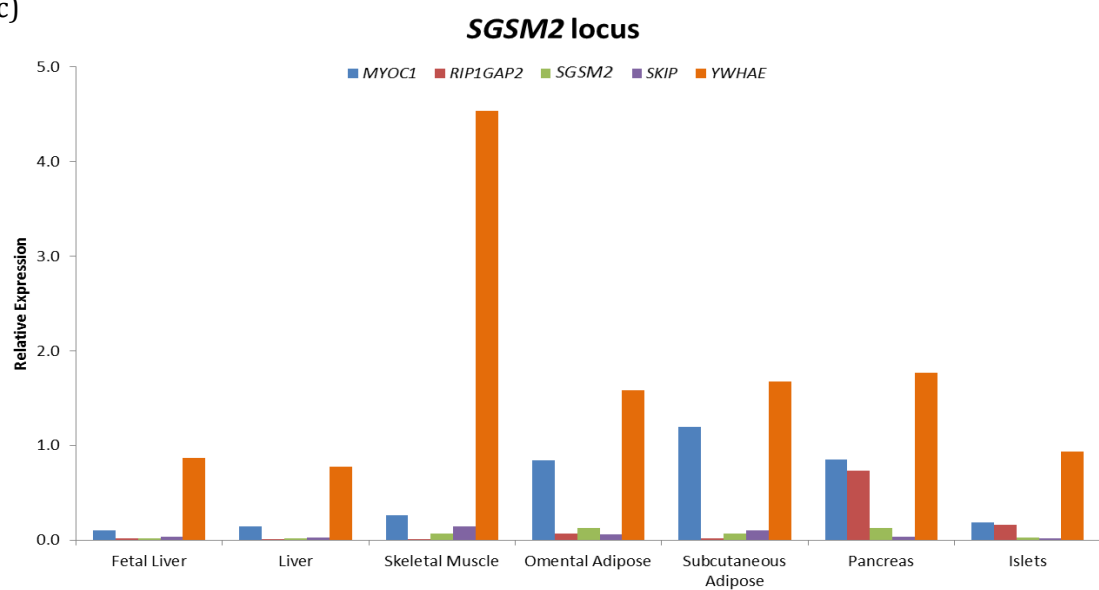
Since *CT62* is only quantifiably expressed in testis tissue, *LARP6* is the only gene present in relevant tissues at its locus. It is expressed in all relevant tissues, most strongly in adipose tissue and least strongly in liver.

b)



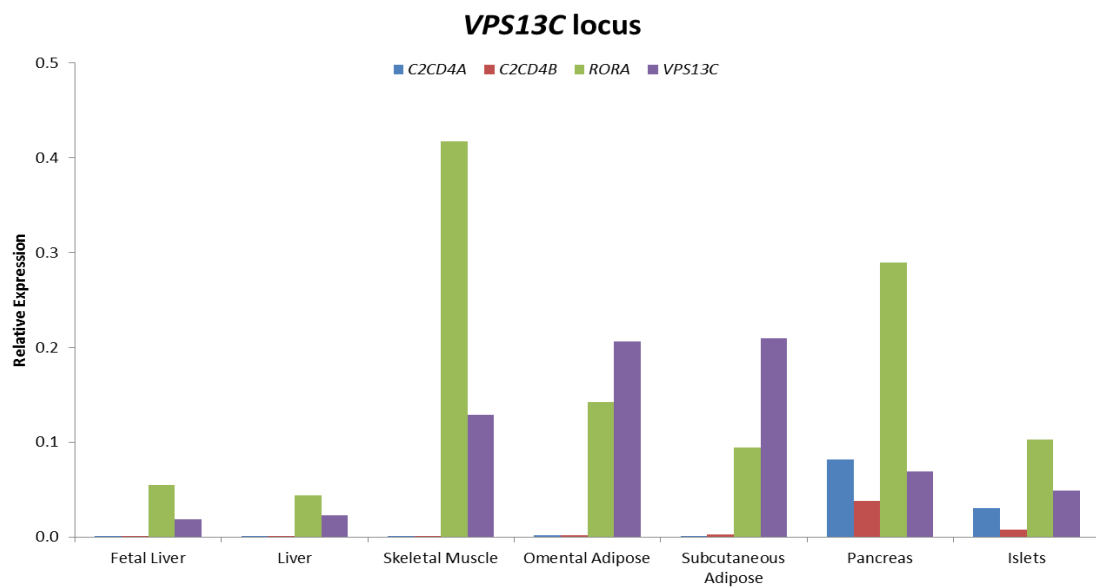
The *MADD* locus shows broadly similar levels of *CELF1*, *MADD*, *MTCH2* and *SLC39A13* expression in pancreas and in islet tissue. *LRP4* is comparatively lowly expressed in these tissues. *CELF1* and *MTCH2* are the predominant transcripts in both omental and subcutaneous adipose tissue. *CELF1* and *MTCH2* are also much more highly expressed than *LRP4*, *MADD* and *SLC39A13* in liver and skeletal muscle, with *MTCH2* being the most strongly expressed in both tissue types.

c)



YWHAE is the most prevalent transcript across all tissues, particularly skeletal muscle, at the *SGSM2* locus. Omental and subcutaneous adipose tissue also show expression of *MYOC1*, whilst pancreas and islets show some expression of *MYOC1* and *RIP1GAP2* but relatively little expression of *SGSM2* or *SKIP*.

d)



At the *VPS13C* locus, liver, skeletal muscle and both adipose tissue types show quantifiable expression of *RORA* and *VPS13C*, but almost no expression of *C2CD4A* or *C2CD4B*. *RORA* is the more prevalent transcript in liver and muscle, whilst *VPS13C* is more highly expressed in omental and subcutaneous adipose tissue. *C2CD4A* and *C2CD4B* are present in pancreas tissue and islets, although at lower levels than *RORA* and *VPS13C*.

e)

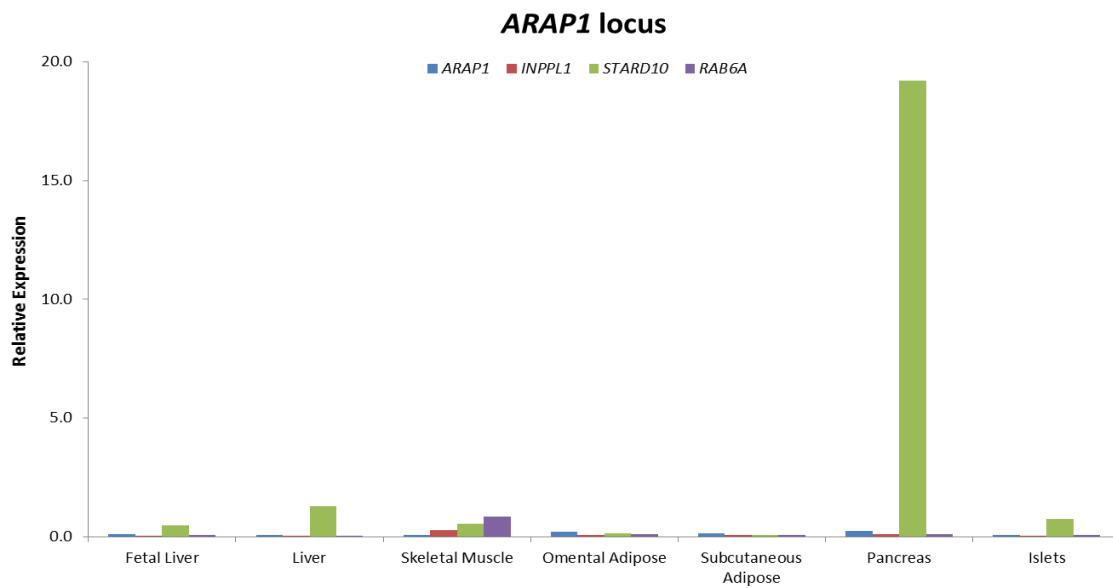


Figure 4.4(a-e): Relative expression levels of genes at each proinsulin-associated locus. Expression levels are presented in metabolically relevant tissues and relative to endogenous control genes. Gene names used to label each locus are arbitrary.

Expression at the *ARAP1* locus is dominated by the extremely high level of *STARD10* expression in pancreas. Interestingly, this strong expression is not reflected to a similar extent in islets, although *STARD10* is the most highly expressed gene in both islet and liver tissue. *RAB6A* is the most highly expressed gene in skeletal muscle.

4.4 Discussion

4.4.1 Comparison of data between pancreas, islets and β -cells

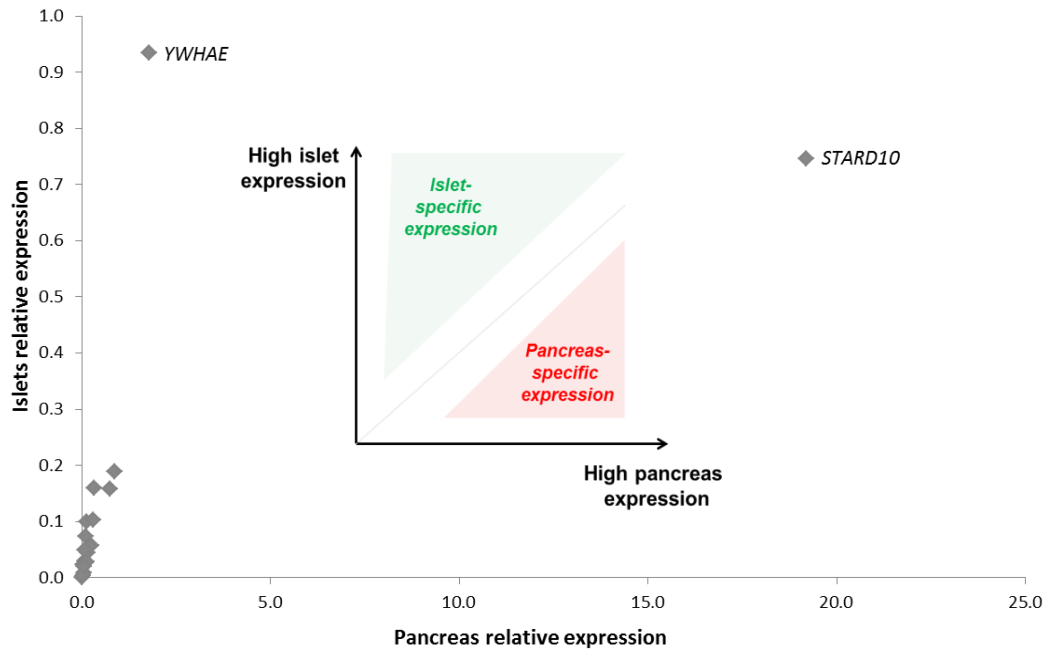
The majority of tested genes were quantifiably expressed in pancreas and islets. It had been anticipated that only a limited subset of genes would be expressed in islets, and therefore suitable for follow-up analysis in enriched β -cell samples. The reality, that few genes could be immediately excluded, is in line with results from other expression profile experiments following genome-wide association analysis (Dupuis et al., 2010; Voight et al., 2010).

However, even if genes are quantifiably expressed in pancreas, islets and enriched β -cells, it is still pertinent to compare the pattern of expression across these tissues. A gene expressed very strongly in pancreas but less so in islets, for example, might be significantly involved in exocrine rather than endocrine pancreas function. Conversely, a gene with stronger expression in islets than pancreas may be particularly relevant to endocrine systems. Similarly, a gene expressed strongly in islets but not in β -cells may be relevant to alpha or delta cell function, whilst a gene expressed more strongly in pure β -cells than whole islets may be particularly relevant to insulin secretion.

Figure 4.5a compares relative expression levels of all genes in islets and pancreas tissue. *STARD10* is immediately evident as a gene expressed extremely strongly in pancreas (11.2 times more highly than *YWHAE*, and 22.5 times more highly than any other gene), but less so in islet tissues (less highly than *YWHAE*, and only 4 times more than any other gene). *YWHAE*, on the other hand, is particularly strongly expressed in islets but not pancreas. Figure 4.5b removes *STARD10* and *YWHAE*, presenting other

genes on a more informative scale. Here, a correlation ($r^2 = 0.42$ for figure 4.5a, 0.77 for figure 4.5b) between expression in pancreas and expression in islets is evident.

a)



b)

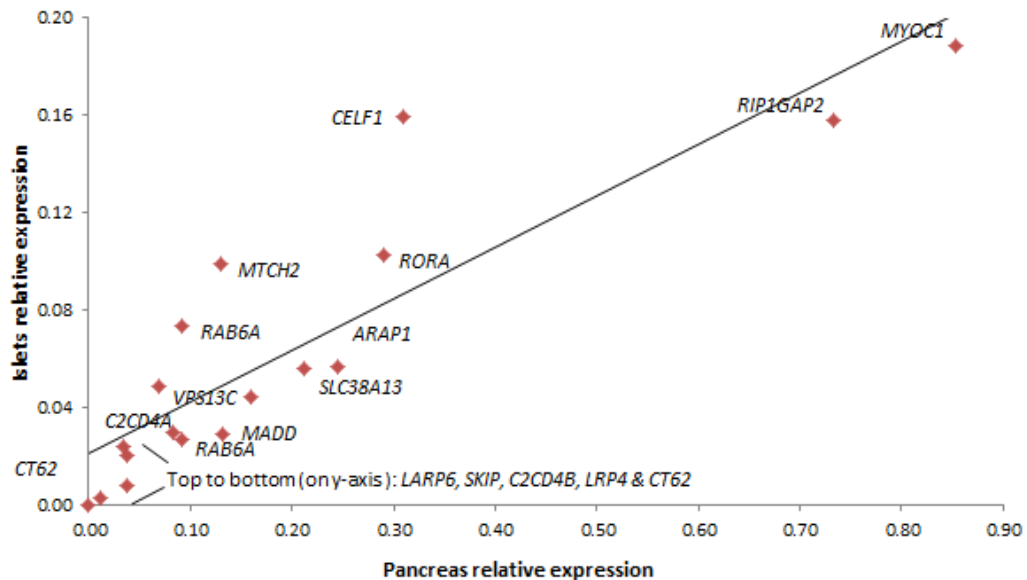


Figure 4.5(a&b): Relative expression levels of tested genes in human pancreas (x-axis) and human islets (y-axis). **a)** shows all genes and schematic to demonstrate islet-and pancreas-specific expression differentiation. **b)** shows the same data with *STARD10* and *YWHAE* removed to present other genes on a clearer scale. Line of correlation between islet and pancreas expression marked.

The genes *RAB6A*, *MTCH2* and *RORA* and fall above this line, indicating an excess of expression in islets over pancreas. Whilst this excess is not in absolute terms, it may still indicate an importance to the endocrine cells of the pancreas. The relevance of each of these genes will be discussed in section 4.4.2.

It is possible that the broadly similar pattern of expression between pancreas tissues and islets could reflect a significant level of exocrine contamination in the assessed islet samples. However, the islet samples used were >80% pure. It therefore seems reasonable to suppose that any significant difference in expression patterns would be evident above any noise caused by contamination. Furthermore, expression patterns in enriched β -cells (>95% pure) very closely resembled patterns in islets (figures 4.4 and 4.6).

Figure 4.6 shows expression data from enriched β -cells, performed by Dr Nabila Bouatia-Naji in Lille, France. The pattern of expression is identical between islets (see

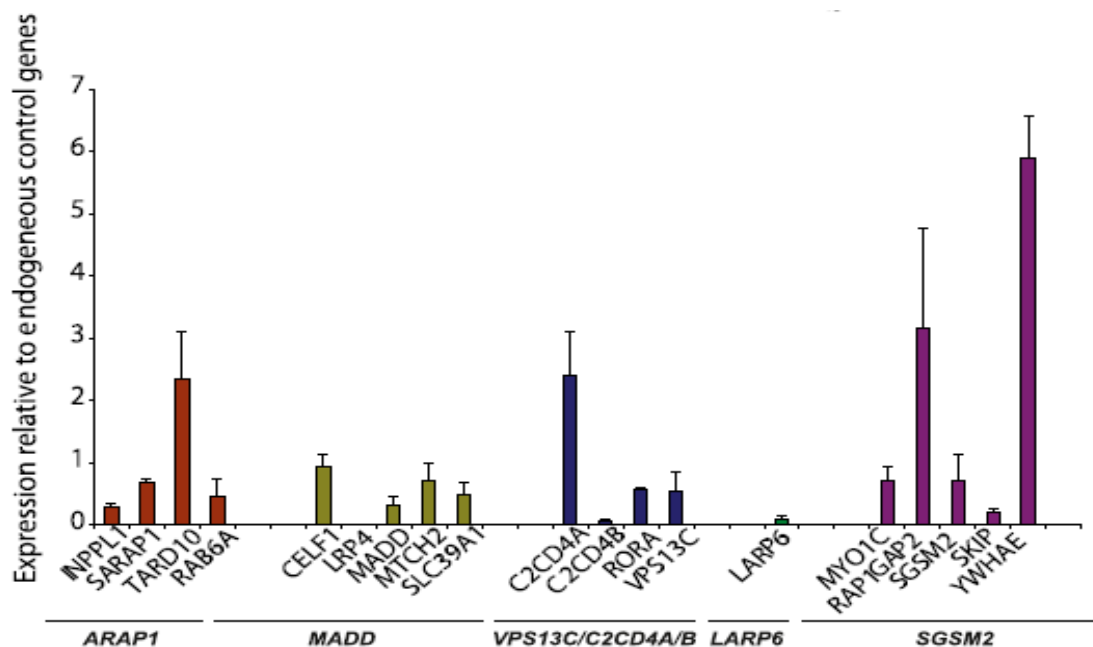
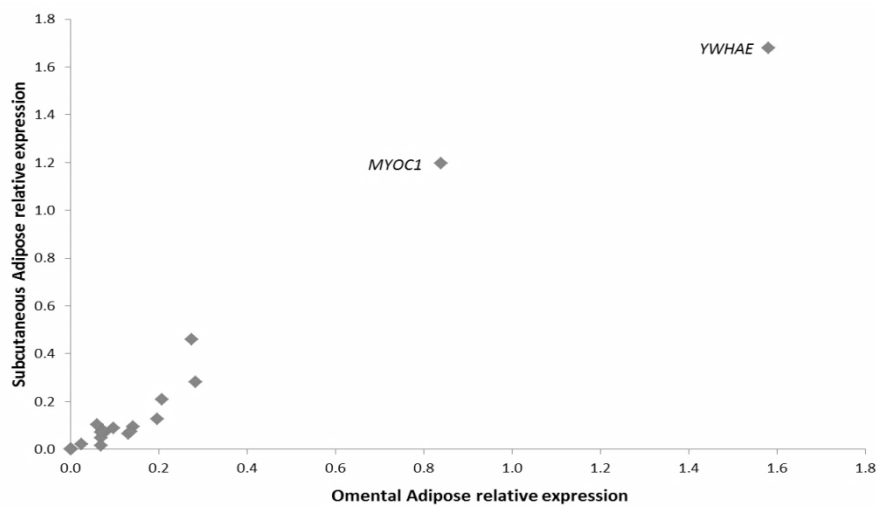


Figure 4.6: Expression levels of genes near proinsulin-associated variants in human FAC-sorted β -cells. Data are mean $\pm$ SD from three donors. Figure taken from (Strawbridge et al., 2011).

figure 4.4) and β -cells at the *LARP6*, *ARAP1* and *MADD* loci. However, at the *SGSM2* locus, *RIP1GAP2* is expressed relatively more highly in β -cells than in islets, whilst at the *VPS13C* locus *C2CD4A* is expressed relatively more strongly in β -cells than in islets. Their selective β -cell expression means that *RIP1GAP2* and *C2CD4A* may be particularly good candidates for mediating the proinsulin effect at their respective loci, and will be discussed in section 4.4.2. The expression patterns between foetal and adult liver, and between omental and subcutaneous adipose tissues were also extremely similar (figure 4.7), indicating a broadly similar expression profile in these tissues.

a)



b)

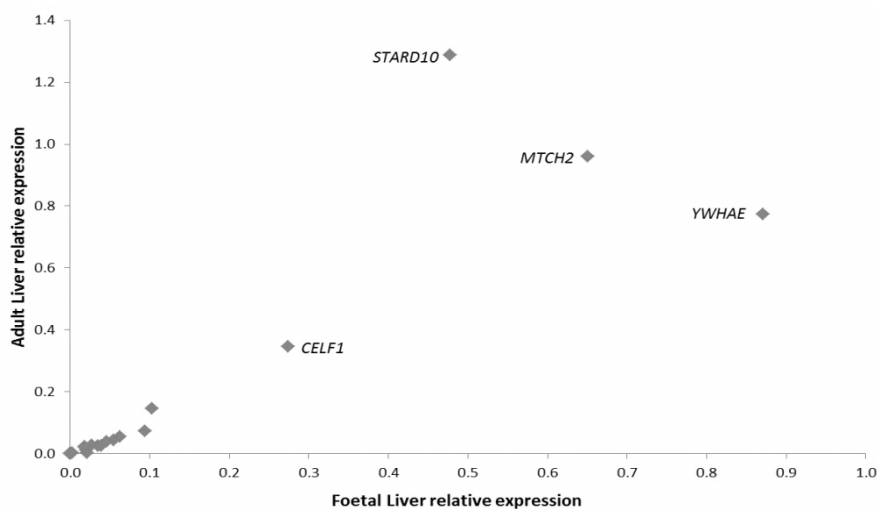


Figure 4.7(a&b): Relative expression levels of tested genes in **a)** subcutaneous and omental adipose tissue, and **b)** in adult and foetal liver.

4.4.2 Consideration of individual loci

In the current absence of islet eQTL data, expression profile analysis is a major route to establishing the relevance of candidate genes in an association signal to disease phenotype. The information presented in sections 4.3.1 and 4.3.2 must be considered together to gain the best insight into each locus. It may be relevant that a given gene is the most highly expressed in islets at a particular locus, but it may also be relevant if a gene is expressed only in islets, even at a comparatively low level. Relative strength and specificity of expression are both important considerations. The MAGIC consortium also conducted analysis of proinsulin-associated variants with other glucose and insulin-related traits (including fasting insulin, fasting glucose, insulin resistance [as measured by HOMA-IR], β -cell function [as measured by HOMA-B] and T2D risk), which will be discussed below.

4.4.2 (a) *LARP6* locus

Gene expression profiling clearly highlighted the gene most likely to be responsible for proinsulin association at the *LARP6* locus. By showing that *CT62* is not expressed in any metabolically relevant tissues (and is in fact expressed exclusively in testis amongst tissues tested), it can be excluded as a relevant gene (figure 4.3a). *LARP6*, as the only remaining candidate gene at the locus, is therefore likely to be the gene through which nearby variants exert their effect on proinsulin levels.

LARP6 encodes a ribonucleoprotein, which is involved in the regulation of translation and subcellular localisation of collagen (Cai et al., 2010). The mechanism through which it may influence proinsulin levels is unclear. However, since the proinsulin-raising allele at this locus is nominally associated with increased fasting insulin and

HOMA-IR ($p=0.02$ and 0.044 respectively), but *increased* HOMA-B ($p=0.007$), the variant may exert its effect on proinsulin via insulin resistance rather than deficient proinsulin processing itself. Insulin resistance places a demand on the β -cell for increased insulin secretion. If the rate of proinsulin transcription is increased but the proinsulin processing pathway is unable to entirely keep pace, more unprocessed proinsulin will be secreted and the PI/I ratio will be raised (Strawbridge et al., 2011).

LARP6 is widely expressed, including strongly in adipose tissue (figure 4.3b), and further studies by the MAGIC consortium found evidence that the proinsulin-raising allele is also associated with lower levels of *LARP6* expression in adipose tissue (Strawbridge et al., 2011).

4.4.2 (b) *MADD* locus

The proinsulin-raising allele at the *MADD* locus is also associated with increased fasting glucose and decreased HOMA-B ($p=3.1 \times 10^{-9}$ and 8.1×10^{-5} respectively), but not with HOMA-IR or fasting insulin. This suggests that associated variants work through proinsulin processing in the β -cell itself, rather than increasing the PI/I ratio through insulin resistance.

All tested genes at the *MADD* locus have a strong claim for involvement in β -cell processes. *MADD* itself encodes an adaptor protein which interacts with MAP kinases to propagate apoptotic signals. The MAP kinase pathway has been implicated in the proliferation and apoptosis of β -cells (Trumper et al., 2001), suggesting that *MADD* could contribute to beta-cell mass and therefore proinsulin processing potential.

SLC39A1C encodes a putative zinc transporter (Fukada et al., 2008), and therefore could be involved in the zinc-insulin crystallization which occurs within secretory vesicles. The low-density lipoprotein receptor-related protein *LRP4* is a negative regulator of Wnt signalling (Li et al., 2010), whilst *MTCH2* encodes a putative mitochondrial carrier protein and harbours variants recently shown to associate with obesity (Kulyte et al., 2011). *CELF1* is involved in mRNA alternative splicing, including the exclusion of exon 11 from *INSR* (encoding the insulin receptor) pre mRNA (O'Leary et al., 2010).

Expression profiles of these genes individually are not particularly informative, in that they are all widely expressed and none shows particularly strong expression in pancreas or islets (relative to other tissues). *MADD* (figure 4.3c) is expressed most strongly in cerebellum, and fairly highly in pancreas although less so in islets. *SLC39A1C* (figure 4.3d) and *MTCH2* (figure 4.3f) are most strongly expressed in testis tissue, whilst *CELF1* (figure 4.3g) and *LRP4* (figure 4.3e) are expressed at highest levels in skeletal muscle and brain respectively.

In islets themselves, *CELF1* is the most strongly expressed regional gene, followed by *MTCH2*, and then *MADD* and *SLC29A13* at similar levels. *LRP4* expression levels are comparatively low (figure 4.4b). This pattern is replicated exactly in enriched β -cells, where *LRP4* expression is sufficiently low to be unquantifiable. *CELF1* and *MTCH2* are also expressed more highly in islets compared to pancreas than most genes tested (fall above the line in figure 4.5b). On the basis of comparative expression data then, *CELF2* and *MTCH2* look the best candidates for mediating a β -cell proinsulin processing effect.

4.4.2 (c) *SGSM2* locus

The proinsulin-raising allele at the *SGSM2* locus is also associated with increased fasting insulin and increased HOMA-IR ($p=0.04$ and 0.023 respectively), but not with HOMA-B or insulinogenic index. It therefore seems likely that the gene through which this variant works raises PI/I ratios through an effect on insulin resistance, increasing demands placed on the β -cell proinsulin processing pathways.

SGSM2 is involved in the sorting and recycling of vacuoles between early endosomes and the plasma membrane (Yang et al., 2007), including recycling of the glucose transporter *GLUT4* and the glucagon receptor (Kessler et al., 2000; Krilov et al., 2008). It is widely expressed, most strongly in brain (figure 4.3h). *YWHAE* belongs to the family of proteins which mediate signal transduction by binding to phosphoserine-containing proteins, possibly regulating insulin sensitivity. *SKIP* (an isoform of *INPP5K*) negatively regulates insulin signalling by inhibiting phosphorylation of protein kinase B (*PKB/AKT*). It has been shown to increase insulin-induced *GLUT4* translocation and glycogen synthesis in rat myocytes (Ijuin and Takenawa, 2003). Consistent with roles in insulin signalling sensitivity, both *YWHAE* and *SKIP* are strongly expressed in skeletal muscle and adipose tissue (figures 4.3i and 4.3j). *MYOC1* is a myosin which utilises energy from ATP to generate mechanical force. It has been previously reported to be highly expressed in adipocytes, where it is part of the insulin signalling pathway which controls the movement of *GLUT4*-containing vesicles to the plasma membrane (Bose et al., 2002). Data presented here (figure 4.3k) supports this strong expression in adipose tissues, and also suggests strong expression in pancreas. Finally, *RAP1GAP2* encodes a GTPase-activating protein which is essential for the regulation of insulin granule dynamics by cAMP (Shibasaki et al., 2007). As might be

expected given this role, it is strongly expressed in pancreas and islet tissues, as well as in placenta (figure 4.3l).

YWHAE is the most strongly expressed gene at this locus in both pancreas and islets, and tissues of insulin action (figure 4.4c). *SGSM2* and *SKIP* are comparatively lowly expressed in these tissues, whilst *MYOC1* is strongly expressed in adipose tissue and pancreas only. *RIP1GAP2* shows similar expression to *MYOC1* in pancreas and islets, but is strikingly more highly expressed in enriched β -cells. If proinsulin-raising alleles at these locus were associated with indices of β -cell function, then this would make *RIP1GAP2* an attractive candidate for the mediating the effect. However, since proinsulin-raising alleles are instead associated with indices of insulin resistance, the genes strongly expressed in the tissues of insulin action – namely *MYOC1* and *YWHAE* – are more likely candidates.

4.4.2 (d) *VPS13C* locus

The proinsulin-raising allele at the *VPS13C* locus is known to also be associated with T2D and impaired β -cell function (Grarup et al., 2011; Yamauchi et al., 2010). The lead SNP in the MAGIC analysis (rs4502156) is in tight LD ($r^2 = 0.94$) with the lead SNP (rs7172432) identified in a T2D GWA study in a Japanese population (Yamauchi et al., 2010). In the MAGIC analysis, rs4502156 was associated with T2D and reduced HOMA-B ($p=0.001$ and 1.4×10^{-7}), but not with HOMA-IR or fasting insulin. As with association at the *MADD* locus, a variant which raises proinsulin but lowers insulin secretion suggests a defect in the β -cell proinsulin processing and secretion pathway.

VPS13C is involved in the trafficking of proteins between the trans-Golgi network and the pre-vacuolar compartment (Velayos-Baeza et al., 2004). It is widely expressed, most strongly in adipose tissue (figure 4.3m). *RORA* encodes a nuclear hormone receptor, which regulates the expression and secretion of fibroblast growth factor 21, a hepatic hormone that regulates peripheral glucose tolerance and hepatic lipid metabolism (Wang et al., 2010b). It is also widely expressed, most strongly in cerebellum and skeletal muscle (figure 4.3n). It is the most strongly expressed gene at this locus in liver (figure 4.4d). *C2CD4A* and *C2CD4B* encode nuclear factors which are upregulated in response to inflammation. It has been hypothesised that they regulate other genes which control cellular architecture (Warton et al., 2004). Both genes are strikingly selectively expressed in pancreas and islets, with barely quantifiable expression elsewhere.

Both *RORA* and *VPS13C* are more highly expressed in islets than *C2CD4A* or *C2CD4B* (figure 4.4d). However, pancreatic and islet-specific expression of *C2CD4A* and *C2CD4B* (figures 4.3o and 4.3p) suggest that they may play a key role in the endocrine functions of the pancreas. Furthermore, *C2CD4A* (but not *C2CD4B*) is the most highly expressed regional gene in enriched β -cells – expressed at higher levels than *RORA* (figure 4.4d). The selectivity and strong β -cell expression of *C2CD4A* therefore makes it an excellent candidate for a proinsulin processing effect. Little is currently known about the function of *C2CD4A*, but, on the basis of these experiments, it certainly warrants further investigation.

4.4.2 (e) *ARAP1* locus

Associations with insulin and glucose-related traits at the *ARAP1* locus are counterintuitive. The proinsulin-raising allele is associated with *lower* fasting plasma glucose ($p=1.7\times 10^{-4}$), *improved* β -cell function ($p=1.4\times 10^{-4}$) and *lower* risk of T2D ($p=7.8\times 10^{-6}$) (reported as *CENTD2* in (Dupuis et al., 2010)). It is, however, also associated with increased fasting insulin ($p=0.02$).

In general, *raised* PI/I ratios are associated with increased T2D risk (see section 4.1.3). It is possible that at this locus, the genetic defect which increases T2D risk causes a general down-regulation of insulin secretion, ensuring that less immature insulin is secreted and thus lowering the PI/I ratio. In any case, the evidence from T2D and other insulin-related traits demonstrates that disproportionate elevations *or* reductions in proinsulin can indicate β -cell dysfunction, and suggests that the responsible gene is likely to play a role in the β -cell.

ARAP1, *RAB6A* and *STARD10* have clear potential relevance to the β -cell. *ARAP1* mediates the formation of filopodia, cytoplasmic projections involved in the sensation of chemotropic cues and cell migration in response to these cues. It could, therefore, be involved in the maturation and migration of the secretory granule. *RAB6A*, meanwhile, is expressed in many exocytotic vesicles (Grigoriev et al., 2007) and regulates the membrane fusion of exocytotic carriers. *STARD10* encodes a lipid transfer protein which regulates lipid transfer activity by preventing membrane association (Olayioye et al., 2007). It may thereby influence lipid composition of the membrane, which in turn can be envisaged to interfere with membrane fusion and vesicle insulin secretion.

INPPL1 is involved in insulin signalling. It encodes a phosphatase which negatively regulates the PI3K (phosphoinositide 3-kinase) pathway, and overexpression reduces insulin-stimulated MAP kinase activation. However, silencing of *INPPL1* does not affect insulin signalling or GLUT4 trafficking. It is widely expressed (figure 4.3t), most strongly in skeletal muscle – one of the main tissues of insulin action.

Amongst the potentially β -cell relevant genes, *STARD10* has the most striking expression profile; its extremely strong expression in pancreas dominates regional expression plots (figures 4.3r and 4.4e). Although it is not expressed so dramatically highly in islets, it remains the most highly expressed regional gene (figure 4.4e). It is also the most highly expressed regional gene in pure β -cells (figure 4.6), which indicates that high expression in islets is unlikely to be the result of exocrine sample contamination. On this basis, it is probably the most interesting candidate gene at the *ARAP1* locus. Neither *RAB6A* nor *ARAP1* are strongly expressed in islets (figures 4.3q and 4.3s), although expression of *RAB6A* in islets is in excess of that which might be predicted basis of whole pancreas expression (figure 4.7), which may suggest endocrine specificity.

4.4.3 Study Limitations

The gene expression profiling data outlined in this chapter has contributed to resolving the complex challenge of interpreting genome-wide association results for proinsulin. It has identified *LARP6* as the probable gene through which the proinsulin associated variants on chromosome 15 work, and identified *CELFI* and *MTCH2*, *MYOC1* and *YWHAE*, *C2CD4A* and *STARD10* as particularly interesting genes within their respective loci. Whether these genes are implicated in proinsulin levels through insulin resistance or proinsulin processing directly, they are also highlighted as potentially relevant to the pathogenesis of T2D.

However, just as the existence of statistical association does not prove causation, neither does the demonstration of strong or specific expression. Whilst expression profiling can suggest that genes are more or less likely to be relevant to a given phenotype, it cannot prove molecular mechanisms. It is best used as one of many tools to add biological evidence to GWA data, highlighting key genes which constitute good targets for functional analysis. Other information may come from the refinement of associated regions through fine-mapping and genotyping in trans-ethnic populations, the identification of rare coding mutations contributing to disease phenotypes in candidate genes and, when islet sample numbers allow, eQTL data. It is hoped that the results reported in this chapter will be useful for the future selection of functional targets.

Chapter 5

Common structural variation and T2D risk

5.1 Introduction

5.1.1 Genome-wide association studies for Copy number variation: the Wellcome Trust Case Control Consortium (WTCCC) CNV study

SNPs within copy number variant sequence were largely excluded from early GWAS efforts, because they tend to fail quality control checks and were therefore excluded from genotyping chips (Lee et al., 2008). For example, a parent carrying an A allele at a variant site on one chromosome and a deletion on the other will appear to be homozygous A, yet their child who inherits the deleted chromosome and a G allele from the other parent will appear to be homozygous G. In this way, SNPs in a copy number variable region may appear to violate Mendelian inheritance. Older genome-wide typing chips (such as the Affymetrix 500k) lack any probes within 75% of large (>1kb) and common (MAF > 0.05) deletions. Newer platforms (for example Illumina 1M and Affymetrix 6.0) have improved coverage and deliberately target some copy number variant sequence by using overall genotyping probe intensity, but 30% of common CNVs remain completely excluded and very few are interrogated by multiple probes (Cooper et al., 2008).

In response to this paucity of existing coverage, the WTCCC initiated the first CNV-specific GWAS for common human disease in 2008. As outlined in section 1.4.1, there was a strong belief that copy number variation could explain some of the missing common disease heritability. The study analysed 10,894 polymorphic CNVs in 19,050 samples, including ~2,000 cases each from breast cancer, coronary artery disease, hypertension, Crohn's disease, bipolar disorder, rheumatoid arthritis, type 1 diabetes (T1D) and T2D disease cohorts, together with ~3,000 shared controls from United

Kingdom Blood Service (UKBS) and 1958 Birth Cohort (58BC) sample sets. CNVs were selected to the basis of a discovery project which used 42 million oligonucleotide microarray probes to generate a comprehensive map of 11,700 CNVs (Conrad et al., 2010). Analysis was conducted on the Agilent Comparative Genomic Hybridization (CGH) platform, which hybridises differentially labelled test and reference samples to an array, before comparing label intensity as an indication of DNA quantity.

Within the WTCCC consortium, my planned role was to follow up the strongest T2D-associated CNVs emerging from the GWAS, and, in preparation for this work, I optimised and compared alternative methods of CNV genotyping.

5.1.2 The necessity and challenge of genotyping copy number variants

Signals identified in GWA studies require confirmation and replication, especially when, as for CNVs, there are significant challenges in determining accurate genotypes. In anticipation of data emerging from the WTCCC CNV scan, laboratories with access to disease-specific replication cohorts needed to establish practical and reliable methods for high-throughput genotyping of copy number variants. The laboratory in which I work holds T2D disease cohorts, and was tasked with validating (using alternative assays) and replicating (using additional samples) interesting signals for association with T2D.

Large-scale genotyping of CNVs is not straightforward. Between 2004 and 2007, whilst the number of CNVs catalogued in discovery efforts grew to over 4,000, the number of CNVs successfully genotyped in replication cohorts reached only a few hundred

(McCarroll et al., 2006). Genotyping multibase (and often multiallelic) CNVs is a fundamentally different problem to genotyping single-base, diallelic SNPs. The challenge is not to discriminate *between* different molecules, but to *count* a single kind of molecule with high precision. Assays are therefore required to measure relative rather than absolute changes and to distinguish between experimental noise and biological signal - increasingly challenging at higher copy numbers where the differences between copy number classes become proportionally smaller. Whilst in CNV discovery projects a high false negative rate (failure to detect variant) is tolerated in order to preserve a low false positive rate (calling a spurious CNV), replication genotyping must achieve low false negative *and* low false positive rates.

There are a variety of approaches to genotyping a CNV. Where it is simple deletion with well-defined breakpoints, genotyping may be possible by means of a PCR-based break-point assay (figure 5.1). A reverse primer is designed on one flank of the deletion, which can amplify in conjunction with two alternative forward primers; one within the deleted region and one outside it. Dependent upon the deletion status of a sample, amplification will occur from either one or both sets of primers, producing amplicons of two sizes which can be visualised via gel electrophoresis (Tuzun et al., 2005).

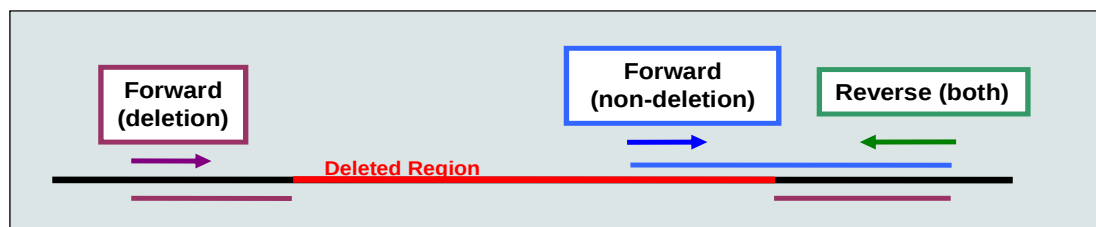


Figure 5.1: Schematic representation of a breakpoint assay. A common reverse primer is amplified in conjunction with a forward primer either within or outside the deleted region, producing amplicons of different sizes.

Where a CNV is more complex, other techniques must be employed. Two methods of assessing relative copy number changes are Quantitative PCR (QPCR) (D'Haene et al., 2010) and Multiplex Amplicon Quantitation (MAQ) (Kumps et al., 2010). Both rely upon amplification of a test sequence within the copy number variable test region and control sequence outside it, followed by a comparative quantitation of test and control amplicons. QPCR methods compare the increasing fluorescence levels from a labelled test and a differentially labelled control amplicon during in real-time during a PCR reaction, whilst MAQ quantifies amplicons via post-PCR fragment analysis.

5.1.3 Experimental aims

Oxford was the designated laboratory for follow-up of T2D signals emerging from the WTCCC CNV scan. My primary objective was to establish and compare robust methods of CNV genotyping, and to deploy them in replication and validation of any CNVs found to be associated with T2D in the WTCCC study.

My experiments aimed to:

- establish, validate and optimise three alternative methods for genotyping CNVs - break-point assay, QPCR and MAQ;
- compare the accuracy and reliability of these three methods using two test CNVs (detailed in section 5.1.4);
- use optimisation and comparison results to inform validation and replication genotyping of T2D-associated CNVs emerging from the WTCCC CNV scan.

5.1.4 Test CNVs

The optimisation and comparison of alternative methods for CNV genotyping requires a “test set” of CNVs for which true copy number is known. Here, two CNVs of differing complexity were used for assay validation and comparison.

A simple deletion (therefore copy number 0/1/2) close to *NEGR1* and associated with BMI (Speliotes et al., 2010; Willer et al., 2009) was used as an ‘easy to cluster’ variant. The ~47kb deletion (chr1:72,537,705-72,585,028) has the advantage of being perfectly tagged in European populations by rs2815752 (Conrad et al., 2010), such that allele A (allele frequency = 0.64) associates with the deletion, whilst allele G (allele frequency = 0.36) is found exclusively with the non-deletion copy number genotype. It was therefore possible to obtain the “true” copy number of 380 control samples (UKBS cohort) by genotyping rs2815752 (figure 5.2). From these results, a set of 30 samples representing a spread of true copy number genotypes (13 homozygous deletion [0 copies], 10 heterozygotes [1 copy] and 7 wild-type [2 copies]) was used for further assay optimisation and validation.

A second, more complex CNV was used as a representative ‘hard to cluster’ variant. Copy number variation at the *FCGR3* region is associated with risk of autoimmune disorders, particularly systemic lupus erythematosus (Willcocks et al., 2008). The region includes a deletion and duplication, meaning that copy number can range from zero to four. However, zero-copy and four-copy genotypes are rare (frequency <0.01) in European populations, meaning that the easily observably copy number classes are one, two and three (Niederer et al., 2010). A 164kb region at the *FCGR3* locus (Chr:159,750,000-159,914,000) was typed in 90 HapMap CEU samples as part of the WTCCC pilot study, and was also typed using MAQ and QPCR methodologies in the

type 1 diabetes replication laboratory located at the Wellcome Trust Sanger Institute. The WTCCC pilot and Sanger results therefore provide two sets of comparative data for experimental validation.

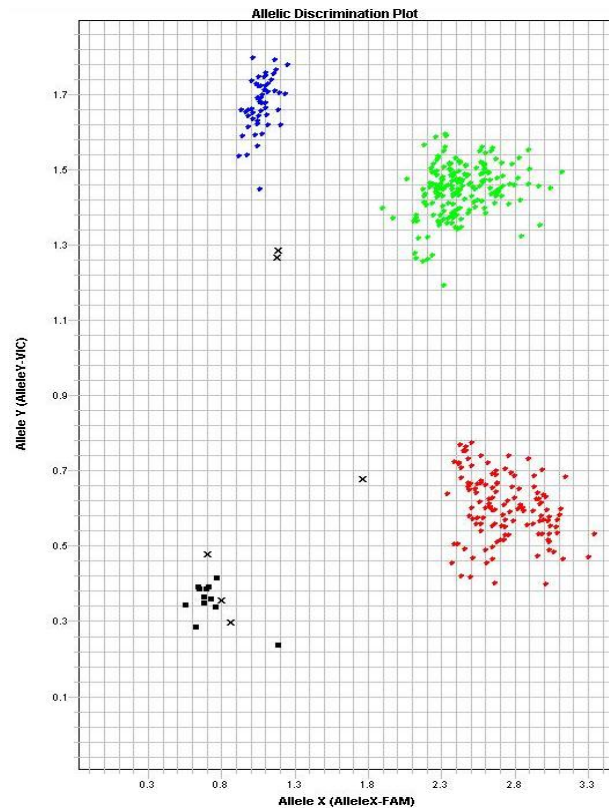


Figure 5.2: Genotyping cluster plot: rs2815752. FAM (A allele, deletion) fluorescence (x-axis) is plotted against VIC (G allele, non-deletion) fluorescence (y-axis). **Red points** represent AA (deletion) homozygous samples, **blue points** represent GG (non-deletion) homozygotes and **green points** represent heterozygotes. Black cross (X) denotes a samples of undetermined genotype, whilst black squares represent non-template controls (dH₂O) which should not amplify.

5.2 Samples and Methods

5.2.1 Samples Used

Assays for the *NEGR1* and *FCGR3* test CNVs were validated using 30 UKBS control and 90 HapMap CEPH samples respectively. Replication genotyping of signals emerging from the WTCCC CNV study was performed in UKBS and 58BC control samples (total 736), and in T2D case samples from the Diabetes UK Warren 2 repository (total 588).

UKBS controls were taken from the UK Blood Service Common Controls (UKBS-CC) collection (ethical approval number 05/Q0106/74). The collection encompasses anonymised DNA samples from 3622 healthy blood donors in England, Scotland and Wales (age 18-69 years), and is currently available to the academic community as a reference resource.

58BC controls were taken from the 1958 birth cohort study (<http://www.b58cgene.sgu.ac.uk/>), which includes all births in England, Wales and Scotland during one week in 1958. Aged 44-45 years, 7,692 cohort members provided blood samples with consent for generation of Epstein-Barr virus (EBV)-transformed cell lines. DNA was previously extracted from cell lines representative of gender and regional populations, and of self-reported European ethnicity.

Diabetes diagnosis (at age 25-75 years) in the Warren 2 repository is based on fasting plasma glucose >7mmol/L, or prescribed treatment with insulin or antidiabetic oral agents. Other forms of diabetes (e.g. maturity-onset diabetes of the young, mitochondrial diabetes and T1D) are excluded by standard clinical criteria.

5.2.2 SNP genotyping

CNV-tagging SNPs were genotyped using inventoried TaqMan genotyping assays (table 5.1), as described in section 2.5.

Variant	SNP	Location	MAF	Assay ID	PCR cycles
NEGR1	rs2815752	Chr1:72812440	0.35	C_26668839	40
CNV1934.1	rs930668	Chr4:62004333	0.03	C_8965824	40

Table 5.1: Genotyping assay details – CNV- tagging SNPs. Locations are b36/hg18. MAF = minor allele frequency (HapMap CEU samples). Assay IDs refer to Applied Biosystems TaqMan genotyping assays.

5.2.3 Breakpoint assays

The test CNV deletion at *NEGR1* was genotyped via a breakpoint assays. A common reverse primer outside (3' end) the deleted region was used in conjunction with a wild-type-specific forward primer (located in the deleted region) or a deletion-specific forward primer (located 5' to the deleted region) (figure 5.3). On a wild-type chromosome, the amplicon produced by the deletion-specific primer would be larger than 48kb, and therefore not amplified in a standard PCR reaction.

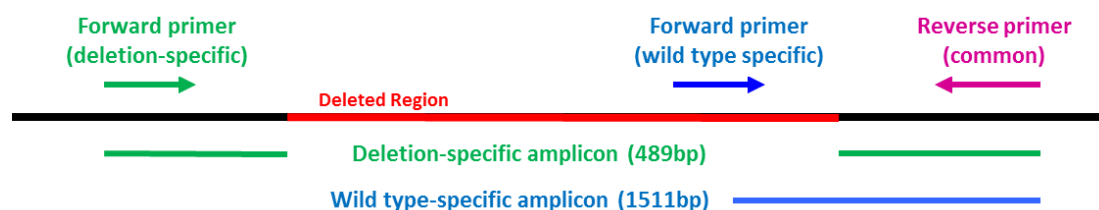


Figure 5.3: Schematic representation of *NEGR1* breakpoint assay. A common reverse primer is amplified in conjunction with a forward primer either within or outside the deleted region, producing amplicons of different sizes.

PCR primers were designed using human genome sequence obtained from the University of California Santa Cruz Genome Bioinformatics website and Primer 3 software, as detailed in section 2.1.1. Primer sequences are provided in table 5.2.

Primer	Primer sequence
Deletion-specific FWD	GAGGAGGTGCCCCATACATA
Wild-type-specific FWD	GGAGTTAGGAGCAACCACCA
Common REV	TTTTCCTCATGAACTCGGAAG

Table 5.2: Primer sequences for *NEGR1* breakpoint assay. FWD = Forward primer, REV = reverse primer.

The *NEGR1* region was amplified using standard Applied Biosystems reagents and thermal cycling parameters (section 2.1.2). Gel electrophoresis was performed on 2% agarose gels (100bp DNA ladder) as detailed in section 2.1.3, with an electrophoresis time of 60min used to improve differentiation between fragment sizes.

Deletion status was determined by the presence or absence of deletion and wild-type-specific amplicons. Deletion genotypes were visually called by myself and independently by a postdoctoral colleague (Dr. Amanda Bennett, OCDEM, University of Oxford). Samples with discrepant calls were excluded from analysis.

5.2.4 Quantitative PCR for copy number analysis

Copy number polymorphisms at the test CNVs *NEGR1* and *FCGR3* were also genotyped using QPCR. QPCR is a standard technique for the quantification of gene expression, which seeks to count or relatively quantify copies of mRNA (reverse transcribed to cDNA). By amplifying gDNA rather than cDNA, the QPCR described here uses a similar technique to quantify the copy number of gDNA sequence.

TaqMan gene expression assays (table 5.3) were designed within the copy number variable sequence using Primer Express software, as detailed in section 2.5.1. Custom assays with FAM-labelled probes were obtained from Applied Biosystems, together with a VIC-labelled two-copy reference gene (*RNaseP*; assay ID 4316844).

Variant		Sequence
<i>NEGR1</i>	FWD primer	GGATAAACTATCCAGGTTGTCATCTCAA
	REV primer	AATATTATGTTTTTTCCTAGGTTGTATATAATTTCTAACT
	Probe	TTGAGCAAGAAAATTG
<i>FCGR3</i>	FWD primer	GCTCAAATCGCTTGGTTAGC
	REV primer	ACCAGAAATTCGGACACAGC
	Probe	TGGGCCTAAAGATGACCTCTCTGGA

Table 5.3: TaqMan assay details for copy number determination

At the time of these experiments, TaqMan assays had not been extensively tested or optimised for copy number analysis. The determination of reaction volumes and components therefore forms part of the assay optimisation and validation reported in results section 5.3.1. After these investigations, the reaction mix used was as detailed in table 5.4.

Reaction component	Volume (μl) in 10μl reaction
DNA	5ng
Gene expression mastermix	5.0
Test assay (FAM labelled)	0.5
RNaseP reference assay (VIC labelled)	0.5
dH ₂ O	4.0

Table 5.4: CNV TaqMan reaction components. For n=1. Reaction mixes scaled up linearly

QPCR reactions were performed in triplicate on an Applied Biosystems 7900 machine as detailed in section 2.7.4. Fluorescence thresholds and C_T values were determined by SDSv2.3 software. Standard curves were generated and assay efficiencies calculated as detailed in section 2.7.3.

Copy number estimates were calculated by comparing C_T values from the test assay (*NEGR1* or *FCGR3*) and the two-copy reference assay (*RNaseP*) for each sample. Replicates were summarised by removing the least consistent replicate (based on distance from the mean), and taking the mean of remaining readings. Normalised signal for each sample was calculated as a ΔC_T value, where ΔC_T is calculated by

$$Ct \text{ test assay} - Ct \text{ reference assay}$$

The resulting values were linearised using the formula:

$$2^{-\Delta C_T}$$

Since the reference assay was a two-copy gene (i.e. a ratio between test and reference of 1:1 would indicate a test assay copy number of two), the resulting comparative values were multiplied by two to produce the final copy number estimate.

5.2.5 Multiplex Amplicon Quantification for copy number analysis

Multiplex MAQ assays were designed and manufactured as a commercial service by Multiplicom (Antwerp, Netherlands), as described in section 2.8.1. Multiplicom were provided with the region of interest and designed 5-7 FAM-labelled amplicons within it (figure 5.4), to multiplex with 7-8 FAM-labelled control amplicons in copy number invariable sequence.

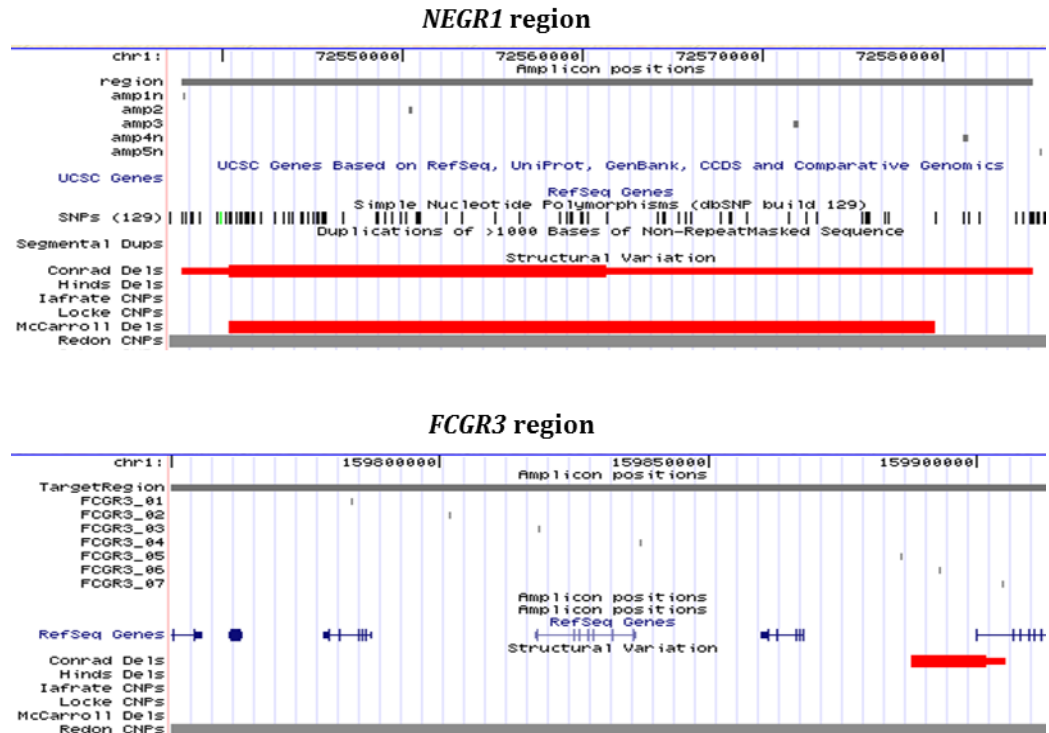


Figure 5.4: Location of MAQ test amplicons. Target region (provided to Multiplicom) is shown as the uppermost grey bar and amplicons designed within it are shown below. Target regions' relationship to previously defined copy number polymorphisms are shown in the custom copy number tracks below. Figured produced using UCSC bioinformatics website (<http://genome.ucsc.edu/>).

Details of amplicons in the *NEGR1* assay are provided in table 5.5. The *FCGR3* assay was provided by collaborators at the Sanger Institute (led by Dr. Kathy Stirrups), whose license agreement with Multiplicom did not permit them to share precise assay details.

Amplicon	Chromosome	Start position	Length (bp)
Test amplicons	1	72537857	105
	2	72550495	225
	3	72572043	339
	4	72581399	263
	5	72585560	120
Reference amplicons	1	126499500	113
	2	14624700	135
	3	119134100	220
	4	38644500	237
	5	68763500	286
	6	234415000	330

Table 5.5: *NEGR1* MAQ assay details. Chromosomal locations based on human genome build 36 (May 2006).

PCR amplification was performed with 50ng DNA using Multiplicom reagents as described in section 2.1.2. 2µl PCR product underwent fragment analysis on an AB3130xl machine, as detailed in section 2.9.

Analysis was conducted using MAQs software (Multiplicom), which quantifies fragment peak areas at sizes defined according to the amplicons included in the assay. Each test amplicon is compared to each control amplicon, producing a set of ratios. These ratios are compared to their equivalents in the reference sample, producing dosage quotients (DQs) according to the formula:

$$\frac{(Control\ sample\ target\ amplicon)}{(Control\ sample\ reference\ amplicon)} \div \frac{(Test\ sample\ target\ amplicon)}{(Test\ sample\ reference\ amplicon)}$$

The set of DQs for each test amplicon is used to produce a final DQ mean and SD.

Although true copy number was known for many samples in the experiments described below, CNV genotyping tasks may have no such positive controls. The mean of all samples was therefore used as an initial reference, which produced three clear copy number groups for both the *NEGR1* and *FCGR3* assays. On the basis of known possible copy number genotypes (0/1/2 for *NEGR1* and 1/2/3 for *FCGR3*), a mean of three randomly selected two-copy samples was used as the reference in further analyses. Final DQ results were therefore multiplied by two to produce a copy number estimate from the test/reference comparison.

5.3 Results

5.3.1 Assay optimisation and validation

5.3.1 (a) Breakpoint assays

Optimisation for a breakpoint assay is simply optimisation of a PCR and gel electrophoresis. A break-point assay designed over the *NEGR1* deletion produced two distinct band sizes according to deletion status (~500bp and ~1500bp for wild-type and deleted state respectively), shown in figure 5.5.

In a simple breakpoint assay, it is not possible to be certain whether an absent band represents true absence of the copy number genotype which would produce that amplicon, or reflects a failed reaction. Duplexing deletion and wild-type primer sets rather than amplifying them in separate reactions would resolve this uncertainty, as well as reducing reagent and DNA consumption. However, the two *NEGR1* amplicons could not be successfully duplexed at a range of $MgCl_2$ concentrations (0.5 – 4mM per reaction) and annealing temperatures (51-73°C). At all conditions, duplexing resulted in loss of the larger amplicon, presumably due to preferential amplification of shorter sequence (figure 5.5).

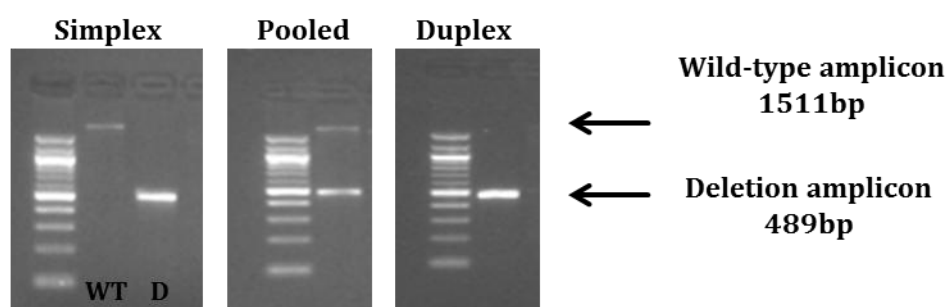


Figure 5.5: Duplexing test (*NEGR1* deletion heterozygote sample). Left: heterozygote sample with deletion and non-deletion primer sets amplified and run separately (WT=wild-type lane, D=deletion lane). Centre: Simplex products as above but pooled for gel electrophoresis. Right: duplexed PCR showing loss of larger band. Samples run with 100bp DNA ladder.

Both *NEGR1* primer sets were also duplexed with six alternative primer sets - previously shown to be reliable and with sufficiently different amplicon sizes to be easily resolvable by gel electrophoresis - to act as a positive control for reaction success. However, these also failed at a range of MgCl₂ concentrations and annealing temperatures. All further experiments were therefore performed in simplex, with the chance of reaction failure in one amplicon minimised by amplifying both primer sets on the same plate and using the same reagent batches. Products from the two simplex reactions were pooled for gel electrophoresis. The certain *NEGR1* deletion status of genotyped samples provides an indication of whether biased amplicon failure presents a problem for the breakpoint assay.

5.3.1 (b) Quantitative PCR

To determine the optimum reaction volume for QPCR analysis of copy number, three alternative volumes were compared (table 5.6). Manufacturer's guidelines recommend 20µl, but previous experience in my laboratory with QPCR for mRNA expression analysis suggests that lower volumes may be more reliable – perhaps because they allow faster temperature change during thermal cycling. Reduced reaction volumes also have the advantage of consuming less reagent and sample DNA.

Reaction component	Volume (µl) in 20µl reaction	Volume (µl) in 10µl reaction	Volume (µl) in 5µl reaction
DNA	10ng	5ng	2.5ng
Gene expression mastermix	10.0	5.0	2.5
Assay	1.0	0.5	0.25
dH ₂ O	9.0	4.5	2.25

Table 5.6: CNV QPCR alternative volume reaction components. For n=1. Reaction mixes scaled up linearly.

These three volumes were used to amplify a pool of 10 test DNA samples, serially diluted to produce a standard curve and allow calculation of reaction efficiency. Table 5.7 shows the correlation coefficient (r^2) for the standard curve, the assay efficiency based on the slope of the standard curve, and the variability between sample replicates at the experimental DNA input (centre point of standard curve).

Reaction Volume	Coefficient of variance			Standard Curve r^2			Efficiency		
	<i>RNaseP</i>	<i>NEGR1</i>	<i>FCGR3</i>	<i>RNaseP</i>	<i>NEGR1</i>	<i>FCGR3</i>	<i>RNaseP</i>	<i>NEGR1</i>	<i>FCGR3</i>
20 μ l	0.05	0.58	0.21	0.99	0.98	0.89	0.99	0.95	0.98
10 μ l	0.13	0.40	0.23	0.99	0.92	0.92	0.98	0.96	0.97
5 μ l	0.31	0.46	1.17	0.97	0.82	0.64	0.72	0.64	0.51

Table 5.7: CNV QPCR alternative volume reaction comparison. Coefficient of variance given for replicates at input DNAs of 10ng, 5ng and 2.5ng (for 20 μ l, 10 μ l and 5 μ l reaction volumes respectively). Reaction efficiency calculated from standard curves. Most advantageous values are highlighted in green; least advantageous values in red.

A volume of 5 μ l clearly impairs reaction efficiency and reliability. However, the difference between 10 μ l and 20 μ l reaction volumes is minimal, and in some cases 10 μ l reactions appear slightly more efficient and reliable. Further experiments were therefore performed using a reaction volume of 10 μ l, since this is almost indistinguishable from the recommended 20 μ l – and more reagent and DNA-efficient.

Just as duplexing would be advantageous (but was unsuccessful) for breakpoint assays, it would be beneficial to duplex target (*NEGR1* or *FCGR3*, FAM-labelled) and reference (*RNaseP*, VIC-labelled) assays in a QPCR reaction. As well as reducing reagent consumption, duplexing would remove the possibility of variability in input DNA concentration between target and reference wells. However, in order for duplexing to be valid, it must have no effect upon assay efficiency.

Using standard curve analysis of a sample pool as described above, figure 5.6 compares the amplification of *RNAseP*, *NEGR1* and *FCGR3* assays in simplex and in duplex with one another across a titration of DNA concentrations (*NEGR1* and *FCGR3* in duplex with *RNAseP*, data shown for *RNAseP* in duplex with *NEGR1*). In all cases, the difference between the gradients of the simplex and duplex slopes is <0.1 , demonstrating that duplexing has minimal effect upon amplification efficiency. Further QPCR copy number analysis was therefore conducted with duplexed target and reference amplicons.

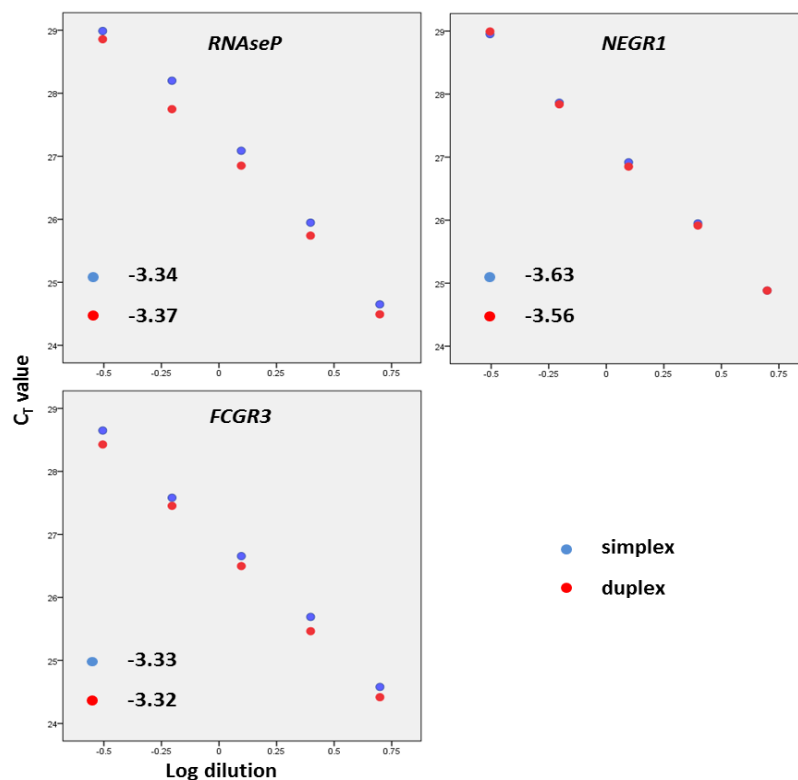


Figure 5.6: QPCR duplexing test. Simplex results in blue, duplex results in red. log DNA concentration (x-axis) is plotted against threshold fluorescence cycle number (y-axis). Values are the gradients of the simplex and duplex slopes respectively. Duplexing has no significant effect upon assay efficiency as revealed by standard curve analysis.

5.3.1 (c) Multiplex Amplicon Quantitation

Experimental optimisation for the *NEGR1* and *FCGR3* MAQ assays was conducted by Multiplicom - the company which markets, designs and produces MAQ as a commercial service. Optimum reaction mixes and cycling conditions were provided with the custom primer mix.

Figure 5.7 shows fragment analysis data from an *NEGR1* test sample, analysed in Multiplicom's MAQs software. The peak quantification regions (highlighted in grey for control amplicons and yellow for test amplicons) are pre-set in the analysis software. This plot provides visual confirmation that the quantification regions are located appropriately and amplicons are of the expected size.

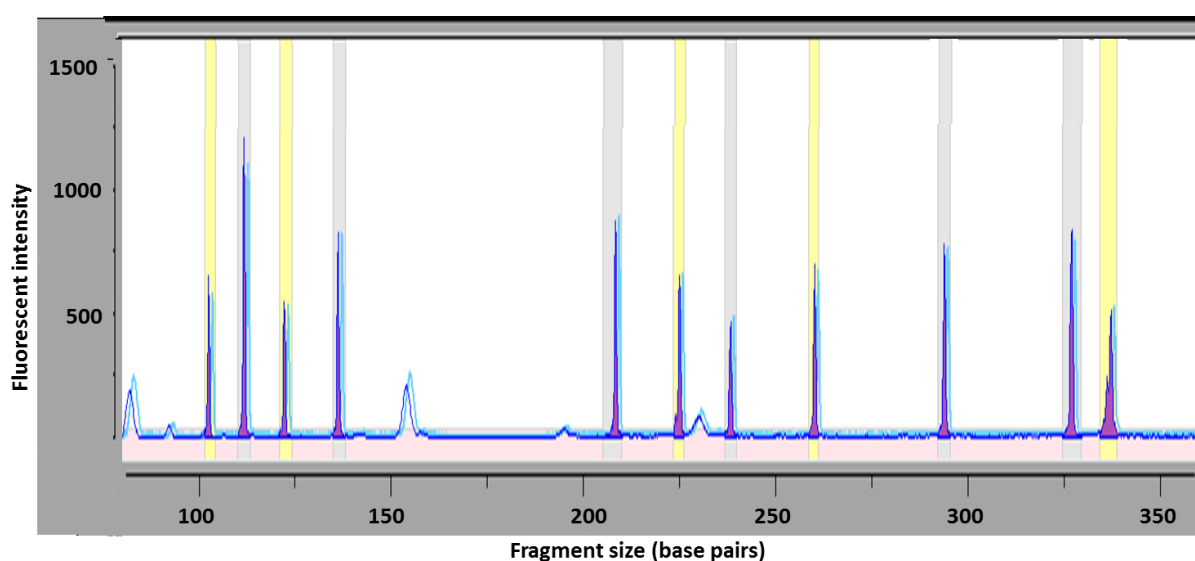


Figure 5.7: Fragment analysis peaks for MAQ analysis (single sample). Fragment size (x-axis) is plotted against fluorescent intensity (y-axis). Peaks regions marked in grey are control amplicons; those marked in yellow are test amplicons. Plot produced in MAQs software.

DQ results from three test DNAs of *NEGR1* deletion copy number zero, one and two (based on tag SNP genotype) are shown in figure 5.8. There are clear differences in DQ between the samples at test amplicons two, three and four, but test amplicons one and

five appear invariant. Given their locations at either end of the proposed copy number variable region (figure 5.4), and the imperfectly defined breakpoints of many CNVs, it is possible that amplicons one and five may lie outside the deleted region. Further analysis was therefore conducted using the mean DQ from amplicons two, three and four only.

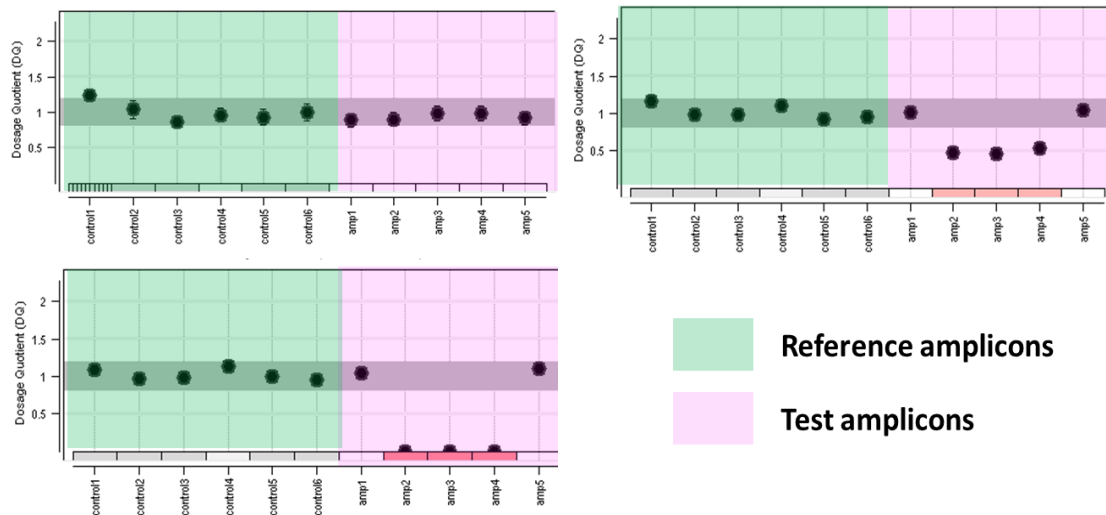


Figure 5.8: Dosage Quotients of amplicons in a MAQ experiment. Each plot represents a single sample and each point an amplicon: left-hand six amplicons (in green) are controls, right-hand five (in pink) are test amplicons located in the copy number variable region. Grey band marks dosage quotient ~ 1 (i.e. copy number = 2, assuming comparison to a 2-copy reference). Top left: sample with no deletion (two copies for test amplicons). Bottom right: sample with heterozygous deletion (one copy for test amplicons). Bottom left: sample with homozygous deletion (0 copies for test amplicons). Note the invariant results for test amplicons 1 and 5, suggesting location outside the variable region.

Similar validation of the *FCGR3* assay performed by collaborators at the Sanger Institute resulted in the exclusion of one out of seven test amplicons from further analysis.

5.3.2 Method comparison

5.3.2 (a) *NEGR1*

All three CNV genotyping methods were broadly successful in assigning copy number classes at the *NEGR1* polymorphism.

A break-point assay was able to determine ‘true’ (as defined by tagging SNP genotype) copy number according to observed amplicon sizes with perfect accuracy for the 25 samples which passed QC. All wild samples produced the expected amplicon at ~1500bp, all homozygous deletion samples produced the expected amplicon at ~500bp, and all heterozygous samples amplified both products (figure 5.9). However, five samples (17%) failed analysis because their genotype calls were uncertain or discrepant between duplicate genotype calling.

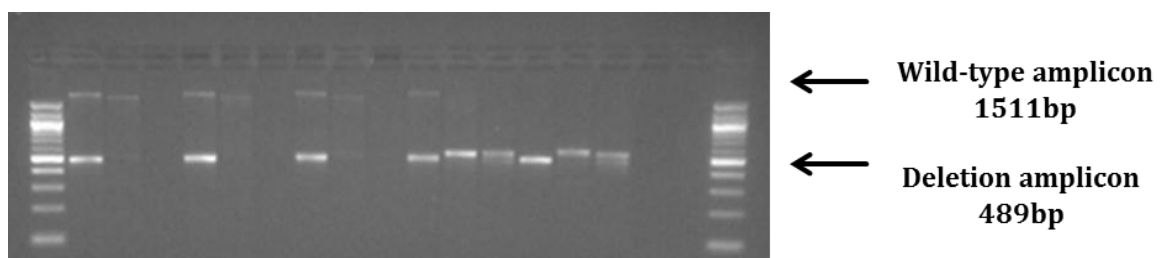


Figure 5.9: Example *NEGR1* deletion breakpoint gel. Visible bands reflect amplicon sizes as expected according to tag-SNP copy number. Gel run with 100bp DNA ladder.

Figure 5.10 compares *NEGR1* copy number estimates from QPCR and MAQ analysis to genotyping results from the CNV-tagging SNP rs2815752, with samples coloured by genotype call as an indicator of ‘true’ copy number. There is a clear positive relationship between genotyped and estimated copy number in both cases, with some indication that MAQ results show less variation within copy number classes (less

variability on the y-axis) and better discrimination between copy number classes (more distance between groups on the y-axis).

