The Role of Genetic Variation in Type 2 Diabetes Predisposition

A thesis submitted to the University of Oxford



In candidature for the degree of Doctor of Philosophy

by

Neelam Hassanali

Green Templeton College, University of Oxford

and

Oxford Centre for Diabetes, Endocrinology, and Metabolism

Radcliffe Department of Medicine, University of Oxford

Trinity term, 2015

Memorandum

The work in this thesis is the original work of the author. All experiments were performed at the Oxford Centre for Diabetes, Endocrinology & Metabolism (OCDEM) under the supervision of Professor Mark McCarthy and Professor Anna Gloyn. Funding was provided by Diabetes UK.

The evaluation of type 2 diabetes risk variants in a South Asian population of Sri Lankan descent (Chapter 3) was completed in collaboration with the University of Exeter and has been published in PLOS ONE. The manuscript (Hassanali et al. 2014) is provided as an appendix. Genetic risk scores were calculated by Dr Maneka G De Silva using genotyping data generated as part of this thesis.

The replication genotyping performed in Sri Lankan samples (Chapter 4) contributed to two large-scale genome-wide association studies. For the first study, coordination and meta-analysis of the data was completed by lead collaborators Professor Jaspal S Kooner and Dr John C Chambers. In the second study, work was led by Dr Richa Saxena and Dr Dharambir K. Sanghera. The results detailed in Chapter 4 form part of the published manuscripts Kooner et al. 2011 and Saxena et al. 2013, both of which are provided in the appendix.

The *HNF1A* capillary sequencing performed in Chapter 5 contributes to the published manuscript Thanabalasingham et al. 2013, which is provided in the appendix section.

The *HNF1A* variants functionally characterised in Chapter 6 were identified as part of a study completed by Dr Saima Atizaz, which evaluated the performance of hsCRP as a biomarker in South Asian individuals.

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 48,500 words.

The Role of Genetic Variation in Type 2 diabetes Predisposition

Neelam Hassanali, Green Templeton College

Submitted in candidature for the degree of Doctor of Philosophy, Trinity term 2015

Genome wide association studies (GWAS) and next generation sequencing technologies (NGS) are proving to be transformative in facilitating our understanding of the genetic component of type 2 diabetes (T2D). These methodologies have identified common, low frequency and rare genetic variants that are associated with T2D predisposition and can provide valuable insight into disease pathophysiology. The work presented in this thesis uses a combination of approaches to understand how common and rare variants influence disease susceptibility.

The initial waves of studies seeking common variant associations with T2D focused on individuals of European ancestry. Evaluating their results in other populations may offer insight into ethnic differences in disease prevalence and presentation as well as identify novel mechanisms for disease susceptibility. My work aimed to determine whether a subset of common variants conferring T2D-risk in Europeans contribute similarly to T2D susceptibility in South Asians, who have a substantially higher occurrence of T2D.

My analysis identified no systematic difference in allelic odds ratios or allele frequencies between samples of Sri Lankan and European origin. This suggests that systematic differences in common variant associations are unlikely to explain the interethnic differences in prevalence and presentation of T2D. As part of a large-scale collaborative GWAS in individuals of South Asian ancestry, my work contributed to the identification seven loci revealing novel mechanisms, which may contribute to diabetes risk.

Genome-wide association study results have also indicated several biomarker profiles (glycan profiles and hsCRP) that might be useful diagnostic tools for discriminating between patients with Maturity-Onset Diabetes of the Young (MODY) caused by mutations in *HNF1A* (HNF1A-MODY) and common forms of diabetes. Correctly, diagnosing HNF1A-MODY has important implications for patients' treatment and disease prognosis, and for their family members. However, cases of HNF1A-MODY are currently often incorrectly diagnosed. Biomarkers may therefore offer a useful way to prioritise patients for genetic screening.

To examine whether glycan profiles and hsCRP could serve as surrogate measures for HNF1A dysfunction, I completed capillary sequencing of *HNF1A* in patients with biomarker profiles suggestive of HNF1A dysfunction. Using an evaluation pipeline comprising family studies, *in silico* evaluation and *in vitro* functional characterisation, I demonstrated that variants within *HNF1A* are associated with a spectrum of phenotypes varying from casual for MODY, increasing T2D risk or neutral.

My work exemplifies the considerable challenges in determining the clinical significance of coding variants identified through NGS on disease risk. The data shows that biomarker profiles are a promising tool for assessing HNF1A dysfunction. However, it is likely that clinical interpretation of variants within *HNF1A* will need to be performed on an individual basis using a combination of factors that includes clinical, biochemical and lifestyle factors.

Acknowledgements

My DPhil experience has not only involved navigating my way through a world filled with complicated science words and experiments, but has also involved me overcoming some personal challenges. This has provided me with the strength and determination that allows me to write this page in my thesis today.

My first thank you goes to my supervisors Professor Mark McCarthy and Professor Anna Gloyn for the opportunity, resources, and guidance to complete my research. Their decision to welcome me into the Diabetes Research Laboratories at OCDEM as a research assistant all those years ago has been a life-changing experience and I will always be thankful that they 'picked me'. A heartfelt thank you goes to Professor Keith Frayn. Your unwavering support and kindness has been a source of comfort through some of the darkest DPhil days and I am incredibly grateful for the fortuitous assignment of you as my college advisor. To all the staff at Green Templeton College, the friendly hellos and random conversations have made GTC a home for me over the last few years -thank you.

My time in the Oxford bubble has been filled with many new adventures and memories, all of which will stay with me forever. Being part of OCDEM and GTC has made my Oxford experience very special and I have been lucky enough to make some wonderful lifelong friends.

A big thank you goes to my lab family: Chris, Mandy, and Carla, your patience whilst fielding my numerous questions over the years and the regular supply of hugs has made the lab a very happy place to be. Amy, you are one in a million and I have had some of my best days at OCDEM working with you. Jana and Anne, thank you for sharing your knowledge and guiding me through the murky world of molecular biology. To my fellow students past and present, Matt R, Natasha and Søren: I could not have asked for a better group of people to share the lab bench with.

To all my friends outside the lab, the past four years has been filled with early morning rows, blades, GTC-GCR committee meetings, squash tournaments, lunches, dinners, trips away and so much more. Thank you for taking me in and I look forward to many more happy times with you all.

Laura, Coops, Paul and Brent, you've introduced me to so many new things (you know what they are) - including what it means to be part of a group of friends, so thank you. Nicola, Susan, and Shauna, thank you for all the wise words and continuing faith that completing this DPhil might actually be possible (especially through the writing process). Martijn, thank you for sharing your office, and introducing me to the world of classical music. You have been a great source of wisdom and incredibly welcoming whenever I have visited your happy home. Mary, my first Oxford friend: thank you for writing to me before I even moved to Oxford. Since then we have never looked back. From bike riding to living together, we have shared so much. You became my person through this DPhil process especially towards the end. I am incredibly lucky and grateful to have you in my life, and I look forward to sharing many more happy adventures with you as we both start new chapters in our lives.

To my family the 'Watford Wallahs', thank you for all your support and putting up with my absence whilst I completed my DPhil. I look forward to family gatherings where I can spend time with you all and make new memories whilst watching all the little ones grow up. Khalil and Nedah, thank you for all the encouragement and support over the past four years. I look forward to spending more time with the both of you. To my parents, thank you for the food packages and for giving me the support to follow my own path.

Thank you.

List of common abbreviations

λs	Sibling relative risk
ADP	Adenosine diphosphate
ATP	Adenosine triphosphate
BMI	Body mass index
bp	Base pair
cDNA	Complementary deoxyribonucleic acid
Chr	Chromosome
ddNTPs	Dideoxyribonucleotide
dH_20	Distilled water
DIAGRAM+	Diabetes genetics replication and meta-analysis consortium
DNA	Deoxyribonucleic acid
dNTPs	Deoxyribonucleotide
eQTL	Expression quantitative trait loci
ESP	Exome Sequencing Project
ExAc	Exome Aggregation Consortium
fpg	Fasting plasma glucose
gDNA	genomic DNA
GAD antibodies	Glutamic acid decarboxylase antibodies
GCK	Glucokinase
GSIS	Glucose stimulated insulin secretion
GWA(S)	Genome-wide association (study)
hsCRP	Highly sensitive C-reactive protein
HNF1A	Hepatocyte Nuclear Factor- 1 alpha
HWE	Hardy Weinberg Equilibrium
Kb	Kilobase
LD	Linkage disequilibrium
MAF	Minor allele frequency
MAGIC	Meta-Analysis of glucose and insulin related traits consortium
miRNA	Micro ribonucleic acid
MODY	Maturity Onset Diabetes of the Young
mRNA	Messenger RNA
NCBI	National centre for Biotechnology Information

NEFAs	Non-esterified fatty acids
NTC	Non-template control
OGTT	Oral glucose tolerance test
OR	Odds Ratio
PBS	Phosphate buffered saline
PCR	Polymerase chain reaction
PNDM	Permanent neonatal diabetes
QC	Quality control
RAF	Risk allele frequency
RNA	Ribonucleic acid
SD	Standard deviation
SDM	Site-directed mutagenesis
SNP(s)	Single nucleotide polymorphism(s)
T1D	Type 1 diabetes
T2D	Type 2 diabetes
TBE	Tris/Borate/EDTA
TBST	Tris buffered saline tween
TNDM	Transient neonatal diabetes mellitus
TNFα	Tumour necrosis factor-α
UCSC	University of California Santa Cruz
WHO	World Health Organisation
WT	Wild-type

Table of Contents

Chapter 1 1

1.1 The global impact of type 2 diabetes 2

1.2 Pathophysiology of type 2 diabetes 3

1.2.1 Diagnosis and clinical presentation 3

1.2.2 Causes of type 2 diabetes 4

1.3 Type 2 diabetes and genetics 7

1.3.1 The role of genetics in type 2 diabetes 7

1.3.2 Methodologies for variant discovery 8

1.3.3 What have we learnt so far and the challenges ahead 12

1.4 Type 2 diabetes in the South Asian population 13

1.4.1 Replication of discovery variants in non-European populations 13

1.4.2 Prevalence and presentation 14

1.4.3 Factors influencing differences in T2D incidence 15

1.5 Monogenic diabetes 17

1.5.1 Different subtypes of Maturity Onset Diabetes of the Young (MODY) 18

1.5.2 Diagnostic and clinical presentation of MODY 19

1.6 The hepatocyte nuclear factor-1-alpha (*HNF1A*) gene 19

1.6.1 HNF1A and animal models 21

1.6.2 HNF1A and the pancreatic transcription factor network 23

1.6.3 Rare and common variants in HNF1A are associated with increased risk for diabetes 24

1.6.4 Challenges in attributing function to variants 25

1.6.5 The role of biomarkers in identifying HNF1A-MODY 26

1.7 Aims of thesis 29

Chapter 2 30

2.1 Subjects studied 31

2.1.1 Sri Lankan type 2 diabetes (T2D) case-control samples 31
2.2 DNA extraction from whole blood 32
2.3 Genomic DNA amplification 32
2.3.1 Primer design 32
2.3.2 Polymerase chain reaction (PCR) 32
2.3.3 PCR product visualisation using agarose gel electrophoresis 33
2.4 Capillary DNA sequencing 33
2.4.1 Clean-up of PCR products 35
2.4.2 Sequencing reactions 35
2.4.3 Sequencing product clean-up 35
2.4.4 Capillary electrophoresis and analysis 37
2.5 Genotyping (Applied Biosystems TaqMan 7900 HT technologies) 37
2.5.1 Assay design 37
2.5.2 TaqMan PCR 37
2.5.3 Genotype determination 38
2.6 Site-directed mutagenesis 38
2.7 Transformation of competent *Escherichia coli* (*E. coli*) 38
2.8 Preparation of agar-antibiotic culture plates 40
2.9 DNA extraction from transformed *Escherichia coli* 40
2.10 DNA quantification 42
2.11 Restriction digest of DNA 42
2.12 Plasmid DNA sequencing 42
2.13 General maintenance of cell lines and seeding cells for experiments 43
2.14 Transient transfections 45
2.15 Transactivation studies 45
2.16 Protein expression studies 47
2.16.1 Transactivation assay lysate western blot analysis 47
2.16.2 Preparation of cytosolic and nuclear protein extracts 47

2.16.3 Protein quantification 48
2.16.4 Western blot analysis 49
2.17 Subcellular localisation-Immunofluorescence 50
2.18 *In silico* prediction tools 51
2.19 Statistical Analysis 51
Chapter 3 52
3.1 Introduction 53
3.1.1 Investigating type 2 diabetes (T2D) in South Asians 53
3.1.2 Diabetes in the Sri Lankan population 54
3.1.3 Experimental aims 54
3.2 Methodology 55
3.2.1 Study samples 55
3.2.2 DNA Extraction of Sri Lankan control samples 55
3.2.3 SNP selection and Genotyping 55
3.2.4 Statistical analysis 58
3.2.5 Genetic risk score 58
3.2.6 Collation of reported odds ratio and allelic frequency from previous studies in European populations 61
3.2.7 Power calculations 61
3.3 Results 61
3.3.1 Genotyping of T2D risk loci 61
3.3.2 Individual associations for the 36 T2D risk loci 61
3.3.3 Collation of reported odds ratio and allele frequency and power calculations 62
3.3.4 Comparison of odds ratios and risk allele frequency between Sri Lankan and European samples 63
3.3.5 Genetic risk score results 63
3.4 Discussion 71
Chapter 4 73

4.1 Introduction 74

4.1.1 Study A: Genome-wide association study in individuals of South Asian ancestry 75

4.1.2 Study B: Genome-wide association study in Sikhs of Punjabi origin from India 76

4.1.3 Investigating South Asian subpopulations could identify ethnic-specific variants 78

4.2 Methodology 80

4.2.1 Study samples 80

4.2.2 Study A: Sri Lankan subjects used in the replication phase of the South Asian GWAS 80

4.2.3 Study B: Sri Lankan subjects used as part of the replication phase for the Sikh Punjabi GWAS 80

4.2.4 Genotyping 81

4.2.5 Statistical analysis 81

4.2.6 Collating the reported allele frequencies and odds ratios for the 24 SNPs in South Asians and European populations 82

4.3 Results 84

4.3.1 Genotyping of 24 SNPs in the Sri Lankan sample set 84

4.3.2 Study A: The identification of six novel T2D susceptibility loci in individuals of South Asian ancestry 84

4.3.3 Study B: Genotyping of 4 SNPs in the Sri Lankan cohort as part of the replication phase of the GWAS performed in Sikh Punjabi individuals from India ... 85

4.3.4 Comparison of T2D risk variants between Sri Lankans, other South Asian cohorts and Europeans 90

4.4 Discussion 101

Chapter 5 106

5.1 Introduction 107

5.1.1 The importance of diagnosing monogenic diabetes in the clinic 107

5.1.2 Using glycan profiles as a biomarker for HNF1A-MODY 108

5.1.3 Determining whether HNF1A variants identified in patients with biomarker profiles suggestive of HNF1A dysfunction cause HNF1A-MODY 111
5.1.4 Experimental aims 112
5.2 Samples and methods 112
5.2.1 Subjects studied 112
5.2.2 PCR and gel electrophoresis of HNF1A exons 113
5.2.3 Capillary sequencing of HNF1A exons 114
5.2.4 Assessing the pathogenicity of HNF1A variants identified 114
5.2.5 In silico prediction tools 114
5.2.6 Genotyping of D526N in the ORCADE population 116
5.2.7 Site-directed mutagenesis and construction of HNF1A mutant proteins ... 116
5.2.8 Transactivation assays 118
5.2.9 Protein expression and subcellular localisation 119
5.2.10 Statistical analysis 120
5.3 Results 121
5.3.1 Identification of HNF1A variants, follow-up family studies and in silico predictions 121
5.3.2 Genotyping of the p.D526N variant in the ORCADES cohort 128
5.3.3 Functional characterisation of HNF1A variant proteins: transactivation studies 128
5.3.4 Protein expression and localisation of HNF1A variant proteins 131
5.3.5 Assessing nuclear localisation of HNF1A variant proteins 137
5.4 Discussion 141
Chapter 6 147
6.1 Introduction 148
6.1.1 Maturity Onset Diabetes of the Young (MODY) in South Asians 148
6.1.2 Using highly sensitive C-reactive protein (hsCRP) as a biomarker for HNF1A-MODY 149
6.1.3 Evaluating hsCRP as a biomarker for HNF1A-MODY in South Asians 151

6.1.4 Experimental aims 151
6.2 Samples and methods 152
6.2.1 Subjects 152
6.2.2 Mutational screening of HNF1A 154
6.2.3 Assessing the pathogenicity of HNF1A variants identified 154
6.2.4 In silico assessment of the HNF1A variants 154
6.2.5 Site-directed mutagenesis and construction of HNF1A mutants 154
6.2.6 Generation of pGL3-CRP reporter construct 156
6.2.7 Transactivation assays 156
6.2.8 Protein expression studies—western blots 157
6.2.9 Cellular localisation—immunofluorescence 157
6.2.10 Statistical Analysis 157
6.3 Results 158
6.3.1 Identification of HNF1A variants in South Asian Subjects 158
6.3.2 c.307 G>A.p.V103M 159
6.3.3 c.1015 G>A.p.G339S 166
6.3.4 c.1135 C>T.p.P379S 166
6.3.5 c.1544 C>T.pT515M 168
6.3.6 c.1751 T>C.p.L584P 168
6.3.7 c.14delT and c.872dupG 168
6.4 Discussion 169
Chapter 7 174
Appendices 182

Chapter 1

General Introduction

1.1 The global impact of type 2 diabetes

The growing prevalence of type 2 diabetes (T2D) is major health concern affecting countries worldwide. In 2014, 387 million people, almost 9% of the adult population, were estimated to have diabetes (Figure 1.1.1) (IDF 2013). It is predicted that by 2035, this figure will increase by 55% to 592 million people (IDF 2013). T2D is one of the world's leading non-communicable diseases (NCDs) accounting for almost 90% of diabetes cases globally (WHO factsheet 312). Together with its impact on human health, T2D also places a considerable economic burden on health care systems and families (IDF 2013). Worldwide, the number of deaths attributed to diabetes was 5.1 million and 10.8% of overall health expenditure was used to treat diabetes in 2013 (IDF 2013).

Many cases of T2D remain undiagnosed, as individuals are unaware of the symptoms, or live in low- and middle-income countries where medical resources are not always easily accessible. Delayed diagnosis and poor management of diabetes can lead to costly and detrimental health complications (IDF 2013). Devising and implementing intervention schemes to reduce the number of individuals with diabetes and encourage appropriate disease management will assist in reducing the global burden associated with T2D.

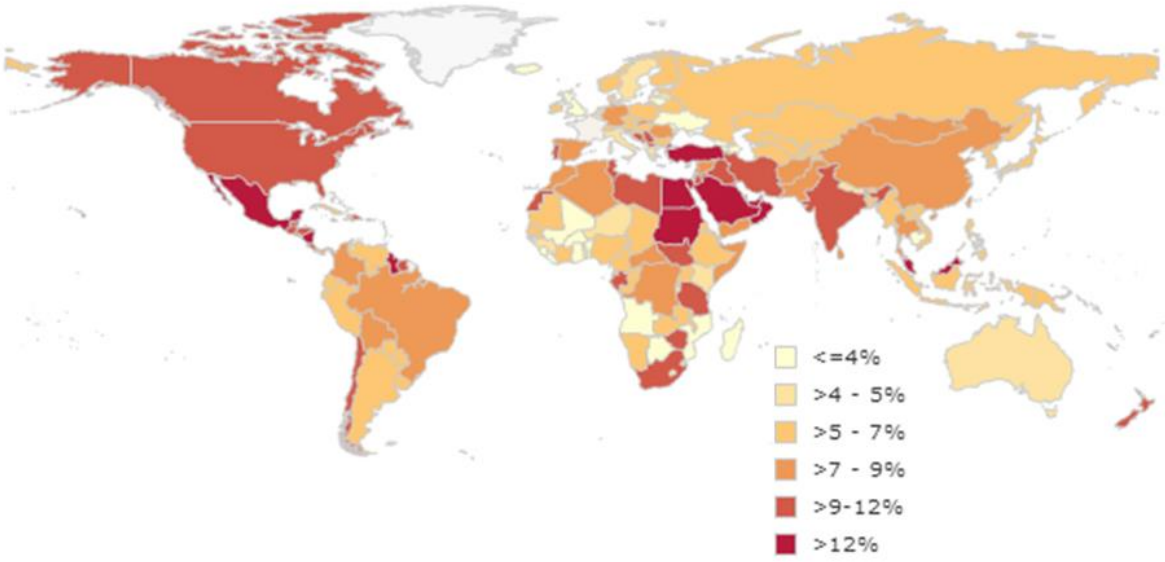


Figure 1.1.1 Global type 2 Diabetes prevalence, 2014. Prevalence (%) of T2D in the adult population (20–79 years). The figure was taken from the International Diabetes Federation Diabetes Atlas (http://www.idf.org/atlasmap/atlasmap?indicator=i1&date=2014).

1.2 Pathophysiology of type 2 diabetes

1.2.1 Diagnosis and clinical presentation

Diabetes is a condition that can affect people in a variety of forms at various stages throughout life. T2D is the most common type of diabetes (described below), followed by type 1 diabetes. T1D is an immune-mediated disease that develops due to the selective destruction of the insulin-producing pancreatic beta-cells. Patients with T1D are usually treated with insulin injections to compensate for the lack of insulin production. Other subtypes of diabetes also exist including monogenic diabetes and this will be discussed further in section 1.5.

T2D is a metabolic disorder characterised by high blood glucose levels (hyperglycemia) due to inadequate amounts of insulin being produced to meet the body's requirements. This insufficient insulin production or/and action is caused by relative impaired insulin secretion, defects in insulin action or both (ADA 2012). Clinical diagnosis is usually made by determining blood plasma glucose levels via a blood test. Assessment is made using a combination of measurements for plasma glucose levels after a fasting state (no caloric intake for at least 8 hours) and 2 hours (2-h) post an ingestion of an oral glucose challenge (ADA 2012). Current World Health Organisation (WHO) guidelines suggest diagnostic criteria of ≥7.0 mmol/L (126mg/dl) for elevated fasting plasma glucose levels and ≥11.1 mmol/L (200 mg/dl) for 2-h plasma glucose (WHO 2012). An intermediate group of individuals with 'impaired' glycaemia at a high risk of developing future diabetes have also been categorised. The cut-off criteria suggested for fasting plasma glucose levels is 6.1 to 6.9 mmol/L (110 to 125 mg/dl) and for 2-h glucose levels is ≥7.8 and <11.1 mmol/L (140 and 200g/dl) (WHO 2012). Glycated haemoglobin (HbA1c) is commonly used by clinicians as a measure of diabetes management as it represents blood glucose levels over a 2- to 3-month period. Levels above ≥48 mmol/mol (≥6.5%) denote glycaemia (ADA 2012). Investigations looking at prolonged glycaemia over time have demonstrated that by reducing and improving HbA1c levels the associated risk of T2D complications can be reduced (Stratton et al. 2000).

The typical symptoms of T2D include increased thirst (polydipsia) and hunger, frequent and excessive urination (polyuria), fatigue, and unexplained weight change. Less common signs include skin irritations and blurred vision. Many of these symptoms develop gradually and often appear mild; consequently, patients may be diagnosed several years after disease onset. T2D is linked to many secondary health complications

as prolonged hyperglycemia can damage blood vessels, nerves, and organs. These are all damaging to health and in some cases can lead to life-threatening complications. Patients with T2D have an increased risk of developing cardiovascular disease (CVD), neuropathy, diabetic retinopathy, and nephropathy (Clarke et al. 2003). Premature mortality in patients with T2D is primarily caused by CVD or renal failure (Roglic and Unwin 2010). One study showed that individuals with diabetes conferred a twofold risk for coronary heart disease and major stroke subtypes (Sarwar et al. 2010). Nearly 70% of all lower-half extremity amputations are performed on patients with diabetes (Bruun et al. 2013). Diabetic retinopathy is one of the leading causes of vision loss in working-age adults (IDF 2013).

1.2.2 Causes of type 2 diabetes

Despite the high prevalence of and damaging health consequences associated with T2D, discovering the underlying functional pathways contributing to disease pathogenesis remains a challenge. Adaptations in lifestyle, such as reduced physical activity, together with dietary changes that include a higher intake of foods that are processed, energy-dense with a high-fat and sugar content, are undoubtedly linked to the marked rise in T2D prevalence. The metabolic modifications that occur because of these environmental alterations combined with genetic influences result in a very heterogeneous disorder (Nolan et al. 2011).

1.2.2.1 Insulin resistance

Insulin, a hormone released from the pancreatic beta-cell, plays a key role in regulating the metabolism of carbohydrates and fats. The balanced interaction between insulin action and insulin secretion is the determining factors that govern glycaemia (Stumvoll et al. 2003). Changes in plasma glucose levels during a postprandial or fasting state are carefully mediated by insulin via a tissue-specific response. The increase in circulating plasma insulin levels promotes the uptake of glucose by skeletal muscle, suppresses lipolysis (breakdown of adipocytes) and reduces glucose output from the liver (Dinneen et al. 1992; Stumvoll et al. 2005). Insulin resistance occurs when the effect of insulin on any of these pathways is insufficient; for example, if glucose clearance by skeletal muscle is affected, or if suppression of endogenous glucose production by the liver is influenced (Stumvoll et al. 2005). Obesity is strongly connected with insulin resistance (Kahn et al. 2006). Adipose tissue is important in metabolism and modulates certain circulating hormones (leptin and adiponectin), the release of non-esterified fatty acids (NEFAs) and pro-inflammatory cytokines. The levels of many of these products are

increased in obesity, influencing insulin action (Kahn et al. 2006; Stumvoll et al. 2005). Higher levels of NEFAs and the inflammatory cytokines tumour necrosis factor-α (TNFα) and interleukin-6 (IL6) are produced by expanded visceral adipose tissue, negatively influencing the insulin-signalling cascade. TNFα encourages adipocyte lipolysis, thus increasing circulating NEFAs (Hotamisligil 2000). Concurrently, NEFAs inhibit insulin-stimulated glucose metabolism in skeletal muscle and stimulate gluconeogenesis (synthesis of new glucose molecules) (Boden and Shulman 2002; Stumvoll et al. 2005).

1.2.2.2 Insulin secretion

Insulin resistance alone is not enough to result in T2D; in some individuals at least, the pancreatic beta-cells are able to compensate by upregulating insulin secretion to maintain balanced glycaemia. However, over time the beta-cells can become unable to meet increased demands, resulting in hyperglycaemia (Stumvoll et al. 2003; Stumvoll et al. 2005).

In healthy individuals, the postprandial increase in plasma glucose levels triggers a rapid uptake of glucose by the beta-cell. Glucose is taken up via several cell surface glucose transporter proteins. The enzyme glucokinase (GCK) is the main glucose sensor for beta-cells and catalyses a reaction that converts glucose to glucose-6-phosphate (Matschinsky et al. 1993). The generation of ATP via glycolysis and the Krebs cycle and the consequential change in ATP/ADP ratios result in the closure of the ATP-sensitive potassium (K_{ATP}) channels. The closure of K_{ATP} channels alters the cell membrane potential and membrane depolarization leads to the opening of L-type calcium channels. The influx of calcium leads to the fusion of insulin secretory granules with the plasma membrane and the release of insulin to the cell membrane via exocytosis (Ashcroft and Rorsman 2012; Fajans et al. 2001). Figure 1.2.1 illustrates the normal process of glucose-stimulated insulin secretion (GSIS).