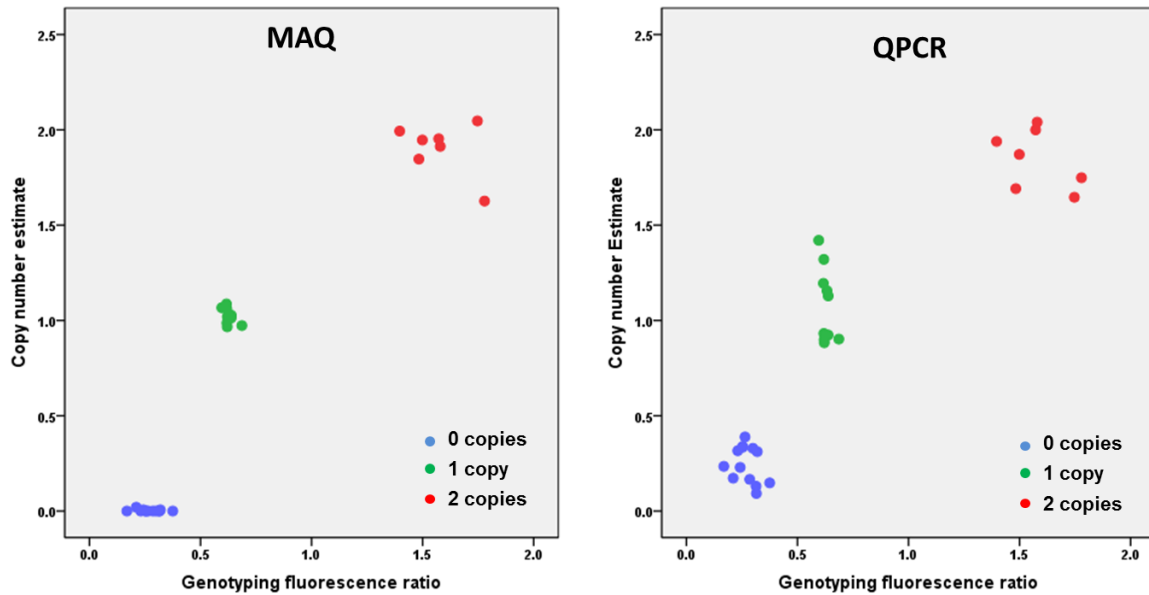


Figure 5.10: Comparison of MAQ and QPCR copy number estimates for *NEGR1* deletion. VIC/FAM genotyping fluorescence ratio (x-axis) is plotted against copy number data from MAQ (left) or QPCR (right) experiments (y-axis). Points are coloured by copy number according to tagging SNP genotype.

This observation is largely borne out by a numerical comparison of the copy number estimate range within each copy number class (table 5.8).

"True" copy number	N samples	QPCR results mean (range)	MAQ results mean (range)
0	13	0.25 (0.09 - 0.39)	0.00 (0.00 - 0.02)
1	10	1.08 (0.88 – 1.42)	1.02 (0.97 – 1.07)
2	7	1.85 (1.65 – 2.04)	1.90 (1.63 – 2.05)

Table 5.8: Mean and range of *NEGR1* copy number estimates within each copy number class for QPCR and MAQ.

The means for MAQ results are closer to the true copy number for all three classes. The range within copy number class “0” is 15 times larger for QPCR data than MAQ data, and five times larger for copy number class “1”. For copy number class “2”, the range for MAQ data is slightly (8%) larger than the range for QPCR data. None of the ranges overlap between copy number classes.

Replication genotyping of a CNV would typically require copy number to be assigned on the basis of QPCR or MAQ data alone. The extent to which data sets divide into clear copy number clusters can be assessed visually, by plotting a histogram of copy number estimates (figure 5.11).

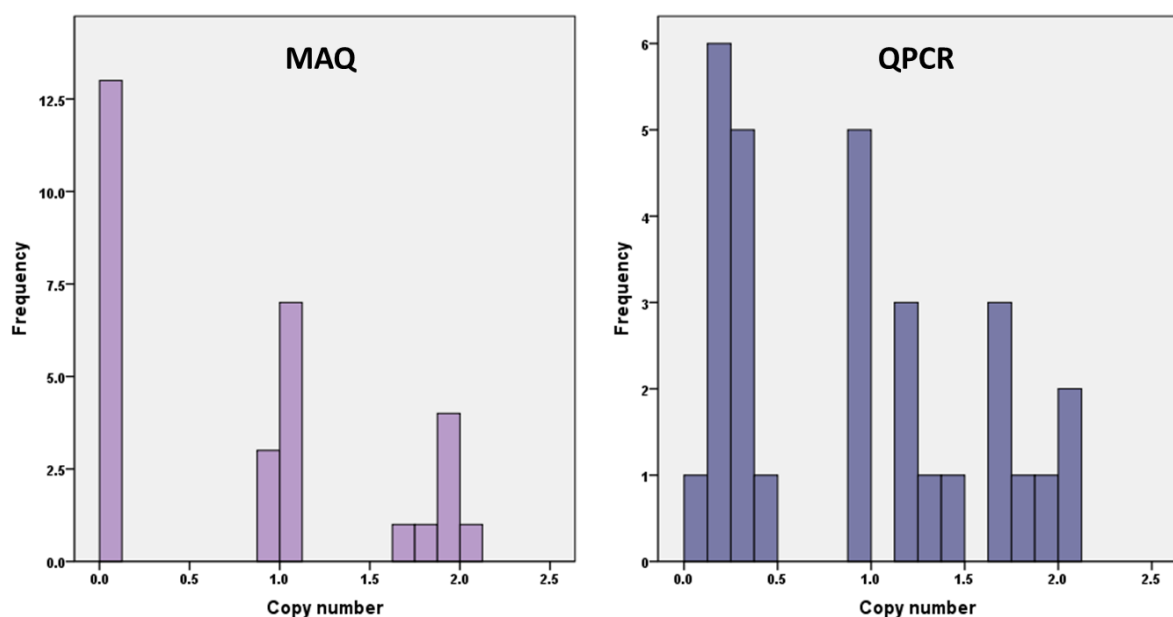


Figure 5.11: Histogram of MAQ and QPCR copy number estimates for *NEGR1* deletion. Histograms are specified to divide into 20 bins.

The MAQ data set separates into three clear groups. QPCR provides clear separation between the 0-copy and 1-copy clusters, but the boundary between 1-copy and 2-copy clusters is less evident.

Assignment of copy number can be achieved computationally through cluster analysis. With the number of clusters fixed at three, k-means cluster analysis assigned samples to copy number classes as shown in figure 5.12. For both MAQ and QPCR data, all samples were correctly and exclusively assigned to copy number classes. The calculated cluster centres were closer to true copy number for MAQ data (0.0, 1.0 and 1.9) than for QPCR data (0.2, 1.1 and 1.8), but both techniques were 100% accurate in assigning samples to copy number classes.

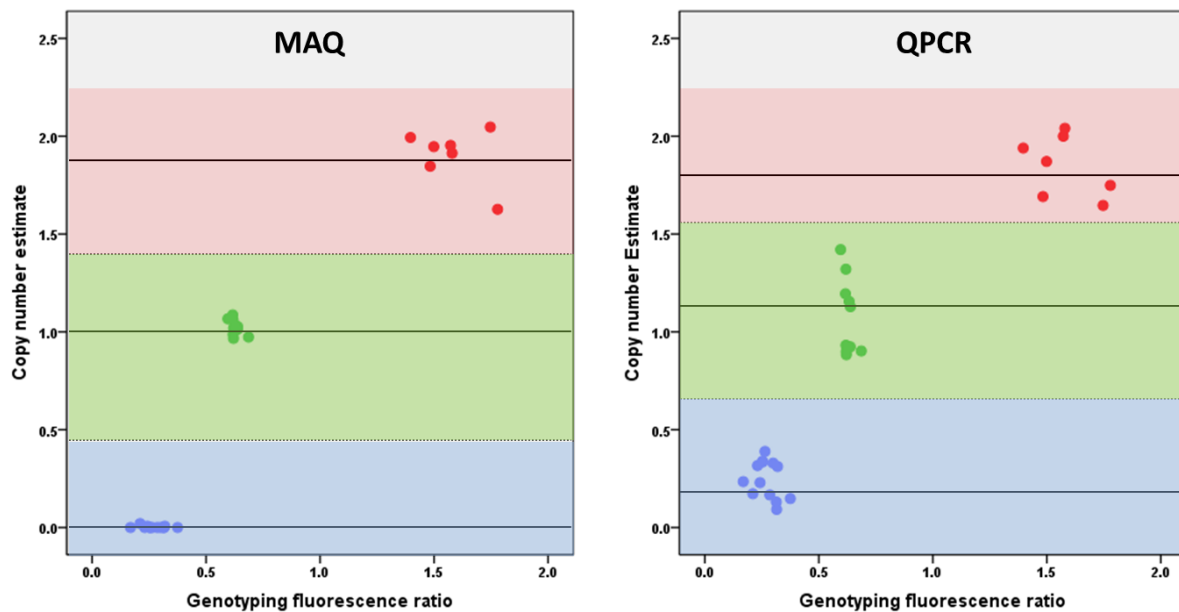


Figure 5.12: Cluster analysis for *NEGR1* deletion using MAQ and QPCR copy number estimates. VIC/FAM genotyping fluorescence ratio (x-axis) is plotted against copy number data from MAQ (left) or QPCR (right) experiments (y-axis). Points are coloured by copy number according to tagging SNP genotype. Solid lines represent cluster centres according to k-means (3 clusters) cluster analysis, dashed lines represent cluster boundaries. Background colours denote the region assigned to the copy number cluster according to cluster analysis. In both these plots, all samples fall within their appropriate colours and are therefore correctly assigned to a copy number class.

5.3.2 (b) *FCGR3*

Clear differences between MAQ and QPCR emerge for the *FCGR3* polymorphism (*FCGR3* is not a simple deletion polymorphism, so could not be typed through a breakpoint assay). Figure 5.13 compares *FCGR3* copy number estimates from QPCR and MAQ analysis with copy number data produced via array CGH as part of the WTCCC CNV pilot study. Samples are coloured by pilot copy number call as an indicator of 'true' copy number. There is a positive relationship between pilot and estimated copy number in both cases, but within-class variation (y-axis spread) is far greater for the QPCR data, and separation between copy number classes is much less evident.

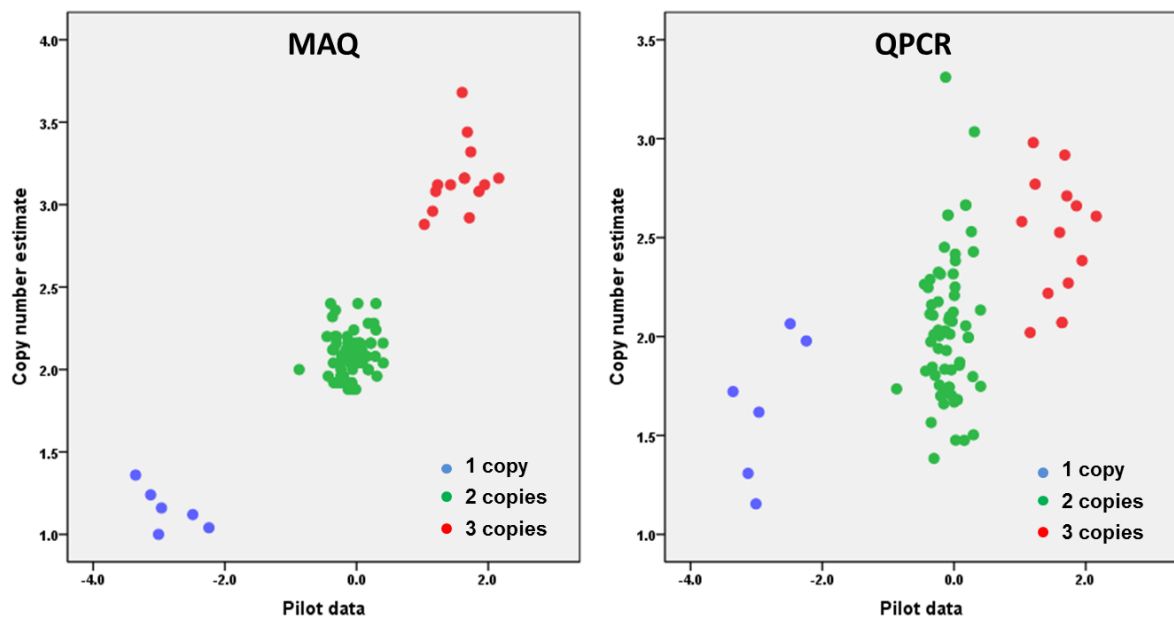


Figure 5.13: Comparison of MAQ and QPCR copy number estimates for *FCGR3* copy number polymorphism. WTCCC pilot data (x-axis) is plotted against copy number data from MAQ (left) or QPCR (right) experiments (y-axis). Points are coloured by copy number according to WTCC pilot calls.

This difference can be demonstrated through the range of data in each copy number class (table 5.9).

"True" copy number	N samples	QPCR results mean (range)	MAQ results mean (range)
1	6	1.64 (1.15 – 2.06)	1.15 (1.00 – 1.36)
2	71	2.04 (1.38 – 3.31)	2.09 (1.88 – 2.40)
3	13	2.48 (2.02 – 2.98)	3.16 (2.88 – 3.68)

Table 5.9: Mean and range of *FCGR3* copy number estimates within each copy number class for QPCR and MAQ.

Although all ranges are larger than those for the *NEGR1* deletion, the range of values in each copy number class is greater for QPCR data than for MAQ data – 2.5 times, 3.7 times and 1.2 times for classes "1", "2" and "3" respectively. Furthermore, there is no overlap between the copy number class ranges for MAQ data, but the range of QPCR data overlaps between classes "1" and "2" and between classes "2" and "3".

Histograms (figure 5.14) show a clear separation into three groups for the MAQ data set, but no visually discernible groups for QPCR data.

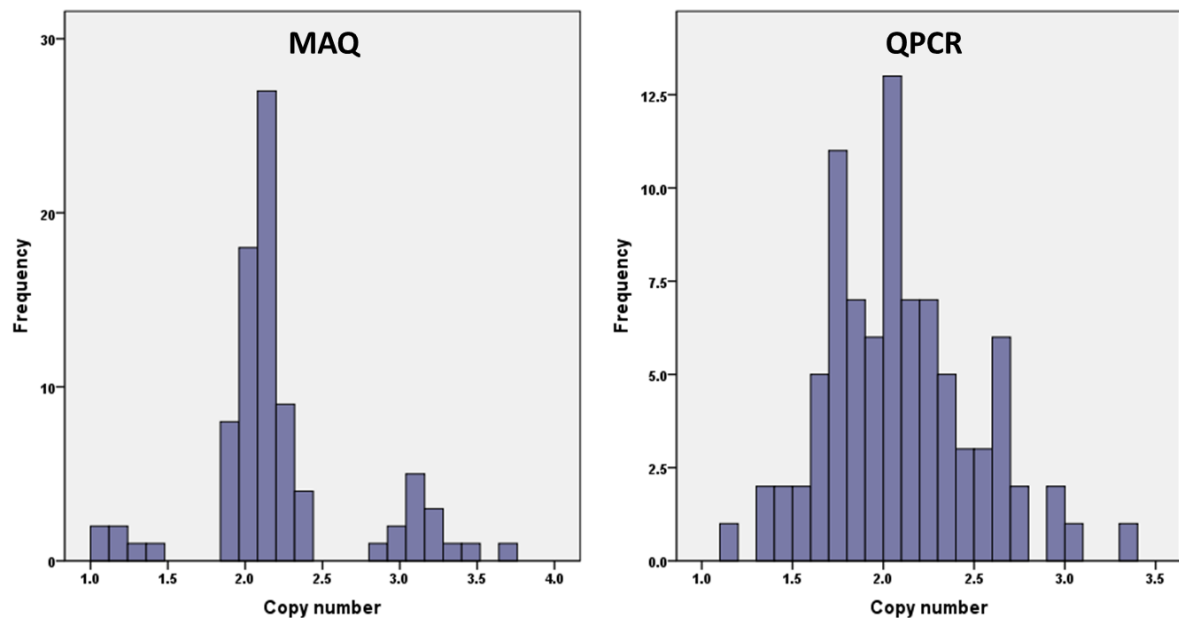


Figure 5.14: Histogram of MAQ and QPCR copy number estimates for *FCGR3* deletion. Histograms are specified to divide into 25 bins.

Computational differentiation of groups (k-means cluster analysis with cluster number set at 3) drew similar conclusions (figure 5.15).

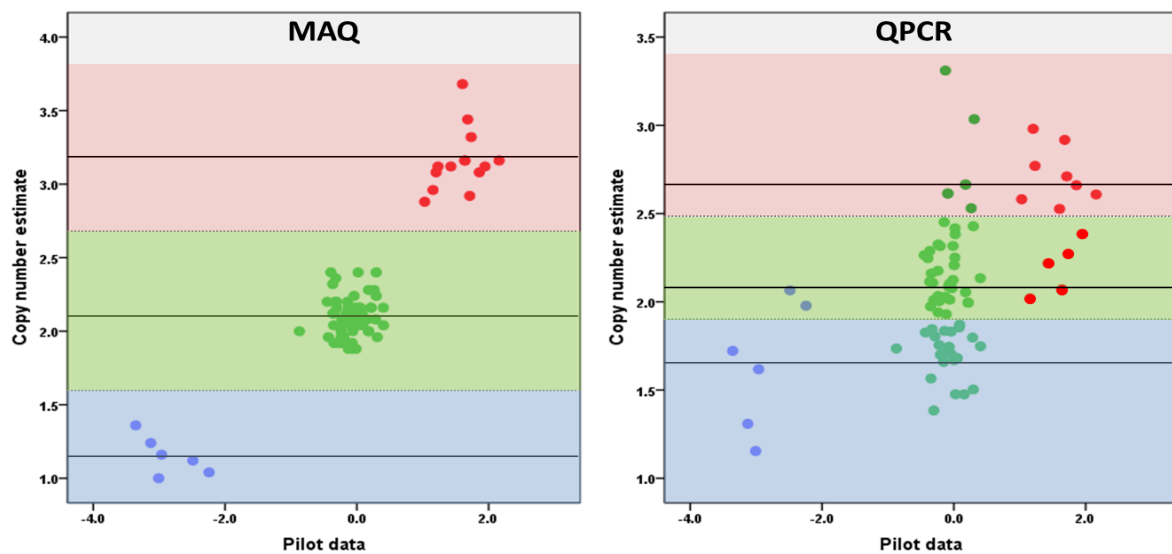


Figure 5.15: Cluster analysis for *NEGR1* deletion using MAQ and QPCR copy number estimates. WTCCC pilot data (x-axis) is plotted against copy number data from MAQ (left) or QPCR (right) experiments (y-axis). Points are coloured by copy number according to WTCC pilot calls. Solid lines represent cluster centres according to k-means (3 clusters) cluster analysis, dashed lines represent cluster boundaries. Background colours denote the region assigned to the copy number cluster according to cluster analysis.

On the basis of MAQ data, all samples were correctly assigned, with calculated cluster centres of 1.2, 2.1 and 3.2. For QPCR data however, 49% of samples were incorrectly assigned to copy number groups and calculated cluster centres were much less accurate at 1.7, 2.1 and 2.7.

Summary of method comparison

All three methods were broadly successful in genotyping the simple deletion at *NEGR1*. MAQ and QPCR both called 100% of samples correctly, whilst all samples which passed breakpoint assay QC (83%) were accurately genotyped. MAQ was more successful than QPCR in genotyping the complex *FCGR3* polymorphism, with 100% compared to 49% accuracy. On the basis of these data, MAQ looks to be the method of choice for genotyping CNVs – particularly for more complex regions.

5.3.3 Follow-up of T2D-associated CNVs arising from WTCCC CNV scan

The WTCCC genome-wide association scan of copy number variation (Craddock et al., 2010) genotyped 10,894 CNVs in 3,000 control samples and ~2000 case samples from each of eight diseases. Genotype calls were polymorphic and passed quality control assessment at 3,432 loci, demonstrating the challenge of accurately determining copy number variation. The majority (88%) of successfully genotyped CNVs showed two or three copy number classes, consistent with the presence of two alleles in the population. Approximately half (56%) were common (MAF > 0.05), and represented 42-50% of all large (>0.5kb) and common copy number variation in populations of European ancestry.

The most striking associations identified were between CNVs in the human leukocyte antigen (HLA) region and autoimmune disorders. Associations with T1D, rheumatoid arthritis and Crohn's disease reached significance values of 8×10^{-153} , 1.4×10^{-39} and 1.2×10^{-5} respectively, conferring disease odds ratios of 0.20, 1.22 and 0.80.

Association analysis with T2D risk was conducted using 2105 case samples and 3,000 shared controls. The strongest disease associations were observed with CNVs 1943.1 and 5583.1 – deletions on chromosome 4 and 12 respectively (figure 5.16 and table 5.10).

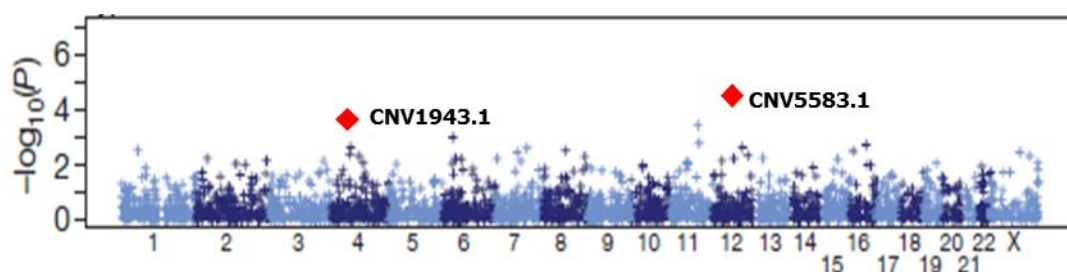


Figure 5.16: Manhattan plot of WTCCC CNV T2D association data, highlighting investigated CNVs. Chromosome location (x-axis) is plotted against strength of association (y-axis), each point representing a genotyped CNV. Figure taken from Craddock et. al., 2010.

Variant	Start location	Size	Type	MAF cases	MAF controls	p-value
CNV1943.1	Chr4:61,681,667	1.4kb	deletion	0.029	0.020	5.68×10^{-4}
CNV5583.1	Chr12:69,818,942	1.0kb	deletion	0.36	0.40	3.9×10^{-5}

Table 5.10: CNVs associated with T2D risk

Neither variant reaches the 5×10^{-8} association threshold typically used for genome-wide significance in SNP association studies - based on correcting a p-value of 0.05 for one million tests. However, the WTCCC CNV study tested fewer than 11,000 CNVs and identified fewer than 4,000 polymorphic sites, so requires a less stringent threshold for significance. Nevertheless, CNVs 1943.1 and 5583.1 did not deviate significantly from the expected distribution of association significance (figure 5.17).

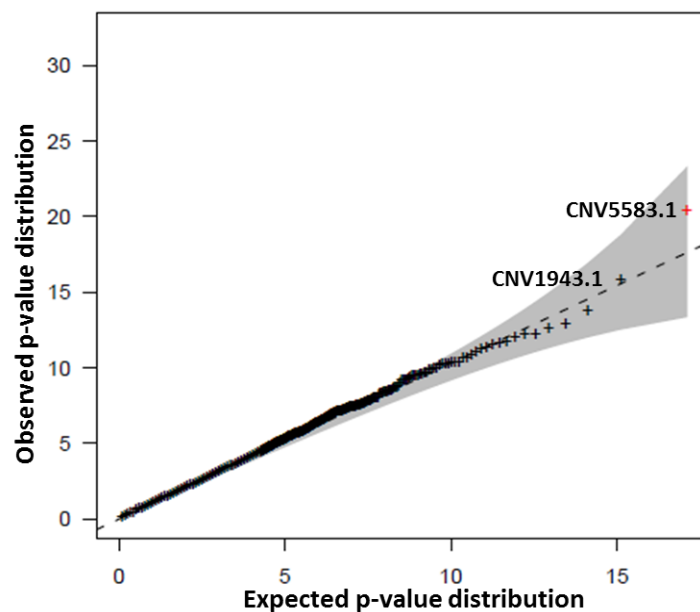


Figure 5.17: Q-Q plot showing CNV association data for T2D. Hatched line represents expected correlation between observed and expected p-values under the null hypothesis of no significant disease association, shaded area represents 95% confidence interval for that correlation Figure taken from Craddock et. al., 2010.

To reduce the possibility of missing genuine associations, a relaxed threshold for follow-up was set at around 1×10^{-4} . CNV1943.1 and CNV5583.1 were therefore taken forward for replication.

5.3.3 (a) CNV1943.1

CNV1943.1 is a ~1.4kb intronic deletion (Chr4:61681667-61683045), identified in the WTCCC discovery sample set (Craddock et al., 2010) (figure 5.18). The closest annotated gene (~70kb away) is latrophilin 3 (*LPHN3*), a neuron-specific G-protein coupled receptor in which SNP variation is associated with susceptibility to Attention Deficit Hyperactivity Disorder (Arcos-Burgos et al., 2010; Silva and Ushkaryov, 2010). The WTCCC CNV scan reports deletion allele frequencies of 0.020 and 0.029 in control and T2D cohorts respectively, providing a moderate p-value for association with T2D of 5.68×10^{-4} .

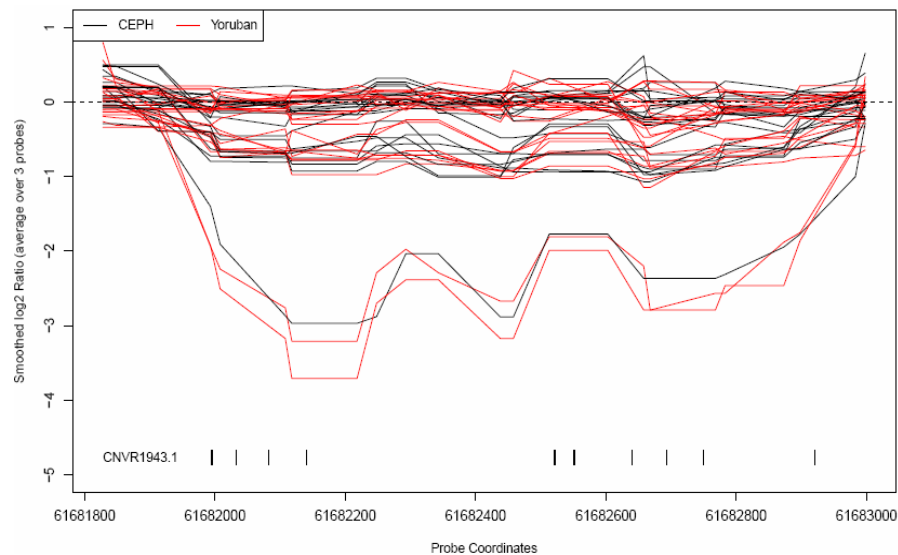


Figure 5.18: CNV1943.1 discovery. Chromosomal position (x-axis) is plotted against normalised probe intensity (y-axis). Each line represents probe intensity data from an individual sample; a region of reduced probe intensity, indicating a deletion, is evident in three samples. Figure taken from (Craddock et al., 2010).

A comparison of WTCCC CNV calls for 1200 United Kingdom Blood Service (UKBS) control samples with SNP data for those same samples (conducted by Dr. Richard Pearson, WTCHG, University of Oxford; personal communication) suggested that rs930668 is a perfectly tagging SNP ($r^2=1$) for CNV1943.1. To confirm the accuracy of

this tagging SNP analysis, I genotyped both rs930668 in an independent sample set comprising 340 samples from the 58BC control cohort (table 5.11).

Genotyped (n)	Genotyping success	GG (n)	AG (n)	AA (n)	MAF
340	95%	307	15	0	0.02

Table 5.11: rs930668 (CNV1943.1 tagging SNP) genotyping results

According to the tagging SNP analysis conducted by Dr. Pearson, chromosomes which carry the A allele at rs930668 always carry the CNV1943.1 deletion, whilst chromosomes which carry the G allele are always wild-type. To confirm this statistical analysis experimentally, I designed a breakpoint assay for CNV1943.1 and genotyped all 15 AG heterozygotes and 15 randomly selected GG homozygotes for the deletion.

CNV1943.1 is located within a region of high sequence homology to multiple locations in the human genome. Any primers located within the deletion region (and in a ~1kb) region 3' of it) were highly non-specific and did not amplify successfully. A single primer set located outside the homologous sequence region was therefore designed to flank the deletion, such that the resulting amplicon would encompass the deleted region (figure 5.19).



Figure 5.19: Schematic representation of CNV1943.1 breakpoint assay. Common forward and reverse primers located on either side of the deleted region, producing amplicons of different sizes according to deletion status.

Primer sequences are provided in table 5.12. The CNV1943.1 region was amplified using PFU polymerase reagents and thermal cycling parameters (section 2.1.2), which are more suited to long amplicons. To aid differentiation of large fragment sizes, gel electrophoresis was conducted for 90min at 110V in 1% agarose gels, with 1kb DNA ladder.

Primer	Sequence
Common FWD	GCAGGGCAATCTGACTTTTC
Common REV	GGGGCCTGCTGTGACTATAA

Table 5.12: Primers for CNV1943.1 breakpoint assay

With the exception of one sample which failed PCR, there was perfect concordance between deletion genotype according to breakpoint assay and deletion genotype inferred from rs930668 genotype. All GG homozygotes produced one amplicon of the size predicted for wild-type (non-deletion) chromosomes, visible as the larger (uppermost) band in figure 5.20. All AG heterozygotes produced one amplicon of a smaller size, visible as the lower band in figure 5.20.

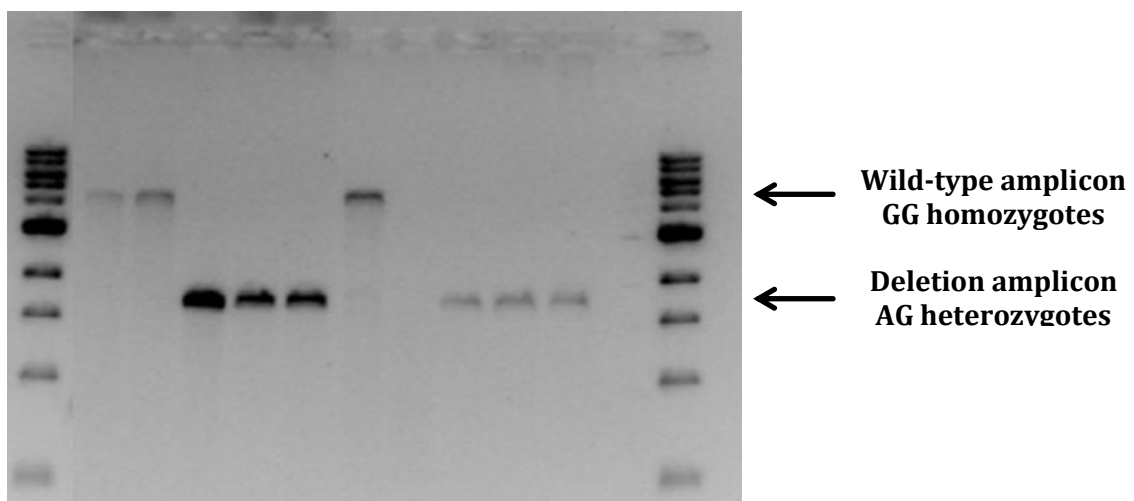


Figure 5.20: CNV1943.1 breakpoint assay gel electrophoresis. Gel image is shown in negative and with a 1kb DNA ladder (bands at 0.5kb, 1kb 1.5kb, 2kb and 1kb intervals thereafter (heavy band at 3kb). Wells 7 and 11 (from left) are non-template controls.

CNV1943.1 was genotyped by a single primer set, flanking the deleted region such that wild-type and deleted chromosomes produce amplicons of differing sizes. It might therefore be expected that heterozygous samples should produce both amplicons simultaneously. In reality however, heterozygous samples produced only the smaller (deletion) band, probably due to preferential amplification of smaller over larger product. In this breakpoint assay, therefore, it is not possible to distinguish between heterozygous samples and deletion homozygotes – both would produce a single amplicon of the smaller size. Although the rarity of the CNV1943.1 deletion means that there were no homozygous deletion samples in the test cohort (and ~2500 samples would be needed to find one), this is a weakness of the assay in principle.

Replication of CNV1943.1 via SNP analysis

Confirmation of rs960668 as a perfectly tagging SNP for CNV1943.1 means that it is unnecessary to directly genotype CNV1943.1 in a replication cohort. Instead, its association with T2D can be explored indirectly through rs930668 in existing SNP-based GWAS (table 5.13).

Study	Participant n (cases/controls)	rs960668 p-value (to 3 s.f)
WTCCC CNV (Craddock et al., 2010)	2,005 / 13,045	0.000568
WTCCC1 (SNP) (WTCCC, 2007)	1,924 / 2,938	0.0149
DIAGRAM* (Voight et al., 2010)	7,192 / 36,028	0.0245

Table 5.13: rs930668 T2D association in various genome-wide association studies. P-values quoted (to three significant figures) are for association with T2D. *The original DIAGRAM meta-analysis includes WTCCC data as one of its eight constituent analyses. The data reported here is from a meta-analysis of the other seven cohorts, excluding the WTCCC.

The WTCCC CNV and WTCCC1 SNP analyses used largely overlapping (by 1711 samples) case samples. The stronger p-value in the CNV analysis is explicable through the addition of shared controls comprising individuals from other disease cohorts included in the study. However, the DIAGRAM meta-analysis includes data from seven independent sample sets (results reported in table 5.13 exclude WTCCC1 data as the eighth cohort), making its effective sample size much larger and results more accurate. The weakened evidence for association of rs930668 with T2D in DIAGRAM therefore suggests that the moderate association CNV1943.1 with T2D was spurious and does not warrant further investigation.

5.3.3 (b) CNV5583.1

CNV5583.1 is a ~1kb deletion (Chr12:69818942-69819942), likely to disrupt an exon of tetraspanin 8 (*TSPAN8*). SNP variation at the *TSPAN8* locus (lead SNP rs4760790) has been previously associated with T2D (Zeggini et al., 2008), raising the possibility that the CNV might be responsible for disease association at regional linked SNPs.

A collaborator at the Wellcome Trust Centre for Human Genetics (Dr. Zamin Iqbal, WTCHG, University of Oxford) identified a 960bp inversion/deletion at a similar location (Chr12:69819116–69820076) through novel assembly of high-throughput sequencing data in a single HapMap CEPH sample (NA12878, figure 5.21).

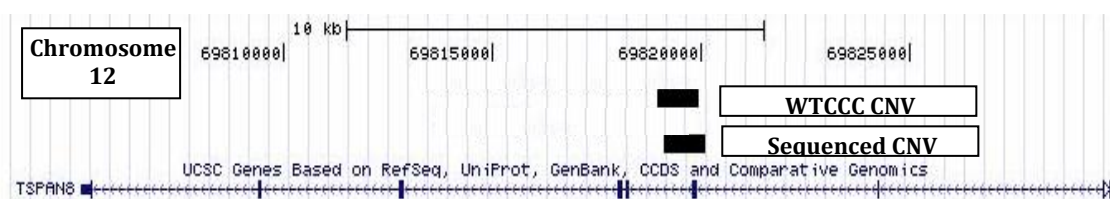


Figure 5.21: CNV locations at *TSPAN8*. WTCCC CNV location data from Craddock et al (2010). Sequenced CNV location data from Zamin Iqbal (Personal Communication). Locations plotted using UCSC custom tracks tool.

To ascertain whether this was the same CNV as typed in the WTCCC, I designed a breakpoint assay to genotype it in a random selection of ~500 cases and ~700 controls included in the WTCCC analysis. It was not possible to design a common deletion and wild-type primer with a suitable annealing temperatures and no self-complementarity; instead, the region was amplified using separate deletion and wild-type primer sets (figure 5.22).

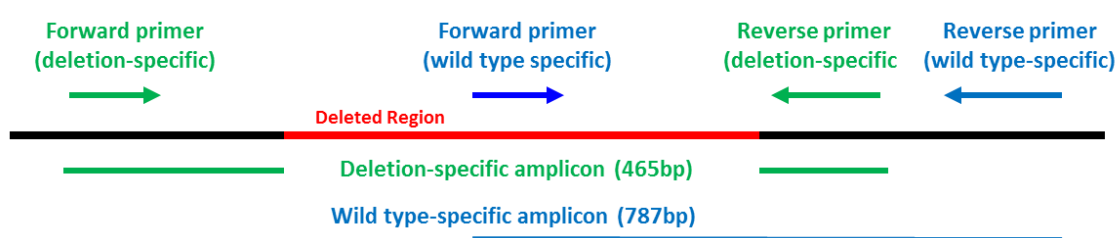


Figure 5.22: Schematic representation of CNV5583.1 breakpoint assay. Independent deletion-specific and wild-type specific primer sets produce amplicons of different sizes according to deletion status.

Primer sequences are provided in table 5.14. The CNV5583.1 region was amplified using standard Applied Biosystems reagents and thermal cycling parameters (section 2.1.2). Gel electrophoresis was performed on 2% agarose gels (100bp DNA ladder) as detailed in section 2.1.3, with an electrophoresis time of 60min used to improve differentiation between fragment sizes.

Primer	Sequence
Wild-type-specific FWD	CTGTCTTGCCCTTTGCTTCT
Wild-type-specific REV	TCTTCCACCCAGCTTTCATC
Variant-specific FWD	TGTCCACAGCAACGTAGGAG
Variant-specific REV	GGCTTCTCAGTTCCCAAACA

Table 5.14: Primers for CNV5583.1 breakpoint assay

As with the *NEGR1* deletion, it was not possible to duplex primer sets, even at a range of $MgCl_2$ concentrations and annealing temperatures. For CNV5583.1, duplexing resulted in the appearance of a spurious amplicon of around ~300 bp, many non-specific bands and disappearance of the bands of interest (figure 5.23). Wild-type and variant primer sets were therefore amplified in simplex and pooled for gel electrophoresis.

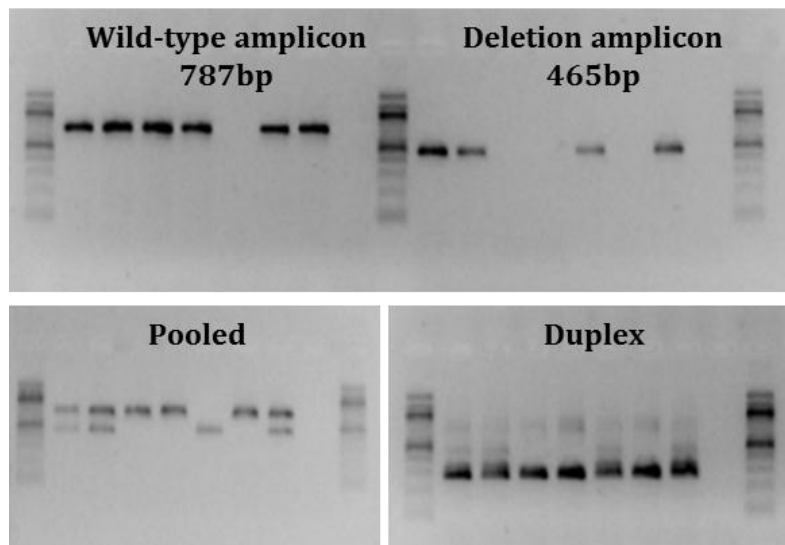


Figure 5.23: Duplexing test (*TSPAN8* CNV). Top panels: seven samples (plus NTC in right hand well) run with deletion (right) and non-deletion (left) primer sets separately. Bottom panels: simplex products as above but pooled for gel electrophoresis (left) and duplexed PCR (right). Note spurious bands in duplexed PCR. Samples run with 100bp DNA ladder, gel image shown in negative.

CNV genotyping results are shown in figure 5.24 and table 5.15. Genotyping success was relatively poor (76%) for the case samples – due both to complete PCR failure, and to uncertain or discrepant CNV genotype calls. Poorer performance amongst case sample DNA is consistent with results from other research projects, and is attributable to older, more frequently thawed and handled DNA. Hardy-Weinberg disequilibrium in the case samples is due to an underrepresentation of heterozygote samples (based on observed allele frequencies, 220 heterozygotes would be expected), a phenomenon

which is explicable if unreliable PCR amplification tends to result in heterozygote samples appearing homozygous due to an absent amplification band.

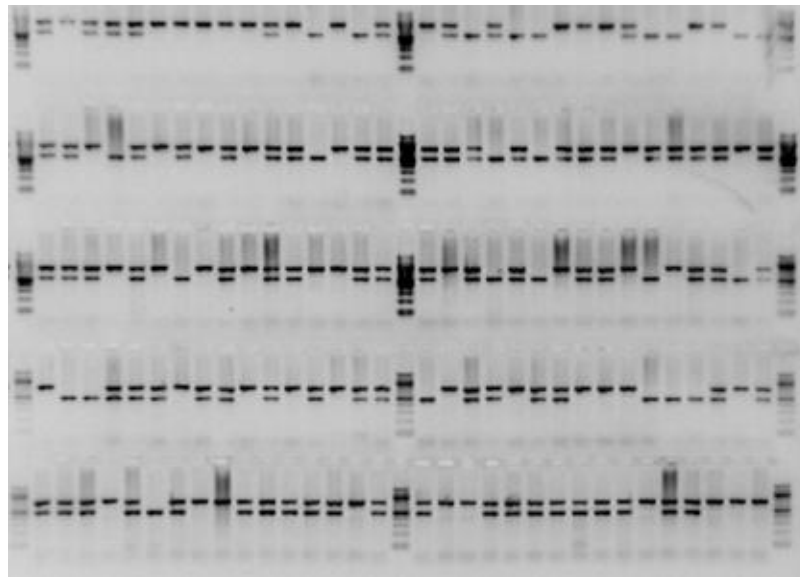


Figure 5.24: *TSPAN8* gel electrophoresis. Gel image is shown in negative and with a 100bp DNA ladder. Smaller amplicon (465bp) represents chromosomes carrying the deletion; larger band (787bp) represents wild-type chromosomes.

	Controls	WTCCC cases
Genotyped N	736	588
Genotyping success	95%	76%
Wild-type N	235	158
Heterozygote N	349	193
Homozygous variant N	114	98
Variant frequency	0.41	0.43
Hardy-Weinberg p-value	0.74	0.03

Table 5.15: *TSPAN8* CNV genotyping results

Because of the unreliability of case sample amplification, concordance between the CNV identified through sequencing (breakpoint assay data) and CNV5583.1 (data from WTCCC) was assessed in control samples only. The two CNVs showed a concordance of 98%, strongly suggesting that the two variants are, in fact, the same polymorphism. All

14 inconsistent samples were SNP-genotyped heterozygotes which appeared to be homozygous in initial breakpoint analysis but were confirmed to be heterozygous upon repeat amplification.

The WTCCC CNV scan reports deletion allele frequencies of 0.36 and 0.40 in control and disease cohorts respectively, providing a moderate p-value for association with T2D of 3.9×10^{-5} . Like CNV1943.1, CNV5583.1 has a tagging SNP (rs1798090) which can be used as proxy to investigate the deletion in existing SNP GWAS studies. In the DIAGRAM meta-analysis (Voight et al., 2010), rs1798090 is less strongly associated with T2D risk than other regional SNPs - notably rs4760790 (table 5.16).

Variant	Odds ratio (95%CI)	p-value
rs1798090 (CNV-tagging SNP)	1.08 (1.04-1.13)	1.5×10^{-4}
rs4760790 (best regional SNP)	1.11 (1.06-1.16)	3.6×10^{-6}

Table 5.16: Comparison of CNV-tagging and best regional SNP at the *TSPAN8* locus from DIAGRAM analysis. Note that results for rs4760790 are reported from stage 1 analysis for comparison with rs1798090. rs4760790 reached genome-wide significance at further analysis stages.

rs1798090 and rs4760790 are in weak LD ($r^2 = 0.23$). The stronger association disease association of rs4760790 suggests that it, rather than CNV CNV5583.1, is likely to be (or to be in strong LD with) the true causal variant. The disease association of CNV5583.1 is therefore likely to be a consequence of weak correlation with the true causal variant. CNV5583.1 itself is therefore unlikely to be responsible for T2D risk at the *TSPAN8* locus, and was not analysed further.

5.4 Discussion

5.4.1 Genotyping method comparison

On the basis of variation at *NEGR1* and *FCGR3*, the superiority of MAQ over QPCR for genotyping CNVs is clear. MAQ data shows better concordance with “true” copy number (as defined by tagging SNP genotype or data from other platforms), and achieves more reliable copy number calling on the basis of cluster analysis. Whilst QPCR was broadly successful in genotyping the simple *NEGR1* deletion, only MAQ successfully classified the more complex *FCGR3* polymorphism. In fact, for the *NEGR1* deletion, MAQ performed better than the breakpoint assay. Both methodologies classified samples with perfect accuracy, but MAQ produced fewer sample failures. Other breakpoint assay concerns – failure to distinguish between heterozygotes and deletion homozygotes for CNV1943.1, biased longer amplicon PCR failure for CNV5583.1 and subjective calling for all assays – suggest that MAQ should be the first choice approach for CNV genotyping, even when breakpoint assays are available.

Similar conclusions were drawn from parallel experiments conducted by Dr. Kathy Stirrups in the T1D replication laboratory at the Wellcome Trust Sanger Institute (Personal Communication). A comparison of Oxford and Sanger data sets for the *FCGR3* variant (figure 5.25) reveals that the two centres produced similarly good separation between copy number classes for MAQ and similarly poor separation for QPCR. Cluster analysis based on the Sanger data closely reflects results from Oxford data, classifying all samples correctly from MAQ, but misclassifying 52% samples from QPCR.

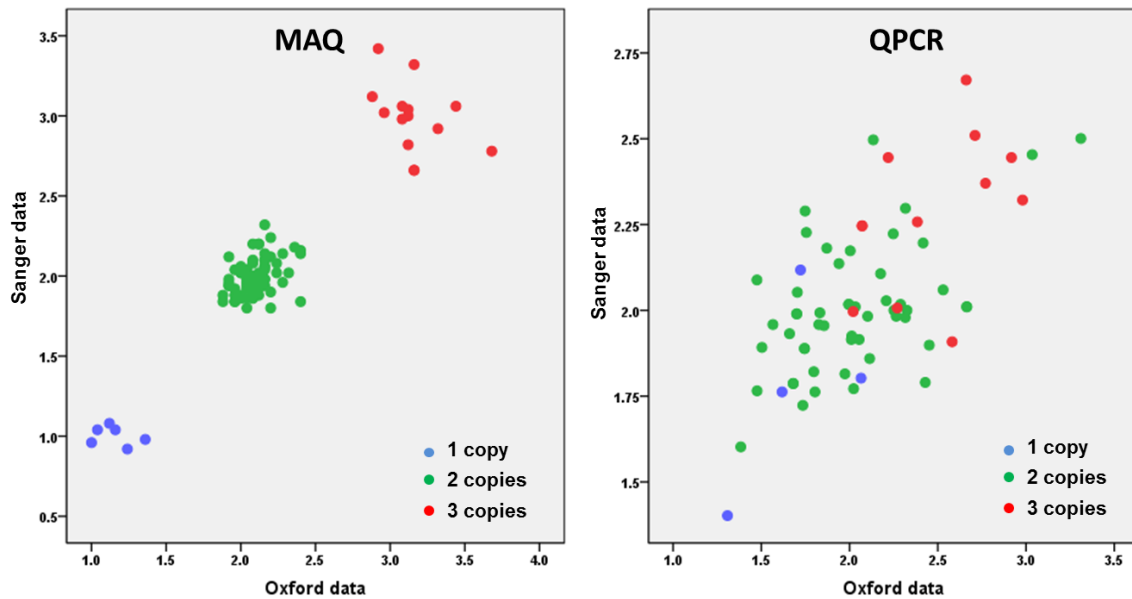


Figure 5.25: Comparison of MAQ and QPCR data for the *FCGR3* copy number polymorphism from laboratories in Oxford and at the Sanger Institute. Points are coloured by copy number according to WTCC pilot calls.

The success of MAQ may, in part, be due its inclusion of multiple amplicons within a region of interest. In effect, it produces multiple mini data-sets per CNV. Modelling of PCR-based quantitative assays suggest that five replicates are required to distinguish one from two copies (with 95% accuracy) and eight replicates are needed to reliably separate two and three copies (Weaver et al., 2010). Here, QPCR was able to distinguish between one and two copies at the *NEGR1* variant with three replicates, but to achieve more redundancy would be practically and financially unfeasible using standard QPCR laboratory techniques.

Microfluidic platforms such as those provided by Fluidigm (San Francisco, US) may facilitate higher replicate numbers, providing more reliable copy number genotyping by QPCR. They also hold the potential for digital PCR, whereby many reactions are performed at very low DNA input concentrations, such that some reactions are amplification-positive and others are negative. The proportion of positive reactions

(compared to a reference) provides an indication of copy number. Digital PCR returns copy number analysis to a simple binary counting exercise, with fewer opportunities for experimental noise and no demands of equivalent assay efficacy. It may prove more accurate than standard PCR analysis (Whale et al., 2012), but requires hundreds of independent reaction chambers per sample (Weaver et al., 2010) and is therefore impractical without specialist equipment.

MAQ analysis demands only standard techniques (PCR and fragment analysis), and therefore seems the best practical option for CNV genotyping in most laboratories. However, cost implications must also be taken into consideration. Whilst the high multiplexing potential of MAQ may render it cost-effective in replicating multiple CNVs in a single reaction, the initial expense of assay design (around €2000) makes it unrealistic for small-scale follow-ups such as those conducted for CNV1943.1 and CNV5583.1. The use of breakpoint assays in my own follow-up analyses demonstrate that, for replication of simple deletions where a high levels of multiplexing is not required, MAQ is not cost-effective when compared to a breakpoint assay.

5.4.2 Linkage Disequilibrium between CNVs and SNPs

CNVs 1943.1 and 5583.1 were both perfectly tagged by SNPs, limiting the necessity for direct copy number genotyping. This is indicative of overall results from the WTCCC CNV scan: well-typed CNVs tend to be well tagged by SNPs. Of the 18 CNVs marked for replication (across all eight diseases), 15 were perfectly tagged by SNP variation. In the study as a whole, 78% of biallelic CNVs with a MAF > 0.1 were well tagged ($r^2 > 0.8$) by at least one SNP.

Other studies have also revealed widespread SNP-tagging of common CNVs. Analyses of 180 and 270 CEU HapMap samples concluded that 77% and 85% respectively of common CNVs ($MAF > 0.05$) are well captured ($r^2 > 0.8$) by HapMap SNPs, and more than 70% are perfectly ($r^2 = 1.0$) tagged (Conrad et al., 2010; McCarroll et al., 2008b). The LD relationships between SNPs and common biallelic CNVs are similar to those between SNPs themselves (McCarroll et al., 2008b), suggesting that common deletions generally represent ancestral mutations which arose in single mutational events.

If most common CNVs are indeed tagged by SNPs, then their contribution to common disease is likely to already have been indirectly explored through SNP-based studies. There seems little to be gained by directly genotyping well-tagged CNVs in replication cohorts, since doing so will provide no more information than genotyping the tagging SNP. However, there may still be value in cataloguing the LD relationships between CNVs and SNPs. A CNV which is in strong LD with established disease-associated SNPs may emerge as a causal candidate for the locus. For example, a 20-kb deletion immediately upstream of *IRGM* is in perfect LD with the most strongly associated regional SNP for Crohn's disease. The tagging SNP is non-coding and all regional coding SNPs are less strongly disease-associated, providing good evidence that the deletion might be the causal variant (McCarroll et al., 2008a). Had the SNPs which tagged CNVs 1943.1 and 5583.1 emerged as the most strongly disease-associated regional variants, the CNVs would have been good candidates for a functional effect.

Where a candidate CNV and SNP are in perfect LD, the causal variant can only be confidently determined by dissociating their effects. Functional assays using recombinant DNA sequence can separate SNP and CNV, but figure 5.26 shows that it may also be possible through genotyping in alternative populations with differing

patterns of LD. In Europeans, the region of perfect LD around rs1798090 is over 10kb wide and encompasses CNV5583.1, but in HapMap YRI samples the interval is narrower and does not include the CNV. As discussed in section 5.3.3 (b), CNV5583.1 is not an interesting CNV for follow-up, but it demonstrates the principle that genotyping SNPs and CNVs independently in samples of African descent may allow their effects to be separated.

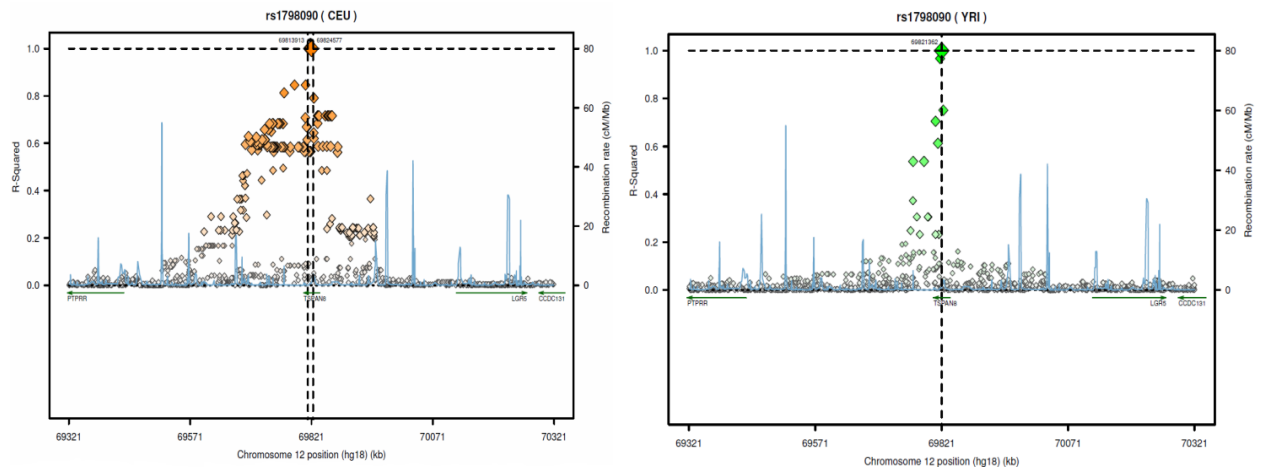


Figure 5.26: LD at the CNV5583.1 (*TSPAN8*) locus in European (left) and Yoruban (right) populations. SNPs are plotted by chromosomal location (x-axis) and r^2 value (y-axis) in relation to the CNV-tagging SNP rs1798090. Vertical dotted lines show the interval encompassed in an r^2 of 1.

5.4.3 Copy number variation and T2D

The WTCCC scan as a whole failed to provide evidence of an extensive role for copy number variation in complex disease susceptibility. In addition to the T2D variants detailed above, variation in the HLA region was associated with autoimmune disorders (Crohn's disease, rheumatoid arthritis and T1D), and the previously-identified (McCarroll et al., 2008a) deletion at *IRGM* was associated with Crohn's disease. The scan captured 42-50% of common ($MAF > 0.05$) CNVs longer than 0.5kb, with 80% power to detect an effect with an odds ratio >1.5 and 50% for an odds ratio >1.24 . On this basis, common CNVs seem unlikely to make a substantial contribution to T2D-susceptibility, or to account for a significant proportion of the missing heritability.

Similarly, an analysis of 275 T2D cases and 496 controls in a Korean population identified no robust associations between CNVs T2D susceptibility, and putative associations showed no overlap with the WTCCC analysis (Bae et al., 2011).

Since 80% of the copy-number differences between any two individuals arise from common CNVs (McCarroll et al., 2008b), the WTCCC scan is likely to have captured a large component of human copy number variation. However, it remains possible that low frequency ($MAF < 0.1$) CNVs, or complex polymorphisms with many copy number classes, may contribute significantly to T2D risk - as is the case for BMI and obesity (Bochukova et al., 2010; Jacquemont et al., 2011; Walters et al., 2010) CNVs with complex mutational origins are less likely to be well-tagged by SNPs on a single haplotype, and also challenging to type and therefore more likely to be excluded from genome-wide analysis. In the WTCCC study, only 22% of rare ($MAF < 0.05$) CNVs were well tagged, and previous analyses also estimate this figure at around 20% (Conrad et al., 2010; McCarroll et al., 2008b). However, a comparison of WTCCC data from tagged and non-tagged CNVs identified no enrichment for disease association amongst poorly tagged variants - suggesting either that rare and complex CNVs do not play a significant role in T2D susceptibility, or that the ones which do remain absent from genome-wide analyses.

The WTCCC CNV analysis has demonstrated that common CNVs are unlikely to make a substantial contribution to T2D susceptibility. The missing heritability of T2D must lie in other kinds of variation - perhaps rare CNVs as yet not interrogated through genome-wide scans, or perhaps rare single nucleotide variants. The search for T2D-associated rare variants forms the subject of chapter six.

Chapter 6

Rare single nucleotide variation and T2D risk

6.1 Introduction

6.1.1 Rare variation and common disease susceptibility analysis

Genome-wide association studies of common single nucleotide variation have been successful in identifying novel T2D susceptibility loci. However, currently identified common SNPs account for a small proportion of the overall variance of T2D, and attention has fallen on other types of genomic variation which may account for the unexplained variance (Manolio et al., 2009, and see section 1.4).

One such alternative source of variation can be found in low frequency (which I define here as $0.005 < \text{MAF} < 0.05$) and rare (defined here as $\text{MAF} < 0.005$) single nucleotide variants. As discussed in section 1.4.2, existing GWA studies largely exclude rare and low frequency variants and are underpowered to detect their disease associations, whilst incomplete penetrance impedes detection by linkage analysis. However, deep resequencing and large-scale whole genome sequencing programmes are uncovering variants in the low and rare frequency ranges, facilitating disease association analysis.

For example, the 1000 Genomes Project is one of the largest whole-genome sequencing programmes currently underway. Pilot phase data, published in 2010, included whole genome sequence from 179 individuals (Consortium, 2010) at low sequence coverage. Subsequent data releases have increased sample numbers and sequencing depth - low coverage whole genome and high coverage exome sequence from 1092 individuals as of March 2012 - and the project aims ultimately to provide sequence data for 2500 samples from five major ancestral groups. Eight million of the 14.5 million variants identified in the pilot project were novel, and most of these were rare. Simulations

suggest that, in the populations it samples, the 1000 Genomes Project will identify almost all variation with MAFs > 0.005 and at least 80% of variation with MAFs > 0.001 (Ionita-Laza et al., 2009).

Sequencing projects such as the 1000 Genomes are not only useful in identifying new variation; they also enable it to be tested for association with disease susceptibility. Obtaining rare variant genotypes through sequencing is still prohibitively expensive on a very large scale, but by uncovering the haplotype backgrounds of low frequency variant alleles, sequencing analysis provides a reference panel which can be used to impute genotypes on the basis of existing common variant SNP data. Variants become less amenable to imputation analysis as allele frequencies decrease, and it should be noted that the LD relationships of private variants ($MAF < 0.00001$) cannot be determined. Imputation software programmes use haplotype patterns in a reference panel to predict unobserved genotypes in a study dataset.

6.1.2 A genome-wide association study for rare variants and T2D risk

As part of the continuing effort to map rare and low frequency variant disease signals, analysis performed by Dr. Andrew Morris at the Wellcome Trust Centre for Human Genetics, University of Oxford, aimed to assess the evidence for association of T2D with rare and low frequency (defined in his analysis as $MAF < 0.01$) genetic variation. This pilot analysis was performed in 1,926 T2D cases and 2,942 controls of European descent, genotyped for common variants (Affymetrix 500K chip) as part of the Wellcome Trust Case Control Consortium SNP GWAS (WTCCC, 2007). Analysis in these WTCCC samples was intended to serve as a pilot study for rare variant imputation in larger sample numbers from the DIAGRAM consortium.

Rare and low frequency (but not private) variant genotypes were imputed using variants discovered and haplotype patterns identified in the 1000 Genomes Project European reference panel (August 2010 release from 560 haplotypes), using IMPUTEv2 software (Howie et al., 2011; Howie et al., 2009). T2D association was tested using GRANVIL (Gene- or Region-based ANalysis of Variants of Intermediate and Low frequency) software, which incorporates uncertainty in genotype imputation (Magi et al., 2011; Morris and Zeggini, 2010). Analysis is therefore based not on definite genotype calls, but on a probability estimate of each genotype for every sample. Only well-imputed SNPs were included, as defined by imputation info scores > 0.4 . Info scores measure the reduction in power of imputed data when compared with direct genotyping; a score of 0.5 represents 50% power to detect disease association compared to direct genotyping data.

GRANVIL models disease status as a function of the proportion of variants ($MAF < 0.01$) within a gene at which an individual carries a minor allele. Gene boundaries were

defined using the UCSC human genome database (build 37), extended 50kb up- and down-stream to incorporate regulatory regions and other functional elements. GRANVIL assesses mutational load within individuals, testing for association of T2D with accumulations of minor alleles within genes. It rests on a model of multiple rare causal variants, each of modest effect size but residing in the same gene. Searching for an accumulation of multiple rare alleles has greater power than analysing rare variants singly, and has proved successful in identifying rare variation contributing to low high-density lipoprotein-cholesterol (HDL-C) levels and T1D (Cohen et al., 2004; Nejentsev et al., 2009).

The strongest signals of association (i.e. accumulation of rare and low frequency variants) in this analysis were identified in *BMP2* (bone morphogenic protein 2) and *IGFL4* (insulin growth factor-like family member 4; figure 6.1). Their associations with T2D risk ($p=1.0 \times 10^{-6}$ and $p=2.4 \times 10^{-6}$ respectively) are genome-wide significant when Bonferroni corrected for 30,000 tested genes. *BMP2* contained 36 rare variants (mean MAF = 0.0043) with odds ratios of 1.26 (1.15-1.38) per minor allele, whilst *IGFL4* contained 12 rare variants (mean MAF = 0.0044) with odds ratios of 1.85 (1.43-2.38) per minor allele (table 6.1).

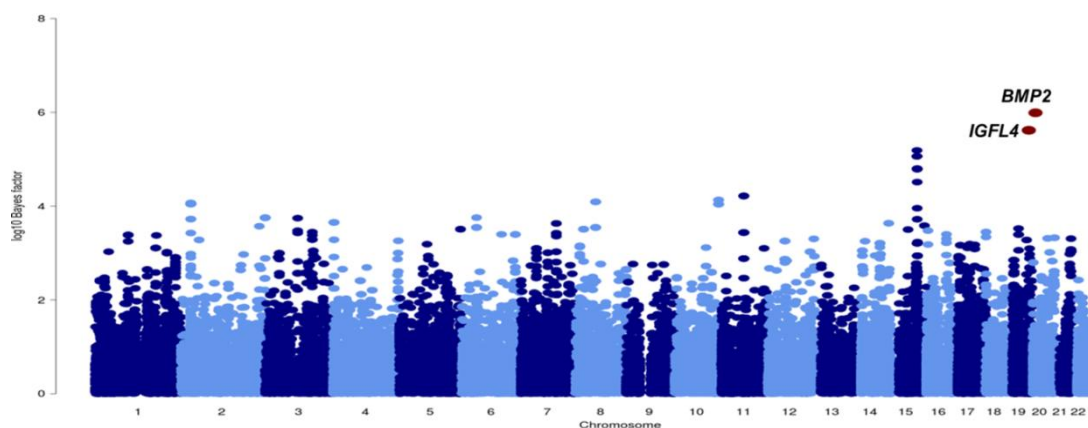


Figure 6.1: Manhattan plot summarising genome-wide rare variant association with T2D. Chromosomal position (x-axis) is plotted against strength of disease association (y-axis). Genes with the strongest accumulation of rare variants in case samples (genome-wide significant when correcting for 30,000 genes) are marked in red.

Gene	Number rare variants	Mean MAF	Mean odds ratio (95% CI) per minor allele	p-value
<i>BMP2</i>	36	0.0043	1.26 (1.15-1.38)	1.0x10 ⁻⁶
<i>IGFL4</i>	12	0.0044	1.85 (1.43-2.38)	2.5x10 ⁻⁶

Table 6.1: Genes with significant evidence ($p < 10^{-5}$) of rare variant association with T2D. Minor allele frequencies (MAF) and odds ratios are averaged across all rare variants within each gene.

BMP2 encodes a growth factor, which regulates the differentiation of pluripotent mesenchymal cells in to bone (osteogenesis), cartilage (chondrogenesis) or fat (adipogenesis) in a dose-dependent manner (Sottile and Seuwen, 2000). It is thought to play a role in β -cell differentiation (Chung et al., 2008; Yew et al., 2005), and is required for normal pancreas development (Ahnfelt-Ronne et al., 2010). Variants within *BMP2* have previously been associated with variation in fat mass (Devaney et al., 2009; McGuigan et al., 2008), providing further credibility as a candidate for T2D susceptibility. The physiological function of *IGFL4* is not clear, but it is strongly expressed in the cerebellum and the family of signalling molecules to which it belongs plays crucial roles in cellular energy metabolism and in growth and development (Emtage et al., 2006).

6.1.3 Experimental aims

The 1000 Genomes reference panel is a new and constantly updating resource. The accuracy of rare variant imputation which it provides is variable, and so disease associations based upon imputed data require experimental validation. My experiments aimed to:

- Directly genotype associated rare and low frequency variants in a subset of the samples used for association analysis, to assess the accuracy of imputation using the 1000 Genomes European reference panel
- Assess the impact of direct rather than imputed genotype calls upon disease association strength

The burden analysis conducted by Dr. Morris models disease status as a function of the *proportion* of rare variant sites at which an individual carries a minor allele. It is therefore not necessarily clear which particular variant or variants within a gene are driving an association.

Of 48 variants close to the two genes, nine (six for *BMP2* and three for *IGFL4*) were individually associated with T2D at $p < 0.005$, in a direction consistent with contributing to the mutational load (table 6.2). These variants were selected for direct genotyping. Two of the variants lie within introns of *BMP2*; the seven others fall within the 50kb window included around genes to capture functional elements. None was coding.

Closest Gene	Variant	Position	Major allele	Minor allele	MAF cases (n=1926)	MAF controls (n=2942)	Info score	T2D OR	T2D p-value
BMP2 (chr:20)	rs75328734	6,747,206	G	A	0.01002	0.00532	0.480	1.89	1.1x10 ⁻⁴
	rs79750773*	6,750,169	A	G	0.00998	0.00527	0.468	1.90	8.8x10 ⁻⁵
	rs11908042*	6,756,100	C	T	0.00991	0.00516	0.449	1.93	4.4x10 ⁻⁵
	rs12625185	6,781,727	C	T	0.00601	0.00281	0.707	2.15	4.8x10 ⁻³
	rs73593469	6,784,103	T	A	0.00578	0.00249	0.753	2.33	3.5x10 ⁻³
	BMP2_1	6,786,330	G	A	0.01066	0.00549	0.501	1.95	1.2x10 ⁻⁴
IGFL4 (chr:19)	rs80276707	46,502,150	G	A	0.01021	0.00543	0.455	1.89	7.6x10 ⁻⁵
	IGFL4_1	46,513,006	C	T	0.00528	0.00302	0.406	1.75	4.8x10 ⁻³
	IGFL4_2	46,561,434	C	T	0.00490	0.00261	0.448	1.88	3.5x10 ⁻³

Table 6.2: Rare and low frequency variants selected for genotyping. Minor allele frequency (MAF), per allele odds ratio (OR) and p-value for T2D association are according to imputed genotype data. Info score provides an indication of imputation confidence; variants with an info score >0.4 are considered well imputed. Positions quoted are hg19/b37. *denotes variants within the gene (intronic) – all other variants fall within the 50kb analysis window around genes. *BMP2_1*, *IGFL4_1* and *IGFL4_2* have not been assigned rs numbers.

6.2 Samples and Methods

6.2.1 Samples used

T2D case samples from the Diabetes UK Warren 2 repository (n=1503) and control samples from the United Kingdom Blood Service (UKBS; n=1428) and 1958 Birth Cohort (58BC; n=1458) collections were genotyped. Genotyped samples were a subset of those used in the imputation analysis described in section 6.1.2. Sample details are provided in section 5.2.1.

6.2.2 Genotyping

Inventoried assays were not available for any of the variants selected for follow up. Where possible, TaqMan genotyping assays were designed using Primer Express software, as described in section 2.5.1. Primer and probe sequences for designed assays are provided in table 6.3, although rs73593469 failed manufacture. Assays for rs79750773, rs11908042, rs12625185 and *IGFL4_1* could not be designed using Primer Express, even when amplicon length, complementarity and annealing temperature parameters were relaxed. These sequences were submitted to Applied Biosystems for assay design using their in-house software. Assays for rs11908042, rs12625185 and *IGFL4_1* were successfully designed and manufactured, but rs79750773 again failed the design process - perhaps because it lies in a very GC-rich region. Primer/probe sequences for assays designed in-house by Applied Biosystems are not available.

Variant	Primer	Sequence
BMP2_1 (-ve strand)	FWD primer	TCCTCCCAACCTCAAGTTATCC
	REV primer	ATTTTTGAGGAATCCAGTTATTGGA
	FAM probe	CCTGTCTTGACCTCC
	VIC probe	CTGTCTTGATCTCCCA
rs75328734 (BMP2) (+ve strand)	FWD primer	CGATAGCTCTGGGAAAGCAGAA
	REV primer	GAGCTGCTGCACGATGAGTCT
	FAM probe	ATCCTGGGAAGGGCAGA
	VIC probe	TGGGAAGGACAGAAA
rs73593469* (BMP2) (+ve strand)	FWD primer	GGAGCACTTATGTATCATACTCTGCATA
	REV primer	CCCGGCTGGGCAAGA
	FAM probe	TGAGGTCTAGAATGGT
	VIC probe	TTGAGGTCTAGAAAGGT
rs80276707 (IGFL4) (+ve strand)	FWD primer	CCAAGGACACACCAAGGAGAA
	REV primer	TGCGGCCTCACACCTTAAG
	FAM probe	AACCCAGAAGTACC
	VIC probe	TGAACCCAGAAATAC
IGFL4_2 (-ve strand)	FWD primer	CCGATACCTGTACCCCCATTG
	REV primer	CACCTAGGAGCCTGTAAAATCAAAA
	FAM probe	CTGGGAAGTAACTAAGTTG
	VIC probe	CTGGGAATAACTAAGTT

Table 6.3: Rare and low frequency variant TaqMan genotyping assay primer and probe sequences, as designed using Primer Express software. *rs73593469 assay failed manufacture.

A total of seven successfully designed and manufactured assays were available for genotyping. Genotyping was initially performed using standard procedures as described in section 2.5.2. Where assays were unsuccessful in differentiating positive control samples (confirmed heterozygotes), a range of optimisation steps were taken. Samples were subjected to an extra 10 (total 50) and an extra 20 (total 60) cycles of PCR amplification. Genotyping was tested with a double concentration of template

DNA (20ng per reaction), increased concentration of assay (0.05µl per reaction) and increased total reaction volume (5µl and 10µl). Two alternative thermal cycling programmes were tested: a shortened annealing and extension step of 60°C for 30 seconds, and a touchdown programme which began with an annealing temperature of 65°C and dropped by 0.5°C per cycle to 55°C.

Genotypes were assigned manually, because standard algorithms often fail when one homozygote genotype cluster is absent. Genotype calls were made independently by myself and a colleague (Dr. Amanda Bennett, OCDEM, University of Oxford), and any discrepant results were excluded from analysis.

6.2.3 Positive control sequencing

Where genotyping assays failed to differentiate samples with a high imputed probability of heterozygosity, those samples were sequenced to confirm their heterozygous status. Primers for PCR amplification across rare variant sites were designed as described in section 2.1.1. Primer sequences are detailed in table 6.4, and were amplified using Qiagen reagents as described in section 2.1.2. Bidirectional sequencing was performed using universal M13 primers as detailed in section 2.2.1 and analysed as outlined in section 2.2.3.

Gene	Variant	Primer	Primer sequence
BMP2	<i>BMP2_1</i>	FWD	GAGACAGGGTTTTGCCATGT
		REV	CCTCAGGGGGAGAGAATAGAA
IGFL4	<i>IGFL4_1</i>	FWD	TGGCTAGGCTGGACCTTAGA
		REV	TGGCTGGGAAGAGGAAGATA
	<i>IGFL4_2</i>	FWD	CTCCATCTTGGCCAGTCTGT
		REV	GCTTGTGCAAGGGAAGTCTC

Table 6.4: Rare variant positive control sequencing primers

6.3 Results

6.3.1 Genotyping results

The seven successfully manufactured TaqMan genotyping assays were validated using a panel of 40 randomly selected UKBS samples, supplemented by six putative positive heterozygote controls. Positive controls were samples with the highest probability of heterozygosity, according to imputation analysis (provided by Dr. Andrew Morris). They were therefore putative rather than confirmed heterozygotes.

Assay validation results are shown in figure 6.2, with positive control samples for each assay highlighted in purple. There is clear separation into common homozygote and heterozygote clusters for rs11908042, rs12625185 and rs80276707, and to a lesser extent for rs75328734. For these assays, positive control samples cluster predominantly into a heterozygote group, providing evidence that the assays successfully distinguish between genotypes and an initial indication of imputation accuracy.

BMP2_1 shows almost no amplification of the VIC fluorescent probe (y-axis), with all samples clustering in a homozygous X location. The single sample which appears to show allele Y amplification is likely to be spurious, since its duplicate partner is in the homozygous X cluster. Positive control samples show slightly lower FAM (x-axis) amplification for *IGFL4_1*, but no difference in VIC amplification (y-axis), and overall insufficient differentiation to make confident genotype calls. There appears to be very little amplification for *IGFL4_2*, with almost no separation between samples and NTCs.

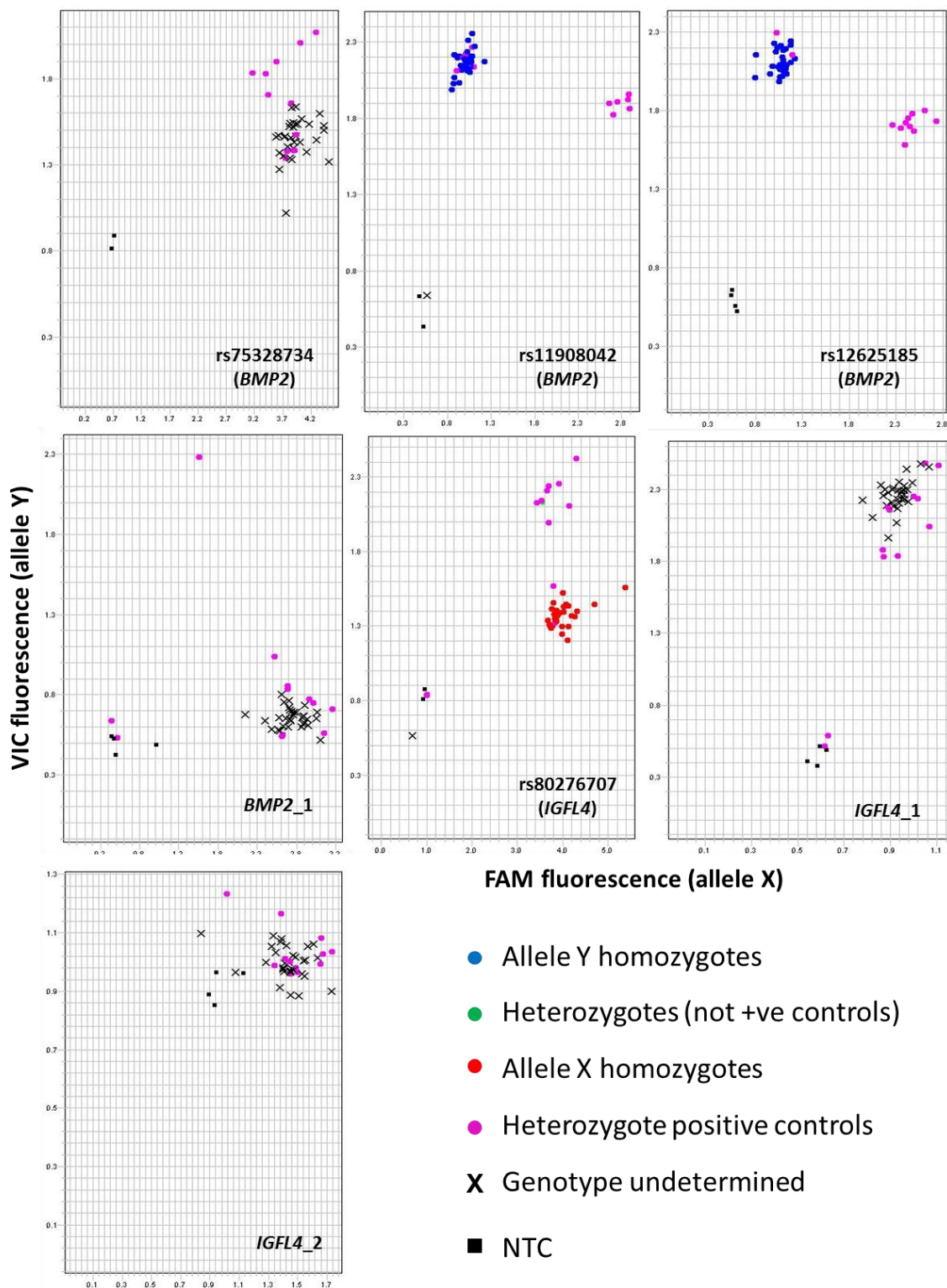


Figure 6.2: Rare variant assay validation plots, as produced by SDSv2.3 software. FAM fluorescence (x-axis) is plotted against VIC fluorescence (y-axis). Red and blue points represent homozygotes, green points represent genotyped heterozygotes and purple points represent heterozygote positive controls (on the basis of imputation). Black cross (X) denotes a samples of undetermined genotype, whilst black squares represent non-template controls (NTCs; dH₂O) which should not amplify. Note that positive control samples were run in duplicate, such that the twelve points represent six samples.

Rs75328734 is in very tight LD with rs11908042 ($r^2=0.97$). Since the genotyping assay for rs11908042 worked well, rs75328734 was not optimised. For *BMP2_1*, *IGFL4_1* and *IGFL4_2*, the heterozygous status of their putative positive control samples was confirmed via capillary sequencing (figure 6.3). At least three of the imputed positive controls were genuine heterozygotes in all cases, confirming that their failure to separate from the main sample cluster was due to assay failure. Genotyping assays for *BMP2_1*, *IGFL4_1* and *IGFL4_2* were tested using alternative template and probe concentrations and alternative amplification conditions (detailed in section 6.2.2), but could not be optimised to successfully distinguish confirmed rare variant heterozygotes from other samples. The possible reasons for assay failure are discussed in section 6.4.

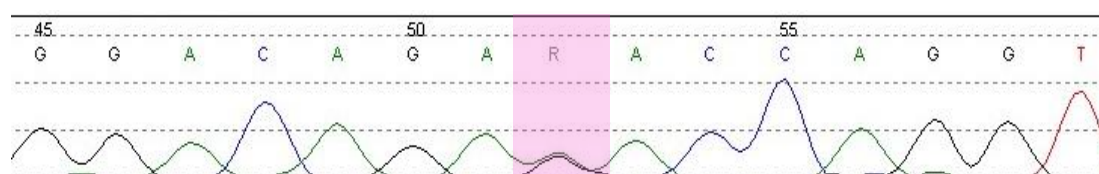


Figure 6.3: Heterozygote positive control Electropherogram. Highlighted variant is *BMP2_1* (G→GA)

The three successful rare variant assays (rs11908042, rs12625185 and rs80276707) were taken forward for genotyping in ~1500 T2D cases and ~2800 control samples, comprising a subset of the samples used for imputation analysis. Each genotyping plate included at least one positive control heterozygote. Figure 6.4 shows example allelic discrimination plots from the three assays, and overall results are provided in table 6.5. Note that the common allele is measured by VIC fluorescence (allele Y) for rs11908042 and rs12625185, but by FAM fluorescence (allele X) for rs80276707.

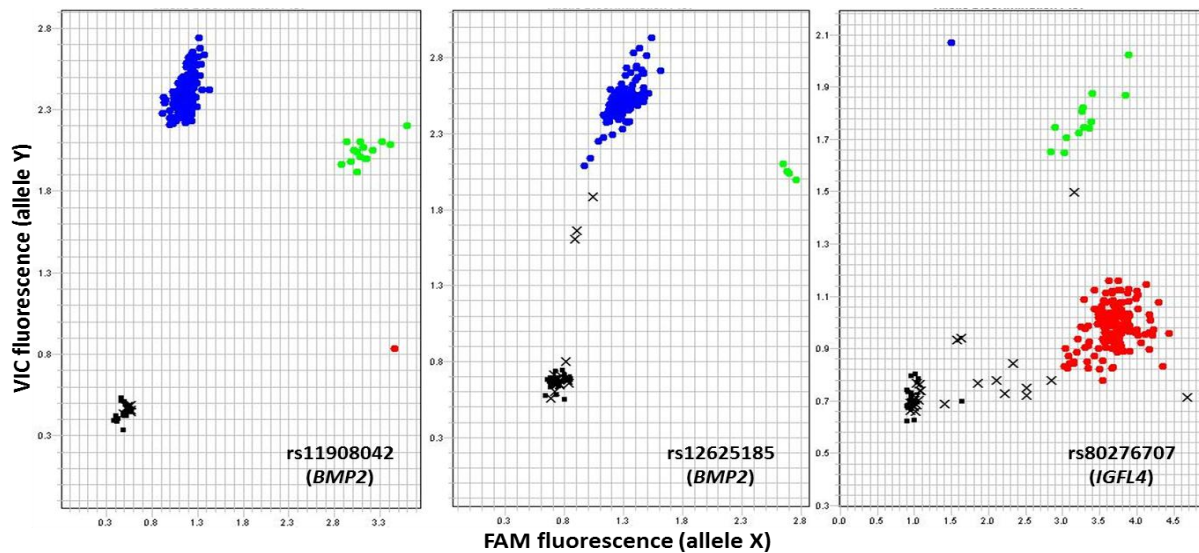


Figure 6.4: Example rare variant genotyping plots, as produced by SDSv2.3 software. FAM fluorescence (x-axis) is plotted against VIC fluorescence (y-axis). Red and blue points represent homozygotes, green points represent heterozygotes. Black cross (X) denotes a samples of undetermined genotype, whilst black squares represent non-template controls (dH₂O) which should not amplify.

Variant		N	Pass rate	N common homozygote	N heterozygote	N rare homozygote	MAF
rs11908042 <i>BMP2</i>	Controls	2840	95%	2639	65	0	0.012
	Cases	1503	90%	1316	43	1	0.017
rs12625185 <i>BMP2</i>	Controls	2886	95%	2723	22	0	0.004
	Cases	1500	82%	1219	12	2	0.006
rs80276707 <i>IGFL4</i>	Controls	2886	95%	2655	92	1	0.017
	Cases	1501	86%	1224	56	4	0.025

Table 6.5: Rare variant genotyping results. Pass rate refers to the proportion of genotyped samples with successful genotype calls.

Poorer genotyping success amongst case samples is consistent with the results presented in chapter four, and with data produced in other research projects using the same samples. It is likely to be attributable to older, more frequently thawed DNA.

6.3.2 Imputation accuracy

Comparing directly tested genotypes with imputed genotype probabilities provides an indication of the accuracy of rare variant imputation using the 1000 Genome European reference panel (August 2010 release). Table 6.6 shows the mean imputed probability of common allele homozygosity and heterozygosity for samples which were directly genotyped as common homozygotes (top row) or heterozygotes (bottom row). The info score is an estimation of imputation accuracy provided by IMPUTEv2 software. Probability estimates range from 0 (least likely) to 1 (most likely).

Variant	Info score	Genotype call	Imputed probability homozygote (mean)	Imputed probability heterozygote (mean)
rs11908042 <i>BMP2</i>	0.449	Homozygote	0.9920	0.0080
		Heterozygote	0.8243	0.1753
rs12625185 <i>BMP2</i>	0.707	Homozygote	0.9971	0.0029
		Heterozygote	0.5027	0.4971
rs80276707 <i>IGFL4</i>	0.455	Homozygote	0.9934	0.0066
		Heterozygote	0.7983	0.2104

Table 6.6: Summary of concordance between direct genotype calls and imputed genotype probabilities at rare variants in *BMP2* and *IGFL4*. For each variant, top row contains samples genotyped as common allele homozygotes and bottom row contains samples genotyped as heterozygotes. Columns represent the mean imputed probabilities of homozygosity and heterozygosity. Info score is an estimation of imputation accuracy from IMPUTEv2 software; scores >0.4 indicate good imputation.

The imputed probability of homozygosity for directly genotyped homozygotes is >99% in all cases, but this not surprising. For a rare variant with a MAF of 0.01, 99% accuracy could be achieved by predicting all samples to be common homozygotes. More instructive is the reduction in imputed probability of homozygosity, and the accompanying increase in imputed probability of heterozygosity, for directly genotyped heterozygotes. For all assays, genotyped heterozygotes exhibit a clear shift towards higher imputed probability of heterozygosity. This is particularly clear for

rs12625185, which may well reflect its higher info score – representing more accurate imputation. However, genotyped heterozygotes retain a higher mean probability of homozygosity than heterozygosity for all assays.

Figure 6.5 gives a graphical representation of this information for all genotyped samples (case and control together). FAM fluorescence is plotted against VIC fluorescence (as in a standard genotyping plot), but samples are coloured and sized according to their imputed probability of heterozygosity. Large red points represent samples with a high probability of heterozygosity according to imputation; small purple points represent samples with a low probability of heterozygosity. Samples with a high probability of heterozygosity are overrepresented in, but by no means restricted to, the heterozygous sample clusters. Similarly, samples with a high probability of common allele homozygosity are primarily, but not exclusively, experimentally genotyped in the homozygous clusters.

Table 6.7 shows the percentage of directly genotyped heterozygotes which are included within the top 5%, 2% and 1% of samples when ranked according to imputed probability of heterozygosity. In other words, the 5% of samples most likely to carry a rare allele according to imputation include 38%, 76% and 52% of the genotyped rare alleles for rs11908042, rs12625185 and rs80276707 respectively.

	Top 5%	Top 2%	Top 1%
rs11908042 (<i>BMP2</i>)	38%	31%	19%
rs12625185 (<i>BMP2</i>)	76%	76%	68%
rs80276707 (<i>IGFL4</i>)	52%	33%	21%

Table 6.7: Inclusion of genotyped heterozygotes in high probability imputed heterozygotes. For each variant, the percentage of genotyped heterozygotes included in the top 5%, 2% and 1% of samples most likely to be heterozygous according to imputation is shown.

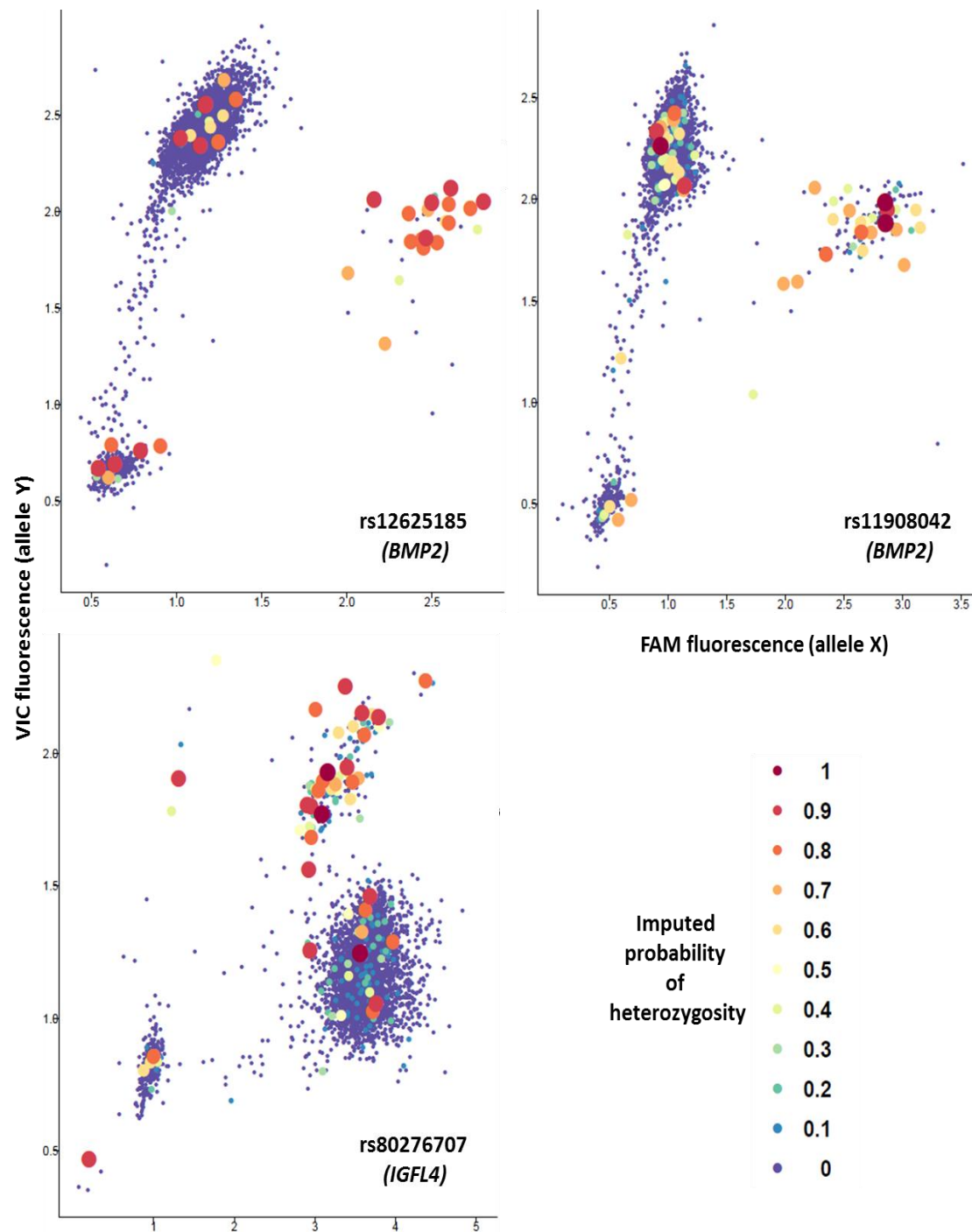


Figure 6.5: Genotyping cluster plots of rare variants, coloured and sized according to imputed probability of heterozygosity. Each sample (case and control) is coloured and sized according to its imputed probability of carrying the rare allele in a heterozygous state. Small purple points represent samples most likely to be homozygous for the common allele; large red points represent samples most likely to be heterozygous.

The imputation software IMPUTEv2 produces probability estimates for each genotype, rather than hard genotype calls – and these probability estimates were used in the original association analysis. However, the accuracy of imputation can also be assessed by comparing direct genotypes with hard imputed genotypes, based on a genotype probability cut-off. Figure 6.6 demonstrates the problem of selecting a suitable cut-off for genotype assignment. The red line shows the percentage of samples above each probability boundary (x-axis) which were shown to be heterozygotes through direct genotyping. The percentage of genuine heterozygotes might be expected to increase with the cut-off value, and this is broadly the case - although there is little improvement above 0.3. The green line shows the percentage of all genotyped heterozygotes which are ‘captured’ above each probability division, beginning at 100% and decreasing as values become more stringent.

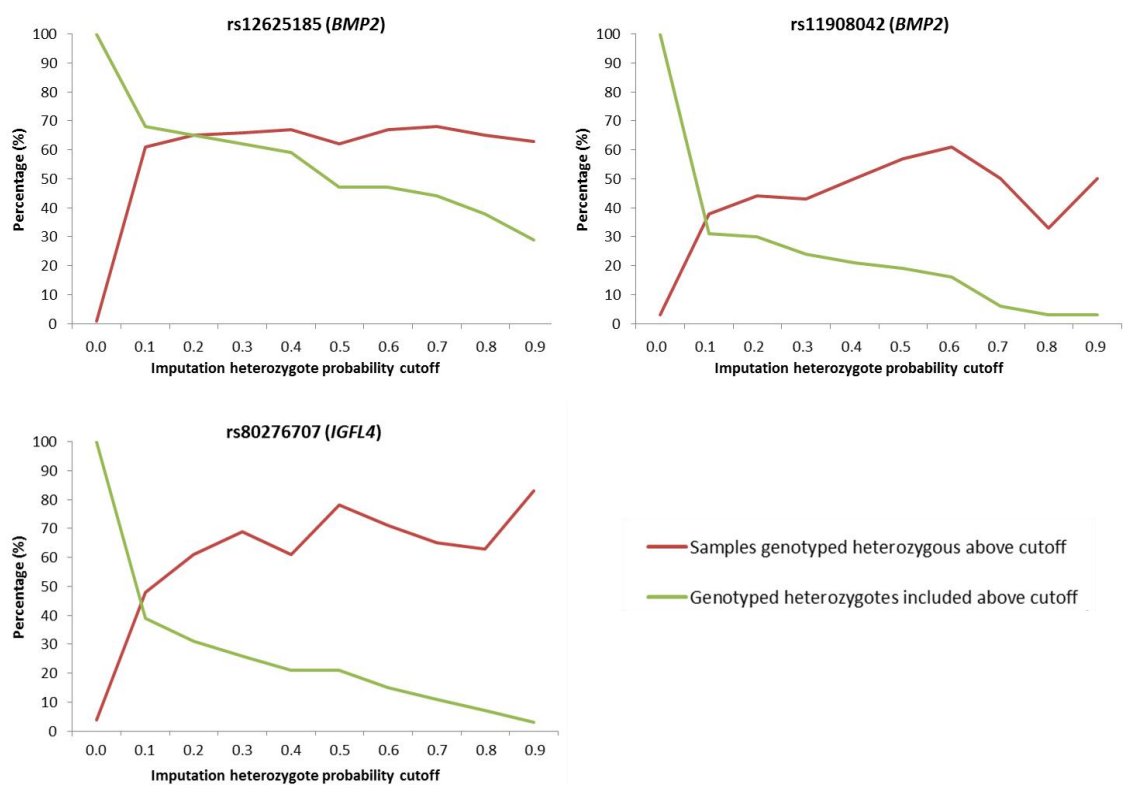


Figure 6.6: Performance of imputation probability cut-offs for genotyping rare variants. Cut-offs for imputed probability of heterozygosity (x-axis) is plotted against percentage of samples above that cut off which were genotyped as heterozygotes (in red) and percentage of all genotyped heterozygotes which are included above that cut off (in green).

The ideal boundary would see both lines at 100%, indicating perfect separation of genotyped heterozygotes and homozygotes. However, no assay comes close to perfect separation at any probability cut-off, and there is a trade-off between the two parameters. For rs80276707, a boundary at 0.9 would mean 83% of imputed heterozygotes are also genotyped heterozygotes, but only 3% of all genotyped heterozygotes are captured. Table 6.8 details these figures for all assays at potential divisions of 0.5, 0.8 and 0.9. The accuracy of genotyping (left-hand columns) is not much improved at more stringent cut-off values, but the inclusion of genotyped heterozygotes decreases dramatically (right-hand columns), particularly for rs11908042 and rs80276707.

	Samples above cut-off that are genotyped heterozygotes			Genotyped heterozygotes included above cut-off		
	0.5	0.8	0.9	0.5	0.8	0.9
rs11908042 (<i>BMP2</i>)	57%	33%	50%	19%	3%	3%
rs12625185 (<i>BMP2</i>)	62%	65%	63%	47%	38%	29%
rs80276707 (<i>IGFL4</i>)	78%	63%	83%	21%	7%	3%

Table 6.8: Performance of imputation probability cut-offs for genotyping rare variants. For each variant, the percentage of samples above genotyped as heterozygotes and the percentage of all genotyped heterozygotes included above imputation heterozygote probability cut-offs of 0.5, 0.8 and 0.9 are provided.

My results show encouraging but by no means perfect concordance between imputed and direct genotypes, and demonstrate the challenge of converting imputed probabilities to hard genotype classes.

6.3.3 T2D association analysis

The association analysis described in section 6.1.2 was conducted on the basis of imputed genotype data. GRANVIL uses imputed genotype probabilities rather than hard calls, taking uncertainty into account but reducing power to detect disease association. The inclusion of direct genotype calls will therefore increase analysis power.

Table 6.9 compares imputed and genotyped rare variant allele frequencies amongst case and control samples. For all variants and all samples, the genotyped MAF is higher than the imputed MAF, suggesting that imputation ‘missed’ more genuine heterozygotes than it did falsely classify homozygotes. However, the tendency for a higher rare allele frequency amongst case samples is retained for all variants. Association p-values are based on experimental genotypes only and calculated using Fishers’ exact test. The trend towards an over-representation of rare alleles in case samples is confirmed by suggestive association p-values. None of the genotyped variants reaches the 0.05 significance threshold, but association at rs80276705 is approaching significance and highly suggestive in a sample set of this size.

Variant	Controls		Cases		p-value for T2D association
	Imputed MAF	Genotyped MAF	Imputed MAF	Genotyped MAF	
rs11908042 (BMP2)	0.00516	0.01202	0.00991	0.01654	0.102
rs12625185 (BMP2)	0.00281	0.00401	0.00601	0.00649	0.110

Table 6.9: Comparison of imputed and genotyped rare variant frequencies for case and control samples. P-value for association with T2D based on direct genotypes and calculated using Fishers’ exact test.

The rare variant association analysis which highlighted *BMP2* and *IGFL4* as potentially interesting genes was not conducted using data from variants individually. Instead, it looked at the total number of minor alleles carried at all rare variants within a gene.

Table 6.10 compares results from the original GRANVIL analysis (performed entirely using imputed genotypes) with results when the analysis is re-run incorporating direct genotypes for rs11908042, rs12625185 and rs80276707. Analysis was conducted by Dr. Andrew Morris. It should be noted that not all samples included in the original analysis were available for direct genotyping, so the re-analysis comprises a mixture of experimental (~60%) and imputed (~40%) genotypes.

Gene	Imputed genotypes		Experimental genotypes	
	p-value	OR (95%CI)	p-value	OR (95%CI)
<i>BMP2</i>	1.0x10 ⁻⁶	1.26 (1.15-1.38)	7.2x10 ⁻⁶	1.26 (1.14-1.39)
<i>IGFL4</i>	2.5x10 ⁻⁶	1.85 (1.43-2.38)	4.2x10 ⁻⁴	1.63 (1.24-2.14)

Table 6.10: Comparison of burden test using imputed experimental genotypes. Odds ratio (OR) quoted per minor allele.

Disease association is weakened by the incorporation of experimental genotyping results for both genes, but does not disappear. For *BMP2*, the disease odds ratio per minor allele is retained, and association strength is altered only marginally. For *IGFL4*, association strength is weakened but retained, and the disease odds ratio is slightly reduced.

6.4 Discussion

Rare variant burden analysis uncovered some evidence for association of T2D with rare variation in two genes. It highlights the potential for identification of rare variant associations using existing GWAS genotyping data, supplemented with imputation from high-density reference panels.

The difficulties which I experienced in designing and amplifying for confirmatory genotyping highlight the challenges of genotyping rare variants. Some sequence regions are not amenable to amplification – they may be highly repetitive, homologous to other genomic locations or have an extreme G/C content. Both the assays which failed manufacture were in highly G/C-rich regions, whilst the *IGFL4_2* (which could not be optimised) was in a region of high homology to multiple other genomic locations. For common variation where LD structures are well-known, failed assays can often be replaced by proxy SNP assays, located in sequence more amenable to amplification. This was the case for rs75328734, which is in very high LD with and could therefore be replaced by rs11908042, but, in general, the LD relationships of rare variants are weaker and less well catalogued, and so there are fewer proxy options when assays fail.

It is also challenging to assign genotypes at rare variants. In all three assays, standard analysis software (SDSv2.3) was unable to make genotype calls in the absence of a rare homozygote cluster. Genotype calls were instead made manually, with the inevitable subjectivity which that entails. I was fortunate in having access to positive heterozygote controls, which could be used to confirm the location of heterozygote sample clusters. Where these are unavailable – or where positive control DNA is

limited – it may be necessary to use synthetic positive controls for TaqMan genotyping, comprising a mixture of oligonucleotides carrying the two possible rare variant alleles.

Assessment of imputation accuracy for rs11908042, rs12625185 and rs80276707 is encouraging in demonstrating a clear link between imputed and direct genotypes, but also shows that imputation is far from perfect. Attempts to select appropriate probability cut-offs demonstrate clearly the challenge of calling genotypes based on imputed probabilities. Similarly, analysis of imputation based on 1000 Genomes Project pilot data found an error rate of 35% for variants with MAFs < 0.01 (Altshuler et al., 2010). Analysis of 8865 samples with Affymetrix 500K SNP genotypes and deep exon sequence for 202 genes showed that, even using a large reference panel of over 3,000 samples, only 38% of variants with a MAF < 0.01 were well imputed ($r^2 > 0.7$ with experimental genotype), falling to 35% for variants with a MAF < 0.005 (Li et al., 2011).

There are several ways in which imputation based on the 1000 Genomes reference panel can be expected to improve. Firstly, the reference panel will expand. Analysis in this chapter was based on the August 2010 data release, which included 566 European haplotypes – but the project will ultimately sequence 2500 samples from five broad ethnic groups. Secondly, SNP data from denser arrays than the Affymetrix 500K chip used in WTCCC1 will facilitate more accurate imputation (Altshuler et al., 2010). In the deep exon analysis described above, the proportion of well imputed variants (MAF < 0.01) increased from 38% to 54% when data from the Illumina 550K rather than the Affymetrix 500K genotyping chip were used (Li et al., 2011). Finally, the eventual inclusion of sequence data from other populations will improve the accuracy of imputation. Analysis based on HapMap 3 has shown that imputation in CEU samples is improved by the addition of Toscani (Italian), Mexican and Gujarati (but not East

Asian) reference samples, independent of reference panel size (Jostins et al., 2011). The same analysis showed that increased reference panel size and diversity has the greatest benefit to imputational accuracy at lower allele frequencies, raising hope that subsequent and more diverse releases of 1000 Genomes Project data will allow more accurate imputation.

The strength of evidence for T2D association at the three variants which I genotyped does not support firm conclusions. Replacing most imputed genotypes included in the original analysis with experimental genotypes weakened disease association, but only slightly (particularly for *BMP2*). Had disease association disappeared or been dramatically attenuated upon inclusion of direct genotypes, the original result would probably have been spurious and due to erroneous imputation. Conversely, strengthened association would have provided strong evidence of a real effect which warranted large-scale follow-up. Given the inconclusive results, it is sensible to wait for imputation based on improving 1000 Genomes reference panels and in other larger sample sets to assess whether similar effects at *BMP2* and *IGFL4* emerge.

If whole genome sequencing becomes dramatically faster and cheaper, imputation will become unnecessary; rare and low frequency variants will be directly typed in all samples through sequencing. But for now, using large and deeply sequenced reference panels in conjunction with dense SNP arrays will greatly increase power to detect rare and low frequency variants associated with complex diseases such as T2D. The experiments reported in this chapter demonstrate that association analysis based on imputed variant data is promising but imperfect, and requires careful experimental validation.

Chapter 7

General Discussion

Research over the past decade has significantly improved our understanding of the mechanisms of T2D susceptibility. Genes encoding the glucokinase regulatory protein (*GCKR*), ATP-sensitive potassium channel genes (*KCNJ11/ABCC8*) and a zinc transporter (*SLC30A8*) have been convincingly implicated in disease pathogenesis through functional analysis of coding variation (Beer et al., 2009; Hamming et al., 2009; Nicolson et al., 2009; Rees et al., 2012), whilst variable expression of the transcription factor *TCF7L2* has been implicated through FAIRE-seq and transcript level analysis (Gaulton et al., 2010; Lyssenko et al., 2007). Almost 60 loci at which common genetic variation influences disease susceptibility have been identified, which, taken together, suggest the primacy of β -cell dysfunction over insulin resistance in the pathogenesis of T2D (Voight et al., 2010).

Despite these successes, our picture of the mechanisms by which genetic variation confers susceptibility to T2D is far from complete. Only a small proportion of disease-relevant variation has been identified (Voight et al., 2010), and the biological effects of most known variants are unclear. The work presented in this thesis has addressed both issues, using a variety of approaches to interrogate genetic variation other than common SNPs which may contribute to T2D susceptibility, and to uncover disease mechanisms at known associated loci.

Chapters three and four aimed to elucidate causal disease mechanisms, demonstrating the utility of highly detailed and more universally applicable approaches respectively. Chapter four described expression profiling of genes within loci associated with proinsulin/insulin ratios, which are a marker of T2D risk (Wareham et al., 1999). Disease-associated loci identified through GWA studies frequently contain multiple genes of equivalent biological plausibility, and my work as part of the MAGIC

consortium demonstrates the utility of expression profiling as a tool to eliminate or highlight multiple genes. Through quantification of mRNA transcript levels of 20 genes in 28 human tissues – including islets - *CT62* was excluded as a potentially causative gene because it is not expressed in any relevant tissues, whilst tissue-specific expression of *CELF2* and *MTCH2*, *MYOC1* and *YWHAE*, and *C2CD4A* and *STARD10* highlighted them as particularly interesting candidates at their respective loci.

Within chapter four, expression data is interpreted in conjunction with physiological data provided by other members of the MAGIC consortium (Strawbridge et al., 2011). For example, SNPs associated with proinsulin/insulin ratios at the *SGSM2* locus are also associated with insulin sensitivity (but not β -cell function), and I therefore looked for specific expression in tissues of insulin action. Conversely, proinsulin-related SNPs at the *VSP13C* and *MADD* loci are also associated with β -cell function (but not insulin sensitivity), increasing the relevance of regional genes expressed specifically in islets and β -cells. Similar syntheses of varied data will be required to confidently determine causal transcripts and mechanisms at disease-associated loci. The transcript profiling work which I performed is not in itself sufficient to determine mechanisms of T2D susceptibility, but will be of assistance in the selection of genes for functional and eQTL analysis.

Whereas the work presented in chapter four uses expression profiling across five associated loci, chapter three focuses intensively on one disease locus. My work on T2D-associated SNPs at *KCNQ1* highlights the challenges in determining molecular mechanisms of disease susceptibility, and the need for detailed consideration of regional biology. For *KCNQ1*, disease-associated SNPs had a parent-of-origin specific risk effect (Kong et al., 2009) and fell within an imprinted region, with established

relevance to prenatal growth syndromes and islet growth phenotypes (Kassem et al., 2001; Lee et al., 1999). By determining which of the regional transcripts are imprinted in developing foetal and adult islets, I was able to narrow the field of disease candidates from the seven transcripts originally tested to *KCNQ1*, *KCNQ1OT1* (both at early developmental stages only), or *CDKN1C* (at early developmental stages or adulthood). Given its established role in β -cell proliferation phenotypes (Kassem et al., 2001), *CDKN1C* is particularly interesting candidate for the disease effect.

Imprinted gene expression at the *KCNQ1* locus is regulated through methylation at an imprinting control centre (Thakur et al., 2004). By demonstrating that T2D-associated SNPs are also associated with increased DNA methylation, my work identifies a potential disease susceptibility mechanism. I was not able to detect any effect of T2D risk genotype status upon gene expression (total or allele-specific), and cannot therefore report conclusive answers regarding disease mechanisms or relevant transcripts. There are, however, several avenues of future work which may be able to more fully determine disease mechanisms. Most obviously, the failure to detect expression effects may simply be a problem of power, rectifiable by a larger sample size. It is also possible that altered transcript half-life or protein level, or, in the case of *CDKN1C*, altered isoform balance – none of which was assessed in my experiments – could influence gene function.

The challenges which I discovered in determining disease mechanisms at the *KCNQ1* locus are indicative of the difficulties faced for all T2D associations which cannot be attributed to coding variation. Annotation of functional regulatory elements is limited, particularly in scarce islet samples, but fuller interrogation of the islet epigenome is underway. Genome-wide chromatin profiling in nine islet samples (Stitzel et al., 2010)

found that 35% of regions hypersensitive to DNase I digestion, marking regions of open chromatin, were islet-specific. No novel transcripts or alternative promoters were identified within known T2D loci, but six loci (*TCF7L2* [previously identified by Gaulton et al., 2010], *IGF2BP2*, *KCNQ1*, *WFS1*, *FTO* and *CDC123/CAMK1D*) contained DNase I hypersensitive sites. This suggests that the risk SNPs may play a role in gene expression, although only the *TCF7L2* and *WFS1* loci showed an effect on enhancer activity in MIN6 and HeLa cell lines. A genome-wide ChIP-seq analysis of histone marks in 12 islet samples (Bhandare et al., 2010) found that 18 out of 113 T2D SNPs (including suggestive as well as confirmed genome-wide significant loci) were close to H3K4me1 markers of transcriptional activation, representing potential enhancer elements which should be given priority for functional testing.

Further epigenomic analysis, including genome-wide methylation analysis, will be a major focus of attempts to uncover the T2D susceptibility mechanisms of regulatory variation in the coming years. However, these analyses, together with all investigations into β -cell disease phenotypes, are hindered by the scarcity of human islet and β -cell samples. The expression profiling which I performed will be largely superseded by whole-transcriptome islet eQTL data when it becomes available – but sample numbers are currently limiting. My work on the *KCNQ1* locus uses the largest sample set of adult human islets published to date, but all studies are limited by power to detect small effects.

Islet isolation and β -cell purification from cadavers is expensive and challenging, and the number of donors is unlikely to dramatically increase in the near future. However, recent development of viable human β -cell lines (McCluskey et al., 2011; Ravassard et al., 2011), and differentiation of β -cells from human induced pluripotent stem (iPS)

cells (Zhang et al., 2009), offer potential solutions to the problem of inadequate β -cell supply for disease mechanism analysis. The EndoC- β H1 cell line (Ravassard et al., 2011) expresses β -cell specific markers, secretes insulin in response to glucose and is stable for at least 80 passages. iPS cells are artificially derived from a non-pluripotent cell (often adult fibroblasts), and can give rise to any foetal or adult cell type (Takahashi et al., 2007; Yu et al., 2007). Pancreatic endocrine lineages (Zhang et al., 2009) produce ~25% insulin positive cells, with a glucose-induced insulin response and gene expression pattern similar to those of adult human islets. Insulin positive cells can also be derived from embryonic stem cells (Kroon et al., 2008), but iPS cells avoid ethical concerns surrounding the use of embryonic tissue.

Although further work is necessary to confidently define disease-relevant transcripts at proinsulin-associated loci, and to further determine the T2D susceptibility mechanism at *KCNQ1*, my work has highlighted interesting transcripts for proinsulin functional follow-up and provided preliminary evidence for a disease mechanism mediated through methylation and gene expression. I hope that forthcoming islet eQTL, mQTL and other epigenomic data, as well as new β -cell sources, will facilitate further disease mechanism analysis at these loci.

Disease mechanisms can only be determined once relevant genetic variation is identified, and the challenge of accounting for unexplained variance in T2D susceptibility is addressed in chapters five and six. The WTCCC genome-wide copy number variant analysis (Craddock et al., 2010) concluded that common CNVs are unlikely to account for a substantial proportion of genetic variance in T2D susceptibility. Chapter five details SNP and CNV genotyping for two T2D-associated variants, confirming that common CNVs are generally well-tagged by SNPs. Any

disease susceptibility which they confer is therefore likely to have already been indirectly explored (in European populations) through SNP GWAS. Nevertheless, common CNVs may be causal for some of the disease association signals currently attributed to SNP variation. Where SNPs and CNVs are in perfect LD in European populations, it may be possible to differentiate their effects through genotyping in different populations with weaker LD structures. Genotyping common CNVs in alternative populations, and rare CNVs which were not comprehensively interrogated by the WTCCC study, are both approaches which may use the genotyping methods which I optimised and compared.

Like copy number variation, the focus for further investigation into single nucleotide variation – at least in European populations – will be in the low and rare frequency ranges. Exome resequencing of candidate genes has already identified rare variants associated with T2D susceptibility in *MTNR1B* and *HNF4A* (Bonnetfond et al., 2012; Jafar-Mohammadi et al., 2011), but common SNP results suggest that many relevant variants might be in unsuspected genes, or in regulatory sequence. Finding them will demand extensive whole exome (for coding variation) and whole genome (for regulatory variation) sequencing. Information from one such whole genome sequencing programme, the 1000 genomes project (Consortium, 2010), was used in chapter six to impute rare variant genotypes for disease association analysis. My direct genotyping highlighted the inaccuracy of rare variant imputation based on current reference panels, and the consequent loss of association analysis power.

Improved imputation of rare variants will require larger reference panels, such as those being produced by the full 1000 Genomes project, the GoT2D programme (2650 control and T2D case samples) and the UK10K scheme (4000 controls with detailed

phenotypes). GoT2D sequenced 1325 control samples and 1325 cases with extreme T2D phenotypes, at low coverage across the whole genome and deep coverage across exomes. The samples were typed at 2.5M SNPs, with the intention of imputing variants across the entire frequency range into 28,000 samples. Direct whole genome and exome sequencing may also reveal genes relevant to T2D pathogenesis through burden tests of private mutations, which cannot be imputed or identified by individual association analysis.

The data presented in this thesis have addressed the two key challenges in understanding mechanisms of T2D susceptibility: identifying relevant genetic variation (chapters five and six) and determining its biological relevance (chapters three and four). My work has contributed to the first systematic analysis of common copy number variation and T2D risk, and highlighted the challenges to be overcome in imputing and genotyping rare single nucleotide variants. It has revealed candidates for functional analysis around proinsulin-associated loci, and made significant progress towards uncovering risk mechanisms at the *KCNQ1* locus. Together, my experiments 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 strategies to combat the global diabetes epidemic.

Appendices

References and manuscripts

References

- ADA, 2010. Diagnosis and classification of diabetes mellitus. *Diabetes Care*, 33 Suppl 1: S62-9.
- Ahnfelt-Ronne, J., Ravassard, P., Pardanaud-Glavieux, C., Scharfmann, R. and Serup, P., 2010. Mesenchymal bone morphogenetic protein signaling is required for normal pancreas development. *Diabetes*, 59(8): 1948-56.
- Akagi, K. et al., 1997. Cre-mediated somatic site-specific recombination in mice. *Nucleic Acids Res*, 25(9): 1766-73.
- Almgren, P. et al., 2011. Heritability and familiarity of type 2 diabetes and related quantitative traits in the Botnia Study. *Diabetologia*, 54(11): 2811-9.
- Altshuler, D. et al., 2000. The common PPARgamma Pro12Ala polymorphism is associated with decreased risk of type 2 diabetes. *Nat Genet*, 26(1): 76-80.
- Altshuler, D.M. et al., 2010. Integrating common and rare genetic variation in diverse human populations. *Nature*, 467(7311): 52-8.
- Andersson, E.A. et al., 2010. Type 2 diabetes risk alleles near ADCY5, CDKAL1 and HHEX-IDE are associated with reduced birthweight. *Diabetologia*, 53(9): 1908-16.
- Arcos-Burgos, M. et al., 2010. A common variant of the latrophilin 3 gene, LPHN3, confers susceptibility to ADHD and predicts effectiveness of stimulant medication. *Mol Psychiatry*, 15(11): 1053-66.
- Arima, T. et al., 2005. ZAC, LIT1 (KCNQ1OT1) and p57KIP2 (CDKN1C) are in an imprinted gene network that may play a role in Beckwith-Wiedemann syndrome. *Nucleic Acids Res*, 33(8): 2650-60.
- Arlt, M.F. et al., 2009. Replication stress induces genome-wide copy number changes in human cells that resemble polymorphic and pathogenic variants. *Am J Hum Genet*, 84(3): 339-50.
- Ashcroft, F.M., Harrison, D.E. and Ashcroft, S.J., 1984. Glucose induces closure of single potassium channels in isolated rat pancreatic beta-cells. *Nature*, 312(5993): 446-8.
- Assmann, A., Hinault, C. and Kulkarni, R.N., 2009. Growth factor control of pancreatic islet regeneration and function. *Pediatr Diabetes*, 10(1): 14-32.
- Babenko, A.P., 2008. A novel ABCC8 (SUR1)-dependent mechanism of metabolism-excitation uncoupling. *J Biol Chem*, 283(14): 8778-82.
- Babenko, A.P. et al., 2006. Activating mutations in the ABCC8 gene in neonatal diabetes mellitus. *N Engl J Med*, 355(5): 456-66.
- Bae, J.S. et al., 2011. The genetic effect of copy number variations on the risk of type 2 diabetes in a Korean population. *PLoS One*, 6(4): e19091.
- Beer, N.L. et al., 2009. 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(21): 4081-8.
- Bell, C.G. et al., 2010. Integrated genetic and epigenetic analysis identifies haplotype-specific methylation in the FTO type 2 diabetes and obesity susceptibility locus. *PLoS One*, 5(11): e14040.

- Bellanne-Chantelot, C. et al., 2005. Large genomic rearrangements in the hepatocyte nuclear factor-1beta (TCF2) gene are the most frequent cause of maturity-onset diabetes of the young type 5. *Diabetes*, 54(11): 3126-32.
- Bellocq, C. et al., 2004. Mutation in the KCNQ1 gene leading to the short QT-interval syndrome. *Circulation*, 109(20): 2394-7.
- Bhandare, R. et al., 2010. Genome-wide analysis of histone modifications in human pancreatic islets. *Genome Res*, 20(4): 428-33.
- Birney, E. et al., 2007. Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. *Nature*, 447(7146): 799-816.
- Bochukova, E.G. et al., 2010. Large, rare chromosomal deletions associated with severe early-onset obesity. *Nature*, 463(7281): 666-70.
- Boden, G. and Shulman, G.I., 2002. Free fatty acids in obesity and type 2 diabetes: defining their role in the development of insulin resistance and beta-cell dysfunction. *Eur J Clin Invest*, 32 Suppl 3: 14-23.
- Bonnefond, A. et al., 2012. Rare MTNR1B variants impairing melatonin receptor 1B function contribute to type 2 diabetes. *Nat Genet*, 44(3): 297-301.
- Bosco, D. et al., 2010. Unique arrangement of alpha- and beta-cells in human islets of Langerhans. *Diabetes*, 59(5): 1202-10.
- Bose, A. et al., 2002. Glucose transporter recycling in response to insulin is facilitated by myosin Myo1c. *Nature*, 420(6917): 821-4.
- Braun, M. et al., 2008. Voltage-gated ion channels in human pancreatic beta-cells: electrophysiological characterization and role in insulin secretion. *Diabetes*, 57(6): 1618-28.
- Byrne, M.M. et al., 1996. Altered insulin secretory responses to glucose in diabetic and nondiabetic subjects with mutations in the diabetes susceptibility gene MODY3 on chromosome 12. *Diabetes*, 45(11): 1503-10.
- Cabrera, O. et al., 2006. The unique cytoarchitecture of human pancreatic islets has implications for islet cell function. *Proc Natl Acad Sci U S A*, 103(7): 2334-9.
- Cai, L., Fritz, D., Stefanovic, L. and Stefanovic, B., 2010. Binding of LARP6 to the conserved 5' stem-loop regulates translation of mRNAs encoding type I collagen. *J Mol Biol*, 395(2): 309-26.
- Casparry, T., Cleary, M.A., Baker, C.C., Guan, X.J. and Tilghman, S.M., 1998. Multiple mechanisms regulate imprinting of the mouse distal chromosome 7 gene cluster. *Mol Cell Biol*, 18(6): 3466-74.
- Chen, K.S. et al., 1997. Homologous recombination of a flanking repeat gene cluster is a mechanism for a common contiguous gene deletion syndrome. *Nat Genet*, 17(2): 154-63.
- Chen, Y.H. et al., 2003. KCNQ1 gain-of-function mutation in familial atrial fibrillation. *Science*, 299(5604): 251-4.
- Cho, Y.S. et al., 2012. Meta-analysis of genome-wide association studies identifies eight new loci for type 2 diabetes in east Asians. *Nat Genet*, 44(1): 67-72.
- Chomczynski, P. and Sacchi, N., 1987. Single-step method of RNA isolation by acid guanidinium thiocyanate-phenol-chloroform extraction. *Anal Biochem*, 162(1): 156-9.
- Chung, W.S., Shin, C.H. and Stainier, D.Y., 2008. Bmp2 signaling regulates the hepatic versus pancreatic fate decision. *Dev Cell*, 15(5): 738-48.

- Chung, W.Y., Yuan, L., Feng, L., Hensle, T. and Tycko, B., 1996. Chromosome 11p15.5 regional imprinting: comparative analysis of KIP2 and H19 in human tissues and Wilms' tumors. *Hum Mol Genet*, 5(8): 1101-8.
- Church, C. et al., 2009. A mouse model for the metabolic effects of the human fat mass and obesity associated FTO gene. *PLoS Genet*, 5(8): e1000599.
- Church, C. et al., 2010. Overexpression of Fto leads to increased food intake and results in obesity. *Nat Genet*.
- Clark, A., Jones, L.C., de Koning, E., Hansen, B.C. and Matthews, D.R., 2001. Decreased insulin secretion in type 2 diabetes: a problem of cellular mass or function? *Diabetes*, 50 Suppl 1: S169-71.
- Cnop, M. et al., 2010. The long lifespan and low turnover of human islet beta cells estimated by mathematical modelling of lipofuscin accumulation. *Diabetologia*, 53(2): 321-30.
- Cohen, J.C. et al., 2004. Multiple rare alleles contribute to low plasma levels of HDL cholesterol. *Science*, 305(5685): 869-72.
- Conrad, D.F., Andrews, T.D., Carter, N.P., Hurles, M.E. and Pritchard, J.K., 2006. A high-resolution survey of deletion polymorphism in the human genome. *Nat Genet*, 38(1): 75-81.
- Conrad, D.F. et al., 2010. Origins and functional impact of copy number variation in the human genome. *Nature*, 464(7289): 704-12.
- Consortium, G.P., 2010. A map of human genome variation from population-scale sequencing. *Nature*, 467(7319): 1061-73.
- Cooper, G.M., Nickerson, D.A. and Eichler, E.E., 2007. Mutational and selective effects on copy-number variants in the human genome. *Nat Genet*, 39(7 Suppl): S22-9.
- Cooper, G.M., Zerr, T., Kidd, J.M., Eichler, E.E. and Nickerson, D.A., 2008. Systematic assessment of copy number variant detection via genome-wide SNP genotyping. *Nat Genet*, 40(10): 1199-203.
- Craddock, N. et al., 2010. Genome-wide association study of CNVs in 16,000 cases of eight common diseases and 3,000 shared controls. *Nature*, 464(7289): 713-20.
- D'Haene, B., Vandesompele, J. and Hellemans, J., 2010. Accurate and objective copy number profiling using real-time quantitative PCR. *Methods*, 50(4): 262-70.
- da Silva Xavier, G. et al., 2009. TCF7L2 regulates late events in insulin secretion from pancreatic islet beta-cells. *Diabetes*, 58(4): 894-905.
- Davies, S.M., 1994. Developmental regulation of genomic imprinting of the IGF2 gene in human liver. *Cancer Res*, 54(10): 2560-2.
- de Cid, R. et al., 2009. Deletion of the late cornified envelope LCE3B and LCE3C genes as a susceptibility factor for psoriasis. *Nat Genet*, 41(2): 211-5.
- Devaney, J.M. et al., 2009. Differences in fat and muscle mass associated with a functional human polymorphism in a post-transcriptional BMP2 gene regulatory element. *J Cell Biochem*, 107(6): 1073-82.
- DGV, 2010. Database of Genomic Variants
<http://dgvbeta.tcag.ca/dgv/app/home?ref=NCBI36/hg18>
- Diabetes UK, 2010. Key Statistics on Diabetes.
http://www.diabetes.org.uk/Documents/Reports/Diabetes_in_the_UK_2010.pdf

- Diaz-Meyer, N. et al., 2003. Silencing of CDKN1C (p57KIP2) is associated with hypomethylation at KvDMR1 in Beckwith-Wiedemann syndrome. *J Med Genet*, 40(11): 797-801.
- Dickson, S.P., Wang, K., Krantz, I., Hakonarson, H. and Goldstein, D.B., 2010. Rare variants create synthetic genome-wide associations. *PLoS Biol*, 8(1): e1000294.
- Dinneen, S., Gerich, J. and Rizza, R., 1992. Carbohydrate metabolism in non-insulin-dependent diabetes mellitus. *N Engl J Med*, 327(10): 707-13.
- Dupuis, J. et al., 2010. New genetic loci implicated in fasting glucose homeostasis and their impact on type 2 diabetes risk. *Nat Genet*, 42(2): 105-16.
- Emtage, P. et al., 2006. IGFL: A secreted family with conserved cysteine residues and similarities to the IGF superfamily. *Genomics*, 88(4): 513-20.
- Eyre-Walker, A., 2010. Evolution in health and medicine Sackler colloquium: Genetic architecture of a complex trait and its implications for fitness and genome-wide association studies. *Proc Natl Acad Sci U S A*, 107 Suppl 1: 1752-6.
- Fitzpatrick, G.V. et al., 2007. Allele-specific binding of CTCF to the multipartite imprinting control region KvDMR1. *Mol Cell Biol*, 27(7): 2636-47.
- Fitzpatrick, G.V., Soloway, P.D. and Higgins, M.J., 2002. Regional loss of imprinting and growth deficiency in mice with a targeted deletion of KvDMR1. *Nat Genet*, 32(3): 426-31.
- Flanagan, S.E. et al., 2007. Mutations in ATP-sensitive K⁺ channel genes cause transient neonatal diabetes and permanent diabetes in childhood or adulthood. *Diabetes*, 56(7): 1930-7.
- Florez, J.C. et al., 2004. Haplotype structure and genotype-phenotype correlations of the sulfonylurea receptor and the islet ATP-sensitive potassium channel gene region. *Diabetes*, 53(5): 1360-8.
- Frayling, T.M. et al., 2007. A common variant in the FTO gene is associated with body mass index and predisposes to childhood and adult obesity. *Science*, 316(5826): 889-94.
- Freathy, R.M. et al., 2009. Type 2 diabetes risk alleles are associated with reduced size at birth. *Diabetes*, 58(6): 1428-33.
- Freathy, R.M. et al., 2010. Variants in ADCY5 and near CCNL1 are associated with fetal growth and birth weight. *Nat Genet*, 42(5): 430-5.
- Freathy, R.M. et al., 2007. Type 2 diabetes TCF7L2 risk genotypes alter birth weight: a study of 24,053 individuals. *Am J Hum Genet*, 80(6): 1150-61.
- Fritsche, A. et al., 2002. Relationships among age, proinsulin conversion, and beta-cell function in nondiabetic humans. *Diabetes*, 51 Suppl 1: S234-9.
- Froguel, P. et al., 1992. Close linkage of glucokinase locus on chromosome 7p to early-onset non-insulin-dependent diabetes mellitus. *Nature*, 356(6365): 162-4.
- Fukada, T. et al., 2008. The zinc transporter SLC39A13/ZIP13 is required for connective tissue development; its involvement in BMP/TGF-beta signaling pathways. *PLoS One*, 3(11): e3642.
- Gardner, R.J. et al., 2000. An imprinted locus associated with transient neonatal diabetes mellitus. *Hum Mol Genet*, 9(4): 589-96.
- Gardner, R.J., Robinson, D.O., Lamont, L., Shield, J.P. and Temple, I.K., 1998. Paternal uniparental disomy of chromosome 6 and transient neonatal diabetes mellitus. *Clin Genet*, 54(6): 522-5.

- Gaulton, K.J. et al., 2010. A map of open chromatin in human pancreatic islets. *Nat Genet*, 42(3): 255-9.
- George, S. et al., 2004. A family with severe insulin resistance and diabetes due to a mutation in AKT2. *Science*, 304(5675): 1325-8.
- Giurgea, I. et al., 2006. The Knudson's two-hit model and timing of somatic mutation may account for the phenotypic diversity of focal congenital hyperinsulinism. *J Clin Endocrinol Metab*, 91(10): 4118-23.
- Glaser, B. et al., 1998. Familial hyperinsulinism caused by an activating glucokinase mutation. *N Engl J Med*, 338(4): 226-30.
- Gloyn, A.L. 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(18): 1838-49.
- Gloyn, A.L. et al., 2005. Relapsing diabetes can result from moderately activating mutations in KCNJ11. *Hum Mol Genet*, 14(7): 925-34.
- Gloyn, A.L. et al., 2003. Large-scale association studies of variants in genes encoding the pancreatic beta-cell KATP channel subunits Kir6.2 (KCNJ11) and SUR1 (ABCC8) confirm that the KCNJ11 E23K variant is associated with type 2 diabetes. *Diabetes*, 52(2): 568-72.
- Gorden, P., Hendricks, C.M. and Roth, J., 1974. Circulating proinsulin-like component in man: increased proportion in hypoinsulinemic states. *Diabetologia*, 10(5): 469-74.
- Gordois, A., Scuffham, P., Shearer, A., Oglesby, A. and Tobian, J.A., 2003. The health care costs of diabetic peripheral neuropathy in the US. *Diabetes Care*, 26(6): 1790-5.
- Gorlov, I.P., Gorlova, O.Y., Frazier, M.L., Spitz, M.R. and Amos, C.I., 2011. Evolutionary evidence of the effect of rare variants on disease etiology. *Clin Genet*, 79(3): 199-206.
- Grant, S.F. et al., 2006. Variant of transcription factor 7-like 2 (TCF7L2) gene confers risk of type 2 diabetes. *Nat Genet*, 38(3): 320-3.
- Grarup, N. et al., 2011. The diabetogenic VPS13C/C2CD4A/C2CD4B rs7172432 variant impairs glucose-stimulated insulin response in 5,722 non-diabetic Danish individuals. *Diabetologia*, 54(4): 789-94.
- Grigoriev, I. et al., 2007. Rab6 regulates transport and targeting of exocytotic carriers. *Dev Cell*, 13(2): 305-14.
- Grunnet, L.G. et al., 2009. Regulation and function of FTO mRNA expression in human skeletal muscle and subcutaneous adipose tissue. *Diabetes*, 58(10): 2402-8.
- Hales, C.N. et al., 1991. Fetal and infant growth and impaired glucose tolerance at age 64. *BMJ*, 303(6809): 1019-22.
- Hamming, K.S. et al., 2009. 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(10): 2419-24.
- Hanson, R.L. et al., 1998. An autosomal genomic scan for loci linked to type II diabetes mellitus and body-mass index in Pima Indians. *Am J Hum Genet*, 63(4): 1130-8.
- Hassanein, M.T. et al., 2010. Fine mapping of the association with obesity at the FTO locus in African-derived populations. *Hum Mol Genet*, 19(14): 2907-16.
- Hastings, P.J., Lupski, J.R., Rosenberg, S.M. and Ira, G., 2009. Mechanisms of change in gene copy number. *Nat Rev Genet*, 10(8): 551-64.

- Hatada, I. et al., 1996. An imprinted gene p57KIP2 is mutated in Beckwith-Wiedemann syndrome. *Nat Genet*, 14(2): 171-3.
- Hattersley, A.T. and Tooke, J.E., 1999. The fetal insulin hypothesis: an alternative explanation of the association of low birthweight with diabetes and vascular disease. *Lancet*, 353(9166): 1789-92.
- Hedley, P.L. et al., 2009. The genetic basis of long QT and short QT syndromes: a mutation update. *Hum Mutat*, 30(11): 1486-511.
- Heijmans, B.T. et al., 2008. Persistent epigenetic differences associated with prenatal exposure to famine in humans. *Proc Natl Acad Sci U S A*, 105(44): 17046-9.
- Helgadottir, A. et al., 2006. A variant of the gene encoding leukotriene A4 hydrolase confers ethnicity-specific risk of myocardial infarction. *Nat Genet*, 38(1): 68-74.
- Helgason, A. et al., 2007. Refining the impact of TCF7L2 gene variants on type 2 diabetes and adaptive evolution. *Nat Genet*, 39(2): 218-25.
- Henrichsen, C.N. et al., 2009. Segmental copy number variation shapes tissue transcriptomes. *Nat Genet*, 41(4): 424-9.
- Holland, P.M., Abramson, R.D., Watson, R. and Gelfand, D.H., 1991. Detection of specific polymerase chain reaction product by utilizing the 5'----3' exonuclease activity of *Thermus aquaticus* DNA polymerase. *Proc Natl Acad Sci U S A*, 88(16): 7276-80.
- Holmkvist, J. et al., 2009. The type 2 diabetes associated minor allele of rs2237895 KCNQ1 associates with reduced insulin release following an oral glucose load. *PLoS One*, 4(6): e5872.
- Horike, S. et al., 2000. Targeted disruption of the human LIT1 locus defines a putative imprinting control element playing an essential role in Beckwith-Wiedemann syndrome. *Hum Mol Genet*, 9(14): 2075-83.
- Hotamisligil, G.S., Arner, P., Caro, J.F., Atkinson, R.L. and Spiegelman, B.M., 1995. Increased adipose tissue expression of tumor necrosis factor- α in human obesity and insulin resistance. *J Clin Invest*, 95(5): 2409-15.
- Howie, B., Marchini, J. and Stephens, M., 2011. Genotype imputation with thousands of genomes. *G3 (Bethesda)*, 1(6): 457-70.
- Howie, B.N., Donnelly, P. and Marchini, J., 2009. A flexible and accurate genotype imputation method for the next generation of genome-wide association studies. *PLoS Genet*, 5(6): e1000529.
- IDF, 2012. International Diabetes Federation Diabetes Atlas.
<http://www.idf.org/diabetesatlas>
- Ijuin, T. and Takenawa, T., 2003. SKIP negatively regulates insulin-induced GLUT4 translocation and membrane ruffle formation. *Mol Cell Biol*, 23(4): 1209-20.
- IKMC, 2012. International Knockout Mouse Consortium.
<http://www.knockoutmouse.org>
- Ingalls, A.M., Dickie, M.M. and Snell, G.D., 1950. Obese, a new mutation in the house mouse. *J Hered*, 41(12): 317-8.
- Ingelsson, E. et al., 2010. Detailed physiologic characterization reveals diverse mechanisms for novel genetic Loci regulating glucose and insulin metabolism in humans. *Diabetes*, 59(5): 1266-75.
- Ionita-Laza, I., Lange, C. and N, M.L., 2009. Estimating the number of unseen variants in the human genome. *Proc Natl Acad Sci U S A*, 106(13): 5008-13.

- Jacquemont, S. et al., 2011. Mirror extreme BMI phenotypes associated with gene dosage at the chromosome 16p11.2 locus. *Nature*, 478(7367): 97-102.
- Jafar-Mohammadi, B. et al., 2011. A role for coding functional variants in HNF4A in type 2 diabetes susceptibility. *Diabetologia*, 54(1): 111-9.
- Janssens, A.C., Gwinn, M., Khoury, M.J. and Subramonia-Iyer, S., 2006. Does genetic testing really improve the prediction of future type 2 diabetes? *PLoS Med*, 3(2): e114; author reply e127.
- Jeon, J.P. et al., 2010. Copy number variation at leptin receptor gene locus associated with metabolic traits and the risk of type 2 diabetes mellitus. *BMC Genomics*, 11: 426.
- Jespersen, T., Grunnet, M. and Olesen, S.P., 2005. The KCNQ1 potassium channel: from gene to physiological function. *Physiology (Bethesda)*, 20: 408-16.
- Jonsson, A. et al., 2009. A variant in the KCNQ1 gene predicts future type 2 diabetes and mediates impaired insulin secretion. *Diabetes*, 58(10): 2409-13.
- Jostins, L., Morley, K.I. and Barrett, J.C., 2011. Imputation of low-frequency variants using the HapMap3 benefits from large, diverse reference sets. *Eur J Hum Genet*, 19(6): 662-6.
- Justice, M.J., Noveroske, J.K., Weber, J.S., Zheng, B. and Bradley, A., 1999. Mouse ENU mutagenesis. *Hum Mol Genet*, 8(10): 1955-63.
- Kahn, S.E. et al., 1995. Proinsulin as a marker for the development of NIDDM in Japanese-American men. *Diabetes*, 44(2): 173-9.
- Kanduri, C. et al., 2002. A differentially methylated imprinting control region within the Kcnq1 locus harbors a methylation-sensitive chromatin insulator. *J Biol Chem*, 277(20): 18106-10.
- Kapoor, R.R. et al., 2011. Hyperinsulinaemic hypoglycaemia and diabetes mellitus due to dominant ABCC8/KCNJ11 mutations. *Diabetologia*, 54(10): 2575-83.
- Kassem, S.A. et al., 2001. p57(KIP2) expression in normal islet cells and in hyperinsulinism of infancy. *Diabetes*, 50(12): 2763-9.
- Kessler, A. et al., 2000. Rab11 is associated with GLUT4-containing vesicles and redistributes in response to insulin. *Diabetologia*, 43(12): 1518-27.
- Kim, Y.D. et al., 2008. Metformin inhibits hepatic gluconeogenesis through AMP-activated protein kinase-dependent regulation of the orphan nuclear receptor SHP. *Diabetes*, 57(2): 306-14.
- Kirchhoff, K. et al., 2008. Polymorphisms in the TCF7L2, CDKAL1 and SLC30A8 genes are associated with impaired proinsulin conversion. *Diabetologia*, 51(4): 597-601.
- Köbberling, J.T., H, 1982. Empirical risk figures for first degree relatives of non-insulin-dependent diabetics. In: J.T. Köbberling, R. (Editor), *The Genetics of Diabetes Mellitus*. Academic Press, London, pp. 201-209.
- Kong, A. et al., 2009. Parental origin of sequence variants associated with complex diseases. *Nature*, 462(7275): 868-74.
- Kooner, J.S. et al., 2011. Genome-wide association study in individuals of South Asian ancestry identifies six new type 2 diabetes susceptibility loci. *Nat Genet*, 43(10): 984-9.
- Krilov, L., Nguyen, A., Miyazaki, T., Unson, C.G. and Bouscarel, B., 2008. Glucagon receptor recycling: role of carboxyl terminus, beta-arrestins, and cytoskeleton. *Am J Physiol Cell Physiol*, 295(5): C1230-7.

- Krishnamurthy, J. et al., 2006. p16INK4a induces an age-dependent decline in islet regenerative potential. *Nature*, 443(7110): 453-7.
- Kroon, E. et al., 2008. Pancreatic endoderm derived from human embryonic stem cells generates glucose-responsive insulin-secreting cells in vivo. *Nat Biotechnol*, 26(4): 443-52.
- Kulkarni, R.N. et al., 1999. Tissue-specific knockout of the insulin receptor in pancreatic beta cells creates an insulin secretory defect similar to that in type 2 diabetes. *Cell*, 96(3): 329-39.
- Kulyte, A. et al., 2011. MTCH2 in human white adipose tissue and obesity. *J Clin Endocrinol Metab*, 96(10): E1661-5.
- Kumps, C. et al., 2010. Multiplex Amplicon Quantification (MAQ), a fast and efficient method for the simultaneous detection of copy number alterations in neuroblastoma. *BMC Genomics*, 11: 298.
- Lee, M.P. et al., 1999. Loss of imprinting of a paternally expressed transcript, with antisense orientation to KVLQT1, occurs frequently in Beckwith-Wiedemann syndrome and is independent of insulin-like growth factor II imprinting. *Proc Natl Acad Sci U S A*, 96(9): 5203-8.
- Lee, S., Kasif, S., Weng, Z. and Cantor, C.R., 2008. Quantitative analysis of single nucleotide polymorphisms within copy number variation. *PLoS One*, 3(12): e3906.
- Lemaire, K. et al., 2009. Insulin crystallization depends on zinc transporter ZnT8 expression, but is not required for normal glucose homeostasis in mice. *Proc Natl Acad Sci U S A*, 106(35): 14872-7.
- Lettre, G. et al., 2011. Genome-wide association study of coronary heart disease and its risk factors in 8,090 African Americans: the NHLBI CARE Project. *PLoS Genet*, 7(2): e1001300.
- Li, L. et al., 2011. Performance of genotype imputation for rare variants identified in exons and flanking regions of genes. *PLoS One*, 6(9): e24945.
- Li, Y. et al., 2010. LRP4 mutations alter Wnt/beta-catenin signaling and cause limb and kidney malformations in Cenani-Lenz syndrome. *Am J Hum Genet*, 86(5): 696-706.
- Lindgren, C.M. et al., 2002. Genomewide search for type 2 diabetes mellitus susceptibility loci in Finnish families: the Botnia study. *Am J Hum Genet*, 70(2): 509-16.
- Ling, C. et al., 2008. Epigenetic regulation of PPARGC1A in human type 2 diabetic islets and effect on insulin secretion. *Diabetologia*, 51(4): 615-22.
- Liu, Y. et al., 2009. Variants in KCNQ1 are associated with susceptibility to type 2 diabetes in the population of mainland China. *Diabetologia*, 52(7): 1315-21.
- Longo, N. et al., 1994. Impaired growth in Rabson-Mendenhall syndrome: lack of effect of growth hormone and insulin-like growth factor-I. *J Clin Endocrinol Metab*, 79(3): 799-805.
- Loos, R.J. et al., 2007. TCF7L2 polymorphisms modulate proinsulin levels and beta-cell function in a British European population. *Diabetes*, 56(7): 1943-7.
- Lupski, J.R., Belmont, J.W., Boerwinkle, E. and Gibbs, R.A., 2011. Clan genomics and the complex architecture of human disease. *Cell*, 147(1): 32-43.
- Lyssenko, V. et al., 2007. Mechanisms by which common variants in the TCF7L2 gene increase risk of type 2 diabetes. *J Clin Invest*, 117(8): 2155-63.

- MacDonald, P.E. and Wheeler, M.B., 2003. Voltage-dependent K(+) channels in pancreatic beta cells: role, regulation and potential as therapeutic targets. *Diabetologia*, 46(8): 1046-62.
- Mackay, D.J. et al., 2008. Hypomethylation of multiple imprinted loci in individuals with transient neonatal diabetes is associated with mutations in ZFP57. *Nat Genet*, 40(8): 949-51.
- Mackay, D.J. and Temple, I.K., 2010. Transient neonatal diabetes mellitus type 1. *Am J Med Genet C Semin Med Genet*, 154C(3): 335-42.
- Magi, R., Kumar, A. and Morris, A.P., 2011. Assessing the impact of missing genotype data in rare variant association analysis. *BMC Proc*, 5 Suppl 9: S107.
- Maher, E.R. and Reik, W., 2000. Beckwith-Wiedemann syndrome: imprinting in clusters revisited. *J Clin Invest*, 105(3): 247-52.
- Mancini-DiNardo, D., Steele, S.J., Ingram, R.S. and Tilghman, S.M., 2003. A differentially methylated region within the gene Kcnq1 functions as an imprinted promoter and silencer. *Hum Mol Genet*, 12(3): 283-94.
- Manolio, T.A. et al., 2009. Finding the missing heritability of complex diseases. *Nature*, 461(7265): 747-53.
- Matschinsky, F.M., 2002. Regulation of pancreatic beta-cell glucokinase: from basics to therapeutics. *Diabetes*, 51 Suppl 3: S394-404.
- McCarroll, S.A. et al., 2006. Common deletion polymorphisms in the human genome. *Nat Genet*, 38(1): 86-92.
- McCarroll, S.A. et al., 2008a. Deletion polymorphism upstream of IRGM associated with altered IRGM expression and Crohn's disease. *Nat Genet*, 40(9): 1107-12.
- McCarroll, S.A. et al., 2008b. Integrated detection and population-genetic analysis of SNPs and copy number variation. *Nat Genet*, 40(10): 1166-74.
- McCarthy, M.I., 2003. Growing evidence for diabetes susceptibility genes from genome scan data. *Curr Diab Rep*, 3(2): 159-67.
- McCluskey, J.T. et al., 2011. Development and functional characterization of insulin-releasing human pancreatic beta cell lines produced by electrofusion. *J Biol Chem*, 286(25): 21982-92.
- McCulloch, L.J. 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(4): 648-53.
- McGuigan, F. et al., 2008. Variation in the bone morphogenetic protein-2 gene: effects on fat and lean body mass in young and elderly women. *Eur J Endocrinol*, 158(5): 661-8.
- McIntyre, N., Holdsworth, C.D. and Turner, D.S., 1964. New Interpretation of Oral Glucose Tolerance. *Lancet*, 2(7349): 20-1.
- Merikangas, K.R. and Risch, N., 2003. Genomic priorities and public health. *Science*, 302(5645): 599-601.
- Meyre, D. et al., 2010. Prevalence of loss-of-function FTO mutations in lean and obese individuals. *Diabetes*, 59(1): 311-8.
- Mohammad, F., Mondal, T., Guseva, N., Pandey, G.K. and Kanduri, C., 2010. Kcnq1ot1 noncoding RNA mediates transcriptional gene silencing by interacting with Dnmt1. *Development*, 137(15): 2493-9.
- MolPAGE, 2012. MolPAGE project details. <http://www.ebi.ac.uk/fg/molpage>

- Monk, D. et al., 2006. Limited evolutionary conservation of imprinting in the human placenta. *Proc Natl Acad Sci U S A*, 103(17): 6623-8.
- Morita, H., Wu, J. and Zipes, D.P., 2008. The QT syndromes: long and short. *Lancet*, 372(9640): 750-63.
- Moritani, M. et al., 2005. Hypoplasia of endocrine and exocrine pancreas in homozygous transgenic TGF-beta1. *Mol Cell Endocrinol*, 229(1-2): 175-84.
- Morris, A.P. and Zeggini, E., 2010. An evaluation of statistical approaches to rare variant analysis in genetic association studies. *Genet Epidemiol*, 34(2): 188-93.
- Musso, C. et al., 2004. Clinical course of genetic diseases of the insulin receptor (type A and Rabson-Mendenhall syndromes): a 30-year prospective. *Medicine (Baltimore)*, 83(4): 209-22.
- Myles, S., Davison, D., Barrett, J., Stoneking, M. and Timpson, N., 2008. Worldwide population differentiation at disease-associated SNPs. *BMC Med Genomics*, 1: 22.
- Nagano, T. and Fraser, P., 2009. Emerging similarities in epigenetic gene silencing by long noncoding RNAs. *Mamm Genome*, 20(9-10): 557-62.
- Nakamura, M. and Yamada, K., 1967. Studies on a diabetic (KK) strain of the mouse. *Diabetologia*, 3(2): 212-21.
- Neel, J.V., 1962. Diabetes mellitus: a "thrifty" genotype rendered detrimental by "progress"? *Am J Hum Genet*, 14: 353-62.
- Nejentsev, S., Walker, N., Riches, D., Egholm, M. and Todd, J.A., 2009. Rare variants of IFIH1, a gene implicated in antiviral responses, protect against type 1 diabetes. *Science*, 324(5925): 387-9.
- Nicolson, T.J. et al., 2009. Insulin storage and glucose homeostasis in mice null for the granule zinc transporter ZnT8 and studies of the type 2 diabetes-associated variants. *Diabetes*, 58(9): 2070-83.
- Niederer, H.A. et al., 2010. Copy number, linkage disequilibrium and disease association in the FCGR locus. *Hum Mol Genet*, 19(16): 3282-94.
- Nielsen, E.M. et al., 2005. Studies of variations of the cyclin-dependent kinase inhibitor 1C and the cyclin-dependent kinase 4 genes in relation to type 2 diabetes mellitus and related quantitative traits. *J Mol Med*, 83(5): 353-61.
- Niemitz, E.L. et al., 2004. Microdeletion of LIT1 in familial Beckwith-Wiedemann syndrome. *Am J Hum Genet*, 75(5): 844-9.
- Nissen, S.E. and Wolski, K., 2007. Effect of rosiglitazone on the risk of myocardial infarction and death from cardiovascular causes. *N Engl J Med*, 356(24): 2457-71.
- O'Leary, D.A. et al., 2010. HTS-Compatible Patient-Derived Cell-Based Assay to Identify Small Molecule Modulators of Aberrant Splicing in Myotonic Dystrophy Type 1. *Curr Chem Genomics*, 4: 9-18.
- O'Moore-Sullivan, T.M. and Prins, J.B., 2002. Thiazolidinediones and type 2 diabetes: new drugs for an old disease. *Med J Aust*, 176(8): 381-6.
- Olayioye, M.A. et al., 2007. Phosphorylation of StarD10 on serine 284 by casein kinase II modulates its lipid transfer activity. *J Biol Chem*, 282(31): 22492-8.
- Orho-Melander, M. et al., 2008. Common missense variant in the glucokinase regulatory protein gene is associated with increased plasma triglyceride and C-reactive protein but lower fasting glucose concentrations. *Diabetes*, 57(11): 3112-21.