Disruptions at any stage of GSIS can alter insulin secretion. Cases of monogenic diabetes demonstrate how mutations in a single gene involved in the GSIS pathway can lead to impaired insulin secretion and progressive beta-cell failure. Monogenic diabetes will be discussed further in section 1.5. Other hormones and neurotransmitters are also involved in insulin secretion and can influence directly or indirectly the effect insulin has on tissues (Ashcroft and Rorsman 2012). Glucagon, a hormone secreted from the pancreatic alpha-cells, works together with insulin to maintain normoglycaemia.

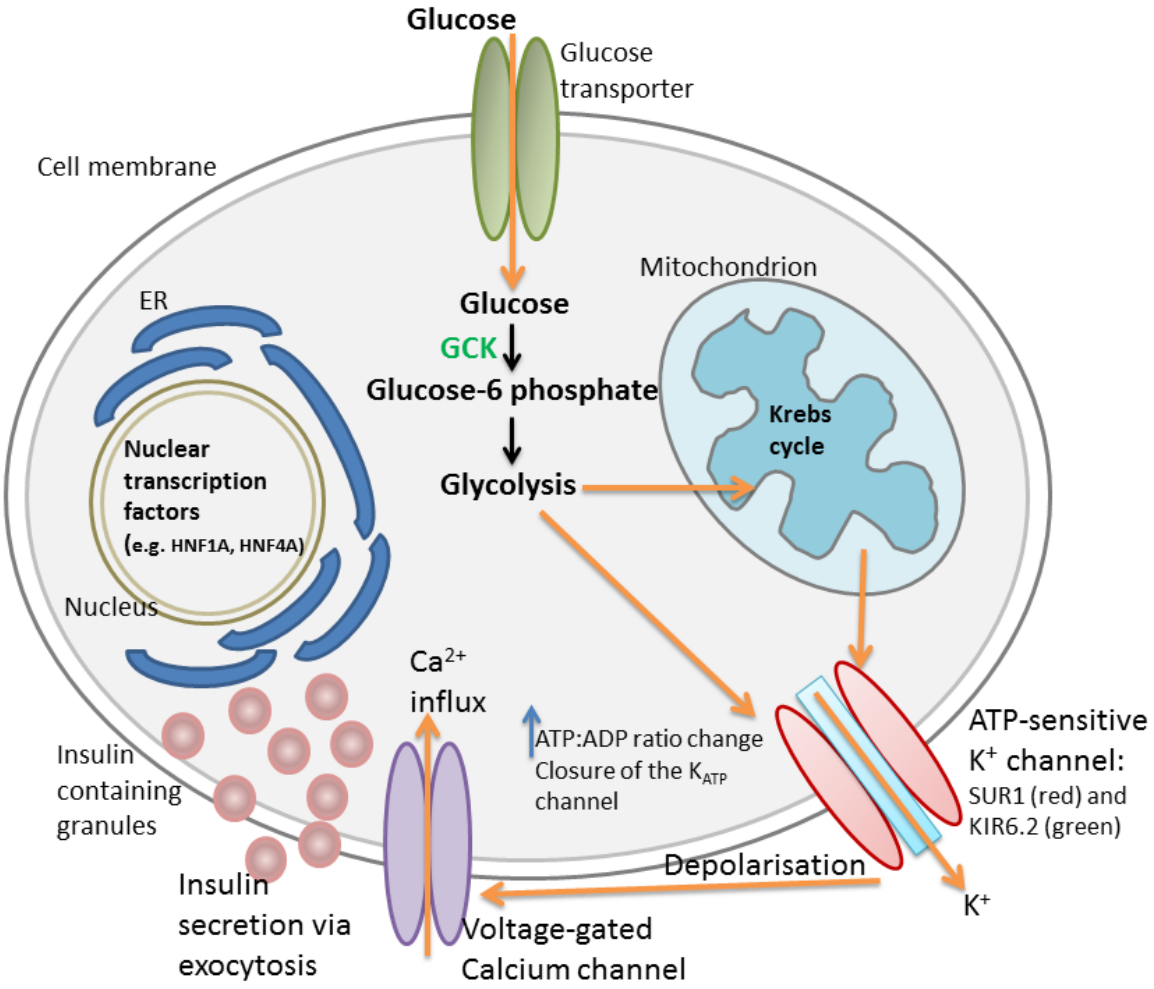


Figure 1.2.1 Schematic representation of insulin secretion from the pancreatic beta-cell. Glucose enters the cell via glucose transporters and is phosphorylated to glucose-6-phospahate by the enzyme glucokinase (GCK). Glucose-6-phospate is further metabolised by glycolysis and the Krebs cycle resulting in the generation of ATP. The change in the ATP:ADP ratio causes the K_{ATP}-sensitive channel to close, resulting in membrane depolarisation and subsequent opening of the voltage-gated calcium channels. The influx of calcium into the cell triggers exocytosis of insulin-containing granules. Figure adapted from (Fajans et al. 2001)

Glucagon levels are suppressed when plasma insulin concentrations increase. The decrease in glucagon levels reduces the output of glucose by the liver by inhibiting glycogenolysis (breakdown of glycogen) and gluconeogenesis (synthesis of new glucose molecules). This reciprocal mechanism ensures that sufficient levels of glucose are released into the systemic circulation when in a postabsorpative state (3–5 h after a meal and thereafter) (Dinneen et al. 1992).

1.3 Type 2 diabetes and genetics

1.3.1 The role of genetics in type 2 diabetes

There are widespread geographical differences in the global prevalence for T2D. For example, among Caucasians in Europe T2D prevalence is less than 5%, whereas amongst Native American Arizona Pima Indians the prevalence for T2D is almost 50% (Diamond 2003). Although it is widely acknowledged that lifestyle factors, dietary changes and increased life expectancy are important contributors to the current T2D epidemic, the indication for a genetic component influencing T2D predisposition comes from several factors. Inconsistencies in T2D prevalence between ethnic groups suggest genetic factors might contribute to disease risk. This is illustrated by differences in T2D prevalence in the United Kingdom (UK) in 2013, between Europeans (1.7%) and South Asians (6.7%) (Diabetes UK 2014). Familial aggregation for T2D has been recognised for many years with evidence from twin family, and population studies. For example, the relative risk for developing diabetes among siblings if one parent is affected is 3.5 and this increases to a sixfold risk if both parents are affected, when compared with the general population (Meigs et al. 2000). In addition, the concordance rate of T2D between monozygotic (MZ) twins is estimated to be around 70% whilst between dizygotic (DZ) twins it is only 20–30% (Medici et al. 1999; Newman et al. 1987; Poulsen et al. 1999). However, there are some limitations to twin studies that might lead to a higher concordance rate in MZ twins—such as a more similar prenatal environment than that with DZ twins—that need to be considered (Barroso 2005). Our understanding of T2D pathogenesis, and the ability to investigate the genetic components contributing to disease risk, has been greatly enhanced by the emergence of powerful genomic technologies, which will be discussed below (Lupski et al. 2011).

1.3.2 Methodologies for variant discovery

1.3.2.1 Linkage analysis and candidate gene association studies

Early approaches employed to identify genetic markers that co-segregated with diabetes were met with limited success for T2D. Family-based linkage studies proved successful for identifying rare, highly penetrant mutations involved in monogenic diabetes (Owen and Hattersley 2001). However, these studies had little impact in identifying genetic variants with low penetrance such as those implicated in T2D, as many regions failed to replicate in subsequent studies.

The candidate gene approach searched for association in genes that were selected based on prior knowledge of biology. These studies were performed in large numbers of unrelated individuals, and revealed two notable success stories, which have been reproducibly associated with T2D. The P12A variant in *PPARG* (Altshuler et al. 2000) and the E23K variant in *KCNJ11* (Gloyn et al. 2003) are both compelling candidates, and are current targets for T2D therapy; thiazolidinedione is a type of insulin-sensitising drug targeting the PPARG receptor, whilst the K_{ATP} channel, specifically SUR1 is the target for sulphonlyureas.

1.3.2.2 Genome-wide association studies (GWAS)

The identification of *TCF7L2* via the candidate gene association studies approach was based on previous studies that suggested linkage at the 10q region (Grant et al. 2006). At present, variants at this locus are the most consistently replicated, and confer the greatest risk for T2D. In addition, *TCF7L2* provided the first example for T2D where a gene was identified without prior knowledge of biological candidacy (Grant et al. 2006).

Over the past decade, there has been an explosion in the number of common single nucleotide polymorphisms (SNPs) that have robustly replicated with complex traits. The advent of genome-wide association studies (GWAS) gave the opportunity to scan the entire genome in a hypothesis-free manner, in large collections of individuals with and without disease (Consortium 2007). Several important developments endorsed the success of GWAS. The completion of the Human Genome Project, and the understanding of linkage disequilibrium (LD) patterns facilitated by the HapMap Project, showed that investigators did not have to genotype all the SNPs in the genome (Consortium 2005; Frazer et al. 2007). As sections of the genome are inherited together, only a proportion of SNPs known as informative SNPs (tagging SNPs) were required to be genotyped, as these SNPs were able to act as representatives for the rest, with only a slight loss in

statistical power (Consortium 2005; Frazer et al. 2007). In addition, the development of new genotyping technologies, the reduction in genotyping costs, the availability of statistical analysis tools and the collection of large well-characterised patient sample sets have all aided GWAS in becoming a powerful platform for detecting common genetic variation (Consortium 2007).

The breakthrough year for T2D was 2007, which saw the initial wave of GWA studies. This confirmed the already-implicated loci *PPARG*, *KCNJ11* and *TCF7L2* and the identification of five novel susceptibility loci 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). The sixth signal, a variant near *FTO*, was found to be indirectly associated with T2D through an effect on body mass index (BMI), hence identifying one of the first loci associated with obesity (Frayling et al. 2007).

In European populations, the DIAbetes Genetics Replication And Meta-analysis (DIAGRAM) combined GWAS data from three studies and identified six novel loci influencing T2D risk (Zeggini et al. 2008). A subsequent meta-analysis merged data from five further studies identifying twelve additional T2D signals (Voight et al. 2010). A GWAS meta-analysis performed using data generated from large-scale genotyping, using a custom-based array Metabochip, identified a further ten T2D susceptibility loci (Morris et al. 2012).

By aggregating studies, and increasing sample sizes, the power to detect variants with smaller effect sizes and more extreme allele frequencies has vastly improved. GWAS have also been extended to individuals of non-European descent. Together, GWAS in Europeans and non-Europeans have led to the successful completion of transethnic meta-analyses and fine-mapping studies (Mahajan et al. 2014) discussed further in section 1.4.1. To date, taken together GWAS across multiple ancestral groups have identified almost ~153 common variants (MAF >5%) which map to >120 loci for T2D (reviewed in (Prasad and Groop 2015).

In instances where the GWAS signal is a non-synonymous variant, located in the coding region of a plausible gene, it has provided a tractable target for functional follow-up and establishing a molecular mechanism. For example, non-synonymous variants in *GCKR* and *SLC30A8*, both credible biological candidates for influencing T2D risk, have since been functionally evaluated. The signal identified on chromosome 2 through a GWAS (Saxena et al. 2007) was found to map to a common variant, p.P446L in *GCKR*, which

encodes the glucokinase regulatory protein (GKRP) (Orho-Melander et al. 2008). Functional characterisation of this p.P446L has shown that the risk allele affects GKRP function through an assortment of molecular mechanisms, resulting in the activation of GCK under conditions of normoglycaemia and hypoglycaemia (Beer et al. 2009; Rees et al. 2012). *SLC30A8*, one of the first loci implicated with T2D (Sladek et al. 2007), encodes a zinc transporter (ZnT8), predominately expressed in the pancreatic beta-cell (Chimienti et al. 2004). The transporter enables the accumulation of zinc from the cytoplasm into intracellular vesicles, crucial for insulin maturation and storage (Chimienti et al. 2004). Characterisation of the missense SLC30A8 variant p.R325W has demonstrated that the variant causes the zinc transporter to be less active, and thus decreases intracellular zinc levels (Kim et al. 2011; Nicolson et al. 2009).

1.3.2.3 Whole genome/exome sequencing

Recent deep sequencing efforts have established that rare variants are in abundance in the human genome due to accelerated and rapid population growth and weakened purifying selection (Keinan and Clark 2012; Tennessen et al. 2012). These studies have further supported the view that rare and low-frequency variants may have an important role in contributing to complex disease risk (Nelson et al. 2012). This has prompted a change in direction and studies are now starting to focus on the search for low- (MAF 1–5%) and rare (<1%) frequency variants. It is anticipated that these variants will have modest effect sizes, bridging the gap between the large effects seen in rare Mendelian forms of diabetes and the small effects seen by common variants associated with T2D (Manolio et al. 2009).

The development of next generation sequencing (NGS) technologies is enabling the search for rare and low-frequency variants via whole genome sequencing (WGS) and whole exome sequencing (WES). Facilitated by decreasing sequencing costs, NGS technologies are now comprehensively detecting both common and rare coding variants in an unbiased manner (Kiezun et al. 2012).

Whole exome sequencing, which targets the protein coding regions of the genome, has shown to be a powerful and cost-effective method for detecting rare, causative alleles, particularly for Mendelian disorders (Bamshad et al. 2011). Studies are now emerging demonstrating the clinical application of using WES for patients. For example, (Bonnefond et al. 2012a) demonstrated that by using WES and focusing on variants that were shared between affected individuals it was possible to identify a single non-

synonymous variant in the *KCNJ11* gene. Accurately diagnosing subtypes of diabetes has important implications for diabetes treatment and management for the patient. Already implicated in neonatal diabetes mellitus (NDM), patients with mutations in *KCNJ11* can be successfully treated with oral sulphonlyureas rather than insulin therapy (Gloyn et al. 2004).

Whole exome sequencing has been useful at providing further insight into beta-cell function. One investigation demonstrated how WES discovered coding, disease-causing mutations in a subset of patients with unexplained pancreatic agenesis. This rare condition is caused by complete or partial failure of the pancreas to develop, manifesting in both neonatal diabetes and exocrine pancreatic insufficiency (Lango Allen et al. 2012). By sequencing the exomes of two affected individuals and their unaffected parents and then applying a filtering strategy to enrich for likely pathogenic mutations Lango Allen et al. identified heterozygous mutations in the *GATA6* (GATA-binding factor 6) gene in each affected proband. Subsequent targeted sequencing of *GATA6* identified another 12 individuals harbouring mutations in *GATA6*. Follow-up studies have shown that sequence alterations in GATA6 cause a spectrum of diabetic phenotype from pancreatic agenesis to adult onset T2D without exocrine insufficiency (Catli et al. 2013; De Franco et al. 2013). Functional studies suggest that inactivating mutations in *GATA6* result in loss of function, causing pancreatic agenesis due to haploinsufficiency.

In a recent study the deCode consortium was able to gather a sample size of nearly 300,000 individuals (Steinthorsdottir et al. 2014). This was possible by combining WES data of 2,630 Icelandic individuals along with imputed data created around Icelandic genealogy were the authors calculated genotype probabilities for 290,256 untyped individuals (Steinthorsdottir et al. 2014). This sufficiently powered the study to detect four previously unreported variants associated with T2D (Steinthorsdottir et al. 2014). Of these, one low-frequency variant found in intron 1 of the *CCND2* gene was found to reduce T2D risk by half (OR=~0.5). Two missense variants were identified in *PAM* that conferred moderate risk for T2D and a rare frameshift mutation in *PDX1*, a gene already implicated in monogenic diabetes (Steinthorsdottir et al. 2014). Low-frequency variants in *PAM* have also been reported to be associated with the insulinogenic index (Huyghe et al. 2013). Although the risk variants detected in *PAM* and *PDX1* were predicted to impact protein function, no direct functional experiments were performed to confirm causality (Steinthorsdottir et al. 2014).

The overlap between Mendelian diabetes genes and those implicated in T2D susceptibility is expanding (*PDX1*, *PPARG*, *HNF1A*, *KCNJ11*, *ABCC8 and INS*). In some of these genes rare loss of function mutations are now being classified as causing a spectrum of diabetes-related phenotypes ranging from neonatal diabetes, MODY (Maturity-Onset Diabetes of the Young), T2D or in some cases with no influence on disease status. As more rare variants with large effect sizes are identified as being associated with T2D, it will become increasingly difficult to define a cut-off point for classification of a mutation as an incompletely penetrant Mendelian variant or a rare T2D mutation with a large effect size in individuals. Functional work of the kind described in this thesis may help to address this issue.

1.3.3 What have we learnt so far and the challenges ahead

Physiological characterisation of a subset of variants at identified GWAS loci in normoglycaemic individuals has revealed that many of the variants exert their effect for disease risk through beta-cell dysfunction rather than increased insulin resistance (Dimas et al. 2014; Voight et al. 2010).

One of the main limitations and current challenges of GWAS has been the slow progress in moving from the discovery of the association signal to insightful biological interpretation. This has been attributed to several factors. First, linkage disequilibrium patterns are extensive, and so the associated variant may only be linked to the causal variant rather than being the actual causal variant(s). Second, many of the signals map to non-coding regions of the genome and so defining the causative transcript for a specific variant(s) and determining the functional effects has been difficult.

The sequence-based studies currently under way offer a more systematic approach of identifying causal alleles. They are likely to discover novel variants at rarer and lower frequencies with larger effect sizes and some of which will map to coding regions. The detection of rare variants in already implicated regions will provide evidence supporting the particular transcript (Bonnefond et al. 2012b; Nejentsev et al. 2009). Furthermore, the finding of rare and low-frequency variants with moderate effect sizes may explain the missing proportion of genetic variation for disease susceptibility (Manolio et al. 2009). Currently, for the ~120 genomic loci identified to influence T2D susceptibility, the effect sizes are modest and, when combined, explain less than 20% of the family aggregation for T2D (Prasad and Groop 2015).

Continued work will involve studies in very large cohorts, especially when searching for rarer, more penetrant variants. These studies will have important outcomes for the discovery of rare and low-frequency variants but also for sequencing data used in a clinical setting for diagnostic and prognostic purposes. For this information to be used effectively for either disease prediction or to inform therapy options, detailed functional work at the cellular or molecular level will be necessary to distinguish between variants that affect gene function from those that are neutral.

1.4 Type 2 diabetes in the South Asian population

1.4.1 Replication of discovery variants in non-European populations

The GWAS already published by the time I started my DPhil had identified many risk alleles associated with T2D susceptibility (Voight et al. 2010; Zeggini et al. 2008). However, all of these studies had been performed using samples of European descendants. Therefore, these left several open questions in populations of non-European descent. First, will the results seen in Europeans be replicated in other populations? Second, will the causal variants have similar allele frequencies and effect sizes contributing to disease risk, in turn providing an improved understanding of disease predisposition in non-European populations (Rosenberg et al. 2010)? The work presented in Chapter 3 of this thesis aims to address these questions in a South Asian population of Sri Lankan descent.

Several developments in technology and statistical methodologies have assisted in the growing number of GWAS that have been completed in non-Europeans (Cho et al. 2012; Palmer et al. 2012; Saxena et al. 2012 ; Tabassum et al. 2013; Yasuda et al. 2008). First, there has been an improvement in genotyping platforms capturing the genetic variation in non-European individuals (Morris 2011). Second, the improvement in the accurate imputation (refers to the statistical inference of individual genotypes at un-typed loci from other marker sets) has been achieved by the availability of matched reference panels such as those derived in the HapMap and 1000 Genomes dataset improving power for association mapping (Marchini et al. 2007; Nothnagel et al. 2009). Chapter 4 of this thesis describes the results of my contribution to two GWAS performed in individuals of South Asian ancestry.

The methodology of GWAS relies on patterns of linkage disequilibrium (LD) to minimise the number of variants that need to be genotyped to capture the entire genome. However, this means that the associated region may be quite large, containing many

possible candidate genes. On confirmation of the signal through replication studies, identifying the causal variant for subsequent functional characterisation is difficult due to patterns of linkage disequilibrium. The increasing availability of GWAS data from diverse populations offers the opportunity to conduct much larger studies on the transethnic level (Morris 2011). It is anticipated that transethnic meta-analysis will increase power for detecting further loci with even more modest effect sizes that are associated with complex traits (Morris 2011). Furthermore, by exploiting differences in allele frequencies and LD patterns between major populations, under the assumption that the causal variant is the same across multiple studies, transethnic fine-mapping studies offer the opportunity to refine the association signal, and potentially to define the causal variant (McCarthy 2011). A recent study that aggregated GWAS meta-analyses from European, East Asian, South Asian and Mexican and Mexican American ancestry found that there was an excess in T2D risk loci, which were directionally consistent across all ethnic groups (Mahajan et al. 2014). In addition, Mahajan et al. observed an improvement in fine-mapping resolution for several T2D susceptibility loci. For example, for the *JAZF1* loci they were able to reduce the genomic interval to include only four SNPs that were in LD with the lead SNP across all ancestral groups (Mahajan et al. 2014).

1.4.2 Prevalence and presentation

Nearly one-fifth of the world's adult population with diabetes lives in the South-East Asia region. This region comprises seven countries, including individuals from India, Pakistan, Bangladesh, Sri Lanka, Nepal, Bhutan and the Maldives (IDF 2013). In 2013, the number of people estimated to have diabetes (20–79 years) within this region was 72.1 million. This figure is predicted to rise to 123 million by 2035, equating to 10.1% of the adult population (IDF 2013).

South Asians are a diverse, multicultural population consisting of many distinct religions, sociocultural practices and dietary preferences (Ramachandran et al. 2010) South Asian populations have an increased propensity for pre-diabetes with a higher prevalence of both impaired fasting glucose and impaired glucose tolerance, leading to a more rapid progression to diabetes (Ramachandran et al. 2006). Compared with other ethnic groups, T2D in South Asians occurs almost a decade younger and at a lower BMI threshold (Ramachandran et al. 1997). Due to the earlier onset of T2D, the overall lifetime risk of developing the associated macrovascular and microvascular complications are therefore increased amongst native and immigrant South Asians (Misra et al. 2014).

1.4.3 Factors influencing differences in T2D incidence

The increased predisposition for T2D in South Asians is due to a complex interaction between lifestyle, nutritional, genetic and acquired risk factors that will be discussed below (Ramachandran et al. 2014).

1.4.3.1 Environmental factors

Rapid globalisation has led to socioeconomic and nutritional transitions in many South Asian countries. The economic shift has vastly improved levels of modernisation and industrialisation (Ramachandran et al. 2010). The rates of urbanisation have also increased, partly to accommodate the growing population and partly because of migration from rural to urban areas (Srinath Reddy et al. 2005). The behavioural changes associated with an urban lifestyle impact heavily on the rising prevalence of many chronic disorders in South Asians including obesity, T2D and cardiovascular disease (Ramachandran et al. 2014). Urbanisation is associated with increased incomes and lifestyle changes that influence levels of physical activity and dietary habits. The subsequent imbalance between diet and reduced energy expenditure has led to the increased prevalence of obesity (Popkin 2006). The urban–rural convergence is influencing a lifestyle transition in rural communities; for example, in many developing countries the number of overweight people is increasing in rural areas (Leon 2008). A study that focused on a rural population in India reported a threefold increase in diabetes prevalence over a 14-year period due to improved standards of living, reduced levels of physical activity and increased prevalence of obesity (Ramachandran et al. 2004).

Societal factors also pose hurdles for disease management and treatment regimes, increasing mortality rates due to diabetes and its related complications. This is partly attributed to differences in cultural practices and religious beliefs. Many cases of T2D remain undiagnosed due to lack of awareness of diabetic symptoms and the resistance in seeking medical help (Misra et al. 2014). Suboptimal glycaemic control due to non-compliance with treatment and lifestyle changes is also a factor (Misra et al. 2014). Moreover, a delayed diagnosis due to a lack of affordable healthcare and treatment for many people in lower socioeconomic groups increases the likelihood of further complications (Ramachandran et al. 2014).

1.4.3.2 Biological factors

In South Asians, the adverse health effects caused by lifestyle transitions are met with an inherent predisposition for insulin resistance and other metabolic features that increase the risk of T2D susceptibility (Ramachandran et al. 2010).

The greater degree of insulin resistance in South Asians is a key feature for the increased susceptibility for T2D. The notion that South Asians require higher levels of insulin to maintain euglycemia has been demonstrated in several studies comparing both normoglycaemic and diabetic Asian Indian patients, and in comparison with individuals of European origin (McKeigue et al. 1991; Mohan et al. 1986). Indian Asians had higher basal levels of insulin in both healthy controls and T2D patients, and required greater levels of insulin in response to an oral glucose load (Mohan et al. 1986). This ethnic-specific, reduced insulin sensitivity is related to South Asians showing a tendency for increased central obesity even though individuals are not considered overweight (McKeigue et al. 1991). For a given BMI, South Asians tend to have a greater amount of total body fat than individuals of European origin (Kalra et al. 2013). This differential body composition and higher absolute levels of both subcutaneous fat and visceral fat are not reflected in their non-obese BMI. The higher levels of visceral fat have been associated with adverse health effects, including insulin resistance and dyslipidaemia (Banerji et al. 1999; Raji et al. 2001). Although (Chandalia et al. 2007) showed no difference in visceral fat depots between South Asians and Europeans, they demonstrated that truncal fat and larger abdominal subcutaneous adipocytes were correlated with insulin resistance in South Asians. Adipocytes secrete hormones or 'adipokines' that control important metabolic processes including insulin action; however, in states such as obesity they become dysfunctional (Antuna-Puente et al. 2008).

There are several evolutionary and developmental hypotheses that have been postulated to account for the increased risk for T2D in South Asians. The 'thrifty' genotype hypothesis suggests that genetic variation aided survival in the past when food supplies were intermittent but now have evolved into disadvantageous traits due to environmental and lifestyle changes (Neel 1962). A recent study took 65 credibly associated T2D risk loci (known at that time) and showed that when considered together these risk alleles were not determined by past positive selection, and so it is unlikely that this hypothesis is a cause for the increasing prevalence for T2D (Ayub et al. 2014). However, this study was performed using samples of African, European and East Asian ancestry only, and further investigations are required in individuals of South Asian origin.

An alternative developmental proposal, the 'thrifty' phenotype, suggests that intrauterine detriments, such as maternal malnutrition during pregnancy, induce fetal reprogramming geared towards coping with future food shortages (Hales and Barker 1992). An investigation that examined body size and fat measurements in newborn babies in rural India compared their findings with European babies born in the UK and found that Indian babies had lower birthweights; and low muscle mass, but fat preservation was similar (Yajnik et al. 2003). The authors suggested that if carried through postnatally and accelerated in childhood through rapid weight gain, it might predispose to an insulin-resistant state (Yajnik et al. 2003).

Although lifestyle transitions including dietary habits and increased physical inactivity have a substantial role in the growing T2D epidemic in South Asians, genetic factors also influence risk. As described above in section 1.4.1 studies examining the genetic predisposition for T2D risk in South Asians have been well underway since the start of this thesis and several of these will be described further in Chapters 3 and 4. Additional genetic studies in South Asians, such as those described in this thesis, will further our understanding of T2D pathophysiology in this high-risk group, aiding better predictive analysis and preventative measures for at-risk individuals.

1.5 Monogenic diabetes

Rare, monogenic forms of diabetes represent a heterogeneous group of disorders caused by defects in a single gene. These can be subclassified into Maturity Onset Diabetes of the Young (MODY), neonatal diabetes, mitochondrial diabetes and syndromic forms of diabetes, of which MODY is the most common (Molven and Njolstad 2011). Studies of the genes implicated in monogenic diabetes have provided an insight into how disruption in a particular gene can influence disease pathogenesis. Many of the gene defects either affect the development of the beta-cell or pancreas and/or the insulin secretory pathway leading to pancreatic beta-cell dysfunction (reviewed by (Schwitzgebel 2014). The insight gained into the molecular and physiological pathways from some of these investigations has led to some invaluable therapeutic advances. For example, both *KCNJ11* and *ABCC8* are crucial for regulating insulin release as they encode the subunits that compose the K_{ATP} channel. Activating mutations in both these genes cause neonatal diabetes mellitus as sensitivity to ATP is diminished. Treatment with an oral sulphonylurea drug has been shown to restore insulin secretion as it promotes closure of the K_{ATP} channel in an ATP-independent manner (Babenko et al. 2006; Gloyn et al. 2004).

1.5.1 Different subtypes of Maturity Onset Diabetes of the Young (MODY)

Maturity Onset Diabetes of the Young along with juvenile onset was first described by (Tattersall and Fajans 1975) to aid differentiation between T1D (insulin-dependent) and T2D (noninsulin-dependant) diabetes. To date, 13 genes have been implicated in causing MODY (with varying degrees of certainty) and these are summarised in Table 1.5.1. For many of these MODY genes, the clinical severity depends on the type of mutation (homozygous or heterozygous), and how deleterious an effect the mutation has on gene function (Vaxillaire et al. 2012). The enzyme glucokinase (GCK) is expressed in multiple tissues, including the liver and the pancreatic beta. In the beta-cell, *GCK* is the main glucose sensor and plays a critical role in glucose homeostasis (Matschinsky et al. 1993). Mutations within *GCK* demonstrate how a spectrum of clinical phenotype exists depending on the type of mutation. Heterozygous mutations in *GCK* alter the glucose threshold for insulin secretion to a higher level and so affected individuals present with mildly elevated hyperglycaemia (Velho et al. 1992). Patients still have an adequate insulin response, and so many patients remain asymptomatic and require minimal treatment (Froguel et al. 1993; Hattersley et al. 1992). A recent investigation compared treated and untreated patients with GCK-MODY and found that there was no difference in HbA1c levels between the two groups, suggesting that the hyperglycaemia caused by the shift in glucose sensing did not respond to pharmacological treatment (Stride et al. 2014). However, either homozygous inactivating mutations or compound heterozygous mutations in *GCK* cause a severe PNDM phenotype and require insulin therapy (Njolstad et al. 2001), whilst activating mutations lead to neonatal hyperinsulinemic hypoglycaemia (Glaser et al. 1998).

Generally, the pathophysiology and phenotypes caused by mutations in the different MODY genes are quite different from one another and these influence management strategies (Owen and Hattersley 2001). For example, patients with HNF1A-MODY have been shown to be sensitive to the hypoglycaemic effects of sulphonylurea as it is able to avoid the defects in glucose metabolism caused by HNF1A deficiency (Pearson et al. 2003). Sulphonylurea binds to the sulphonylurea receptor subunit on the K_{ATP} channel, resulting in the channel's closure. Consequently, this induces a rise in membrane potential, causing the voltage-gated calcium channels to open, calcium to enter the β-cell and insulin to be released (Pearson et al. 2003). This targeted drug response by sulphonylurea, resulting in an increased pancreatic insulin secretory response, and the improved sensitivity to the insulin released highlights why it is necessary to identify the

type of diabetes. Therefore, correctly diagnosing HNF1A-MODY has direct implications on patient care and treatment.

A review by (Murphy et al. 2008) indicated that due to the distinct genetic aetiologies and consequential clinical heterogeneity it may be inappropriate to classify these subtypes of diabetes into one group. (Murphy et al. 2008) proposed that MODY subtypes should be sub classified according to the gene that is implicated in the disease. Nevertheless, the term MODY is still commonly used to describe this subgroup of diabetes.

1.5.2 Diagnostic and clinical presentation of MODY

The classical criteria used to identify patients with MODY are autosomal dominant inheritance (at least two generations), an age of onset usually before aged 45 years, insulin independence (insulin may be required for optimal control), absence of features of insulin resistance and absence of beta-cell autoimmunity (Thanabalasingham and Owen 2011). The subtype of MODY determines the clinical presentation and treatment response, so an accurate diagnosis has important implications for disease prognosis and family screening (Thanabalasingham and Owen 2011). Currently, the rate at which MODY is diagnosed is low and many cases are missed or misdiagnosed. This is attributed to several factors: it can be difficult to distinguish clinically different subtypes of MODY from common forms of diabetes; there could be the lack of physician awareness about MODY and the importance of making an accurate diagnosis; and the sequencing of every MODY gene is expensive (Shields et al. 2010).

1.6 The hepatocyte nuclear factor-1-alpha (*HNF1A*) gene

HNF1A (TCF1) located on chromosome 12q24.31 is a 10-exon gene that encodes for a full-length homeoprotein of 631 amino acids (Bach et al. 1990). The protein can act as either a homodimer or heterodimer with HNF1β. The consensus recognition site for HNF1A proteins is a 13-base-pair palindromic sequence GTTAATnATTAAC (Bach et al. 1990; Tomei et al. 1992; Vaxillaire et al. 1999). The protein consists of a dimerization domain (amino acids 1–33), a DNA binding domain consisting of two parts (a pseudo-POU-like domain, amino acids 100–184 and homeodomain amino acids 198–281) and a transactivation domain (amino acids 282–631) (Bjorkhaug et al. 2003; Ryffel 2001; Vaxillaire et al. 1999). Figure 1.6.1 is a schematic representation of *HNF1A*, illustrating which exons encode its dimerization, DNA binding and transactivation domains. HNF1A encodes three different isoforms (HNF1A (A) to HNF1A (C)) generated by alternative

Gene	Protein	Inheritance	Phenotype/Syndrome	References
HNF4A	Hepatocyte nuclear factor-4-alpha	Autosomal dominant	Adolescent or early adulthood familial onset diabetes Macrosomia, neonatal hypoglycaemia Linked to T2D risk	(Yamagata et al. 1996a) (Winckler et al. 2007)
GCK	Glucokinase	Autosomal dominant/recessive	Mild non-progressive hyperglycemia (a frequent cause of PNDM (homozygote carriers) Linked to T2D risk	(Froguel et al. 1993) (Winckler et al. 2007)
HNF1A (TCF1)	Hepatocyte nuclear factor-1-alpha	Autosomal dominant	Adolescent or early adulthood familial onset diabetes (a frequent cause of MODY 3) Linked to T2D risk	(Yamagata et al. 1996b) (Holmkvist et al. 2006) (Voight et al. 2010)
PDX1/IPF	Pancreatic/duodenum homeobox 1	Autosomal dominant	Adolescent or early adulthood onset diabetes Pancreatic agenesis (PNDM Rare)	(Stoffers et al. 1997)
HNF1B (TCF2)	Hepatocyte nuclear factor-1-beta Transcription factor	Autosomal dominant *De novo*	Diabetes in early adulthood TNDM/PNDM Developmental disorders effecting the pancreas, kidneys Linked to T2D risk	(Horikawa et al. 1997) (Voight et al. 2010)
NEUROD1	Neurogenic differentiation factor 1	Autosomal dominant/recessive	(Rare) Diabetes in early adulthood PNDM Cerebral hypoplasia, deafness, retinal dystrophy	(Malecki et al. 1999)
KLF11	Krueppel-like factor 11	Autosomal dominant	Adult onset diabetes	(Neve et al. 2005)
CEL	Carboxyl Ester Lipase (Bile salt-activated lipase)	Autosomal dominant	Progressive adult onset diabetes	(Raeder et al. 2006)
PAX4	Paired box protein Pax4	Autosomal dominant	Adult onset diabetes	(Plengvidhya et al. 2007)
INS	Insulin	Autosomal dominant	PNDM TNDM Progressive beta-cell destruction	(Edghill et al. 2008)
BLK	B lymphoid tyrosine kinase	Autosomal dominant	Adult onset diabetes	(Borowiec et al. 2009)
ABBC8	SUR1 Sulfonylurea receptor-1 Subunit of K_{ATP} channel	Mostly *de novo* or autosomal dominant	Reduced beta-cell function PNDM (most common) TNDM (less often) DEND (rare)	(Babenko et al. 2006)
KCNJ11	Kir6.2 Inward rectifier K+ Channel pore-forming subunit	Mostly *de novo* or autosomal dominant	Reduced beta-cell function PNDM (most common) TNDM (less often) DEND (rare) Linked to T2D risk	(Gloyn et al. 2004) (Bonnefond et al. 2012a)

Table 1.5.1 List of genes known to harbour mutations that cause MODY. PNDM, permanent neonatal diabetes mellitus; TNDM, transient neonatal diabetes mellitus; DEND, developmental delay, epilepsy, neonatal diabetes; T2D, type 2 diabetes.

splicing and differential use of polyadenylation sites. HNF1A (A) encodes for the full-length protein, HNF1A (B) exons 1–6 only and HNF1A (C) includes exons 1–7 (Bach and Yaniv 1993) (Figure 1.6.1). All three isoforms contain altered C-terminal domains and display distinct spatiotemporal expression patterns (Bach and Yaniv 1993; Harries et al. 2006).

All 3 isoforms have been shown to be active with HNF1A (B) and HNF1A(C) showing fivefold increase in transcriptional activity compared to HNF1A (A) (Bach and Yaniv 1993; Harries et al. 2006). Furthermore, previous studies have demonstrated no difference in DNA binding affinity or subcellular localisation between isoforms as this region is consistent across all three transcripts (Bach and Yaniv 1993; Bjorkhaug et al. 2005). Studies completed by (Harries et al. 2006), which looked at the expression of all three isoforms across different tissues, showed that although the total *HNF1A* expression is comparable amongst tissues there is a difference in the distribution of each isomer. For example, in the kidney and liver the major transcript is *HNF1A* (A), whereas in total pancreas and islets, *HNF1A* (B) was the predominant transcript. Differences in expression patterns have also been observed between fetal and adult pancreas. *HNF1A* (A) was most abundant in the fetal pancreas whilst *HNF1A* (B) showed the highest expression in the mature pancreas (Harries et al. 2006).

1.6.1 HNF1A and animal models

Animal models are a useful way of gaining insight into disease mechanism and pathogenesis. Most of our understanding of the complex interaction for *HNF1A* in the pancreatic beta-cell has arisen from work on mouse models. In general, humans demonstrate a diabetic phenotype with heterozygous mutations in *HNF1A* whilst heterozygous ($Hnf1a^{+/-}$) mice are indistinguishable from wild-type mice (Pontoglio et al. 1998). In humans, there has been no example of *HNF1A* homozygote mutations carriers and so is presumed to be lethal.

Mice with homozygous ($Hnf1a^{-/-}$) mutations display a severe phenotype resulting in diabetes, due to defects in insulin secretion as a result of errors in glycolytic signalling (Dukes et al. 1998; Pontoglio et al. 1998). Furthermore, in contrast to humans, mice also develop dwarfism, hepatic dysfunction and phenylketonuria (Lee et al. 1998; Pontoglio et al. 1996). Some mice exhibit progressive renal Fanconi Syndrome, causing death shortly after weaning (Pontoglio et al. 1996). It has been suggested that the

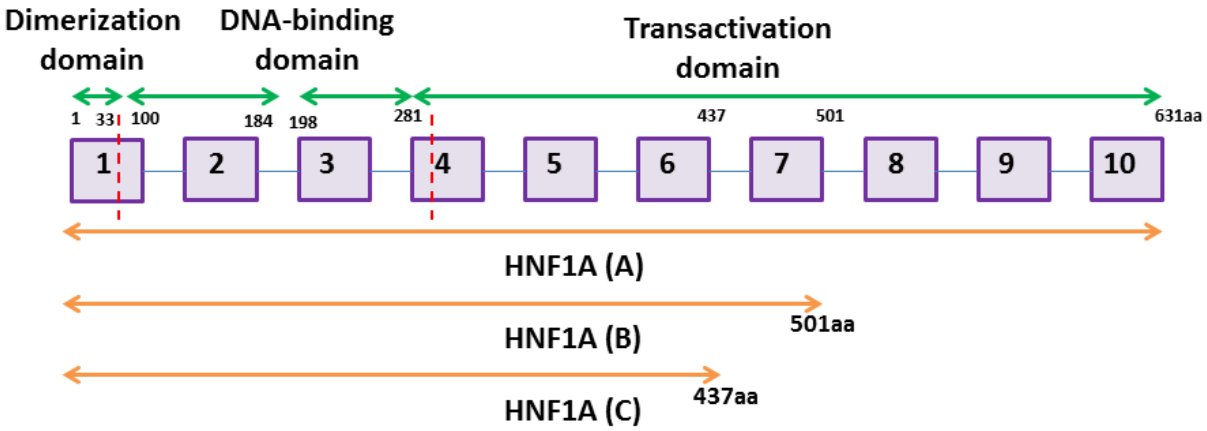


Figure 1.6.1 Schematic diagram of *HNF1A* illustrating the 3 functional domains and the 3 differential isoforms. Exons 1–6 are common to all 3 isoforms (HNF1A ABC), exon 7 is common to 2 isoforms (HNF1A AB), and exons 8–10 are present in only HNF1A (A).

discrepancies between phenotypes seen in rodent models and humans are due to differences in expression of *HNF1A* (Harries et al. 2009).

In comparison to humans, rodents only express a single isoform of *HNF1A (A)*. This may be explained by rodents not possessing the required genomic sequence in the area adjacent to the polyadenylation signal, and so once transcribed the partially processed *HNF1A* (B) and *HNF1A* (C) isoforms become unstable due to the lack of the poly-A tail (Harries et al. 2004). Furthermore, levels of *HNF1A* mRNA transcript have been found to be higher in rodents than in humans (Harries et al. 2009). Consequently, variants in *HNF1A* could alter the activity of the transcription network but the tolerance of the network might be species-specific due to the differential expression (Harries et al. 2009). As mutations in *HNF1A* are thought to cause haploinsufficiency, the higher levels in rat and mice tissues might be above the required threshold even in the heterozygous state and so do not manifest a diabetic phenotype (Harries et al. 2009).

1.6.2 HNF1A and the pancreatic transcription factor network

HNF1A belongs to a family of transcription factors, the hepatocyte nuclear family (HNF) that also includes genes encoding for *HNF1β*, *HNF4A* and *HNF6.* Transcription factors are regulatory proteins that have a key role in developing and maintaining cellular functions in mammalian cells. They contain a DNA binding domain, which recognises a specific DNA sequence (for the gene it regulates), thereby controlling transcription by acting as either an activator or a repressor (Ryffel 2001). When initially identified, this network of transcription factors was thought to be restricted to the liver and only regulate liver-specific genes (Bach et al. 1990). It is now well recognised that these transcription factors have a key role in other tissues, including the pancreas and kidney (Blumenfeld et al. 1991). Despite belonging to the same family and being expressed in the same tissues, the role and impact of these transcriptional regulators differ between tissues (Ferrer 2002).

Much of our understanding of the function of HNF1A comes from investigations using mice. Studies have shown that *Hnf1a* is expressed from early embryogenesis but, despite this, the absence of *Hnf1a* does not affect pancreatic development, and diabetes manifests mainly from young adulthood (Nammo et al. 2008). During early embryogenesis in mice, *Hnf1a* is mainly expressed in epithelial cells and, after birth, expression is predominantly localised to the exocrine cells with a lowered expression in islet cells (Nammo et al. 2008). The HNF1A transcription factor has a key role in the

glucose-stimulated insulin secretory response (GSIS), as it regulates the transcription of important genes such as the insulin gene. Furthermore, HNF1A is involved in the regulation of genes that encode for enzymes involved in the transport and metabolism of glucose, in addition to proteins required for normal beta-cell function (Fajans et al. 2001). The proposed mechanism for MODY is haploinsufficiency and that *HNF1A* has specific functions in the beta-cell that are vulnerable to decrease gene dosage (Harries et al. 2004; Thomas et al. 2002). Studies in mice have shown that the main function of Hnf1a is to ensure that all the key machinery required for beta-cells to respond to glucose and stimuli are all in one place (Ferrer 2002). All members of this family work together in a complex transcriptional network, for example, the relationship between *HNF1A* and *HNF4A*. In hepatocytes, transcription of *HNF1A* is controlled by HNF4A whilst in the pancreatic beta-cell, this relationship is reversed and *HNF1A* regulates *HNF4A* via a tissue-specific HNF4A-P2 promoter (Boj et al. 2001; Thomas et al. 2001).

1.6.3 Rare and common variants in HNF1A are associated with increased risk for diabetes

Genetic variants at different frequencies (common, low and rare) associated with susceptibility for diabetes have all been identified within *HNF1A* through family-based linkage studies, GWAS and genome sequencing (Flannick et al. 2013; Steinthorsdottir et al. 2014; Voight et al. 2010; Yamagata et al. 1996b). Identifying which variants are responsible for causing MODY and which are associated with an increased risk for T2D is important, as different variants are casual for different phenotypes and influences treatment and disease management for patients and their families.

Mutations in *HNF1A* are the most common cause of MODY and account for 52% of patients with a molecular diagnosis of MODY in the UK (Shields et al. 2010). A review in 2013 reported 414 mutations in *HNF1A*, with missense, nonsense and frameshift mutations all being described to be causal for MODY (Colclough et al. 2013). The highest proportions of mutations are found in exon 2 and 4, with the most frequent being the c.872dupC reported in at least 234 families. In addition, missense mutations are the most common in the dimerization and the DNA binding domain (Colclough et al. 2013). As described above, only the first six exons overlap across the three distinct isoforms. The marked difference in spatial and temporal expression patterns of the three isoforms has led to proposals that mutations affecting exons 8–10 of *HNF1A* may manifest in a milder diabetic phenotype due to compensation from other isoforms (Harries et al. 2006). Investigations performed by Harries et al. (2006) and Bellanné-Chantelot et al.

(2008) have demonstrated that the position and type of mutation within *HNF1A* can influence the age of onset of diabetes. For example, patients with missense mutations in exons 8–10 have shown to be diagnosed with diabetes 12 years later than in patients with missense sequence alterations in exons 1–6 (Bellanne-Chantelot et al. 2008; Harries et al. 2006).

Several population-specific rare (<1%), and low-frequency variants (1–5%) have been identified within *HNF1A* conferring risk for T2D. For example, the p.G319S allele has been shown to disrupt mRNA splicing, resulting in decreased HNF1A activity, and reduced insulin secretion (Harries et al. 2008). Homozygote carriers from the Canadian Oji-Cree population demonstrated a fourfold increase for T2D risk (Hegele et al. 1999b, 1999a; Triggs-Raine et al. 2002). Recently, a low-frequency variant, p.E508K in exon 8 of HNF1A, was associated with an increased T2D prevalence in the Latino population. Functional studies compared the transactivation potential of p.E508K alongside other *HNF1A* variants identified in patients with either T2D or HNF1A-MODY (Estrada et al. 2014). The results demonstrated that p.E508K had lower transcriptional activity than wild-type HNF1A and the variant identified in a T2D patient. In contrast, those variants associated with HNF1A-MODY had greater reductions in transactivation potential (Estrada et al. 2014).

1.6.4 Challenges in attributing function to variants

Early studies performed in small groups of patients with extreme phenotypes have successfully identified variants casual for disease phenotypes. However, this has led towards a bias for larger effect sizes to be reported, which may not be representative across the population (Begg 2002). The completion of studies using appropriately sized population-based control samples have now begun to change this, and more realistic effect sizes are now being determined for disease-causing variants (Flannick et al. 2013).

The development of powerful sequencing tools are now enabling whole exome and genome sequencing projects to be completed in adequately sized sample sets in participants with disease and those without. However, the clinical importance of many of these variants on human health and disease remains to be determined. Therefore, a current challenge is distinguishing between rare missense alleles that cause protein dysfunction and affect human health and those alleles that are neutral and have no detrimental impact on protein activity or disease status (Thormaehlen et al. 2015). Until

this can be improved, sequenced-based diagnostics remains elusive as it risks inappropriate medical interventions and decisions (Ormond 2008).

Furthermore, a recent study in which several MODY genes (including *HNF1A*) were sequenced in population cohorts found members of the general population to harbour the same rare coding variants previously implicated in causing MODY in subjects who remained euglycemic even at middle age (Flannick et al. 2013). This suggests that a rare variant in a MODY gene should not be assumed as causal for the diabetic phenotype, and supplementary evidence for determining pathogenicity of the variant will be necessary for some variants. Therefore, additional clinical, biochemical and functional investigations could be useful for measuring protein function and subsequently be used to understand an individual's risk for disease.

Exploring whether there are differences between biochemical profiles of patients with varying subtypes of diabetes might be a way of elucidating whether patients carry functional variants in certain genes associated with diabetes susceptibility. For example, for HNF1A-MODY, non-genetic biomarkers are currently being investigated to establish whether they can discriminate between HNF1A-MODY and other common forms of diabetes. If successful, biomarkers will assist in identifying and prioritising patients for genetic screening in the clinic. This will be discussed further in section 1.6.5 and in Chapters 5 and 6 of the thesis.

Functional characterisation of variants *in vitro* will be informative on whether the variants affect protein function. Chapters 5 and 6 of this thesis outline *in vitro* experiments performed to determine whether HNF1A variant proteins affect activity. Although these assays are informative, they would be difficult to perform in a rapid, high-throughput manor. Reports of high-throughput cell-based methodologies are now emerging, which are anticipated to address the issue of discriminating between disruptive versus non-disruptive missense alleles in a systematic unbiased manor (Majithia et al. 2014; Thormaehlen et al. 2015). Future studies are likely to include a combination of both sequencing and biological studies to understand the impact of novel variants uncovered by studies sequencing the genome (Thormaehlen et al. 2015).

1.6.5 The role of biomarkers in identifying HNF1A-MODY

Biomarkers are biological molecules that can be detected and measured, and can be used as an indicator of normal or diseased conditions or as a pharmacological response to therapeutic interventions (Strimbu and Tavel 2010). Prospective biomarkers should be

able to identify cases (sensitivity) and distinguish them from healthy individuals (specific), whilst being inexpensive and reliable for measuring.

The difficulties in clinically distinguishing different subtypes of MODY from common forms of diabetes and the high costs linked with genetic screening have encouraged the search for suitable non-genetic biomarkers for detecting HNF1A-MODY. It is likely that biomarkers might be used in the clinic as a screening tool to prioritise patients for genetic testing as an adjunct to conventional clinical diagnostic criteria. A definitive molecular diagnosis for HNF1A-MODY has important implications for the patient and other family members. Patients can be successfully treated and managed with oral low-dose sulphonylueras improving overall quality of life (Pearson et al. 2003; Shepherd et al. 2009), whilst other family members with diabetes can be screened.

Several efforts using a variety of biochemical profiles to differentiate between subtypes of diabetes has been explored with limited success so far. Urinary metabolic profiling of MODY and young onset T2D patients did not discriminate between HNF1A, other MODY subtypes and T2D (Gloyn et al. 2012). Cystatin C, which is used as a marker for glomerular filtration rate (GFR) and kidney function, was unsuccessful at discriminating for HNF1A-MODY (Nowak et al. 2013). Studies investigating Apolipoprotein M (apoM) have shown that apoM was unable to discriminate between HNF1A-MODY and T2D (Cervin et al. 2010). One investigation showed that apoM can distinguish between T1D and HNF1A-MODY but requires further investigation as a clinical diagnostic tool (Mughal et al. 2013). Findings from GWAS have highlighted novel biological pathways that are being explored as potential biomarkers for HNF1A-MODY and these provide the basis of Chapters 5 and 6 of the thesis.

Several independent GWAS have described that common variation near *HNF1A* resulted in small alterations in levels of highly sensitive C-reactive protein (hsCRP) (Okada et al. 2011; Reiner et al. 2008; Ridker et al. 2008). Since then, studies performed in European cohorts have reported that levels of hsCRP are reduced in patients with HNF1A-MODY compared to those with other forms of diabetes and non-diabetic controls (McDonald et al. 2011; Owen et al. 2010; Thanabalasingham et al. 2011). The evaluation of whether hsCRP is a good marker for HNF1A-MODY in non-European cohorts is still needed. Chapter 6 of this thesis explores whether hsCRP is a biomarker that can be used in South Asians.

Following the result of a GWAS, which revealed that *HNF1A* plays a role in the regulation of fucosylation (Lauc et al. 2010), it was hypothesised that patients with mutations in *HNF1A* may have altered levels of plasma glycans. Chapter 5 of this thesis explores glycans in more detail and describes the studies used to validate glycan profiles as a potential biomarker for HNF1A-MODY, and then goes on to select patients based on biomarker levels to screen for variants in *HNF1A.* These variant proteins are then functionally characterised to evaluate whether they result in *HNF1A* dysfunction.

A recent study reported that two circulating microRNAs, miR-103 and miR-224, are upregulated in the serum of patients with HNF1A-MODY (Bonner et al. 2013). MicroRNAs (miRNAs) are non-coding RNA molecules, typically 19–25 nucleotides in length, and are involved in regulating the cleavage or translational repression of target mRNAs (Bartel 2004). (Bonner et al. 2013) demonstrated that it is possible to discriminate between HNF1A-MODY carriers and individuals with T2D based on differences in miR-103 expression between the two groups. Their study demonstrated that HNF1A-MODY patients have a sixfold increase in mean expression levels of miR-103 compared to T2D patients. To date, this is the first study to report miR-103 as a potential biomarker for HNF1A-MODY but further clinical studies are required in much larger cohorts to validate the findings. The analysis of miRNAs could be a practical and economically feasible option for clinical translation (Bonner et al. 2013). However, at present, there are severe limitations linked with measuring miRNAs robustly. For example, assessing the quality of the miRNA isolated from biological preparations will be important as degrading mRNA can interfere with the quantification of miRNAs (Etheridge et al. 2011). Furthermore, a lack of standardization of protocols makes it difficult to validate results across different platforms. These challenges need to be addressed before clinical translation can be considered (Etheridge et al. 2011).

At present, biomarkers are not being used routinely in the clinic as a screening tool for HNF1A-MODY. Each of the above-mentioned biomarkers has advantages and disadvantages that still need to be overcome. For example, hsCRP is particularly useful as it is cheap and widely available. However, hsCRP levels are also influenced by inflammation and thus may be elevated because of concurrent disease. Glycan profiles seem not to be affected by inflammation but currently no high-throughput screen is available. Chapters 5 and 6 of this thesis explore whether glycan profiles and hsCRP, respectively, can be used as a measure of HNF1A function, which could then be used to as discriminatory tool for HNF1A-MODY in the clinic.

1.7 Aims of thesis

This Introduction has outlined how the application of ever-increasing powerful genetic and genomic methodologies is transforming our understanding of genetic variation and the role in T2D predisposition. Although more difficult than anticipated, functional studies for interrogating statistically associated variants are now providing valuable insights into molecular mechanisms and biological pathways that contribute to disease pathophysiology. Furthermore, it is hoped that these findings will help lead to clinical translation, to guide appropriate diagnosis and therapeutic decisions for at-risk individuals.

The work presented in this thesis uses a combination of approaches to understand how common and rare variants influence disease susceptibility. Chapters 3 and 4 investigate common variants in individuals of South Asian ancestry, an ethnic group with an increased risk for T2D. Functional studies in Chapters 5 and 6 evaluate the potential of biomarkers as a clinical tool for identifying rare variants in a gene implicated in causing young onset diabetes. More specifically, it will:

- Explore the contribution to T2D predisposition in South Asians, of known common risk variants, most of which have been identified in Europeans, and to determine whether risk is shared or ethnic specific (Chapter 3);
- Identify novel T2D risk variants in South Asian populations through collaborative genome-wide association studies (Chapter 4);
- Determine whether non-synonymous variants of uncertain pathogenicity identified in patients with biomarker profiles suggestive of HNF1A dysfunction have altered transcription function using *in vitro* assays (Chapter 5); and
- Functionally characterise HNF1A variant proteins identified in patients using hsCRP and evaluate whether hsCRP is a good marker for diagnosing diabetes in South Asian individuals (Chapter 6).