- Pagnamenta, A.T. et al., 2009. A 15q13.3 microdeletion segregating with autism. *Eur J Hum Genet*, 17(5): 687-92.
- Pan, Q. et al., 2010. KCNQ1 loss-of-function mutation impairs gastric acid secretion in mice. *Mol Biol Rep*, 37(3): 1329-33.
- Pawitan, Y., Seng, K.C. and Magnusson, P.K., 2009. How many genetic variants remain to be discovered? *PLoS One*, 4(12): e7969.
- Pearson, E.R. et al., 2003. Genetic cause of hyperglycaemia and response to treatment in diabetes. *Lancet*, 362(9392): 1275-81.
- Perl, S. et al., 2010. Significant human beta-cell turnover is limited to the first three decades of life as determined by in vivo thymidine analog incorporation and radiocarbon dating. *J Clin Endocrinol Metab*, 95(10): E234-9.
- Perry, J.R. and Frayling, T.M., 2008. New gene variants alter type 2 diabetes risk predominantly through reduced beta-cell function. *Curr Opin Clin Nutr Metab Care*, 11(4): 371-7.
- Pfaffl, M.W., 2001. A new mathematical model for relative quantification in real-time RT-PCR. *Nucleic Acids Res*, 29(9): e45.
- Postic, C. et al., 1999. Dual roles for glucokinase in glucose homeostasis as determined by liver and pancreatic beta cell-specific gene knock-outs using Cre recombinase. *J Biol Chem*, 274(1): 305-15.
- Poulsen, P., Kyvik, K.O., Vaag, A. and Beck-Nielsen, H., 1999. Heritability of type II (non-insulin-dependent) diabetes mellitus and abnormal glucose tolerance--a population-based twin study. *Diabetologia*, 42(2): 139-45.
- Prokopenko, I. et al., 2009. Linkage disequilibrium mapping of the replicated type 2 diabetes linkage signal on chromosome 1q. *Diabetes*, 58(7): 1704-9.
- Psiachou, H. et al., 1993. Leprechaunism and homozygous nonsense mutation in the insulin receptor gene. *Lancet*, 342(8876): 924.
- Ramagopalan, S.V. et al., 2011. Rare variants in the CYP27B1 gene are associated with multiple sclerosis. *Ann Neurol*, 70(6): 881-6.
- Ravassard, P. et al., 2011. A genetically engineered human pancreatic beta cell line exhibiting glucose-inducible insulin secretion. *J Clin Invest*, 121(9): 3589-97.
- Redon, R. et al., 2006. Global variation in copy number in the human genome. *Nature*, 444(7118): 444-54.
- Redrup, L. et al., 2009. The long noncoding RNA Kcnq1ot1 organises a lineage-specific nuclear domain for epigenetic gene silencing. *Development*, 136(4): 525-30.
- Rees, M.G. et al., 2012. Cellular characterisation of the GCKR P446L variant associated with type 2 diabetes risk. *Diabetologia*.
- Reich, D.E. and Lander, E.S., 2001. On the allelic spectrum of human disease. *Trends Genet*, 17(9): 502-10.
- Reynisdottir, I. et al., 2003. Localization of a susceptibility gene for type 2 diabetes to chromosome 5q34-q35.2. *Am J Hum Genet*, 73(2): 323-35.
- Ripsin, C.M., Kang, H. and Urban, R.J., 2009. Management of blood glucose in type 2 diabetes mellitus. *Am Fam Physician*, 79(1): 29-36.
- Rivas, A. and Francis, H.W., 2005. Inner ear abnormalities in a Kcnq1 (Kvlqt1) knockout mouse: a model of Jervell and Lange-Nielsen syndrome. *Otol Neurotol*, 26(3): 415-24.

- Rivas, M.A. et al., 2011. Deep resequencing of GWAS loci identifies independent rare variants associated with inflammatory bowel disease. *Nat Genet*, 43(11): 1066-73.
- Roder, M.E., Porte, D., Jr., Schwartz, R.S. and Kahn, S.E., 1998. Disproportionately elevated proinsulin levels reflect the degree of impaired B cell secretory capacity in patients with noninsulin-dependent diabetes mellitus. *J Clin Endocrinol Metab*, 83(2): 604-8.
- Rosenblum, B.B. et al., 1997. New dye-labeled terminators for improved DNA sequencing patterns. *Nucleic Acids Res*, 25(22): 4500-4.
- Rosengren, A.H. et al., 2012. Reduced Insulin Exocytosis in Human Pancreatic beta-cells With Gene Variants Linked to Type 2 Diabetes. *Diabetes*.
- Saad, M.F. et al., 1990. Disproportionately elevated proinsulin in Pima Indians with noninsulin-dependent diabetes mellitus. *J Clin Endocrinol Metab*, 70(5): 1247-53.
- Sabbah, E. et al., 2000. Genetic, autoimmune, and clinical characteristics of childhood- and adult-onset type 1 diabetes. *Diabetes Care*, 23(9): 1326-32.
- Sandovici, I. et al., 2011. Maternal diet and aging alter the epigenetic control of a promoter-enhancer interaction at the *Hnf4a* gene in rat pancreatic islets. *Proc Natl Acad Sci U S A*, 108(13): 5449-54.
- Sanger, F., Nicklen, S. and Coulson, A.R., 1977. DNA sequencing with chain-terminating inhibitors. *Proc Natl Acad Sci U S A*, 74(12): 5463-7.
- Saxena, R. et al., 2012. Large-Scale Gene-Centric Meta-Analysis across 39 Studies Identifies Type 2 Diabetes Loci. *Am J Hum Genet*, 90(3): 410-425.
- Saxena, R. et al., 2007. Genome-wide association analysis identifies loci for type 2 diabetes and triglyceride levels. *Science*, 316(5829): 1331-6.
- Scott, L.J. et al., 2007. A genome-wide association study of type 2 diabetes in Finns detects multiple susceptibility variants. *Science*, 316(5829): 1341-5.
- Sebat, J. et al., 2007. Strong association of de novo copy number mutations with autism. *Science*, 316(5823): 445-9.
- Sebat, J. et al., 2004. Large-scale copy number polymorphism in the human genome. *Science*, 305(5683): 525-8.
- Sempoux, C. et al., 1998. Pancreatic B-cell proliferation in persistent hyperinsulinemic hypoglycemia of infancy: an immunohistochemical study of 18 cases. *Mod Pathol*, 11(5): 444-9.
- Sharp, A.J. et al., 2005. Segmental duplications and copy-number variation in the human genome. *Am J Hum Genet*, 77(1): 78-88.
- Shibasaki, T. et al., 2007. Essential role of *Epac2/Rap1* signaling in regulation of insulin granule dynamics by cAMP. *Proc Natl Acad Sci U S A*, 104(49): 19333-8.
- Shulman, G.I., 2000. Cellular mechanisms of insulin resistance. *J Clin Invest*, 106(2): 171-6.
- Silva, J.P. and Ushkaryov, Y.A., 2010. The latrophilins, "split-personality" receptors. *Adv Exp Med Biol*, 706: 59-75.
- Simmons, R.A., 2007. Developmental origins of diabetes: the role of epigenetic mechanisms. *Curr Opin Endocrinol Diabetes Obes*, 14(1): 13-6.
- Sladek, R. et al., 2007. A genome-wide association study identifies novel risk loci for type 2 diabetes. *Nature*, 445(7130): 881-5.

- Small, K.S. et al., 2011. Identification of an imprinted master trans regulator at the KLF14 locus related to multiple metabolic phenotypes. *Nat Genet*, 43(6): 561-4.
- Smilnich, N.J. et al., 1999. A maternally methylated CpG island in KvLQT1 is associated with an antisense paternal transcript and loss of imprinting in Beckwith-Wiedemann syndrome. *Proc Natl Acad Sci U S A*, 96(14): 8064-9.
- Smith, L.M. et al., 1986. Fluorescence detection in automated DNA sequence analysis. *Nature*, 321(6071): 674-9.
- Sottile, V. and Seuwen, K., 2000. Bone morphogenetic protein-2 stimulates adipogenic differentiation of mesenchymal precursor cells in synergy with BRL 49653 (rosiglitazone). *FEBS Lett*, 475(3): 201-4.
- Speliotes, E.K. et al., 2010. Association analyses of 249,796 individuals reveal 18 new loci associated with body mass index. *Nat Genet*, 42(11): 937-48.
- Stamler, J., Vaccaro, O., Neaton, J.D. and Wentworth, D., 1993. Diabetes, other risk factors, and 12-yr cardiovascular mortality for men screened in the Multiple Risk Factor Intervention Trial. *Diabetes Care*, 16(2): 434-44.
- Steck, A.K. and Winter, W.E., 2011. Review on monogenic diabetes. *Curr Opin Endocrinol Diabetes Obes*, 18(4): 252-8.
- Stefansson, H. et al., 2008. Large recurrent microdeletions associated with schizophrenia. *Nature*, 455(7210): 232-6.
- Steinthorsdottir, V. et al., 2007. A variant in CDKAL1 influences insulin response and risk of type 2 diabetes. *Nat Genet*, 39(6): 770-5.
- Stitzel, M.L. et al., 2010. Global epigenomic analysis of primary human pancreatic islets provides insights into type 2 diabetes susceptibility loci. *Cell Metab*, 12(5): 443-55.
- Stolerman, E.S. et al., 2009. TCF7L2 variants are associated with increased proinsulin/insulin ratios but not obesity traits in the Framingham Heart Study. *Diabetologia*, 52(4): 614-20.
- Strawbridge, R.J. et al., 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(10): 2624-34.
- Stride, A. et al., 2002. The genetic abnormality in the beta cell determines the response to an oral glucose load. *Diabetologia*, 45(3): 427-35.
- Stumvoll, M., Goldstein, B.J. and van Haeften, T.W., 2008. Type 2 diabetes: pathogenesis and treatment. *Lancet*, 371(9631): 2153-6.
- Stumvoll, M., Tataranni, P.A., Stefan, N., Vozarova, B. and Bogardus, C., 2003. Glucose allostasis. *Diabetes*, 52(4): 903-9.
- Takahashi, K. et al., 2007. Induction of pluripotent stem cells from adult human fibroblasts by defined factors. *Cell*, 131(5): 861-72.
- Tan, J.T. et al., 2010. Polymorphisms identified through genome-wide association studies and their associations with type 2 diabetes in Chinese, Malays, and Asian-Indians in Singapore. *J Clin Endocrinol Metab*, 95(1): 390-7.
- Tan, J.T. et al., 2009. Genetic variation in KCNQ1 associates with fasting glucose and beta-cell function: a study of 3,734 subjects comprising three ethnicities living in Singapore. *Diabetes*, 58(6): 1445-9.
- Thakur, N. et al., 2004. An antisense RNA regulates the bidirectional silencing property of the Kcnq1 imprinting control region. *Mol Cell Biol*, 24(18): 7855-62.

- Thanabalasingham, G. and Owen, K.R., 2011. Diagnosis and management of maturity onset diabetes of the young (MODY). *BMJ*, 343: d6044.
- Thomas, P., Ye, Y. and Lightner, E., 1996. Mutation of the pancreatic islet inward rectifier Kir6.2 also leads to familial persistent hyperinsulinemic hypoglycemia of infancy. *Hum Mol Genet*, 5(11): 1809-12.
- Thomas, P.M. et al., 1995. Mutations in the sulfonylurea receptor gene in familial persistent hyperinsulinemic hypoglycemia of infancy. *Science*, 268(5209): 426-9.
- Torgerson, D.G. et al., 2012. Resequencing candidate genes implicates rare variants in asthma susceptibility. *Am J Hum Genet*, 90(2): 273-81.
- Trumper, A. et al., 2001. Glucose-dependent insulinotropic polypeptide is a growth factor for beta (INS-1) cells by pleiotropic signaling. *Mol Endocrinol*, 15(9): 1559-70.
- Tsai, F.J. et al., 2010. A genome-wide association study identifies susceptibility variants for type 2 diabetes in Han Chinese. *PLoS Genet*, 6(2): e1000847.
- Tuzun, E. et al., 2005. Fine-scale structural variation of the human genome. *Nat Genet*, 37(7): 727-32.
- Ullrich, S. et al., 2005. Effects of I(Ks) channel inhibitors in insulin-secreting INS-1 cells. *Pflugers Arch*, 451(3): 428-36.
- Unoki, H. et al., 2008. SNPs in KCNQ1 are associated with susceptibility to type 2 diabetes in East Asian and European populations. *Nat Genet*, 40(9): 1098-102.
- van Vliet-Ostaptchouk, J.V. et al., 2012. Common Variants in the Type 2 Diabetes KCNQ1 Gene Are Associated with Impairments in Insulin Secretion During Hyperglycaemic Glucose Clamp. *PLoS One*, 7(3): e32148.
- Vandesompele, J. et al., 2002. Accurate normalization of real-time quantitative RT-PCR data by geometric averaging of multiple internal control genes. *Genome Biol*, 3(7): RESEARCH0034.
- Velayos-Baeza, A., Vettori, A., Copley, R.R., Dobson-Stone, C. and Monaco, A.P., 2004. Analysis of the human VPS13 gene family. *Genomics*, 84(3): 536-49.
- Voight, B.F. et al., 2010. Twelve type 2 diabetes susceptibility loci identified through large-scale association analysis. *Nat Genet*, 42(7): 579-89.
- Volkmar, M. et al., 2012. DNA methylation profiling identifies epigenetic dysregulation in pancreatic islets from type 2 diabetic patients. *EMBO J*, 31(6): 1405-26.
- Walsh, T. et al., 2008. Rare structural variants disrupt multiple genes in neurodevelopmental pathways in schizophrenia. *Science*, 320(5875): 539-43.
- Walters, R.G. et al., 2010. A new highly penetrant form of obesity due to deletions on chromosome 16p11.2. *Nature*, 463(7281): 671-5.
- Wang, H. et al., 2009. Phenotypic and molecular evaluation of a chromosome 1q region with linkage and association to type 2 diabetes in humans. *J Clin Endocrinol Metab*, 94(4): 1401-8.
- Wang, K. et al., 2010a. Large copy-number variations are enriched in cases with moderate to extreme obesity. *Diabetes*, 59(10): 2690-4.
- Wang, Y., Solt, L.A. and Burris, T.P., 2010b. Regulation of FGF21 expression and secretion by retinoic acid receptor-related orphan receptor alpha. *J Biol Chem*, 285(21): 15668-73.

- Wardle, J., Carnell, S., Haworth, C.M. and Plomin, R., 2008. Evidence for a strong genetic influence on childhood adiposity despite the force of the obesogenic environment. *Am J Clin Nutr*, 87(2): 398-404.
- Wareham, N.J., Byrne, C.D., Williams, R., Day, N.E. and Hales, C.N., 1999. Fasting proinsulin concentrations predict the development of type 2 diabetes. *Diabetes Care*, 22(2): 262-70.
- Warton, K., Foster, N.C., Gold, W.A. and Stanley, K.K., 2004. A novel gene family induced by acute inflammation in endothelial cells. *Gene*, 342(1): 85-95.
- Waters, K.M. et al., 2010. Consistent association of type 2 diabetes risk variants found in europeans in diverse racial and ethnic groups. *PLoS Genet*, 6(8).
- Weaver, S. et al., 2010. Taking qPCR to a higher level: Analysis of CNV reveals the power of high throughput qPCR to enhance quantitative resolution. *Methods*, 50(4): 271-6.
- Whale, A.S. et al., 2012. Comparison of microfluidic digital PCR and conventional quantitative PCR for measuring copy number variation. *Nucleic Acids Res*.
- WHO, 2011a. World Health Organisation Diabetes Fact Sheet.
<http://www.who.int/mediacentre/factsheets/fs312/en/>.
- WHO, 2011b. Use of glycated haemoglobin (HbA1c) in the diagnosis of diabetes mellitus. http://www.who.int/diabetes/publications/diagnosis_diabetes2011/en/index.html.
- Wijesekara, N. et al., 2010. Beta cell-specific Znt8 deletion in mice causes marked defects in insulin processing, crystallisation and secretion. *Diabetologia*, 53(8): 1656-68.
- Willcocks, L.C. et al., 2008. Copy number of FCGR3B, which is associated with systemic lupus erythematosus, correlates with protein expression and immune complex uptake. *J Exp Med*, 205(7): 1573-82.
- Willer, C.J. et al., 2009. Six new loci associated with body mass index highlight a neuronal influence on body weight regulation. *Nat Genet*, 41(1): 25-34.
- Wiltshire, S. et al., 2001. 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(3): 553-69.
- Winckler, W. et al., 2007. Evaluation of common variants in the six known maturity-onset diabetes of the young (MODY) genes for association with type 2 diabetes. *Diabetes*, 56(3): 685-93.
- WTCCC, 2007. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature*, 447(7145): 661-78.
- Xu, B. et al., 2008. Strong association of de novo copy number mutations with sporadic schizophrenia. *Nat Genet*, 40(7): 880-5.
- Yamauchi, T. et al., 2010. A genome-wide association study in the Japanese population identifies susceptibility loci for type 2 diabetes at UBE2E2 and C2CD4A-C2CD4B. *Nat Genet*, 42(10): 864-8.
- Yang, H., Sasaki, T., Minoshima, S. and Shimizu, N., 2007. Identification of three novel proteins (SGSM1, 2, 3) which modulate small G protein (RAP and RAB)-mediated signaling pathway. *Genomics*, 90(2): 249-60.

- Yang, T.L. et al., 2008. Genome-wide copy-number-variation study identified a susceptibility gene, UGT2B17, for osteoporosis. *Am J Hum Genet*, 83(6): 663-74.
- Yasuda, K. et al., 2008. Variants in KCNQ1 are associated with susceptibility to type 2 diabetes mellitus. *Nat Genet*, 40(9): 1092-7.
- Yatsuki, H. et al., 2002. Domain regulation of imprinting cluster in Kip2/Lit1 subdomain on mouse chromosome 7F4/F5: large-scale DNA methylation analysis reveals that DMR-Lit1 is a putative imprinting control region. *Genome Res*, 12(12): 1860-70.
- Yew, K.H. et al., 2005. Cross-talk between bone morphogenetic protein and transforming growth factor-beta signaling is essential for exendin-4-induced insulin-positive differentiation of AR42J cells. *J Biol Chem*, 280(37): 32209-17.
- Yoshioka, M., Kayo, T., Ikeda, T. and Koizumi, A., 1997. A novel locus, Mody4, distal to D7Mit189 on chromosome 7 determines early-onset NIDDM in nonobese C57BL/6 (Akita) mutant mice. *Diabetes*, 46(5): 887-94.
- Yoshioka, N., Kuzuya, T., Matsuda, A., Taniguchi, M. and Iwamoto, Y., 1988. Serum proinsulin levels at fasting and after oral glucose load in patients with type 2 (non-insulin-dependent) diabetes mellitus. *Diabetologia*, 31(6): 355-60.
- Yu, J. et al., 2007. Induced pluripotent stem cell lines derived from human somatic cells. *Science*, 318(5858): 1917-20.
- Zeggini, E. et al., 2008. Meta-analysis of genome-wide association data and large-scale replication identifies additional susceptibility loci for type 2 diabetes. *Nat Genet*, 40(5): 638-45.
- Zeggini, E. et al., 2007. Replication of genome-wide association signals in UK samples reveals risk loci for type 2 diabetes. *Science*, 316(5829): 1336-41.
- Zhang, D. et al., 2009. Highly efficient differentiation of human ES cells and iPS cells into mature pancreatic insulin-producing cells. *Cell Res*, 19(4): 429-38.
- Zimmet, P., Alberti, K.G. and Shaw, J., 2001. Global and societal implications of the diabetes epidemic. *Nature*, 414(6865): 782-7.
- Zollino, M. et al., 2009. A case of Beckwith-Wiedemann syndrome caused by a cryptic 11p15 deletion encompassing the centromeric imprinted domain of the BWS locus. *J Med Genet*.
- Zuk, O., Hechter, E., Sunyaev, S.R. and Lander, E.S., 2012. The mystery of missing heritability: Genetic interactions create phantom heritability. *Proc Natl Acad Sci U S A*, 109(4): 1193-8.

Insights Into the Molecular Mechanism for Type 2 Diabetes Susceptibility at the *KCNQ1* Locus From Temporal Changes in Imprinting Status in Human Islets

Mary E. Travers,¹ Deborah J. G. Mackay,² Marloes Dekker Nitert,³ Andrew P. Morris,⁴ Cecilia M. Lindgren,⁴ Andrew Berry,⁵ Paul R. Johnson,^{1,6,7} Neil Hanley,⁵ Leif C. Groop,³ Mark I. McCarthy,^{1,4,6} and Anna L. Gloyn^{1,6}

The molecular basis of type 2 diabetes predisposition at most established susceptibility loci remains poorly understood. *KCNQ1* maps within the 11p15.5 imprinted domain, a region with an established role in congenital growth phenotypes. Variants intronic to *KCNQ1* influence diabetes susceptibility when maternally inherited. By use of quantitative PCR and pyrosequencing of human adult islet and fetal pancreas samples, we investigated the imprinting status of regional transcripts and aimed to determine whether type 2 diabetes risk alleles influence regional DNA methylation and gene expression. The results demonstrate that gene expression patterns differ by developmental stage. *CDKN1C* showed monoallelic expression in both adult and fetal tissue, whereas *PHLDA2*, *SLC22A18*, and *SLC22A18AS* were biallelically expressed in both tissues. Temporal changes in imprinting were observed for *KCNQ1* and *KCNQ1OT1*, with monoallelic expression in fetal tissues and biallelic expression in adult samples. Genotype at the type 2 diabetes risk variant rs2237895 influenced methylation levels of regulatory sequence in fetal pancreas but without demonstrable effects on gene expression. We demonstrate that *CDKN1C*, *KCNQ1*, and *KCNQ1OT1* are most likely to mediate diabetes susceptibility at the *KCNQ1* locus and identify temporal differences in imprinting status and methylation effects, suggesting that diabetes risk effects may be mediated in early development.

The translation of established type 2 diabetes risk variants into an improved understanding of disease pathology is challenging. Progress has been made primarily at the few loci where causal alleles are coding (1–4), but disease mechanisms are unclear for the majority of loci mapping outside coding regions.

The *KCNQ1* locus harbors at least two independent regions of association with type 2 diabetes risk (intron 10 and intron 15), both acting through impaired islet function (5–11). *KCNQ1* itself encodes the K_v7.1 voltage-gated potassium channel subunit, which is expressed in human β -cells (12) but plays an uncertain role in insulin secretion. Neither patients with cardiac arrhythmia caused by *KCNQ1* mutations nor *Kcnq1*-null mice demonstrates hyperglycemia or glucose intolerance (13,14), whereas *KCNQ1* knockdown in human islets does not alter insulin secretion (15).

In accordance with the location of *KCNQ1* at the imprinted 11p15.5 region, associated alleles at both signals confer disease risk only when maternally inherited (16). It has been demonstrated, primarily through studies of the syntenic region of mouse chromosome 7, that regional gene expression is regulated by differential methylation at the promoter of *KCNQ1* overlapping transcript 1 (*KCNQ1OT1*), a nontranslated antisense RNA that regulates maternal-specific expression of downstream genes (17) (Fig. 1).

Disruption of genomic architecture at the 11p15.5 chromosomal region has a well-established role in Beckwith-Wiedemann syndrome (BWS), a congenital overgrowth syndrome often associated with hypoglycemia (18). Furthermore, unbalanced placental expression of two regional genes, *PHLDA2* and *CDKN1C*, is associated with intrauterine growth retardation (19). We hypothesize that type 2 diabetes risk may be mediated through disruption of methylation and imprinted gene expression within the imprinted cluster. In this study, we perform the first assessment of 11p15.5 regional imprinting in adult human islets and fetal pancreas and investigate the effect of risk genotype status on DNA methylation and imprinted gene expression.

RESEARCH DESIGN AND METHODS

Islet and fetal pancreas isolation and DNA/RNA extraction. Human islets of deceased donors of European descent were obtained (with research consent) from the Oxford Diabetes Research & Wellness Foundation Human Islet Isolation Facility ($n = 30$) and the Human Tissue Laboratory at Lund University Diabetes Centre ($n = 42$). Fetal pancreas samples ($n = 18$) were

From the ¹Oxford Centre for Diabetes Endocrinology & Metabolism, University of Oxford, Oxford, U.K.; the ²Faculty of Medicine, University of Southampton, Southampton, U.K.; the ³Lund University Diabetes Centre, Malmö, Sweden; the ⁴Wellcome Trust Centre for Human Genetics, University of Oxford, Oxford, U.K.; the ⁵Endocrinology and Diabetes, Faculty of Medical & Human Sciences, University of Manchester, Manchester, U.K.; the ⁶Oxford NIHR Biomedical Research Centre, Churchill Hospital, Oxford, U.K.; and the ⁷Nuffield Department of Surgical Sciences, University of Oxford, Oxford, U.K. Corresponding author: Anna L. Gloyn, anna.gloyn@dr1.ox.ac.uk. Received 19 June 2012 and accepted 19 August 2012.

DOI: 10.2337/db12-0819

This article contains Supplementary Data online at <http://diabetes.diabetesjournals.org/lookup/suppl/doi:10.2337/db12-0819/-/DC1>.

The funder played no role in the conduct of the study, collection of data, management of the study, analysis of data, interpretation of data, or preparation of the manuscript.

© 2012 by the American Diabetes Association. Readers may use this article as long as the work is properly cited, the use is educational and not for profit, and the work is not altered. See <http://creativecommons.org/licenses/by-nc-nd/3.0/> for details.

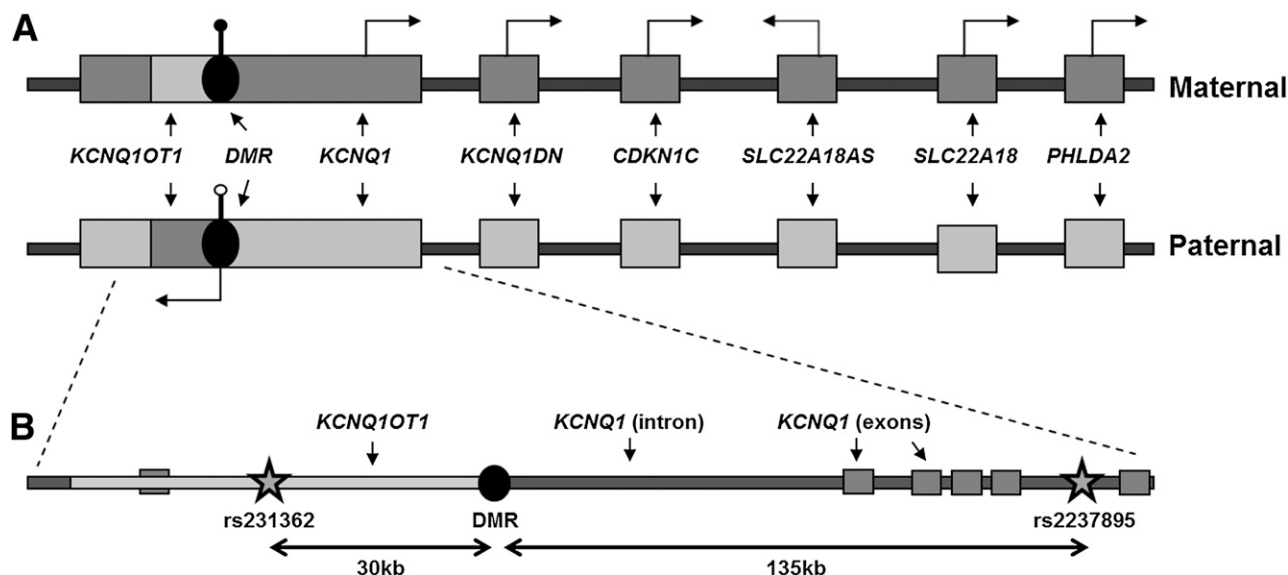


FIG. 1. Schematic representation of imprinting control [as described in (17)] and type 2 diabetes-associated SNPs at 11p15.5. **A:** Chromosome 11:2,450,000–2,960,000. The closed circle at the DMR represents a high level of methylation; the open circle represents a low level of methylation. The transcribed sequence is shown in dark gray, and untranscribed is shown in light gray. Arrows indicate direction of transcription. **B:** Chromosome 11:2,650,000–2,880,000. The smaller region is distinguished by exonic (boxes) and intronic regions of *KCNQ1* (dark gray), a region of *KCNQ1OT1* transcription (light gray), and relative positions of top disease-associated SNPs (rs231362, chr11:2,691,471, and rs2237985, chr11:2,857,194). All genomic coordinates are b37/hg19 (graphics not to scale).

obtained with informed consent and ethical approval from the North West Regional Ethics Committee. All islet preparations were >80% pure, with RNA integrity numbers >7. Donor and purity details are provided in Supplementary Table 1. DNA and RNA were extracted using TRIzol (Life Technologies).

cDNA synthesis. cDNA was generated through random-primed first-strand synthesis from 1 μ g RNA in accordance with standard protocols, including treatment with DNase I.

Genotyping. Type 2 diabetes-associated single-nucleotide polymorphisms (SNPs) (rs231362 and rs2237895) were selected as the lead SNPs (strongest evidence for association) in each of the independent signals. Reporter coding SNPs for imprinting analysis were selected to have the highest possible minor allele frequencies, maximizing heterozygous samples capable of differentiating mRNA products from homologous chromosomes. Genotyping was performed using TaqMan chemistry and SDS2.3 allelic discrimination software (Applied Biosystems). All SNPs reached genotyping pass rates of >95%, were present in accordance with expected (HapMap Centre d'Etude du Polymorphisme Humain) minor allele frequencies, and did not depart from Hardy-Weinberg equilibrium.

Sequencing. Primers were designed using web-based software (<http://frodo.wi.mit.edu>). For cDNA sequencing, primers were designed across exons to prevent amplification of residual genomic DNA (gDNA). Samples were amplified using AmpliTaq Gold DNA polymerase enzyme and sequenced on an ABI 3700 genetic analyzer machine using the BigDye Terminator v1.1 Cycle Sequencing Kit (all from Applied Biosystems). Results were analyzed using Mutation Surveyor version 3.4 software (SoftGenetics).

Bisulphite treatment and pyrosequencing. DNA (500 ng) was bisulfite treated using the EZ DNA Methylation-Gold Kit (Zymo Research). Treated DNA was amplified in duplicate using biotinylated primers and Titanium Taq reagents (Clontech) before pyrosequencing on a PSQ 96MA machine (Qiagen). Analysis was performed using PyroMark version 2.0 software, including cytosine bases outside CpG dinucleotides to confirm efficiency of conversion.

Fragment analysis for indel analysis. PCR was performed using HotStar Taq and Q-Solution reagents (Qiagen) with FAM-labeled primers. Products were analyzed on an Applied Biosystems 3130xl machine (Dye set D) and visualized using Peak Scanner version 1.0 software (Applied Biosystems).

Gene expression. Genes were selected for analysis on the basis of inclusion in the GeneImprint database (www.geneimprint.com) and expression regulation by *KCNQ1OT1*'s differentially methylated promoter. Quantitative PCR was performed in triplicate on an AB 7900HT machine, using inventoried (where available) or custom (designed using Primer Express version 3.0 software) TaqMan expression assays and cDNA samples diluted 1:10 in 0.01 mol/L Tris. Total expression reactions multiplexed a FAM-labeled test assay and a VIC-labeled endogenous control assay (one of *RNAseP*, *HPRT*, or *PPIA*). Analysis was performed using the $\Delta\Delta C_t$ method. Allele-specific expression analysis

compared the amplification cycle number at which VIC and FAM fluorescence levels crossed a critical threshold, where VIC/FAM probes annealed over the two alleles of a coding SNP.

Statistical analysis. Methylation and expression in islet samples obtained from Lund ($n = 42$) and Oxford ($n = 30$) were not significantly different ($P \geq 0.2$, Kolmogorov-Smirnov independent samples test). All islet samples, therefore, were combined for further analysis. Analysis was conducted by linear regression of expression or methylation on genotype group, with donor age, donor sex, sample purity, and center of origin (for adult islets) or days post-conception (for fetal pancreas) included as covariates. Analysis was also performed using a Kruskal-Wallis analysis of ranks, which retained significance in all cases.

RESULTS

Genes at the 11p15.5 cluster show temporal differences in imprinting status. *KCNQ1*, *KCNQ1OT1*, *CDKN1C*, *PHLDA2*, *SLC22A18*, and *SLC22A18AS* (but not *KCNQ1DN* or *TRPM5*) were quantifiably expressed in both adult islets ($n = 72$) and fetal pancreas ($n = 18$). To determine the imprinting status for each gene, we used common coding SNPs or a coding indel (for *CDKN1C* only) to distinguish between mRNA products from homologous chromosomes (Supplementary Table 2). Imprinted or nonimprinted expression was indicated by mono- or biallelic expression of these SNPs in cDNA from samples that were heterozygous in gDNA.

CDKN1C was monoallelically expressed in both adult and fetal tissues (Fig. 2A), as demonstrated in all samples heterozygous for the coding indel. Imprinted expression, however, was not universal among other genes in the region. *PHLDA2* and *SLC22A18* were expressed biallelically in all adult and all fetal samples (Fig. 2B). For *SLC22A18AS*, where there were relatively few heterozygous samples, cDNA sequencing demonstrated retained heterozygosity at all coding SNPs (Supplementary Fig. 1). *KCNQ1* and *KCNQ1OT1* showed temporal changes in imprinting status. Coding SNPs in both genes showed monoallelic expression in all fetal pancreas samples but were biallelically

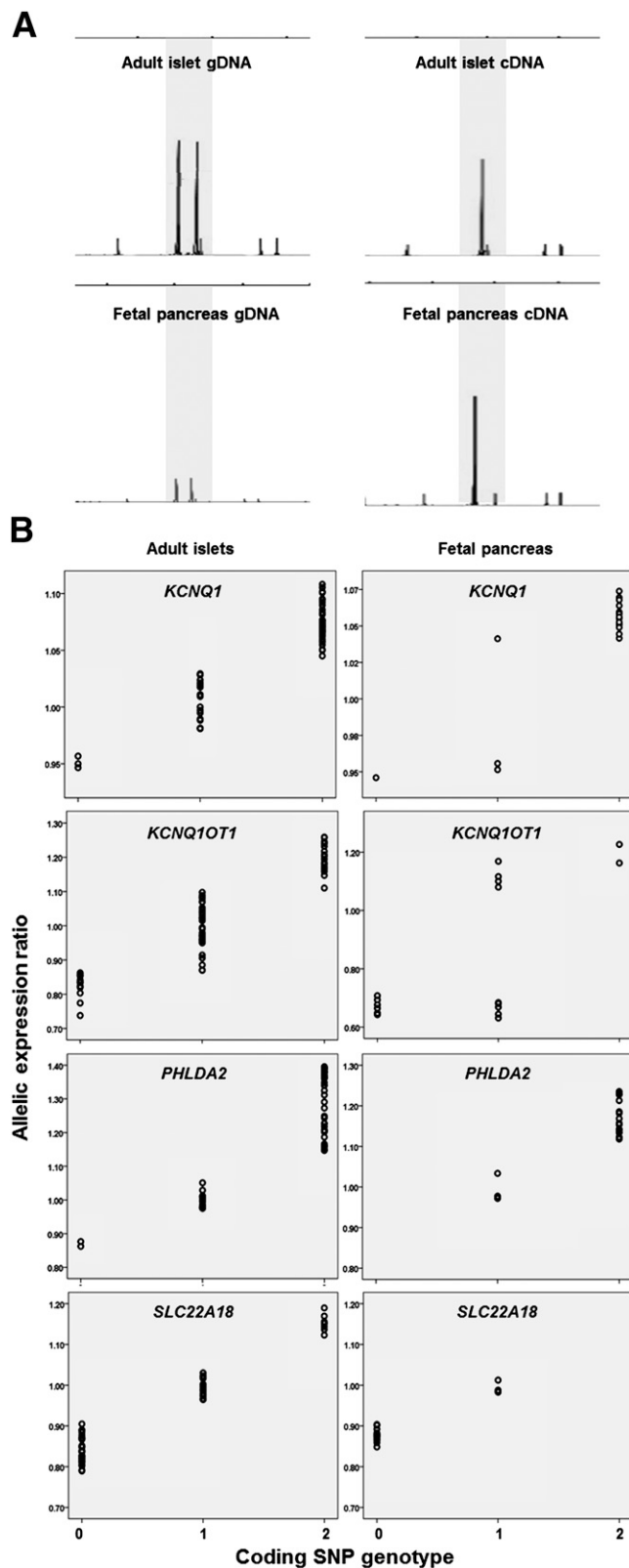


FIG. 2. A: Imprinting status of genes in the 11p15.5 cluster *CDKN1C*. Fragment analysis traces demonstrate monoallelic (imprinted) expression of *CDKN1C* in adult islet (*top*) and fetal pancreas (*bottom*) samples. *Left:* genomic DNA, with the two size peaks characteristic of a heterozygote for del171APVA highlighted. *Right:* cDNA from the same heterozygous samples. In both cases, only one size peak is evident, indicating the presence of mRNA from only one chromosome. Every sample heterozygous for del171APVA at the gDNA level appeared homozygous at the cDNA level, indicating monoallelic expression. **B:** Imprinting status of genes in the 11p15.5 cluster *KCNQ1*, *KCNQ1OT1*,

expressed in all adult islet samples (Fig. 2B). These results were confirmed by sequencing cDNA from all heterozygous samples (Supplementary Fig. 2).

Type 2 diabetes risk genotype status influences regional methylation in a developmentally flexible manner. To test the hypothesis that risk SNPs influence gene expression through an effect on DNA methylation, we quantified methylation at five representative regions of sequence. One assay widely used for the clinical diagnosis of BWS (20) maps within the differentially methylated region (DMR), and two were located at CTCF and putative PLAGL1 transcription factor binding sites within the DMR (21,22). Two other assays were designed at the DMR boundary on the basis of proximity to numerous CpG dinucleotides and disease-associated SNPs (rs231354 and rs2283202, respectively). To ascertain risk genotype status, DNA from each sample was genotyped for SNPs representing the intron 10 (rs231362) and intron 15 (rs2237895) signals.

rs2237895 was nominally associated ($P < 0.1$) with changes in methylation status at three of the sites tested and at differing developmental stages (Fig. 3). At the diagnostic DMR region and CTCF binding site, methylation was higher (7.3% and 5.6%, $P = 0.02$ and 0.08 respectively) in fetal samples homozygous for the risk allele than in fetal non-risk homozygotes. This effect was not apparent in adult islets, where risk genotype had no effect on methylation levels at these same sites ($P > 0.25$). Conversely, an effect was seen only in adult tissues at the PLAGL1 binding site. Here, methylation was 1.6% higher in risk genotype homozygotes than in nonrisk homozygotes ($P = 0.006$) in adult islets, but no effect was seen in fetal pancreas ($P = 0.36$). No effects were identified from the top disease-associated SNP within intron 10 of *KCNQ1* (rs231362, $P > 0.1$) (Supplementary Fig. 3) or from either SNP at the two candidate assays ($P > 0.1$) (Supplementary Fig. 4).

Type 2 diabetes risk genotype and total and allele-specific gene expression. We identified no relationship between rs2237895 risk allele number and total gene expression of *KCNQ1*, *KCNQ1OT1*, *CDKN1C*, *PHLDA2*, *SLC22A18*, or *SLC22A1AS* in the samples available for study ($P > 0.1$ for all genes) (Fig. 4). There was also no detectable relationship between rs231362 risk allele number and total gene expression (Supplementary Fig. 5) or between methylation at any of the five tested sites and total gene expression ($P > 0.1$ in all cases). Power for this analysis was reduced by undetermined risk allele parent of origin. We also examined the effect of risk genotype on

PHLDA2, and *SLC22A18*. Plots demonstrate flexibility of imprinting at the 11p15.5 cluster. The x-axis represents genotypes of reporter coding SNPs within the relevant gene (numbering arbitrary). The left- and right-hand clusters contain the two homozygote groups, whereas the center cluster contains heterozygous samples. The y-axis represents the ratio of mRNA expression level of a fluorescent probe specific to one allele of the coding SNP against the expression level of a differently labeled probe specific to the alternative allele. The two homozygous groups are therefore expected to have relatively high and low ratios, representing substantially more amplification of the probe specific for one allele. Under biallelic expression (all adult islet samples and *PHLDA2* and *SLC22A18* fetal pancreas samples), heterozygous samples form a central cluster with a ratio of ~1, indicating equal expression from homologous chromosomes. Under monoallelic (imprinted) expression (*KCNQ1* and *KCNQ1OT1* fetal pancreas samples), heterozygote samples separate into two distinct groups, corresponding to the two homozygous clusters and indicating expression from only one chromosome.

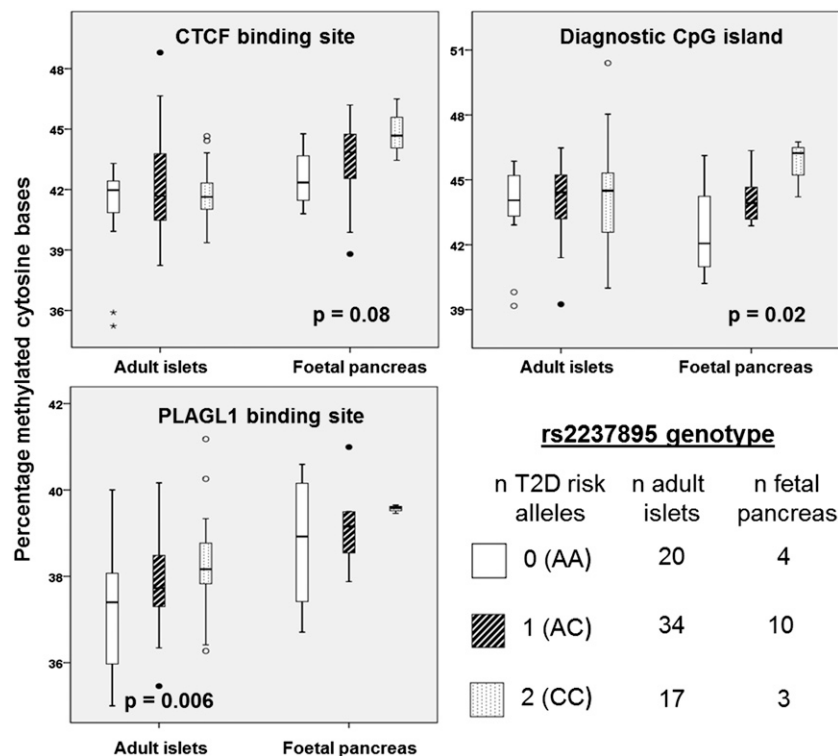


FIG. 3. Methylation according to rs2237895 type 2 diabetes risk genotype. Box plots show the effect of risk allele number (*x*-axis) on methylation level (*y*-axis) separated by tissue type. Boxes represent quartiles; whiskers encompass values within 1.5 times the interquartile range. Risk genotype is significantly associated with increased methylation levels at the CTCF binding site and Beckwith-Wiedemann diagnostic CpG island in fetal samples only and at the PLAGL1 binding site in adult islets only. T2D, type 2 diabetes.

allele-specific expression levels but identified no impact on the balance of expression between chromosomes for any tested gene ($P > 0.2$).

DISCUSSION

There is compelling evidence that diabetes risk at the *KCNQ1* locus is mediated through a gene with imprinted expression (16). By demonstrating that *PHLDA2*, *SLC22A18*, and *SLC22A18AS* are biallelically expressed in both adult and fetal pancreas and islets, we show that they are unlikely to be involved in a proximal molecular mechanism for diabetes risk. Likewise, any diabetes susceptibility mechanism working through *KCNQ1* or *KCNQ1OT1* is likely to be early in islet development because these transcripts are imprinted in fetal pancreas but not in adult islets.

The cyclin-dependent kinase inhibitor *CDKN1C* (encoding p57^{KIP2}), imprinted at both developmental time points, emerges as a particularly strong regional candidate. *CDKN1C* is expressed by 30–40% of β -cells in healthy individuals but never concurrently with the Ki67 marker of cell proliferation (23). *CDKN1C* expression is abolished in the hyperproliferative pancreatic lesions of focal hyperinsulinemia, suggesting a key role in regulating pancreatic β -cell proliferation (23). Loss-of-function mutations in *CDKN1C* cause BWS and hypoglycemia, whereas gain-of-function mutations in the PCNA binding domain have recently been shown to cause congenital undergrowth (18,24). A higher level of DNA methylation in individuals carrying more risk alleles is consistent with a disease model of reduced *KCNQ1OT1* transcription, diminished repressive histone modifications (25), and increased

CDKN1C expression, leading to impaired islet proliferation or development.

We demonstrate that risk genotype status is related to DNA methylation in a developmentally variable manner. Although the relationship between genotype and methylation is statistically stronger in adult than in fetal samples (probably attributable to improved power from a larger sample size), the magnitude of effect appears larger in fetal samples. It is noteworthy that the true allele-specific effect size is likely to be underestimated because data were obtained by pyrosequencing PCR products amplified from both chromosomes. Further work in larger numbers of human islet samples, when they become available, will be required to explore in more detail the relationship between DNA methylation and expression.

By use of the largest cohort of human islets currently available, we have performed the first assessment of imprinting status at 11p15.5 in human adult islets and fetal pancreas. The data provide insights into the complexity of imprinting at 11p15.5, highlighting the necessity of performing functional studies in relevant tissues and at appropriate developmental stages. The data have significant implications for the molecular mechanism by which associated variants in this region exert their effect on diabetes risk.

ACKNOWLEDGMENTS

The work in Oxford was supported by the Medical Research Council (G0700222/81696 to A.L.G.), the Wellcome Trust (095101/Z/10/Z to A.L.G.), and the European Foundation for the Study of Diabetes (Albert Reynold Travel Fellowship to M.E.T.). The work in Oxford and Manchester

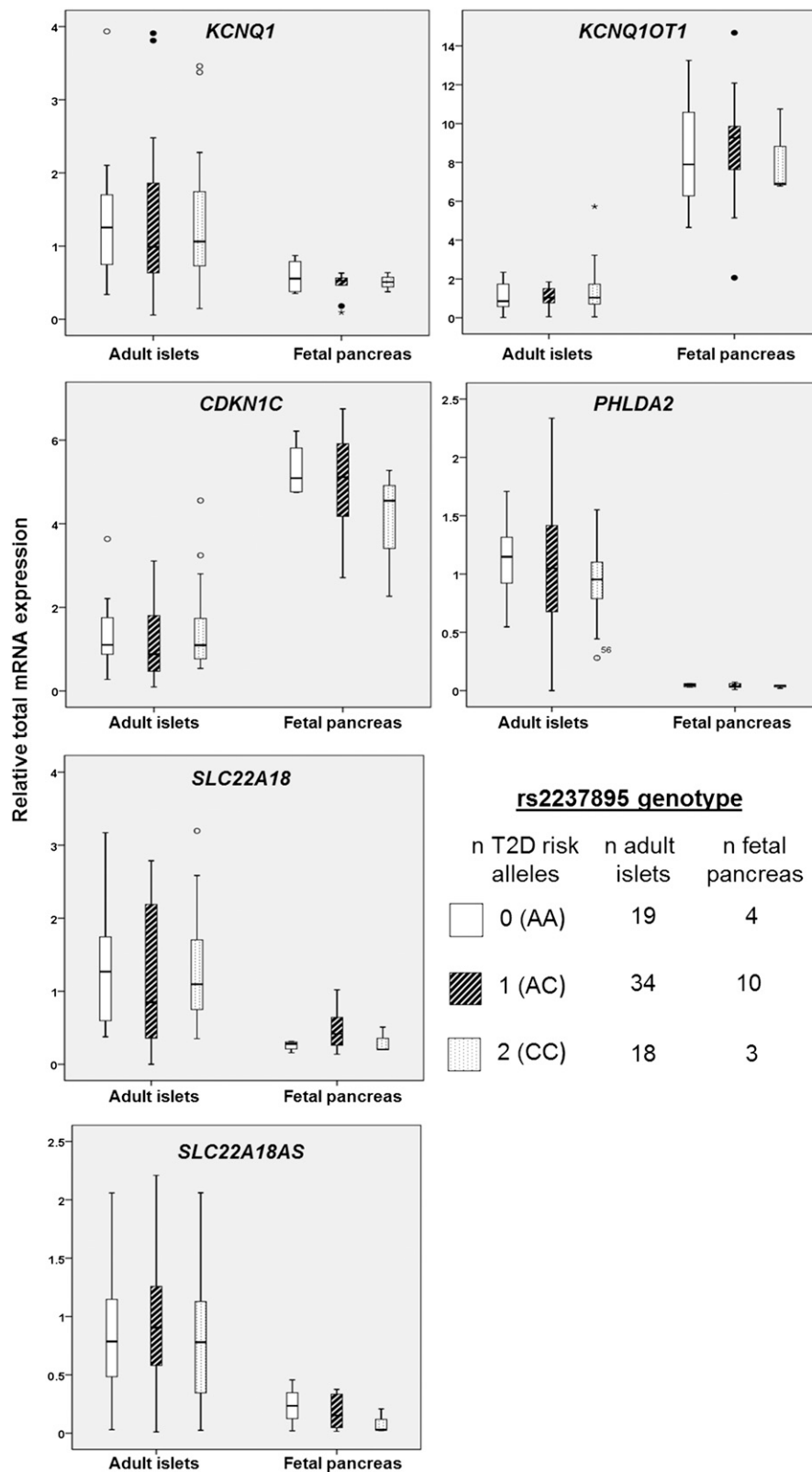


FIG. 4. Total expression according to rs2237895 type 2 diabetes risk genotype. Box plots show the effect of risk allele number (*x*-axis) on total mRNA expression level (*y*-axis), separated by tissue type. Boxes represent quartiles; whiskers encompass values within 1.5 times the interquartile range. There was no evidence for an effect of risk allele number on expression levels of any of the tested genes in either tissue type ($P > 0.05$ in all cases). T2D, type 2 diabetes.

was funded by the National Institute of Health Research Biomedical Research Centres in Oxford and Manchester, respectively. The work in Southampton was funded by Diabetes UK (BDA:RD08/0003611 to D.J.G.M.). The work at Lund University Diabetes Centre was funded by the Swedish Research Council (DNr 521-2007-4037, DNr 2009-1039, and DNr 349-2008-6589 to L.C.G.). M.E.T. was funded by a Medical Research Council studentship. M.D.N. was funded by the Swedish Foundation for Medical Research. N.H. is a Wellcome Trust Senior Fellow in Clinical Science. A.L.G. is a Wellcome Trust Senior Fellow in Basic and Biomedical Research.

No potential conflicts of interest relevant to this article were reported.

M.E.T., D.J.G.M., A.P.M., C.L.M., M.I.M., and A.L.G. designed the project. P.R.J. and L.C.G. provided human islet samples. N.H. provided fetal pancreas samples. M.E.T., M.D.N., and A.B. performed the research. M.E.T., M.I.M., and A.L.G. analyzed the data. M.E.T., M.I.M., and A.L.G. wrote the manuscript. D.J.G.M., M.D.N., A.P.M., C.L.M., A.B., N.H., and L.C.G. contributed to the manuscript. A.L.G. is the guarantor of this work and, as such, had full access to all the data in the study and takes responsibility for the integrity of the data and the accuracy of the data analysis.

Parts of this study were presented at the 46th Annual Meeting of the European Association for the Study of Diabetes (EASD), Stockholm, Sweden, 20–24 September 2010.

The authors thank the Nordic Network for Clinical Islet Transplantation (Juvenile Diabetes Research Foundation award 31-2008-413) and the tissue isolation teams and human tissue laboratory within EXODIAB/Lund University Diabetes Centre. The authors also thank the Islet Isolation team in Oxford, particularly Stephen Hughes and Sarah Cross, for providing human islets for this work.

REFERENCES

- Hamming KS, Soliman D, Maternisz LC, 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* 2009;58:2419–2424
- Nicolson TJ, Bellomo EA, Wijesekara N, et al. Insulin storage and glucose homeostasis in mice null for the granule zinc transporter ZnT8 and studies of the type 2 diabetes-associated variants. *Diabetes* 2009;58:2070–2083
- Rees MG, Wincovitch S, Schultz J, et al. Cellular characterisation of the *GCKR* P446L variant associated with type 2 diabetes risk. *Diabetologia* 2012;55:114–122
- Beer NL, Tribble ND, McCulloch LJ, et al. The P446L variant in *GCKR* associated with fasting plasma glucose and triglyceride levels exerts its effect through increased glucokinase activity in liver. *Hum Mol Genet* 2009;18:4081–4088
- Tsai FJ, Yang CF, Chen CC, et al. A genome-wide association study identifies susceptibility variants for type 2 diabetes in Han Chinese. *PLoS Genet* 2010;6:e1000847
- Unoki H, Takahashi A, Kawaguchi T, et al. SNPs in *KCNQ1* are associated with susceptibility to type 2 diabetes in East Asian and European populations. *Nat Genet* 2008;40:1098–1102
- Voight BF, Scott LJ, Steinthorsdottir V, et al.; MAGIC investigators; GIANT Consortium. Twelve type 2 diabetes susceptibility loci identified through large-scale association analysis. *Nat Genet* 2010;42:579–589
- Yasuda K, Miyake K, Horikawa Y, et al. Variants in *KCNQ1* are associated with susceptibility to type 2 diabetes mellitus. *Nat Genet* 2008;40:1092–1097
- Holmkvist J, Banasik K, Andersen G, et al. The type 2 diabetes associated minor allele of rs2237895 *KCNQ1* associates with reduced insulin release following an oral glucose load. *PLoS ONE* 2009;4:e5872
- Jonsson A, Isomaa B, Tuomi T, et al. A variant in the *KCNQ1* gene predicts future type 2 diabetes and mediates impaired insulin secretion. *Diabetes* 2009;58:2409–2413
- Tan JT, Nurbaya S, Gardner D, Ye S, Tai ES, Ng DP. Genetic variation in *KCNQ1* associates with fasting glucose and β -cell function: a study of 3,734 subjects comprising three ethnicities living in Singapore. *Diabetes* 2009;58:1445–1449
- MacDonald PE, Wheeler MB. Voltage-dependent K(+) channels in pancreatic beta cells: role, regulation and potential as therapeutic targets. *Diabetologia* 2003;46:1046–1062
- Jespersen T, Grunnet M, Olesen SP. The *KCNQ1* potassium channel: from gene to physiological function. *Physiology (Bethesda)* 2005;20:408–416
- Pan Q, Ma J, Zhou Q, et al. *KCNQ1* loss-of-function mutation impairs gastric acid secretion in mice. *Mol Biol Rep* 2010;37:1329–1333
- Rosengren AH, Braun M, Mahdi T, et al. Reduced insulin exocytosis in human pancreatic β -cells with gene variants linked to type 2 diabetes. *Diabetes* 2012;61:1726–1733
- Kong A, Steinthorsdottir V, Masson G, et al.; DIAGRAM Consortium. Parental origin of sequence variants associated with complex diseases. *Nature* 2009;462:868–874
- Thakur N, Tiwari VK, Thomassin H, et al. An antisense RNA regulates the bidirectional silencing property of the *Kcnq1* imprinting control region. *Mol Cell Biol* 2004;24:7855–7862
- Weksberg R, Shuman C, Beckwith JB. Beckwith-Wiedemann syndrome. *Eur J Hum Genet* 2010;18:8–14
- McMinn J, Wei M, Schupf N, et al. Unbalanced placental expression of imprinted genes in human intrauterine growth restriction. *Placenta* 2006;27:540–549
- Mackay DJ, Callaway JL, Marks SM, et al. Hypomethylation of multiple imprinted loci in individuals with transient neonatal diabetes is associated with mutations in *ZFP57*. *Nat Genet* 2008;40:949–951
- Arima T, Kamikihara T, Hayashida T, et al. *ZAC*, *LIT1* (*KCNQ1OT1*) and *p57^{KIP2}* (*CDKN1C*) are in an imprinted gene network that may play a role in Beckwith-Wiedemann syndrome. *Nucleic Acids Res* 2005;33:2650–2660
- Fitzpatrick GV, Pugacheva EM, Shin JY, et al. Allele-specific binding of CTCF to the multipartite imprinting control region KvDMR1. *Mol Cell Biol* 2007;27:2636–2647
- Kassem SA, Ariel I, Thornton PS, et al. *p57^{KIP2}* expression in normal islet cells and in hyperinsulinism of infancy. *Diabetes* 2001;50:2763–2769
- Arboleda VA, Lee H, Parnaik R, et al. Mutations in the PCNA-binding domain of *CDKN1C* cause IMAGE syndrome. *Nat Genet* 2012;44:788–792
- Redrup L, Branco MR, Perdeaux ER, et al. The long noncoding RNA *Kcnq1ot1* organises a lineage-specific nuclear domain for epigenetic gene silencing. *Development* 2009;136:525–530

Type 2 diabetes and obesity: genomics and the clinic

Mary E. Travers · Mark I. McCarthy

Received: 30 March 2011 / Accepted: 26 May 2011 / Published online: 7 June 2011
© Springer-Verlag 2011

Abstract Type 2 diabetes (T2D) and obesity represent major challenges for global public health. They are at the forefront of international efforts to identify the genetic variation contributing to complex disease susceptibility, and recent years have seen considerable success in identifying common risk-variants. Given the clinical impact of molecular diagnostics in rarer monogenic forms of these diseases, expectations have been high that genetic discoveries will transform the prospects for risk stratification, development of novel therapeutics and personalised medicine. However, so far, clinical translation has been limited. Difficulties in defining the alleles and transcripts mediating association effects have frustrated efforts to gain early biological insights, whilst the fact that variants identified account for only a modest proportion of observed familiarity has limited their value in guiding treatment of individual patients. Ongoing efforts to track causal variants through fine-mapping and to illuminate the biological mechanisms through which they act, as well as sequence-based discovery of lower-frequency alleles (of potentially larger effect), should provide welcome acceleration in the capacity for clinical translation. This review will

summarise recent advances in identifying risk alleles for T2D and obesity, and existing contributions to understanding disease pathology. It will consider the progress made in translating genetic knowledge into clinical utility, the challenges remaining, and the realistic potential for further progress.

Introduction: genetic diseases of global impact

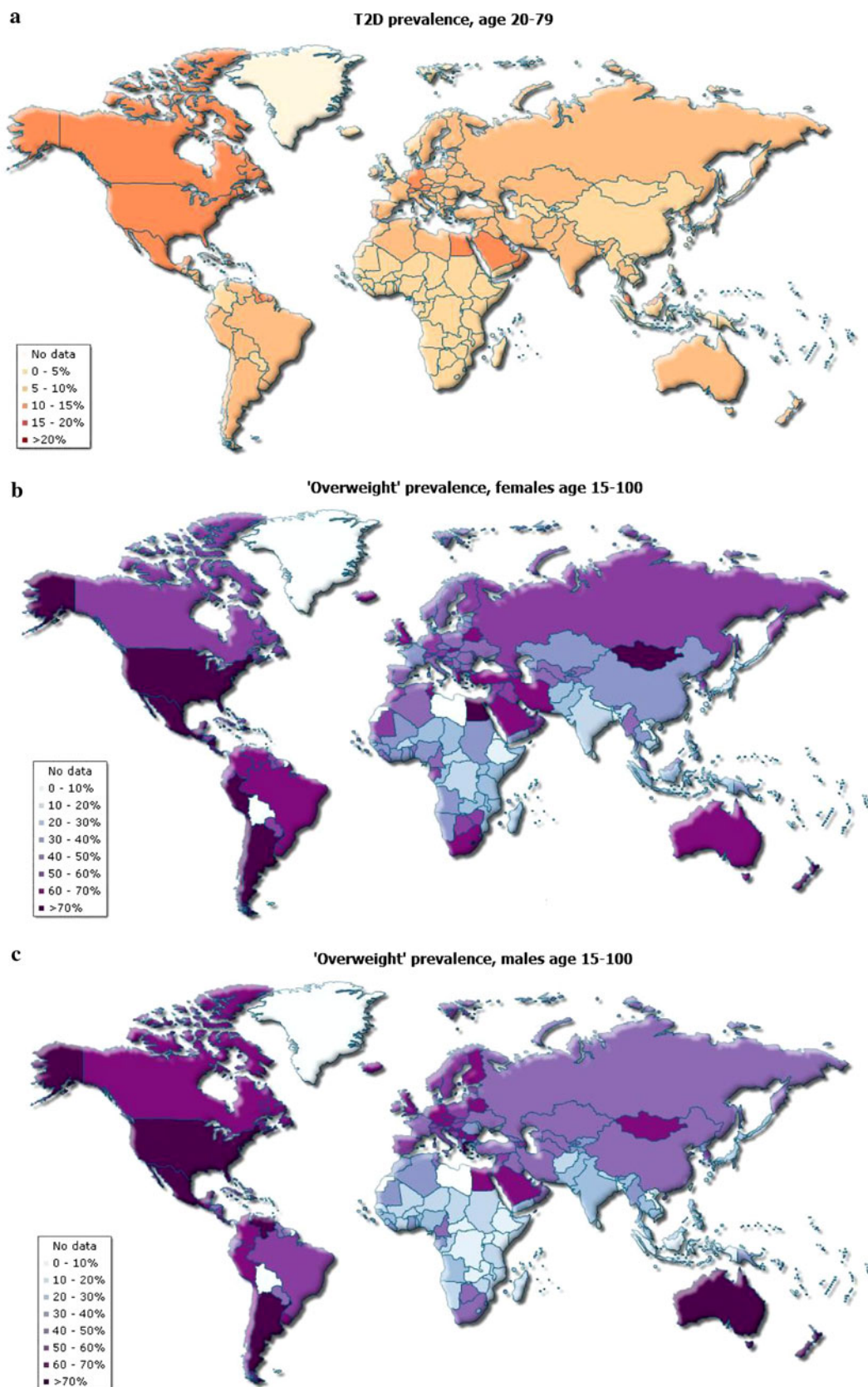
The rising prevalences of obesity and type 2 diabetes (T2D) indicate a crisis in global health (IDF Diabetes Atlas 2010). Worldwide, there are more than 400 million adults with a body mass index (BMI) exceeding 30 kg/m² (defining “obesity”) and 220 million with T2D (Fig. 1), figures which are projected to rise to 700 million and 366 million, respectively, by 2030 (World Health Organisation 2010; International Diabetes Federation 2010). Both diseases have substantial implications for mortality: in 2004, over 112,000 deaths in the United States were attributed to increased cardiovascular disease (CVD) resulting from obesity (Flegal et al. 2007), and in the same year, diabetes-related complications were estimated to account for 5% of all global mortality (World Health Organisation 2010). In 2006, for the first time, more people died as a result of being overweight than underweight, whilst treatment of T2D and obesity-related complications was responsible for 8% of all healthcare costs in the European Economic region (World Health Organisation 2010).

Despite these sobering figures and their ramifications for individuals, families and healthcare systems, current understanding of the basic pathophysiology of T2D and obesity remains rudimentary. Whilst both diseases have monogenic and syndromic counterparts—including maturity onset diabetes of the young (MODY), neonatal diabetes

M. E. Travers · M. I. McCarthy (✉)
Oxford Centre for Diabetes, Endocrinology and Metabolism,
Churchill Hospital, University of Oxford, Old Road,
Headington, Oxford OX3 7LJ, UK
e-mail: mark.mccarthy@dr1.ox.ac.uk

M. I. McCarthy
Wellcome Trust Centre for Human Genetics,
University of Oxford, Roosevelt Drive,
Headington, Oxford OX3 7BN, UK

M. I. McCarthy
Oxford NIHR Biomedical Research Centre, Churchill Hospital,
Old Road, Headington, Oxford OX3 7LJ, UK



◀ **Fig. 1** Prevalence of type 2 diabetes and ‘overweight’ by country (as of 2010). **a** Colour intensity represents percentage of individuals aged 20–79 with diabetes (fasting plasma glucose >7.0 mmol/L). **b** Colour intensity represents percentage of females aged 15–100 with BMI > 25 kg/m². **c** Colour intensity represents percentage of males aged 15–100 with BMI > 25 kg/m² (“overweight” data from World Health Organisation 2010: <https://apps.who.int/infobase/>; diabetes data from International Diabetes Federation: <http://www.diabetesatlas.org/datatables/>)

and Prader–Willi syndrome (PWS)—which have enjoyed significant progress in the characterisation of causal mechanisms, the overwhelming majority of cases of T2D and obesity have a more complex aetiological basis. Typically, an individual’s risk of disease development reflects the intersection of inherited variation at many genetic sites and exposure to modern environmental stressors, including increased energy intake and decreased physical activity (Stumvoll et al. 2008).

Clearly, the current explosion in T2D and obesity prevalence must be due primarily to environmental change—the timescales involved are far too short for shifts in susceptibility variant frequency. But not everyone exposed to the increasingly pervasive “obesogenic” environment seems at equivalent risk. Quantifying the genetic component of complex disease is not straightforward, but most estimates place the heritability of T2D at around 25% (Poulsen et al. 1999) and that of obesity between 50 and 80% (Maes et al. 1997). Heritability estimates are likely to fluctuate with time and space (reflecting changes in environmental variance and overall disease prevalence), but there is little empirical evidence of an attenuation of heritability in more contemporary studies (Wardle et al. 2008).

The clinical benefits of genomics: lessons from monogenic obesity and diabetes

Thanks to their high penetrance, the alleles responsible for rare, monogenic forms of non-autoimmune diabetes and obesity were relatively easily identified through linkage analysis (reviewed in Owen and Hattersley 2001; O’Rahilly and Farooqi 2006). These discoveries have led to molecular classifications of disease with demonstrable prognostic and therapeutic relevance. For example, individuals with maturity onset diabetes of the young (MODY) due to mutations in *HNF1A* respond particularly well to treatment with sulfonylureas, whilst those with mutations in glucokinase (*GCK*) can often come off medication entirely given their relatively benign prognosis (Schnyder et al. 2005; Pearson et al. 2003). Infants with neonatal diabetes due to mutations in the *KCNJ11* gene, conventionally treated with insulin, have typically shown substantial improvements in diabetes control when oral sulfonylureas are substituted (Pearson et al.

2006; Gloyn et al. 2004). Meanwhile, identification of mutations in the leptin gene (*LEP*) causing severe early-onset obesity (Montague et al. 1997) has resulted in the development of recombinant leptin therapy as a life-saving treatment for affected children (Farooqi et al. 1999).

As a consequence of such advances in genetic understanding and classification, molecular diagnostics and personalised therapy are now standard components of clinical care for patients with these monogenic forms of disease and for their families. The ambition, then, is for an improved understanding of the genetic basis of common forms of T2D and obesity to inspire similar insights into disease biology and to underpin future developments in clinical care.

A brief history of susceptibility gene discovery

The first and unavoidable step towards genomic medicine lies in the identification of genetic variants robustly associated with the disease of interest. However, the multifactorial nature of complex diseases such as T2D and obesity has rendered even this initial stage an enormous challenge.

Family-based linkage approaches, so successful in identifying the mutations responsible for monogenic and syndromic subtypes of obesity and T2D, proved poorly suited to revealing the variants of lower penetrance implicated in more typical forms of the diseases. By 2006, the Human Obesity Gene Map (Rankinen et al. 2006) listed 253 loci “linked” to obesity, but very few of these had been replicated in multiple studies. A meta-analysis of linkage data from >31,000 individuals and ~10,000 families failed to reveal any convincing BMI-influencing loci (Saunders et al. 2007).

Attention turned instead to association approaches in larger, unrelated samples sets (Merikangas and Risch 2003). Association analyses, however, rely upon typing the causal variant or a closely correlated proxy, and hence, initial efforts were constrained by practical limitations of genotyping cost and capacity to the evaluation of variants within pre-defined candidate genes. Nonetheless, this approach heralded the first wave of robustly associated variants. For T2D, non-synonymous variants in genes encoding the targets of two drugs widely used in T2D management [P12A in *PPARG* (Altshuler et al. 2000) for thiazolidinediones and E23K in *KCNJ11* (Gloyn et al. 2003) for sulfonylureas] showed consistent, though modest (per-allele odds ratios of ~1.2), evidence of association with disease risk. For obesity, variants within two genes already known to harbour mutations implicated in monogenic obesity—*MC4R* (V103I, I251L) and *PCSK1* (N221D, Q665E-S690T)—were shown to be associated with common obesity risk (Heid et al. 2005; Geller et al. 2004; Benzinou et al. 2008).

However, the candidate gene approach is restricted by its intrinsic reliance upon prior knowledge and expectation. When, as with T2D and obesity, our understanding of disease pathogenesis is imperfect, there is a manifest need to extend the search for susceptibility variants across the entire genome in an unbiased, hypothesis-free manner. The first gene to be implicated in T2D susceptibility without prior biological candidacy was *TCF7L2*, discovered following systematic association analysis across a region of previously identified linkage (Grant et al. 2006). The most strongly associated variants at this locus have the greatest effect on T2D susceptibility of any common variant so far identified.

The advent of genome-wide association studies (GWAS) proved transformative for the field. For T2D, the first wave of GWAS in 2007 confirmed the known loci at *PPARG*, *KCNJ11* and *TCF7L2*, but added a further six novel loci including signals near *CDKAL1*, *HHEX*, *SLC30A8*, *IGF2BP2* and *CDKN2A* (Frayling et al. 2007; Saxena et al. 2007; Scott et al. 2007; Sladek et al. 2007; Zeggini et al. 2007; Steinthorsdottir et al. 2007; WTCCC 2007). At the sixth locus, near *FTO* (Frayling et al. 2007), association with T2D was predicated entirely on case–control differences in adiposity, serendipitously revealing *FTO* as the first common variant signal for body mass index and risk of obesity.

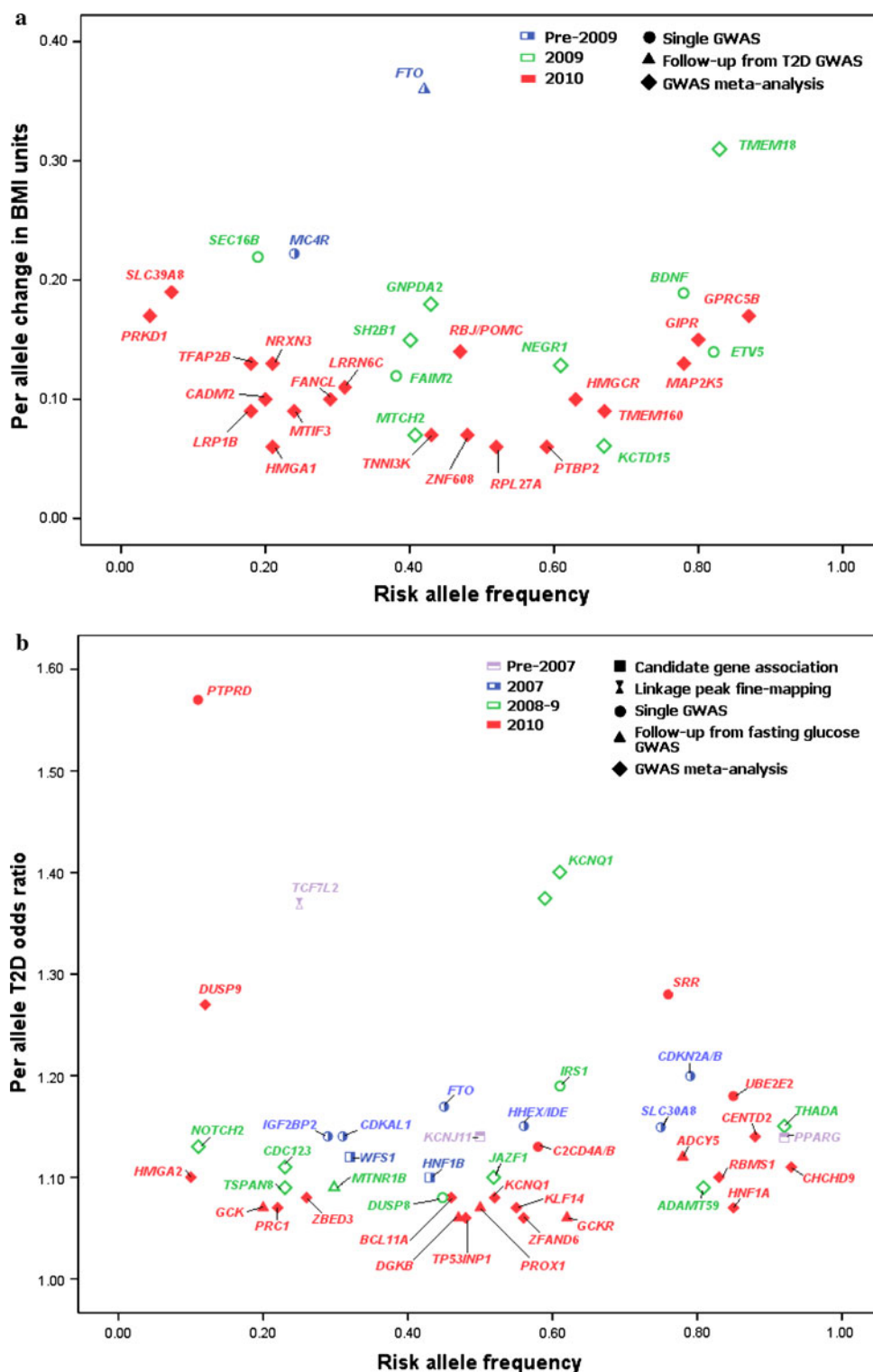
Successive rounds of GWA meta-analysis (Zeggini et al. 2008; Willer et al. 2009; Voight et al. 2010; Speliotes et al. 2010; Loos et al. 2008; Heid et al. 2010) have brought the count of confirmed common variant signals for T2D to more than 40, and for BMI and obesity to over 30 (McCarthy 2010). As expected, the improvement in power derived from increasing sample size has been particularly beneficial in exposing variants of smaller effect and more extreme risk allele frequency (Fig. 2). For T2D, the Diabetes Genetics Replication and Meta-analysis (DIAGRAM) Consortium first combined data from three published GWAS to reveal six novel loci (Zeggini et al. 2008) and subsequently aggregated data from an additional five GWAS to capture a further 12 signals (Voight et al. 2010). The Genomic Investigation of Anthropometric Traits (GIANT) Consortium first combined data from 15 GWAS cohorts to reveal six new loci contributing to variation in BMI, as well as replicating the by-then established common variant signals at *FTO* and *MC4R* (Willer et al. 2009). Almost in parallel, the deCODE group reported ten BMI-influencing loci (Thorleifsson et al. 2009). The synthesis of these two efforts, involving genetic analysis of almost 250,000 individuals, confirmed 14 existing loci and revealed 18 novel signals for BMI and obesity (Speliotes and Johnson 2010). Common copy number variants (CNVs), as opposed to SNPs, have only been implicated at

one of these loci, *NEGR1*, where a 45-kb deletion upstream of the gene is in perfect LD with a strongly associated SNP (Willer and Havulinna 2009). The role of rare CNVs in obesity has not been well examined so far, but rare deletions at chromosome 16p11.2 have been shown to have high penetrance for obesity and mental retardation (Walters et al. 2010).