Chapter 2

Materials and Methods

2.1 Subjects studied

2.1.1 Sri Lankan type 2 diabetes (T2D) case-control samples

Participants genotyped as part of the studies described in Chapters 3 and 4 were ascertained from two previously published independent collections from Sri Lanka (Katulanda et al. 2008b; Katulanda et al. 2008a).

The control subjects were participants in the Sri Lankan Diabetes Cardiovascular Study (SLDCS), a cross-sectional, nationally representative, epidemiological investigation that recruited 4388 subjects (40% male) across seven Sri Lankan provinces (Katulanda et al. 2008a). DNA collection was initiated partway through the study and DNA samples were available for 1769 subjects. Among recruited subjects, normoglycaemia was determined based on the results of a 75g oral glucose tolerance test (OGTT), interpreted using American Diabetes Association (ADA) and World Health Organisation (WHO) criteria (FPG ≤5.6mmol/L or 2-hr post-OGTT plasms glucose was <7.8mmol/L) (Alberti and Zimmet 1998). The mean (standard deviation, SD) age of the SLDCS participants was 45.2 (14.7) years with a body mass index (BMI) of 21.2 (3.9) kgm^{-2}.

The T2D cases were mainly derived from the Sri Lankan Young Diabetes Study (SLYDS) (Katulanda et al. 2008b). Cases (n = 1001, 44% male) were recruited from private and government diabetes clinics. Age of diagnosis for diabetes was between 16 and 40 years and all participants were under the age of 45 years at recruitment. The mean (SD) age of the SLYDS subjects was 36.7 (5.8) years and BMI was 24.7 (4.0) kgm^{-2}. Within these individuals, T2D status was defined if glutamic acid decarboxylase autoantibodies (GADA) titre was ≤14units/ml, and if the interval between diagnosis and the initiation of insulin therapy was at least six months (Katulanda et al. 2008b). The SLYDS cases were augmented with 176 T2D cases from SLDCS. The mean (SD) age for these 176 T2D cases was 49.7 (12.9) years, with a mean BMI of 23.7 (3.9) kgm^{-2}. Additional exclusion criteria were applied to the combined set of cases that included ages ≥80 years and missing sex information (Katulanda et al. 2008a).

At the time of recruitment, participants in both studies were classified according to major ethnic groups present in Sri Lanka (Sinhalese, Sri Lankan Tamil, Indian Tamil, Muslim, Burgher, other) based on a combination of spoken language, religion/cultural identification and surname. The Ethical Review Committee of the University of Colombo approved participant collection. All participants provided informed written consent (Katulanda et al. 2008b; Katulanda et al. 2008a).

A summary of the clinical characteristics of subjects genotyped for each discrete study is provided in the relevant chapters.

2.2 DNA extraction from whole blood

DNA extractions were performed primarily for the control (SLDCS) samples used in Chapters 3 and 4. Prior to extraction, whole blood samples were stored at -80°C. DNA was extracted from whole blood samples using the automated Maxwell 16 system (Promega, Madison, WI, USA) with a whole blood purification kit following manufacturer guidelines. The kit utilises magnetic particle-based methodologies allowing purification of genomic DNA. Samples were passed through the kit twice with DNA being eluted into 300µl of elution buffer each time. Working stock DNA was stored at 4°C; and stock DNA were stored at -20°C.

2.3 Genomic DNA amplification

2.3.1 Primer design

Primer sequences for *HNF1A* amplification used in Chapter 5 were ordered based on previously published sequences (Ellard et al. 2008). Primers were designed for sequencing purposes and all had additional universal M13 tags appended to the 5′ end for sequencing ease (forward $^{5'}$tgtaaaacgacggccagt$^{3'}$, reverse $^{5'}$caggaaacagctatgacc$^{3'}$). All primers were purchased from Eurofins MWG Operon (Ebersberg, Germany).

Details of all primers are described below or in the Methods section of experimental Chapter 5.

2.3.2 Polymerase chain reaction (PCR)

As part of experimental Chapter 5, the polymerase chain reaction (PCR) was used to amplify fragments of *HNF1A* genomic DNA that were subsequently used to perform Sanger sequencing with dye-terminator chemistry. Unless stated, standardised in-house PCR conditions were employed using AmpliTaq Gold (Applied Biosystems, Warrington, UK). Each reaction contained 40ng of patient DNA, 10× PCR buffer, 25mM magnesium chloride (Applied Biosystems), 8mM deoxynucleoside triphosphates (dNTPs) (Promega UK Ltd, Southampton, UK), 5pM forward and reverse primers (Operon), 0.1U AmpliTaq Gold (Applied Biosystems) and sterile water. The complete reagent master mix and volumes used for one PCR reaction are detailed in Table 2.3.1.

All PCR reactions were performed in 96-well, skirted plates (ThermoScientific, Cramlington, UK) and amplified on a PTC-225 Peltier thermal cycler (MJ Research,

Waltham, MA, USA) using the PCR conditions listed in Table 2.3.2. On completion of the PCR, samples were stored at -20°C until required.

2.3.3 PCR product visualisation using agarose gel electrophoresis

To confirm successful DNA amplification of the required PCR products, all samples were checked on an agarose gel electrophoresis. A standard 2% agarose gel was made by dissolving 5g of agarose (Sigma Aldrich Company Ltd, Gillingham, UK) into 250ml of Tris/borate/ethylenediaminetetraacetic acid (TBE) buffer (NBS Biologicals, Huntingdon, UK).

To assist dissolution, the mixture was carefully heated using an 800W microwave for 2 minutes. Once cooled, but prior to solidification, 5µl of the intercalating agent ethidium bromide (Sigma Aldrich) (0.002% of the final solution) was added to the agarose-TBE mixture. This was then poured into an appropriately sized gel electrophoresis tank containing correctly spaced well combs and allowed to solidify at room temperature. Once set, the tank was filled with 1× TBE buffer. Equal volumes of PCR product (5µl) were mixed with orange-G loading buffer (4g sucrose in 10ml distilled water (dH_20) and 20mg Orange G [Sigma Aldrich]) and briefly vortexed. PCR samples with orange-G were then loaded onto the gel alongside 6µl of a 100bp ladder (New England Biolabs Ltd, Hitchin, UK) or 1 kilobase (kb) DNA ladder (Promega UK Ltd), depending on the size of the PCR product to be resolved. For amplicon sizes of 300–600bp, gels were run at 120V for approximately 45 minutes; however, the voltage and duration were increased for larger sized PCR products. Gels were viewed using ultra violet light (UV) on a Gel Doc 2000 illuminator system and Quantity One v4.3.1 software (BioRad, Hemel Hempstead, UK).

2.4 Capillary DNA sequencing

Sequencing was performed to identify nucleotide changes in patient samples that had biological biomarker (glycan and hsCRP) profiles indicative of genetic alterations within the *HNF1A* gene (Chapters 5 and 6). The sequences of the PCR-amplified products were obtained via a modified version of the chain-termination (Sanger) sequencing method using the Applied Biosystems BigDye v3.1 Cycle sequencing kit. The kit is composed of fluorescently labelled dideoxynucleotide triphosphates (ddNTPs), with each of the four ddNTPs emitting light at different wavelengths. In addition to ddNTPs, each sequencing reaction contains unlabelled dNTPs, DNA polymerase, PCR product template, and product-specific primers.

Reagent	Volume (µl)
10× PCR buffer II	2.5
25mM Magnesium chloride	1.5
8mM dNTPs	1.0
5pM Forward primer	1.0
5pM Reverse primer	1.0
AmpliTaq Gold	0.1
Distilled water	15.9
DNA (20ng/µl)	2.0

Table 2.3.1 PCR reagents and volumes used in a standard 25µl reaction volume (*n* = 1). These volumes represent the standard conditions. Any variations on these are described in the relevant chapters.

Section	Number of cycles	Temperature (°C)	Time (minutes)
DNA denaturation	1	95	15
		94	1
Primer annealing and	35	62	1
extension		72	1
Elongation	1	72	10

Table 2.3.2 Standard PCR thermal cycler reaction conditions. Any deviations from these are stated in the relevant experimental chapters.

During the sequencing reaction, the ddNTPs become incorporated into the elongating DNA strand; however, as they lack a 3'-hydroxyl group they are unable to form a phosphodiester bond with subsequent nucleotide, thus terminating elongation of the DNA strand. The different fluorescence emission spectrum of each ddNTP allows the premature chain termination at each nucleotide to be determined by capillary electrophoresis (Rosenblum et al. 1997; Sanger and Coulson 1975; Sanger et al. 1977; Smith et al. 1986). Sequencing reactions involved three distinct steps that are described below.

2.4.1 Clean-up of PCR products

Exonuclease 1 (Exo) (New England Biolabs) and Thermosensitive Alkaline Phosphatase (TSAP) (Promega UK Ltd) were pooled in a 1:1 volume ratio equivalent to 10 and 0.5U/µl of each enzyme, respectively. One microliter of Exo-TSAP was added to 1µl of PCR product and centrifuged briefly. Samples were then incubated at 37°C for 30 minutes to allow the enzymes to remove any residual PCR contaminants (including excess primer and dNTPs), after which samples were heated to 80°C to inactivate the enzymes.

2.4.2 Sequencing reactions

The sequencing reaction contained 2µl of the cleaned-up PCR-Exo-TSAP mix, along with 8µl of mastermix, described in Table 2.4.1. Sequencing primers were complementary to the M13 tails added to the PCR primers (section 2.1.1).

Samples were spun down, then placed on a 96-well thermal cycler to undergo a sequencing reaction, as described in Table 2.4.2.

2.4.3 Sequencing product clean-up

On completion of the sequencing reaction, samples were centrifuged at 1000*g* for 2 minutes, after which 5µl of distilled water was added to each reaction, giving a total volume of 15µl. To remove any residual sequencing reagents, samples were passed through a 96-well Performa DTR V3 short clean-up plate (VH Bio, Gateshead, UK) following manufacturer guidelines.

The combined sequencing products and water were passed through the columns of the plate by centrifugation (850*g* for 5 minutes). Eluent was collected in an empty 96-well non-skirted plate (ThermoScientific) and heated for 90 minutes at 70°C to remove any water using a thermocycler.

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 2.4.1 DNA sequencing reagent for one reaction (*n* = 1). Sequencing mix and dilution buffer were supplied from Applied Biosystems as part of the BigDye v3.1 cycling sequencing kit.

Section	Number of cycles	Temperature (°C)	Time
Denaturation	1	96	10 seconds
Primer annealing	25	50	10 seconds
and extension		5	4 minutes

Table 2.4.2 Thermal cycler conditions for DNA sequencing. Any deviations from these are stated in the relevant experimental chapters.

2.4.4 Capillary electrophoresis and analysis

Samples were sequenced using capillary electrophoresis (ABI 3730xl DNA Analyser, standard run module, 15-second injections, v1.1 dye set and analysis module) at the university's core sequencing facility based at the Department of Zoology. Fluorescence signals emitted were converted to .ab1 output files, which were then analysed and compared with appropriate reference sequences using Mutation Surveyor v3.4 (Soft Genetics, Cambridge, UK).

2.5 Genotyping (Applied Biosystems TaqMan 7900 HT technologies)

DNA samples were genotyped using TaqMan technology (Applied Biosystems). Each assay contained primers, allowing specific amplification of the genomic region of interest in the presence of differentially labelled fluorescent probes (VIC/FAM) specific to the two possible alleles. The genotyping methodologies described below were used for studies performed in results Chapters 3, 4, and 5.

2.5.1 Assay design

All genotyping assays were predesigned and commercially available from Applied Biosystems (Life Technologies, Paisley, UK) where possible. The assays are designed to align with the genome precisely targeting the allele of interest. If no assay was commercially available, a custom assay was designed using relevant genomic sequences from the National Centre for Biotechnology Information (NCBI) and Primer Express v3.0 (Applied Biosystems). Details of the genotyping assays are provided in the relevant experimental Chapters 3, 4 and 5.

2.5.2 TaqMan PCR

Firstly, 10ng of DNA in a maximum volume of 2µl was dried down into 384 well plates (4titude, Wotten, UK) by incubating at 80°C for 10 minutes. Secondly, genotyping PCRs were performed in total reaction volume of 2µl. This comprised genotyping mastermix (containing polymerase and dNTPs; KBiosciences, Hoddesdon, UK), genotyping assay and distilled water. The volumes for one reaction are given in Table 2.5.1.

Samples were briefly centrifuged and then heated on the PTC-225 Peltier thermal cycler (MJ Research) using the conditions listed in Table 2.5.2.

2.5.3 Genotype determination

After PCR completion, samples were briefly centrifuged. The genotype calls were then determined through a post-read allelic discrimination on an Applied Biosystems 7900HT machine (Life Technologies) using SDS v2.3 software and standard analysis settings. All genotypes were assigned according to pre-programmed algorithms but were also visually checked. Some genotype re-assignments were made; these were independently agreed between a colleague (Dr Amanda Bennett, OCDEM, University of Oxford) and myself. Results were discarded if more than 5% of samples required re-assignment. Non-template control (NTC) wells containing dH_20 were also included on each plate. Results were also discarded if the NTCs showed amplification.

2.6 Site-directed mutagenesis

Site-directed mutagenesis (SDM) was performed to introduce specific nucleotide changes in the cDNA sequence of *HNF1A* to investigate the functional impact of genetic variants identified in experimental chapters 5 and 6. Primer sequences were designed using the QuikChange Primer Design (Agilent Technologies Ltd, Berkshire, UK), which is a specific programme for designing mutagenic primers. Reactions were performed using the QuikChange II site-directed mutagenesis kit (Agilent Technologies) according to the manufacturer's guidelines. Multiple reactions were set up using 50ng of template DNA. Table 2.6.1 lists all the reagents contained in the reaction mixes and Table 2.6.2 lists the thermal cycler conditions used for the SDM and template amplification. On completion of the PCR, the SDM reactions were cooled on ice for 2 minutes. Samples were then subjected to *Dpn1*-mediated digestion. This process removed any template DNA that had not been mutagenised. Before samples were digested with *Dpn1*, 10µl of the 50µl SDM reaction product was removed and placed in a clean 0.2ml PCR tube (Alpha Laboratories Ltd, Eastleigh, UK). These samples were labelled 'undigested PCR product'.

2.7 Transformation of competent *Escherichia coli* (*E. coli*)

Plasmid DNA was introduced into competent *E. coli* cells through transformation. Reactions were performed in 1.5ml microcentrifuge tubes that were pre-chilled on ice before commencement of the procedure. Five microlitres of plasmid DNA was added to 50µl of XL1-blues supercompetent cells (Stratagene, Agilent Technologies) for SDM reactions. For routine transformations, DH5-α competent cells (Life Technologies) were used.

Reagent	Volume in reaction (µl)
Genotyping Master Mix	1.00
Genotyping assay	0.02
Distilled water	0.98

Table 2.5.1 Genotyping reagents (*n* = 1). All volumes were scaled up linearly.

Section	Number of cycles	Temperature (°C)	Time
Polymerase activation	1	95	10 minutes
Primer annealing and	40 or 50 cycles	92	15 seconds
Extension		60	1 minute

Table 2.5.2 Thermal cycler conditions for genotyping. Any deviations from these standard conditions are provided in the relevant experimental chapters.

Reactions were gently mixed by flicking the tube and then incubated on ice for 30 minutes. Following incubation, reactions were transformed via 'heat-shock'. Tubes were placed on a pre-heated heat block set at 42°C for 45 seconds and then immediately transferred to ice for a minimum of 2 minutes to allow the cells to recover. To obtain maximal transformation efficiency, 500µl of super optimal broth (SOC) media was added (Sigma Aldrich) to each tube and the reactions were incubated for 90 minutes at 37°C, in an orbital incubator at 250rpm (Stuart Orbital Incubator S150, Jencons Scientific Ltd, Bedfordshire, UK). Post-incubation, transformation reactions were spun down at 2938*g* using a benchtop centrifuge (Force micro 1618, Labnet International, USA) for 1 minute. Three hundred microlitres of supernatant was carefully removed and discarded, the remaining pellet was then resuspended, and 100µl was added to and spread across agar-ampicillin plates. Plates were incubated overnight at 37°C. The next day plates were checked for bacterial colonies, wrapped in parafilm and stored at 4°C.

2.8 Preparation of agar-antibiotic culture plates

Agar-antibiotic plates were prepared by adding 17.5g of agar (Sigma Aldrich) into 500ml of distilled water in a sterile 1L glass conical flask (Fisher Scientific Ltd, Loughbourgh, UK). Flasks containing the agar solution were then sterilised by heating to 120°C for 20 minutes followed by 60°C for 1 hour in a steam autoclave (Priorclave Ltd, London, UK). Once completed, the flasks were allowed to cool at room temperature. Once below 35°C, ampicillin at a concentration of 100mg/ml was added to the agar-sterilised solutions. The flasks were mixed by gentle 'swirling' and then poured into sterile plastic culture plates (Appleton Woods Ltd, Birmingham, UK) and allowed to set at room temperature as a ~1cm-deep gel. Once solid, the plates were sealed with parafilm laboratory sealing tape (Fisher Scientific Ltd, Loughborough, UK) and stored at 4ºC.

2.9 DNA extraction from transformed *Escherichia coli*

A commercially available strain of bacteria, SCS1 *Escherichia coli*, was routinely used to amplify plasmid DNA (Stratagene, Europe). This was used because of both the endonuclease and recombination-deficient properties, resulting in less degradation of extracted DNA and increased stability of the transformed plasmid, respectively.

Reagent	Volume added to reaction (µl)
10× reaction buffer	5.0
dNTPs	1.0
Forward mutagenesis primer (125nM)	1.0
Reverse mutagenesis primer (125nM)	1.0
Pfu Turbo (2.5U/µl)	1.0
dsDNA template (5–50ng)	As appropriate
ddH_20	Final volume up to 50µl

Table 2.6.1 SDM reactions (*n* = 1). 10× buffer, dNTPs, and Pfu Turbo were all supplied with the QuikChange II kit (Agilent Technologies Ltd, UK). Primers were ordered from Eurofins MWG Operon (Ebersberg, Germany). All reactions were performed in 0.2ml PCR tubes (Alpha Laboratories, Eastleigh, UK).

Segment	Cycles	Temperature	Time
DNA denaturation	1	95°C	2 minutes
Primer annealing	18	95°C	1 minute
		55°C	1 minute
Extension		68°C	7 minutes
End		4°C	10 minutes

Table 2.6.2 SDM thermal cycler conditions. Any deviations from these standard conditions are provided in the relevant experimental chapters.

Transformed colonies, as described in section 2.7, were picked from agar-antibiotic plates, and used to inoculate 10ml LB-antibiotic medium and cultured overnight at 37°C, in an orbital incubator at 250rpm (Stuart Orbital Incubator S150). The next day, bacteria were pelleted at 14,000*g* for 10 minutes in a Force 1618 desktop microcentrifuge (Labnet International, Rutland, UK). Post centrifugation, the supernatant was discarded and DNA was extracted using the QIAprep Spin Miniprep kit according to manufacturer's guidelines with elution of DNA in 50µl of distilled water (QIAgen Ltd, Crawley, UK).

2.10 DNA quantification

A Nanodrop 2000 (ThermoScientific) spectrophotometer, set to the nucleic acid-DNA-50 programme, was used to measure the concentration of DNA by measuring the sample absorbance at a wavelength of 260nm. In addition, sample purity was assessed by measuring the absorbance of protein and phenol/organic compounds at 280 and 230nm, respectively. Prior to commencement, the spectrophotometer was 'blanked' using 1.2µl nuclease-free water, as this contained everything except the compound (DNA) of interest. Thus, this blanking meant that any subsequent measured absorbance was due to the presence of DNA. DNA concentration and purity were then determined by measuring 1.2µl of each sample.

2.11 Restriction digest of DNA

All restriction digests performed contained the same basic reagents listed in Table 2.11.1. Unless otherwise stated, all reactions were incubated at 37°C for 2 hours. Restriction enzymes and their related buffers were obtained from either New England Biolabs UK or Promega (Madison, WI, USA). Details of specific restrictions on enzymes used are provided in the results chapters 5 and 6.

2.12 Plasmid DNA sequencing

All SDM reactions performed in experimental chapters 5 and 6 were verified by Sanger sequencing to confirm that nucleotide mutagenesis had occurred at the desired position and no other aberrant nucleotide insertions or deletions were present. Primers used for sequencing were designed manually to span the *HNF1A* complementary DNA (cDNA) sequences. Primer pairs were selected based on their length (18–28bp), melting temperature (55–65°C), size of amplicon generated (300–700bp), guanine and cytosine (GC) content (40–60%) and whether they fell outside polymorphic sites. All primers were purchased from Eurofins MWG Operon. All sequencing samples were sent to Source Bioscience (Nottingham, UK). All samples were sequenced with primers SP6 and

M13 forward supplied by Source Bioscience and the relevant *HNF1A* primers. Primer sequences are listed in Table 2.12.1.

2.13 General maintenance of cell lines and seeding cells for experiments

In vitro assays functionally evaluating *HNF1A* variants described in experimental chapters 5 and 6 were performed in HeLa and INS-1 cells. These two cell lines were chosen as HeLa cells lack endogenous *HNF1A* whilst INS-1 cells express endogenous *HNF1A*. HeLa cells are a human cell line derived from cervical cancer cells (Jones 1997; Lucey et al. 2009). Cells were split twice weekly (1:10) and maintained in T75 or T175 flasks (Sarstedt, Leicester, UK). Cells were grown under the following conditions: Dulbecco's Modified Eagle Medium (DMEM) media (Gibco, Life Technologies) supplemented with 10% (*v/v*) foetal calf serum (Gibco, Life Technologies) and 1% (*v/v*) penicillin/streptomycin (Life Technologies).

All cells were grown and incubated at 37°C and 5% CO_2. INS-1 cells are a rodent insulinoma cell line (Asfari et al. 1992). Again, cells were split twice weekly (1:5) and maintained in T75 flasks (Sarstedt). Cells were grown under the following culturing conditions: Roswell Park Memorial Institute-1640 (RPMI-1640) media (Sigma Aldrich) supplemented with 10% (*v/v*) foetal calf serum, 1% penicillin/streptomycin, 1% (*v/v*) L-glutamine (Life Technologies), 1% (*v/v*) sodium pyruvate, 1% HEPES and 0.1% (*v/v*) β-mercaptoethanol (Life Technologies). All cells were grown and incubated at 37°C and 5% CO_2. Cells were seeded 24 hours prior to carrying out transient transfections. Cell densities for transactivation studies, western blot analysis, and immunofluorescence are detailed in the relevant sections below.

Briefly, media was removed from the flask and cells were washed with 1× phosphate buffered saline (PBS) (Sigma Aldrich). For a T75, 1ml of TrypLE Express (Life Technologies) was added to the flask after which cells were then incubated at 37°C for ~3–5 minutes. Cells were then viewed under the tissue culture microscope (Nikon Eclipse T5100) to check cells had detached. To neutralise the TrypLE-Express, 5ml of media (specific to the cell line) was mixed with the cells in the flask. The cells were then transferred to a 15ml falcon tube (Sarstedt) and pelleted by spinning at 240*g* for 5 minutes using a benchtop centrifuge (Beckman GS-15R [S4180], Beckman Coulter, High Wycombe, UK).

Reagent	Volume in reaction (µl)
Template DNA	5.0
Bovine serum albumin (BSA) (10×)	0.5
Restriction enzyme	1.0
Restriction enzyme buffer (10×)	2.0
Distilled water	To make final volume 20µl

Table 2.11.1 Restriction digest reagents for one reaction (*n* = 1). Any deviations from these standard conditions are provided in the relevant experimental chapters.

Primer label	Primer Sequence
SP6 promoter	5′-atttaggtgacactatag-3′
M13 Forward	5′-tgtaaaacgacggccagt-3′
HNF1A_F1	5′-ggtttctaaactgagccagc-3′
HNF1A_F2	5′-gtggcgaagatggtcaagtc-3′
HNF1A_F3	5′-aggaggagcgagagacgcta-3′
HNF1A_F4	5′-tagtgacagtgtctacacc-3′
HNF1A_F5	5′-cagttctcccagccgctgca-3′
HNF1A_F6	5′-ttcacacgccggcatctcag-3′

Table 2.12.1 Primer sequences for primers used to verify SDM. Primers SP6, M13 forward, HNF1A_F2, HNF1A_F4 and HNF1A_F6 in most cases covered the entire *HNF1A* cDNA region.

On completion, the supernatant was removed and the cell pellet was re-suspended in 5ml of fresh media. Cells were counted using a cell counter (Nexcelom Bioscience, Cellometer Auto T4, Sarisbury Green, UK), then the required number of cells were plated, depending on the requirements for the experiment.

2.14 Transient transfections

Transient transfections were completed for all *in vitro* assays described in experimental chapters 5 and 6. All transfections were completed in biological safety cabinets in the tissue culture facilities at OCDEM. Transient transfections were performed using Lipofectamine 2000 transfection reagent following manufacturer guidelines. The ratio of DNA to Lipofectamine differed for each cell line. HeLa cells used a 1:3 ratio and INS-1 cells used a 1:5 (DNA: Lipofectamine). Transfections were performed 24 hours after cells were plated. Prior to transfection, culture media corresponding to the appropriate cell line was replaced on cells with media containing no foetal calf serum or antibiotics. Five hours post transfection culture media was removed and replaced with fresh complete media for the appropriate cell line. Cells were incubated at 37°C at 5% CO_2 and cells were harvested 24 hours after transfections.

2.15 Transactivation studies

Chapters 5 and 6 focus on evaluating the transcriptional activity of wild-type and variant HNF1A. This was assessed by measuring the activity of the human albumin promoter, human HNF4A P2 promoter and human CRP promoter as reporter constructs using the Dual Luciferase Assay System (Promega, Madison, WI, USA). Full details of constructs are provided in relevant experimental chapters. All reporter assays were performed in two cell lines, one lacking endogenous *HNF1A* (HeLa cells) and the other expressing endogenous *HNF1A* (INS-1 cells).

HeLa cells (5.5×10^4) and INS-1 cells (3.5×10^5) were plated (as described in section 2.13) in 24 well plates (Sarstedt) 24 hours prior to transfection. Cells were transfected using Lipofectamine 2000 reagent (Life Technologies) together with 0.25µg of either wild-type or mutant pRK5-HNF1A, 0.5µg of promoter/pGL3-basic the luciferase reporter construct (Promega, Madison, WI, USA) and 0.01µg of pRL-SV40 (Promega, Madison, WI, USA), which was used as the internal control vector. To make up the final plasmid DNA composition to 1µg per well, 0.24µg of pGL3-empty vector was used as a balance control.

Transcriptional activity was measured 24 hours post transfection using the Dual Luciferase Assay System (Promega, Madison, WI, USA). Briefly, media was removed from the plates and cells were lysed using 100µl per well of 1× passive lysis buffer (PLB) which was provided as part of the luciferase assay kit. Plates were placed on an orbital shaker for 15 minutes to aid lysis. Once lysed, samples were kept on ice. The luciferase assay was performed in white 96-well plates (Corning, Appleton Woods) using 20µl of cell lysate per well. Seventy-five microliters of Luciferase Assay Reagent (LAR II) and 75µl of Stop & Glo Reagent was used per well (Promega, Madison, WI, USA). All luciferase experiments were read using the Verities Microplate Luminometer (Promega Madison, WI, USA).

A dual reporter system was used as it utilises the distinct bioluminescent properties of firefly and *Renilla* luciferase, allowing for the simultaneous expression and measurement of the reporter constructs. For all three constructs (albumin, HNF4A-P2 and CRP) the proximal promoter sequences were cloned into a promoter-less pGL3-basic vector which contains the firefly luciferase gene. This was known as the 'experimental' construct; its activity is the result of HNF1A overexpression, and transactivation of one of the promoter constructs. *Renilla* is constitutively expressed, and acts as an internal experimental control. Luminescence values for the experimental control were normalised to *Renilla* values, thus minimising experimental variability due to differences in cell viability and transfection efficiency.

Three different *Renilla* constructs of varying promoter strength were tested to find the most appropriate one for this experimental system: herpes simplex virus (HSV)-thymidine kinase (pRL-TK), a relatively weak promoter; the early Simian virus 40 (SV40) enhancer/promoter region (pRL-SV40); and the strongest cytomegalovirus (CMV) immediate early enhancer/promoter region (pRL-CMV). All three were assessed in HeLa cells. Results of these assays indicated that the early SV40 enhancer/promoter region (pRL-SV40) generated luciferase values of a similar magnitude to those observed for the experimental construct, and so was selected for all further transactivation assays. At least three independent experiments on consecutive days were performed in triplicate for each cell line and promoter. Results were calculated as percentage transcriptional activity by normalising to the wild-type HNF1A protein.

2.16 Protein expression studies

Expression of variant HNF1A protein is assessed in Chapters 5 and 6 using cell lysates collected from the transactivation studies, and using prepared cytosolic and nuclear protein extracts. An in-frame N-terminal FLAG tag was used to detect wild-type and variant HNF1A protein expression. Further details of the HNF1A construct are provided in Chapter 5, section 5.2.9. β-Tubulin was used as a control for protein expression across all samples.

2.16.1 Transactivation assay lysate western blot analysis

Protein expression studies were performed using the cell lysates collected from the transactivation assays. Representative western blots were completed using 10µl of cell lysate collected from one triplicate well from two distinct luciferase assays. Together, 10µl of cell lysate, 4µl of 4× loading buffer (Bio-Rad Laboratories Ltd, Hertfordshire, UK) and 1µl of β-mercaptoethanol (Sigma Aldrich) were heated to 72°C for 15 minutes and then loaded on to a Bio-Rad gel as described below in section 2.16.4.

Membranes were simultaneously incubated overnight at 4°C with two primary antibodies, the first monoclonal anti-FLAG (Sigma Aldrich) used at 1:50,000 concentration to detect HNF1A protein expression and the second β-Tubulin (Santa Cruz, CA, USA) 1:2000. The next day following 4× 10-minute washes in Tris buffered saline Tween (TBST), the membrane was first blocked, using goat anti-rabbit IgG secondary antibody, horseradish peroxidase (HRP) conjugate (Thermo Scientific, Rockford, IL, USA) at 1:2500 concentration, with shaking at room temperature. The membrane was then visualised as described in section 2.16.4. Subsequently, after 3× 10-minute washes with TBST the membrane was blocked with rabbit anti-mouse IgG (H+L) secondary antibody HRP conjugate (Thermo Scientific) at a 1:2500 concentration for 1 hour at room temperature. Finally, the membrane was visualised using a Bio-Rad imager.

2.16.2 Preparation of cytosolic and nuclear protein extracts

HeLa cells (2.5×10^5) and INS-1 cells (1.2×10^6) were transfected with Lipofectamine 2000 along with 2.5µg of either wild-type or mutant pRK5-HNF1A construct. Twenty-four hours post transfection whole cells extracts were collected. The protocol was performed on ice to preserve protein integrity as much as possible. Briefly, media was removed and cells were washed in 1000µl 1× PBS (Sigma Aldrich). Cells were then incubated with 500µl of Tris ethylenediaminetetraacetic acid (EDTA) NaCl (TEN) for 5 minutes. Using cell scrapers, cells were harvested and collected into labelled pre-chilled 1.5ml

microcentrifuge tubes. Collected cells were spun down at 80*g* (Allegra 64R centrifuge, Beckman Coulter) for 10 minutes, at 4°C. Afterwards, supernatant was removed without dislodging the pellet. Pellets were snap frozen in dry ice before storing at -80°C until ready for fractionation.

Pellets were separated into cytoplasmic and nuclear fractions as described previously (Lees and Whitelaw 1999). Slight modifications to the protocol were made which included the removal of Nonidet P-40-Ficoll lysis buffer from the hypotonic buffer. Detergent-free hypotonic buffer composed of 10mM HEPES (pH 7.9), 1.5mM $MgCl_2$, 10 mM KCl and 10% glycerol (*v/v*). The nuclear extract buffer consisted of 20 mM HEPES (pH 7.9), 1.5mM $MgCl_2$, and 0.5mM EDTA, 0.42M KCl and 20% glycerol. Both buffers, once made, were stored at 4°C. On the day of fractionation, the required volume of each buffer was aliquoted, and to each buffer fresh 1mM dithiothreitol (DTT) [1µl of 1M/mol/L/ml] and protease inhibitor cocktail [10µl of 100× stock/ml] (Sigma Aldrich) was added.

Pellets were resuspended in 3× pellet volumes of hypotonic buffer and then incubated on ice for 5 minutes, then centrifuged (Allegra 64R centrifuge, Beckman Coulter) for 30 minutes at 10,960*g*, 4ºC. The supernatant was transferred to pre-chilled 1.5ml microcentrifuge tubes and stored as the cytosolic fraction. To remove any cytoplasmic contamination, pellets were washed in 1× pellet volume of hypotonic buffer and centrifuged for 10 minutes at 10,960*g*, 4°C. This was repeated four times, discarding the supernatant each time. Next, pellets were resuspended in 2× pellet volumes of nuclear buffer. Tubes were incubated for 45 minutes on a rotating wheel at 4°C. Tubes were then centrifuged for 30 minutes, 10,960*g*, 4°C. The supernatant was transferred to pre-chilled 1.5ml microcentrifuge tube and stored as the nuclear fraction. Both fractions were snap frozen on dry ice before storing at -80°C.

2.16.3 Protein quantification

Cytosolic and nuclear protein samples were quantified prior to western blot analysis (Chapter 5). Protein quantification was performed using standard Bradford reagent (Bio-Rad Laboratories) and colorimetric assays. The assay utilises the change in absorbance of the Coomassie brilliant blue dye when bound to protein. When bound to protein, the Bradford reagent stabilises the blue form of the dye; this can then be measured at an absorbance of 595nm, allowing an estimation of the protein concentration in the complex. To produce a standard curve, sixteen 0–2.2mg/ml protein standards were

made up using BSA (Sigma Aldrich) and 1× PBS. Each sample was diluted 1:10 for cytoplasmic fractions and 1:4 for nuclear fractions to ensure sample concentrations fell in the middle of the standard curve. Enough Bradford reagent was made up for a 96-well plate, by diluting 1 volume Bradford reagent to 5 volumes distilled water. Using the same tips, 190µl of Bradford reagent was added to each well. Ten microlitres of protein standards and samples was added in duplicate on a 96-well spectrophotometer plate, and absorbance at 595nm recorded using a Versamax microplate reader and Softmax Pro v5.2 software (both Molecular Devices Ltd, Wokingham, UK).

2.16.4 Western blot analysis

All protein samples were detected using western blot analysis (Chapters 5 and 6). The Bio-Rad Chemi Doc system was used to run western blots (Bio-Rad Laboratories). Prior to western blotting, all cytosolic and nuclear samples were diluted to a starting concentration of 5µg to ensure equal protein loading in all wells. To 10µl of protein lysate 5µl of 4× loading buffer (Bio-Rad Laboratories) and 1µl of β-mercaptoethanol (Sigma Aldrich) were added. Samples were heated to 75°C for 15 minutes to ensure reduction and denaturing of the proteins. Subsequently, heated samples were spun down for 1 minute at 2040*g* using a benchtop centrifuge (Force micro 1618; Labnet International).

A Bio-Rad Criterian TGX stain-free pre-cast gel (1.0 mm, 26 wells) was set up using a Bio-Rad western blot chamber. One litre of running buffer (1× Tris/glycine/sodium dodecyl sulphate (SDS) (TGS) buffer consisting of 25mM Tris, 192mM glycine and 0.1% SDS pH8.3) was used to fill a medium-sized tank. At least one lane per gel was used for a molecular weight marker (Precision Plus, All blue standards, Bio-Rad laboratories). Using gel-loading tips, each sample was loaded. Gels were run for approximately 25 minutes at 300V.

Prior to transfer of proteins to the membrane, the gel was activated using the Chemi Doc Imager (Bio-Rad Laboratories). Proteins were transferred to polyvinylidene difluoride (PVDF) membrane using the Bio-Rad Trans-Blot Turbo pack, and the Turbo 7 Minute programme. Following the transfer process, the membrane was viewed using a Chemi Doc Imager using the stain-free blot setting to ensure that protein transfer from gel to membrane had occurred.

Membranes were blocked with 20ml 5% non-fat milk powder and TBST for 1 hour before being incubated overnight at 4°C with monoclonal anti-FLAG (Sigma Aldrich) 5%

milk powder and TBST. The following day, membranes were washed 4 times for 10 minutes with TBST before being blocked for 1 hour using rabbit anti-mouse IgG (H+L) secondary antibody HRP conjugate (Thermo Scientific). After washing for 5 × 5 minutes with TBST, membranes were incubated for 5 minutes using clarity western ECL substrate (Bio-Rad Laboratories). Membranes were then visualised using a Chemi Doc Imager.

As a quality control for the fractionation process membranes were probed with two additional antibodies. For cytosolic fractions, the primary antibody anti-GAPDH (Santa Cruz, CA, USA) was used and for the nuclear fractions, membranes were probed with Topoisomerase IIα antibody (Cell Signalling Technology, Danvers, MA, USA). These antibodies were probed on subsequent days and all images were captured using the Chemi Doc imager. Specific antibody concentrations are provided in the relevant experimental chapters.

2.17 Subcellular localisation-Immunofluorescence

Immunostaining was performed to determine the subcellular localisation of each *HNF1A* variant discovered in experimental chapters 5 and 6. HeLa cells (5.5×10^4) and INS-1 cells (3.5×10^5) were seeded on Falcon chamber slides (VWR International Ltd, Northern Ireland). Cells were transfected with Lipofectamine 2000 and 0.8µg of either wild-type or variant HNF1A. Cells were fixed with 500µl of 4% paraformaldehyde 24 hours post transfection. To permeabilize the cell membranes, cells were incubated with 200µl PBS-Triton for up to 1 hour. Cells were then blocked using 200µl of 10% BSA-phosphate buffered saline Tween (PBST) for 30 minutes. An overnight incubation at 4°C with 200µl primary antibody, monoclonal anti-FLAG [1:2000] (Sigma Aldrich) together with 10% BSA-PBST was performed. The next day, cells were washed 5 × 5 minutes with 200µl PBST per well. The next stage of the procedure was performed in the dark. Cells were blocked for 1 hour with 200µl of secondary antibody [1:100] anti-mouse Alexa Fluor-488 labelled with green fluorescence. Cells were then washed with 200µl PBST 5 × for 5 minutes. To stain the nucleus, cell permeant nuclear stain bright far-red (NucBlue Live 647, Thermo Fisher Scientific) was used as per manufactures guidelines. Slides were then sealed using commercially available nail varnish; slides were then stored at 4°C. All images were viewed using a confocal microscope (Nikon model, Eclipse, E600FN) using Bio-Rad coloured filters.

2.18 *In silico* prediction tools

In silico prediction tools were used to evaluate whether the *HNF1A* variants identified in Chapters 5 and 6 affected protein function. The programmes used were Polymorphism Phenotype [http://genetics.bwh.harvard.edu/pph2/](Adzhubei et al. 2010), Sorting Intolerant From Tolerant (SIFT) [http://sift.jcvi.org/] (Kumar et al. 2009) and CONsensus DELeteriousness score of missense SNVs (Condel) [http://bg.upf.edu/fannsdb/] (Gonzalez-Perez and Lopez-Bigas 2011). Both PolyPhen and SIFT make predictions by comparing sequence homology of related proteins and by looking at the degree of conservation through evolution at the position of change. Additionally, PolyPhen uses UniProt [http://www.uniprot.org/] entries to assess if the amino acid substitution is located in important structural or functional sites (Flanagan et al. 2010; Ng and Henikoff 2001, 2003). Condel integrates the results of five different computational tools (including PolyPhen and SIFT) and calculates an average weighted score based on results of all the programmes (Gonzalez-Perez and Lopez-Bigas 2011).

2.19 Statistical Analysis

In result chapters 3 and 4, assessment of association between each individual genotyped single nucleotide polymorphism (SNP) and T2D status was conducted using logistic regression assuming a log additive model. All associations were adjusted for covariates age and sex. All analyses were conducted in PLINK v1.07 (http://pngu.mgh.harvard.edu/~purcell/plink/ ; Purcell et al. 2007).

In experimental chapters 5 and 6, an analysis of variance (ANOVA) was applied to compare mean fold activation between wild-type HNF1A and each variant. Average luminescence values for each transactivation experiment were log transformed to normalise the data prior to running a univariate analysis. An ANOVA was used as it allowed one single test to compare multiple *HNF1A* variants to wild-type. IBM SPSS v20.0 was used to perform all analysis. Furthermore, an unpaired student t-test was completed to compare differences in protein levels for wild-type and variant HNF1A. Statistical significance was defined as $P<0.05$ for all tests.

Chapter 3

Evaluation of common type 2 diabetes risk variants in South Asian samples of Sri Lankan descent

Below is the reference to the publication that the results presented in this chapter contribute. The complete manuscript is provided in the appendix section:

- Hassanali, N., et al. (2014), 'Evaluation of common type 2 diabetes risk variants in a South Asian population of Sri Lankan descent', *PLoS One,* 9 (6), e98608.

3.1 Introduction

3.1.1 Investigating type 2 diabetes (T2D) in South Asians

Over the past three decades the prevalence of T2D in South Asians has shown a particularly dramatic increase (Gujral et al. 2013; Jayawardena et al. 2012), prompted by profound changes in socioeconomic factors and lifestyle. It is estimated that almost 72 million adults have diabetes and 24 million have impaired glucose tolerance (IGT) in the South East Asia region (Ramachandran et al. 2014). It is widely accepted that T2D is a complex disorder and individual risk reflects the influence of environmental factors on a background of genetic predisposition.

Compared to European counterparts, South Asians tend to be diagnosed with diabetes earlier, with a lower BMI and display a more rapid decline in glycaemic control over time (Jayawardena et al. 2012). The increased prevalence of T2D extends to South Asian groups living outside their native countries and this suggests that there may also be an underlying biological predisposition in addition to environmental and lifestyle factors (Jayawardena et al. 2012; Saxena et al. 2013).

The advent of large-scale genome-wide association studies (GWAS) has led to the identification of over 120 genetic loci that contribute to T2D risk (Dupuis et al. 2010; Frayling et al. 2007; Johansen et al. 2010; Kooner et al. 2011; Lyssenko et al. 2009; Mahajan et al. 2014; Morris et al. 2012; Saxena et al. 2012; Scott et al. 2007; Sladek et al. 2007; Steinthorsdottir et al. 2007; Zeggini et al. 2007; Zeggini et al. 2008). Most of the early studies were conducted in Europeans, but increasingly, similar approaches are being deployed in samples of South Asian, East Asian and African origin (Chauhan et al. 2010; Chidambaram et al. 2010; Kooner et al. 2011; Mahajan et al. 2014; Orozco et al. 2010; Palmer et al. 2012; Rees et al. 2011; Sanghera et al. 2008; Saxena et al. 2012). These studies have revealed novel signals (Kooner et al. 2011; Saxena et al. 2013; Tabassum et al. 2013), but have also shown appreciable overlap with associations first discovered in European groups (Chauhan et al. 2010; Chidambaram et al. 2010; Rees et al. 2011; Sanghera et al. 2008).

In this chapter, I extend these studies to individuals from the island of Sri Lanka. As elsewhere in South Asia, the incidence of T2D is increasing and it is predicted that by 2030

approximately 14% of the adult population will have the condition, many of them undiagnosed (Whiting et al. 2011).

3.1.2 Diabetes in the Sri Lankan population

Sri Lanka is an island situated in the Indian Ocean with a population of over 20 million (Jayawardena et al. 2014). The major ethnic group are the Sinhalese, comprising almost 72% of the population. This is followed by the Sri Lankan Tamils and Indian Tamils, which due to significant intermarriage are very similar in their sociocultural practises. The ethnic groups the Burghers, Moors and Malays are descendants of intermarriages between the Sinhalese and Tamils to Dutch, Portuguese and Arab traders (Amarasekara et al. 2014; Kshatriya 1995; Papiha et al. 1996).

Similar to other South Asian communities, the number of cases of T2D is increasing in Sri Lankans with an estimated national prevalence of 10.3% with the frequency of affected people growing in urban areas, particularly amongst the educated and highest earners (Jayawardena et al. 2012; Pubudu De Silva et al. 2012). It is estimated that within Sri Lanka one in five of the population has diabetes or prediabetes (Katulanda et al. 2008a). Studies investigating the occurrence of obesity, hypertension and metabolic syndrome have all indicated that there is an increased prevalence of these diseases in the adult Sri Lankan population compared with European counterparts (Katulanda et al. 2010; Katulanda et al. 2012; Katulanda et al. 2014). Together, all of these health complications increase risk for T2D or are a result of secondary complications due to diabetes. A study examining health trends in Sri Lanka projected that in 2010 there will be an increase in the proportion of patients hospitalised due to non-communicable diseases; for example, 36% of patients were being hospitalised due to diabetes (Jayasekara and Schultz 2007). It is expected that this will place an inevitable burden on the countries health care system, especially as these diseases require long-term management. Therefore, preventative strategies to minimise disease risk are necessary (Jayasekara and Schultz 2007). Despite the growing incidence of T2D, relatively little is known about the genetic predisposition of T2D in Sri Lankans, and whether the risk of T2D is similar to that of other South Asian regions.

3.1.3 Experimental aims

The aims of this study were

- To test known T2D risk variants in Sri Lankan samples to establish their role in T2D in Sri Lankans; and

- To determine if there are any systematic differences in allele frequency or odds ratio that could contribute to differences in prevalence or presentation for T2D.

3.2 Methodology

3.2.1 Study samples

A detailed cohort description is provided in Chapter 2 (section 2.1.1). Cases were from the Sri Lankan Young Diabetes Study (SLYDS). Of the 1001 cases, 965 had DNA samples available for genotyping. Twenty-nine samples were excluded due to missing GADA data, 48 for having insulin treatment within 6 months of diagnosis and 10 individuals for a positive diagnosis of mitochondrial diabetes (mt3243 A>G;), leaving 878 T2D cases available for inclusion (Figure 3.2.1) (Katulanda et al. 2008b).

Control subjects were participants in the Sri Lankan Diabetes Cardiovascular Study (SLDCS). Of the available DNA samples 1523 individuals who were confirmed as normoglycaemic based on oral glucose tolerance data (interpreted according to then-current ADA and WHO criteria), and under the age of 80 years were included in this study (Figure 3.2.1) (Alberti and Zimmet 1998).

Table 3.2.1 provides a summary of the clinical characteristics of the participants available for genotyping.

3.2.2 DNA Extraction of Sri Lankan control samples

Prior to genotyping, I extracted DNA from whole blood for the control samples. Extraction was performed using the automated Maxwell 16 system whole blood purification kit as described in Chapter 2 (section 2.4).

3.2.3 SNP selection and Genotyping

Thirty-seven T2D risk single nucleotide polymorphisms (SNPs) were selected for genotyping based on studies published as of mid-2010. All chosen risk SNPs had previously reached genome-wide significance (p=5.0×10^{-8}) in previous GWAS, mainly in studies using samples of European ancestry (McCarthy 2010; Voight et al. 2010). All assays were predesigned by Applied Biosystems and I carried out the genotyping using Applied Biosystems TaqMan 7900 HT technology.

Characteristics	SLYDS T2D subjects	SLDCS normoglycaemic subjects
Number of males/females (*n*)	359/519	653/870
Age (years)	37.3 (5.3)	44.3 (14.2)
Age at diagnosis (years)	32.5 (5.3)	N/A
Weight (kg)	62.8 (12.2)	51.9 (10.8)
Height (cm)	158.2 (10.7)	156.5 (8.7)
BMI (kg/m^2)	25.0 (3.9)	21.2 (3.9)
Waist circumference (cm)	88.2 (9.3)	74.5 (10.9)*
FPG (mmol/L)	9.1 (3.5)	4.7 (0.5)

Table 3.2.1 Summary of the clinical characteristics of subjects available for genotyping. All values are mean (standard deviation) except (*n*). BMI: body mass index; FPG: fasting plasma glucose. *Waist circumference data were only available for 1522 control individuals.

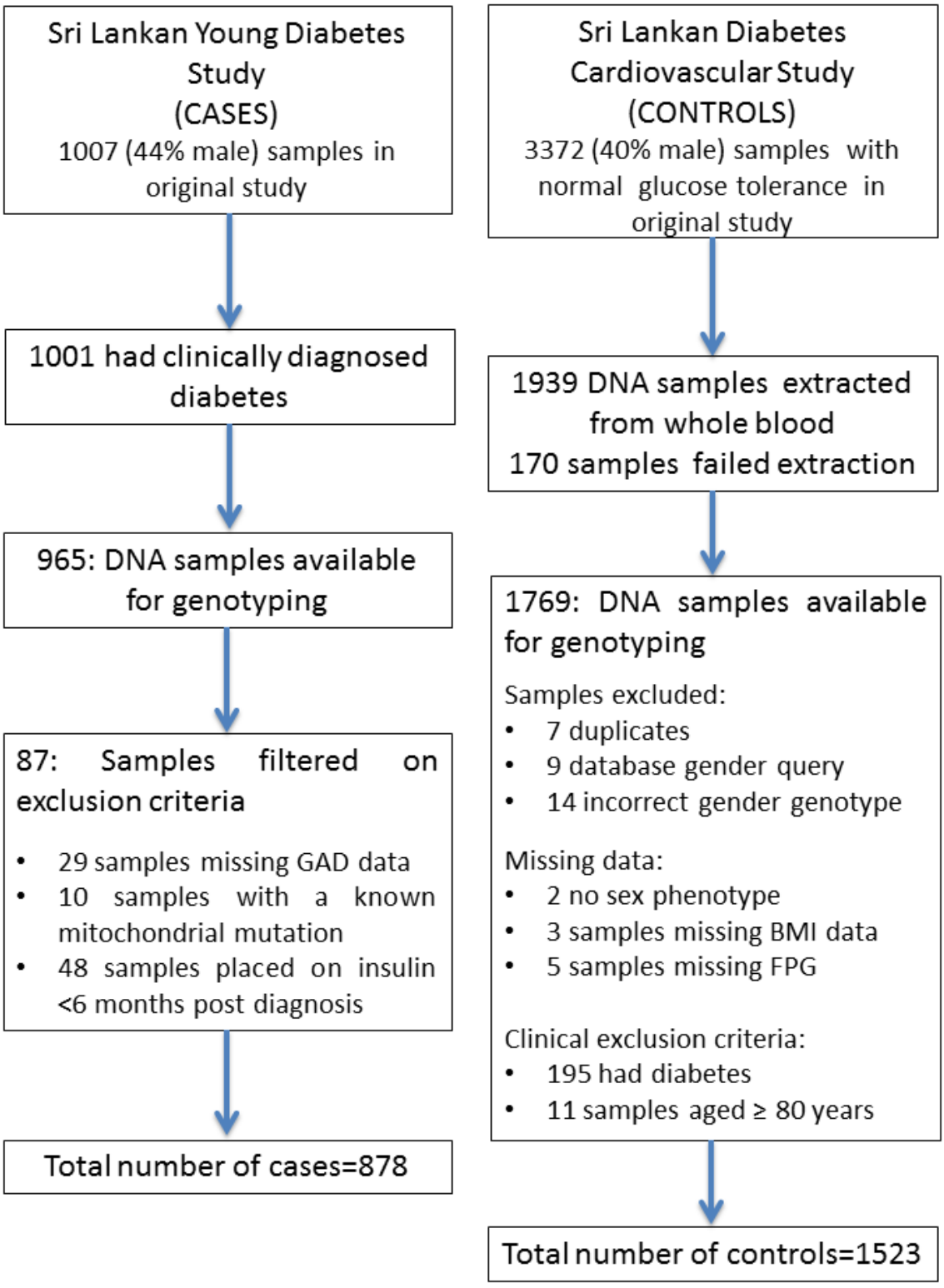


Figure 3.2.1: Sri Lankan studies: Exclusion criteria applied to samples used in the replication of common type 2 diabetes risk variants in a Sri Lankan population study.

A summary of the selected SNPs is shown in Table 3.2.2. A total of 878 T2D cases and 1523 normoglycaemic controls were genotyped for each SNP. Each assay was labelled with fluorescent probes (Vic/Fam) to allow for discrimination between the two possible alleles. A PCR reaction was completed and genotype data were determined as described in Chapter 2 (section 2.5). On completion of genotyping for all SNPs data were sent to the Wellcome Trust Centre of Human Genetics where colleagues Dr Will Rayner and Neil Robertson (WTCHG, University of Oxford) collated the genotyping data for each SNP and generated PED and MAP files that I used for analysis. PED files comprised the genotyping information for all the samples and the MAP files contained the SNP information and location.

As part of the quality check for the genotyping, genotyping call rate was checked for each individual. In addition, each SNP was assessed for departure from Hardy Weinberg equilibrium (HWE) in the control samples. Any SNP with a $P \leq 0.001$ in control samples was excluded from further analysis.

3.2.4 Statistical analysis

Firstly, a logistic regression was used to assess the association between each individual SNP and T2D status assuming a log additive model. All associations were adjusted for age and sex. In addition, the case–control data were reanalysed after stratifying for self-identified ethnic subgroup using the Cochran–Mantel–Haenszel (CMH) test. The individual SNP association analyses were undertaken in PLINK v1.07 (http://pngu.mgh.harvard.edu/~purcell/plink/ ; Purcell et al. 2007).

3.2.5 Genetic risk score

The current study was part of a collaborative effort with colleagues at the University of Exeter. Once I completed the genotyping and the PED and MAP files were generated these data were sent to Dr Maneka De Silva, where she completed the analysis for the genetic risk scores (GRS). Briefly, logistic regression was used to test the significance of GRS that combines the information from the 36 T2D-associated SNPs. The SNPs were coded as 0, 1 or 2, corresponding to the number of T2D risk-increasing alleles in Europeans, except for the X chromosome SNP rs5945326 at *DUSP9*, where male genotypes were coded as 0 or 1, and female genotypes as 0, 0.5 or 1 (to reflect random X inactivation). These analyses were completed using Stata/SE version 10.1 for Windows (StataCorp, Brownsville, TX). The GRS were created by using individuals with genotypes available from at least 29 of the 36 type 2 diabetes SNPs (i.e. 80% of the SNPs genotyped).