Most of the early GWAS involved individuals of European descent, but a growing number of discoveries are being made in other ethnic groups. For example, BMI-associated variants in *MC4R* have been confirmed in Indian Asians (Chambers et al. 2008; Been et al. 2010), whilst GWAS in East Asians have revealed several novel associations for T2D (Yasuda et al. 2008; Unoki et al. 2008; Tsai et al. 2010; Yamauchi et al. 2010). By and large, although differences in allele frequency and effect size (combined with chance) may mean that initial discoveries are more likely to be made in one or other ethnic group (Myles et al. 2008), almost all the common variant signals for T2D so far examined have a consistent effect on diabetes risk across multiple ethnic groups (Waters et al. 2010). This provides strong evidence that common variant signals are driven by causal variants that are themselves frequent (and of rather ancient origin), rather than by “synthetic” effects of multiple rare causal variants—as has sometimes been proposed (Dickson et al. 2010). Further evidence of the utility of studies in diverse ethnic groups is provided by fine-mapping of obesity association at the *FTO*-locus in African-derived populations. Thanks to their weaker linkage disequilibrium patterns, the list of potentially causative variants has begun to be narrowed (Hasanein et al. 2010; Adeyemo et al. 2010).

The majority of GWAS have approached T2D as a binary disease phenotype, comparing allele frequencies between case and control groups. In contrast, most studies have tackled obesity through its cognate quantitative trait, BMI. The few studies that have taken the complementary approach have been especially instructive. The ‘Meta-Analyses of Glucose and Insulin-related traits Consortium’ (MAGIC; Dupuis et al. 2010) investigated the genetic basis of normal physiological variation in continuous glycaemic measure such as fasting glucose and insulin. Evaluating the resulting common variant signals with respect to T2D risk showed a wide disparity in consequences, ranging from relatively strong (e.g. *MTNR1B*) to negligible (e.g. *G6PC2*) effects on diabetes susceptibility. This range demonstrates that the mechanisms influencing physiological and pathophysiological variation in glucose homeostasis are only partially overlapping. Similarly, case–control studies of extreme obesity (Scherag et al. 2010; Meyre et al. 2009) have identified loci that seem quite distinct from those shown to influence population-level variation in BMI.

Fig. 2 Allele frequencies and effect sizes of known BMI-influencing and T2D-susceptibility loci. Risk allele frequencies and effect sizes are shown in European populations unless otherwise stated. Note how most recently identified loci (in red, filled markers) from ever-larger and better-powered meta-analyses have revealed loci with smaller effect sizes (bottom of plot) and/or more extreme risk allele frequencies (far left and right of plot). Gene names represent best candidate transcripts on the basis of location and biological plausibility, not necessarily proven to be causal. **a** BMI-influencing loci, coloured and filled by year of identification and with symbols according to experimental method of discovery. Note that *NEGR1*, *TMEM18*, *SH2B1* and *KCTD15* were identified simultaneously by single GWAS (Thorleifsson et al. 2009) and GWAS meta-analysis (Willer et al. 2009). **b** T2D-susceptibility loci, coloured and filled by year of identification and with symbols according to experimental method of discovery. *KCNQ1*: two SNPs in low LD in European populations are likely to represent independent association signals. *UBE2E2*, *C2CD4A/B*, *KCNQ1* (odds ratio ~ 1.4): odds ratios from Japanese populations and allele frequencies from HapMap-JPT. *SRR*, *PTPRD*: odds ratios from Chinese populations and allele frequencies from HapMap-CHB. *DUSP8*: odds ratio shown across both maternally and paternally inherited alleles—there is evidence that the risk effect may be larger from paternally derived alleles and protective from maternally derived alleles



The first step to translation: from associated variants to biological mechanisms

The discoveries of the past few years have necessarily focused, for both theoretical and practical reasons, on the identification of common susceptibility-alleles, with

emphasis rightly placed on obtaining robust statistical support for association. From a clinical perspective, the value of these discoveries lies primarily in the opportunities they provide for enhanced understanding of disease biology, and the major challenge lies in unlocking the mechanisms whereby these associated variants influence

disease progression. Success in this endeavour is not unduly influenced by the relatively modest effect sizes of many of the loci identified: in the case of *KCNJ11* and *PPARG*, for example, the therapeutic consequences of manipulating protein function (with sulfonylureas or thiazolidinediones, respectively) are out of all proportion to the limited effects of T2D-susceptibility variants within their encoding genes.

For most of the common variant loci revealed by GWAS, the transition from association signal to causal mechanism has proved far from straightforward. It has been challenging, because of the extensive linkage disequilibrium in most human populations, to tie the association signal to a single causal allele. With so many signals mapping to non-coding sequence, it has been difficult to define which of many regional transcripts is likely to be mediating the association. Both issues represent obstacles to the translation of genetic discoveries into an improved understanding of disease biology.

Fortunately, there are exceptions. At a subset of T2D-susceptibility loci, including *GCKR*, *PPARG* and *SLC30A8*, there is substantial statistical and biological evidence to support particular coding sequence variants as causal. For example, the T2D association signal on chromosome 2p23 (Saxena et al. 2007) can, through a combination of genetic and functional approaches, reasonably be attributed to the P446L variant in the glucokinase regulatory protein *GCKR*, one of 17 genes mapping to the original 420 kb interval of association (Orho-Melander et al. 2008). Functional characterisation has shown that the T2D-risk allele alters fructose-6-phosphate-mediated regulation of GKR, with consequences for glycolytic flux which explain the variant's effects on both glucose and lipid metabolism (Beer et al. 2009). The R325W variant in *SLC30A8* (Sladek et al. 2007) represents a further functionally active missense polymorphism which appears causal for its local association. *SLC30A8* encodes a zinc transporter, ZnT8, known to be expressed in the pancreatic islet and implicated in the proper function of β cell insulin granules (Nicolson et al. 2009). In mice, β cell-specific knockouts of *Znt8* are glucose intolerant and display defects in insulin production, crystallisation, packaging and secretion (Wijesekara et al. 2010), whilst the variant protein shows reduced zinc transport activity (Nicolson et al. 2009).

However, even the most obvious candidate variants must be treated with some caution, as the story of the T2D-associated E23K variant in *KCNJ11* demonstrates (Gloyn et al. 2003). All the right indications were there: a missense mutation in an excellent biological candidate gene, encoding a subunit of the K_{ATP} channel which is central to glucose-stimulated insulin secretion. However, fine-mapping efforts have demonstrated that in European, West African

and East Asian populations, E23K is in perfect LD with a second nonsynonymous variant (A1369S) within the adjacent gene *ABCC8*, which happens to encode the second protein component of the same K_{ATP} channel (Florez et al. 2004). Recent functional studies suggest that the *ABCC8* variant may, in fact, be the stronger candidate for mediating a T2D-risk effect (Hamming et al. 2009).

Progress has also been made at loci for which no obvious causal coding variant can be identified. At *TCF7L2*, fine-mapping studies have converged upon the intronic SNP rs7903146 as the most compelling candidate variant (Helgason et al. 2007). ChIP-Seq (chromatin immunoprecipitation sequencing) studies have shown that this variant maps within a region of islet-specific open chromatin, and that the two alleles differ in their capacity to achieve or maintain this state (Gaulton et al. 2010). Whilst alterations in chromatin state have not yet been shown to alter *TCF7L2* transcription, this seems the most likely mechanism. *TCF7L2* mRNA levels in human pancreatic islets increase with number of risk alleles and are fivefold higher in human islets isolated from T2D patients than those isolated from controls (Lyssenko et al. 2007). Over-expression of *TCF7L2* leads to reduced glucose-stimulated insulin secretion (Lyssenko et al. 2007) and, perhaps contradictorily, reduced apoptosis (Shu et al. 2008).

However, relatively few T2D- and obesity-susceptibility loci have been so obliging. The T2D association signal on chromosome 10q, for example, contains three genes—*HHEX*, *KIF11* and *IDE*—each with biological credibility for a role in T2D pathogenesis. Even the much-studied BMI association signal near *FTO* comprises a 47-kb linkage disequilibrium block which may be involved in regulation of the adjacent gene, *RPGRIP1L*, as well as *FTO* itself (Stratigopoulos et al. 2011). Whilst there is evidence from rodents that manipulation of *Fto* expression does indeed influence adiposity (Church et al. 2009, 2010), data from human studies are less persuasive (Meyre et al. 2010; Boissel et al. 2009). The T2D-susceptibility signal on chromosome 9p maps some 200 kb from the coding sequence of *CDKN2A* and *CDKN2B*, genes considered, on the basis of known biology, the most plausible candidates (Krishnamurthy et al. 2006; Harismendy et al. 2011). Whilst the discovery of a noncoding RNA (variously called *ANRIL* or *CDKN2BAS*), transcribed from the region of maximal diabetes association and thought to influence *CDKN2B* expression (Pasmant et al. 2007), highlights a potential disease mechanism, empirical evidence is so far not available. To be so reliant upon prior biology can feel a somewhat uncomfortable necessity for an experimental approach which explicitly set out to be biology-agnostic.

Despite challenges in establishing the relevant mechanism at each associated locus, the unbiased genome-wide

approach has offered considerable insight into the broad pathophysiological processes of disease pathogenesis. Analyses in normoglycaemic individuals have shown that most T2D-associated loci exert their primary effects on disease risk through reduced insulin secretion rather than increased insulin resistance (Perry and Frayling 2008; Voight et al. 2010), helping to address the long-standing debate over the relative roles of these processes in diabetes pathogenesis. Genes implicated in cell cycle regulation are overrepresented within T2D-associated regions, adding weight to the notion that control of β cell mass is central to the maintenance of normal glucose homeostasis (Voight et al. 2010). Amongst loci where the T2D-risk alleles are associated with reduced insulin sensitivity, only the signal at *FTO* seems to be driven by an effect on body mass index. At the other loci, including *KLF14* and *ADAMTS9*, association evidence has the potential to highlight entirely novel mechanisms connecting adipocyte and hepatic dysfunction to insulin resistance and diabetes.

Equivalent insights into the pathogenesis of obesity and the regulation of body mass index have been harder to derive. The fact that many of the most obvious positional candidates at BMI and obesity-associated loci have proven or suspected roles in the function of the central nervous system (CNS) is consistent with the known role of the hypothalamus in appetite regulation (Willer et al. 2009; Speliotes et al. 2010). For example, *NEGR1* is involved in neuronal growth (Marg et al. 1999), whilst *SH2B1* is involved in hypothalamic leptin signalling (Ren et al. 2007). *Sh2b1* knockout mice are obese, but the phenotype can be rescued by targeted expression of *Sh2b1* in neurons (Ren et al. 2007). These findings reinforce the view of common obesity as a disorder of behaviour rather than metabolism, mediated through hypothalamic dysregulation. In contrast, equivalent studies of fat distribution, rather than overall adiposity (Lindgren et al. 2009; Heid et al. 2010; Heard-Costa et al. 2009) have highlighted candidate transcripts implicated in the regulation of adipocyte development and function.

The growing power of techniques for genetic and functional evaluation of regional targets is likely to catalyse further successes in characterising causal variants and connecting them to the genes, pathways and networks they modulate. For example, transethnic fine mapping approaches, particularly in samples of African origin, should help to pin down the causal variants within common GWAS signals. Resequencing efforts will reveal novel loci, as well as rare, obviously functional, coding alleles in transcripts close to known loci that can be causally linked to the disease of interest (Nejentsev et al. 2009). At the same time, ever richer genomic characterisation of key tissues will help tie disease-associated variants to local transcript expression and other regulatory functions, whilst functional

studies in cellular systems and animal models will expand the ability to explore the physiological effects of genetic variation.

Nevertheless, for diseases such as diabetes and obesity, limited access to the tissues most obviously implicated in disease pathogenesis—the pancreatic β cell and hypothalamus, respectively—represents a serious obstacle to such studies. Advances in stem cell science offer the exciting prospect of overcoming this limitation through re-differentiation of patient-derived induced pluripotent stem (iPS) cells to generate authentic cellular models of key tissues. In parallel, ongoing large-scale sequencing studies are likely to reveal novel low frequency and rare risk alleles in coding sequence, some with larger effects than those encountered by existing GWAS. The expectation is that these will be inherently more amenable to experimental follow-up, accelerating the pace of functional discovery and delivering biological insights that will underpin the development of novel diagnostic and therapeutic options.

The second step: from biology to clinical practice

Public funding for genetic research is made available on the premise that the knowledge gained will improve our capacity to prevent and treat the conditions that afflict us, and it is against such translational advances that the success of research will ultimately be judged. Given that the interval between initial discovery and subsequent translational implementation typically exceeds 20 years (Contopoulos-Ioannidis et al. 2008), and since so many of the major genetic discoveries have emerged in the past few years, any attempt to draw up a “translational scoresheet” must be regarded as premature and provisional.

Broadly speaking, there are two principal ways in which genetic discoveries, and the biological insights they engender, can foster translational benefits. The first, and arguably more important, lies in using an improved understanding of disease pathogenesis as the basis for development of novel approaches for the diagnosis, monitoring, treatment and prevention of diabetes and obesity. With respect to drug targets, the fact that variants within *PPARG* and *KCNJ11* (encoding the targets of glitazones and sulfonylureas, two of the major classes of diabetes therapeutics), consistently emerge from genome wide association scans provides confirmation that unbiased genome-wide discovery efforts can reveal pathways capable of useful clinical manipulation. It is of course premature to expect newly discovered loci to have made this transition, not least because for most we are still unclear about the responsible transcript, but some promising candidates have emerged and excited interest from the pharmaceutical and biotechnical communities.

For example, variation in the melatonin receptor 1B gene (*MTNR1B*) is associated with insulin secretion, fasting glucose and T2D risk (Lyssenko et al. 2009; Prokopenko et al. 2009). *MTNR1B* expression is localised to the β cell within human islets and shows altered expression in islets from type 2 diabetic donors, whilst the receptor it encodes mediates the inhibitory effect of melatonin on glucose-stimulated insulin response (Lyssenko et al. 2009). Inhibition of this melatonin-ligand receptor system is therefore an attractive therapeutic option for T2D, particularly given that a dopamine receptor agonist which regulates melatonin content, bromocriptine (Zawilska and Iuvone 1990), is already approved for T2D therapy. In a similar vein, the demonstration that variants within the *SLC30A8* gene (encoding the islet Zn transporter ZnT8) are associated with reduced insulin secretion has highlighted the importance of zinc as a modulator of islet function and prompted efforts to explore the potential for pharmaceutical and public health approaches to treatment and prevention (Sun et al. 2009).

The second route to translational advance lies in the capacity to use genetic variation as a tool to explore inherited predisposition at the level of the individual and to use such information to deliver personalised medicine through more accurate diagnosis, better prognostication and/or therapeutic optimisation. To date, successful applications of personalised medicine in the clinical management of patients with diabetes and obesity are restricted to the highly familial, monogenic forms of disease.

The fundamental problem for efforts to build equivalent diagnostic and prognostic tools for more typical forms of diabetes and obesity lies in the modest effect sizes of the common variants so far discovered and therefore the limited proportion of heritable variance which they explain. *FTO*, the largest effect locus for obesity, is associated with an increase in adiposity of 0.3 BMI units (kg/m^2) per risk allele and explains only 0.34% of overall heritable variation (Willer et al. 2009). Consideration of all 32 currently known BMI-influencing loci increases this figure to only 1.45% (Speliotes et al. 2010). In combination, the 40 or so common variant loci implicated in T2D susceptibility explain a sibling relative risk of ~ 1.2 (Voight et al. 2010), well below epidemiological estimates that range between two and three (Köbberling 1982) and therefore define little more than 10% of the observed familial aggregation.

For useful clinical prediction, genetic testing must be both sensitive and specific in discriminating between those who will and will not develop diabetes or become obese on follow-up. The standard measure of this discriminative capacity is provided by receiver operating characteristic (ROC) curves, which plot the performance of a given diagnostic test in terms of those two factors. An area under the curve (AUC) of 0.5 indicates that the test performs no

better than chance, whilst an AUC of 1.0 describes a perfect test: in many clinical settings, an AUC of 0.8–0.9 is considered a “good” test. On these criteria, the discriminative performance of currently known markers is distinctly unimpressive (Tables 1, 2). For T2D, an AUC of approximately 0.60 for genetic information alone is typical and compares unfavourably with values close to 0.80 for conventional risk factors—age, BMI, gender—alone (Lango et al. 2008). Adding genetic information to a model that already incorporates those conventional risk factors leads to little further improvement in discriminative performance (Cornelis et al. 2009). The situation for obesity is little better: genetic factors alone produce an AUC of around 0.57 and add little to age and sex (Li et al. 2010).

However, it is clear that genetic variants so far identified do not substantially improve the discriminative accuracy of disease prediction based on clinical characteristics. Even genetic models which incorporate thousands of additional putative common variant association signals are likely to offer limited improvement (Evans et al. 2009). In some studies, genetic prediction has been shown to be slightly more effective in certain groups, such as the young (Meigs et al. 2008; de Miguel-Yanes et al. 2011) or those with increasing duration of follow-up (Lyssenko et al. 2008), but at the individual level, accuracy still falls well below any reasonable threshold for clinical utility. Enumeration of the extent and nature of any statistical interactions between genetic variation and environmental factors may allow for some improvement in prediction. However, the evidence to date for gene–environment interaction is limited, and arguably, genome-wide association meta-analysis will favour the detection of loci without appreciable heterogeneity of effects as a result of gene–environment interactions. Similarly, as data from a wider range of ethnicities become available, an awareness of the interaction of differing genetic and environmental risk factors will help to reveal how relative risk effects may vary across populations (Helgadóttir et al. 2006).

If individual prediction is not yet feasible, it is certainly true that risk allele scores have the capacity to highlight groups at either end of the risk distribution. Individuals carrying more than 12 T2D-risk alleles (in the highest quintile of the population according to loci known at the time) have twice the disease risk that would be predicted on the basis of BMI alone (Lyssenko et al. 2008). There is a fourfold difference in T2D risk between the top and bottom 1% of individuals in terms of risk allele score (Lango et al. 2008), and the 7–10% of children with three or four *FTO* and *MC4R* risk alleles have a threefold increased risk of childhood obesity compared to the 20–24% of individuals who carry no risk alleles (Cauchi et al. 2009). Consequently, existing genetic tests may have some utility in providing risk stratification at the group level, leading, for

Table 1 Studies assessing the contribution of genetic risk variant information to prediction of T2D disease status, as quantified by area under the curve (AUC) in receiver-operating characteristic (ROC) curve studies

Cohort	Population	Study type	Sample size	Variables included	Clinical data included	AUC clinical data only	AUC genetic variants only	AUC clinical plus genetic
Weedon et al. (2006)	Warren 2 and others	Caucasian (UK)	6,077	3	–	–	0.58	–
Cauchi et al. (2009)	SU.VI.MAX, DIAB2.NEPIR O.-GENE	Caucasian (French)	8,827	15	Age, sex, BMI	–	–	0.86
Lango et al. (2008)	GoDARTS	Caucasian (Scotland)	4,907	18	Age, sex, BMI	0.78	0.60	0.80
Lyssenko et al. (2008)	MPP, Botnia	Swedish/Finnish	18,831	16	Age, sex, BMI, diabetes family history, blood pressure, triglycerides, fasting glucose	0.74	0.62	0.75
Meigs (2008)	Framingham offspring	European ancestry (US)	2,377	18	Age, sex, BMI, diabetes family history, fasting glucose, blood pressure, HDL cholesterol, triglycerides	0.900	–	0.901
Vaxillaire et al. (2008)	DESIR	Caucasian, age 30–65	3,877	3 (out of 19 tested)	Age, sex, BMI	0.83	0.56	0.83
van Hoek et al. (2008)	Rotterdam study	Caucasian, age 55+	6,544	18	Age, sex, BMI	0.66	0.60	0.68
Cornelis et al. (2009)	HPFS, NHS	European ancestry (US)	6,310	10	Age, sex, BMI, diabetes family history, smoking, alcohol intake, physical activity	0.78	–	0.79
Lin et al. (2009)	CoLaus	Swiss	5,360	15	Age, BMI, family history, WHR, triglyceride/HDL-cholesterol ratio	0.86	0.59	0.87
Miyake et al. (2009)	Chinese/Japanese	Cross-sectional	4,686	11 (out of 23 tested)	Age, sex, BMI	0.68	0.63	0.72
Schulze et al. (2009)	EPIC-Potsdam	German	2,500	20	Age, weight, height, lifestyle factors, A1C, fasting glucose	0.89	–	0.89
Sparso et al. (2009)	Inter99, ADDITION	Danish	9,395	19	Age, sex, BMI	0.92	0.6	0.93
de Miguel-Yanes et al. (2011)	Framingham offspring	European ancestry (US)	3,471	40	Age, sex, BMI, diabetes family history, fasting glucose, triglycerides, blood pressure, HDL cholesterol	0.90	–	0.91

Table 2 Studies quantifying the contribution of genetic risk variant information to prediction of obesity (BMI ≥ 30 kg/m²) status, as quantified by area under the curve (AUC) in receiver-operating characteristic (ROC) curve studies

	Cohort	Population	Study type	Sample size	Variants included	Clinical data included	AUC clinical data only	AUC genetic variants only	AUC clinical plus genetic
Renstrom et al. (2009)	NSHED	Swedish	Cross-sectional	4,923	9	–	–	0.58	–
Cheung et al. (2010) ^a	CRISPS and others	Chinese	Cross-sectional	1,170	13	–	–	0.59	–
Li et al. (2010)	EPIC	European (UK)	Cross-sectional	20,431	12	Age, sex	0.57	0.57	0.60
Sandholt et al. (2010)	Inter99	Danish	Cross-sectional	6,514	20	Age, sex, diet, physical activity, smoking, education, employment, obesity drugs	0.67	0.58	0.69
Speliotes et al. (2010)	ARIC	Caucasian	Cross-sectional	8,120	32	Age, sex	0.52	0.57	0.58
Peterson (2011)	MGS-C	European and African-American	Cross-sectional	3,626	56 ^b	Age, sex, ethnicity	–	–	0.59

^a Obesity defined as BMI ≥ 27.5 kg/m²

^b Including variants of suggestive as well genome-wide significance from Willer et al. (2009) and Thorleifsson et al. (2009)

example, to design of more efficient clinical trials (Schork and Topol 2010). Since lifestyle interventions are often seen to be beneficial even for those at highest genetic risk (Hivert et al. 2011; Florez et al. 2006), identification of high risk groups could support targeting of preventative public health measures.

However, the main hope for improved prognostic and diagnostic precision, as for biological insight, lies with more complete enumeration of the genetic component of predisposition and the integration of such data with relevant information from other sources (including pertinent environmental exposures and epigenetic changes). In particular, a great deal of hope rests on the expectation that resequencing studies will reveal causal variants in the low (MAF 0.005–0.05) and rare (MAF < 0.005) frequency ranges. Whilst causal allele in the lowest MAF ranges will demand alternative discovery, analysis and translational approaches (Gloyn and McCarthy 2010), optimism remains that a proportion of the variants within them will have larger effect sizes than the common variant signals found to date and will therefore provide valuable boosts to predictive performance. Whereas over 400 common (MAF of 0.3) variants with allelic odds ratios between 1.05 and 1.10 would fail to provide a discriminative test with an AUC > 0.8, the inclusion of just ten rarer variants (MAF of 0.05) with odds ratios of 3.0 would achieve this benchmark (Janssens et al. 2006).

It is worth remembering that the most effective risk stratification strategies may not require genetic testing at all. In situations where genetic discoveries reveal a pathogenetic process that can be captured through a serum or urine biomarker, there may be considerable advantage (for instance, the ability to bypass the challenges to molecular diagnostics of locus and allelic heterogeneity) in focusing clinical attention on that instead. A measurement of serum LDL-cholesterol levels is far easier and has better predictive power than obtaining sequence and genotype data at the many loci now known to influence lipid levels. An early example of this approach comes from the identification of C-reactive protein (CRP) levels as a diagnostic marker for HNF1A-MODY, a specific subtype of monogenic diabetes. GWA studies (Ridker et al. 2008; Reiner et al. 2008) had shown that common variants near *HNF1A* have a strong influence on CRP levels, raising the possibility that CRP levels might be more dramatically disturbed in individuals carrying rare, large-effect *HNF1A* mutations causal for MODY. Confirmation that this is indeed the case (Owen et al. 2010) has exposed CRP as a diagnostic tool that can be used to screen individuals with early-onset diabetes and increase the confidence with which HNF1A-MODY patients can be passed for definitive molecular diagnostic testing.

If clinical prediction has not yet achieved general clinical utility, what of the hopes for therapeutic optimisation?

Monogenic forms of diabetes and obesity already provide some of the most dramatic success stories for pharmacogenetics: the transfer of children with permanent neonatal diabetes caused by mutations in *KCNJ11* and *ABCC8* from insulin to high-dose sulfonylureas (Sagen et al. 2004; Rafiq et al. 2008) and the treatment of morbid obesity due to leptin deficiency with recombinant leptin (Farooqi et al. 1999).

To date, pharmacogenetic studies in common forms of obesity and T2D have not offered such dramatic applications. The field is in its infancy, and with some notable exceptions, much of the work to date has been candidate driven, small scale, and has not displayed the inferential rigour seen in discovery GWAS. Even variation within *CYP2C9*, a long-established rate limiting enzyme for the metabolism of sulfonylureas, has been difficult to tie down to an effect upon sulfonylurea treatment outcomes (reviewed in Pearson 2009).

However, there have been some successes. In T2D, the presence of common polymorphisms in known diabetes drug targets has presented obvious candidates for pharmacogenetic analysis. Evidence of a relationship between *ABCC8/KCNJ11* genotype and sulfonylurea response is encouraging. Recent analyses in large cohorts have reported, for example, a 45% increased risk of glibenclamide treatment failure amongst risk compared to non-risk allele homozygotes (Sesti et al. 2006) and a greater decrease in fasting plasma glucose following gliclazide treatment amongst risk allele carriers (Feng et al. 2008). An effect upon gliclazide response is consistent with functional data which demonstrates that the risk variant K_{ATP} channel has 3.5 times increased sensitivity to gliclazide inhibition (Hamming et al. 2009).

Studies relating effects of *PPARG* variants on thiazolidinedione treatment have yielded inconsistent results (reviewed in Pearson 2009). More convincing are data demonstrating that genotype at the *TCF7L2* T2D risk locus is associated with variation in response to sulfonylurea treatment. In a retrospective observational study, patients carrying two risk alleles at *TCF7L2* were almost twice as likely to fail treatment objectives as those carrying no risk alleles, with an intermediate effect for heterozygotes (Pearson et al. 2007).

For metformin, polymorphisms within organic cation transporter 1 (*OCT1*) and the multidrug and toxin extrusion (MATE) 1 protein (*SLC47A1*) have been significantly associated with drug response (Shu et al. 2007; Jablonski et al. 2010; Becker et al. 2009). Most recently, a GWA study provided compelling evidence that variants near the *ATM* gene are significantly associated with glycaemic response to metformin therapy (Zhou et al. 2011). As with the *TCF7L2* association, the effect sizes are too modest to suggest clinical utility from the perspective of individual

therapeutic optimisation, not least because they only explain 2.5% of total variance in metformin response (Zhou et al. 2011), but implication of *ATM* may provide valuable clues to the mechanisms through which metformin exerts its beneficial effects on glucose metabolism.

Conclusion: where to from here?

In the past few years, human genetic research has begun to make progress in characterising the genetic basis of common forms of diabetes and obesity. In fact, the scale and power of the most recent association studies and meta-analyses mean that, as far as common variants are concerned, it is likely that future discoveries will be limited to alleles of rather small effect size: there are almost certainly no *FTOs* or *TCF7L2s* left to find, in European subjects at least. The first challenge to be faced lies in extending these genome-wide surveys to encompass a wider diversity of ethnic groups and more complete range of variation types, examined across the full allele frequency spectrum. Thanks to ever growing collaboration, and the falling costs and growing capacity of resequencing technologies, such studies are already well underway. We will soon be in possession of a far more complete and systematic view of the relationship between DNA sequence variation and individual predisposition to diseases such as diabetes and obesity. Application of these same tools within prospective studies and clinical trials should provide equivalent insights into the genetic basis of individual differences in the speed of disease progression, risk of disease-specific complications and response to therapeutic or preventative interventions.

This information brings with it new challenges. A list of associated loci and SNPs is of little value unless we can synthesise those findings into an improved model of disease biology. That means identifying the causal variants, and unravelling both the proximal (i.e. molecular) and distal (i.e. cellular and physiological) mechanisms whereby they execute their effects on disease predisposition. Given the large number of loci likely to be discovered, obtaining these mechanistic insights will require far better integration between high-throughput science (providing, for example, tissue-specific functional annotation of regulatory sequence, or large-scale siRNA manipulation of transcripts of interest) and the more detailed knowledge of “molecule-specific” domain experts.

All of this activity should be motivated by the need for clinical translation, particularly for diseases such as obesity and diabetes which represent major causes of global morbidity and mortality, and for which currently available therapeutic and preventative options are manifestly inadequate. In time, it is reasonable to expect that the more

complete model of disease pathogenesis which is emerging from human genetics discovery will lead to novel therapeutic, diagnostic and preventative approaches that are of widespread benefit.

The future of personalised medicine is less secure. At this stage, we simply do not know the circumstances under which knowledge of an individual's genome sequence will be of genuine clinical benefit. The answer is likely to depend on the genetic architecture of the disease in question and the relative contributions of inherited genetic variation (as opposed to other factors such as environment, somatic mutations and epigenetic changes) to predisposition. It will also depend on the health status of the individual: the value of a genome sequence in a perfectly healthy 40 year old is likely to be very different to that of a child afflicted by some serious genetic disease. Fortunately, such considerations need not be the subject of conjecture for too many years since the gathering pace of sequence data accumulation will soon provide empirical resolution.

Reservations about the current value of genetic testing have not, of course, prevented the commercialisation of personal genome analysis. Whilst decisions regarding access to such personal data are best left in the hands of the individuals concerned, the limitations of currently available tests (for now based around common variant arrays) are all too obvious. For instance, individual estimates of disease risk are not stable and often change as additional variants are discovered and incorporated into risk models (Mihaescu et al. 2009). In some ways, the relatively poor performance of current reagents has served to protect individuals and their health-care contacts from the more difficult decisions that are likely to come with access to sequence-level data. These include the appropriate personal response to the identification of rare alleles, particularly those that lie in genes which may invite medical action (e.g. *BRCA1*) or those with significant prognostic importance but for which therapeutic options are limited (e.g. *APOE* or *HTT*).

In our view, such concerns provide no justification for proscriptive regulation of access, but they do call for greater dialogue between the relevant parties and for aggressive educational efforts at all levels. In a recent survey, 48% of doctors in the United States reported that they were very likely to use an FDA-approved genetic test for a complex disease, 39% even before published evidence of clinical efficacy (Grant et al. 2009). Such enthusiasm sits poorly with the limited training of most current medical practitioners in the interpretation of genetic findings and the small numbers of the specialists (clinical geneticists, genetic counsellors) that might be expected to fill this gap.

From the perspective of the population at large, the omens seem good. In the same US study, over 70% of respondents said a “high risk” or “good responder” result

was very likely to improve their motivation to adopt lifestyle changes and adhere to medication regimes, whilst only 1.3% said they would be less motivated by a “low risk” result. These claims deserve to be treated with some degree of scepticism given that very sound diet and lifestyle advice is frequently ignored, but it is also possible that genetic data are more psychologically influential. Surveys of those who have sought and obtained genetic risk information based on common variant array genotyping have generally reported a positive experience, with little evidence of any of the adverse outcomes that some have predicted (Grant et al. 2009; Bloss et al. 2011).

In short, we have just begun a phase of genetic discovery which has the potential to transform the ways in which we manage diseases of global impact such as diabetes and obesity. The precise ways in which this transformation will play out are difficult to predict, but the ever-accelerating pace of human genetics discovery will reveal the landscape on which these developments will take place.

Acknowledgments MET is funded by a Medical Research Council studentship. MMC is supported by the Oxford NIHR Biomedical Research Centre.

Conflict of interest The authors declare that they have no conflict of interest.

References

- Adeyemo A, Chen G, Zhou J, Shriner D, Doumatey A, Huang H, Rotimi C (2010) FTO genetic variation and association with obesity in West Africans and African Americans. *Diabetes* 59:1549–1554
- Altshuler D, Hirschhorn JN, Klannemark M, Lindgren CM, Vohl MC, Nemesh J, Lane CR, Schaffner SF, Bolk S, Brewer C, Tuomi T, Gaudet D, Hudson TJ, Daly M, Groop L, Lander ES (2000) The common PPARgamma Pro12Ala polymorphism is associated with decreased risk of type 2 diabetes. *Nat Genet* 26:76–80
- Becker ML, Visser LE, van Schaik RH, Hofman A, Uitterlinden AG, Stricker BH (2009) Genetic variation in the multidrug and toxin extrusion 1 transporter protein influences the glucose-lowering effect of metformin in patients with diabetes: a preliminary study. *Diabetes* 58:745–749
- Been LF, Nath SK, Ralhan SK, Wander GS, Mehra NK, Singh J, Mulvihill JJ, Sanghera DK (2010) Replication of association between a common variant near melanocortin-4 receptor gene and obesity-related traits in Asian Sikhs. *Obesity* 18:425–429
- Beer NL, Tribble ND, McCulloch LJ, Roos C, Johnson PR, Orholm-Melander M, Gloyn AL (2009) 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* 18:4081–4088
- Benzinou M, Creemers JW, Choquet H, Lobbens S, Dina C, Durand E, Guerardel A, Boutin P, Jouret B, Heude B, Balkau B, Tichet J, Marre M, Potoczna N, Horber F, Le Stunff C, Czernichow S, Sandbaek A, Lauritzen T, Borch-Johnsen K, Andersen G, Kiess W, Korner A, Kovacs P, Jacobson P, Carlsson LM, Walley AJ, Jorgensen T, Hansen T, Pedersen O, Meyre D, Froguel P (2008)

- Common nonsynonymous variants in PCSK1 confer risk of obesity. *Nat Genet* 40:943–945
- Bloss CS, Schork NJ, Topol EJ (2011) Effect of direct-to-consumer genome-wide profiling to assess disease risk. *N Engl J Med* 364:524–534
- Boissel S, Reish O, Proulx K, Kawagoe-Takaki H, Sedgwick B, Yeo GS, Meyre D, Golzio C, Molinari F, Kadhon N, Etchevers HC, Saudek V, Farooqi IS, Froguel P, Lindahl T, O'Rahilly S, Munnich A, Colleaux L (2009) Loss-of-function mutation in the diacylglycerol-acyltransferase-encoding FTO gene causes severe growth retardation and multiple malformations. *Am J Hum Genet* 85:106–111
- Cauchi S, Stutzmann F, Cavalcanti-Proença C, Durand E, Pouta A, Hartikainen AL, Marre M, Vol S, Tammelin T, Laitinen J, Gonzalez-Izquierdo A, Blakemore AI, Elliott P, Meyre D, Balkau B, Jarvelin MR, Froguel P (2009) Combined effects of MC4R and FTO common genetic variants on obesity in European general populations. *J Mol Med* 87:537–546
- Chambers JC, Elliott P, Zabaneh D, Zhang W, Li Y, Froguel P, Balding D, Scott J, Kooner JS (2008) Common genetic variation near MC4R is associated with waist circumference and insulin resistance. *Nat Genet* 40:716–718
- Cheung CY, Tso AW, Cheung BM, Xu A, Ong KL, Fong CH, Wat NM, Janus ED, Sham PC, Lam KS (2010) Obesity susceptibility genetic variants identified from recent genome-wide association studies: implications in a Chinese population. *J Clin Endocrinol Metab* 95:1395–1403
- Church C, Lee S, Bagg EA, McTaggart JS, Deacon R, Gerken T, Lee A, Moir L, Mecnovic J, Quail MM, Schofield CJ, Ashcroft FM, Cox RD (2009) A mouse model for the metabolic effects of the human fat mass and obesity associated FTO gene. *PLoS Genet* 5:e1000599
- Church C, Moir L, McMurray F, Girard C, Banks GT, Teboul L, Wells S, Brüning JC, Nolan PM, Ashcroft FM, Cox RD (2010) Overexpression of Fto leads to increased food intake and results in obesity. *Nat Genet* 42(12):1086–1092
- Contopoulos-Ioannidis DG, Alexiou GA, Gouvas TC, Ioannidis JP (2008) Medicine. Life cycle of translational research for medical interventions. *Science* 321:1298–1299
- Cornelis MC, Qi L, Zhang C, Kraft P, Manson J, Cai T, Hunter DJ, Hu FB (2009) Joint effects of common genetic variants on the risk for type 2 diabetes in U.S. men and women of European ancestry. *Ann Intern Med* 150:541–550
- de Miguel-Yanes JM, Shrader P, Pencina MJ, Fox CS, Manning AK, Grant RW, Dupuis J, Florez JC, D'Agostino RB Sr, Cupples LA, Meigs JB (2011) Genetic risk reclassification for type 2 diabetes by age below or above 50 years using 40 type 2 diabetes risk single nucleotide polymorphisms. *Diabetes Care* 34:121–125
- Dickson SP, Wang K, Krantz I, Hakonarson H, Goldstein DB (2010) Rare variants create synthetic genome-wide associations. *PLoS Biol* 8:e1000294
- Dupuis J, Langenberg C, Prokopenko I, Saxena R, Soranzo N, Jackson AU, Wheeler E, Glazer NL, Bouatia-Naji N, Gloyn AL, Lindgren CM, Mägi R, Morris AP, Randall J, Johnson T, Elliott P, Rybin D, Thorleifsson G, Steinthorsdottir V, Henneman P, Grallert H, Dehghan A, Hottenga JJ, Franklin CS, Navarro P, Song K, Goel A, Perry JR, Egan JM, Lajunen T, Grarup N, Sparso T, Doney A, Voight BF, Stringham HM, Li M, Kanoni S, Shrader P, Cavalcanti-Proença C, Kumari M, Qi L, Timpson NJ, Gieger C, Zabena C, Rocheleau G, Ingelsson E, An P, O'Connell J, Luan J et al (2010) New genetic loci implicated in fasting glucose homeostasis and their impact on type 2 diabetes risk. *Nat Genet* 42:105–116
- Evans DM, Visscher PM, Wray NR (2009) Harnessing the information contained within genome-wide association studies to improve individual prediction of complex disease risk. *Hum Mol Genet* 18:3525–3531
- Farooqi IS, Jebb SA, Langmack G, Lawrence E, Cheetham CH, Prentice AM, Hughes IA, McCamish MA, O'Rahilly S (1999) Effects of recombinant leptin therapy in a child with congenital leptin deficiency. *N Engl J Med* 341:879–884
- Feng Y, Mao G, Ren X, Xing H, Tang G, Li Q, Li X, Sun L, Yang J, Ma W, Wang X, Xu X (2008) Ser1369Ala variant in sulfonylurea receptor gene ABCC8 is associated with antidiabetic efficacy of gliclazide in Chinese type 2 diabetic patients. *Diabetes Care* 31:1939–1944
- Flegal KM, Graubard BI, Williamson DF, Gail MH (2007) Cause-specific excess deaths associated with underweight, overweight, and obesity. *JAMA* 298:2028–2037
- Florez JC, Burt N, de Bakker PI, Almgren P, Tuomi T, Holmkvist J, Gaudet D, Hudson TJ, Schaffner SF, Daly MJ, Hirschhorn JN, Groop L, Altshuler D (2004) Haplotype structure and genotype-phenotype correlations of the sulfonylurea receptor and the islet ATP-sensitive potassium channel gene region. *Diabetes* 53:1360–1368
- Florez JC, Jablonski KA, Bayley N, Pollin TI, de Bakker PI, Shuldiner AR, Knowler WC, Nathan DM, Altshuler D (2006) TCF7L2 polymorphisms and progression to diabetes in the Diabetes Prevention Program. *N Engl J Med* 355:241–250
- Frayling TM, Timpson NJ, Weedon MN, Zeggini E, Freathy RM, Lindgren CM, Perry JR, Elliott KS, Lango H, Rayner NW, Shields B, Harries LW, Barrett JC, Ellard S, Groves CJ, Knight B, Patch AM, Ness AR, Ebrahim S, Lawlor DA, Ring SM, Ben-Shlomo Y, Jarvelin MR, Sovio U, Bennett AJ, Melzer D, Ferrucci L, Loos RJ, Barroso I, Wareham NJ, Karpe F, Owen KR, Cardon LR, Walker M, Hitman GA, Palmer CN, Doney AS, Morris AD, Smith GD, Hattersley AT, McCarthy MI (2007) A common variant in the FTO gene is associated with body mass index and predisposes to childhood and adult obesity. *Science* 316:889–894
- Gaulton KJ, Nammo T, Pasquali L, Simon JM, Giresi PG, Fogarty MP, Panhuis TM, Mieczkowski P, Secchi A, Bosco D, Berney T, Montanya E, Mohlke KL, Lieb JD, Ferrer J (2010) A map of open chromatin in human pancreatic islets. *Nat Genet* 42:255–259
- Geller F, Reichwald K, Dempfle A, Illig T, Vollmert C, Herpertz S, Siffert W, Platzer M, Hess C, Gudermann T, Biebermann H, Wichmann HE, Schafer H, Hinney A, Hebebrand J (2004) Melanocortin-4 receptor gene variant I103 is negatively associated with obesity. *Am J Hum Genet* 74:572–581
- Gloyn AL, McCarthy MI (2010) Variation across the allele frequency spectrum. *Nat Genet* 42:648–650
- Gloyn AL, Weedon MN, Owen KR, Turner MJ, Knight BA, Hitman G, Walker M, Levy JC, Sampson M, Halford S, McCarthy MI, Hattersley AT, Frayling TM (2003) 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* 52:568–572
- Gloyn AL, Pearson ER, Antcliff JF, Proks P, Bruining GJ, Slingerland AS, Howard N, Srinivasan S, Silva JM, Molnes J, Edghill EL, Frayling TM, Temple IK, Mackay D, Shield JP, Sumnik Z, van Rhijn A, Wales JK, Clark P, Gorman S, Aisenberg J, Ellard S, Njolstad PR, Ashcroft FM, Hattersley AT (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
- Grant SF, Thorleifsson G, Reynisdottir I, Benediktsson R, Manolescu A, Sainz J, Helgason A, Stefansson H, Emilsson V, Helgadóttir A, Styrkarsdóttir U, Magnússon KP, Walters GB, Palsdóttir E, Jonsdóttir T, Gudmundsdóttir T, Gylfason A, Saemundsdóttir J, Wilensky RL, Reilly MP, Rader DJ, Bagger Y, Christiansen C, Gudnason V, Sigurdsson G, Thorsteinsdóttir U, Gulcher JR,

- Kong A, Stefansson K (2006) Variant of transcription factor 7-like 2 (TCF7L2) gene confers risk of type 2 diabetes. *Nat Genet* 38:320–323
- Grant RW, Hivert M, Pandiscio JC, Florez JC, Nathan DM, Meigs JB (2009) The clinical application of genetic testing in type 2 diabetes: a patient and physician survey. *Diabetologia* 52:2299–2305
- Hamming KS, Soliman D, Maternisz LC, Niazi O, Lang Y, Gloyn AL, Light PE (2009) 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* 58:2419–2424
- Harismendy O, Notani D, Song X, Rahim NG, Tanasa B, Heintzman N, Ren B, Fu XD, Topol EJ, Rosenfeld MG, Frazer KA (2011) 9p21 DNA variants associated with coronary artery disease impair interferon-gamma signalling response. *Nature* 470:264–268
- Hassanein MT, Lyon HN, Nguyen TT, Akyzbekova EL, Waters K, Lettre G, Tayo B, Forrester T, Sarpong DF, Stram DO, Butler JL, Wilks R, Liu J, Le Marchand L, Kolonel LN, Zhu X, Henderson B, Cooper R, McKenzie C, Taylor HA Jr, Haiman CA, Hirschhorn JN (2010) Fine mapping of the association with obesity at the FTO locus in African-derived populations. *Hum Mol Genet* 19:2907–2916
- Heard-Costa NL, Zillikens MC, Monda KL, Johansson A, Harris TB, Fu M, Haritunians T, Feitosa MF, Aspelund T, Eiriksdottir G, Garcia M, Launer LJ, Smith AV, Mitchell BD, McArdle PF, Shuldiner AR, Bielinski SJ, Boerwinkle E, Brancati F, Demerath EW, Pankow JS, Arnold AM, Chen YD, Glazer NL, McKnight B, Psaty BM, Rotter JJ, Amin N, Campbell H, Gyllenstein U, Pattaro C, Pramstaller PP, Rudan I, Struchalin M, Vitart V, Gao X, Kraja A, Province MA, Zhang Q, Atwood LD, Dupuis J, Hirschhorn JN, Jaquish CE, O'Donnell CJ, Vasan RS, White CC, Aulchenko YS, Estrada K, Hofman A, Rivadeneira F, Uitterlinden AG, Witteman JC, Oostra BA, Kaplan RC, Gudnason V, O'Connell JR, Borecki IB, van Duijn CM, Cupples LA, Fox CS, North KE (2009) NRXN3 is a novel locus for waist circumference: a genome-wide association study from the CHARGE Consortium. *PLoS Genet* 5:e1000539
- Heid IM, Vollmert C, Hinney A, Doring A, Geller F, Lowel H, Wichmann HE, Illig T, Hebebrand J, Kronenberg F (2005) Association of the 1031 MC4R allele with decreased body mass in 7937 participants of two population based surveys. *J Med Genet* 42:e21
- Heid IM, Jackson AU, Randall JC, Winkler TW, Qi L, Steinthorsdottir V, Thorleifsson G, Zillikens MC, Speliotes EK, Mägi R, Workalemahu T, White CC, Bouatia-Naji N, Harris TB, Berndt SI, Ingelsson E, Willer CJ, Weedon MN, Luan J, Vedantam S, Esko T, Kilpeläinen TO, Kutalik Z, Li S, Monda KL, Dixon AL, Holmes CC, Kaplan LM, Liang L, Min JL, Moffatt MF, Molony C, Nicholson G, Schadt EE, Zondervan KT, Feitosa MF, Ferreira T, Allen HL, Weyant RJ, Wheeler E, Wood AR, MAGIC, Estrada K, Goddard ME, Lettre G, Mangino M, Nyholt DR, Purcell S, Smith AV, Visscher PM, Yang J, McCarroll SA, Nemesh J, Voight BF et al (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
- Helgadóttir A, Manolescu A, Helgason A, Thorleifsson G, Thorsteinsdóttir U, Gudbjartsson DF, Gretarsdóttir S, Magnusson KP, Gudmundsson G, Hicks A, Jonsson T, Grant SF, Sainz J, O'Brien SJ, Sveinbjornsdóttir S, Valdimarsson EM, Matthiasson SE, Levey AI, Abramson JL, Reilly MP, Vaccarino V, Wolfe ML, Gudnason V, Quyyumi AA, Topol EJ, Rader DJ, Thorgerirsson G, Gulcher JR, Hakonarson H, Kong A, Stefansson K (2006) A variant of the gene encoding leukotriene A4 hydrolase confers ethnicity-specific risk of myocardial infarction. *Nat Genet* 38:68–74
- Helgason A, Palsen S, Thorleifsson G, Grant SF, Emilsson V, Gunnarsdóttir S, Adeyemo A, Chen Y, Chen G, Reynisdóttir I, Benediktsson R, Hinney A, Hansen T, Andersen G, Borch-Johnsen K, Jorgensen T, Schafer H, Faruque M, Doumatey A, Zhou J, Wilensky RL, Reilly MP, Rader DJ, Bagger Y, Christiansen C, Sigurdsson G, Hebebrand J, Pedersen O, Thorsteinsdóttir U, Gulcher JR, Kong A, Rotimi C, Stefansson K (2007) Refining the impact of TCF7L2 gene variants on type 2 diabetes and adaptive evolution. *Nat Genet* 39:218–225
- Hivert MF, Jablonski KA, Perreault L, Saxena R, McAteer JB, Franks PW, Hamman RF, Kahn SE, Haffner S, Meigs JB, Altshuler D, Knowler WC, Florez JC (2011) Updated genetic score based on 34 confirmed type 2 diabetes loci is associated with diabetes incidence and regression to normoglycemia in the Diabetes Prevention Program. *Diabetes* 60:1340–1348
- International Diabetes Federation (2010) <http://www.diabetesatlas.org/datatables>
- Jablonski KA, McAteer JB, de Bakker PI, Franks PW, Pollin TI, Hanson RL, Saxena R, Fowler S, Shuldiner AR, Knowler WC, Altshuler D, Florez JC (2010) Common variants in 40 genes assessed for diabetes incidence and response to metformin and lifestyle intervention in the diabetes prevention program. *Diabetes* 59:2672–2681
- Janssens AC, Aulchenko YS, Elefante S, Borsboom GJ, Steyerberg EW, van Duijn CM (2006) Predictive testing for complex diseases using multiple genes: fact or fiction? *Genet Med* 8:395–400
- Köbberling JT, Tillil H (1982) Empirical risk figures for first degree relatives of non-insulin-dependent diabetics. In: Köbberling JT, Tattersall R (eds) *The genetics of diabetes mellitus*. Academic Press, London, pp 201–209
- Krishnamurthy J, Ramsey MR, Ligon KL, Torrice C, Koh A, Bonner-Weir S, Sharpless NE (2006) p16INK4a induces an age-dependent decline in islet regenerative potential. *Nature* 443:453–457
- Lango H, Palmer CN, Morris AD, Zeggini E, Hattersley AT, McCarthy MI, Frayling TM, Weedon MN (2008) Assessing the combined impact of 18 common genetic variants of modest effect sizes on type 2 diabetes risk. *Diabetes* 57:3129–3135
- Li S, Zhao JH, Luan J, Luben RN, Rodwell SA, Khaw KT, Ong KK, Wareham NJ, Loos RJ (2010) Cumulative effects and predictive value of common obesity-susceptibility variants identified by genome-wide association studies. *Am J Clin Nutr* 91:184–190
- Lin X, Song K, Lim N, Yuan X, Johnson T, Abderrahmani A, Vollenweider P, Stirnadel H, Sundseth SS, Lai E, Burns DK, Middleton LT, Roses AD, Matthews PM, Waeber G, Cardon L, Waterworth DM, Mooser V (2009) Risk prediction of prevalent diabetes in a Swiss population using a weighted genetic score—the CoLaus Study. *Diabetologia* 52:600–608
- Lindgren CM, Heid IM, Randall JC, Lamina C, Steinthorsdottir V, Qi L, Speliotes EK, Thorleifsson G, Willer CJ, Herrera BM, Jackson AU, Lim N, Scheet P, Soranzo N, Amin N, Aulchenko YS, Chambers JC, Drong A, Luan J, Lyon HN, Rivadeneira F, Sanna S, Timpson NJ, Zillikens MC, Zhao JH, Almgren P, Bandinelli S, Bennett AJ, Bergman RN, Bonnycastle LL, Bumpstead SJ, Chanock SJ, Cherkas L, Chines P, Coin L, Cooper C, Crawford G, Doering A, Dominiczak A, Doney AS, Ebrahim S, Elliott P, Erdos MR, Estrada K, Ferrucci L, Fischer G, Forouhi NG, Gieger C, Grallert H, Groves CJ, Grundy S, Guiducci C, Hadley D, Hamsten A et al (2009) Genome-wide association scan meta-analysis identifies three Loci influencing adiposity and fat distribution. *PLoS Genet* 5:e1000508
- Loos RJ, Lindgren CM, Li S, Wheeler E, Zhao JH, Prokopenko I, Inouye M, Freathy RM, Attwood AP, Beckmann JS, Berndt SI,

- Prostate, Lung, Colorectal, and Ovarian (PLCO) Cancer Screening Trial, Jacobs KB, Chanock SJ, Hayes RB, Bergmann S, Bennett AJ, Bingham SA, Bochud M, Brown M, Cauchi S, Connell JM, Cooper C, Smith GD, Day I, Dina C, De S, Dermizakis ET, Doney AS, Elliott KS, Elliott P, Evans DM, Sadaf Farooqi I, Froguel P, Ghorri J, Groves CJ, Gwilliam R, Hadley D, Hall AS, Hattersley AT, Hebebrand J, Heid IM, KORA, Lamina C, Gieger C, Illig T, Meitinger T, Wichmann HE, Herrera B, Hinney A, Hunt SE, Jarvelin MR, Johnson T, Jolley JD, Karpe F, Keniry A, Khaw KT et al (2008) Common variants near MC4R are associated with fat mass, weight and risk of obesity. *Nat Genet* 40:768–775
- Lyssenko V, Lupi R, Marchetti P, Del Guerra S, Orho-Melander M, Almgren P, Sjogren M, Ling C, Eriksson KF, Lethagen AL, Mancarella R, Berglund G, Tuomi T, Nilsson P, Del Prato S, Groop L (2007) Mechanisms by which common variants in the TCF7L2 gene increase risk of type 2 diabetes. *J Clin Invest* 117:2155–2163
- Lyssenko V, Jonsson A, Almgren P, Pulizzi N, Isomaa B, Tuomi T, Berglund G, Altshuler D, Nilsson P, Groop L (2008) Clinical risk factors, DNA variants, and the development of type 2 diabetes. *N Engl J Med* 359:2220–2232
- Lyssenko V, Nagorny CL, Erdos MR, Wierup N, Jonsson A, Spiegel P, Bugliani M, Saxena R, Fex M, Pulizzi N, Isomaa B, Tuomi T, Nilsson P, Kuusisto J, Tuomilehto J, Boehnke M, Altshuler D, Sundler F, Eriksson JG, Jackson AU, Laakso M, Marchetti P, Watanabe RM, Mulder H, Groop L (2009) Common variant in MTNR1B associated with increased risk of type 2 diabetes and impaired early insulin secretion. *Nat Genet* 41:82–88
- Maes HH, Neale MC, Eaves LJ (1997) Genetic and environmental factors in relative body weight and human adiposity. *Behav Genet* 27:325–351
- Marg A, Sirim P, Spaltmann F, Plagge A, Kauselmann G, Buck F, Rathjen FG, Brummendorf T (1999) Neurotractin, a novel neurite outgrowth-promoting Ig-like protein that interacts with CEPU-1 and LAMP. *J Cell Biol* 145:865–876
- McCarthy MI (2010) Genomics, type 2 diabetes, and obesity. *N Engl J Med* 363:2339–2350
- Meigs JB, Shrader P, Sullivan LM, McAteer JB, Fox CS, Dupuis J, Manning AK, Florez JC, Wilson PW, D'Agostino RB Sr, Cupples LA (2008) Genotype score in addition to common risk factors for prediction of type 2 diabetes. *N Engl J Med* 359:2208–2219
- Merikangas KR, Risch N (2003) Genomic priorities and public health. *Science* 302:599–601
- Meyre D, Delplanque J, Chevre JC, Lecoer C, Lobbens S, Gallina S, Durand E, Vatin V, Degraeve F, Proenca C, Gaget S, Korner A, Kovacs P, Kiess W, Tichet J, Marre M, Hartikainen AL, Horber F, Potoczna N, Hercberg S, Levy-Marchal C, Pattou F, Heude B, Tauber M, McCarthy MI, Blakemore AI, Montpetit A, Polychronakos C, Weill J, Coin LJ, Asher J, Elliott P, Jarvelin MR, Visvikis-Siest S, Balkau B, Sladek R, Balding D, Walley A, Dina C, Froguel P (2009) Genome-wide association study for early-onset and morbid adult obesity identifies three new risk loci in European populations. *Nat Genet* 41:157–159
- Meyre D, Proulx K, Kawagoe-Takaki H, Vatin V, Gutierrez-Aguilar R, Lyon D, Ma M, Choquet H, Horber F, Van Hul W, Van Gaal L, Balkau B, Visvikis-Siest S, Pattou F, Farooqi IS, Saudek V, O'Rahilly S, Froguel P, Sedgwick B, Yeo GS (2010) Prevalence of loss-of-function FTO mutations in lean and obese individuals. *Diabetes* 59:311–318
- Mihaescu R, van Hoek M, Sijbrands EJ, Uitterlinden AG, Witteman JC, Hofman A, van Duijn CM, Janssens AC (2009) Evaluation of risk prediction updates from commercial genome-wide scans. *Genet Med* 11:588–594
- Miyake K, Yang W, Hara K, Yasuda K, Horikawa Y, Osawa H, Furuta H, Ng MC, Hirota Y, Mori H, Ido K, Yamagata K, Hinokio Y, Oka Y, Iwasaki N, Iwamoto Y, Yamada Y, Seino Y, Maegawa H, Kashiwagi A, Wang HY, Tanahashi T, Nakamura N, Takeda J, Maeda E, Yamamoto K, Tokunaga K, Ma RC, So WY, Chan JC, Kamatani N, Makino H, Nanjo K, Kadowaki T, Kasuga M (2009) Construction of a prediction model for type 2 diabetes mellitus in the Japanese population based on 11 genes with strong evidence of the association. *J Hum Genet* 54:236–241
- Montague CT, Farooqi IS, Whitehead JP, Soos MA, Rau H, Wareham NJ, Sewter CP, Digby JE, Mohammed SN, Hurst JA, Cheetham CH, Earley AR, Barnett AH, Prins JB, O'Rahilly S (1997) Congenital leptin deficiency is associated with severe early-onset obesity in humans. *Nature* 387:903–908
- Myles S, Davison D, Barrett J, Stoneking M, Timpson N (2008) Worldwide population differentiation at disease-associated SNPs. *BMC Med Genomics* 1:22
- Nejentsev S, Walker N, Riches D, Egholm M, Todd JA (2009) Rare variants of IFIH1, a gene implicated in antiviral responses, protect against type 1 diabetes. *Science* 324:387–389
- Nicolson TJ, Bellomo EA, Wijesekara N, Loder MK, Baldwin JM, Gyulkhandanyan AV, Koshkin V, Tarasov AI, Carzaniga R, Kronenberger K, Taneja TK, da Silva Xavier G, Libert S, Froguel P, Scharfmann R, Stetsyuk V, Ravassard P, Parker H, Gribble FM, Reimann F, Sladek R, Hughes SJ, Johnson PR, Masseboeuf M, Burcelin R, Baldwin SA, Liu M, Lara-Lemus R, Arvan P, Schuit FC, Wheeler MB, Chimienti F, Rutter GA (2009) Insulin storage and glucose homeostasis in mice null for the granule zinc transporter ZnT8 and studies of the type 2 diabetes-associated variants. *Diabetes* 58:2070–2083
- O'Rahilly S, Farooqi IS (2006) Genetics of obesity. *Philos Trans R Soc Lond B Biol Sci* 361:1095–1105
- Orho-Melander M, Melander O, Guiducci C, Perez-Martinez P, Corella D, Roos C, Tewhey R, Rieder MJ, Hall J, Abecasis G, Tai ES, Welch C, Arnett DK, Lyssenko V, Lindholm E, Saxena R, de Bakker PI, Burt N, Voight BF, Hirschhorn JN, Tucker KL, Hedner T, Tuomi T, Isomaa B, Eriksson KF, Taskiran MR, Wahlstrand B, Hughes TE, Parmell LD, Lai CQ, Berglund G, Peltonen L, Vartiainen E, Jousilahti P, Havulinna AS, Salomaa V, Nilsson P, Groop L, Altshuler D, Ordovas JM, Kathiresan S (2008) Common missense variant in the glucokinase regulatory protein gene is associated with increased plasma triglyceride and C-reactive protein but lower fasting glucose concentrations. *Diabetes* 57:3112–3121
- Owen K, Hattersley AT (2001) Maturity-onset diabetes of the young: from clinical description to molecular genetic characterization. *Best Pract Res Clin Endocrinol Metab* 15:309–323
- Owen KR, Thanabalasingham G, James TJ, Karpe F, Farmer AJ, McCarthy MI, Gloyn AL (2010) Assessment of high-sensitivity C-reactive protein levels as diagnostic discriminator of maturity-onset diabetes of the young due to HNF1A mutations. *Diabetes Care* 33:1919–1924
- Pasmant E, Laurendeau I, Heron D, Vidaud M, Vidaud D, Bieche I (2007) Characterization of a germ-line deletion, including the entire INK4/ARF locus, in a melanoma-neural system tumor family: identification of ANRIL, an antisense noncoding RNA whose expression coclusters with ARF. *Cancer Res* 67:3963–3969
- Pearson ER (2009) Pharmacogenetics in diabetes. *Curr Diab Rep* 9:172–181
- Pearson ER, Starkey BJ, Powell RJ, Gribble FM, Clark PM, Hattersley AT (2003) Genetic cause of hyperglycaemia and response to treatment in diabetes. *Lancet* 362:1275–1281
- Pearson ER, Flechtner I, Njolstad PR, Malecki MT, Flanagan SE, Larkin B, Ashcroft FM, Klimes I, Codner E, Iotova V,

- Slingerland AS, Shield J, Robert JJ, Holst JJ, Clark PM, Ellard S, Sovik O, Polak M, Hattersley AT (2006) Switching from insulin to oral sulfonylureas in patients with diabetes due to Kir6.2 mutations. *N Engl J Med* 355:467–477
- Pearson ER, Donnelly LA, Kimber C, Whitley A, Doney AS, McCarthy MI, Hattersley AT, Morris AD, Palmer CN (2007) Variation in TCF7L2 influences therapeutic response to sulfonylureas: a GoDARTs study. *Diabetes* 56:2178–2182
- Perry JR, Frayling TM (2008) New gene variants alter type 2 diabetes risk predominantly through reduced beta-cell function. *Curr Opin Clin Nutr Metab Care* 11:371–377
- Peterson RE, Maes HH, Holmans P, Sanders AR, Levinson DF, Shi J, Kendler KS, Gejman PV, Webb BT (2011) Genetic risk sum score comprised of common polygenic variation is associated with body mass index. *Hum Genet* 129:221–230
- Poulsen P, Kyvik KO, Vaag A, Beck-Nielsen H (1999) Heritability of type II (non-insulin-dependent) diabetes mellitus and abnormal glucose tolerance—a population-based twin study. *Diabetologia* 42:139–145
- Prokopenko I, Langenberg C, Florez JC, Saxena R, Soranzo N, Thorleifsson G, Loos RJ, Manning AK, Jackson AU, Aulchenko Y, Potter SC, Erdos MR, Sanna S, Hottenga JJ, Wheeler E, Kaakinen M, Lyssenko V, Chen WM, Ahmadi K, Beckmann JS, Bergman RN, Bochud M, Bonnycastle LL, Buchanan TA, Cao A, Cervino A, Coin L, Collins FS, Crisponi L, de Geus EJ, Dehghan A, Deloukas P, Doney AS, Elliott P, Freimer N, Gateva V, Herder C, Hofman A, Hughes TE, Hunt S, Illig T, Inouye M, Isomaa B, Johnson T, Kong A, Krestyaninova M, Kuusisto J, Laakso M, Lim N, Lindblad U, Lindgren CM, McCann OT, Mohlke KL, Morris AD, Naitza S et al (2009) Variants in MTNR1B influence fasting glucose levels. *Nat Genet* 41:77–81
- Rafiq M, Flanagan SE, Patch AM, Shields BM, Ellard S, Hattersley AT (2008) Effective treatment with oral sulfonylureas in patients with diabetes due to sulfonylurea receptor 1 (SUR1) mutations. *Diabetes Care* 31:204–209
- Rankinen T, Zuberi A, Chagnon YC, Weisnagel SJ, Argyropoulos G, Walts B, Perusse L, Bouchard C (2006) The human obesity gene map: the 2005 update. *Obesity* 14:529–644
- Reiner AP, Barber MJ, Guan Y, Ridker PM, Lange LA, Chasman DI, Walston JD, Cooper GM, Jenny NS, Rieder MJ, Durda JP, Smith JD, Novembre J, Tracy RP, Rotter JJ, Stephens M, Nickerson DA, Krauss RM (2008) Polymorphisms of the HNF1A gene encoding hepatocyte nuclear factor-1 alpha are associated with C-reactive protein. *Am J Hum Genet* 82:1193–1201
- Ren D, Zhou Y, Morris D, Li M, Li Z, Rui L (2007) Neuronal SH2B1 is essential for controlling energy and glucose homeostasis. *J Clin Invest* 117:397–406
- Renstrom F, Payne F, Nordstrom A, Brito EC, Rolandsson O, Hallmans G, Barroso I, Nordstrom P, Franks PW (2009) Replication and extension of genome-wide association study results for obesity in 4923 adults from northern Sweden. *Hum Mol Genet* 18:1489–1496
- Ridker PM, Pare G, Parker A, Zee RY, Danik JS, Buring JE, Kwiatkowski D, Cook NR, Miletich JP, Chasman DI (2008) Loci related to metabolic-syndrome pathways including LEPR, HNF1A, IL6R, and GCKR associate with plasma C-reactive protein: the Women's Genome Health Study. *Am J Hum Genet* 82:1185–1192
- Sagen JV, Raeder H, Hathout E, Shehadeh N, Gudmundsson K, Baevre H, Abuelo D, Phornphutkul C, Molnes J, Bell GI, Gloyn AL, Hattersley AT, Molven A, Sovik O, Njolstad PR (2004) Permanent neonatal diabetes due to mutations in KCNJ11 encoding Kir6.2: patient characteristics and initial response to sulfonylurea therapy. *Diabetes* 53:2713–2718
- Sandholt CH, Sparso T, Grarup N, Albrechtsen A, Almind K, Hansen L, Toft U, Jorgensen T, Hansen T, Pedersen O (2010) Combined analyses of 20 common obesity susceptibility variants. *Diabetes* 59:1667–1673
- Saunders CL, Chiodini BD, Sham P, Lewis CM, Abkevich V, Adeyemo AA, de Andrade M, Arya R, Berenson GS, Blangero J, Boehnke M, Borecki IB, Chagnon YC, Chen W, Comuzzie AG, Deng HW, Duggirala R, Feitosa MF, Froguel P, Hanson RL, Hebebrand J, Huezo-Dias P, Kissebah AH, Li W, Luke A, Martin LJ, Nash M, Ohman M, Palmer LJ, Peltonen L, Perola M, Price RA, Redline S, Srinivasan SR, Stern MP, Stone S, Stringham H, Turner S, Wijmenga C, Collier DA (2007) Meta-analysis of genome-wide linkage studies in BMI and obesity. *Obesity* 15:2263–2275
- Saxena R, Voight BF, Lyssenko V, Burt NP, de Bakker PI, Chen H, Roix JJ, Kathiresan S, Hirschhorn JN, Daly MJ, Hughes TE, Groop L, Altshuler D, Almog P, Florez JC, Meyer J, Ardlie K, Bengtsson Bostrom K, Isomaa B, Lettre G, Lindblad U, Lyon HN, Melander O, Newton-Cheh C, Nilsson P, Orho-Melander M, Rastam L, Speliotes EK, Taskinen MR, Tuomi T, Guiducci C, Berglund A, Carlson J, Gianniny L, Hackett R, Hall L, Holmkvist J, Laurila E, Sjogren M, Sterner M, Surti A, Svensson M, Tewhey R, Blumenstiel B, Parkin M, Defelice M, Barry R, Brodeur W, Camarata J, Chia N, Fava M, Gibbons J, Handsaker B, Healy C, Nguyen K, Gates C, Sougnez C, Gage D, Nizzari M, Gabriel SB, Chirn GW, Ma Q, Parikh H, Richardson D, Ricke D, Purcell S (2007) Genome-wide association analysis identifies loci for type 2 diabetes and triglyceride levels. *Science* 316:1331–1336
- Scherag A, Dina C, Hinney A, Vatin V, Scherag S, Vogel CI, Muller TD, Grallert H, Wichmann HE, Balkau B, Heude B, Jarvelin MR, Hartikainen AL, Levy-Marchal C, Weill J, Delplanque J, Korner A, Kiess W, Kovacs P, Rayner NW, Prokopenko I, McCarthy MI, Schafer H, Jarick I, Boeing H, Fisher E, Reinehr T, Heinrich J, Rzehak P, Berdel D, Borte M, Biebermann H, Krude H, Roskopf D, Rimbach C, Rief W, Fromme T, Klingenspor M, Schurmann A, Schulz N, Nothen MM, Muhl-eisen TW, Erbel R, Jockel KH, Moebus S, Boes T, Illig T, Froguel P, Hebebrand J, Meyre D (2010) Two new loci for body-weight regulation identified in a joint analysis of genome-wide association studies for early-onset extreme obesity in French and German study groups. *PLoS Genet* 6:e1000916
- Schnyder S, Mullis PE, Ellard S, Hattersley AT, Fluck CE (2005) Genetic testing for glucokinase mutations in clinically selected patients with MODY: a worthwhile investment. *Swiss Med Wkly* 135:352–356
- Schork NJ, Topol EJ (2010) Genotype-based risk and pharmacogenetic sampling in clinical trials. *J Biopharm Stat* 20:315–333
- Schulze MB, Weikert C, Pischon T, Bergmann MM, Al-Hasani H, Schleicher E, Fritsche A, Haring HU, Boeing H, Joost HG (2009) Use of multiple metabolic and genetic markers to improve the prediction of type 2 diabetes: the EPIC-Potsdam Study. *Diabetes Care* 32:2116–2119
- Scott LJ, Mohlke KL, Bonnycastle LL, Willer CJ, Li Y, Duren WL, Erdos MR, Stringham HM, Chines PS, Jackson AU, Prokunina-Olsson L, Ding CJ, Swift AJ, Narisu N, Hu T, Pruim R, Xiao R, Li XY, Conneely KN, Riebow NL, Sprau AG, Tong M, White PP, Hetrick KN, Barnhart MW, Bark CW, Goldstein JL, Watkins L, Xiang F, Saramies J, Buchanan TA, Watanabe RM, Valle TT, Kinnunen L, Abecasis GR, Pugh EW, Doheny KF, Bergman RN, Tuomilehto J, Collins FS, Boehnke M (2007) A genome-wide association study of type 2 diabetes in Finns detects multiple susceptibility variants. *Science* 316:1341–1345
- Sesti G, Laratta E, Cardellini M, Andreozzi F, Del Guerra S, Irace C, Gnasso A, Grupillo M, Lauro R, Hribal ML, Perticone F, Marchetti P (2006) The E23K variant of KCNJ11 encoding the pancreatic beta-cell adenosine 5'-triphosphate-sensitive potassium channel subunit Kir6.2 is associated with an increased risk