GENE	CHR	SNP	Assay ID	Method of identification in Europeans	Reference
NOTCH2	1	rs10923931	C___1188816_10	GWA meta-analysis	Zeggini et al. (2007)
PROX1	1	rs340874	C___1022779_10	GWAS MAGIC	Voight et al. (2010)
THADA	2	rs7578597	C__32653841_10	GWA meta-analysis	Zeggini et al. (2007)
BCL11A	2	rs243021	C___2363509_10	GWAS Diagram	Voight et al. (2010)
IRS1	2	rs7578326	C__29270128_10	GWA	Voight et al. (2010)
GCKR	2	rs780094	C___2862873_10	GWAS MAGIC	Dupuis et al. (2010)
PPARG	3	rs1801282	C___1129864_10	Candidate gene study	Altshuler et al. (2000)
IGF2BP2	3	rs4402960	C___2165199_10	GWA	Zeggini et al. (2007)
ADAMTS9	3	rs4607103	C__26129590_10	GWA meta-analysis	Zeggini et al. (2008)
ADCY5	3	rs11708067	C___8989067_10	GWAS MAGIC	Dupuis et al.(2010)
WFS1	4	rs10010131	C__30473796_10	Candidate gene study	Sandhu et al. (2007)
ZBED3	5	rs4457053	C__26380786_10	GWAS Diagram	Voight et al. (2010)
CDKAL1	6	rs10946398	C__31635230_10	GWA	Zeggini et al. (2007)
JAZF1	7	rs864745	C___7601116_20	GWA meta-analysis	Zeggini et al. (2008)
KLF14	7	rs972283	C___7614445_10	GWAS Diagram	Voight et al. (2010)
GCK	7	rs4607517	C__27917691_10	GWAS MAGIC	Dupuis et al.(2010)
DGKB-TMEM195	7	rs2191349	C__16153073_20	GWAS MAGIC	Dupuis et al. (2010)
SLC30A8	8	rs13266634	C____357888_10	GWA	Sladek et al. (2007)
TP53INP1	8	rs896854	C___8792773_10	GWAS Diagram	Voight et al. (2010)

CDKN2A/B	9	rs10811661	C__31288917_10	GWA	Zeggini et al. (2007)
CHCHD9	9	rs13292136	C__11871167_10	GWAS Diagram	Voight et al. (2010)
TCF7L2	10	rs7903146	C__29347861_10	Linkage study	Grant et al. (2006)
HHEX/IDE	10	rs1111875	C__11214581_10	GWA	Sladek et al. (2007)
CDC123/CAMK1D	10	rs12779790	C__26360618_10	GWA meta-analysis	Zeggini et al. (2008)
KCNJ11	11	rs5219	C__11654065_10	Candidate gene study	Gloyn (2004)
KCNQ1	11	rs2237892	C__16171025_10	GWA	Yasuda et al. (2008)
KCNQ1	11	rs231362	C___3075844_1_	GWAS Diagram	Voight et al. (2010)
MTNR1B	11	rs10830963	C___3256858_10	GWA	Prokopenko (2009)
CENTD2	11	rs1552224	C___1953903_10	GWAS Diagram	Voight et al. (2010)
TSPAN8/LGR5	12	rs7961581	C____121473_10	GWA meta-analysis	Zeggini et al. (2008)
HMGA2	12	rs1531343	C___1692218_10	GWAS Diagram	Voight et al. (2010)
HNF1A	12	rs7957197	C__29094181_10	GWAS Diagram	Voight et al. (2010)
ZFAND6	15	rs11634397	C__11527673_10	GWAS Diagram	Voight et al. (2010)
PRC1	15	rs8042680	C___2999711_10	GWAS Diagram	Voight et al. (2010)
FTO	16	rs9939609	C__30090620_10	GWA	Frayling et al. (2007)
HNF1B (TCF2)	17	rs757210	C____543332_10	Candidate gene study	Winckler et al. (2007)
DUSP9	X	rs5945326	C__11232180_10	GWAS Diagram+	Voight et al. (2010)

Table 3.2.2 Summary of lead SNPs chosen for genotyping. Genotyping assays were acquired from Applied Biosystems (Life Technologies). All SNPs were identified in studies using European samples with the exception of SNP rs2237892 at the KCNQ1 loci.

3.2.6 Collation of reported odds ratio and allelic frequency from previous studies in European populations

To compare effect size estimates for Sri Lankan case–control samples with those observed in Europeans, I compiled odds ratio (OR) estimates for each locus for European case–control data from the literature (Altshuler et al. 2000; Dupuis et al. 2010; Frayling et al. 2007; Gloyn et al. 2004; Grant et al. 2006; Prokopenko et al. 2009; Sandhu et al. 2007; Voight et al. 2010; Zeggini et al. 2007; Zeggini et al. 2008). To minimise inflation of these estimates in initial genome-wide association discovery samples (the "winner's curse"), I used OR values from replication samples wherever possible (Kraft 2008). Allele frequencies were acquired using data from the CEU (Utah residents with Northern and Western European ancestry) component of the International HapMap Project.

3.2.7 Power calculations

Quanto was used to calculate power under assumptions of a log-additive model, a disease prevalence of 10% and a significance threshold (α) of 0.05. Power was calculated for each individual SNP using allelic odds ratios (for European case–control comparisons) collated from the literature and effect allele frequencies from the CEU component of HapMap (http://biostats.usc.edu/Quanto.html).

3.3 Results

3.3.1 Genotyping of T2D risk loci

Seventy-four samples with a low (<80%) overall genotyping call rate were removed from further analysis, leaving 2327 individuals (830 cases, 1497 controls) for final analysis. The average genotyping call rate for these 2327 individuals was 97%. All SNPs were assessed for departure from Hardy Weinberg equilibrium. All 37 SNPs were in HWE except for the SNP rs2237892 in the *KCNQ1* locus (P<0.001 in controls) which was removed from subsequent analyses.

3.3.2 Individual associations for the 36 T2D risk loci

Individual SNP associations for the 36 SNPs in 830 T2D patients and 1497 controls are summarised in Table 3.3.1. Nominal associations with T2D ($P \leq 0.05$) were observed for 16 of the 36 SNPs tested when adjusted for age and sex. The strongest effects were seen at *TCF7L2* (OR 1.38 [95% CI 1.21–1.58], $p=2.8\times10^{-6}$) and *SLC30A8* (OR 1.32 [95% CI 1.14–1.55], $P=3.9\times10^{-4}$). The observation of *TCF7L2* demonstrating the strongest association is consistent with previous reports in both European and other South Asian subgroups (Mahajan et al. 2014; Morris et al. 2012), with the effect size observed for *TCF7L2* in the Sri

Lankan samples (OR 1.38) being comparable to that observed in Europeans (OR 1.37) (Grant et al. 2006).

At 15 of the 16 loci that showed nominal significance, the direction of effect was consistent with that reported in the original discovery study, with the exception being rs7961581 (*TSPAN8/LGR5:* OR 0.87 [95% CI 0.77–0.99], *P*=0.049). In all, 31 of the 36 SNPs tested showed evidence of association that was directionally consistent with previous studies in Europeans (i.e. the same allele increases T2D risk in the two populations) (binomial $P=3.2\times10^{-6}$).

Given some case–control imbalance with respect to 'self-identified' ethnic group, case–control analyses were repeated in stratified samples using the CMH test. The results of the CMH test are summarised in Table 3.3.1.

The results obtained were broadly comparable: 14 of the 16 loci were also associated (*P*<0.05) in the CMH test with *TCF7L2* still showing the strongest association (OR 1.35 [95% CI 1.19–1.54], $P=6.1\times10^{-6}$). Two loci *IRS1* (*P*=0.023) and *JAZF1* (*P*=0.049) were no longer statistically associated in the CMH test. Three other loci *PROX1*, *PPARG* and *FTO* became significant (*P*=<0.05) for the first time. Together, 17 variants showed nominal significance in the stratified analysis. Overall, the similarity between the non-stratified and stratified analysis suggests that ethnic group did not influence the results of the T2D SNP associations. Therefore, the subsequent analyses were performed (unless otherwise stated) on the combined (all case–controls irrespective of self-identified ethnic group) dataset.

3.3.3 Collation of reported odds ratio and allele frequency and power calculations

PubMed was used to search the literature for the relevant publications for the SNPs of interest, to gather the odd ratios and allele frequency data used for the European data (Table 3.3.2). These data were then used to compare the effect size estimates and RAF between the Sri Lankan samples and Europeans.

Power was calculated as described above using the sample size of 830 cases and 1497 controls. Table 3.3.2 displays the results of the power calculation at each individual SNP and provides a summary of the studies that were used to collate the information for the European samples.

3.3.4 Comparison of odds ratios and risk allele frequency between Sri Lankan and European samples

Given the considerable differences in prevalence and presentation of T2D between European and South Asian populations, I looked for evidence for consistent differences in effect size or allele frequency between the present study and previous reports from European studies. The results of the Pearson correlation coefficient found significant correlations between Sri Lankan and European samples for both the allelic odds ratio point estimates (Figure 3.3.1; r=0.50, P=1.8×10^{-3}) and risk allele frequencies (Figure 3.3.2: r=0.64, P=2.3×10^{-5}); however, there were no suggestions of systematic differences in either. Allelic OR point estimates were greater in Sri Lankans than Europeans at 19 of 36 loci, and risk allele frequencies at 20 of 36. As the magnitude of effect or frequency of the risk allele appears to be no different to that seen in Europeans, it is unlikely that these SNPs explain the differences in prevalence and presentation for T2D in the Sri Lankan population.

3.3.5 Genetic risk score results

The results of the genetic risk score demonstrated that individuals carrying greater numbers of (weighted) T2D risk-increasing alleles had increased T2D risk (Figure 3.3.3), with an allelic OR of 1.10 (95%CI: 1.08–1.13, P=1.2×10^{-17}). Individuals in the highest quintile of the genetic risk score had more than threefold higher odds of T2D (3.44 [95%CI: 2.56–4.63] P=2.6×10^{-16}) when compared to individuals in the lowest quintile.

Nearest gene	CHR	SNP	Alleles	T2D risk allele in Sri Lankans	T2D risk allele in Europeans	RAF in Sri Lankans	Odds ratio (95% CI) adjusted for age and sex	*P*-value adjusted for age and sex	Odds ratio for CMH test (95% CI) adjusted for age, sex	CMH test *P*-value adjusted for age, sex
NOTCH2	1	rs10923931	G/T	T	T	0.25	1.04 (0.90–1.20)	0.566	1.04 (0.90–1.20)	0.610
PROX1	1	rs340874	C/T	C	C	0.51	1.13 (1.00–1.29)	0.053	1.16 (1.02–1.31)	0.021
BCL11A	2	rs243021	G/A	A	A	0.49	1.23 (1.09–1.41)	1.1×10^{-3}	1.25 (1.11–1.41)	3.6×10^{-4}
GCKR	2	rs780094	C/T	C	C	0.20	0.87 (0.75–1.03)	0.110	0.91 (0.78–1.06)	0.21
IRS1	2	rs7578326	A/G	A	A	0.81	1.20 (1.03–1.41)	0.023	1.17 (1.00–1.37)	0.055
THADA	2	rs7578597	T/C	T	T	0.89	1.17 (0.96–1.44)	0.118	1.10 (0.91–1.35)	0.321
ADAMTS9	3	rs4607103	C/T	C	C	0.46	1.02 (0.89–1.15)	0.781	0.99 (0.87–1.11)	0.815
ADCY5	3	rs11708067	A/G	A	A	0.81	1.22 (1.04–1.45)	0.016	1.28 (1.09–1.50)	2.7×10^{-3}
IGF2BP2	3	rs4402960	G/T	T	T	0.49	1.10 (0.97–1.26)	0.125	1.10 (0.97–1.25)	0.124
PPARG	3	rs1801282	C/G	C	C	0.91	1.21 (0.97–1.52)	0.093	1.30 (1.04–1.62)	0.021
WFS1	4	rs10010131	G/A	G	G	0.76	1.24 (1.07–1.45)	5.6×10^{-3}	1.25 (1.07–1.45)	4.1×10^{-3}
ZBED3	5	rs4457053	A/G	G	G	0.20	1.25 (1.08–1.46)	3.6×10^{-3}	1.23 (1.06–1.43)	6.7×10^{-3}
CDKAL1	6	rs10946398	T/C	C	C	0.24	1.09 (0.94–1.27)	0.240	1.07 (0.92–1.24)	0.390
DGKB/TMEM195	7	rs2191349	T/G	T	T	0.66	0.99 (0.86–1.12)	0.821	0.99 (0.87–1.13)	0.855
GCK	7	rs4607517	G/A	A	A	0.12	0.92 (0.75–1.13)	0.441	0.93 (0.76–1.13)	0.449
JAZF1	7	rs864745	C/T	T	T	0.79	1.17 (1.00–1.37)	0.049	1.14 (0.98–1.33)	0.082
KLF14	7	rs972283	G/A	G	G	0.63	1.01 (0.89–1.16)	0.834	1.03 (0.91–1.17)	0.633
SLC30A8	8	rs13266634	C/T	C	C	0.77	1.32 (1.14–1.55)	3.9×10^{-4}	1.33 (1.14–1.55)	3.0×10^{-4}
TP53INP1	8	rs896854	C/T	T	T	0.40	1.03 (0.90–1.17)	0.66	1.02 (0.90–1.16)	0.761
CDKN2A/B	9	rs10811661	T/C	T	T	0.85	1.34 (1.12–1.61)	1.2×10^{-3}	1.32 (1.10–1.57)	2.0×10^{-3}
CHCHD9	9	rs13292136	C/T	C	C	0.84	1.29 (1.08–1.54)	4.4×10^{-3}	1.31 (1.10–1.56)	2.4×10^{-3}

CDC123/CAMK1D	10	rs12779790	A/G	G	G	0.14	1.09 (0.91–1.31)	0.327	1.09 (0.91–1.30)	0.346
HHEX/IDE	10	rs1111875	C/T	C	C	0.36	1.14 (1.01–1.30)	0.048	1.17 (1.03–1.32)	0.017
TCF7L2	10	rs7903146	C/T	T	T	0.34	1.38 (1.21–1.59)	2.8×10^{-6}	1.35 (1.19–1.54)	6.1×10^{-6}
CENTD2	11	rs1552224	A/C	A	A	0.82	1.12 (0.95–1.33)	0.174	1.11 (0.95–1.31)	0.194
KCNJ11	11	rs5219	C/T	T	T	0.34	1.15 (1.01–1.31)	0.035	1.16 (1.02–1.31)	0.026
KCNQ1	11	rs231362	A/G	G	G	0.76	1.03 (0.89–1.19)	0.728	1.03 (0.90–1.20)	0.644
MTNR1B	11	rs10830963	C/G	G	G	0.84	1.12 (0.99–1.27)	0.067	1.10 (0.97–1.25)	0.132
HMGA2	12	rs1531343	G/C	C	C	0.19	1.23 (1.05–1.45)	8.8×10^{-3}	1.22 (1.04–1.43)	0.013
HNF1A	12	rs7957197	T/A	T	T	0.96	0.94 (0.67–1.33)	0.735	1.02 (0.73–1.42)	0.907
TSPAN8/LGR5	12	rs7961581	T/C	C	C	0.35	0.87 (0.77–0.99)	0.049	0.84 (0.74–0.96)	0.012
PRC1	15	rs8042680	A/C	A	A	0.74	1.18 (1.03–1.37)	0.020	1.24 (1.08–1.43)	0.003
ZFAND6	15	rs11634397	G/A	G	G	0.53	1.00 (0.89–1.14)	0.933	1.00 (0.88–1.13)	0.996
FTO	16	rs9939609	A/T	A	A	0.36	1.11 (0.98–1.27)	0.089	1.14 (1.00–1.30)	0.044
HNF1B(TCF2)	17	rs757210	C/T	T	T	0.28	1.01 (0.88–1.16)	0.851	0.99 (0.87–1.14)	0.931
DUSP9	X	rs5945326	A/G	A	A	0.57	1.15 (1.03–1.29)	0.010	1.29 (1.14–1.47)	6.1×10^{-5}

Table 3.3.1 Individual SNP association results for T2D risk in the Sri Lankan case–control samples

Locus	Chr.	SNP	Power	Effect allele frequency (HapMap/CEU)	Odds ratio European studies	Reference for European odds ratio
NOTCH2	1	rs10923931	0.15	0.11	1.09	Zeggini et al. (2008) (2nd stage replication)
PROX1	1	rs340874	0.20	0.5	1.07	Dupuis et al. (2010) (stage 1 replication data)
BCL11A	2	rs243021	0.24	0.46	1.08	Voight et al. (2010) (2nd stage replication)
GCKR	2	rs780094	0.15	0.62	1.06	Dupuis et al. (2010) (stage 1 replication data)
IRS1	2	rs2943641	0.33	0.61	1.1	Voight et al. (2010) (2nd stage replication)
THADA	2	rs7578597	0.23	0.92	1.15	Zeggini et al. (2008) (2nd stage replication)
ADAMT59	3	rs4607103	0.23	0.81	1.1	Zeggini et al. (2008) (2nd stage replication)
ADCY5	3	rs11708067	0.33	0.78	1.12	Dupuis et al. (2010) (stage 1 replication data)
IGF2BP2	3	rs4402960	0.25	0.29	1.09	Zeggini et al. (2007) replication meta-analysis
PPARG	3	rs1801282	0.18	0.92	1.13	Altshuler et al. (2000)
WFS1	4	rs1001013	0.41	0.32	1.12	Sandhu et al. (2007)
ZBED3	5	rs4457053	0.16	0.26	1.07	Voight et al. (2010) (2nd stage replication)
CDKAL1	6	rs10946398	0.53	0.34	1.14	Zeggini et al. (2007) (stage 1 replication data)
DGKB	7	rs2191349	0.16	0.47	1.06	Dupuis et al. (2010) (stage 1 replication data)
GCK	7	rs4607517	0.15	0.2	1.07	Dupuis et al. (2010) (stage 1 replication data)
JAZF1	7	rs864745	0.24	0.52	1.08	Zeggini et al. (2008) (2nd stage replication)
KLF14	7	rs972283	0.16	0.55	1.06	Voight et al. (2010) (2nd stage replication)
SLC30A8	8	rs13266634	0.50	0.75	1.15	Sladek et al. (2007)
TP53INP1	8	rs896854	0.13	0.48	1.05	Voight et al. (2010) (2nd stage replication)
CDKN2A/B	9	rs10811661	0.58	0.79	1.18	Zeggini et al. (2007) (stage 1 replication data)
CHCHD9	9	rs13292136	0.10	0.93	1.08	Voight et al. (2010) (2nd stage replication)
CDC123	10	rs12779790	0.31	0.23	1.11	Zeggini et al. (2008) (2nd stage replication)
HHEX/IDE	10	rs1111875	0.24	0.56	1.08	Sladek et al. (2007)
TCF7L2	10	rs7903146	1.00	0.25	1.37	Grant et al. (2006)
CENTD2	11	rs1552224	0.27	0.88	1.14	Voight et al. (2010) (2nd stage replication)
KCNJ11	11	rs5219	0.57	0.5	1.14	Gloyn et al. (2004)

KCNQ1	11	rs231362	0.20	0.52	1.07	Voight et al. (2010) (2nd stage replication)
MTNR1B	11	rs10830963	0.22	0.3	1.08	Voight et al. (2010) (2nd stage replication)
HMGA2	12	rs1531343	0.12	0.1	1.08	Voight et al. (2010) (2nd stage replication)
HNF1A	12	rs7957197	0.09	0.85	1.05	Voight et al. (2010) (2nd stage replication)
TSPAN8	12	rs7961581	0.13	0.23	1.06	Zeggini et al. (2008) (2nd stage replication)
PRC1	15	rs8042680	0.12	0.22	1.06	Voight et al. (2010) (2nd stage replication)
ZFAND6	15	rs11634397	0.12	0.56	1.05	Voight et al. (2010) (2nd stage replication)
FTO	16	rs8050136	0.73	0.45	1.17	Frayling et al. (2007)
HNF1B	17	rs757210	0.45	0.43	1.12	Winckler et al. (2007)
DUSP9	X	rs5945326	0.87	0.12	1.32	Voight et al. (2010) (2nd stage replication)

Table 3.3.2 Summaries of the reported allele frequency and odds ratio in Europeans for the T2D SNPs investigated and power calculations. Power calculations were performed assuming a disease prevalence of 10% and a significance threshold (α) of 0.05. Effect allele frequencies used were from the CEU component of HapMap and allelic odds ratio from European populations using the sample size of 830 cases and 1497 controls.

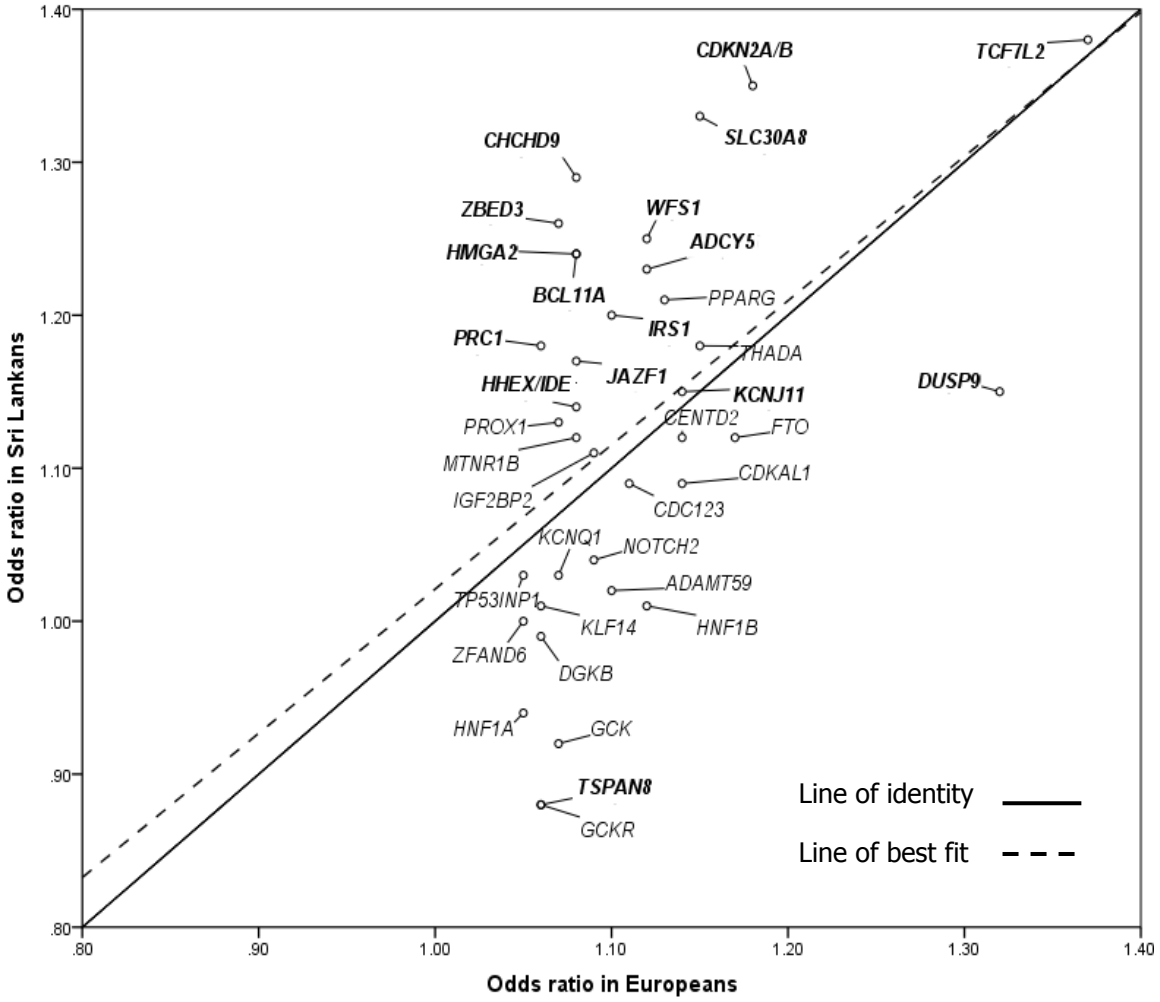


Figure 3.3.1 Comparisons of allelic ORs between the Sri Lankans and Europeans. Loci labelled in bold are variants that showed nominal significance (P<0.05) in Sri Lankan subjects. European ORs used were derived from previously reported studies.

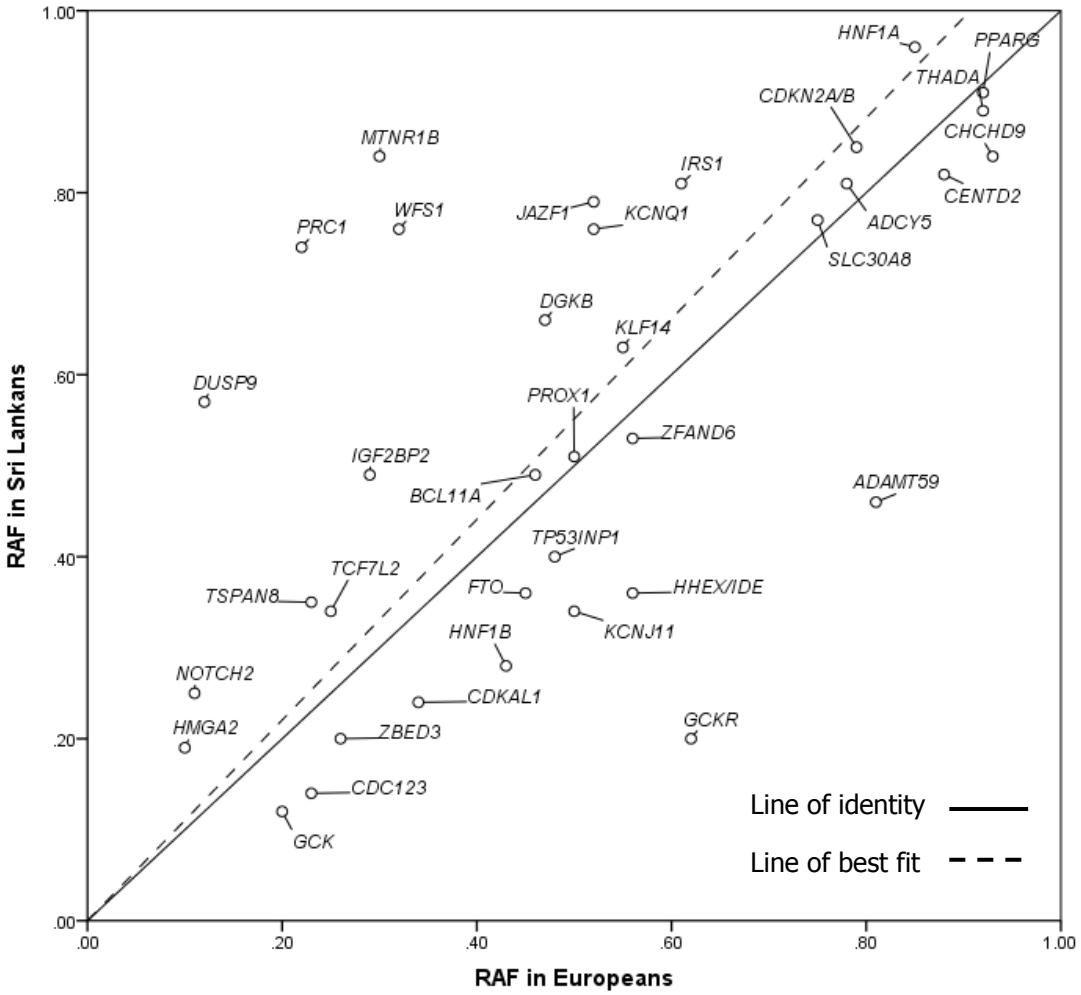


Figure 3.3.2 Comparison of risk allele frequencies (RAF) between the Sri Lankan samples included in this study (cases and controls combined) and previously reported RAF for Europeans (from HapMap CEU).

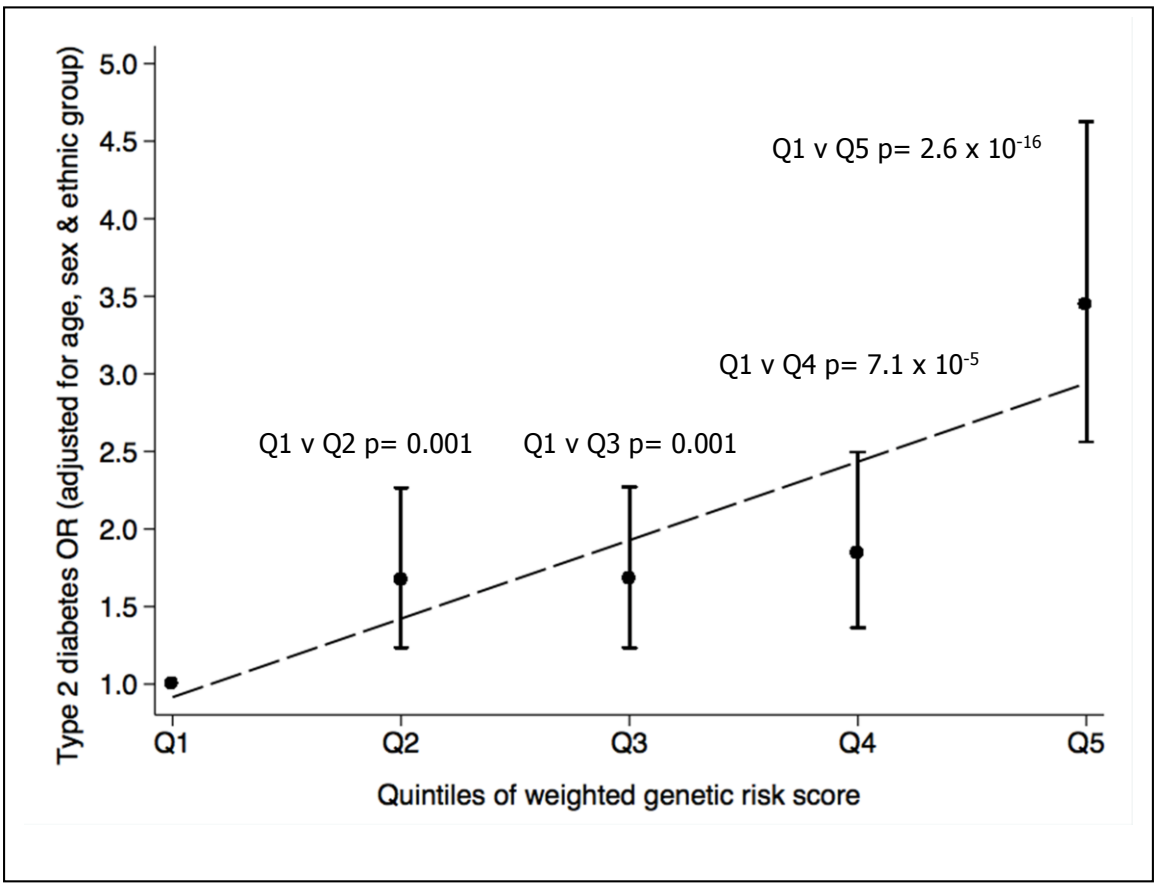


Figure 3.3.3 The combined impact of the 36 T2D-associated SNPs on T2D risk in T2D cases and controls from Sri Lanka. Subjects were grouped into quintiles of the weighted genetic risk score. Circles represent the T2D odds ratio (adjusted for age, sex and ethnic group) when comparing each quintile group to the group in the lowest quintile (Q1). The capped lines represent the 95% CI of the T2D odds ratios. Analysis and figure completed by Dr Maneka De Silva.

3.4 Discussion

This study has shown that established T2D risk variants, most of them initially identified in European samples, show strong enrichment of association in T2D cases and controls from Sri Lanka. This pattern of enrichment, along with the absence of any systematic difference in risk-allele frequency or odds ratios between Sri Lankan and European samples has several important corollaries.

These data provide further evidence of transethnic consistency in allelic patterns for common T2D variants, which builds on similar findings observed in a variety of ethnic groups including others of South Asian origin (Chauhan et al. 2010; Chidambaram et al. 2010; Mahajan et al. 2014; Rees et al. 2011; Sanghera et al. 2008; Waters et al. 2010). This is consistent with the view that, at most of these association signals, the underlying causal variants tagged by GWAS studies are shared across different ethnic groups and most likely arose many tens of thousands of years ago, before the major human migrations out of Africa (Mahajan et al. 2014; Marigorta and Navarro 2013). This pattern of transethnic consistency provides further evidence against the promoted hypothesis that common variant GWA signals arise due to the combined ('synthetic') effects of multiple causal rare alleles at each locus (Dickson et al. 2010; Goldstein 2009). Rare variants are predominately of recent origin and thus likely to be ethnic-specific and not shared across populations (Anderson et al. 2011). Therefore, if this synthetic association model were true for most loci, these consistent patterns of common variant association across populations would not be expected (Dickson et al. 2010; Goldstein 2009). Whilst the findings in this study indicate that the synthetic model does not apply across T2D-risk loci, they do necessarily exclude it at all loci. To answer whether the synthetic association model is applicable to any of the T2D-risk loci, comprehensive deep whole genome sequencing in large numbers of individuals of diverse ethnicities will be required.

South Asians have an increased prevalence for T2D and this occurs at a younger age and at lower BMI thresholds compared with European counterparts (McKeigue et al. 1991). Despite differences in both the prevalence and presentation of T2D between Sri Lanka and Europe, we observed no systematic differences in either risk allele frequency or effect size. This finding is supported by two recent publications that both showed no significant difference in effect size estimates or RAF for the common T2D risk SNPs between the two populations (Mahajan et al. 2014; Sohani et al. 2014).

South Asians have an increased predisposition for insulin resistance (Dhawan et al. 1994) and this plays a key role in the increased prevalence for T2D observed in South Asians. Currently, many of the identified loci associated with T2D susceptibility are linked to diminished insulin secretion and β-cell function. Therefore, differences in effect for these loci are unlikely to be to very different between the two populations (Bakker et al. 2013). Consequently, it is likely that the ethnic differences in epidemiological and physiological patterns cannot be credited to differences in common variant predisposition. Further investigations such as whole exome and genome sequencing may lead to the discovery of casual low frequency or rare variants that might contribute to the increased risk for diabetes between the two ethnic groups (Sohani et al. 2014).

Although the general patterns are clear, the relatively modest sample sizes available in this study limit the interpretations that can be made at any individual locus. None of the variants tested reached stringent genome-wide significance, and only about half of the loci reached nominal significance (i.e. $p \leq 0.05$). Of the 20 SNPs not reaching nominal significance ($P > 0.05$), 15 were directionally consistent with the associations reported in Europeans, indicating that many of these are likely to be false-negatives reflecting the limited power of the study. Most of these loci have very modest odds ratios, and therefore would require larger samples sizes to be detected. Several loci amongst the 15 loci which showed no association ($P > 0.05$) in this study but did display directional consistency with data from Europeans have shown evidence of association in subsequent South Asian case–control studies. For example, variants at the *GCKR* and *CDC123* loci were not associated with T2D in the present study, but were associated in far larger meta-analyses of South Asian samples (Kooner et al. 2011).

In conclusion, this study is the first one that has comprehensively assessed T2D susceptibility variants in Sri Lankans. The current study has shown that common T2D risk variants identified in Europeans have a similar genetic risk in Sri Lankans, adding further to the evidence that South Asians and Europeans share many overlapping common variants, which contribute to T2D risk. Further studies are required in larger samples sizes, which will increase power for determining whether those variants with small effect sizes are true disease susceptibility variants in Sri Lankans.

Chapter 4

Identification of novel type 2 diabetes risk variants in individuals of South Asian ancestry using genome-wide association studies

Below are the references to the publications that the results presented in this chapter contribute. The complete manuscript is provided in the appendix section:

- 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.

- Saxena, R., et al. (2013), 'Genome-wide association study identifies a novel locus contributing to type 2 diabetes susceptibility in Sikhs of Punjabi origin from India', *Diabetes,* 62 (5), 1746-55.

4.1 Introduction

The initial waves of genome-wide association studies (GWAS) focusing on the genetic basis of type 2 diabetes (T2D) were performed in individuals of European ancestry (Scott et al. 2007; Voight et al. 2010; Zeggini et al. 2007; Zeggini et al. 2008). This was primarily due to the availability of large numbers of individuals that were well characterised phenotypically and genotyping platforms that catalogued genetic markers predominately based on the genetic variability in Europeans (Rosenberg et al. 2010). The subsequent collections of T2D cases and normoglycaemic controls (or sometimes population control samples) of other ethnic origins and improvements in genotyping arrays capturing genetic variants in non-European populations have facilitated the number of GWAS in other ancestral groups (Mahajan et al. 2014).

The increased prevalence of T2D in the South Asian population makes this group particularly interesting to investigate. The Indian subcontinent is home to many ethnic groups, cultures, and languages (Chaubey et al. 2007). Studies examining the genetic diversity of the people in India suggest that there were two major 'out-of-Africa' migration routes and the majority of people are a mixture of two divergent populations, the Ancestral North Indians (ANI) and the Ancestral South Indians (ASI) (Reich et al. 2009; Xing et al. 2010). The individuals of ANI ancestry are genetically related to Europeans, Middle Easterners and Central Asians (Indo-Europeans). The individuals of ASI ancestry are distinct from the ANIs with a suggestion that ASIs are predominantly Southern Dravidian speakers and may be distantly related to indigenous Andaman Islanders (Reich et al. 2009).

Investigations of the Y chromosome and mitochondrial DNA and more recently genome sequencing have reported evidence of significant admixture between the two groups (Moorjani et al. 2013). Furthermore, archaeological and linguistic analysis also supports mixture of two distinct populations (Moorjani et al. 2013). Indo-European languages such as Hindi and Sanskrit are spoken mainly in northern India whilst Dravidian languages (including Tamil and Telugu) are mainly spoken in southern India. Linguistic studies have shown that some of the oldest Hindu text written in archaic Sanskrit contains 'loan' words from the Dravidian languages, which are not found in Indo-European languages outside the Indian subcontinent, suggesting contact between the divergent populations (Moorjani et al. 2013). Analysis of the current situation in mainland India shows considerable admixture between the two groups with ANI

ancestry ranging from 39 to 71% in most Indian groups (Moorjani et al. 2013; Reich et al. 2009).

This chapter describes two separate collaborative GWA studies. Both studies were performed in populations of South Asian ancestry, focusing on identifying novel T2D susceptibility loci. My contribution was to generate and contribute data from the Sri Lankan samples as part of the replication phase of each study.

4.1.1 Study A: Genome-wide association study in individuals of South Asian ancestry

This GWAS was a multistage process led by collaborators at Imperial College London. Figure 4.1.1 illustrates an overview of the study design employed for this GWAS. The GWAS scans were coordinated by the lead collaborators and were performed using Illumina genotyping arrays and genotypes were called using GeneCall or Illumina algorithms. The initial discovery phase (stage 1) GWAS was performed in 5561 cases and 14,458 controls. Subjects comprised three South Asian cohorts, the London life Sciences Population (LOLIPOP) study, the Pakistan Risk of Myocardial Infarction Study (PROMIS) and the Singapore Indian Eye (SINDI) study. South Asians were defined as those individuals with ancestry from India, Pakistan, Sri Lanka and Bangladesh. Although previously described studies have suggested varying degrees of population substructure, principal component analysis showed that in this study there was no population stratification between the three cohorts (Kooner et al. 2011).

All meta-analysis data were analysed and compiled by lead collaborators unless otherwise stated. Analysis at the discovery stage tested the association of directly genotyped single nucleotide polymorphisms (SNPs) only, as at the time no suitable South Asian reference panel was in place for imputation (Kooner et al. 2011). Quality checks were performed and SNPs were excluded if they had a call rate of <97%, a MAF <1% or deviated from Hardy Weinberg equilibrium (HWE) ($P < 1.0 \times 10^{6}$). Samples were removed if they had a SNP call rate <95%. Primary SNP association testing was performed separately in men and women using logistic regression and an additive genetic model. The data of the separate analyses were combined by using a fixed-effects inverse variance meta-analysis (Kooner et al. 2011). The results of the combined association testing showed that the lead SNP reaching genome-wide association ($P < 5 \times 10^{-8}$) was at the *TCF7L2* locus. A further 59 independent SNPs showed suggestive association ($P \leq 1 \times 10^{-4}$) with T2D risk and of these 50 had not been previously described in other T2D GWAS.

Additional secondary analysis was performed which included the following five GWAS meta-analyses (Kooner et al. 2011):

- Male-gender specific;
- Female-gender specific;
- Adjusted for body mass index;
- Lean T2D cases (BMI<25kg/m^2) compared to overweight controls (BMI>25kg/m^2); and
- GWAS analysis, which contained imputed data for missing genotypes using the HapMap 2 reference panel (2,646,472 SNPs).

To prioritise a subset of SNPs for stage 2 replication testing in other South Asian cohorts, the following three approaches were employed (Figure 4.1.1):

- All SNPs (n=7) associated with T2D risk ($P \leq 1\times10^{-5}$) in the primary South Asian GWAS analysis using the directly genotyped SNPs;
- All SNPs (n=1 identified in the female-specific analysis) associated with T2D ($P < 5\times10^{-8}$) in any of the five secondary GWAS analysis; and
- Those SNPs (n=12) which showed disease association ($P > 1\times10^{-5}$ and $P \leq 1\times10^{-4}$) in the stage 1 of the South Asian GWAS analysis but also had a P-value $\leq 1\times10^{-3}$ in an analysis which combined results from the GWAS stage of the DIAbetes Genetics Replication And Meta-analysis plus study (DIAGRAM+ or DIAGRAM v2).

The DIAGRAM consortium focuses on characterising the genetic basis of T2D primarily in individuals of European ancestry and has been involved in identifying many novel T2D loci, which contribute to disease risk (Voight et al. 2010). Therefore, these 12 SNPs showed association with T2D risk in both South Asians and Europeans (Kooner et al. 2011). The replication of these 20 SNPs was then performed as part of stage 2 in multiple centres, resulting in 13,170 cases and 25,398 controls of South Asian ancestry being genotyped.

4.1.2 Study B: Genome-wide association study in Sikhs of Punjabi origin from India

The GWAS included multiple meta-analysis phases to maximise the possibility of detecting novel population-specific and multiethnic T2D risk loci. This study was an independent effort to that described above and the stage 1 samples were not part of study A. A summary of the study design is outlined in Figure 4.1.2. The initial discovery

cohorts (stage 1) comprised 1616 individuals (842 cases and 774 controls) ascertained from the Asian Indian Diabetic Hearth Study (AIDHS), otherwise known as the Sikh Diabetes Study (SDS) (Saxena et al. 2013). The AIDHS/SDS cohort is well characterised and were all recruited from the same geographical location, the State of Punjab. Medical records, use of medication and fasting plasma glucose levels ≥7.0 mmol/L (≥126 mg/dL) were used to define T2D status. Control subjects were selected based on fasting glucose of <100.8 mg/dL or a 2-h glucose <141.0 mg/dL (Saxena et al. 2013).

Genotyping and meta-analysis described below was completed by lead authors of the study unless otherwise stated. Genotyping (genome-wide) for stage 1 was performed using a Human Illumina 660-W-Quad BeadChip panel. Assessment of population structure showed no outliers, and samples were removed if they showed relatedness when verified through identity by descent sharing (Saxena et al. 2013). SNPs were excluded if they deviated from HWE (P <1.0×10^{6}), had a call rate of <95% or had a MAF <1% (Saxena et al. 2013). Association testing for SNPs with T2D in stage 1 was performed using a logistic regression and an additive genetic model. Analysis was carried out for directly genotyped SNP to increase genome coverage; genotypes were imputed using the multiethnic HapMap 3 reference panel, which had not been available for the previous study (Saxena et al. 2013). The multiethnic reference panel consisted of ~1.5 million autosomal SNPs, from 1011 individuals from Africa, Asia, Europe and the Americas. Of these, 100 individuals were from the Gujarati Indians in Houston, Texas (GIH) population (Saxena et al. 2013). A combined analysis of directly genotyped and imputed SNPs adjusted for covariates BMI, age, sex and five principal components identified 513 independent novel candidate signals for T2D risk (P<10^{-3}) in the discovery cohort.

Stage 2 involved a two-step replication process to try to optimise the detection of population-specific variants. The lead association signals from the discovery stage 1 (r^2 <0.25 and P<10^{-3}) were replicated and this was divided into stages 2a and 2b (Figure 4.1.2). Stage 2a consisted of the SNPs undergoing *in silico* replication in the Punjabi Sikh counterpart of the LOLIPOP cohort, in line with the discovery cohort, whilst stage 2b entailed replication in the non-Sikh component of the LOLIPOP and PROMIS cohorts. A subsequent meta-analysis of stages 1, 2a and 2b used a fixed effects inverse-variance method to merge the results of the individual studies, which identified 66 novel potential T2D signals (P<10^{-4}) (Saxena et al. 2013).

To increase further the sample size and to detect multiethnic T2D loci, a stage 3 multilevel (a–d) replication phase was performed (Saxena et al. 2013):

- Stage 3a involved directly genotyping the 66 SNPs from stage 2 in a further 2894 Sikh individuals using the Sequenom MassARRAY platform.
- Stage 3b replication process had two main goals. The first was to replicate those SNPs from an interim analysis of all South Asians from the stage 2b component. This consisted of 31 SNPs ($P<10^{4}$) and the replication was performed in other South Asian (non-Sikh) samples, one of which was the Sri Lankan samples. The second goal was to determine whether the signals specific to the Sikh ethnic group overlapped with other South Asian ethnicities. This involved the testing of signals ($P<10^{-4}$) based on the result of the meta-analysis of stage 1, 2a and 3a.
- Stage 3c, *in silico* replication of the 31 SNPs, was carried out in East Asians (AGEN cohort).
- Stage 3d, *in silico* replication of the 31 SNPs, was performed in the Europeans (DIAGRAM v2).

4.1.3 Investigating South Asian subpopulations could identify ethnic-specific variants

Extending GWAS to populations of non-European ancestry serves several different purposes. It gives the opportunity to identify new disease susceptibility variants driven by differences in disease presentation and prevalence between populations. T2D in South Asians is more prevalent than in other major populations. Part of the increased prevalence of T2D in South Asians is likely driven by the increased propensity for insulin resistance (McKeigue et al. 1991). Currently, the majority of the associated T2D risk loci are implicated in influencing indices that affect beta-cell function (Dimas et al. 2014; Voight et al. 2010). Therefore, examining South Asian individuals could further reveal novel loci, especially those that influence pathways linked to insulin resistance.

Multiethnic data also contribute useful information about linkage disequilibrium (LD) patterns for existing disease susceptibility alleles across major ethnic groups, and aids fine mapping studies to refine the variant within confirmed loci (McCarthy et al. 2008). For example, a recent study combining individuals from European and African American ancestry demonstrated how this approach could successfully refine loci associated with lipids (Bentley et al. 2014). Provided causal variants are shared across ancestral groups, the greater prevalence of T2D in South Asians might aid detection of shared variants through trans-ethnic meta-analysis.

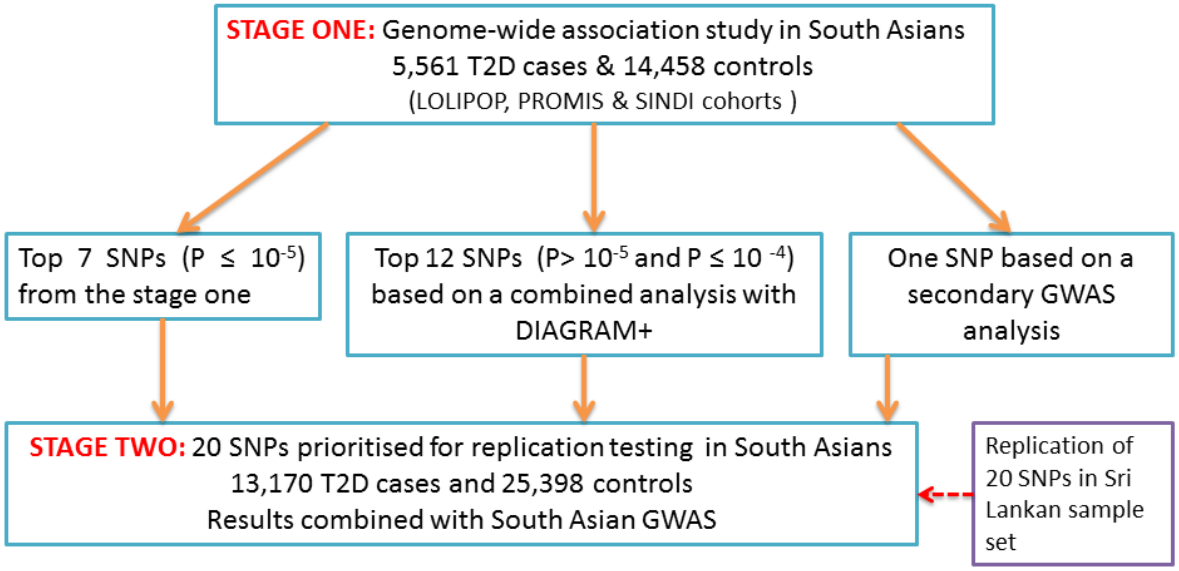


Figure 4.1.1 Outline of the study design used for the South Asian GWAS (Study A). Figure adapted from the original publication (Kooner et al. 2011).

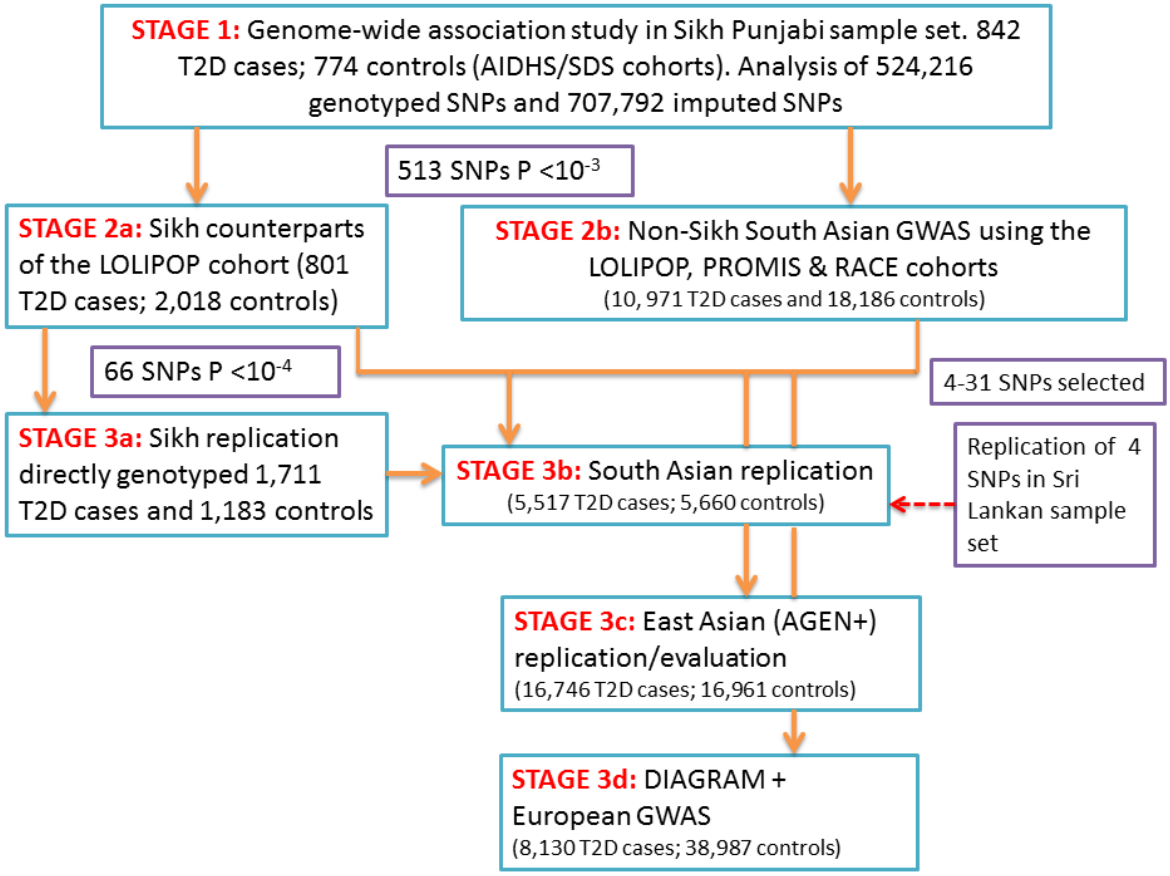


Figure 4.2.1 Outline of the study design used for the Sikh-Punjabi GWAS (Study B). Figure adapted from the original publication (Saxena et al. 2013).

4.1.3.1 Experimental aims

The aims of the work described in this chapter were the following:

- To complete the genotyping and association analysis of 24 SNPs in Sri Lankan samples. These SNPs were selected by collaborators as part of the replication phase of large-scale genome-wide association studies in populations of South Asian ancestry.
- To use my association analysis to identify whether these 24 T2D risk variants are ethnic-specific or shared between individuals of Sri Lankan origin and South Asian and European cohorts.

4.2 Methodology

4.2.1 Study samples

Full details of the studies from which the Sri Lankan case–control subjects were ascertained are described in Chapter 2 (sections 2.1).

4.2.2 Study A: Sri Lankan subjects used in the replication phase of the South Asian GWAS

As part of the replication phase for study A, 20 SNPs were genotyped in the Sri Lankan samples. After removal of samples based on the exclusion criteria described in Chapter 2 (section 2.1) and DNA samples that failed based on a per sample genotype call rate ≤90% a total of 841 cases (40% male) and 1471 (43% male) controls were taken forward for association testing. The mean (SD) age of the SLYDS cases were 39.0 (8.1) years, BMI was 24.7 (4.0) kgm^{-2} and mean FPG was 8.9 (3.8) mmol/L. The SLDCS control samples had a mean age 44.7 (14.4) years, BMI 21.2 (4.2) kgm^{-2} and FPG 4.8 (1.2) mmol/L.

4.2.3 Study B: Sri Lankan subjects used as part of the replication phase for the Sikh Punjabi GWAS

As part of study B, four SNPs were selected by the lead collaborators to genotype in the Sri Lankan samples. Samples were removed based on the exclusion criteria descried in Chapter 2 (section 2.1) and if DNA samples failed on a per sample genotyping call rate of ≤90%. In total 879 cases (41% males) and 1613 controls (43% males) participants were taken forward for association testing. The mean (SD) age of the SLYDS cases were 37.3 (5.2) years, BMI was 24.9 (4.0) kgm^{-2} and mean FPG was 9.1 (3.5) mmol/L. The SLDCS control samples had a mean age 44.8 (14.4) years, BMI 21.3 (4.0) kgm^{-2} and FPG 4.8 (1.0) mmol/L.