- of secondary failure to sulfonylurea in patients with type 2 diabetes. *J Clin Endocrinol Metab* 91:2334–2339
- Shu Y, Sheardown SA, Brown C, Owen RP, Zhang S, Castro RA, Ianculescu AG, Yue L, Lo JC, Burchard EG, Brett CM, Giacomini KM (2007) Effect of genetic variation in the organic cation transporter 1 (OCT1) on metformin action. *J Clin Invest* 117:1422–1431
- Shu L, Sauter NS, Schulthess FT, Matveyenko AV, Oberholzer J, Maedler K (2008) Transcription factor 7-like 2 regulates beta-cell survival and function in human pancreatic islets. *Diabetes* 57:645–653
- Sladek R, Rocheleau G, Rung J, Dina C, Shen L, Serre D, Boutin P, Vincent D, Belisle A, Hadjadj S, Balkau B, Heude B, Charpentier G, Hudson TJ, Montpetit A, Pshezhetsky AV, Prentki M, Posner BI, Balding DJ, Meyre D, Polychronakos C, Froguel P (2007) A genome-wide association study identifies novel risk loci for type 2 diabetes. *Nature* 445:881–885
- Sparso T, Grarup N, Andreasen C, Albrechtsen A, Holmkvist J, Andersen G, Jorgensen T, Borch-Johnsen K, Sandbaek A, Lauritzen T, Madsbad S, Hansen T, Pedersen O (2009) Combined analysis of 19 common validated type 2 diabetes susceptibility gene variants shows moderate discriminative value and no evidence of gene-gene interaction. *Diabetologia* 52:1308–1314
- Speliotes EK, Willer CJ, Berndt SI, Monda KL, Thorleifsson G, Jackson AU, Allen HL, Lindgren CM, Luan J, Mägi R, Randall JC, Vedantam S, Winkler TW, Qi L, Workalemahu T, Heid IM, Steinthorsdottir V, Stringham HM, Weedon MN, Wheeler E, Wood AR, Ferreira T, Weyant RJ, Segrè AV, Estrada K, Liang L, Nemesh J, Park JH, Gustafsson S, Kilpeläinen TO, Yang J, Bouatia-Naji N, Esko T, Feitosa MF, Kutalik Z, Mangino M, Raychaudhuri S, Scherag A, Smith AV, Welch R, Zhao JH, Aben KK, Absher DM, Amin N, Dixon AL, Fisher E, Glazer NL, Goddard ME, Heard-Costa NL, Hoesel V, Hottenga JJ, Johansson A, Johnson T et al (2010) Association analyses of 249,796 individuals reveal 18 new loci associated with body mass index. *Nat Genet* 42:937–948
- Steinthorsdottir V, Thorleifsson G, Reynisdottir I, Benediktsson R, Jonsdottir T, Walters GB, Styrkarsdottir U, Gretarsdottir S, Emilsson V, Ghosh S, Baker A, Snorraddottir S, Bjarnason H, Ng MC, Hansen T, Bagger Y, Wilensky RL, Reilly MP, Adeyemo A, Chen Y, Zhou J, Gudnason V, Chen G, Huang H, Lashley K, Doumatey A, So WY, Ma RC, Andersen G, Borch-Johnsen K, Jorgensen T, van Vliet-Ostapchouk JV, Hofker MH, Wijmenga C, Christiansen C, Rader DJ, Rotimi C, Gurney M, Chan JC, Pedersen O, Sigurdsson G, Gulcher JR, Thorsteinsdottir U, Kong A, Stefansson K (2007) A variant in CDKAL1 influences insulin response and risk of type 2 diabetes. *Nat Genet* 39:770–775
- Stratigopoulos G, Leduc CA, Cremona ML, Chung WK, Leibel RL (2011) Cut-like homeobox 1 (CUX1) regulates expression of the fat mass and obesity-associated and retinitis pigmentosa GTPase regulator-interacting protein-1-like (RPGRIP1L) genes and coordinates leptin receptor signaling. *J Biol Chem* 286:2155–2170
- Stumvoll M, Goldstein BJ, van Haeften TW (2008) Type 2 diabetes: pathogenesis and treatment. *Lancet* 371:2153–2156
- Sun Q, van Dam RM, Willett WC, Hu FB (2009) Prospective study of zinc intake and risk of type 2 diabetes in women. *Diabetes Care* 32:629–634
- Thorleifsson G, Walters GB, Gudbjartsson DF, Steinthorsdottir V, Sulem P, Helgadottir A, Styrkarsdottir U, Gretarsdottir S, Thorlacius S, Jonsdottir I, Jonsdottir T, Olafsdottir EJ, Olafsdottir GH, Jonsson T, Jonsson F, Borch-Johnsen K, Hansen T, Andersen G, Jorgensen T, Lauritzen T, Aben KK, Verbeek AL, Roeleveld N, Kampman E, Yanek LR, Becker LC, Tryggvadottir L, Rafnar T, Becker DM, Gulcher J, Kiemenev LA, Pedersen O, Kong A, Thorsteinsdottir U, Stefansson K (2009) Genome-wide association yields new sequence variants at seven loci that associate with measures of obesity. *Nat Genet* 41:18–24
- Tsai FJ, Yang CF, Chen CC, Chuang LM, Lu CH, Chang CT, Wang TY, Chen RH, Shiu CF, Liu YM, Chang CC, Chen P, Chen CH, Fann CS, Chen YT, Wu JY (2010) A genome-wide association study identifies susceptibility variants for type 2 diabetes in Han Chinese. *PLoS Genet* 6:e1000847
- Unoki H, Takahashi A, Kawaguchi T, Hara K, Horikoshi M, Andersen G, Ng DP, Holmkvist J, Borch-Johnsen K, Jorgensen T, Sandbaek A, Lauritzen T, Hansen T, Nurbaya S, Tsunoda T, Kubo M, Babazono T, Hirose H, Hayashi M, Iwamoto Y, Kashiwagi A, Kaku K, Kawamori R, Tai ES, Pedersen O, Kamatani N, Kadowaki T, Kikkawa R, Nakamura Y, Maeda S (2008) SNPs in KCNQ1 are associated with susceptibility to type 2 diabetes in East Asian and European populations. *Nat Genet* 40:1098–1102
- van Hoek M, Dehghan A, Witteman JC, van Duijn CM, Uitterlinden AG, Oostra BA, Hofman A, Sijbrands EJ, Janssens AC (2008) Predicting type 2 diabetes based on polymorphisms from genome-wide association studies: a population-based study. *Diabetes* 57:3122–3128
- Vaxillaire M, Veslot J, Dina C, Proenca C, Cauchi S, Charpentier G, Tichet J, Fumeron F, Marre M, Meyre D, Balkau B, Froguel P (2008) Impact of common type 2 diabetes risk polymorphisms in the DESIR prospective study. *Diabetes* 57:244–254
- Voight BF, Scott LJ, Steinthorsdottir V, Morris AP, Dina C, Welch RP, Zeggini E, Huth C, Aulchenko YS, Thorleifsson G, McCulloch LJ, Ferreira T, Grallert H, Amin N, Wu G, Willer CJ, Raychaudhuri S, McCarroll SA, Langenberg C, Hofmann OM, Dupuis J, Qi L, Segrè AV, van Hoek M, Navarro P, Ardlie K, Balkau B, Benediktsson R, Bennett AJ, Blagieva R, Boerwinkle E, Bonnycastle LL, Bengtsson Boström K, Bravenboer B, Bumpstead S, Burtt NP, Charpentier G, Chines PS, Cornelis M, Couper DJ, Crawford G, Doney AS, Elliott KS, Elliott AL, Erdos MR, Fox CS, Franklin CS, Ganser M, Gieger C, Grarup N, Green T, Griffin S, Groves CJ et al (2010) Twelve type 2 diabetes susceptibility loci identified through large-scale association analysis. *Nat Genet* 42:579–589
- Walters RG, Jacquemont S, Valsesia A, de Smith AJ, Martinet D, Andersson J, Falchi M, Chen F, Andrieux J, Lobbens S, Delobel B, Stutzmann F, El-Sayed Moustafa JS, Chevre JC, Lecoeur C, Vatin V, Bouquillon S, Buxton JL, Boute O, Holder-Espinasse M, Cuisset JM, Lemaitre MP, Ambresin AE, Brioschi A, Gaillard M, Giusti V, Fellmann F, Ferrari N, Hadjikhani N, Campion D, Guilmatre A, Goldenberg A, Calmels N, Mandel JL, Le Caignec C, David A, Isidor B, Cordier MP, Dupuis-Girod S, Labalme A, Sanlaville D, Beri-Dexheimer M, Jonveaux P, Leheup B, Ounap K, Bochukova EG, Henning E, Keogh J, Ellis RJ, Macdermot KD, van Haelst MM, Vincent-Delorme C, Plessis G, Touraine R, Philippe A, Malan V, Mathieu-Dramard M, Chiesa J, Blaumeiser B, Kooy RF, Caiazzo R, Pigeyre M, Balkau B, Sladek R, Bergmann S, Mooser V, Waterworth D, Reymond A, Vollenweider P, Waeber G, Kurg A, Palta P, Esko T, Metspalu A, Nelis M, Elliott P, Hartikainen AL, McCarthy MI, Peltonen L, Carlsson L, Jacobson P, Sjostrom L, Huang N, Hurles ME, O'Rahilly S, Farooqi IS, Mannik K, Jarvelin MR, Pattou F, Meyre D, Walley AJ, Coin LJ, Blakemore AI, Froguel P, Beckmann JS (2010) A new highly penetrant form of obesity due to deletions on chromosome 16p11.2. *Nature* 463:671–675
- Wardle J, Carnell S, Haworth CM, Plomin R (2008) Evidence for a strong genetic influence on childhood adiposity despite the force of the obesogenic environment. *Am J Clin Nutr* 87:398–404
- Waters KM, Stram DO, Hassanein MT, Le Marchand L, Wilkens LR, Maskarinec G, Monroe KR, Kolonel LN, Altshuler D, Henderson BE, Haiman CA (2010) Consistent association of type 2

- diabetes risk variants found in Europeans in diverse racial and ethnic groups. *PLoS Genet* 6
- Weedon MN, McCarthy MI, Hitman G, Walker M, Groves CJ, Zeggini E, Rayner NW, Shields B, Owen KR, Hattersley AT, Frayling TM (2006) Combining information from common type 2 diabetes risk polymorphisms improves disease prediction. *PLoS Med* 3:e374
- Wijesekara N, Dai FF, Hardy AB, Giglou PR, Bhattacharjee A, Koshkin V, Chimienti F, Gaisano HY, Rutter GA, Wheeler MB (2010) Beta cell-specific Znt8 deletion in mice causes marked defects in insulin processing, crystallisation and secretion. *Diabetologia* 53:1656–1668
- Willer CJ, Speliotes EK, Loos RJ, Li S, Lindgren CM, Heid IM, Berndt SI, Elliott AL, Jackson AU, Lamina C, Lettre G, Lim N, Lyon HN, McCarroll SA, Papadakis K, Qi L, Randall JC, Roccascella RM, Sanna S, Scheet P, Weedon MN, Wheeler E, Zhao JH, Jacobs LC, Prokopenko I, Soranzo N, Tanaka T, Timpson NJ, Almgren P, Bennett A, Bergman RN, Bingham SA, Bonnycastle LL, Brown M, Burt NP, Chines P, Coin L, Collins FS, Connell JM, Cooper C, Smith GD, Dennison EM, Deodhar P, Elliott P, Erdos MR, Estrada K, Evans DM, Gianniny L, Gieger C, Gillson CJ, Guiducci C, Hackett R, Hadley D, Hall AS, Havulinna AS et al (2009) Six new loci associated with body mass index highlight a neuronal influence on body weight regulation. *Nat Genet* 41:25–34
- World Health Organisation (2010) <https://apps.who.int/infobase/>
- WTCCC (2007) Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature* 447:661–678
- Yamauchi T, Hara K, Maeda S, Yasuda K, Takahashi A, Horikoshi M, Nakamura M, Fujita H, Grarup N, Cauchi S, Ng DP, Ma RC, Tsunoda T, Kubo M, Watada H, Maegawa H, Okada-Iwabu M, Iwabu M, Shojima N, Shin HD, Andersen G, Witte DR, Jorgensen T, Lauritzen T, Sandbaek A, Hansen T, Ohshige T, Omori S, Saito I, Kaku K, Hirose H, So WY, Beury D, Chan JC, Park KS, Tai ES, Ito C, Tanaka Y, Kashiwagi A, Kawamori R, Kasuga M, Froguel P, Pedersen O, Kamatani N, Nakamura Y, Kadowaki T (2010) A genome-wide association study in the Japanese population identifies susceptibility loci for type 2 diabetes at UBE2E2 and C2CD4A–C2CD4B. *Nat Genet* 42:864–868
- Yasuda K, Miyake K, Horikawa Y, Hara K, Osawa H, Furuta H, Hirota Y, Mori H, Jonsson A, Sato Y, Yamagata K, Hinokio Y, Wang HY, Tanahashi T, Nakamura N, Oka Y, Iwasaki N, Iwamoto Y, Yamada Y, Seino Y, Maegawa H, Kashiwagi A, Takeda J, Maeda E, Shin HD, Cho YM, Park KS, Lee HK, Ng MC, Ma RC, So WY, Chan JC, Lyssenko V, Tuomi T, Nilsson P, Groop L, Kamatani N, Sekine A, Nakamura Y, Yamamoto K, Yoshida T, Tokunaga K, Itakura M, Makino H, Nanjo K, Kadowaki T, Kasuga M (2008) Variants in KCNQ1 are associated with susceptibility to type 2 diabetes mellitus. *Nat Genet* 40:1092–1097
- Zawilska J, Iuvone PM (1990) Alpha-2 adrenergic activity of bromocriptine and quinpirole in chicken pineal gland. Effects on melatonin synthesis and [3H]rauwolscine binding. *J Pharmacol Exp Ther* 255:1047–1052
- Zeggini E, Weedon MN, Lindgren CM, Frayling TM, Elliott KS, Lango H, Timpson NJ, Perry JR, Rayner NW, Freathy RM, Barrett JC, Shields B, Morris AP, Ellard S, Groves CJ, Harries LW, Marchini JL, Owen KR, Knight B, Cardon LR, Walker M, Hitman GA, Morris AD, Doney AS, McCarthy MI, Hattersley AT (2007) Replication of genome-wide association signals in UK samples reveals risk loci for type 2 diabetes. *Science* 316:1336–1341
- Zeggini E, Scott LJ, Saxena R, Voight BF, Marchini JL, Hu T, de Bakker PI, Abecasis GR, Almgren P, Andersen G, Ardlie K, Bostrom KB, Bergman RN, Bonnycastle LL, Borch-Johnsen K, Burt NP, Chen H, Chines PS, Daly MJ, Deodhar P, Ding CJ, Doney AS, Duren WL, Elliott KS, Erdos MR, Frayling TM, Freathy RM, Gianniny L, Grallert H, Grarup N, Groves CJ, Guiducci C, Hansen T, Herder C, Hitman GA, Hughes TE, Isomaa B, Jackson AU, Jorgensen T, Kong A, Kubalanza K, Kuruvilla FG, Kuusisto J, Langenberg C, Lango H, Lauritzen T, Li Y, Lindgren CM, Lyssenko V, Marvelle AF, Meisinger C, Midthjell K, Mohlke KL, Morken MA, Morris AD, Narisu N, Nilsson P, Owen KR, Palmer CN, Payne F, Perry JR, Pettersen E, Platou C, Prokopenko I, Qi L, Qin L, Rayner NW, Rees M, Roix JJ, Sandbaek A, Shields B, Sjogren M, Steinthorsdottir V, Stringham HM, Swift AJ, Thorleifsson G, Thorsteinsdottir U, Timpson NJ, Tuomi T, Tuomilehto J, Walker M, Watanabe RM, Weedon MN, Willer CJ, Illig T, Hveem K, Hu FB, Laakso M, Stefansson K, Pedersen O, Wareham NJ, Barroso I, Hattersley AT, Collins FS, Groop L, McCarthy MI, Boehnke M, Altshuler D (2008) Meta-analysis of genome-wide association data and large-scale replication identifies additional susceptibility loci for type 2 diabetes. *Nat Genet* 40:638–645
- Zhou K, Bellenguez C, Spencer CC, Bennett AJ, Coleman RL, Tavendale R, Hawley SA, Donnelly LA, Schofield C, Groves CJ, Burch L, Carr F, Strange A, Freeman C, Blackwell JM, Bramer E, Brown MA, Casas JP, Corvin A, Craddock N, Deloukas P, Dronov S, Duncanson A, Edkins S, Gray E, Hunt S, Jankowski J, Langford C, Markus HS, Mathew CG, Plomin R, Rautanen A, Sawcer SJ, Samani NJ, Trembath R, Viswanathan AC, Wood NW, Harries LW, Hattersley AT, Doney AS, Colhoun H, Morris AD, Sutherland C, Hardie DG, Peltonen L, McCarthy MI, Holman RR, Palmer CN, Donnelly P, Pearson ER (2011) Common variants near ATM are associated with glycemic response to metformin in type 2 diabetes. *Nat Genet* 43:117–120

Reduced Insulin Exocytosis in Human Pancreatic β -cells With Gene Variants Linked to Type 2 Diabetes

Anders H. Rosengren,¹ Matthias Braun,^{2,3} Taman Mahdi,¹ Sofia A. Andersson,¹ Mary E. Travers,² Makoto Shigeto,² Enming Zhang,¹ Peter Almgren,¹ Claes Ladenvall,¹ Annika S. Axelsson,¹ Anna Edlund,¹ Morten Gram Pedersen,¹ Anna Jonsson,¹ Reshma Ramracheya,² Yunzhao Tang,^{1,4} Jonathan N. Walker,^{2,5} Amy Barrett,² Paul R.V. Johnson,^{2,5,6} Valeriya Lyssenko,¹ Mark I. McCarthy,^{2,5,7} Leif Groop,¹ Albert Salehi,¹ Anna L. Gloyn,^{2,5} Erik Renström,¹ Patrik Rorsman,^{2,3,5} and Lena Eliasson¹

The majority of genetic risk variants for type 2 diabetes (T2D) affect insulin secretion, but the mechanisms through which they influence pancreatic islet function remain largely unknown. We functionally characterized human islets to determine secretory, biophysical, and ultrastructural features in relation to genetic risk profiles in diabetic and nondiabetic donors. Islets from donors with T2D exhibited impaired insulin secretion, which was more pronounced in lean than obese diabetic donors. We assessed the impact of 14 disease susceptibility variants on measures of glucose sensing, exocytosis, and structure. Variants near *TCF7L2* and *ADRA2A* were associated with reduced glucose-induced insulin secretion, whereas susceptibility variants near *ADRA2A*, *KCNJ11*, *KCNQ1*, and *TCF7L2* were associated with reduced depolarization-evoked insulin exocytosis. *KCNQ1*, *ADRA2A*, *KCNJ11*, *HHEX/IDE*, and *SLC2A2* variants affected granule docking. We combined our results to create a novel genetic risk score for β -cell dysfunction that includes aberrant granule docking, decreased Ca^{2+} sensitivity of exocytosis, and reduced insulin release. Individuals with a high risk score displayed an impaired response to intravenous glucose and deteriorating insulin secretion over time. Our results underscore the importance of defects in β -cell exocytosis in T2D and demonstrate the potential of cellular phenotypic characterization in the elucidation of complex genetic disorders.

The rapid increase in type 2 diabetes (T2D) incidence results from a combination of lifestyle factors and genetics. Recent genome-wide association studies have identified close to 50 loci associated with T2D risk (1–4). Many of the risk variants are associated with reduced insulin secretion in vivo and are located in the vicinity of genes that are expressed in

pancreatic β -cells (1–4). The insulin secretory defect in T2D is multifactorial and may involve a reduction in β -cell mass, impaired β -cell glucose sensing, and defects in the β -cell secretory machinery downstream of glucose metabolism (5–7). These defects are not easily distinguished in vivo studies, and the specific disease mechanisms coupled to T2D risk loci are therefore incompletely understood.

A comprehensive understanding of the pathophysiology of T2D requires detailed phenotype characterization at the cellular level in addition to the systemic profiling traditionally investigated in genetic association studies. Such investigations may result in the identification of patient subgroups with distinct cellular defects that can be targeted by specific therapies. An obstacle to achieving this goal is our limited knowledge about human β -cells and how they become perturbed in T2D. Most mechanistic studies published to date have relied on animal models or cell lines. However, the recent demonstration of important differences between human and rodent β -cells in terms of electrophysiological and secretory properties highlights the potential pitfalls of extrapolating observations in model systems to man (8).

In this study, we have correlated the function and genotype of human islets obtained from diabetic and nondiabetic (ND) donors. We have analyzed a panel of 14 gene variants robustly associated with T2D susceptibility identified by recent genetic association studies. We have identified four genetic variants that confer reduced β -cell exocytosis and six variants that interfere with insulin granule distribution. Based on these observations, we calculate a genetic risk score for islet dysfunction leading to T2D that involves decreased docking of insulin-containing secretory granules, impaired insulin exocytosis, and reduced insulin secretion.

RESEARCH DESIGN AND METHODS

Human islets. Experimental procedures were approved by the local ethical committees. Islets were obtained (with research consent) at Lund University Diabetes Centre and the Oxford Centre for Islet Transplantation from deceased donors. Donors with known T2D or $\text{HbA}_{1c} > 6.0\%$ and no GAD antibodies were defined as having T2D. Islets were cultured for 1–9 days in CMRL1066 (ICN Biomedicals, Irvine, CA) supplemented with 10 mmol/L HEPES, 2 mmol/L L-glutamine, 50 $\mu\text{g}/\text{mL}$ gentamicin, 0.25 $\mu\text{g}/\text{mL}$ Fungizone (GIBCO, New York, NY), 20 $\mu\text{g}/\text{mL}$ ciprofloxacin (Bayer, Leverkusen, Germany), and 10 mmol/L nicotinamide at 37°C.

Study populations. ND individuals (604) from the Botnia study (9), characterized with intravenous glucose tolerance tests (IVGTTs), were genotyped. The prospective part included 2,770 ND family members (9). The corrected insulin response (CIR) was estimated from an oral glucose tolerance test (10).

Islet isolation and in vitro insulin release. Insulin secretion was measured as described previously (11). In brief, batches of 10–12 islets were pre-incubated in Krebs-Ringer bicarbonate buffer (KRBB) supplemented with 2.8 (Lund) or 1 mmol/L glucose (Oxford) followed by a 1-h incubation in KRBB with 16.7 (Lund) or 20 mmol/L glucose (Oxford).

From the ¹Lund University Diabetes Centre, Department of Clinical Sciences Malmö, Lund University, Malmö, Sweden; the ²Oxford Centre for Diabetes, Endocrinology, and Metabolism, University of Oxford, Churchill Hospital, Oxford, U.K.; the ³Alberta Diabetes Institute, University of Alberta, Edmonton, Alberta, Canada; the ⁴Key Laboratory of Hormones and Development, Ministry of Health, China, Metabolic Diseases Hospital, Tianjin Medical University, Tianjin, China; the ⁵Oxford National Institute of Health Research Biomedical Research Centre, Churchill Hospital, Oxford, U.K.; the ⁶Nuffield Department of Surgery, University of Oxford, Oxford, U.K.; and the ⁷Wellcome Trust Centre for Human Genetics, University of Oxford, Oxford, U.K. Corresponding author: Anders H. Rosengren, anders.rosengren@med.lu.se, or Lena Eliasson, lena.eliasson@med.lu.se.

Received 27 October 2011 and accepted 25 February 2012.

DOI: 10.2337/db11-1516

This article contains Supplementary Data online at <http://diabetes.diabetesjournals.org/lookup/suppl/doi:10.2337/db11-1516/-/DC1>.

© 2012 by the American Diabetes Association. Readers may use this article as long as the work is properly cited, the use is educational and not for profit, and the work is not altered. See <http://creativecommons.org/licenses/by-nc-nd/3.0/> for details.

Islet cell preparation and electrophysiology. The electrophysiological measurements were conducted as described previously (8). β -cells were identified based on their size (8). For linopirdine experiments, cells were incubated with 50 $\mu\text{mol/L}$ of the compound for 1 h prior to experiments but linopirdine was not present during the recordings.

Transmission electron microscopy. Cultured islets were preincubated in KRBB with 2.8 mmol/L glucose for 30 min and then incubated in KRBB containing 16.7 mmol/L glucose for 1 h, fixed in glutaraldehyde, and added to freshly prepared Millonig buffer. Millonig buffer contains 2.26% NaH_2PO_4 and 2.52% NaOH (pH 7.2). Islets were postfixated in osmium tetroxide (1%) and embedded for electron microscopy as previously described (12). The insulin granule diameter was 285 ± 11 nm ($n = 6,957$ granules from 22 individuals).

ADP-ATP and insulin content measurements. The ADP-ATP ratio was measured in human islets as previously reported (13). In brief, islets were incubated in Connaught Medical Research Laboratories medium, washed in PBS, and then mixed with a nucleotide-releasing reagent. ATP levels were measured using a luminometer. The ADP was then converted to ATP by adding an ADP-converting reagent followed by ATP measurements. The insulin-DNA ratio was measured as previously described (14) and expressed as picomoles per microgram. Islets were sonicated and transferred to acidic ethanol, followed by insulin quantification by ELISA (Mercodia). DNA was quantified using Picogreen (Invitrogen).

RNA interference. Islets were transfected with *KCNQ1* small interfering RNA (siRNA) according to the protocol described previously (11). Gene silencing was assessed by quantitative PCR using TaqMan (Applied Biosystems).

Genotyping. Seventeen single-nucleotide polymorphisms (SNPs) were genotyped from genomic DNA on an ABI 7900 (Applied Biosystems).

Statistical procedures. The experimental data showed non-Gaussian distributions. Gaussian distribution was obtained using logarithm transformation. All statistical analyses were therefore performed using $\log_{10}$ -transformed data. Linear regression was used for the insulin secretion comparisons using the average secretion from each individual. The single-cell data, including electrophysiological recordings and granule distribution, were analyzed using a linear model in which all single-cell recordings were included as discrete observations but were grouped using donor ID as the subject cluster variable. β values for genotype-trait associations were obtained from the linear model. Capacitance recordings were obtained from at least four different cells per donor (median seven cells). Electron micrographs were analyzed from at least three different cells per donor (median seven cells). Complete phenotype data on insulin secretion, electrophysiology, and granule distribution were not available from all donors included in the study, which is why the number of donors used for the different trait analyses vary (see n for the different series). Exocytosis normalized to cell size was in the same range in donors from Oxford and Lund, but to account for any systematic differences, a binary covariate denoting the center was used in all analyses. The data presented denote nominal P values without correction for multiple comparisons. There were no interaction effects of the SNPs on the traits investigated. The risk score was calculated by adding the number of risk alleles. A linear model was used for the association analyses with the risk score and the traits. All statistical analyses were performed using SPSS Statistics (version 18.0).

RESULTS

Characterization of human islets from ND and diabetic donors. Human pancreatic islets from 72 ND and 28 T2D donors were collected at two international centers (Supplementary Tables 1 and 2). We attempted to measure as many parameters as possible, but islet numbers set a limit to the number of experiments that could be performed on each preparation. Insulin secretion from size-matched islets was measured under static conditions. In ND donors, high glucose stimulated insulin secretion 5.1-fold ($n = 42$ donors) over that seen at basal conditions (Fig. 1A). In islets from T2D donors, there was an $\sim 50\%$ reduction of glucose-induced insulin secretion ($P = 0.003$; $n = 42$ ND and 17 T2D donors) (Fig. 1A).

β -cell mass may be reduced in T2D (7,15). Indeed, we observed a tendency toward a reduced insulin content in T2D islets compared with those from ND control subjects; islet insulin content averaged 5.6 ± 0.5 ($n = 41$) vs. 4.0 ± 0.6 pmol insulin per μg DNA ($n = 12$) in ND and T2D islets, respectively. Although insulin content is not equivalent to β -cell mass, the $\sim 30\%$ decrease is close to the 35% reduction in islet β -cell mass recently reported (7) but did not reach

statistical significance ($P = 0.11$). However, insulin release at high glucose was still reduced (by 50%) from T2D islets after correction for islet insulin content ($P = 0.008$). Thus, both insulin secretion and content are reduced in T2D.

It has been suggested that insulin secretion in vivo may be upregulated in obese individuals to compensate for insulin resistance (7,15). We therefore divided the donors into those with a BMI above or below the upper quartile (31 kg/m^2). There was no significant difference in glucose-induced insulin secretion between nonobese and obese donors in the combined dataset of all donors (1.4 ± 0.1 ng/islet/h vs. 1.9 ± 0.6 ng/islet/h for nonobese and obese donors, respectively; not statistically different). However, when stratified by diabetes status, the nonobese diabetic donors ($n = 12$) displayed reduced insulin release compared with islets from obese diabetic subjects ($n = 5$; $P = 0.034$) (Fig. 1A).

Characterization of single β -cell exocytosis. Insulin is released in a characteristic biphasic manner: a rapid and transient first phase is followed by a slowly developed but sustained second phase. The consensus model for glucose-induced insulin secretion suggests that glucose triggers insulin secretion via increased cytosolic ATP, closure of ATP-sensitive potassium channels (K_{ATP} channels), initiation of electrical activity, and Ca^{2+} -dependent exocytosis from the β -cells (Supplementary Fig. 1) (16). In addition, several mechanisms fine-tune the release competence and intracellular distribution of the insulin granules. These two processes are referred to as the triggering and amplifying pathways. A current hypothesis postulates that first-phase insulin secretion reflects the release of granules docked beneath the plasma membrane, whereas second phase requires physical recruitment of granules from the cell interior and/or chemical modification (priming) of granules already in place beneath the membrane (16), although this hypothesis is not uncontested (17). There is evidence that T2D interferes with both the proximal (triggering) and distal (amplifying) mechanisms (18).

We examined exocytosis by measurements of cell capacitance that monitor the increase in cell area that results when secretory granules fuse with the plasma membrane. The standard whole-cell configuration was used, and the cytoplasmic ATP and cAMP levels were thus clamped to the concentrations of the pipette-filling solutions, allowing the exocytotic capacity to be analyzed without the confounding effect of glucose metabolism and variation in receptor signaling mediated by cAMP (e.g., glucagon-like peptide 1). Despite the clear difference in insulin secretion between islets from T2D and non-T2D donors, there was no overall change in β -cell exocytosis between the two groups (Fig. 1B).

We also analyzed electron micrographs of human islets. There was a strong correlation between BMI and the number of lipid droplets in β -cells ($P = 0.008$; $n = 14$ ND and 10 T2D donors). In agreement with a recent study (19), there was a correlation between the amount of lipofuscin bodies in β -cells and donor age ($P = 0.012$). However, no differences were observed in the total granule number or number of granules docked below the plasma membrane between ND and T2D islets (Fig. 1C). The islet ADP-ATP ratio at 5.6 mmol/L glucose was not different between ND and T2D islets and averaged 0.08 ± 0.01 ($n = 42$) vs. 0.07 ± 0.01 ($n = 14$), respectively.

Analysis of islet physiology in genetic subgroups. T2D may result from a combination of genetic factors that differentially affect glucose sensing, exocytosis, and β -cell

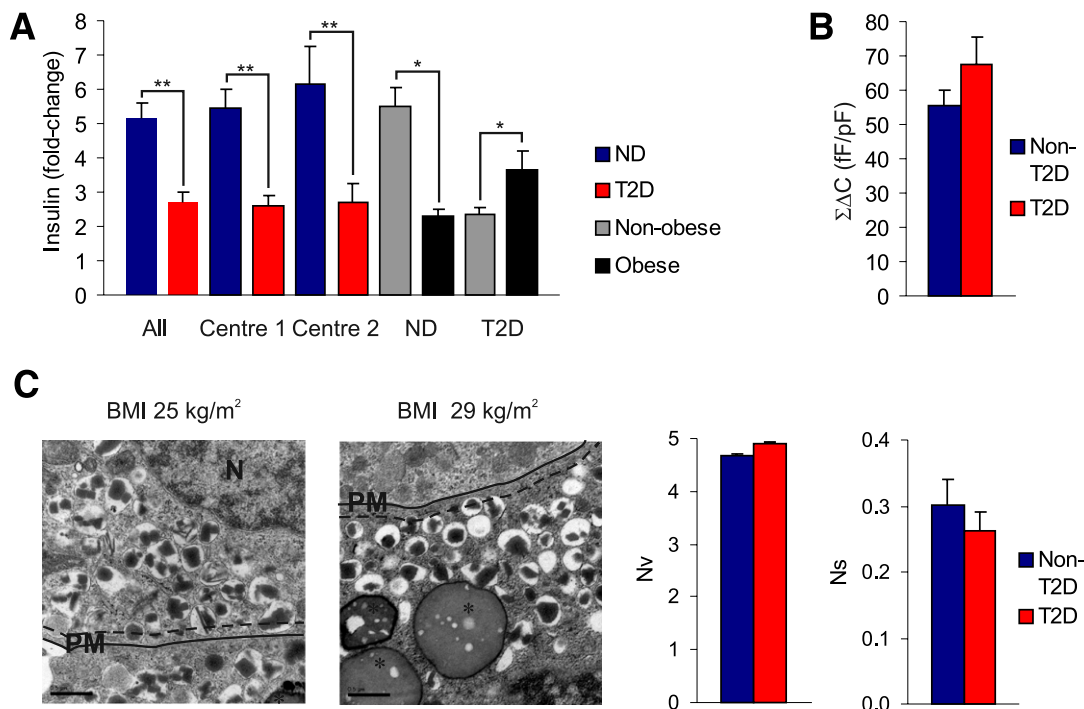


FIG. 1. Characterization of islets from diabetic and ND individuals. **A:** Fold-stimulation of insulin secretion in batch-incubated islets from non-T2D ($n = 42$) and T2D ($n = 17$) donors, nonobese ($n = 37$) and obese ($n = 5$) ND donors, as well as nonobese ($n = 12$) and obese T2D donors ($n = 5$). Fold-stimulation of insulin secretion in non-T2D and T2D donors separated by center (1, Lund; 2, Oxford) is also displayed. **B:** Total exocytosis evoked by a train of 10 depolarizations from -70 to 0 mV ($\Sigma \Delta C$) ($n = 189$ and 60 cells from 28 non-T2D and 8 T2D donors, respectively). **C:** Electron micrographs of human islets from non-T2D donors with low (25 kg/m^2) and high (29 kg/m^2) BMI. Lipid droplets (*), nucleus (N), and plasma membrane (PM, solid line) are indicated; dashed lines show 150-nm distance from PM. Scale bars, $0.5 \mu\text{m}$. The histograms show the total granule number as volume density (Nv; granules/ μm^3) and the number of docked granules as surface density (Ns; granules/ μm^2) in β -cells from non-T2D and T2D donors ($n = 75$ and 42 cells from 12 and 6 donors, respectively). A granule was defined as docked when the center of the granule was located within 150 nm from the plasma membrane. Data are presented as means $\pm$ SEM. * $P < 0.05$ and ** $P < 0.01$.

mass. Disease heterogeneity might thus explain our inability to detect any specific defect accounting for the lowered glucose-induced insulin secretion when analyzing the entire cohort of diabetic donors. Therefore, we examined whether stratification of the donors based on genetic susceptibility could facilitate the identification of defects in single-cell phenotypes.

Fourteen SNPs (Supplementary Table 3) were selected based on their association with reduced insulin secretion in vivo in previous genetic studies (1–4,11). There are multiple ways by which these variants could potentially affect insulin secretion, and it is currently unclear whether some T2D susceptibility variants reduce insulin secretion through direct effects on insulin exocytosis in human β -cells.

There was no interaction of the genotype-phenotype associations with T2D status among the donors as analyzed by a linear interaction model. Based on that observation and for reasons of statistical power, both ND and T2D islets were included in the genotype analyses. However, diabetes status was used as a binary covariate in all analyses. BMI and age were also used as covariates. Associations were analyzed by a linear model where the β values indicate the effect of genotype on the studied parameters. **Genetic variants associated with insulin secretion.** First we investigated the effect of T2D-associated alleles on glucose-stimulated insulin secretion in isolated islets. Human islets from donors with the *TCF7L2* rs7903146 risk genotype, the genetic variant with the largest effect on T2D susceptibility known to date (20,21), displayed significantly reduced insulin secretion ($P = 0.002$; $\beta = -0.199$ [95% CI -0.324 to -0.075]; $n = 47$ donors) (Fig. 2A and Table 1).

There was also a significant reduction in insulin secretion in islets from risk carriers for *ADRA2A* rs553668 ($P = 0.017$; $\beta = -0.165$ [-0.300 to -0.030]; $n = 50$) (Fig. 2B and Table 1), in accordance with previous data (11).

Identification of genetic variants influencing β -cell exocytosis and insulin granule docking. Exocytosis and insulin granule distribution were studied in β -cells from different genotype carriers. Total exocytosis was reduced in *TCF7L2* rs7903146 risk allele carriers ($P = 0.007$; $n = 249$ cells from 36 donors) (Fig. 2C, Table 1, and Supplementary Table 4), whereas intracellular granule distribution was unaffected (Table 1, Supplementary Table 5, and Supplementary Fig. 2), echoing recent studies in rodent cells (22). The *TCF7L2* risk allele did not affect the depolarization-evoked integrated Ca^{2+} current ($P = 0.5$; $\beta = -0.031$ [95% CI -0.114 to -0.052]; $n = 188$ cells from 28 donors) (Supplementary Table 4). We therefore conclude that the *TCF7L2* risk variant interferes with exocytosis at a late stage. In keeping with this idea, the Ca^{2+} sensitivity of exocytosis (estimated from the ratio between exocytosis and the integrated Ca^{2+} current during the first depolarization) was reduced in nonobese (BMI $< 31 \text{ kg/m}^2$) risk allele carriers ($P = 0.002$; $\beta = -0.321$ [-0.526 to -0.116]; $n = 90$ cells from 13 donors) (Supplementary Table 4).

The *ADRA2A* rs553668 risk allele tended to reduce exocytosis ($P = 0.059$; $n = 249$ cells from 36 donors) (Fig. 2D and Table 1), an effect more pronounced in β -cells from nonobese donors where it attained statistical significance ($P = 2 \times 10^{-6}$; $n = 170$ cells from 24 donors; $\beta = -0.433$ compared with -0.160 for the entire cohort) (Supplementary Table 4). Both rapid exocytosis (estimated as the

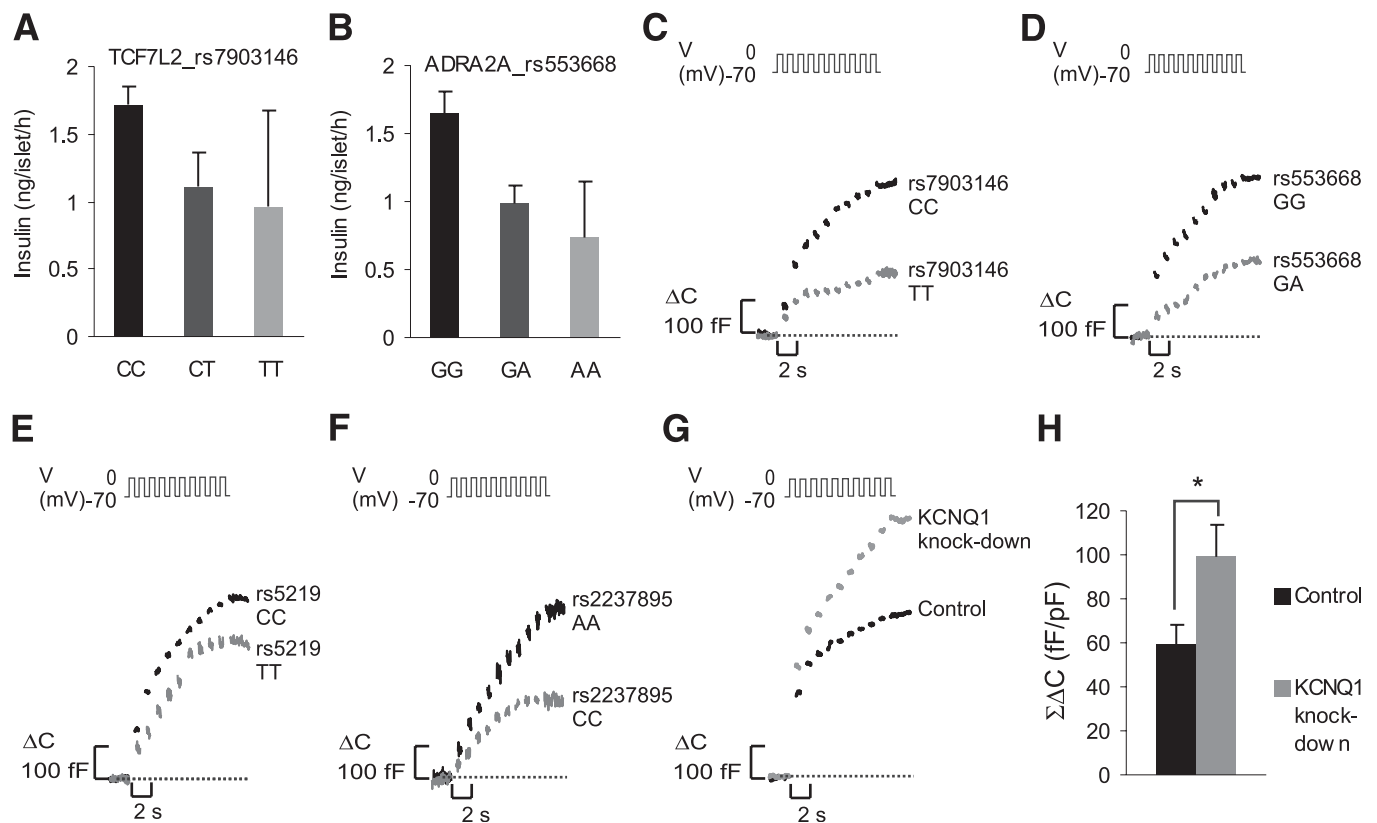


FIG. 2. Genotype effects on insulin secretion and β -cell exocytosis. **A:** Glucose-stimulated insulin secretion in islets from different genotype carriers of rs7903146 (T is risk allele). Data from 28 CC, 17 CT, and 2 TT carriers. **B:** Glucose-stimulated insulin secretion in islets from donors with different genotype for rs553668 (A is risk allele). Data from 36 GG, 12 GA, and 2 AA carriers. **C–F:** Depolarization-evoked increase in cell capacitance (ΔC) in β -cells representative for individuals carrying risk or nonrisk alleles for rs7903146 (*TCF7L2*), rs553668 (*ADRA2A*), rs5219 (*KCNJ11*), and rs2237895 (*KCNQ1*). **G** and **H:** Exocytotic response in human β -cells treated with control siRNA or siRNA targeting *KCNQ1*. Histogram shows average total exocytosis ($\Sigma\Delta C$) from 22 and 21 cells per group, respectively. * $P < 0.05$.

response to the first two depolarizations in the train) and slow exocytosis (the response to pulses 3–10), proposed to correlate with first- and second-phase insulin secretion (16), were affected ($P = 0.004$ and $P = 8 \times 10^{-8}$). There was no

change in the integrated Ca^{2+} currents, whereas the Ca^{2+} sensitivity of exocytosis was reduced in nonobese risk allele carriers ($P = 3 \times 10^{-5}$) (Supplementary Table 4). Furthermore, the risk allele was associated with impaired docking

TABLE 1
Genotype effects on β -cell phenotypes in the complete cohort

SNP	Nearest gene	Insulin secretion ($n = 47$ –50 donors)		Total exocytosis* ($n = 249$ cells from 36 donors)		Ns† ($n = 97$ cells from 18 donors)	
		P value	Effect‡	P value	Effect	P value	Effect
rs7903146§	<i>TCF7L2</i>	0.002	−0.530	0.007	−6.40	0.40	−0.004
rs553668	<i>ADRA2A</i>	0.017	−0.575	0.059	−15.4	0.03	−0.14
rs5219	<i>KCNJ11</i>	0.776	0.109	0.021	−7.00	0.05	−0.12
rs2237895	<i>KCNQ1</i>	0.484	−0.171	0.016	−10.1	0.04	−0.10
rs10946398	<i>CDKAL1</i>	0.153	−0.159	0.414	−1.73	0.187	0.081
rs2191349	<i>DGKB</i>	0.229	0.401	0.301	−14.3	0.381	0.101
rs560887	<i>G6PC2</i>	0.519	0.226	0.676	−4.97	0.06	−0.108
rs4607517	<i>GCK</i>	0.213	0.333	0.667	8.33	0.74	0.002
rs10423928	<i>GIPR</i>	0.972	−0.053	0.849	−4.56	0.8	0.004
rs1111875	<i>HHEX/IDE</i>	0.164	0.121	0.086	−10.9	0.016	−0.054
rs231362	<i>KCNQ1</i>	0.312	0.130	0.319	6.03	0.011	0.073
rs10830963	<i>MTNR1B</i>	0.851	0.001	0.655	0.129	0.074	0.095
rs11920090	<i>SLC2A2</i>	0.829	0.040	0.57	13.4	0.037	0.058
rs13266634	<i>SLC30A8</i>	0.662	0.084	0.819	−1.71	0.72	0.018

*Exocytosis measured as the total increase in single β -cell capacitance in response to 10 depolarizations. †Number of docked insulin granules. ‡Effect of each additional risk allele on insulin secretion (in ng/islet/h), exocytosis (fF/pF), or granule distribution (Ns, granules/ μm^2) estimated from the linear model. §SNPs in boldface were used to calculate the genetic risk score.

of insulin granules studied by electron microscopy ($P = 0.025$; $n = 97$ cells from 18 donors) (Table 1 and Supplementary Fig. 2). These data add novel insights into the effect of *ADRA2A* rs553668 on human β -cell function by demonstrating that the risk variant interferes with late-stage exocytosis via impaired granule docking. The defect was particularly pronounced in nonobese individuals, which is in line with patient data showing a stronger effect of rs553668 on T2D risk in lean subjects (11).

Carriers of the *KCNJ11* rs5219 T allele display reduced β -cell exocytosis ($P = 0.021$; $n = 249$ cells from 36 donors), particularly during the first two pulses of the train (Fig. 2E and Table 1). In addition, the Ca^{2+} sensitivity of exocytosis was reduced ($P = 0.06$ for entire cohort and $P = 0.004$ for nonobese donors; $n = 24$) (Supplementary Table 4). *KCNJ11* encodes Kir6.2, which forms part of the K_{ATP} channel that regulates β -cell membrane potential. Activating mutations in *KCNJ11* are a cause of neonatal diabetes (23), and the nonsynonymous rs5219 risk allele of the channel (E23K) combined with the S1369A risk allele of the neighboring *ABCC8* (encoding the sulfonylurea receptor 1) exhibits decreased ATP inhibition (24). Since membrane potential is clamped in the capacitance measurements of exocytosis, our findings cannot be explained by immediate effects on electrical activity. Rather, we favor the idea that the risk allele results in long-term changes in electrical activity that may perturb granule distribution. The findings that β -cells from risk genotype donors had a decreased number of docked granules (35% reduction per risk allele; $P = 0.05$; $n = 97$ cells from 18 donors) (Table 1, Supplementary Table 5, and Supplementary Fig. 2) and that the exocytotic defect was restricted to the first two pulses support this hypothesis. There is some evidence that K_{ATP} channels may locate to the secretory granules, a finding that may be related to the proposal that sulfonylureas modulate insulin release by a direct effect on exocytosis (25).

The risk allele of rs2237895 (26,27), located in intron 15 of *KCNQ1*, was also associated with impaired exocytosis ($P = 0.016$) (Fig. 2F and Table 1), an effect that was particularly pronounced among nonobese risk allele carriers ($P = 0.001$) (Supplementary Table 4). The reduction was stronger for rapid exocytosis ($P = 0.011$ for pulses 1 and 2 and $P = 0.15$ for pulses 3–10 in the entire cohort and $P = 0.003$ and 0.008 , respectively, in nonobese donors). Moreover, β -cells from risk allele carriers exhibited a reduced number (25% per risk allele) of docked granules ($P = 0.037$; $n = 97$ cells from 18 donors) (Table 1 and Supplementary Fig. 2). Another SNP in *KCNQ1* (rs231362), which is independently associated with T2D risk (4), did not affect exocytosis and increased (30%) rather than reduced the number of docked granules (Table 1).

KCNQ1 encodes a voltage-gated K^+ channel, mutations of which cause both long- and short-QT syndromes (28). Previous investigations into its role in insulin secretion are contradictory (29,30). Preincubation with the *KCNQ1* antagonist linopirdine stimulated exocytosis in human β -cells ($P = 0.03$ for rapid exocytosis; $n = 21$). siRNA silencing of *KCNQ1* ($64 \pm 3\%$ [$n = 3$] reduction of mRNA, 74% protein reduction as measured by immunocytochemistry, and 54% reduction as measured by Western blot) (Supplementary Fig. 3) likewise enhanced exocytosis from 59 ± 9 fF/pC to 99 ± 15 fF/pC ($P = 0.011$) (Fig. 2G and H). Collectively, these data suggest that *KCNQ1* is expressed in human β -cells and that activity of the channel regulates exocytosis by an effect unrelated to any impact this channel may have on action potential firing.

In addition, the risk allele for rs1111875 (*HHEX/IDE*) significantly decreased the number of docked granules ($P = 0.016$), an effect that correlated with a tendency for reduced exocytosis ($P = 0.086$). Moreover, risk genotype for rs11920090 (*SLC2A2*) tended to increase the number of docked granules ($P = 0.037$) (Table 1). Rs13266634 (*SLC30A8*) is the second most strongly associated T2D variant identified to date (1), but previous population studies have shown inconsistent effects of the variant on insulin secretion (31,32). Studies reporting global knockout of this gene have shown abnormalities in zinc accumulation, insulin granule morphology, and insulin secretion from isolated β -cells, but effects on in vivo glucose homeostasis have been less consistent (33,34). In contrast, β -cell-specific *Slc30a8* knockout mice have shown defects in insulin secretion both in vivo and in vitro (35). However, we found no association between the risk variant and either insulin secretion or exocytosis in human islets (Table 1).

mRNA levels of the four genes associated with defects in β -cell function were measured in human islets (Supplementary Table 6). Interestingly, *KCNJ11* expression was reduced by 30% in islets from donors with T2D ($n = 15$) compared with those from normoglycemic donors ($n = 56$), whereas *ADRA2A* and *KCNQ1* transcripts did not differ between islets from ND and T2D donors. For *TCF7L2*, previous reports on its expression in islets from donors with T2D are inconsistent (21,36,37); both silencing and overexpression of this gene have been associated with impaired glucose tolerance (38,39). Of note, we did not detect any change in the levels of *TCF7L2* expression, including a newly discovered isoform (40), in T2D donors. **Genetic risk score for β -cell dysfunction.** Finally, we investigated whether we could combine our findings to construct a genetic risk score for defective β -cell function. For each individual, we added the number of risk alleles for the four SNPs associated with secretion/exocytosis phenotypes (highlighted in bold in Table 1). Thus, each donor was assigned a score from 0 to 8 (an individual being homozygous for all risk alleles would have a score of 8). There was a significant association between the risk score and reduced glucose-stimulated insulin secretion in islets ($P = 0.008$; $n = 47$ donors) (Table 2) whereas there was no effect on insulin content ($P = 0.8$). The risk score was linearly associated with reduced β -cell exocytosis ($P = 5 \times 10^{-6}$) (Fig. 3A), which was more pronounced in nonobese individuals ($P = 4 \times 10^{-8}$). Both rapid and slow exocytosis were decreased (Table 2), an effect not associated with reduced Ca^{2+} influx. Donors with the highest risk score showed reduced Ca^{2+} sensitivity of exocytosis ($P = 0.003$), which was paralleled by a strong correlation between the risk score and a reduced number of docked insulin granules ($P = 0.0004$) (Fig. 3B; for quantification see Table 2). For reasons of statistical power, both T2D and ND donors were included in the genotype analyses, and diabetes status was used as a covariate. When only ND donors were analyzed, all associations exhibited similar trends with respect to the β values, but the P values in most cases did not reach statistical significance (Supplementary Table 7). However, the risk score was associated with reduced glucose-stimulated insulin release ($P = 0.013$; $\beta = -0.068$; $n = 33$ donors), decreased β -cell exocytosis ($P = 0.001$; $\beta = -0.060$; $n = 199$ cells from 29 donors), and reduced number of docked granules ($P = 2 \times 10^{-5}$; $\beta = -0.182$; $n = 60$ cells from 12 donors) also when considering ND donors alone.

To investigate if this risk score could predict insulin secretion capacity in an independent patient cohort, we

TABLE 2
Effect of genetic score on cellular phenotypes

Parameter	P value	Effect*
Glucose-induced insulin secretion (ng/islet/h)	0.008	-0.235
Total exocytosis (fF/pF)	5×10^{-6}	-6.09
Rapid exocytosis (fF/pF)	2×10^{-4}	-6.82
Slow exocytosis (fF/pF)	4×10^{-4}	-4.70
Integrated Ca^{2+} current (pC)	0.70	-0.008
Exocytosis/charge [†] (fF/pC)	0.003	-1.34
Ns [‡] (granules/ μm^2)	4×10^{-4}	-0.080

*A linear model was used to analyze the association between the genetic risk score (0–8) and physiological measures in islets from the donors. The effect of each additional risk allele on the phenotypes is given in units as indicated. [†]The capacitance increase in response to the first depolarization normalized to the charge, indicating the Ca^{2+} sensitivity of exocytosis. [‡]Number of docked granules.

studied the four T2D variants in 604 ND individuals for whom IVGTT data were available. Interestingly, the genetic risk score was associated with reduced fasting insulin ($P = 0.034$) and impaired insulin secretion in response to an IVGTT (Fig. 3C and Supplementary Table 8). Moreover, in 2,770 subjects from the Botnia Prospective cohort, individuals with a high risk score exhibited reduced CIR over time, with suppressed CIR at the follow-up visit ($P = 0.035$; $\beta = -0.030$).

DISCUSSION

This study is the first to investigate the association between T2D genetic risk variants and detailed β -cell phenotypes. The data demonstrate the potential of investigating genetic subgroups of patients in combination

with in vitro cellular phenotypes to better characterize the mechanisms underlying reduced systemic insulin levels. It is of interest that the negative impact of T2D loci on β -cell function was particularly evident in islets from nonobese individuals. This suggests that the functional effects of the T2D-associated SNPs identified to date may be more pronounced in lean than in obese individuals. It may seem counterintuitive that obese individuals with T2D exhibit greater insulin secretion than their lean counterparts. It is likely that this reflects compensation, albeit insufficient to prevent diabetes, in the obese donors. In addition, lean donors that have developed T2D are likely to be those with the lowest insulin secretory capacity or else they would not have become diabetic.

A number of mechanisms could contribute to the reduced insulin secretion in vivo that has been associated with several T2D susceptibility variants. Dissection of the underlying cellular pathology requires 1) access to relevant human tissues from nonrisk and risk genotype carriers, which facilitates the correct translation of association signals compared with studying genetically modified animals, and 2) characterization of the effect of genotype on detailed cellular phenotypes. There are fundamental electrophysiological and secretory differences between human and rodent β -cells, making the study of human islets essential to investigate the influence of T2D susceptibility variants on β -cell function. The biophysical and ultrastructural examination of human β -cells in the current study identified four T2D variants that were associated with reduced exocytosis and enabled characterization of the mechanisms for the exocytotic impairment. The results shed new light on the pathophysiology linked with these risk variants, near *TCF7L2*, *ADRA2A*, *KCNJ11*, and *KCNQ1*, and demonstrate that defective β -cell exocytosis can be an important pathogenic mechanism in genetic subgroups of T2D. The data

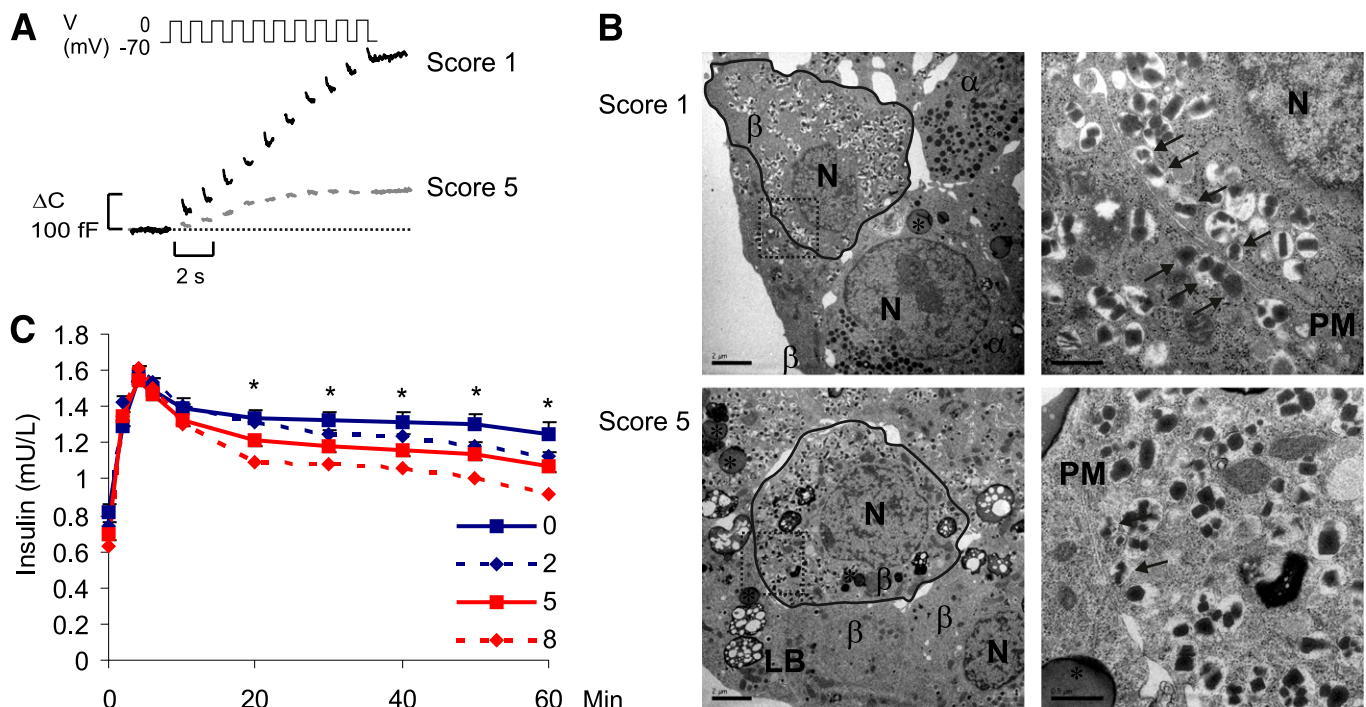


FIG. 3. Effects of the genetic risk score. A: Depolarization-evoked exocytosis in β -cells from donors with different scores. B: Electron micrographs of human islet sections from an individual with low (top) and high (bottom) score. Docked insulin granules (arrows), lipid droplets (*), nuclei (N), lipofuscin bodies (LB), and plasma membrane (PM) are indicated. Left scale bars, 2 μm ; right, 0.5 μm . C: Effects of the risk score on insulin secretion during IVGTT in 604 individuals. Data are means $\pm$ SEM. * $P < 0.05$.

suggest that there may be considerable heterogeneity in the cellular pathways that lead to reduced insulin secretion, which may explain why the reduction of exocytosis is evident only in genetic subgroups and not in the entire T2D cohort. Stratification based on genetic variants may therefore be useful to better resolve the disease mechanisms. Similar approaches may therefore be valuable to study the T2D susceptibility variants that were not associated with defective β -cell exocytosis in the current study (Table 1) and may instead impair systemic insulin release through effects on β -cell mass and/or glucose sensing or indirectly via incretins and innervation.

We acknowledge that although this study is the largest to date (totally including islets from 100 donors), the number of donors and consequently the statistical power is limited compared with that of genetic population studies. However, the combination of genetics and cellular physiology identified four risk alleles that associate with impaired β -cell exocytosis and enabled us to form a novel genetic risk score for single β -cell dysfunction that involves impaired granule docking and defective Ca^{2+} sensitivity of exocytosis. These findings give new insights on the pathophysiology of T2D that may open up possibilities for identifying subgroups of patients who would benefit from treatment specifically aimed at improving β -cell exocytosis. In this context, it may be relevant that glucagon-like peptide 1, via elevation of intracellular cAMP (41), amplifies Ca^{2+} -induced exocytosis in human β -cells.

ACKNOWLEDGMENTS

This study was supported by the Novo Nordisk Foundation, the Wallenberg Foundation, and the Sigrid Juselius Foundation as well as by the Swedish Research Council through research positions (A.H.R., E.R., and L.E.), research grants, a Linnaeus grant, and a Strategic Research Grant (Exodiab). Human islets were provided in Sweden by Prof. Olle Korsgren, the Nordic Network for Clinical Islet Transplantation. This work was funded in Oxford by the Medical Research Council (MRC81696), the National Institute of Health Research, and the Wellcome Trust (095101/Z/10/Z and 095531/Z/11/Z). A.L.G. is a Wellcome Trust Senior Fellow in Basic Biomedical Science (095101/Z/10/Z) and P.R. is a Wellcome Trust Senior Investigator. During the preparation of this manuscript P.R. held a Canada Excellence Research Chair.

No potential conflicts of interest relevant to this article were reported.

A.H.R., M.B., and L.E. designed the project, performed the research, analyzed the data, and wrote the paper. T.M., M.S., E.Z., A.S.A., A.E., A.J., R.R., Y.T., J.N.W., and A.B. performed the research. S.A.A. and M.G.P. performed the research, analyzed the data, and contributed to the manuscript. M.E.T. performed the research and wrote the paper. P.A. and C.L. designed the project and analyzed the data. P.R.V.J. provided human islets. V.L. analyzed the data. M.I.M., L.G., A.L.G., E.R., and P.R. designed the project and wrote the paper. A.S. designed the project, performed the research, and analyzed the data. A.H.R. is the guarantor of this work and, as such, had full access to all the data in the study and takes responsibility for the integrity of the data and the accuracy of the data analysis.

REFERENCES

- Sladek R, Rocheleau G, Rung J, et al. A genome-wide association study identifies novel risk loci for type 2 diabetes. *Nature* 2007;445:881–885

- Saxena R, Voight BF, Lyssenko V, et al.; Diabetes Genetics Initiative of Broad Institute of Harvard and MIT, Lund University, and Novartis Institutes of BioMedical Research. Genome-wide association analysis identifies loci for type 2 diabetes and triglyceride levels. *Science* 2007;316:1331–1336
- Zeggini E, Scott LJ, Saxena R, et al.; Wellcome Trust Case Control Consortium. Meta-analysis of genome-wide association data and large-scale replication identifies additional susceptibility loci for type 2 diabetes. *Nat Genet* 2008;40:638–645
- Voight BF, Scott LJ, Steinthorsdottir V, et al., MAGIC investigators; GIANT Consortium. Twelve type 2 diabetes susceptibility loci identified through large-scale association analysis. *Nat Genet* 2010;42:579–589
- Del Guerra S, Lupi R, Marselli L, et al. Functional and molecular defects of pancreatic islets in human type 2 diabetes. *Diabetes* 2005;54:727–735
- Deng S, Vatamaniuk M, Huang X, et al. Structural and functional abnormalities in the islets isolated from type 2 diabetic subjects. *Diabetes* 2004;53:624–632
- Butler AE, Janson J, Bonner-Weir S, Ritzel R, Rizza RA, Butler PC. Beta-cell deficit and increased beta-cell apoptosis in humans with type 2 diabetes. *Diabetes* 2003;52:102–110
- Braun M, Ramracheya R, Bengtsson M, et al. Voltage-gated ion channels in human pancreatic beta-cells: electrophysiological characterization and role in insulin secretion. *Diabetes* 2008;57:1618–1628
- Groop L, Forsblom C, Lehtovirta M, et al. Metabolic consequences of a family history of NIDDM (the Botnia study): evidence for sex-specific parental effects. *Diabetes* 1996;45:1585–1593
- Hanson RL, Pratley RE, Bogardus C, et al. Evaluation of simple indices of insulin sensitivity and insulin secretion for use in epidemiologic studies. *Am J Epidemiol* 2000;151:190–198
- Rosengren AH, Jokubka R, Tojjar D, et al. Overexpression of $\alpha 2A$ -adrenergic receptors contributes to type 2 diabetes. *Science* 2010;327:217–220
- Vikman J, Jimenez-Feltr  m J, Nyman P, Thelin J, Eliasson L. Insulin secretion is highly sensitive to desorption of plasma membrane cholesterol. *FASEB J* 2009;23:58–67
- Goto M, Holgersson J, Kumagai-Braesch M, Korsgren O. The ADP/ATP ratio: a novel predictive assay for quality assessment of isolated pancreatic islets. *Am J Transplant* 2006;6:2483–2487
- Wennberg L, Song Z, Bennet W, et al. Diabetic rats transplanted with adult porcine islets and immunosuppressed with cyclosporine A, mycophenolate mofetil, and leflunomide remain normoglycemic for up to 100 days. *Transplantation* 2001;71:1024–1033
- Rahier J, Guiot Y, Goebbels RM, Sempoux C, Henquin JC. Pancreatic beta-cell mass in European subjects with type 2 diabetes. *Diabetes Obes Metab* 2008;10(Suppl. 4):32–42
- Rorsman P, Renstr  m E. Insulin granule dynamics in pancreatic beta cells. *Diabetologia* 2003;46:1029–1045
- Shibasaki T, Takahashi H, Miki T, et al. Essential role of Epac2/Rap1 signaling in regulation of insulin granule dynamics by cAMP. *Proc Natl Acad Sci USA* 2007;104:19333–19338
- Ashcroft F, Rorsman P. Type 2 diabetes mellitus: not quite exciting enough? *Hum Mol Genet* 2004;13(Spec No 1):R21–R31
- Cnop M, Hughes SJ, Igoillo-Esteve M, et al. The long lifespan and low turnover of human islet beta cells estimated by mathematical modelling of lipofuscin accumulation. *Diabetologia* 2010;53:321–330
- Grant SF, Thorleifsson G, Reynisdottir I, et al. Variant of transcription factor 7-like 2 (TCF7L2) gene confers risk of type 2 diabetes. *Nat Genet* 2006;38:320–323
- Lyssenko V, Lupi R, Marchetti P, et al. Mechanisms by which common variants in the TCF7L2 gene increase risk of type 2 diabetes. *J Clin Invest* 2007;117:2155–2163
- da Silva Xavier G, Loder MK, McDonald A, et al. TCF7L2 regulates late events in insulin secretion from pancreatic islet beta-cells. *Diabetes* 2009;58:894–905
- Gloyn AL, Pearson ER, Antcliff JF, et al. Activating mutations in the gene encoding the ATP-sensitive potassium-channel subunit Kir6.2 and permanent neonatal diabetes. *N Engl J Med* 2004;350:1838–1849
- Hamming KS, Soliman D, Maternisz LC, 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 $\text{K}(+)$ channel. *Diabetes* 2009;58:2419–2424
- Eliasson L, Renstr  m E, Amm  la C, et al. PKC-dependent stimulation of exocytosis by sulfonylureas in pancreatic beta cells. *Science* 1996;271:813–815
- Unoki H, Takahashi A, Kawaguchi T, et al. SNPs in KCNQ1 are associated with susceptibility to type 2 diabetes in East Asian and European populations. *Nat Genet* 2008;40:1098–1102

27. Yasuda K, Miyake K, Horikawa Y, et al. Variants in KCNQ1 are associated with susceptibility to type 2 diabetes mellitus. *Nat Genet* 2008;40:1092–1097
28. Bellocq C, van Ginneken AC, Bezzina CR, et al. Mutation in the KCNQ1 gene leading to the short QT-interval syndrome. *Circulation* 2004;109:2394–2397
29. Ullrich S, Su J, Ranta F, et al. Effects of I(Ks) channel inhibitors in insulin-secreting INS-1 cells. *Pflugers Arch* 2005;451:428–436
30. Zhang M, Houamed K, Kupersmidt S, Roden D, Satin LS. Pharmacological properties and functional role of Kslow current in mouse pancreatic beta-cells: SK channels contribute to Kslow tail current and modulate insulin secretion. *J Gen Physiol* 2005;126:353–363
31. Boesgaard TW, Zilinskaite J, Vanttinen M, et al.; EUGENE 2 Consortium. The common SLC30A8 Arg325Trp variant is associated with reduced first-phase insulin release in 846 non-diabetic offspring of type 2 diabetes patients—the EUGENE2 study. *Diabetologia* 2008;51:816–820
32. Kirchhoff K, Machicao F, Haupt A, et al. Polymorphisms in the TCF7L2, CDKAL1 and SLC30A8 genes are associated with impaired proinsulin conversion. *Diabetologia* 2008;51:597–601
33. Nicolson TJ, Bellomo EA, Wijesekara N, et al. Insulin storage and glucose homeostasis in mice null for the granule zinc transporter ZnT8 and studies of the type 2 diabetes-associated variants. *Diabetes* 2009;58:2070–2083
34. Lemaire K, Ravier MA, Schraenen A, et al. Insulin crystallization depends on zinc transporter ZnT8 expression, but is not required for normal glucose homeostasis in mice. *Proc Natl Acad Sci USA* 2009;106:14872–14877
35. Wijesekara N, Dai FF, Hardy AB, et al. Beta cell-specific Znt8 deletion in mice causes marked defects in insulin processing, crystallisation and secretion. *Diabetologia* 2010;53:1656–1668
36. Osmark P, Hansson O, Jonsson A, Rönn T, Groop L, Renström E. Unique splicing pattern of the TCF7L2 gene in human pancreatic islets. *Diabetologia* 2009;52:850–854
37. Shu L, Matveyenko AV, Kerr-Conte J, Cho JH, McIntosh CH, Maedler K. Decreased TCF7L2 protein levels in type 2 diabetes mellitus correlate with downregulation of GIP- and GLP-1 receptors and impaired beta-cell function. *Hum Mol Genet* 2009;18:2388–2399
38. Shu L, Sauter NS, Schulthess FT, Matveyenko AV, Oberholzer J, Maedler K. Transcription factor 7-like 2 regulates beta-cell survival and function in human pancreatic islets. *Diabetes* 2008;57:645–653
39. Savic D, Ye H, Aneas I, Park SY, Bell GI, Nobrega MA. Alterations in TCF7L2 expression define its role as a key regulator of glucose metabolism. *Genome Res* 2011;21:1417–1425
40. Locke JM, Da Silva Xavier G, Rutter GA, Harries LW. An alternative polyadenylation signal in TCF7L2 generates isoforms that inhibit T cell factor/lymphoid-enhancer factor (TCF/LEF)-dependent target genes. *Diabetologia* 2011;54:3078–3082
41. Gromada J, Bokvist K, Ding WG, Holst JJ, Nielsen JH, Rorsman P. 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* 1998;47:57–65

Genome-Wide Association Identifies Nine Common Variants Associated With Fasting Proinsulin Levels and Provides New Insights Into the Pathophysiology of Type 2 Diabetes

Rona J. Strawbridge,¹ Josée Dupuis,^{2,3} Inga Prokopenko,^{4,5} Adam Barker,⁶ Emma Ahlqvist,⁷ Denis Rybin,⁸ John R. Petrie,⁹ Mary E. Travers,⁴ Nabila Bouatia-Naji,^{10,11} Antigone S. Dimas,^{5,12} Alexandra Nica,^{12,13} Eleanor Wheeler,¹⁴ Han Chen,² Benjamin F. Voight,^{15,16} Jalal Taneera,⁷ Stavroula Kanoni,^{13,17} John F. Peden,^{5,18} Fabiola Turrini,^{7,19} Stefan Gustafsson,²⁰ Carina Zabena,^{21,22} Peter Almgren,⁷ David J.P. Barker,²³ Daniel Barnes,⁶ Elaine M. Dennison,²⁴ Johan G. Eriksson,^{25,26,27,28} Per Eriksson,¹ Elodie Eury,^{10,11} Lasse Folkersen,²⁹ Caroline S. Fox,^{3,30} Timothy M. Frayling,³¹ Anuj Goel,^{5,18} Harvest F. Gu,³² Momoko Horikoshi,^{4,5} Bo Isomaa,^{27,33} Anne U. Jackson,³⁴ Karen A. Jameson,²⁴ Eero Kajantie,^{25,35} Julie Kerr-Conte,^{10,36} Teemu Kuulasmaa,³⁷ Johanna Kuusisto,³⁷ Ruth J.F. Loos,⁶ Jian'an Luan,⁶ Konstantinos Makrakis,³⁸ Alisa K. Manning,² María Teresa Martínez-Larrad,^{21,22} Narisu Narisu,³⁹ Maria Nastase Mannila,¹ John Öhrvik,¹ Clive Osmond,²⁴ Laura Pascoe,⁴⁰ Felicity Payne,¹⁴ Avan A. Sayer,²⁴ Bengt Sennblad,¹ Angela Silveira,¹ Alena Stančáková,³⁷ Kathy Stirrups,¹³ Amy J. Swift,³⁹ Ann-Christine Syvänen,⁴¹ Tiinamaija Tuomi,^{27,42} Ferdinand M. van 't Hooft,¹ Mark Walker,⁴³ Michael N. Weedon,³¹ Weijia Xie,³¹ Björn Zethelius,⁴⁴ the DIAGRAM Consortium,* the GIANT Consortium,* the MuTHER Consortium,* the CARDIoGRAM Consortium,* the C4D Consortium,* Halit Ongén,^{5,18,45} Anders Mälarstig,¹ Jemma C. Hopewell,⁴⁶ Danish Saleheen,^{47,48} John Chambers,^{49,50} Sarah Parish,⁴⁶ John Danesh,⁴⁷ Jaspal Kooner,^{50,51} Claes-Göran Östenson,³² Lars Lind,⁴¹ Cyrus C. Cooper,²⁴ Manuel Serrano-Ríos,^{21,22} Ele Ferrannini,⁵² Tom J. Forsen,^{28,53} Robert Clarke,⁴⁶ Maria Grazia Franzosi,⁵⁴ Udo Seedorf,⁵⁵ Hugh Watkins,^{5,18} Philippe Froguel,^{10,11,56} Paul Johnson,^{4,57} Panos Deloukas,¹³ Francis S. Collins,⁵⁸ Markku Laakso,³⁷ Emmanouil T. Dermizakis,¹² Michael Boehnke,³⁴ Mark I. McCarthy,^{4,5,59} Nicholas J. Wareham,⁶ Leif Groop,⁷ François Pattou,^{10,36} Anna L. Gloyn,⁴ George V. Dedoussis,¹⁷ Valeriya Lyssenko,⁷ James B. Meigs,^{60,61} Inês Barroso,^{14,62} Richard M. Watanabe,^{63,64} Erik Ingelsson,²⁰ Claudia Langenberg,⁶ Anders Hamsten,¹ and Jose C. Florez^{15,16,61}

From the ¹Atherosclerosis Research Unit, Department of Medicine Solna, Karolinska Institutet, Karolinska University Hospital Solna, Stockholm, Sweden; the ²Department of Biostatistics, Boston University School of Public Health, Boston, Massachusetts; the ³National Heart, Lung, and Blood Institute's Framingham Heart Study, Framingham, Massachusetts; the ⁴Oxford Centre for Diabetes Endocrinology and Metabolism, University of Oxford, Oxford, U.K.; the ⁵Wellcome Trust Centre for Human Genetics, University of Oxford, Oxford, U.K.; the ⁶MRC Epidemiology Unit, Institute of Metabolic Science, Addenbrooke's Hospital, Cambridge, U.K.; the ⁷Department of Clinical Sciences, Diabetes and Endocrinology, University Hospital and Malmö, Lund University, Malmö, Sweden; the ⁸Boston University Data Coordinating Center, Boston, Massachusetts; the ⁹BHF Cardiovascular Research Centre, University of Glasgow, Glasgow, U.K.; the ¹⁰Université Lille-Nord de France, Lille, France; the ¹¹CNRS UMR 8199, Institut Pasteur de Lille, Lille, France; the ¹²Department of Genetic Medicine and Development, University of Geneva Medical School, Geneva, Switzerland; the ¹³Wellcome Trust Sanger Institute, Wellcome Trust Genome Campus, Hinxton, U.K.; the ¹⁴Metabolic Disease Group, Wellcome Trust Sanger Institute, Hinxton, Cambridge, U.K.; the ¹⁵Program in Medical and Population Genetics, Broad Institute, Cambridge, Massachusetts; the ¹⁶Center for Human Genetic Research and Diabetes Research Center (Diabetes Unit), Massachusetts General Hospital, Boston, Massachusetts; the ¹⁷Department of Dietetics-Nutrition, Harokopio University, Athens, Greece; the ¹⁸Department of Cardiovascular Medicine, University of Oxford, Oxford, U.K.; the ¹⁹Department of Medicine, University of Verona, Verona, Italy; the ²⁰Department of Medical Epidemiology and Biostatistics, Karolinska Institutet, Stockholm, Sweden; the ²¹CIBER de Diabetes y Enfermedades Metabólicas Asociadas (CIBERDEM), Madrid, Spain; the ²²Fundación Investigación Biomédica del Hospital Clínico San Carlos, Madrid, Spain; the ²³Heart Research Center, Oregon Health and Science University, Portland, Oregon; the ²⁴MRC Lifecourse Epidemiology Unit, University of Southampton, Southampton General Hospital, Southampton, U.K.; the ²⁵National Institute for Health

and Welfare, Helsinki, Finland; the ²⁶Helsinki University Central Hospital, Unit of General Practice, Helsinki, Finland; the ²⁷Folkhälsan Research Centre, Helsinki, Finland; the ²⁸Department of General Practice and Primary Health Care, University of Helsinki, Helsinki, Finland; the ²⁹Experimental Cardiovascular Research Unit, Department of Medicine Solna, Karolinska Institutet, Stockholm, Sweden; the ³⁰Division of Endocrinology, Diabetes, and Hypertension, Brigham and Women's Hospital, Harvard Medical School, Boston, Massachusetts; the ³¹Institute of Biomedical and Clinical Sciences, Peninsula Medical School, University of Exeter, Exeter, U.K.; the ³²Endocrinology and Diabetes Unit, Department of Molecular Medicine and Surgery, Karolinska Institutet, Stockholm, Sweden; the ³³Malmö Municipal Health Care Center and Hospital, Jakobstad, Finland; the ³⁴Center for Statistical Genetics, Department of Biostatistics, School of Public Health, University of Michigan, Ann Arbor, Michigan; the ³⁵Hospital for Children and Adolescents, Helsinki University Central Hospital and University of Helsinki, Helsinki, Finland; the ³⁶INSERM UMR 859, Lille, France; the ³⁷Department of Medicine, University of Kuopio and Kuopio University Hospital, Kuopio, Finland; the ³⁸First Department of Propædæutic Medicine, Laiko General Hospital, Athens University Medical School, Athens, Greece; the ³⁹National Human Genome Research Institute, National Institutes of Health, Bethesda, Maryland; the ⁴⁰Institute of Cell and Molecular Biosciences, Newcastle University, Newcastle, U.K.; the ⁴¹Department of Medical Sciences, Molecular Medicine, Science for Life Laboratory, Uppsala University, Uppsala, Sweden; the ⁴²Department of Medicine, Helsinki University Central Hospital, and Research Program of Molecular Medicine, University of Helsinki, Helsinki, Finland; the ⁴³Institute of Cellular Medicine, Newcastle University, Newcastle, U.K.; the ⁴⁴Department of Public Health and Caring Sciences, Uppsala University, Uppsala, Sweden; the ⁴⁵Department of Cardiovascular Medicine, University of Oxford, John Radcliffe Hospital, Headington, Oxford, U.K.; the ⁴⁶Clinical Trial Service Unit, University of Oxford, Oxford, U.K.; the ⁴⁷Department of Public Health and Primary Care, University of Cambridge, Cambridge, U.K.; the ⁴⁸Center for Non-Communicable Diseases Pakistan, Karachi,

OBJECTIVE—Proinsulin is a precursor of mature insulin and C-peptide. Higher circulating proinsulin levels are associated with impaired β -cell function, raised glucose levels, insulin resistance, and type 2 diabetes (T2D). Studies of the insulin processing pathway could provide new insights about T2D pathophysiology.

RESEARCH DESIGN AND METHODS—We have conducted a meta-analysis of genome-wide association tests of ~ 2.5 million genotyped or imputed single nucleotide polymorphisms (SNPs) and fasting proinsulin levels in 10,701 nondiabetic adults of European ancestry, with follow-up of 23 loci in up to 16,378 individuals, using additive genetic models adjusted for age, sex, fasting insulin, and study-specific covariates.

RESULTS—Nine SNPs at eight loci were associated with proinsulin levels ($P < 5 \times 10^{-8}$). Two loci (*LARP6* and *SGSM2*) have not been previously related to metabolic traits, one (*MADD*) has been associated with fasting glucose, one (*PCSK1*) has been implicated in obesity, and four (*TCF7L2*, *SLC30A8*, *VPS13C/C2CD4A/B*, and *ARAP1*, formerly *CENTD2*) increase T2D risk. The proinsulin-raising allele of *ARAP1* was associated with a lower fasting glucose ($P = 1.7 \times 10^{-4}$), improved β -cell function ($P = 1.1 \times 10^{-5}$), and lower risk of T2D (odds ratio 0.88; $P = 7.8 \times 10^{-6}$). Notably, *PCSK1* encodes the protein prohormone convertase 1/3, the first enzyme in the insulin processing pathway. A genotype score composed of the nine proinsulin-raising alleles was not associated with coronary disease in two large case-control datasets.

CONCLUSIONS—We have identified nine genetic variants associated with fasting proinsulin. Our findings illuminate the biology underlying glucose homeostasis and T2D development in humans and argue against a direct role of proinsulin in coronary artery disease pathogenesis. *Diabetes* 60:2624–2634, 2011

Genome-wide association studies (GWAS) have uncovered dozens of common genetic variants associated with risk for type 2 diabetes (T2D; reviewed in [1]). Known associated variants in these loci account for only a small proportion of the heritable component of T2D (1), suggesting that additional loci await discovery. The Meta-Analyses of Glucose and Insulin-related traits Consortium (MAGIC) was created under the premise that genome-wide analysis of continuous diabetes-related traits could not only identify loci regulating variation in these glycemic traits, but also yield additional T2D susceptibility loci and insights into the underlying physiology of these loci (2–5). In addition, the genetic study of T2D endophenotypes may help clarify the pathophysiologic heterogeneity of this disease by elucidating the respective roles of β -cell function, insulin

secretion, processing and sensitivity, and glucose metabolism (6).

Discovery of novel genetic determinants of insulin secretion and action has primarily focused on insulin levels (3,4,7,8). Proinsulin is the molecular precursor for insulin and has relatively low insulin-like activity, and its enzymatic conversion into mature insulin and C-peptide is a critical step in insulin production and secretion (Supplementary Fig. 1). Although hyperinsulinemia typically denotes insulin resistance, high proinsulin in relation to circulating levels of mature insulin can indicate β -cell stress as a result of insulin resistance, impaired β -cell function, and/or insulin processing and secretion abnormalities (9) (Supplementary Fig. 2). There is good evidence that higher proinsulin predicts future T2D (10) and coronary artery disease (CAD) (11–13), even after taking fasting glucose levels into account. Interestingly, some loci previously associated with fasting glucose levels (*MADD*) or risk of T2D (*TCF7L2*, *SLC30A8*, *CDKAL1*) are also associated with higher circulating proinsulin (6,14–17). Therefore, genome-wide analysis of proinsulin levels could reveal additional novel loci increasing susceptibility for T2D and perhaps CAD.

Thus, to identify novel loci influencing proinsulin processing and secretion and potentially increasing susceptibility for T2D, we performed a meta-analysis of ~ 2.5 million directly genotyped or imputed autosomal single nucleotide polymorphisms (SNPs) from four GWAS of fasting proinsulin levels (adjusted for concomitant fasting insulin) including 10,701 nondiabetic adult men and women of European descent. Follow-up of 23 lead SNPs from the most significant association signals in up to 16,378 additional individuals of European ancestry detected nine genome-wide significant associations with proinsulin levels, including two novel signals in or near *LARP6* and *SGSM2*, and the known glycemic loci *ARAP1*, *MADD* (two independent signals), *TCF7L2*, *VPS13C/C2CD4A/B*, *SLC30A8*, and *PCSK1*. Here we describe these genetic associations, perform fine-mapping to identify potential causal variants, assess gene expression in human tissues, and define their impact on other glycemic quantitative traits and risk of both T2D and CAD.

RESEARCH DESIGN AND METHODS

Cohort/study description. Four cohorts contributed to the discovery meta-analysis through the contribution of phenotypic and GWAS data. These included the Framingham Heart Study ($n = 5,759$), Precocious Coronary Artery Disease (PROCARDIS) ($n = 3,259$), the Fenland study ($n = 1,372$), and the Diabetes Genetics Initiative (DGI) ($n = 311$), for a total of 10,701 participants.

Pakistan; ⁴⁹Epidemiology and Biostatistics, Imperial College London, Norfolk Place, London, U.K.; ⁵⁰Cardiology, Ealing Hospital NHS Trust, Middlesex, U.K.; the ⁵¹National Heart and Lung Institute, Imperial College London, London, U.K.; the ⁵²Department of Internal Medicine and CNR Institute of Clinical Physiology, University of Pisa School of Medicine, Pisa, Italy; the ⁵³Vaasa Health Care Center, Vaasa, Finland; the ⁵⁴Department of Cardiovascular Research, Mario Negri Institute for Pharmacological Research, Milan, Italy; the ⁵⁵Leibniz Institute for Arteriosclerosis Research, University of Münster, Münster, Germany; the ⁵⁶Department of Genomics of Common Disease, School of Public Health, Imperial College London, Hammersmith Hospital, London, U.K.; the ⁵⁷DRWF Human Islet Isolation Facility and Oxford Islet Transplant Programme, University of Oxford, Oxford, U.K.; the ⁵⁸National Institutes of Health, Bethesda, Maryland; the ⁵⁹Oxford NIHR Biomedical Research Centre, Churchill Hospital, Oxford, U.K.; the ⁶⁰General Medicine Division, Massachusetts General Hospital, Boston, Massachusetts; the ⁶¹Department of Medicine, Harvard Medical School, Boston, Massachusetts; the ⁶²University of Cambridge Metabolic Research Laboratories, Institute of Metabolic Science, Addenbrooke's Hospital, Cambridge, U.K.; the ⁶³Department of Preventive Medicine, Keck

School of Medicine, University of Southern California, Los Angeles, California; and the ⁶⁴Department of Physiology and Biophysics, Keck School of Medicine, University of Southern California, Los Angeles, California.

Corresponding authors: Jose C. Florez, jcflorez@partners.org; Claudia Langenberg, claudia.langenberg@mrc-epid.cam.ac.uk; and Anders Hamsten, anders.hamsten@ki.se.

Received 23 March 2011 and accepted 29 June 2011.

DOI: 10.2337/db11-0415

This article contains Supplementary Data online at <http://diabetes.diabetesjournals.org/lookup/suppl/doi:10.2337/db11-0415/-/DC1>.

R.J.S., J.Du., I.P., A.B., and E.A. contributed equally to this work.

*A full list of the DIAGRAM Consortium, the GIANT Consortium, the MuTHER Consortium, the CARDIoGRAM Consortium, and the C4D Consortium investigators is provided in the Supplementary Data.

© 2011 by the American Diabetes Association. Readers may use this article as long as the work is properly cited, the use is educational and not for profit, and the work is not altered. See <http://creativecommons.org/licenses/by-nc-nd/3.0/> for details.

Eleven cohorts contributed to the follow-up efforts; these included Metabolic Syndrome in Men (METSIM) ($n = 5,122$), Botnia Prevalence, Prediction and Prevention of diabetes (Botnia-PPP) ($n = 2,280$), Helsinki Birth Cohort Study (HBCS) ($n = 1,649$), the Ely study ($n = 1,568$), the Hertfordshire study ($n = 1,016$), Uppsala Longitudinal Study of Adult Men (ULSAM) ($n = 939$), Relationship between Insulin Sensitivity and Cardiovascular disease (RISC) ($n = 914$), Prospective Investigation of the Vasculature in Uppsala Seniors (PIVUS) ($n = 912$), Segovia ($n = 911$), the Greek Health Randomized Aging Study (GHRAS) ($n = 668$), and Stockholm Diabetes Prevention Program (SDPP) ($n = 399$), for a total of 16,378 participants (with maximal sample for any one SNP of 15,898). We excluded individuals with known diabetes, on antidiabetic treatment, or with fasting glucose ≥ 7 mmol/L (3); all participants were of European descent.

Proinsulin and insulin measurements. Proinsulin (pmol/L) was measured from fasting whole blood, plasma, or serum or a combination of these using enzyme-linked immunosorbent or immunometric assays. Fasting insulin (pmol/L) was measured using either enzyme-linked immunosorbent, immuno-fluorescent, or radioimmunoassays (Supplementary Table 1).

Genotyping. Genome-wide commercial arrays (Affymetrix 500K, MIPS 50K, and Illumina Human1M/610K) were used by the four discovery cohorts as described in Supplementary Table 1. Imputation and quality control methods are described in the Supplementary Data.

Statistical analyses. We aimed to identify genetic variants associated with high proinsulin levels relative to an individual's fasting insulin levels. This can be done by examining proinsulin-to-insulin ratios or by statistically adjusting proinsulin for fasting insulin. We chose the latter because the adjusted trait has comparable predictive value (18) and displayed better statistical performance in pilot studies and adequate heritability in the Framingham Heart Study, one of the larger cohorts examined here ($h^2 = 0.36$ vs. 0.34 for the proinsulin-to-insulin ratio). In Framingham, correlation between the adjusted

trait and the ratio was 0.95, and the quantile-quantile GWAS plots were comparable.

We used a linear regression model with natural log transformed fasting proinsulin as the dependent variable and genotypes as predictors, with adjustment for natural-log transformed fasting insulin values, sex, age, geographical covariates (if applicable), and age squared (Framingham only) to evaluate the association under an additive genetic model. Association analysis was performed by individual studies using SNPTTEST (19), STATA (20), PLINK (21), or LMEKIN (R kinship package) software (22). Genome-wide association inflation coefficients were estimated for each discovery cohort using the genomic control (GC) method (23) and applied subsequently to each individual SNP association test statistics to correct for cryptic relatedness. The λ GC value for the final meta-analysis of proinsulin adjusted for fasting insulin was 1.01. The inverse-variance fixed effects meta-analysis method was used to evaluate the pooled regression estimates for additively coded SNPs using METAL (24). Sex interaction effects were evaluated with a function in the GWAMA software (25).

Follow-up SNP selection and analysis. We carried forward to stage 2 the most significant SNP from each of 21 independent loci that showed association with proinsulin in stage 1 analyses at $P < 1 \times 10^{-5}$. Additionally, two SNPs near the $P < 1 \times 10^{-5}$ threshold (in *ASAP2* and a gene desert region) were carried forward as a result of biological plausibility (*ASAP2* is involved in vesicular transport) and/or consistency of direction of effect in all discovery stage 1 studies (both loci). We genotyped these 23 variants in 11 additional stage 2 studies totaling 16,378 nondiabetic participants of European ancestry (Supplementary Table 1; genotyping assays and conditions are available upon request). We meta-analyzed stage 1 and stage 2 results using inverse-variance weighted fixed effects meta-analysis methods, including up to 27,079 participants.

Additional analyses of expression and expression quantitative trait loci (eQTL) studies are described in the Supplementary Data.

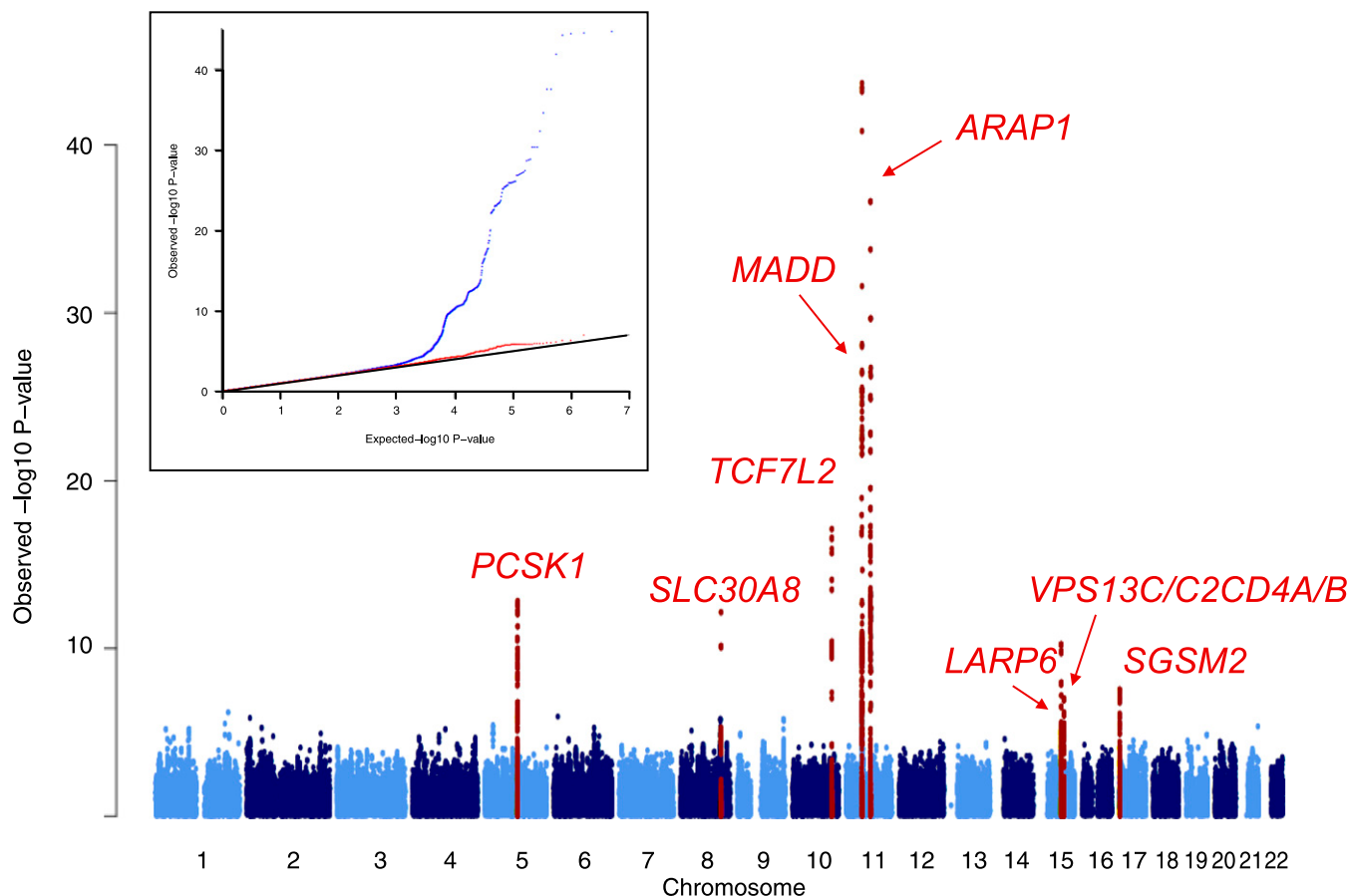


FIG. 1. Manhattan plot of the association P values for fasting proinsulin adjusted for fasting insulin. Directly genotyped and imputed SNPs are plotted with their meta-analysis P values (as $-\log_{10}$ values) as a function of genomic position (NCBI Build 36). The SNPs that achieved genome-wide significance ($P < 5 \times 10^{-8}$) on follow-up are shown in red. *Insert:* Quantile-quantile (Q-Q) plot for fasting proinsulin adjusted for fasting insulin. The expected null distribution is plotted along the diagonal, the entire distribution of observed P values is plotted in blue, and a distribution that excludes the nine novel findings is plotted in red.

TABLE 1
Loci associated with fasting proinsulin levels at genome-wide levels of statistical significance

SNP	Nearest gene	CHR	Position	Allele (effect/other)	Freq.	Discovery P value ($n = 10,538$ – $10,701$)	Replication β (SE) ($n = 5,894$ – $15,898$)	Replication β (SE), BMI adjusted ($n = 5,893$ – $15,890$)	Replication β (SE), BMI + FG adjusted ($n = 5,744$ – $15,741$)	Combined P value ($n = 19,168$ – $26,599$)	Heterogeneity I^2 in % (Q -test P value)
rs11603334	<i>ARAP1</i>	11	72110633	A/G	0.15	3.3×10^{-38}	0.0928 (0.0053)	0.0943 (0.0056)	0.0938 (0.0052)	3.2×10^{-102}	$77.2 (1.9 \times 10^{-7})$
rs10501320	<i>MADD</i>	11	47250375	G/C	0.72	2.1×10^{-45}	0.0805 (0.0064)	0.0748 (0.0069)	0.0775 (0.0064)	1.1×10^{-88}	$68.1 (8.7 \times 10^{-4})$
rs6235	<i>PCSK1</i>	5	95754654	G/C	0.28	7.7×10^{-14}	0.0394 (0.005)	0.0407 (0.0052)	0.0438 (0.0049)	9.8×10^{-27}	$30.2 (0.14)$
rs7903146	<i>TCF7L2</i>	10	114748339	T/C	0.30	3.5×10^{-18}	0.0321 (0.007)	0.0330 (0.0076)	0.0303 (0.0063)	2.3×10^{-20}	$66.2 (1.6 \times 10^{-3})$
rs4502156	<i>VPS13C/C2CD4A/B</i>	15	60170447	T/C	0.58	3.8×10^{-11}	0.0294 (0.0044)	0.0278 (0.0046)	0.0260 (0.0043)	3.5×10^{-20}	$45.1 (0.03)$
rs11558471	<i>SLC30A8</i>	8	118254914	A/G	0.69	4.2×10^{-13}	0.0280 (0.0049)	0.0280 (0.0053)	0.0273 (0.0048)	3.1×10^{-18}	$58.5 (0.01)$
rs10838687	<i>MADD</i> (2nd signal)*	11	47269468	T/G	0.80	8.6×10^{-7}	0.0253 (0.0050)	0.0240 (0.0049)	0.0246 (0.0049)	6.9×10^{-12}	$18.4 (0.26)$
rs1549318	<i>LARP6</i>	15	68896201	T/C	0.61	8.0×10^{-8}	0.0192 (0.0051)	0.0209 (0.0056)	0.0180 (0.005)	2.4×10^{-10}	$52.7 (0.02)$
rs4790333	<i>SGSM2</i>	17	2209453	T/C	0.45	2.2×10^{-8}	0.0154 (0.0043)	0.0205 (0.0046)	0.0179 (0.0042)	3.0×10^{-9}	$56.8 (3.5 \times 10^{-3})$
rs9727115	<i>SNX7</i> **	1	98949841	G/A	0.64	5.3×10^{-6}	0.0133 (0.0051)	0.0130 (0.005)	0.0134 (0.0045)	2.4×10^{-7}	$0.00 (0.65)$

β -Coefficients for the effect allele are shown after adjustments for sex, age, geographic covariates (if applicable), and age squared (Framingham only), besides the additional adjustments denoted in the column headings. The combined P values are shown for the joint analysis with the basic adjustments listed above: CHR, chromosome; Freq, frequency. *This second SNP in *MADD* was selected for further testing based on a suggestive P value (8.6×10^{-7}) when analyses for chromosome 11 were conditioned for the top two signals (*MADD* and *ARAP1*). The conditional analysis results are shown. ***SNX7* only reached genome-wide significance after adjusting for fasting glucose ($P = 5.4 \times 10^{-9}$). Another signal (rs306549 in *DDX31*) reached genome-wide significance in women ($P = 2.0 \times 10^{-8}$) but not in men ($P = 0.17$); see text for details.

RESULTS

Genome-wide association meta-analysis (stage 1). We conducted a two-stage association study in individuals of European descent (total $n = 27,079$, with $n = 10,701$ in the discovery stage). Cohort and phenotype information can be found in Supplementary Table 1, and the study design is outlined in Supplementary Fig. 3. A total of 21 independent variants (including two SNPs identified during conditional analyses, see below) met our statistical threshold for follow-up ($P < 1 \times 10^{-5}$; Fig. 1). The clean dataset showed no systematic deviation from the null expectation, with the exception of the tail of the distribution (Fig. 1, insert).

Follow-up studies (stage 2) and global (stage 1 + stage 2) meta-analysis for 23 loci. We followed up 23 SNPs (the 21 mentioned above plus 2 others that approached our significance threshold and were selected as a result of biological plausibility; see RESEARCH DESIGN AND METHODS) in 11 cohorts totaling up to 16,378 nondiabetic individuals of European descent (Table 1 and Supplementary Table 2). Joint meta-analysis of discovery and follow-up cohorts ($n = 27,079$) revealed nine signals at eight loci reaching genome-wide significance ($P < 5 \times 10^{-8}$), of which two are novel (*SGSM2*, *LARP6*), five have previously been associated with glucose metabolism and/or T2D (*TCF7L2*, *SLC30A8*, *MADD*, *VPS13C/C2CD4A/B*, and *ARAP1*), and one (*PCSK1*) has been previously implicated in obesity and associated with proinsulin levels, although not at genome-wide significance (Table 1 and Fig. 2). Adjusting for BMI, fasting glucose, or both did not attenuate these signals. Of note, when adjusting for fasting glucose or both fasting glucose and BMI (but not BMI alone), one other locus, *SNX7*, reached genome-wide significance ($P = 5.4 \times 10^{-9}$ and 1.5×10^{-8} , respectively).

Conditional analyses on the two strongest signals revealed that the *MADD* locus harbors two independent signals 19 kb apart (rs10501320 and rs10838687; $r^2 = 0.068$ in HapMap CEU), whereas a second independent signal near *ARAP1* did not replicate (Fig. 2B, Table 1, and Supplementary Table 2). Among the nine replicated SNPs, individual loci explained between 0.2 and 1.4% of the variance in proinsulin in the discovery samples and up to 2.3% of the variance in the follow-up samples. Together, the nine genome-wide significant SNPs explained between 5.4 and 7.7% of the proinsulin variance in the discovery samples and 8.1% of the variance in the RISC cohort, one of the few follow-up cohorts with genotypes available for all nine SNPs.

Heterogeneity and sex-stratified analyses. We noted some degree of heterogeneity in our joint meta-analyses (Table 1). Part of the heterogeneity arose from the METSIM sample, which enrolled only men; exclusion of this cohort from our meta-analysis reduced the heterogeneity. We also stratified our analyses by sex and tested for a SNP $\times$ sex interaction (26). Our overall findings remained essentially unchanged after sex stratification, and heterogeneity was attenuated (e.g., $I^2 = 77.2\%$, heterogeneity $P = 1.9 \times 10^{-7}$ for combined men and women, whereas $I^2 = 64.6\%$, heterogeneity $P = 4.5 \times 10^{-4}$ [men] and $I^2 = 55.6\%$, heterogeneity $P = 0.01$ [women] in stratified analyses). Furthermore, tests for interaction with sex among SNPs that reached our follow-up significance threshold revealed a locus (rs306549 in *DDX31*) where a genome-wide significant association was seen in women ($P = 2.0 \times 10^{-8}$; Supplementary Fig. 4A) but not men ($P = 0.17$; Supplementary Fig. 4B; sex interaction $P = 8.9 \times 10^{-5}$). Although removal

of the METSIM cohort improved the heterogeneity score and produced nominal significance for the association in men ($P = 0.02$), the effect size remained threefold stronger in women than in men (β -coefficient 0.0427 vs. 0.0165, respectively).

To provide further reassurance regarding any residual heterogeneity, we repeated our meta-analyses based on P values (rather than β -coefficients) and meta-analyzed the resulting z scores. Our findings were essentially unchanged, suggesting that heterogeneity in the β -estimates across cohorts has not produced spurious results.

Exploration of proinsulin processing mechanisms.

Proinsulin is initially cleaved to 32,33-split proinsulin and further to insulin and C-peptide before secretion (Supplementary Fig. 1); we were therefore interested in the effects of the nine top SNPs on these traits. The proinsulin-raising alleles of each SNP were consistently associated with higher 32,33-split proinsulin levels, with effect sizes following the rank order of proinsulin effect sizes. Nearly all associations reached nominal conventional levels of statistical significance in this smaller dataset of 4,103–6,343 individuals with measures of 32,33-split proinsulin levels (all $P < 1.5 \times 10^{-3}$, with the exception of the conditional signal at *MADD*). The insulinogenic index (27), which measures dynamic insulin secretion during the first 30 min after an oral glucose load and was available in 14,956 subjects, showed nominal associations for four loci. Of these, the proinsulin-raising alleles were associated with a lower insulinogenic index at *VPS13C/C2CD4A/B*, *TCF7L2*, and *SLC30A8* and higher at *ARAP1* (Table 2).

We detected no nominal associations with fasting C-peptide ($P > 0.05$). Given the differences in hepatic clearance of insulin and C-peptide, we also performed sensitivity analyses to account for any possible impact this may have had on our results. We adjusted proinsulin levels for fasting C-peptide rather than fasting insulin in two cohorts (Ely and Botnia-PPP); comparison of β -estimates showed that the majority of loci had very similar effect sizes and the same rank order was preserved, arguing against noticeable discrepancies between the two adjustment schemes.

Association with other glycemic traits. To clarify potential mechanisms, the top nine signals (*ARAP1*, two at *MADD*, *PCSK1*, *TCF7L2*, *VPS13C/C2CD4A/B*, *SLC30A8*, *LARP6*, and *SGSM2*) were also examined in relation to other glucometabolic traits (fasting and 2-h postload glucose and insulin, homeostasis model assessment estimates of β -cell function [HOMA-B] and insulin resistance [HOMA-IR] [28], glycated hemoglobin [A1C], T2D, and BMI [Table 3]). We investigated results available from MAGIC meta-analyses of GWAS of glycemic traits (3–5) and obtained T2D and BMI results in collaboration with the Diabetes Genetics Replication And Meta-analysis (DIAGRAM) (29) and Genomewide Investigation of Anthropometric measures (GIANT) (30) consortia, respectively. Nominal associations ($P < 0.05$) were found for fasting glucose (with the proinsulin-raising allele increasing fasting glucose levels at *MADD*, *SLC30A8*, *TCF7L2*, and *VPS13C/C2CD4A/B* and decreasing fasting glucose levels at *ARAP1* and *PCSK1*), fasting insulin (increased levels at *ARAP1*, *LARP6*, and *SGSM2* and decreased levels at *TCF7L2*), HOMA-B (decreased at *MADD*, *SLC30A8*, *VPS13C/C2CD4A/B*, and *TCF7L2* and increased at *PCSK1*, *ARAP1*, and *LARP6*), insulin resistance as measured by HOMA-IR (increased at *LARP6* and *SGSM2* and decreased at *TCF7L2*), and 2-h postload glucose (decreased at *SLC30A8* and *VPS13C/C2CD4A/B* and increased at *ARAP1* and *TCF7L2*).

We detected no significant associations for 2-h postload insulin or insulin sensitivity as estimated by the Matsuda index (31) (Table 3).

Associations with T2D were confirmed for four known T2D loci (*SLC30A8*, *ARAP1*, *VPS13C/C2CD4A/B*, and *TCF7L2*; Table 3). Counterintuitively, the proinsulin-raising allele of *ARAP1* (formerly known as *CENTD2* and reported as such in DIAGRAM+) (29) was associated with a lower fasting glucose (0.019 mg/dL per A allele; $P = 1.7 \times 10^{-4}$), lower A1C (0.023%; $P = 0.02$), and a lower risk of T2D (odds ratio [OR] 0.88; $P = 7.8 \times 10^{-6}$; Table 3). The two novel loci (*LARP6* and *SGSM2*) did not show significant associations with T2D (OR [95% CI]: 1.01 [0.95–1.07] and 1.01 [0.96–1.08], respectively), indicating that if they increase T2D risk they do so to an extent confined within the bounds of narrow 95% CI.

Fine-mapping, copy number variants, and tissue expression. We used MACH (32) or IMPUTE (19) applied to the 1000 Genomes CEU reference panel (www.1000genomes.org) to carry out imputation of ~8 million autosomal SNPs with minor allele frequency >1%. Analysis of 1000 Genomes-imputed data in the four discovery cohorts indicates that although there are low-frequency (1–5%) genetic variants that influence levels of circulating proinsulin, these are found in the same loci that contain common proinsulin-influencing variants, and none of them yield substantially stronger signals than the index SNP at each locus (Supplementary Fig. 5).

Using current databases of copy number variants (33) and the SNAP software (http://www.broadinstitute.org/mpg/snap/index.php; CEU, HapMap release 22), we checked whether any of the proinsulin-associated SNPs were within 500 kb and in linkage disequilibrium (LD) with any of the SNPs known to tag copy number variants in the human genome. No copy number variant tag SNPs with $r^2 > 0.3$ were found within 500 kb of our lead SNPs.

To guide identification of the gene responsible for each association signal, we also examined the gene expression profile of selected genes in each associated region across a range of human tissues, including islets and fluorescence-activated cell (FAC)-sorted β -cells (Fig. 3A–F and Supplementary Fig. 6). We defined 1-Mb intervals around the lead SNP at each locus and prioritized biologically plausible genes as gleaned from the literature (see Box in Supplementary Data). We were able to demonstrate β -cell expression of most genes examined (Fig. 3F). However, at the *LARP6* locus, *CT62* is expressed exclusively in testis, likely excluding it as a relevant gene in this context. At the *ARAP1* locus, *STARD10* is expressed more strongly in pancreatic and islet tissue than any other tissue type; similarly, at the *VPS13C* locus both *C2CD4A* and *C2CD4B* demonstrate higher expression in pancreas and islets than all other tissue types.

We also studied the expression of the transcript for the gene closest to the index SNP at each of the nine replicated loci in human islets isolated from 55 nondiabetic and 9 diabetic individuals. Of the nine loci, *PCSK1* ($P = 0.02$) and *MADD* ($P = 0.07$) demonstrated 35–45% lower expression in subjects with T2D compared with control subjects.

Functional exploration. We evaluated whether any of the associated SNPs was in strong LD with a potentially causal variant. We used SNPper (34) to classify all SNPs in strong LD with the lead SNP ($r^2 \geq 0.8$) within a 1-MB region. We found that *PCSK1* rs6235 codes for a non-synonymous variant (S690T), which is in perfect LD with rs6234, another missense variant (Q665E); both were

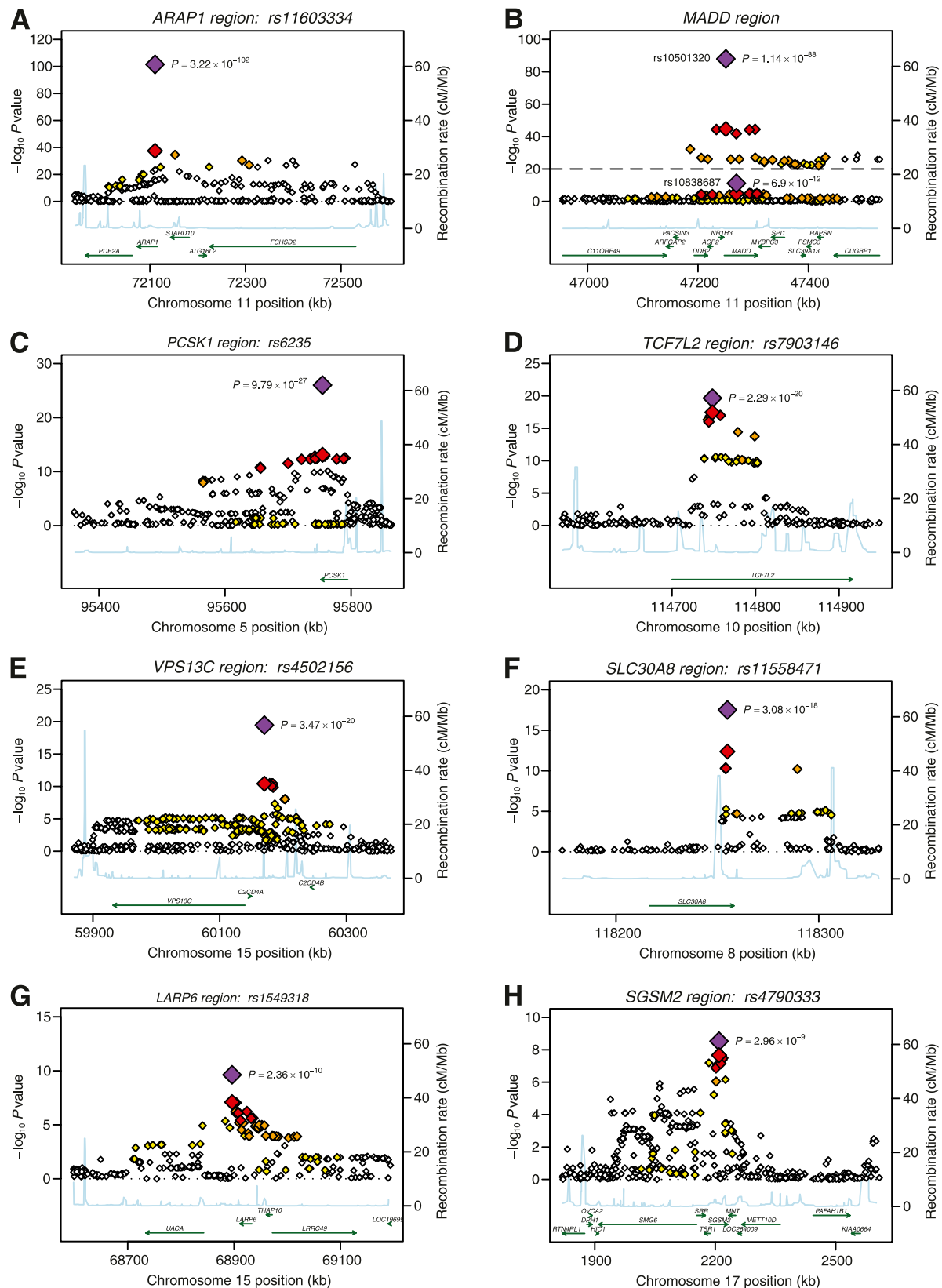


FIG. 2. Regional plots of eight genomic regions containing novel genome-wide significant associations. For each region, directly genotyped and imputed SNPs are plotted with their meta-analysis P values (as $-\log_{10}$ values) as a function of genomic position (NCBI Build 36). In each panel, the stage 1 discovery SNP taken forward to stage 2 follow-up is represented by a purple diamond (with global meta-analysis P value), with its stage 1 discovery P value denoted by a red diamond with bolded borders. Estimated recombination rates (taken from HapMap) are plotted to reflect the local LD structure around the associated SNPs and their correlated proxies (according to a white to red scale from $r^2 = 0$ to 1, based on pairwise r^2 values from HapMap CEU). Gene annotations were taken from the University of California Santa Cruz genome browser. **A:** *ARAP1* region; **B:** *MADD* region; **C:** *PCSK1* region; **D:** *TCF7L2* region; **E:** *VPS13C/C2CD4A/B* region; **F:** *SLC30A8* region; **G:** *LARP6* region; **H:** *SGSM2* region.

TABLE 2
Association of proinsulin loci with insulin-processing traits

SNP	Nearest gene	Alleles proinsulin-raising/ other	32,33-split proinsulin (<i>n</i> = 4,103–6,343)		Fasting insulin (<i>n</i> = 36,775–38,355)		HOMA-B (<i>n</i> = 35,046–36,606)		Insulinogenic index (<i>n</i> = 10,902–14,956)		C-peptide (<i>n</i> = 1,537–2,956)	
			β	<i>P</i> value	β	<i>P</i> value	β	<i>P</i> value	β	<i>P</i> value	β	<i>P</i> value
rs11603334	ARAP1	A/G	0.146	1.2×10^{-25}	0.012	0.02	0.016	1.4×10^{-4}	0.049	1.1×10^{-5}	0.017	0.22
rs10501320	MADD	G/C	0.124	1.4×10^{-20}	-0.003	0.45	-0.015	8.1×10^{-5}	-0.005	0.67	-3.0×10^{-4}	0.98
rs6235	PCSK1	G/C	0.057	4.5×10^{-6}	0.001	0.78	0.008	0.02	-0.004	0.70	-0.008	0.46
rs7903146	TCF7L2	T/C	0.042	1.5×10^{-3}	-0.012	0.005	-0.020	1.4×10^{-7}	-0.063	9.6×10^{-7}	0.000	0.98
rs4502156	VPS13C/C2CD4A/B	T/C	0.039	4.5×10^{-4}	-0.002	0.67	-0.010	0.004	-0.053	3.5×10^{-9}	-0.017	0.11
rs11558471	SLC30A8	A/G	0.059	1.1×10^{-5}	-0.005	0.21	-0.017	1.4×10^{-6}	-0.032	0.001	0.003	0.81
rs10838687	MADD (2nd signal)	T/G	0.006	0.71	0.006	0.22	0.007	0.06	0.0029	0.79	-0.0084	0.47
rs1549318	LARP6	T/C	0.043	1.2×10^{-4}	0.009	0.02	0.009	0.007	-0.001	0.88	-0.014	0.18
rs4790333	SGSM2	T/C	0.040	3.2×10^{-4}	0.008	0.04	0.002	0.55	0.009	0.33	-0.015	0.15

β -Coefficients are adjusted for age, sex, and study-specific covariates (if applicable).

predicted to be nondamaging by Polyphen (35) and SIFT (36). At *SLC30A8*, the proinsulin-associated SNP rs11558471 is a perfect proxy for the known T2D-associated SNP rs13266634, encoding R325W. The T allele (encoding tryptophan) is predicted to be benign by PolyPhen, but damaging by SIFT. We found no other strong ($r^2 > 0.8$) correlations in HapMap CEU with potentially functional SNPs within 1 Mb of the lead signals.

We also tested whether any of the proinsulin-associated SNPs might influence proximal (*cis*) expression of human transcripts, in tissues available to us that had been paired to genetic data. We found a significant association ($P = 0.01$ permutation threshold) of rs1549318 with expression levels of *LARP6* in adipose tissue. SNP rs1549318 is located ~37 kb from *LARP6*, and the proinsulin-raising T allele is associated with lower levels of expression. Analysis of an eQTL database from human liver indicated that the proinsulin-raising A allele of the lead SNP at the *SGSM2* locus (rs179456) was associated with increased liver expression of *TRPS1* ($P = 0.004$).

CAD. We constructed unweighted and weighted genotype scores composed of the nine genome-wide significant proinsulin-raising alleles, with weights defined by the β -coefficients from our replication meta-analysis, and tested the association of these scores with CAD in the Coronary Artery Disease Genome-wide Replication And Meta-Analysis (CARDIoGRAM) (37) ($n = \sim 22,000$ CAD case subjects and 60,000 control subjects) and C4D (38) ($n = 15,420$ CAD case subjects and 15,062 control subjects) datasets. Neither weighted nor unweighted genotype scores reached nominal significance in either dataset ($P = 0.47$ and 0.81 for unweighted and weighted scores in CARDIoGRAM, respectively; $P = 0.60$ and 0.43 for unweighted and weighted scores in C4D, respectively).

DISCUSSION

We report the first meta-analysis of genome-wide association datasets for circulating fasting proinsulin. We adjusted proinsulin for fasting insulin levels, aiming to capture an increase in proinsulin relative to the nonspecific activation of the insulin processing pathway induced by generalized insulin resistance (Supplementary Fig. 2). Loci that simply influence insulin resistance are typically sought by a GWAS for fasting insulin or more sophisticated measures of insulin sensitivity (3,4,6). Thus, we hoped to identify loci that indicate the inability of the β -cell to process proinsulin adequately in response to metabolic demands.

We have identified nine signals at eight loci associated with higher proinsulin levels (see Box in Supplementary Data). Two of these loci (*LARP6* and *SGSM2*) have not been previously related to metabolic traits. A 10th signal emerged after sex-stratified analyses; an explanation for the female-specific genome-wide significant association at *DDX31* requires fine-mapping to identify the causal gene. Although the function of the *DDX31* gene product is unknown, other members of the DEAD-box protein family have been implicated in sex-specific processes such as spermatogenesis (39). We have also replicated at the genome-wide level previously reported nominal associations of *MADD*, *TCF7L2*, *VPS13C/C2CD4A/B*, *SLC30A8*, and *PCSK1* with proinsulin (6,14–17,40). The knowledge that *TCF7L2*, *SLC30A8*, *VPS13C/C2CD4A/B*, and *ARAP1* are established T2D loci provides reassurance that a quest for genetic determinants of proinsulin can serve to identify disease-associated signals. Interestingly, the proinsulin-raising

TABLE 3
Association of proinsulin loci with other glycemic traits

SNP	Nearest gene	Alleles proinsulin- raising/other	Fasting glucose (n = 44,601- 46,186)	HOMA-IR (n = 35,512- 37,072)	Matsuda index (n = 7,055- 9,561)	2-h glucose (n = 15,088- 15,252)	2-h insulin (n = 6,923- 7,083)	HbA _{1c} (n = 33,736- 44,731)	T2D (n = 8,130 case subjects/38,987 control subjects)	BMI (n = 123,260- 123,865)
rs11603334	<i>ARAP1</i>	A/G	β /OR P value -0.019 1.7×10^{-4}	0.016 0.105 0.025	-0.014 0.226 0.009	0.009 0.043 -0.001	-0.050 0.294 -0.012	-0.023 0.016 0.004	0.88 7.8×10^{-6} 1.01	0.012 0.065 0.014
rs10501320	<i>MADD</i>	G/C	β /OR P value 3.1 $\times 10^{-9}$	-0.015 0.775 0.001	0.009 0.413 0.008	-0.001 0.605 0.034	-0.012 0.459 0.015	0.004 0.595 0.013	1.01 0.629 1.02	0.014 0.012 0.016
rs6235	<i>PCSK1</i>	G/C	β /OR P value 4.8 $\times 10^{-4}$	0.001 0.837 -0.010	-0.008 0.435 0.014	0.034 0.109 0.108	0.273 0.401 0.004	0.053 0.015 0.015	1.02 1.40 1.40	0.016 0.002 -0.019
rs7903146	<i>TCF7L2</i>	T/C	β /OR P value 2.8 $\times 10^{-8}$	0.023 0.034 0.002	0.023 0.244 -0.004	0.108 4.3 $\times 10^{-7}$ -0.071	0.004 0.772 0.012	0.029 0.029 0.002	2.2 $\times 10^{-51}$ 1.07 1.07	-0.019 2.4 $\times 10^{-4}$ -0.003
rs4502156	<i>VPS13C/C2CD4A/B</i>	T/C	β /OR P value 8.4 $\times 10^{-8}$	0.002 0.648 -0.017	-0.004 0.658 0.014	-0.001 2.1 $\times 10^{-4}$ -0.002	0.012 0.321 0.080	0.779 0.079 0.013	1.07 0.001 1.14	-0.003 0.528 -0.012
rs11558471	<i>SLC30A8</i>	A/G	β /OR P value 2.6 $\times 10^{-11}$	0.027 0.714 0.004	-0.017 0.197 -0.006	-0.002 1.4 $\times 10^{-4}$ 0.014	0.080 0.275 0.003	0.013 0.080 -0.001	1.11 3.3 $\times 10^{-8}$ 0.98	-0.012 0.027 -0.013
rs10838687	<i>MADD</i> (2nd signal)	T/G	β /OR P value 0.014	0.383 0.009 0.009	0.578 0.001 0.001	0.512 0.008 -0.010	0.830 0.010 -0.010	0.876 -0.005 -0.005	0.383 1.01 1.01	0.020 0.003 0.583
rs1549318	<i>LARP6</i>	T/C	β /OR P value 0.698	0.044 0.009 0.004	0.906 0.009 0.009	0.621 -0.0002 0.008	0.556 0.008 0.008	0.444 -0.005 -0.005	0.667 1.01 0.626	-0.583 -0.005 0.261
rs4790333	<i>SGSM2</i>	T/C	β /OR P value 0.328	0.023 0.023	0.338	0.992	0.519	0.425		

β -Coefficients are adjusted for age, sex, and study-specific covariates (if applicable). OR is for T2D only as a dichotomous trait.

alleles at *TCF7L2*, *SLC30A8*, and *VPS13C/C2CD4A/B* cause impairment of β -cell function, as estimated by HOMA-B and the insulinogenic index. By raising proinsulin but lowering insulin secretion, these loci point to defects in the insulin processing and secretion pathway, distal to the first enzymatic step. Such a hypothesis is consistent with postulated modes of action for *TCF7L2* (41) and *SLC30A8* (42); *VPS13C*, by influencing protein trafficking across membrane compartments, could also affect the same process. Further fine-mapping and functional experiments will be required to establish the precise mechanism at this locus.

ARAP1, which harbors the strongest proinsulin association, provides an intriguing counterpoint. Under its previous designation of *CENTD2* it was recently associated with T2D (29); however, the T2D-associated allele is associated with lower proinsulin levels, as well as lower β -cell function (HOMA-B and insulinogenic index). This suggests that the genetic defect that gives rise to T2D at this locus causes a generalized downregulation of insulin secretion (e.g., through a reduction in β -cell mass/function or very early defects in insulin processing) and stands in contrast with *TCF7L2*, *SLC30A8*, and *VPS13C/C2CD4A/B*. A corollary of the divergent effect of these loci on T2D is that both disproportionate elevations and reductions in proinsulin can indicate β -cell dysfunction. Of the genes that lie within 1 Mb of the *ARAP1* association signal, we have demonstrated islet expression in the four strong biological candidates we examined (*ARAP1*, *INPL1*, *STARD10*, and *RAB6A*); however, expression of *STARD10* was much higher in pancreas than in any other human tissue, and of all genes tested at the *ARAP1* locus *STARD10* was expressed most strongly in islets, indicating that the role of its protein product in the transfer of phospholipids to membranes may be particularly relevant to this cell type.

LARP6 is a ribonucleoprotein identified in the current study as a novel locus associated with increased fasting proinsulin levels. It is involved in the regulation of translation and subcellular localization of collagen I, in a manner dependent upon both the RNA-binding and La domains (43). The associated SNP rs1549318 is located within a region of high LD, which spans the gene and includes a number of SNPs within the RNA-binding domain. Although the link between *LARP6* and proinsulin levels is not clear, it is nominally associated with fasting insulin and HOMA-IR, but not T2D. It may therefore represent a marker of insulin resistance and perhaps other related common dysmetabolic conditions.

In previous publications we have reported the association of *C2CD4B* with fasting glucose (3) and that of the nearby locus *VPS13C* with 2-h glucose (4); *C2CD4B* is also associated with T2D in Japanese (44), with supportive evidence found in Europeans (3,44). Here we show that the same genomic region is associated with fasting proinsulin. The strongest association with proinsulin reported here (rs4502156) and those associated with fasting glucose and 2-h glucose may represent independent signals, since they are all in relatively weak LD in HapMap CEU Europeans: rs4502156 versus rs11071657 (best fasting glucose signal), $r^2 = 0.306$; rs4502156 vs. rs17271305 (best 2-h glucose signal), $r^2 = 0.450$; and rs11071657 versus rs17271305, $r^2 = 0.287$. On the other hand, in Europeans our proinsulin-associated SNP is in strong LD ($r^2 = 0.967$) with the T2D-associated SNP reported by Yamauchi et al. (44). Although four strong biological candidates (*C2CD4A*, *C2CD4B*, *VPS13C*, and *RORA*, a gene that encodes a

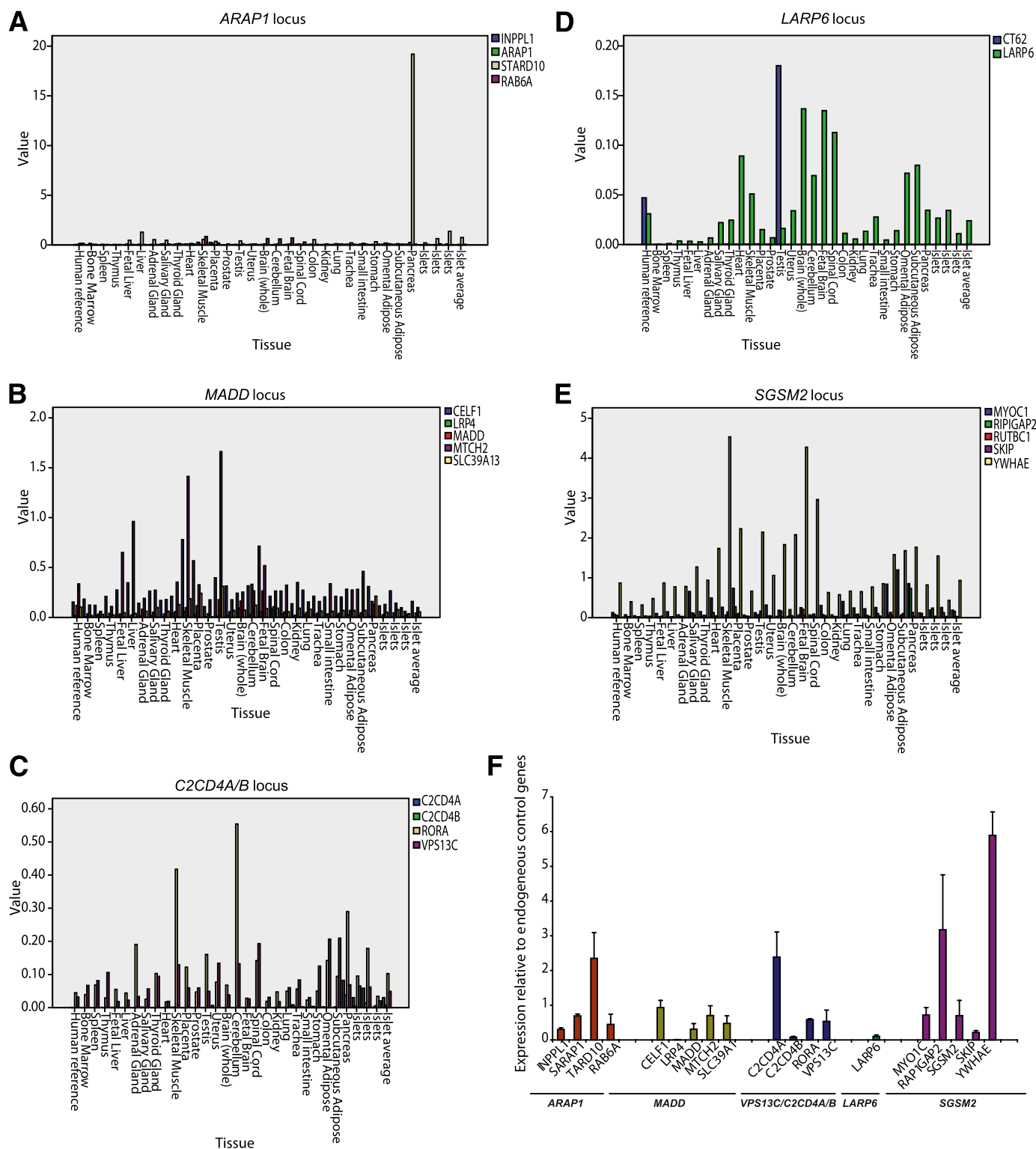


FIG. 3. Expression profiles of biologically plausible genes within each associated locus across a range of human tissue types, including islet preparations from three donors. Expression levels determined with respect to the geometric mean of three endogenous control assays. A: *ARAP1* region; B: *MADD* region; C: *VPS13C/C2CD4A/B* region; D: *LARP6* region; E: *SGSM2* region. F: Expression levels of genes near the proinsulin-associated variants in human FAC-sorted β -cells. Data are expression means $\pm$ SD of the relative expression measured by quantitative PCR obtained from three human nondiabetic donors.

member of the NR1 subfamily of nuclear hormone receptors) are expressed in FAC-sorted β -cells, the relative expression of the first two is much higher in islets than in other human tissues, again suggesting that these two genes, encoding nuclear factors that are upregulated in

response to inflammation, may be particularly relevant to endocrine pancreatic function.

The genome-wide association of a missense variant in *PCSK1* with fasting proinsulin also serves as a positive control. *PCSK1* encodes the protein prohormone convertase

1/3 (PC1), which is the first enzyme in the proinsulin processing pathway, where it cleaves proinsulin to 32,33-split proinsulin (Supplementary Fig. 1). A related enzyme, PC2, acts on 32,33-split proinsulin in the second processing step. People deficient in PC1 become obese at an early age and exhibit pituitary hypofunction because of the lack of several mature peptide hormones (45), whereas PC2-null mice demonstrate increased levels of 32,33-split proinsulin (46). The rs6235 SNP reported here results in the substitution of a serine residue for threonine at position 690 of the molecule; the minor allele (Thr) is associated with higher proinsulin levels. A nominal association of the same allele with higher proinsulin levels has recently been reported (40); its association with higher BMI is only nominal here, but confirms a previous report (47). This specific amino acid change has been shown not to affect enzyme catalysis or maturation of the protein *in vitro* (47), but the COOH terminus of the protein (where S690T is located, adjacent to a conserved proline residue) is known to direct the correct subcellular targeting of the protein as well as stabilizing and partially inhibiting PC1. Although one might expect lower levels of the reaction product (32,33-split proinsulin) in carriers of the risk allele, the potential diversion of the substrate down its alternate path (giving rise to 65,66-split proinsulin, whose assay typically has 60% cross-reactivity with 32,33-split proinsulin) requires further study. Alternatively, if changes in the activity of PC1 also affect that of PC2 (for instance, by competing for inhibitory peptides) one might see reductions in the catalytic function of both enzymes and accumulation of both proinsulin and 32,33-split proinsulin.

Because of the reported relationship between proinsulin levels and coronary events (11–13), the identification of genetic determinants of proinsulin levels might help shed light on whether hyperproinsulinemia is a mediator of CAD or a byproduct of a shared etiological mechanism. If hyperproinsulinemia is causally associated with an increased risk of CAD, one might expect that SNPs that specifically and selectively raise proinsulin levels should increase the risk of CAD given an adequately powered study. We have not observed such an effect for a genotype score constructed with the genome-wide significant proinsulin association signals. Assuming conservative approximations of the reported effect sizes of proinsulin on CAD (OR ~1.5 per 1-SD increase in proinsulin) (12,13), and of the nine SNPs reported here on circulating proinsulin (5%), a CAD cohort like CARDIoGRAM has 99% power to detect an effect of proinsulin SNPs on CAD. The absence of statistical significance argues against a direct etiological role of proinsulin on CAD.

In summary, we have identified nine loci that associate with fasting proinsulin levels. Several of these loci increase risk of T2D; interestingly, both proinsulin-raising and lowering alleles can lead to T2D through decreases in insulin secretion, indicating defects distal or proximal to the first enzymatic step in proinsulin conversion, respectively. Other genetic determinants of proinsulin levels do not necessarily lead to higher T2D risk, suggesting that it is not a mere elevation in proinsulin, but rather the specific impairment in proinsulin processing and the reaction of the β -cell to this defect that determine whether ultimately β -cell insufficiency will cause pathological hyperglycemia. The direct elevation of fasting proinsulin out of proportion to fasting insulin does not seem to increase risk of CAD.

ACKNOWLEDGMENTS

Please see the Supplementary Data.

REFERENCES

- Billings LK, Florez JC. The genetics of type 2 diabetes: what have we learned from GWAS? *Ann N Y Acad Sci* 2010;1212:59–77
- Prokopenko I, Langenberg C, Florez JC, et al. Variants in *MTNR1B* influence fasting glucose levels. *Nat Genet* 2009;41:77–81
- Dupuis J, Langenberg C, Prokopenko I, et al.; DIAGRAM Consortium; GIANT Consortium; Global BPgen Consortium; A. Hamsten on behalf of Procardis Consortium; MAGIC investigators. New genetic loci implicated in fasting glucose homeostasis and their impact on type 2 diabetes risk. *Nat Genet* 2010;42:105–116
- Saxena R, Hivert MF, Langenberg C, et al.; GIANT consortium; MAGIC investigators. Genetic variation in GIPR influences the glucose and insulin responses to an oral glucose challenge. *Nat Genet* 2010;42:142–148
- Soranzo N, Sanna S, Wheeler E, et al.; WTCCC. Common variants at 10 genomic loci influence hemoglobin A_{1c} (C) levels via glycemic and non-glycemic pathways. *Diabetes* 2010;59:3229–3239
- Ingelsson E, Langenberg C, Hivert MF, et al.; MAGIC investigators. Detailed physiologic characterization reveals diverse mechanisms for novel genetic loci regulating glucose and insulin metabolism in humans. *Diabetes* 2010;59:1266–1275
- Rich SS, Goodarzi MO, Palmer ND, et al. A genome-wide association scan for acute insulin response to glucose in Hispanic-Americans: the Insulin Resistance Atherosclerosis Family Study (IRAS FS). *Diabetologia* 2009;52:1326–1333
- Palmer ND, Langefeld CD, Ziegler JT, et al. Candidate loci for insulin sensitivity and disposition index from a genome-wide association analysis of Hispanic participants in the Insulin Resistance Atherosclerosis (IRAS) Family Study. *Diabetologia* 2010;53:281–289
- Røder ME, Porte D Jr, Schwartz RS, Kahn SE. Disproportionately elevated proinsulin levels reflect the degree of impaired B cell secretory capacity in patients with noninsulin-dependent diabetes mellitus. *J Clin Endocrinol Metab* 1998;83:604–608
- Wareham NJ, Byrne CD, Williams R, Day NE, Hales CN. Fasting proinsulin concentrations predict the development of type 2 diabetes. *Diabetes Care* 1999;22:262–270
- Lindahl B, Dinesen B, Eliasson M, et al. High proinsulin concentration precedes acute myocardial infarction in a nondiabetic population. *Metabolism* 1999;48:1197–1202
- Yudkin JS, May M, Elwood P, Yarnell JW, Greenwood R, Davey Smith G; aarperhilly Study. Concentrations of proinsulin like molecules predict coronary heart disease risk independently of insulin: prospective data from the Caerphilly Study. *Diabetologia* 2002;45:327–336
- Zethelius B, Byberg L, Hales CN, Lithell H, Berne C. Proinsulin is an independent predictor of coronary heart disease: report from a 27-year follow-up study. *Circulation* 2002;105:2153–2158
- Loos RJF, Franks PW, Francis RW, et al. TCF7L2 polymorphisms modulate proinsulin levels and beta-cell function in a British Europid population. *Diabetes* 2007;56:1943–1947
- Kirchoff K, Machicao F, Haupt A, et al. Polymorphisms in the *TCF7L2*, *CDKAL1* and *SLC30A8* genes are associated with impaired proinsulin conversion. *Diabetologia* 2008;51:597–601
- González-Sánchez JL, Martínez-Larrad MT, Zabena C, Pérez-Barba M, Serrano-Ríos M. Association of variants of the *TCF7L2* gene with increases in the risk of type 2 diabetes and the proinsulin:insulin ratio in the Spanish population. *Diabetologia* 2008;51:1993–1997
- Stolerman ES, Manning AK, McAteer JB, et al. *TCF7L2* variants are associated with increased proinsulin/insulin ratios but not obesity traits in the Framingham Heart Study. *Diabetologia* 2009;52:614–620
- Hanley AJ, D'Agostino R Jr, Wagenknecht LE, et al.; Insulin Resistance Atherosclerosis Study. Increased proinsulin levels and decreased acute insulin response independently predict the incidence of type 2 diabetes in the insulin resistance atherosclerosis study. *Diabetes* 2002;51:1263–1270
- Marchini J, Howie B, Myers S, McVean G, Donnelly P. A new multipoint method for genome-wide association studies by imputation of genotypes. *Nat Genet* 2007;39:906–913
- StataCorp. *Stata Statistical Software: Release 10*. College Station, TX, StataCorp LP, 2007
- Purcell S, Neale B, Todd-Brown K, et al. PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am J Hum Genet* 2007;81:559–575
- R Development Core Team. *R: A language and environment for statistical computing*. Vienna, Austria, R Foundation for Statistical Computing, 2007

23. Devlin B, Roeder K. Genomic control for association studies. *Biometrics* 1999;55:997–1004
24. Willer CJ, Li Y, Abecasis GR. METAL: fast and efficient meta-analysis of genomewide association scans. *Bioinformatics* 2010;26:2190–2191
25. Mägi R, Morris AP. GWAMA: software for genome-wide association meta-analysis. *BMC Bioinformatics* 2010;11:288
26. Magi R, Lindgren CM, Morris AP. Meta-analysis of sex-specific genome-wide association studies. *Genet Epidemiol* 2010;34:846–853
27. Seltzer HS, Allen EW, Herron AL Jr, Brennan MT. Insulin secretion in response to glycemic stimulus: relation of delayed initial release to carbohydrate intolerance in mild diabetes mellitus. *J Clin Invest* 1967;46:323–335
28. Matthews DR, Hosker JP, Rudenski AS, Naylor BA, Treacher DF, Turner RC. Homeostasis model assessment: insulin resistance and beta-cell function from fasting plasma glucose and insulin concentrations in man. *Diabetologia* 1985;28:412–419
29. Voight BF, Scott LJ, Steinthorsdottir V, et al.; MAGIC investigators; GIANT Consortium. Twelve type 2 diabetes susceptibility loci identified through large-scale association analysis. *Nat Genet* 2010;42:579–589
30. Speliotes EK, Willer CJ, Berndt SI, et al.; MAGIC; Procardis Consortium. Association analyses of 249,796 individuals reveal 18 new loci associated with body mass index. *Nat Genet* 2010;42:937–948
31. Matsuda M, DeFronzo RA. Insulin sensitivity indices obtained from oral glucose tolerance testing: comparison with the euglycemic insulin clamp. *Diabetes Care* 1999;22:1462–1470
32. Li Y, Abecasis GR. Mach 1.0: rapid haplotype reconstruction and missing genotype inference (Abstract). *Am J Hum Genet* 2006;S79:2290
33. Craddock N, Hurles ME, Cardin N, et al.; Wellcome Trust Case Control Consortium. Genome-wide association study of CNVs in 16,000 cases of eight common diseases and 3,000 shared controls. *Nature* 2010;464:713–720
34. Riva A, Kohane IS. A SNP-centric database for the investigation of the human genome. *BMC Bioinformatics* 2004;5:33
35. Sunyaev S, Ramensky V, Koch I, Lathe W 3rd, Kondrashov AS, Bork P. Prediction of deleterious human alleles. *Hum Mol Genet* 2001;10:591–597
36. Ng PC, Henikoff S. Predicting deleterious amino acid substitutions. *Genome Res* 2001;11:863–874
37. Schunkert H, König IR, Kathiresan S, et al.; Cardiogenics; CARDIOGRAM Consortium. Large-scale association analysis identifies 13 new susceptibility loci for coronary artery disease. *Nat Genet* 2011;43:333–338
38. Peden JF, Hopewell JC, Saleheen D, et al.; Coronary Artery Disease (C4D) Genetics Consortium. A genome-wide association study in Europeans and South Asians identifies five new loci for coronary artery disease. *Nat Genet* 2011;43:339–344
39. Dufau ML, Tsai-Morris C, Tang P, Khanum A. Regulation of steroidogenic enzymes and a novel testicular RNA helicase. *J Steroid Biochem Mol Biol* 2001;76:187–197
40. Heni M, Haupt A, Schäfer SA, et al. Association of obesity risk SNPs in *PCSK1* with insulin sensitivity and proinsulin conversion. *BMC Med Genet* 2010;11:86
41. da Silva Xavier G, Loder MK, McDonald A, et al. TCF7L2 regulates late events in insulin secretion from pancreatic islet beta-cells. *Diabetes* 2009;58:894–905
42. Nicolson TJ, Bellomo EA, Wijesekara N, et al. Insulin storage and glucose homeostasis in mice null for the granule zinc transporter ZnT8 and studies of the type 2 diabetes-associated variants. *Diabetes* 2009;58:2070–2083
43. Cai L, Fritz D, Stefanovic L, Stefanovic B. Binding of LARP6 to the conserved 5' stem-loop regulates translation of mRNAs encoding type I collagen. *J Mol Biol* 2010;395:309–326
44. Yamauchi T, Hara K, Maeda S, 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* 2010;42:864–868
45. Farooqi IS, Volders K, Stanhope R, et al. Hyperphagia and early-onset obesity due to a novel homozygous missense mutation in prohormone convertase 1/3. *J Clin Endocrinol Metab* 2007;92:3369–3373
46. Furuta M, Carroll R, Martin S, et al. Incomplete processing of proinsulin to insulin accompanied by elevation of Des-31,32 proinsulin intermediates in islets of mice lacking active PC2. *J Biol Chem* 1998;273:3431–3437
47. Benzinou M, Creemers JW, Choquet H, et al. Common nonsynonymous variants in *PCSK1* confer risk of obesity. *Nat Genet* 2008;40:943–945

Genome-wide association study of CNVs in 16,000 cases of eight common diseases and 3,000 shared controls

The Wellcome Trust Case Control Consortium*

Copy number variants (CNVs) account for a major proportion of human genetic polymorphism and have been predicted to have an important role in genetic susceptibility to common disease. To address this we undertook a large, direct genome-wide study of association between CNVs and eight common human diseases. Using a purpose-designed array we typed ~19,000 individuals into distinct copy-number classes at 3,432 polymorphic CNVs, including an estimated ~50% of all common CNVs larger than 500 base pairs. We identified several biological artefacts that lead to false-positive associations, including systematic CNV differences between DNAs derived from blood and cell lines. Association testing and follow-up replication analyses confirmed three loci where CNVs were associated with disease—*IRGM* for Crohn's disease, *HLA* for Crohn's disease, rheumatoid arthritis and type 1 diabetes, and *TSPAN8* for type 2 diabetes—although in each case the locus had previously been identified in single nucleotide polymorphism (SNP)-based studies, reflecting our observation that most common CNVs that are well-typed on our array are well tagged by SNPs and so have been indirectly explored through SNP studies. We conclude that common CNVs that can be typed on existing platforms are unlikely to contribute greatly to the genetic basis of common human diseases.

Genome-wide association studies (GWAS) have been extremely successful in associating SNPs with susceptibility to common diseases, but published SNP associations account for only a fraction of the genetic component of most common diseases, and there has been considerable speculation about where the 'missing heritability'¹ might lie. Chromosomal rearrangements can cause particular rare diseases and syndromes², and recent reports have suggested a role for rare CNVs, either individually or in aggregate, in susceptibility for a range of common diseases, notably neurodevelopmental diseases^{3–6}. So far, there have been relatively few reported associations between common diseases and common CNVs (see for example refs 7–11), which might simply reflect incomplete catalogues of common CNVs or the lack of reliable assays for their large-scale typing. Here we report the results of our direct association study, identify the population properties of the set of CNVs studied, describe novel analytical methods to facilitate robust analyses of CNV data, and document artefacts that can afflict CNV studies.

We designed an array to measure copy number for the majority of a recently compiled inventory of CNVs from an extensive discovery experiment¹², and several other sources. We then used the array to type 3,000 common controls and 2,000 cases of each of the diseases: bipolar disorder, breast cancer, coronary artery disease, Crohn's disease, hypertension, rheumatoid arthritis, type 1 diabetes and type 2 diabetes. These eight diseases make a major impact on public ill health¹³, cover a range of aetiologies and genetic predispositions, and have been extensively studied via SNP-based GWAS, including our earlier Wellcome Trust Case Control Consortium (WTCCC) study¹⁴.

Pilot experiment, array content, assay and samples

Pilot experiment. We undertook a pilot experiment to compare three different platforms for assaying CNVs and to assess the merits

of different experimental design parameters (see Supplementary Information for full details). On the basis of the pilot data, we chose the Agilent Comparative Genomic Hybridization (CGH) platform, and aimed to target each CNV with ten distinct probes, although in the analyses below we include any CNV targeted by at least one probe (Supplementary Fig. 9). Our analysis of the pilot CGH data indicated that the quality of the copy number signal for genotyping (rather than for discovery) at a CNV is reduced when the reference sample is homozygous deleted, in effect because the reference channel then just measures noise. To minimize this effect we used a fixed pool of DNAs as the reference sample throughout our main experiment.

Array content. Informed by our pilot experiment, we designed the CNV-typing array in a collaboration with the Genome Structural Variation Consortium (GSV) in which a preliminary set of candidate CNVs was shared at an early stage with the WTCCC. Table 1 summarizes the design content of the array, and Fig. 1 illustrates the various categories of designed loci unsuitable for association analysis. (See Methods for further details.)

Assay. In brief (see Supplementary Information for further details), the Agilent assay differentially labels parallel aliquots of the test sample and reference DNA (a pool of genomic UK lymphoblastoid cell-line DNAs from nine males and one female prepared in a single batch for all experiments) and then combines them, hybridizes to the array, washes and scans. Intensity measurements for the two different labels are made at each probe separately for the test and reference DNA. These act as surrogates for the amount of DNA present, with analyses typically relying on the ratio of test to reference intensity measurements at each probe.

Samples. A total of 19,050 case-control samples were sent for assaying: ~2,000 for each of the eight diseases and ~3,000 common controls (these were equally split between the 1958 British Birth Cohort

*Lists of authors and their affiliations appear at the end of the paper.

Table 1 | Discovery source for regions targeted on the genotyping array

Source of loci	Number of loci targeted	Number of loci analysed	Number of loci polymorphic with good calls
CNVs			
GSV discovery project	10,835	10,217	3,096
Affymetrix 500k	18	14	12
Affymetrix 6.0	83	81	47
Illumina 1M	82	81	18
WTCCC CNV loci	231	209	108
Novel sequence			
Novel insert regions	292	292	151
Total	11,541	10,894	3,432

GSV CNVs were prioritized according to extent of polymorphism in European discovery samples. See Methods for full details of other sources.

(58C) and the UK Blood Services (UKBS) controls). These were augmented by 270 HapMap1 samples (see ref. 12 for additional analyses of the HapMap data) and 610 duplicate samples for quality control purposes. About 80% of samples from the WTCCC SNP GWAS were used here. (See Supplementary Information for further details of sample collections, inclusion criteria, and so on.)

Data pre-processing, CNV calling and quality control

Data pre-processing. For each sample, raw data from the CNV experiment consist of intensity measurements for the test and reference sample for each probe. There are numerous choices at the data pre-processing stage, including how to normalize data to reduce inter-individual variation, and how to combine the information across the set of probes within a CNV. Several novel analytical tools substantially improved data quality, but no single approach works well for every CNV, so we carried through 16 pre-processing pipelines to maximize the number of CNVs that can be tested for association. (See Supplementary Information Section 4 for illustrations and a sense of the challenges.)

CNV calling. The objective in CNV calling at each CNV is to assign each assayed sample to a diploid copy-number class, which represents the sum of copy numbers on each allele. This step is analogous

to, but typically considerably more challenging than, calling genotypes from SNP-chip data. Available assays for SNPs are more robust and have better signal-to-noise properties than do available assays for CNVs¹⁵. We used two different statistical methods ('CNVtools', which is available as a Bioconductor package, and 'CNVCALL') in parallel to estimate the number of copy-number classes at each CNV and assign individuals to these classes. (See Supplementary Information for further details.) Figure 2 illustrates three multi-allelic CNVs that have attracted attention in the literature in part due to the difficulties in obtaining reliable data.

Quality control. After the application of quality control metrics to each sample and each CNV (see Methods), 17,304 case-control samples (of 19,050 initially) were available for association testing. There were 3,432 CNVs with more than one copy-number class which passed quality control and were included in subsequent analyses. At these CNVs, concordance of calls between pairs of duplicate samples was 99.7%.

Properties of CNVs

Single-class CNVs. Of the 10,894 distinct putative CNVs typed on the array after removal of detectable redundancies, 60% are called with a single copy-number class, and so cannot be tested for association. After detailed analyses (see Methods) we estimate that just under half of these are probably not polymorphic. For the remainder, the combination of the experimental assay and analytical methods we have used do not allow separate copy-number classes to be distinguished.

Multi-class CNVs. A total of 4,326 CNVs were called with multiple classes. Of these, 3,432 passed quality control filters, which in practice means that the classes were well separated and thus that it was possible to assign individuals to copy-number classes with high confidence. Most of these CNVs (88%) have two or three copy-number classes, consistent with their having only two variants, or alleles, present in the population (we refer to these as bi-allelic CNVs). Note that some loci involving both duplications and deletions could be called with only three classes if both homozygote classes are very rare.

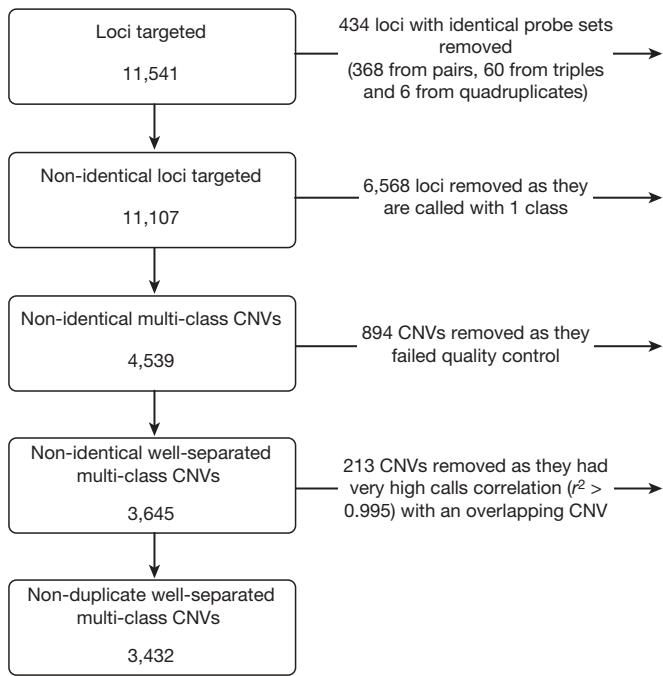


Figure 1 | Flowchart showing which CNVs are included on the array. The chart shows the reasons for CNVs being removed from consideration (the column of arrows and text to the right of the figure) from those originally targeted on the array, and the number of CNVs remaining at each stage of filtering.

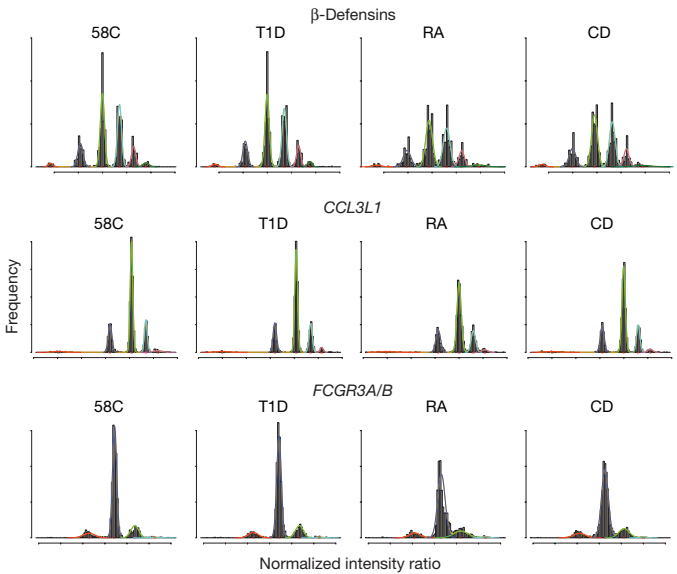


Figure 2 | Illustrative CNVs. Histograms of three multi-allelic CNVs (one per row) previously reported to be associated with autoimmune diseases: β -defensin (CNVR3771.10), *CCL3L1* (CNVR7077.12) and *FCGR3A/B* (CNVR383.1), showing 6, 5 and 4 fitted copy number classes, respectively. The histogram of normalized intensity ratios is shown for one control and the three autoimmune collections. Histograms are overlaid by the fitted distribution used to model each class (variously the red, blue, light-green, cyan, magenta and dark green curves). In all such figures, the area under the fitted curve of a particular colour is the same for all collections at the same CNV. 58C, 1958 British Birth Cohort; CD, Crohn's disease; RA, rheumatoid arthritis; T1D, type 1 diabetes.

Allele frequencies. Supplementary Fig. 21 shows the distribution of minor allele frequency (MAF) for bi-allelic CNVs passing quality control. For example, 44% of autosomal CNVs passing quality control had $\text{MAF} < 5\%$. This is shifted towards lower MAFs compared to commonly used SNP chips. One consequence is that for given sample sizes association studies will tend to have lower power than for SNP studies. (See Supplementary Fig. 22 for power estimates.) Extrapolating from analyses described in ref. 12 gives an estimate that the 3,432 CNVs we directly tested represent 42–50% of common ($\text{MAF} > 5\%$) CNVs greater than 0.5 kilobases (kb) in length which are polymorphic in a population with European ancestry.

Tagging by SNPs. In the literature discussing the possible role of common CNVs in human disease there has been controversy over the extent to which CNVs will be in linkage disequilibrium with SNPs. If linkage disequilibrium between CNVs and SNPs were similar to that between SNPs, SNPs typed in GWAS would act as tags not only for untyped SNPs but also for untyped CNVs, and in turn SNP-based GWAS would have indirectly explored CNVs for association with disease. (See refs 16 and 17 for opposite views.) Our large-scale genotyping of an extensive CNV catalogue allows us to settle this question. In fact, CNVs that are typed well in our experiment are in general well-tagged by SNPs—almost to the same extent that SNPs are well-tagged by SNPs (Supplementary Fig. 20). Among variable 2- and 3-class CNVs passing quality control with $\text{MAF} > 10\%$, 79% have $r^2 > 0.8$ with at least one SNP; for those with $\text{MAF} < 5\%$, 22% have $r^2 > 0.8$ with at least one SNP. This is consistent with the vast majority having arisen from unique mutational events at some time in the past. It follows that genetic variation in the form of common CNVs which can be typed on our array, has already been explored indirectly for association with common human disease through the SNP-based GWAS. In passing, we note that the high correlations between our CNV calls and SNP genotypes provide strong indirect evidence that our CNV calls are capturing real variation. It is possible that the CNVs that we cannot type well are systematically different from those that we can type, for example in having many more copy-number classes, and hence perhaps that they arise from repeated mutational events in the same region, in which case their linkage disequilibrium properties with SNPs could also be systematically different from the CNVs that we can type. We have no data that bear on this question, and it seems likely that such CNVs will be difficult to type genome wide on any currently available platforms.

Association testing

We performed association testing at each of the CNVs that passed quality control, in two parallel approaches. First, we applied a frequentist likelihood ratio association test that combines calling (using CNVtools) and testing into a single procedure, using an extension of an approach previously described¹⁸. Second, we undertook Bayesian association analyses in which the posterior probabilities from CNVCALL were used to calculate a Bayes factor to measure strength of association with the disease phenotypes. Important features of both sets of analyses are that they correctly handle uncertainty in assignment of individuals to copy-number classes, and by allowing for some systematic differences in intensities between cases and controls, that they provide robustness against certain artefacts which could arise from differences in data properties between cases and controls. There were no substantial differences between the broad conclusions from the frequentist and Bayesian approaches.

Our association analyses were based on a model in which a single parameter quantifies the increase in disease risk between successive copy-number classes, analogous to that underlying the trend test for SNP data. Various analyses of the robustness of our procedure, adequacy of the model, and lack of population structure were encouraging (see Methods and Supplementary Information). For example, Supplementary Fig. 23 shows quantile–quantile plots for the primary comparison of each case collection against the combined controls, and for the analogous comparisons between the two control

groups. These show generally good agreement with the expectation under the null hypothesis.

Careful analysis of our association testing revealed several sophisticated biological artefacts that can lead to false-positive associations. These include dispersed duplications, whereby the variation at a CNV is not in the chromosomal location in the reference sequence to which the probes in the CNV uniquely match, and a DNA source effect whereby particular CNVs, and genome-wide intensity data, can look systematically different according to whether the assayed DNA was derived from blood or cell lines. (See Box 1 for illustrations and further details.)

Independent replication of putative association signals is a routine and essential aspect of SNP-based association studies. Particularly in view of the differences in data quality between SNP assays and CNV assays, and the wide range of possible artefacts in CNV studies, replication is even more important in the CNV context. Several possible approaches to replication are available. When a CNV is well tagged by a SNP (or SNPs), replication can be undertaken by assessment of the signal at the tag SNP(s) in an independent sample, either by typing the SNP or by reference to published data. Where no SNP tag is available, direct typing of the CNV in independent samples is necessary, either using a qualitative breakpoint assay or a quantitative DNA dosage assay. In most cases there will be a choice of assays. Notably, replication via SNPs was possible for 15 out of 18 of the CNVs for which we undertook replication based on analysis of our penultimate data freeze.

Figure 3 plots *P* values for the primary frequentist analysis for each CNV in each collection. Table 2 provides details of the top, replicated, association signals in our experiment after visual inspection of cluster plots to detect artefacts not removed by earlier quality control. Cluster plots for each CNV in Table 2 are shown in Supplementary Figs 18 and 19, and Supplementary Files 2 and 3.

There is one positive control for the diseases we studied, namely the known CNV association at the *IRGM* locus in Crohn's disease⁷. Reassuringly, our study found this association ($P = 1 \times 10^{-7}$, odds ratio (OR) = 0.68; throughout, all ORs are with respect to increasing copy number).

We identified three loci—HLA for Crohn's disease, rheumatoid arthritis and type 1 diabetes; *IRGM* for Crohn's disease; and *TSPAN8* for type 2 diabetes—at which CNVs seemed to be associated with disease, all of which we convincingly replicated through previously typed SNPs that tag the CNV, and a fourth locus (CNV7113.6) at which there is suggestive evidence for association and replication in both Crohn's disease and type 1 diabetes.

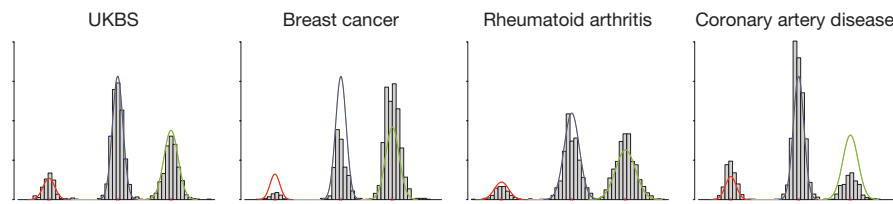
We observed CNVs in the HLA region associated variously with Crohn's disease (CNVR2841.20, $P = 1.2 \times 10^{-5}$, OR = 0.80), rheumatoid arthritis (CNVR2845.14, $P = 1.4 \times 10^{-39}$, OR = 1.77) and type 1 diabetes (CNVR2845.46, $P = 8 \times 10^{-153}$, OR = 0.2). Copy number variation has previously been documented on various HLA haplotypes¹⁹ and owing to the extensive linkage disequilibrium in the region it is perhaps not unexpected to have found CNV associations in our direct study. Linkage disequilibrium across the HLA region has hampered attempts to fine-map causal variation across this locus, and we have no evidence that suggests that the HLA CNVs associated with autoimmune diseases in this study represent signals independent of the known associated haplotypes.

We identified two distinct CNVs 22 kb apart upstream of the *IRGM* gene, both of which are associated with Crohn's disease. The longer CNV (CNVR2647.1, $P = 1.0 \times 10^{-7}$, OR = 0.68) has previously been identified⁷ as a possible causal variant on an associated haplotype first identified through SNP GWAS¹⁴, and acted as our positive control; however, the association of the smaller CNV (CNVR2646.1, $P = 1.1 \times 10^{-7}$, OR = 0.68, located < 2 kb downstream from a different gene, *C5orf62*) is a novel observation. Although direct experimental evidence links the associated haplotypes with variation in expression of the *IRGM* gene, it does not bear on the question of which of the two CNVs or the associated SNPs

Box 1 | Some artefacts in CNV association testing

Some types of artefacts, such as population structure and calling artefacts, are very similar to those seen in SNP studies. Others, related to differences in data properties between cases and controls, can be potentially more serious for CNVs^{26,27}. In this box we draw attention to some specific artefacts of biological interest that we observed and which researchers should consider as explanations of putative disease-relevant associations. We note that, for the unwary, some of these artefacts could easily survive 'replication' of an association.

First, we consider dispersed CNVs. Box 1 Fig. 1 shows cluster plots for a particular CNV (CNVR2664.1) that shows a strong case-control association signal for breast cancer cases ($P = 5 \times 10^{-143}$, higher copy number for disease) with a related signal for rheumatoid arthritis ($P = 3 \times 10^{-27}$), and a signal in the opposite direction for coronary artery disease ($P = 4 \times 10^{-30}$). The right-hand class (green curve) has a higher frequency in breast cancer (and rheumatoid arthritis), and a lower frequency in coronary heart disease. (The area under the green curve is the same for each collection.) This turned out to be an artefact caused by differences in sex ratio in the various case and control samples (breast cancer, 100% female; rheumatoid arthritis, 74% female; coronary artery disease, 22% female; controls, 50% female). Comparing breast cancer cases against female controls abolished the signal. The CNV is annotated as being on chromosome 5 and all 10 probes in the CNV map uniquely to chromosome 5 in the human reference sequence. However, we found that SNPs which tagged the variation at this CNV all mapped to the X chromosome and that the region containing the probes for this CNV is present on the X chromosome in the Venter genome. We conclude that the CNV is a dispersed duplication, with the variation actually occurring on the X chromosome, and not on chromosome 5. We found one similar example, of a CNV (CNVR1065.1, featured in Table 2 as a replicated association) annotated as mapping uniquely to chromosome 2 that shows a strong signal in type 1 diabetes and rheumatoid arthritis. Careful examination shows it to be another dispersed duplication where the polymorphism is located in the HLA region, and is well tagged by HLA SNPs known to be associated with both diseases. Supplementary Fig. 27 shows the clear evidence from interchromosomal linkage disequilibrium that these two loci are dispersed duplications.



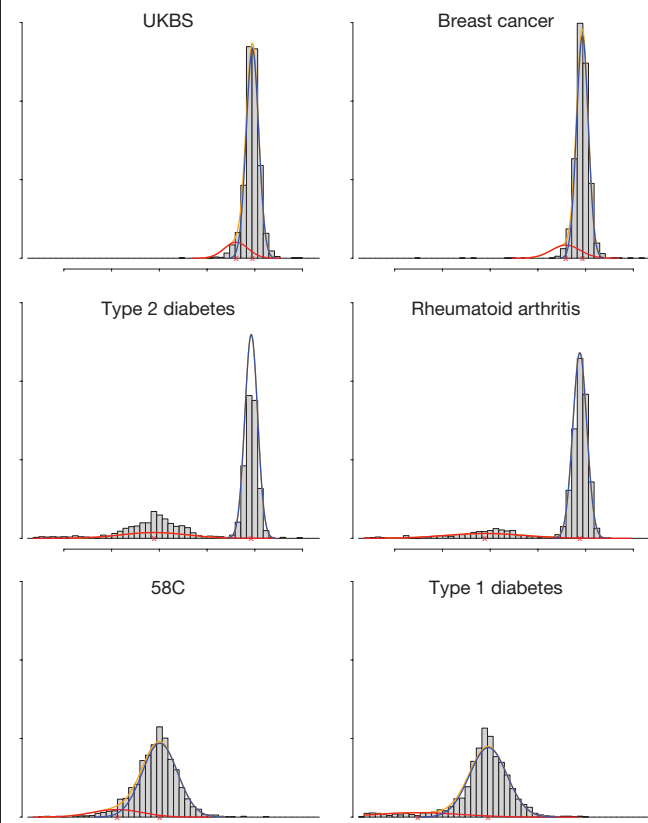
Box 1 Fig. 1 | Dispersed duplications leading to false-positive associations.

Second, we consider variation in DNA source. Box 1 Fig. 2 shows cluster plots for a different CNV (CNVR866.8) with marked differences in type 2 diabetes as compared with the UKBS controls (or against just the 58C controls). The plots show histograms of normalized intensity ratios for six collections. Examination of the pattern across collections is interesting. The collections in the top row show a single tight peak towards the right of the plot. Those in the bottom row show a single, more dispersed peak to the left. The collections in the middle row show evidence of both peaks. It turns out that for collections with the tight peak all DNA samples were derived from blood whereas all samples in the two collections with the single dispersed peak had DNA derived from cell lines. The remaining collections contain some DNAs derived from both sources. This CNV (and many others) thus exhibit systematically different behaviour depending on the DNA source. Box 1 Fig. 3 shows a plot of the second (PC2) and third (PC3) principal components of the array-wide intensity data (plot created using all samples after quality control from all ten collections using data from all CNVs, with each point representing one sample, with the points coloured according to whether that sample was derived from blood (red) or cell lines (blue)). It is clear that these two components can almost perfectly classify samples according to the source of the DNA.

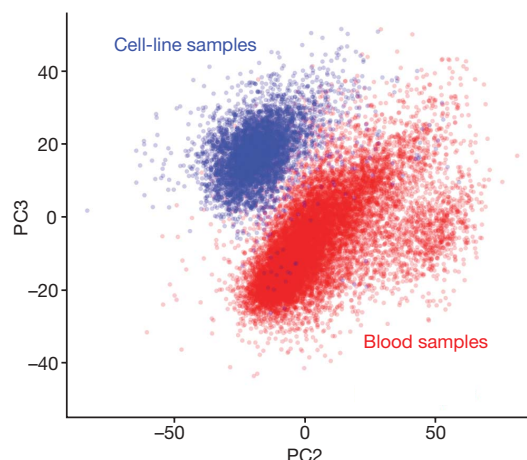
Lymphoblastoid cell lines are typically grown from transformed B cells, whereas DNA extracted from blood comes largely from a mixture of white blood cells. One specific feature of B cells is that each B cell has been subject to its own pattern of rearrangements around the immunoglobulin genes via the process

of V(D)J recombination²⁸. This suggests a natural candidate for our observed DNA source effect, and indeed the CNV illustrated in Box 1 Fig. 2 is located close to one of the immunoglobulin genes, as are the other instances we have found of similar gross DNA source effects. But it is not the whole story. Principal components analysis of genome-wide intensity data with any probe mapping to within 1 megabase of an immunoglobulin gene excluded from analysis (Supplementary Fig. 29) shows reasonably clear discrimination by DNA source (although less clear than when all probes are included), with many probes, genome-wide, contributing to the discrimination.

Dispersed duplications and DNA source effects represent interesting biological artefacts. We also observed more prosaic effects. As one example, Supplementary Fig. 30 shows that there are systematic effects on probe intensity of the row of the plate in which a sample was run.



Box 1 Fig. 2 | DNA source effect leading to false-positive associations.



Box 1 Fig. 3 | Principal component analysis showing DNA source effect.

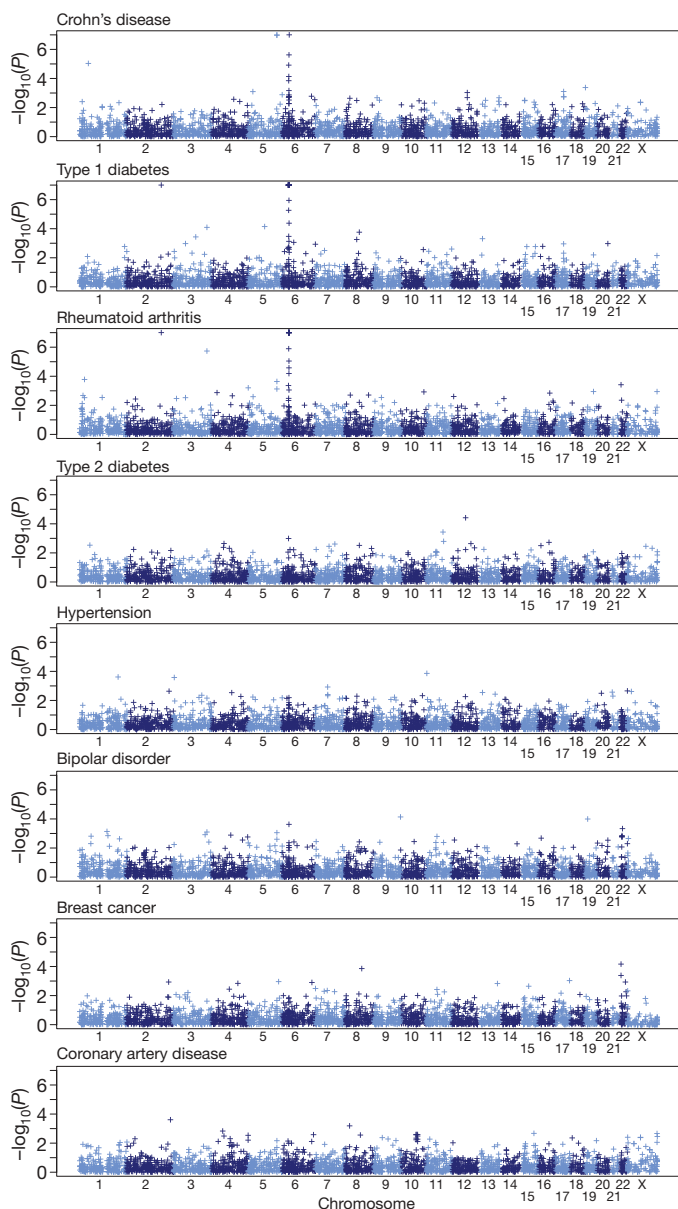


Figure 3 | Genome-wide association results. Distribution of $-\log_{10}(P)$ along the 23 chromosomes where P is the P -value for the one degree-of-freedom test of association for each disease. The x-axis shows the chromosomes numbered from 1 (on the left) to X (on the right). CNVs included in these plots were filtered on the basis of a clustering quality score (see Supplementary Information for details) and manual inspection of the most significant associations. The two apparent associations on chromosome 2 for rheumatoid arthritis and type 1 diabetes result from a dispersed duplication in which the variation is actually located within the HLA locus (see Box 1).

might be driving this variation⁷. Our conditional regression analyses on the two CNVs and SNPs on this haplotype do not point significantly to any one of these as being more strongly associated.

SNP variation in the *TSPAN8* locus was recently shown to be reproducibly associated with type 2 diabetes²⁰, but the potential role of a CNV is a novel observation. This CNV (CNVR5583.1, $P = 3.9 \times 10^{-5}$, OR = 0.85) potentially encompasses part or all of an exon of *TSPAN8* and so is a plausible causal variant. The most significantly associated SNP identified in the recent meta-analysis is only weakly correlated with the CNV as originally tested ($r^2 = 0.17$), and so the CNV may simply be weakly correlated with the true causal variant. Closer examination of probe-level data at this CNV indicates a series of different events (including an inverted duplication and a deletion) resulting in more complex haplotypes than those tested for

association by our automated approach. With this more refined definition of haplotypes the signal is stronger. (See Supplementary Information for details.)

CNVR7113.6 lies within a cluster of segmentally duplicated sequences that demarcate one end of a common 900-kb inversion polymorphism on chromosome 17 that has previously been shown to be associated with number of children and higher meiotic recombination in females²¹. The CNV shows weak evidence for association with Crohn's disease ($P = 1.8 \times 10^{-3}$, OR = 1.15) and type 1 diabetes ($P = 1.1 \times 10^{-3}$, OR = 1.13), but is in extremely high linkage disequilibrium ($r^2 = 1$) with SNPs known to tag the inversion, and so is in tight linkage disequilibrium with a long haplotype spanning many possible causal variants. This CNV encompasses at least one spliced transcript, but no high-confidence gene annotations. Fine-mapping the causal variant within such a long, tightly linked haplotype is likely to prove challenging.

In addition to the loci in Table 2, we undertook replication on 13 other loci, detailed in Supplementary Table 13, for which there was some evidence of association ($P < 1 \times 10^{-4}$ or $\log_{10}(\text{Bayes factor (BF)}) > 2.1$) in our analysis of the penultimate data freeze. Replication results were negative for all these loci. Several other loci for which there is weak evidence ($P < 1 \times 10^{-4}$ or $\log_{10}(\text{BF}) > 2.6$) for association in our final data analysis are listed in Supplementary Table 14.

To investigate further the potential role of CNVs as pathogenically relevant variants underlying published SNP associations, we took 94 association intervals in type 1 diabetes, Crohn's disease and type 2 diabetes (excluding the HLA), and for the index SNP in each association interval assessed its correlation with our calls at 3,432 CNVs. We identified two index SNPs as being correlated with an r^2 of greater than 0.5 with a called CNV. The SNPs were: rs11747270 with both CNVR2647.1 and CNVR2646.1 (*IRGM*), and rs2301436 with CNVR3164.1 (*CCR6*), both for Crohn's disease. Both of these association intervals were also identified in an independent analysis using CNV calls on HapMap samples by ref. 12.

As a further test of our approach, we examined three multi-allelic CNVs that have attracted attention in the literature, both for the challenges of obtaining reliable data and for putative associations with a range of autoimmune diseases: *CCL3L1* (our CNVR7077.12); β -defensins (CNVR3771.10); and *FCGR3A/B* (CNVR383.1)^{10,22–24}. Encouragingly, all three CNVs pass quality control and give good quality data. Figure 2 shows cluster plots for these CNVs in our experiment. The best calls for the three CNVs required the use of two analysis pipelines (sets of choices about normalization and probe summaries) different from our standard pipeline. None of the CNVs shows significant association with the three autoimmune diseases in our study after allowance for multiple testing. In particular, we do not see formally significant evidence to replicate the reported association for *CCL3L1* and rheumatoid arthritis²⁴ (nominal $P = 0.058$).

We also assessed whether CNVs that delete all or part of exons might be enriched among disease susceptibility loci, even if our study were not well-powered enough to see statistically significant evidence of association for individual CNVs. To do so, we compared the 53 exonic deletion CNVs¹² that passed quality control with collections of CNVs of the same size, matched for MAF and numbers of classes. We used a (two-sided) Wilcoxon signed-rank test²⁵ to ask whether the strength of signal for association (measured by Bayes Factors) was systematically different for the exon-deletion CNVs as compared to the matched CNVs. We found no evidence that deletion of an exon systematically changed evidence for association (see Supplementary Information). In a related analysis, we compared CNVs passing quality control that were well tagged by SNPs ($r^2 > 0.8$) to those passing quality control that were not, again matching for MAF and number of classes (excluding low MAF CNVs and those failing Hardy–Weinberg equilibrium tests to avoid calling artefacts). There was no evidence that CNVs passing quality control that are not well tagged by SNPs are enriched for stronger signals of association compared to those which were well tagged (see Supplementary Information).

Table 2 | Replicated CNV associations and those at replicated loci

Disease	Chr.	Start (bp) (CNV)	Length (kb)	Locus	Fitted no. classes*	Combined controls (P)†	Extended reference (P)	Combined controls reference (log ₁₀ (BF))‡	Extended reference (log ₁₀ (BF))	Combined controls (OR)§	Extended reference (OR)	MAF		Replication size		Replication size (P)
												Ctrl¶	Cases#	Ctrl	Cases	
T2D	12	69,818,942 (CNVR5583.1)	1.0	TSPAN8	3	3.9×10 ⁻⁵	2.5×10 ⁻⁶	2.8	4.3	0.85	0.85	0.40	0.36	5,579	4,549☆	3.9×10 ⁻⁵
CD	5	150,157,836 (CNVR2646.1)	3.9	IRGM	3	1.1×10 ⁻⁷	5.5×10 ⁻⁵	5.8	4.1	0.68	0.75	0.07	0.10	7,977	6,894☆	7.5×10 ⁻¹¹
CD	5	150,183,562 (CNVR2647.1)	20.1	IRGM	3	1.0×10 ⁻⁷	4.3×10 ⁻⁵	6.1	3.8	0.68	0.76	0.07	0.10	7,977	6,894☆	3.9×10 ⁻¹⁰
CD	6	31,416,574 (CNVR2841.20)	5.1	HLA	3	1.7×10 ⁻⁵	1.1×10 ⁻⁵	3.6	3.9	0.80	0.82	0.19	0.23	NA	NA	NA
T1D	6	32,582,950 (CNVR2845.46)	6.7	HLA	2	8.0×10 ⁻¹⁵³	2.1×10 ⁻¹⁹⁶	125.5	154.4	0.20	0.26	0.14	0.01	NA	NA	NA
RA	6	32,609,209 (CNVR2845.14)	4.0	HLA	4	1.4×10 ⁻³⁹	8.1×10 ⁻⁶⁰	51.5	73.5	1.77	1.83	NA	NA	NA	NA	NA
RA	2→6	179,004,449 (CNVR1065.1)	0.8	HLA	3	6.8×10 ⁻⁴⁹	1.6×10 ⁻⁶⁹	51.0	73.7	1.85	1.94	0.36	0.49	NA	NA	NA
T1D	2→6	179,004,449 (CNVR1065.1)	0.8	HLA	3	1.3×10 ⁻²⁹	1.1×10 ⁻³⁹	28.0	38.4	1.62	1.61	0.36	0.47	NA	NA	NA
RA	NA	NA (AC_000138.1_44)	5.6	HLA	3	8.3×10 ⁻⁴	1.1×10 ⁻⁵	1.3	2.7	0.87	0.86	0.25	0.28	2,743	3,398	1.1×10 ⁻³
T1D	NA	NA (AC_000138.1_44)	5.6	HLA	3	2.0×10 ⁻³¹	2.7×10 ⁻⁴⁵	31.0	45.1	0.59	0.57	0.25	0.36	2,649	3,883	7.3×10 ⁻⁵⁰
CD	17	40,930,407 (CNVR7113.6)	33.9	Chr17inv	3	1.2×10 ⁻³	5.8×10 ⁻⁴	1.4	1.6	1.15	1.14	0.24	0.21	6,069	4,978☆	8.6×10 ⁻⁵
T1D	17	40,930,407 (CNVR7113.6)	33.9	Chr17inv	3	1.6×10 ⁻³	7.5×10 ⁻⁴	1.0	1.2	1.13	1.12	0.24	0.21	9,395	7,911☆	4.6×10 ⁻⁶

Only one of the several associated CNVs mapping to the HLA in the reference sequence is shown for each of rheumatoid arthritis, type 1 diabetes and Crohn's disease. Further details of replication assays and methods are given in Supplementary Information. AC_000138.1_44 is a novel sequence insertion present in the Venter genome sequence but not in the reference sequence and hence no chromosomal location is presented. Minor allele frequency is only estimated for CNVs with three or fewer copy number classes. CD, Crohn's disease; RA, rheumatoid arthritis; T1D, type 1 diabetes; T2D, type 2 diabetes.

*The number of diploid copy-number classes.

†P value from the frequentist association test combining UKBS and 58C as controls.

‡The log₁₀ of the Bayes factor from the Bayesian association analysis combining UKBS and 58C as controls.

§The odds ratio estimated for each additional copy of the CNV based on both UKBS and 58C as controls.

||Extended reference refers to the analogous quantities calculated in comparing cases of the disease in question with UKBS, 58C and aetiologically unrelated cases.

¶The minor allele frequency in controls (UKBS plus 58C).

#The minor allele frequency in cases.

☆Replication sample includes WTCCC samples.

Discussion

We have undertaken a genome-wide association study of common copy-number variation in eight diseases by developing a novel array targeting most of a recently discovered set of CNVs. Our findings inform understanding of the genetic contributions to common disease, offer methodological insights into CNV analysis, and provide a resource for human genetics research.

One major conclusion is that considerable care is needed in analysing copy-number data from array CGH experiments. Choices of normalization, probe summary and probe weighting can make major differences to data quality and utility in association testing. Notably, the optimal choices vary greatly across the CNVs we studied.

A second major conclusion is that CNV association analyses are susceptible to a range of artefacts that can lead to false-positive associations. Some are a consequence of the less-robust nature of the data compared to SNP chips. But others, such as systematic differences depending on DNA source (for example, blood versus cell lines) and dispersed duplications, are more subtle. Several artefacts could survive replication studies. Simultaneously studying eight diseases helped greatly in identifying these artefacts, and stringent quality control was invaluable in eliminating false-positive associations. At least for currently available CNV-typing platforms, we recommend considerable care in interpreting putative CNV associations combined with independent replication on a different experimental platform.

Despite the important technical challenges and potential artefacts discussed above, we have demonstrated that high-confidence CNV calls can be assigned in large, real-world case-control samples for a substantial proportion of the common CNVs estimated to be present in the human genome. We have identified directly several CNV loci that are associated with common disease. Such loci could contribute to disease pathogenesis. However, the loci identified are well tagged by SNPs and, hence, the associations can be, and were, detected indirectly via SNP association studies.

There is a marked difference between the number of confirmed, replicated associations from our CNV study (3 loci) and that from the comparably sized WTCCC1 SNP GWAS of seven diseases and its immediate follow-up (~24 loci). (In assessing the importance of CNVs in disease, it is the absolute number of associations, rather than the proportion among loci tested, that is important.) Following ref. 12 we estimated that our study directly tests approximately half of all autosomal CNVs >500 bp long, with MAF >5%. For such CNVs, our power averages over 80% for effects with odds ratios >1.4, and ~50% for odds ratio = 1.25 (Supplementary Fig. 22). We conclude that at least for the eight diseases studied, and probably more generally, there are unlikely to be many associated CNVs with effects of this magnitude.

Might there be many more common disease-associated CNVs each of small effect, in the way that we now know to be the case with SNP associations for many diseases? The total number of CNVs over 500 bp with MAF >5% is limited (estimated to be under 4,000 (ref. 12)), so unless many of these simultaneously affect many different diseases (something for which we saw no evidence outside of the HLA region) there would seem to be insufficient such CNVs for hundreds to be associated with each of many common diseases. In addition, most common CNVs (MAF >5%) are well tagged by SNPs, and thus amenable to indirect study by SNP GWAS. Examining the large meta-analyses of SNP GWAS for Crohn's disease, type 1 diabetes and type 2 diabetes, there were 95 published associated loci of which only 3, including HLA, had the property that CNVs correlated with the associated SNPs; two of these were detected in our direct study.

We conclude that common CNVs typable on current platforms are unlikely to have a major role in the genetic basis of common diseases, either through particular CNVs having moderate or large effects (odds ratios >1.3, say) or through many such CNVs having small effects. In particular, such common CNVs seem unlikely to account for a substantial proportion of the 'missing heritability' for these diseases. Among the CNVs that we could type well, those not well

tagged by SNPs have the same overall association properties as those which are well tagged. We saw no enrichment of association signals among CNVs involving exonic deletions.

We have argued elsewhere¹⁴ that the concept of 'genome-wide significance' is misguided, and that under frequentist and Bayesian approaches it is not the number of tests performed but rather the prior probability of association at each locus that should determine appropriate *P* value thresholds. Here, to reduce the possibility of missing genuine associations, we deliberately set relaxed thresholds for taking CNVs into replication studies. Having completed these analyses the hypothesis that, a priori, an arbitrary common CNV is much more likely than an arbitrary common SNP to affect disease susceptibility is not supported by our data.

Limitations. Our findings should be interpreted within the context of several limitations. First, despite our successes in robustly testing some of the previously noted challenging CNVs in the genome, for some CNVs we could not reliably assign copy-number classes from our assay. We estimate that just under half of these were not polymorphic in our data, being either false positives in the discovery experiment, or very rare in the UK population. For the remainder, we were also unable to perform reliable association analyses based directly on intensity measurements (that is, without first assigning individuals to copy number classes; data not shown). Such CNVs might plausibly be systematically different from those that we do type successfully, in which case it is not possible to extrapolate from our results to their potential role in human disease. Second, we note that we have not studied CNVs of sequences not present in the reference assembly, high-copy-number repeats such as LINE elements, or most polymorphic tandem repeat arrays, and our findings may not generalize to such variation. Finally, our experiment was powered to detect associations with common copy number variation and our observations and conclusions do not necessarily generalize to the study of rare copy number variants. Different approaches will be necessary to investigate the contribution of such variation to common disease.

METHODS SUMMARY

Pilot study. A total of 384 samples spanning a range of DNA quality were assayed for 156 previously identified CNVs on each of three different platforms: Agilent CGH, NimbleGen CGH and Illumina iSelect. The pilot experiment contained many more probes per CNV than we anticipated using in the main study, and replicates of these probes, to allow an assessment of data quality as a function of the number of probes per CNV and of the merits of replicating probes predicted in advance to perform well, compared to using distinct probes.

Sample selection. Case samples came from previously established UK collections. Control samples came from two sources: half from the 1958 Birth Cohort and half from a UK Blood Service sample. Approximately 80% of samples had been included within the WTCCC SNP GWAS study. The 610 duplicate samples were drawn from all collections.

Array design. The main study used an Agilent CGH array comprising 105,072 long oligonucleotide probes. Probes were selected to target CNVs identified mainly through the GSV discovery experiment¹², with some coming from other sources. Ten non-polymorphic regions of the X chromosome were assayed for control purposes.

Array processing. Arrays were run at Oxford Gene Technology (OGT). The samples were processed in batches of 47 samples drawn from two different collections, with each batch containing one control sample for quality control purposes. These batches were randomized to protect against systematic biases in data characteristics between collections.

Data analysis. Primary data and low-level summary statistics were produced at OGT. All substantive data analyses were undertaken within the consortium. Plates failing quality control metrics were rerun, as were 1,709 of the least well-performing samples. Details of the common CNVs assayed in this study, including any tag SNP, are given at http://www.wtccc.org.uk/wtcccplus_cnv/supplemental.shtml.

Full Methods and any associated references are available in the online version of the paper at www.nature.com/nature.

Received 16 October 2009; accepted 5 March 2010.

- Manolio, T. A. *et al.* Finding the missing heritability of complex diseases. *Nature* **461**, 747–753 (2009).

- Zhang, F., Gu, W., Hurler, M. E. & Lupski, J. R. Copy number variation in human health, disease, and evolution. *Annu. Rev. Genomics Hum. Genet.* **10**, 451–481 (2009).
- Sebat, J. *et al.* Strong association of *de novo* copy number mutations with autism. *Science* **316**, 445–449 (2007).
- Stankiewicz, P. & Beaudet, A. L. Use of array CGH in the evaluation of dysmorphology, malformations, developmental delay, and idiopathic mental retardation. *Curr. Opin. Genet. Dev.* **17**, 182–192 (2007).
- Stefansson, H. *et al.* Large recurrent microdeletions associated with schizophrenia. *Nature* **455**, 232–236 (2008).
- The International Schizophrenia Consortium. Rare chromosomal deletions and duplications increase risk of schizophrenia. *Nature* **455**, 237–241 (2008).
- McCarroll, S. A. *et al.* Deletion polymorphism upstream of IRGM associated with altered IRGM expression and Crohn's disease. *Nature Genet.* **40**, 1107–1112 (2008).
- Willer, C. J. *et al.* Six new loci associated with body mass index highlight a neuronal influence on body weight regulation. *Nature Genet.* **41**, 25–34 (2009).
- de Cid, R. *et al.* Deletion of the late cornified envelope *LCE3B* and *LCE3C* genes as a susceptibility factor for psoriasis. *Nature Genet.* **41**, 211–215 (2009).
- Hollox, E. J. *et al.* Psoriasis is associated with increased β -defensin genomic copy number. *Nature Genet.* **40**, 23–25 (2008).
- Diskin, S. J. *et al.* Copy number variation at 1q21.1 associated with neuroblastoma. *Nature* **459**, 987–991 (2009).
- Conrad, D. F. *et al.* Origins and functional impact of copy number variation in the human genome. *Nature*. doi:10.1038/nature08516 (7 October 2009).
- Murray, C. J. & Lopez, A. D. Evidence-based health policy—lessons from the Global Burden of Disease Study. *Science* **274**, 740–743 (1996).
- The Wellcome Trust Case Control Consortium. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature* **447**, 661–678 (2007).
- McCarroll, S. A. & Altshuler, D. M. Copy-number variation and association studies of human disease. *Nature Genet.* **39**, S37–S42 (2007).
- Locke, D. P. *et al.* Linkage disequilibrium and heritability of copy-number polymorphisms within duplicated regions of the human genome. *Am. J. Hum. Genet.* **79**, 275–290 (2006).
- McCarroll, S. A. *et al.* Common deletion polymorphisms in the human genome. *Nature Genet.* **38**, 86–92 (2006).
- Barnes, C. *et al.* A robust statistical method for case-control association testing with copy number variation. *Nature Genet.* **40**, 1245–1252 (2008).
- Horton, R. *et al.* Variation analysis and gene annotation of eight MHC haplotypes: the MHC Haplotype Project. *Immunogenetics* **60**, 1–18 (2008).
- Zeggini, E. *et al.* Meta-analysis of genome-wide association data and large-scale replication identifies additional susceptibility loci for type 2 diabetes. *Nature Genet.* **40**, 638–645 (2008).
- Stefansson, H. *et al.* A common inversion under selection in Europeans. *Nature Genet.* **37**, 129–137 (2005).
- Fanciulli, M. *et al.* *FCGR3B* copy number variation is associated with susceptibility to systemic, but not organ-specific, autoimmunity. *Nature Genet.* **39**, 721–723 (2007).
- Mamtani, M. *et al.* *CCL3L1* gene-containing segmental duplications and polymorphisms in *CCR5* affect risk of systemic lupus erythematosus. *Ann. Rheum. Dis.* **67**, 1076–1083 (2008).
- McKinney, C. *et al.* Evidence for an influence of chemokine ligand 3-like 1 (*CCL3L1*) gene copy number on susceptibility to rheumatoid arthritis. *Ann. Rheum. Dis.* **67**, 409–413 (2008).
- Wilcoxon, F. Individual comparisons by ranking methods. *Biom. Bull.* **1**, 80–83 (1945).
- Clayton, D. G. *et al.* Population structure, differential bias and genomic control in a large-scale, case-control association study. *Nature Genet.* **37**, 1243–1246 (2005).
- Field, S. F. *et al.* Experimental aspects of copy number variant assays at *CCL3L1*. *Nature Med.* **15**, 1115–1117 (2009).
- Lieber, M. R., Yu, K. & Raghavan, S. C. Roles of nonhomologous DNA end joining, V(D)J recombination, and class switch recombination in chromosomal translocations. *DNA Repair* **5**, 1234–1245 (2006).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements The principal funder of this project was the Wellcome Trust. Many individuals, groups, consortia, organizations and funding bodies have made important contributions to sample collections and coordination of the scientific analyses. Details are provided in Supplementary Information Section 11. We are indebted to all those who participated within the sample collections.

Author Contributions are listed in Supplementary Information.

Author Information Summary information for the CNVs studied, including genomic locations, numbers of classes and SNP tags on different platforms is available at http://www.wtccc.org.uk/wtcccplus_cnv/supplemental.shtml. Full data are available, under a data access mechanism, from the European Genome-phenome Archive (<http://www.ebi.ac.uk/ega/page.php>). Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Correspondence and requests for materials should be addressed to P.D. (peter.donnely@well.ox.ac.uk).

The Wellcome Trust Case Control Consortium

Nick Craddock^{1*}, Matthew E. Hurles^{2*}, Niall Cardin³, Richard D. Pearson⁴, Vincent Plagnol⁵, Samuel Robson², Damjan Vukcevic⁴, Chris Barnes², Donald F. Conrad², Eleni Giannoulatou³, Chris Holmes³, Jonathan L. Marchini³, Kathy Stirrups², Martin D. Tobin⁶, Louise V. Wain⁶, Chris Yau³, Jan Aerts², Tariq Ahmad⁷, T. Daniel Andrews², Hazel Arbury², Anthony Attwood^{2,8,9}, Adam Auton³, Stephen G. Ball¹⁰, Anthony J. Balmforth¹⁰, Jeffrey C. Barrett², Inês Barroso², Anne Barton¹¹, Amanda J. Bennett¹², Sanjeev Bhaskar², Katarzyna Blaszczyk¹³, John Bowes¹¹, Oliver J. Brand¹⁴, Peter S. Braund¹⁵, Francesca Bredin¹⁶, Gerome Breen^{17,18}, Morris J. Brown¹⁹, Ian N. Bruce¹¹, Jaswinder Bull²⁰, Oliver S. Burren⁵, John Burton², Jake Byrnes⁴, Sian Caesar²¹, Chris M. Clee², Alison J. Coffey², John M. C. Connell²², Jason D. Cooper⁵, Anna F. Dominiczak²², Kate Downes⁵, Hazel E. Drummond²³, Darshana Dudakia²⁰, Andrew Dunham², Bernadette Ebbs²⁰, Diana Eccles²⁴, Sarah Edkins², Cathryn Edwards²⁵, Anna Elliot²⁰, Paul Emery²⁶, David M. Evans²⁷, Gareth Evans²⁸, Steve Eyre¹¹, Anne Farmer¹⁸, I. Nicol Ferrier²⁹, Lars Feuk^{30,31}, Tomas Fitzgerald², Edward Flynn¹¹, Alistair Forbes³², Liz Forty¹, Jayne A. Franklyn^{14,33}, Rachel M. Freathy³⁴, Polly Gibbs²⁰, Paul Gilbert¹¹, Omer Gokumen³⁵, Katherine Gordon-Smith^{1,21}, Emma Gray², Elaine Green¹¹, Chris J. Groves¹², Detelina Grozeva¹, Rhian Gwilliam², Anita Hall²⁰, Naomi Hammond², Matt Hardy⁵, Pile Harrison³⁶, Neelam Hassanali¹², Husam Hebaishi², Sarah Hines²⁰, Anne Hinks¹¹, Graham A. Hitman³⁷, Lynne Hocking³⁸, Eleanor Howard², Philip Howard³⁹, Joanna M. M. Howson⁵, Debbie Hughes²⁰, Sarah Hunt², John D. Isaacs⁴⁰, Mahim Jain⁴, Derek P. Jewell⁴¹, Toby Johnson³⁹, Jennifer D. Jolley^{8,9}, Ian R. Jones¹, Lisa A. Jones²¹, George Kirov¹, Cordelia F. Langford², Hana Lango-Allen³⁴, G. Mark Lathrop⁴², James Lee¹⁶, Kate L. Lee³⁹, Charlie Lees²³, Kevin Lewis², Cecilia M. Lindgren^{4,12}, Meeta Maisuria-Armer⁵, Julian Maller⁴, John Mansfield⁴³, Paul Martin¹¹, Dunecan C. O. Massey¹⁶, Wendy L. McArdle⁴⁴, Peter McGuffin¹⁸, Kirsten E. McClay², Alex Mentzer⁴⁵, Michael L. Mimmack², Ann E. Morgan⁴⁶, Andrew P. Morris⁴, Craig Mowat⁴⁷, Simon Myers³, William Newman²⁸, Elaine R. Nimmo²³, Michael C. O'Donovan¹, Abiodun Onipinla³⁹, Ifejinelo Onyiah², Nigel R. Ovington⁵, Michael J. Owen¹, Kimmo Palin², Kirstie Parnell³⁴, David Pernet²⁰, John R. B. Perry³⁴, Anne Phillips⁴⁷, Dalila Pinto³⁰, Natalie J. Prescott¹³, Inga Prokopenko^{4,12}, Michael A. Quail², Suzanne Rafelt¹⁵, Nigel W. Rayner^{4,12}, Richard Redon^{2,48}, David M. Reid³⁸, Anthony Renwick²⁰, Susan M. Ring⁴⁴, Neil Robertson^{4,12}, Ellie Russell¹, David St Clair¹⁷, Jennifer G. Sambrook^{8,9}, Jeremy D. Sanderson⁴⁵, Helen Schuilenburg⁵, Carol E. Scott², Richard Scott²⁰, Sheila Seal²⁰, Sue Shaw-Hawkins³⁹, Beverley M. Shields³⁴, Matthew J. Simmonds¹⁴, Debbie J. Smyth⁵, Eliian Somaskantharajah², Katarina Spanova²⁰, Sophia Steer⁴⁹, Jonathan Stephens^{8,9}, Helen E. Stevens⁵, Millicent A. Stone^{50,51}, Zhan Su³, Deborah P. M. Symmons¹¹, John R. Thompson⁵, Wendy Thomson¹¹, Mary E. Travers¹², Clare Turnbull²⁰, Armand Valsesia², Mark Walker⁵², Neil M. Walker⁵, Chris Wallace⁵, Margaret Warren-Perry²⁰, Nicholas A. Watkins⁵³, John Webster⁵³, Michael N. Weedon³⁴, Anthony G. Wilson⁵⁴, Matthew Woodburn⁵⁵, B. Paul Wordsworth⁵⁵, Allan H. Young^{29,56}, Eleftheria Zeggini^{2,4}, Nigel P. Carter², Timothy M. Frayling³⁴, Charles Lee³⁵, Gil McVean³, Patricia B. Munroe³⁹, Aarno Palotie², Stephen J. Sawcer⁵⁷, Stephen W. Scherer^{30,58}, David P. Strachan⁵⁹, Chris Tyler-Smith², Matthew A. Brown^{55,60}, Paul R. Burton⁶, Mark J. Caulfield³⁹, Alastair Compston⁵⁷, Martin Farrall⁶¹, Stephen C. L. Gough^{14,33}, Alistair S. Hall¹⁰, Andrew T. Hattersley^{34,62}, Adrian V. S. Hill⁴, Christopher G. Mathew¹³, Marcus Pembrey⁶³, Jack Satsangi²², Michael R. Stratton^{2,20}, Jane Worthington¹¹, Panos Deloukas², Audrey Duncanson⁶⁴, Dominic P. Kwiatkowski^{2,4}, Mark I. McCarthy^{4,12,65}, Willem H. Ouwehand^{2,8,9}, Miles Parkes¹⁶, Nazneen Rahman²⁰, John A. Todd⁵, Nilesh J. Samani^{15,66} & Peter Donnelly^{4,3}

*These authors contributed equally to this work.

¹MRC Centre for Neuropsychiatric Genetics and Genomics, School of Medicine, Cardiff University, Heath Park, Cardiff CF14 4XN, UK. ²The Wellcome Trust Sanger Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge CB10 1SA, UK. ³Department of Statistics, University of Oxford, 1 South Parks Road, Oxford OX1 3TG, UK. ⁴The Wellcome Trust Centre for Human Genetics, University of Oxford, Roosevelt Drive, Oxford OX3 7BN, UK. ⁵Juvenile Diabetes Research Foundation/Wellcome Trust Diabetes and Inflammation Laboratory, Department of Medical Genetics, Cambridge Institute for Medical Research, University of Cambridge, Wellcome Trust/MRC Building, Cambridge CB2 0XY, UK. ⁶Departments of Health Sciences and Genetics, University of Leicester, 217 Adrian Building, University Road, Leicester LE1 7RH, UK. ⁷Genetics of Complex Traits, Peninsula College of Medicine and Dentistry University of Exeter, Exeter EX1 2LU, UK. ⁸Department of Haematology, University of Cambridge, Long Road, Cambridge CB2 0PT, UK. ⁹National Health Service Blood and Transplant, Cambridge Centre, Long Road, Cambridge CB2 0PT, UK. ¹⁰Multidisciplinary Cardiovascular Research Centre (MCRC), Leeds Institute of Genetics, Health and Therapeutics (LIGHT), University of Leeds, Leeds LS2 9JT, UK. ¹¹arc Epidemiology Unit, Stopford Building, University of Manchester, Oxford Road, Manchester M13 9PT, UK. ¹²Oxford Centre for Diabetes, Endocrinology and Medicine, University of Oxford, Churchill Hospital, Oxford OX3 7LJ, UK. ¹³Department of Medical and Molecular Genetics, King's College London School of Medicine, 8th Floor Guy's Tower, Guy's Hospital, London SE1 9RT, UK. ¹⁴Centre for Endocrinology, Diabetes and Metabolism, Institute of Biomedical Research,

University of Birmingham, Birmingham B15 2TT, UK. ¹⁵Department of Cardiovascular Sciences, University of Leicester, Glenfield Hospital, Groby Road, Leicester LE3 9QP, UK. ¹⁶IBD Genetics Research Group, Addenbrooke's Hospital, Cambridge CB2 0QQ, UK. ¹⁷University of Aberdeen, Institute of Medical Sciences, Foresterhill, Aberdeen AB25 2ZD, UK. ¹⁸SGDP, The Institute of Psychiatry, King's College London, De Crespigny Park, Denmark Hill, London SE5 8AF, UK. ¹⁹Clinical Pharmacology Unit, University of Cambridge, Addenbrooke's Hospital, Hills Road, Cambridge CB2 2QQ, UK. ²⁰Section of Cancer Genetics, Institute of Cancer Research, 15 Cotswold Road, Sutton SM2 5NG, UK. ²¹Department of Psychiatry, University of Birmingham, National Centre for Mental Health, 25 Vincent Drive, Birmingham B15 2FG, UK. ²²BHF Glasgow Cardiovascular Research Centre, University of Glasgow, 126 University Place, Glasgow G12 8TA, UK. ²³Gastrointestinal Unit, Division of Medical Sciences, School of Molecular and Clinical Medicine, University of Edinburgh, Western General Hospital, Edinburgh EH4 2XU, UK. ²⁴Academic Unit of Genetic Medicine, University of Southampton, Southampton SO16 5YA, UK. ²⁵Endoscopy Regional Training Unit, Torbay Hospital, Torbay TQ2 7AA, UK. ²⁶Academic Unit of Musculoskeletal Disease, University of Leeds, Chapel Allerton Hospital, Leeds, West Yorkshire LS7 4SA, UK. ²⁷MRC Centre for Causal Analyses in Translational Epidemiology, Department of Social Medicine, University of Bristol, Bristol BS8 2BN, UK. ²⁸Department of Medical Genetics, Manchester Academic Health Science Centre (MAHSC), University of Manchester, Manchester M13 0JH, UK. ²⁹School of Neurology, Neurobiology and Psychiatry, Royal Victoria Infirmary, Queen Victoria Road, Newcastle upon Tyne NE1 4LP, UK. ³⁰The Centre for Applied Genomics and Program in Genetics and Genomic Biology, The Hospital for Sick Children, MaRS Centre—East Tower, 101 College St, Room 14-701, Toronto, Ontario M5G 1L7, Canada. ³¹Department of Genetics and Pathology, Rudbeck Laboratory, Uppsala University, Uppsala 75185, Sweden. ³²Institute for Digestive Diseases, University College London Hospitals Trust, London NW1 2BU, UK. ³³University Hospital Birmingham NHS Foundation Trust, Birmingham B15 2TT, UK. ³⁴Genetics of Complex Traits, Peninsula College of Medicine and Dentistry, University of Exeter, Magdalen Road, Exeter EX1 2LU, UK. ³⁵Department of Pathology, Brigham and Women's Hospital and Harvard Medical School, Boston, Massachusetts 02115, USA. ³⁶University of Oxford, Institute of Musculoskeletal Sciences, Botnar Research Centre, Oxford OX3 7LD, UK. ³⁷Centre for Diabetes and Metabolic Medicine, Barts and The London, Royal London Hospital, Whitechapel, London E1 1BB, UK. ³⁸Bone Research Group, Department of Medicine and Therapeutics, University of Aberdeen, Aberdeen AB25 2ZD, UK. ³⁹Clinical Pharmacology and Barts and The London Genome Centre, William Harvey Research Institute, Barts and The London School of Medicine and Dentistry, Queen Mary University of London, Charterhouse Square, London EC1M 6BQ, UK. ⁴⁰Institute of Cellular Medicine, Musculoskeletal Research Group, 4th Floor, Catherine Cookson Building, The Medical School, Framlington Place, Newcastle upon Tyne NE2 4HH, UK. ⁴¹Gastroenterology Unit, Radcliffe Infirmary, University of Oxford, Oxford OX2 6HE, UK. ⁴²Centre National de Genotypage, 2 Rue Gaston Cremieux, Evry, Paris 91057, France. ⁴³Department of Gastroenterology and Hepatology, University of Newcastle upon Tyne, Royal Victoria Infirmary, Newcastle upon Tyne NE1 4LP, UK. ⁴⁴ALSPAC Laboratory, Department of Social Medicine, University of Bristol, Bristol BS8 2BN, UK. ⁴⁵Division of Nutritional Sciences, King's College London School of Biomedical and Health Sciences, London SE1 9NH, UK. ⁴⁶NIHR-Leeds Musculoskeletal Biomedical Research Unit, University of Leeds, Chapel Allerton Hospital, Leeds, West Yorkshire LS7 4SA, UK. ⁴⁷Department of General Internal Medicine, Ninewells Hospital and Medical School, Ninewells Avenue, Dundee DD1 9SY, UK. ⁴⁸INSERM UMR915, L'Institut du Thorax, Nantes 44035, France. ⁴⁹Clinical and Academic Rheumatology, Kings College Hospital National Health Service Foundation Trust, Denmark Hill, London SE5 9RS, UK. ⁵⁰University of Toronto, St Michael's Hospital, 30 Bond Street, Toronto, Ontario M5B 1W8, Canada. ⁵¹University of Bath, Claverton, Norwood House, Room 5.11a, Bath, Somerset BA2 7AY, UK. ⁵²Diabetes Research Group, School of Clinical Medical Sciences, Newcastle University, Framlington Place, Newcastle upon Tyne NE2 4HH, UK. ⁵³Medicine and Therapeutics, Aberdeen Royal Infirmary, Foresterhill, Aberdeen, Grampian AB9 2ZB, UK. ⁵⁴School of Medicine and Biomedical Sciences, University of Sheffield, Sheffield S10 2JF, UK. ⁵⁵Nuffield Department of Orthopaedics, Rheumatology and Musculoskeletal Sciences, Nuffield Orthopaedic Centre, University of Oxford, Windmill Road, Headington, Oxford OX3 7LD, UK. ⁵⁶UBC Institute of Mental Health, 430-5950 University Boulevard Vancouver, British Columbia V6T 1Z3, Canada. ⁵⁷Department of Clinical Neurosciences, University of Cambridge, Addenbrooke's Hospital, Hills Road, Cambridge CB2 2QQ, UK. ⁵⁸Department of Molecular Genetics, University of Toronto, Toronto, Ontario M5S 1A8, Canada. ⁵⁹Division of Community Health Sciences, St George's, University of London, London SW17 0RE, UK. ⁶⁰Diamantina Institute of Cancer, Immunology and Metabolic Medicine, Princess Alexandra Hospital, University of Queensland, Ipswich Road, Woolloongabba, Brisbane, Queensland 4102, Australia. ⁶¹Cardiovascular Medicine, University of Oxford, Wellcome Trust Centre for Human Genetics, Roosevelt Drive, Oxford OX3 7BN, UK. ⁶²Genetics of Diabetes, Peninsula College of Medicine and Dentistry, University of Exeter, Barrack Road, Exeter EX2 5DW, UK. ⁶³Clinical and Molecular Genetics Unit, Institute of Child Health, University College London, 30 Guilford Street, London WC1N 1EH, UK. ⁶⁴The Wellcome Trust, Gibbs Building, 215 Euston Road, London NW1 2BE, UK. ⁶⁵Oxford NIHR Biomedical Research Centre, Churchill Hospital, Oxford OX3 7LJ, UK. ⁶⁶Leicester NIHR Biomedical Research Unit in Cardiovascular Disease, Glenfield Hospital, Leicester LE3 9QP, UK.

METHODS

Pilot experiment. Full details of Methods are given in the Supplementary Information, but in brief a total of 384 samples from four different collections spanning the range of DNA quality encountered in our previous WTCCC SNP-based association study¹⁴ were assayed for 156 previously identified CNVs on each of three different platforms: Agilent Comparative Genomic Hybridization (CGH), and NimbleGen CGH (run in service laboratories) and Illumina iSelect (run at the Sanger Institute). The pilot experiment contained many more probes per CNV (40–90 depending on platform) than we anticipated using in the main study, and replicates of these probes, to allow an assessment of data quality as a function of the number of probes per CNV and of the merits of replicating probes predicted in advance to perform well, compared to using distinct probes.

The Agilent CGH platform performed best in our pilot and we settled on an array that comprised 105,072 long oligonucleotide probes. On the basis of the pilot data we aimed to target each CNV with 10 distinct probes. Actual numbers of probes per CNV on the array varied from this for several reasons (see Supplementary Information and Supplementary Fig. 9), and we included in our analyses any CNV with at least one probe on the array.

Array content, assay and samples for the main experiment. Array content: the GSV discovery experiment¹² involved 20 HapMap Utah residents with European ancestry (CEU) and 20 HapMap Yoruban (YRI) individuals, and 1 Polymorphism Discovery Resource individual, assayed via 20 NimbleGen arrays containing a total of 42,000,000 probes tiled across the assayable portion of the human reference genome. We prioritized CNVs for our experiment based on their frequency in the discovery sample, with those identified in CEU individuals given precedence. A total of 10,835 out of 11,700 CNVs were included from the list provided by the GSV, with those not included on the array being present in discovery in only 1 YRI individual and not overlapping genes or highly conserved elements. This list was augmented by any common CNVs not present among the GSV list found from analyses of Affymetrix SNP 6.0 data in HapMap 2 samples (83 CNVs), Illumina 1M data in HapMap 3 samples (82 CNVs), analyses of Affymetrix 500K samples (18 CNVs)^{7,29,30}, and from our own analyses of WTCCC1 SNP data (231 CNVs). In addition, we sought to identify CNVs not present in the human reference sequence through analyses of published^{31,32} novel sequence insertions (292 CNVs in total). Thus in total, our array targeted 11,541 putative CNVs. Ten non-polymorphic regions of the X chromosome were also assayed for control purposes.

Most loci targeted on the CNV-typing array derive from microarray-based CNV discovery, which is inherently imprecise. In contrast to SNP discovery by sequencing, arrays do not provide nucleotide-level resolution, nor do they locate additional copies of a sequence in the genome. As a result, when CNVs called in different individuals overlap, but are not identical, these could be called as one or two different CNVs, and where discovered CNVs involve probes which map to multiple places in the reference genome, they might be called as CNVs in each of these locations. Interpretation of counts of CNVs from discovery experiments is thus not straightforward. Data on CNVs across thousands of individuals provide added power to refine CNV definitions and derive a non-redundant set of CNVs. In addition, our CNV-typing array draws together CNVs from different sources, and additional redundancy between these, although minimized during array design, can be identified and removed. Analyses of the final array design revealed 434 of the 11,541 CNVs to be redundant because they were targeted by exactly the same probes as other CNVs on the array, and analysis of our array data revealed a further 213 of 562 CNVs to be redundant from instances where overlapping CNVs passing quality control were called as distinct in discovery yet had effectively identical copy-number calls. See Supplementary Information Section 3.1 for further details on array content.

Assay: arrays were run at Oxford Gene Technology (OGT), with each plate containing one control sample for quality control purposes. Primary data and low-level summary statistics were produced at OGT. All substantive data analyses were undertaken within the consortium. Plates that failed pre-specified quality control metrics were rerun on the array, and in addition we repeated 1,709 of the least well-performing samples, chosen according to our own quality control analyses. (See Supplementary Information for further details.)

Samples: the WTCCC CNV study analysed cases from eight common diseases (breast cancer, bipolar disorder, coronary artery disease, Crohn's disease, hypertension, rheumatoid arthritis, type I diabetes, and type 2 diabetes) and two control cohorts (1958 Birth Cohort (58C) and the UK Blood Service collection (UKBS)). The number of subjects from each cohort that were analysed and the numbers that passed each phase of the quality control procedures within this study are shown in Supplementary Table 7. For bipolar disorder, coronary heart disease, Crohn's disease, hypertension, rheumatoid arthritis, type 1 diabetes, type 2 diabetes and the two control cohorts, a large proportion of the subjects studied in this experiment were the same as those in the WTCCC1 SNP GWAS (Supplementary

Table 2). Where sufficient DNA was not available for the original WTCCC1 individuals, additional new samples from the same cohorts were used, selected using the same approaches used for the WTCCC1 samples. Any samples that failed any of the relevant quality control metrics in WTCCC1 were excluded from consideration for this experiment. The breast cancer cohort was not included in the WTCCC1 SNP GWAS.

Data pre-processing, CNV calling and quality control. Data pre-processing: for each of the targeted loci, the subset of probes that target the locus of interest (at least 1-bp overlap) while also targeting the least number of additional CNVs was selected for assaying (see Supplementary Information Section 4.2 for more details). A total of 16 different analysis 'pipelines' were used to create one-dimensional intensity summaries for each CNV. First, a range of different methods were used to create single intensity measurements for each probe from the red channel (test DNA) and green channel (reference DNA) intensity data. This included different methods for normalization of the signals (see Supplementary Information Section 4.3 for details). Second, some pipelines incorporated a new method called probe variance scaling (PVS) that weights probes based on information derived from duplicate samples (see Supplementary Information Section 4.5 for details). Third, some pipelines used the first principal component of the normalized probe intensities to summarize the probe-level data to CNV-level data, whereas other pipelines used the mean of the probe intensities. Finally, some pipelines additionally used a linear discriminant function (LDF) to refine further the summaries based on information from an initial round of genotype calling (see Supplementary Information Section 4.4 for details).

CNV calling: algorithmic details of the two calling methods used (CNVtools and CNVCALL) are provided in Supplementary Information Section 6. Each method was applied separately to the intensity summaries created from each of the 16 pre-processing pipelines for each CNV locus.

Quality control: samples were excluded on the basis of sample handling errors, evidence of non-European ancestry, evidence of sample mixing, manufacturer's recommendations on data quality, outlying values of various pre-calling and post-calling quality metrics, and identity or close relatedness to other samples (see Supplementary Information Section 5.1 for further details). To choose which pipeline to use for a given CNV we used the pipeline that gave the highest number of classes and the highest average posterior probability in cases where more than one pipeline gave the same maximum number of classes. CNVs were excluded that had identical probe sets to other CNVs, that were called with one class in all pre-processing pipelines, that had low average posterior calls in all pre-processing pipelines, or that had a high calls correlation with an overlapping CNV (see Supplementary Information Section 5.2 for further details).

Properties of CNVs. Single class CNVs: Supplementary Table 15 shows the proportion of the single-class CNVs from the GSV discovery project broken down according to the number of individuals and population(s) in which they were discovered. Of the GSV CNVs discovered in CEU, 52% are single class in our data, whereas a higher proportion (74%) of GSV CNVs discovered exclusively in YRI are single class, as would be expected. CNVs at which distinct copy number classes cannot be distinguished might result because: (1) although polymorphic, the signal to noise ratio at that CNV does not allow reliable identification of distinct copy-number classes; (2) the copy-number variant has an extremely low population frequency; or (3) the CNV was a false positive in a discovery experiment and is in fact monomorphic. In a genuinely polymorphic CNV, the intensity measurements within a pair of duplicates should be more similar than between a random pair of distinct individuals. Intensity comparisons between duplicates and random pairs of individuals thus allow estimates of the proportion of single-class CNVs which are not copy-number variable in our data (see Supplementary Information). These estimates range from ~23% for single-class CNVs discovered in two or more CEU individuals to ~50% of single-class CNVs discovered exclusively in YRI (see Supplementary Information for details). We estimate that 18% of GSV CNVs discovered in CEU do not exhibit polymorphism in our UK sample. This figure is similar to the GSV estimate for false positives in discovery of 15%¹². Overall, considering CNVs on the array from all sources, we estimate that 26% do not exhibit polymorphism, so that just under half of the CNVs that seem in our data to have a single class are likely not to be polymorphic. Not all of these will be false positives in discovery; some represent CNVs that are either unique to the individual in which they were discovered or are extremely rare in the UK population.

Multi-class CNVs: a companion study¹² estimated that 83% of the bi-allelic CNVs it genotyped represent deletions, with the remainder being duplications. Supplementary Table 7 compares the number of copy-number classes estimated by the two calling algorithms used in the analyses for each of the CNVs passing quality control. Most differences in numbers of called classes between the algorithms arise from CNVs where one class is very rare and is handled differently by the algorithms (for example, called as a separate class in one algorithm but classed as outlier samples or merged with a larger class by the other).

These 3,432 CNVs include 80% of the CNVs genotyped on the Affymetrix 6.0 array that are common (MAF >5%) in a population with European ancestry³³; conversely only 15% of the common CNVs we called could be called using the Affymetrix 6.0 array.

Allele frequencies: we calculated minor allele frequencies (MAFs) for 2- and 3-class CNVs by assuming that these CNVs were biallelic and using the expected posterior genotype counts (see Supplementary Information Section 7.3 for further details).

Tagging by SNPs: to determine how well tagged the CNVs analysed in our experiment were by SNPs, we carried out correlation analyses using control samples that were common to the current studies and other WTCCC studies. We analysed three different collections of SNPs. We used imputed HapMap2 SNP calls in the WTCCC1 study that used the Affymetrix 500k array, and actual calls from the WTCCC2 study using both the Affymetrix 6.0 array and a custom Illumina 1.2M array. In all cases we used samples from the UKBS collection (see Supplementary Information Section 7.1 for further details).

Geographical variation: geographical information, at the level of 13 pre-defined regions of the UK, was available for 82% of the samples in our study and we undertook analyses for differences in copy-number class frequencies between regions. The results, shown in Supplementary Fig. 24, confirm that there is no major genome-wide population structure, but that, unsurprisingly, there is differentiation at CNVs within HLA. It does not seem easy to determine whether other regions with low *P* values in this test represent genuine departures from the null hypothesis of no differentiation, rather than chance effects, although we note that the third most regionally differentiated CNV outside the HLA (CNVR7722.1, $P = 3 \times 10^{-5}$, 12-d.f.) is a deletion located within the gene *LILRA3*, which may act as soluble receptor for class I MHC antigens, and so would be consistent with the observed HLA stratification. This deletion is also the subject of a reported disease association³⁴ in multiple sclerosis, a finding that may require some caution given the level of geographical stratification at this CNV in our data. (See Supplementary Information Section 9.1 for further details.)

Association testing. Diagnostic plots such as quantile–quantile and cluster plots were created using R. Cluster plots were visually inspected for all CNVs with putative associations.

Principal component analysis (PCA) was applied to the summarized intensity levels for all CNVs, and for all samples that passed quality control. Plots of the first ten principal components were coloured by various sample parameters and these revealed some of the artefacts described in Box 1.

Where possible, replication was carried out by using data from other studies for SNPs that tag the CNVs of interest. Where there was no SNP tag available, breakpoint or direct quantitative CNV assays were designed (see Supplementary Information Section 9 for further details).

We used a two-sided Wilcoxon signed-rank test to test for differences between distributions of Bayes factors between different subsets of CNVs (those that delete all or part of an exon versus those that do not, and CNVs that are well-tagged by SNPs versus those that are not well tagged). (See Supplementary Information Section 9.5 for further details.)

Testing for population stratification: all our samples are from within the UK, and we have excluded any for which the genetic data suggest evidence of non-European ancestry. All collections in this study, apart from breast cancer, were involved in the WTCCC SNP GWAS, and across these collections 80% of samples coincided between the two studies. Analysis of the WTCCC SNP data¹⁴ established that population structure was not a major factor confounding association testing. Similar analyses using SNP data available for the breast cancer samples yielded similar results (data not shown). These SNP results reinforce the evidence from the quantile–quantile plots in Supplementary Fig. 23 and our geographical analyses of the CNV data.

Expanded reference group analysis: in addition to our primary case-control analyses, following ref. 14 we also undertook expanded reference group analyses, in which copy-number class frequencies in cases for a particular disease are compared with those for controls and the other diseases with no aetiological or known genetic connection (see Supplementary Table 10 for details).

Other analyses. We used information on variability between duplicate samples to determine whether CNVs called with one class show signals of polymorphism (details are given in Supplementary Information Section 9.2).

We used estimates of the number of common autosomal CNVs segregating in a population of European ancestry from ref. 12 to estimate the coverage of common autosomal CNVs in our study (see Supplementary Information Section 9.3 for further details).

We designed a series of PCR primers to analyse further the complex signals associated with CNVR5583.1 found in the *TSPAN8* region. (See Supplementary Information Section 9.4 for further details.)

29. International HapMap Project. (<http://hapmap.ncbi.nlm.nih.gov/>) (2010).
30. Redon, R. *et al.* Global variation in copy number in the human genome. *Nature* **444**, 444–454 (2006).
31. Levy, S. *et al.* The diploid genome sequence of an individual human. *PLoS Biol.* **5**, e254 (2007).
32. Kidd, J. M. *et al.* Mapping and sequencing of structural variation from eight human genomes. *Nature* **453**, 56–64 (2008).
33. McCarroll, S. A. *et al.* Integrated detection and population-genetic analysis of SNPs and copy number variation. *Nature Genet.* **40**, 1166–1174 (2008).
34. Koch, S. *et al.* Association of multiple sclerosis with ILT6 deficiency. *Genes Immun.* **6**, 445–447 (2005).