4.2.4 Genotyping

In total, 24 SNPs were genotyped using the Applied Biosystems 7900HT platform across both studies. Twenty SNPs were part of the replication for study A and four SNPs were part of study B. All genotyping assays were labelled with florescent probes (Vic/Fam) and were predesigned by Applied Biosystems (Table 4.2.1). Genotyping experimental procedures are described in Chapter 2 (section 2.4). On completion of the genotyping, the data were sent to the Wellcome Centre of Human Genetics where the data were collated by colleagues Dr Will Rayner and Neil Robertson (WTCHG, University of Oxford). The generated data were sent back in PED and Map file formats. The PED file contained all the genotyping information for all SNPs along with IDs of the individuals; the Map file provided information on each of the SNPs.

4.2.5 Statistical analysis

The association testing I performed for the 24 SNPs was based on agreed specifications set by the central analysis team for each study. Prior to analysis as part of the quality check for the genotyping within the Sri Lankan sample set, the genotyping call rate was checked for each SNP (<95%). In addition, departure from HWE was assessed in the control samples; those SNPs with $P \leq 0.001$ were excluded. PLINK v.107 was used to carry out the individual SNP associations (http://pngu.mgh.harvard.edu/~purcell/plink/ ; Purcell et al. 2007).

For study A, association testing was performed using a logistic regression assuming an additive model. Analysis was carried out in two stages for the 20 SNPs. The first was of the combined data (both males and females) adjusting for the following covariates: (a) age and sex; (b) age, sex and BMI; (c) age, sex and waist–hip ratio (WHR).The second analysis was a separate analysis in males only and females only adjusting for no covariates.

For study B, association testing was performed using a logistic regression assuming an additive model. The four SNPs were analysed adjusting for covariates age and sex, followed by a second adjustment of age, sex and BMI.

On completion of association analysis for the 24 replication SNPs, data were sent to the lead collaborators. The data for study A (20 SNPs) was sent to the lead collaborators John C Chambers (Imperial College, London) and for study B (4 SNPs) data were sent to Dharambir K. Sanghera (University of Oklahoma, Oklahoma City), where they performed

an inverse variance meta-analysis to incorporate the additional replication data and combine it with stage 1 data.

4.2.6 Collating the reported allele frequencies and odds ratios for the 24 SNPs in South Asians and European populations

The genotyping of 24 SNPs as part of a large multiethnic replication study in the Sri Lankan sample set gave the opportunity to compare and identify whether the allele frequencies and ORs of the SNPs are shared between Sri Lankans, other South Asians and European cohorts or whether the risk of each SNP is population specific. To do this, allele frequencies of the 24 genotyped SNPs were compared between my Sri Lankan dataset and South Asians (SAS) and Europeans (based on CEU [Utah Residents (CEPH) with Northern and Western European ancestry]) from 1000 Genomes phase III (1000 Genomes release 16—Oct 2014, http://browser.1000genomes.org/index.html). I could also compare my Sri Lankan data to frequency data available for individual South Asian sub-populations within the 1000 Genomes data; these were listed as BEB (Bengali in Bangladesh), GIH (Gujarati Indians in Houston, Texas), ITU (Indian Telugu in the UK) and PJL (Punjabi in Lahore, Pakistan).

Risk allele frequencies as well as effect size estimates were also compared between those from my Sri Lankan samples and published GWAS data from the combined South Asian cohort in the study by Kooner et al. (2011) and from Europeans samples in DIAGRAM (Genomes). For these analyses, risk allele frequency was compared for all 24 SNPs. Effect size estimates compared for 19 of the 24 Sri Lankan SNPs were present in both GWAS datasets and could be used. Of the five SNPs that were excluded, one associated only in females in study A (rs17052370), whereas the other four (rs700952, rs11968879, rs4527850, rs12454246) did not replicate. Risk allele frequencies were obtained using HapMap GIH and CEU for these five SNPs.

SNP	CHR	Position	Nearest Gene	Assay ID
rs17031236	2	105538433	*FHL2*	C__33117321_20
rs3923113	2	165210095	*GRB14*	C____245608_10
rs700952	2	5428205	—	C___1044858_10
rs11711167	3	126521709	*ZNF148*	C___8988599_20
rs16861329	3	188149155	*STGAL1*	C__33187550_10
rs6835176	4	40199309	*RBM47*	C__31801773_10
rs9496773	6	144183238	*PHACTR2*	C__30344152_10
rs11968879	6	150202935	*LRP11*	C___1803659_10
rs2111815	7	132349786	*CHCHD3*	C___2199735_10
rs16923500	8	59551647	*CYP7A1*	C__34058315_10
rs4527850	8	134266031	*WISP1*	C__27256625_10
rs1562042	9	105747435	*SMC2*	C___9307574_10
rs1802295	10	70601480	*VPS26A*	C___1531914_10
rs2616646	10	78559423	*KCNMA1*	C___1665108_10
rs4287347	11	132191217	*OPCML*	C____215055_20
rs6576426	15	23254656	*UBE3A*	C__29191865_20
rs620739	15	47852704	*ATP8B4*	C___3013246_10
rs7178572	15	75534245	*HMG20A*	C__29028179_10
rs2028299	15	88175261	*AP3S2*	C___1381378_1_
rs4392054	15	20630271	*ACSM1*	C__11857741_10
rs17052370	17	21660480	*UBBP4*	C__32858855_10
rs1912379	18	66460710	*RPS2P6*	C__11899666_10
rs12454246	18	13834232	—	C___3108175_10
rs4812829	20	42422681	*HNF4A*	C___1199181_30

Table 4.2:1 Summary of SNPs prioritised for genotyping as part of the replication phase of the GWAS studies performed in the South Asian and in the Sikh Punjabi sub-population. Genotyping assays were acquired from Applied Biosystems (Life Technologies). SNPs listed in black text were part of study A; those SNPs listed in purple were part of study B.

4.3 Results

4.3.1 Genotyping of 24 SNPs in the Sri Lankan sample set

In the next sections, I will describe the results of the replication of 24 SNPs from studies A and B, detailed in the Introduction of this chapter, in the Sri Lankan sample set. Table 4.3.1 shows a summary of the quality checks for the 20 SNPs from study A and Table 4.3.2 for the 4 SNPs from study B.

4.3.2 Study A: The identification of six novel T2D susceptibility loci in individuals of South Asian ancestry

The replication phase for study A involved genotyping of 20 SNPs in 13,170 T2D cases and 25,395 controls across 10 South Asian cohorts worldwide, one of which was the Sri Lankan cohort (Kooner et al. 2011). Tables 4.3.3 and 4.3.4 detail the results of the association testing in the Sri Lankan sample set as requested by the central analysis team. Table 4.3.3 shows the association results combined (both males and females) adjusted for the covariates (a) age and sex; (b) age, sex and BMI; and (c) age, sex and WHR. Table 4.3.4 displays the association results separately in men and women with no adjustment for covariates.

In the Sri Lankan data, three SNPs (rs16861329, rs16923500 and rs2028299) showed nominal association ($P \leq 0.05$) with T2D when adjusted for age and sex as well as age, sex and BMI. These signals were located near *ST6GAL1*, *CYP7A1* and *AP3S2*. However, when adjusting for WHR rather than BMI, the association at *CYP7A1* and *AP3S2* was lost. The direction of effect for *ST6GAL1* (OR 1.19 [1.04–1.37]) and *AP3S2* (OR 1.14 [1.00–1.30]) for the age- and sex-adjusted analysis was consistent with the overall study, but was inconsistent for the *CYP7A1* (OR 0.86 [0.75–1.00]) locus. Overall 9 of the 20 SNPs genotyped in the Sri Lankan cohort were directionally consistent with stage 1 of the South Asian GWAS. However, the 95% confidence intervals for the effect size overlapped between the Sri Lankan and combined South Asian GWAS data for the majority of signals (14/20 SNPs).

Individual T2D SNP associations, completed by each of the collaborating centres, were collated by the lead analysis centre (Kooner et al. 2011). Data from the stage 1 South Asian GWAS were combined with the replication data (stage 2), resulting in the six SNPs reaching genome-wide significance ($P < 5 \times 10^{-8}$). These SNPs were found in or near *GRB14*, *VPS26A*, *HMG20A*, *ST6GAL1*, *AP3S2* and *HNF4A*. Of these loci, *GRB14*, *VPS26A* and *HMG20A* were the most strongly associated with T2D. The lead genotyped SNP at

all but the *HNF4A* locus was either the imputed sentinel SNP or in perfect LD with it. At the *HNF4A* locus there was only partial LD in South Asians (r^2 =0.29) (Kooner et al. 2011).

Further analysis by the core analysis team investigated whether there was a link between sentinel SNPs with cis-expression quantitative trait loci (eQTL) in various tissues, including adipose tissue, liver and peripheral blood leucocytes (Kooner et al. 2011). One interesting example was at the *GRB14* locus (rs3923113). The lead SNP is in high LD (r^2=0.85) with a SNP reported to be associated with central adiposity and increased expression of *GRB14* (Kooner et al. 2011).

Additionally, I performed a gender-specific analysis in the Sri Lankan samples (Table 4.3.4) without adjusting for any covariates. Two SNPs showed nominal significance (P<0.05) rs16861329 (P=0.047) and rs4287347 (P=0.013) in the male-only analysis. In females-only, one SNP (rs4392054) showed nominal significance (P=0.039). However, none of these three SNPs could be replicated in the larger South Asian GWAS.

4.3.3 Study B: Genotyping of 4 SNPs in the Sri Lankan cohort as part of the replication phase of the GWAS performed in Sikh Punjabi individuals from India

This GWA study was designed to maximise the prospect of detecting novel population-specific variants and multiethnic loci. Multiple replication stages were performed to facilitate this approach. The genotyping of the Sri Lankan sample set was part of stage 3b (Figure 4.1.2) replication phase where a selection of SNPs were chosen based on analysis performed in the non-Sikh component of South Asian cohorts (LOLIPOP, PROMIS and RACE) to find association signals common to both Sikh-Punjabis and other South Asian populations.

Table 4.3.5 provides a summary of the results of the association testing for the four SNPs genotyped in the Sri Lankan sample set. Two SNPs rs11968879 (OR 1.10 [0.90–1.35]) and rs700952 (OR 1.07 [0.85–1.35]) had an effect that was directionally consistent with that seen in the final meta-analysis. However, none of the four SNPs I genotyped showed statistical significance (P<0.05) with T2D in the Sri Lankan cohort. This was confirmed by the Sikh and South Asian meta-analysis as none of these SNPS reached genome-wide significance.

SNP	CHR	Position	Nearest gene	SNP genotyping call rate (%)	HWE *P*-value in control samples
rs17031236	2	105538433	*FHL2*	97.8	0.175
rs3923113	2	165210095	*GRB14*	97.7	0.496
rs11711167	3	126521709	*ZNF148*	97.8	0.629
rs16861329	3	188149155	*STGAL1*	98.9	0.314
rs6835176	4	40199309	*RBM47*	98.5	0.398
rs9496773	6	144183238	*PHACTR2*	99.1	0.283
rs2111815	7	132349786	*CHCHD3*	99.1	0.023
rs16923500	8	59551647	*CYP7A1*	98.6	0.786
rs1562042	9	105747435	*SMC2*	99.3	0.645
rs1802295	10	70601480	*VPS26A*	98.1	0.197
rs2616646	10	78559423	*KCNMA1*	98.2	0.310
rs4287347	11	132191217	*OPCML*	98.3	0.344
rs6576426	15	23254656	*UBE3A*	99.1	0.802
rs620739	15	47852704	*ATP8B4*	98.3	0.334
rs7178572	15	75534245	*HMG20A*	95.4	0.081
rs2028299	15	88175261	*AP3S2*	97.0	0.228
rs4392054	15	20630271	*ACSM1*	96.4	0.905
rs17052370	17	21660480	*UBBP4*	96.2	0.058
rs1912379	18	66460710	*RPS2P6*	98.7	0.297
rs4812829	20	42422681	*HNF4A*	98.2	0.074

Table 4.3:1: Summary of the quality checks for the 20 SNPs genotyped in study A. All SNPs had to pass a call rate of >95% and SNPs were required to have with a HWE of $P \geq 0.001$ in controls.

SNP	CHR	Position	Nearest gene	SNP genotyping call rate (%)	HWE *P*-value in control samples
rs700952	2	5428205	—	96.0	0.163
rs11968879	6	150202935	*LRP11*	96.4	0.586
rs4527850	8	134266031	*WISP1*	97.0	0.449
rs12454246	18	13834232	—	98.0	0.724

Table 4.3:2: Summary of the quality checks for the 4 SNPs genotyped in study B. All SNPs had to pass a call rate of >95% and SNPs were required to have with a HWE of $P \geq 0.001$ in controls.

SNP	CHR	Nearest gene	Alternate/risk allele	Sri Lankan RAF	Odds ratio (95% CI) adjusted for age and sex	*P*-value adjusted for age and sex	Odds ratio (95% CI) adjusted for age, sex & BMI	*P*-value adjusted for age, sex & BMI	Odds ratio (95% CI) adjusted for age, sex & WHR	*P*-value adjusted for age, sex & WHR
rs17031236	2	*FHL2*	C/A	0.64	1.04 (0.91–1.19)	0.565	1.09 (0.94–1.26)	0.242	1.06 (0.91–1.23)	0.434
rs3923113	2	*GRB14*	C/A	0.83	0.92 (0.78–1.08)	0.307	0.97 (0.81–1.16)	0.713	0.89 (0.74–1.08)	0.236
rs11711167	3	*ZNF148*	A/G	0.33	1.05 (0.92–1.20)	0.481	1.02 (0.89–1.18)	0.753	1.02 (0.88–1.18)	0.830
rs16861329	3	*ST6GAL1*	A/G	0.72	1.19 (1.04–1.37)	0.011	1.21 (1.04–1.40)	0.011	1.22 (1.05–1.42)	0.012
rs6835176	4	*RBM47*	G/A	0.03	1.17 (0.83–1.65)	0.359	1.11 (0.77–1.61)	0.577	1.02 (0.69–1.50)	0.927
rs9496773	6	*PHACTR2*	G/A	0.64	0.97 (0.86–1.10)	0.660	0.95 (0.83–1.09)	0.464	0.94 (0.81–1.09)	0.439
rs2111815	7	*CHCHD3*	G/A	0.45	0.97 (0.85–1.09)	0.589	0.98 (0.86–1.12)	0.773	1.00 (0.87–1.15)	0.981
rs16923500	8	*CYP7A1*	G/A	0.25	0.86 (0.75–1.00)	0.045	0.83 (0.71–0.97)	0.019	0.88 (0.75–1.04)	0.135
rs1562042	9	*SMC2*	A/G	0.72	0.96 (0.84–1.10)	0.582	1.02 (0.87–1.18)	0.819	0.98 (0.84–1.15)	0.816
rs1802295	10	*VPS26A*	G/A	0.23	1.07 (0.93–1.24)	0.357	1.05 (0.89–1.22)	0.587	1.04 (0.88–1.23)	0.638
rs2616646	10	*KCNMA1*	A/G	0.15	0.90 (0.75–1.07)	0.233	0.84 (0.70–1.02)	0.078	0.86 (0.70–1.05)	0.139
rs4287347	11	*OPCML*	A/G	0.52	1.12 (0.99–1.27)	0.065	1.15 (1.00–1.32)	0.045	1.09 (0.95–1.26)	0.234
rs6576426	15	*UBE3A*	A/G	0.19	0.98 (0.84–1.15)	0.810	0.97 (0.82–1.16)	0.768	0.97 (0.81–1.16)	0.713
rs620739	15	*ATP8B4*	A/G	0.28	0.91 (0.80–1.05)	0.193	0.87 (0.75–1.01)	0.066	0.92 (0.78–1.07)	0.274
rs7178572	15	*HMG20A*	A/G	0.43	1.06 (0.94–1.20)	0.352	1.08 (0.94–1.24)	0.271	1.10 (0.95–1.27)	0.189
rs2028299	15	*AP3S2*	A/C	0.34	1.14 (1.00–1.30)	0.045	1.16 (1.01–1.34)	0.038	1.09 (0.94–1.26)	0.267
rs4392054	16	*ACSM1*	G/A	0.12	0.87 (0.72–1.06)	0.160	0.86 (0.69–1.06)	0.155	0.95 (0.76–1.19)	0.655
rs17052370	17	*UBBP4*	G/A	0.30	0.99 (0.86–1.13)	0.859	0.97 (0.84–1.13)	0.731	1.02 (0.87–1.18)	0.805
rs1912379	18	*RPS2P6*	A/G	0.25	0.93 (0.80–1.07)	0.286	0.94 (0.81–1.10)	0.442	0.94 (0.79–1.11)	0.449
rs4812829	20	*HNF4A*	G/A	0.33	1.07 (0.94–1.21)	0.330	1.08 (0.94–1.24)	0.304	1.09 (0.94–1.27)	0.238

Table 4.3:3 Association testing for the 20 SNPs in the Sri Lankan sample set. Results are provided as odds ratio (OR) (95% confidence interval) per copy of the risk allele. RAF, Risk allele frequency.

				MALES				FEMALES			
				Effect per copy of allele A (risk allele)				Effect per copy of allele A (risk allele)			
SNP	**Chr**	**Alt allele**	**Risk allele**	**RAF**	**Beta**	**SE**	**P**	**RAF**	**Beta**	**SE**	**P**
rs17031236	2	C	A	0.654	0.079	0.103	0.442	0.628	0.048	0.085	0.570
rs3923113	2	C	A	0.836	-0.039	0.130	0.765	0.833	-0.096	0.107	0.369
rs11711167	3	A	G	0.690	-0.103	0.104	0.320	0.663	-0.025	0.084	0.763
rs16861329	3	A	G	0.272	-0.211	0.106	0.047	0.288	-0.164	0.089	0.064
rs6835176	4	G	A	0.037	0.469	0.243	0.054	0.031	-0.133	0.241	0.581
rs9496773	6	G	A	0.626	-0.029	0.097	0.764	0.643	0.007	0.083	0.932
rs2111815	7	G	A	0.455	-0.138	0.094	0.142	0.442	0.052	0.081	0.523
rs16923500	8	G	A	0.240	-0.131	0.116	0.260	0.258	-0.133	0.091	0.144
rs1562042	9	A	G	0.274	-0.064	0.110	0.563	0.278	0.107	0.087	0.219
rs1802295	10	G	A	0.230	0.100	0.113	0.374	0.229	0.054	0.094	0.566
rs2616646	10	A	A	0.861	0.078	0.142	0.583	0.845	0.123	0.111	0.267
rs4287347	11	A	G	0.489	-0.239	0.096	0.013	0.481	-0.070	0.080	0.381
rs6576426	15	A	G	0.808	0.034	0.124	0.783	0.809	0.034	0.100	0.734
rs620739	15	G	A	0.723	-0.006	0.106	0.952	0.725	0.143	0.089	0.108
rs7178572	15	A	G	0.575	-0.176	0.096	0.066	0.562	0.046	0.082	0.573
rs2028299	15	A	C	0.661	-0.100	0.101	0.320	0.659	-0.147	0.084	0.080
rs4392054	16	G	A	0.127	0.048	0.144	0.741	0.117	-0.272	0.132	0.039
rs17052370	17	G	A	0.314	0.141	0.102	0.166	0.296	-0.118	0.088	0.181
rs1912379	18	A	G	0.752	0.214	0.113	0.058	0.757	0.013	0.091	0.886
rs4812829	20	G	A	0.325	0.048	0.102	0.640	0.339	0.089	0.082	0.278

Table 4.3:4 Association results for males and females for the 20 SNPs in the Sri Lankan sample set. SNPs were prioritised as part of the second stage replication process in a collaborative GWAS performed in South Asians (study A). RAF, Risk allele frequency.

4.3.4 Comparison of T2D risk variants between Sri Lankans, other South Asian cohorts and Europeans

The genotyping of 24 SNPs as part of a large multiethnic replication study in the Sri Lankan sample set gave the opportunity to compare and identify whether the allele frequencies and ORs of the SNPs are similar between Sri Lankans, other South Asians and European cohorts or whether the risk SNPs are population specific. To do this, risk allele frequencies (RAFs) for the 24 SNPS were compared between Sri Lankans and South Asians, Sri Lankan and Europeans and South Asians and Europeans, followed by the effect size estimates for 19 of the SNPs.

The Sri Lankan samples were genotyped as part of the replication phase in the Kooner et al. (2011) dataset. They made up 6.4% (841 of 13,170) of the total cases and 5.8% (1471 of 25,398) of the controls in the overall South Asian GWAS replication dataset used for the comparison (Kooner et al. 2011). As the percentage contribution of the Sri Lankan samples to the overall South Asian GWAS replication dataset was small, it was anticipated this would not substantially affect the analysis.

As described in the Introduction, the South Asian subcontinent consists of individuals of both ANI and ASI ancestry with an uncertain level of admixture between the two groups (Moorjani et al. 2013; Reich et al. 2009). To determine whether the Sri Lankan samples were genetically similar to other South Asian samples, I used data available from the 1000 Genomes phase III (Table 4.3.6) (Genomes) to compare the common allele frequency of the 24 SNPs in the Sri Lankan samples to four other South Asian subgroups (Figure 4.3.1, plots B–E), as well as the overall South Asian population (Figure 4.3.1, plot A).

The plots illustrate that there is a strong (r^2>0.85) correlation between the distribution of the common allele frequencies of the 24 SNPs in Sri Lankans and that of other South Asian samples. By performing a direct comparison between various South Asian samples, it gave the opportunity to identify those SNPs that might have different allele frequencies between groups. Overall, these data demonstrate that for the 24 SNPs compared, the frequency of the common allele is shared across the different South Asian ancestry samples. This suggests that the Sri Lankan samples are genetically similar to other South Asian samples, although this is based on 24 SNPs only and this would require further testing with a larger set of SNPs to draw a firm conclusion.

SNP	CHR	Nearest gene	A/R allele	RAF	OR (L95–U95 CI) Adjusted for age & sex	*P* (adjusted for age & sex)	OR (L95–U95 CI) adjusted for age, sex & BMI	*P*-value adjusted for age, sex & BMI
rs700952	2	—	T/C	0.07	1.07 (0.85–1.35)	0.520	1.09 (0.85–1.40)	0.463
rs11968879	6	*LRP11*	G/A	0.10	1.10 (0.90–1.35)	0.325	1.12 (0.90–1.41)	0.293
rs4527850	8	*WISP1*	C/T	0.68	0.96 (0.84–1.09)	0.558	0.98 (0.85–1.13)	0.796
rs12454246	18	—	G/A	0.96	0.79 (0.56–1.12)	0.192	0.88 (0.60–1.28)	0.512

Table 4.3:5 Association results of the four SNPs in the Sri Lankan cohort chosen as part of the third stage replication phase in a collaborative GWAS performed in Sikh Punjabi of Indian cohort. For the allele, A indicates the alternate allele, R indicates the risk allele; CHR, chromosome.

One suggestion for the higher prevalence of T2D seen in South Asians than in other ancestries might be due to greater genetic predisposition. This could be owed to (a) larger risk estimates for SNPs, (b) increased frequency of risk alleles or (c) different risk alleles yet to be determined (Sohani et al. 2014). I compared the risk allele frequency and the effect size estimates between South Asians and Europeans to determine whether these were different between the two ethnic groups for the 24 SNPs genotyped. Data for the RAF in South Asians and Europeans were attained using previous GWAS studies (Table 4.3.7) (Kooner et al. 2011; Saxena et al. 2013). Figure 4.3.2 illustrates the comparison of RAF between the Sri Lankan samples, South Asians and Europeans.

As expected, from the previous analysis using the 1000 Genomes data and because the Sri Lankans samples were part of the South Asian GWAS dataset, the strongest correlation (r^2=0.973) was observed between the Sri Lankan samples and the South Asians samples. Next, I compared the Sri Lankans samples with Europeans followed by South Asian and Europeans. The comparison between the Sri Lankan and Europeans samples gave greater granularity for each SNP, to see whether the frequency of the risk allele was greater in Sri Lankans only. If there were any deviations in RAF between the South Asian samples and Europeans, this might be due to the Sri Lankan samples. The correlation between the Sri Lankan samples and Europeans (Figure 4.3.2, plot B was weaker (r^2=0.824). However, there was no SNP with a notable difference in allele frequency to suggest that it occurs more often in either group. The comparison between South Asians and Europeans showed a stronger correlation (r^2=0.916), indicating that the frequency of the risk allele of each SNP is shared in both populations and these SNPs are unlikely to contribute to the difference in prevalence for T2D seen between South Asians and Europeans.

Effect size estimates for 19 SNPs (all from study A) were compared between the Sri Lankans, South Asians and Europeans to assess whether the direction and size of effect of the risk allele was the same across the three groups (Table 4.3.8, Figure 4.3.3). Only nine SNPs were directionally consistent across all three groups. The differences in sample size between the final South Asian GWAS and DIAGRAM+ datasets are much greater than that of the Sri Lankan cohort ($\sim n$=2300 cases and controls), which meant more precise point estimates and tighter confidence intervals were observed in the South Asians and Europeans.

SNP	Common allele in Sri Lankans	Freq. of common allele in Sri Lankans	Freq. of common allele in SAS	Freq. of common allele in BEB	Freq. of common allele in GIH	Freq. of common allele in ITU	Freq. of common allele in PJL	Freq. of common allele in STU
rs17031236	T	0.64	0.68	0.68	0.68	0.72	0.66	0.67
rs3923113	A	0.83	0.79	0.85	0.76	0.78	0.8	0.76
rs11711167	T	0.67	0.66	0.65	0.68	0.68	0.7	0.59
rs16861329	C	0.72	0.75	0.78	0.75	0.69	0.78	0.77
rs6835176	G	0.97	0.96	0.94	0.97	0.96	0.97	0.95
rs9496773	A	0.64	0.61	0.54	0.56	0.6	0.68	0.64
rs2111815	G	0.55	0.6	0.55	0.58	0.61	0.64	0.64
rs16923500	G	0.75	0.74	0.72	0.76	0.76	0.72	0.79
rs1562042	G	0.72	0.74	0.72	0.75	0.75	0.76	0.72
rs1802295	C	0.77	0.74	0.73	0.69	0.75	0.73	0.8
rs2616646	A	0.85	0.87	0.91	0.88	0.81	0.88	0.89
rs4287347	G	0.52	0.47	0.53	0.48	0.45	0.41	0.51
rs6576426	A	0.81	0.81	0.86	0.78	0.86	0.75	0.82
rs620739	A	0.72	0.67	0.64	0.67	0.67	0.68	0.69
rs7178572	A	0.57	0.52	0.51	0.58	0.59	0.4	0.5
rs2028299	A	0.66	0.66	0.58	0.64	0.7	0.67	0.7
rs4392054	C	0.88	0.88	0.88	0.83	0.9	0.92	0.88
rs17052370	C	0.70	0.71	0.7	0.71	0.68	0.73	0.74
rs1912379	A	0.76	0.74	0.76	0.73	0.73	0.79	0.68
rs4812829	G	0.67	0.7	0.69	0.69	0.72	0.73	0.66
rs700952	T	0.93	0.93	0.95	0.91	0.95	0.91	0.95
rs11968879	G	0.90	0.9	0.89	0.9	0.88	0.93	0.91
rs4527850	T	0.68	0.68	0.6	0.69	0.67	0.71	0.7
rs12454246	A	0.96	0.98	0.97	0.98	0.99	0.98	0.98

Table 4.3:6 Common allele frequencies for the 24 SNPs using the common allele in the Sri Lankan in as the reference allele. Data for the **c**ommon allele frequencies was collated using 1000 Genome Project phase III. SA, South Asian; BEB, Bengali in Bangladesh; GIH, Gujarati Indian in Houston, Texas; ITU, Indian Telugu in the UK; PJL, Punjabi in Lahore, Pakistan; Freq, frequency.

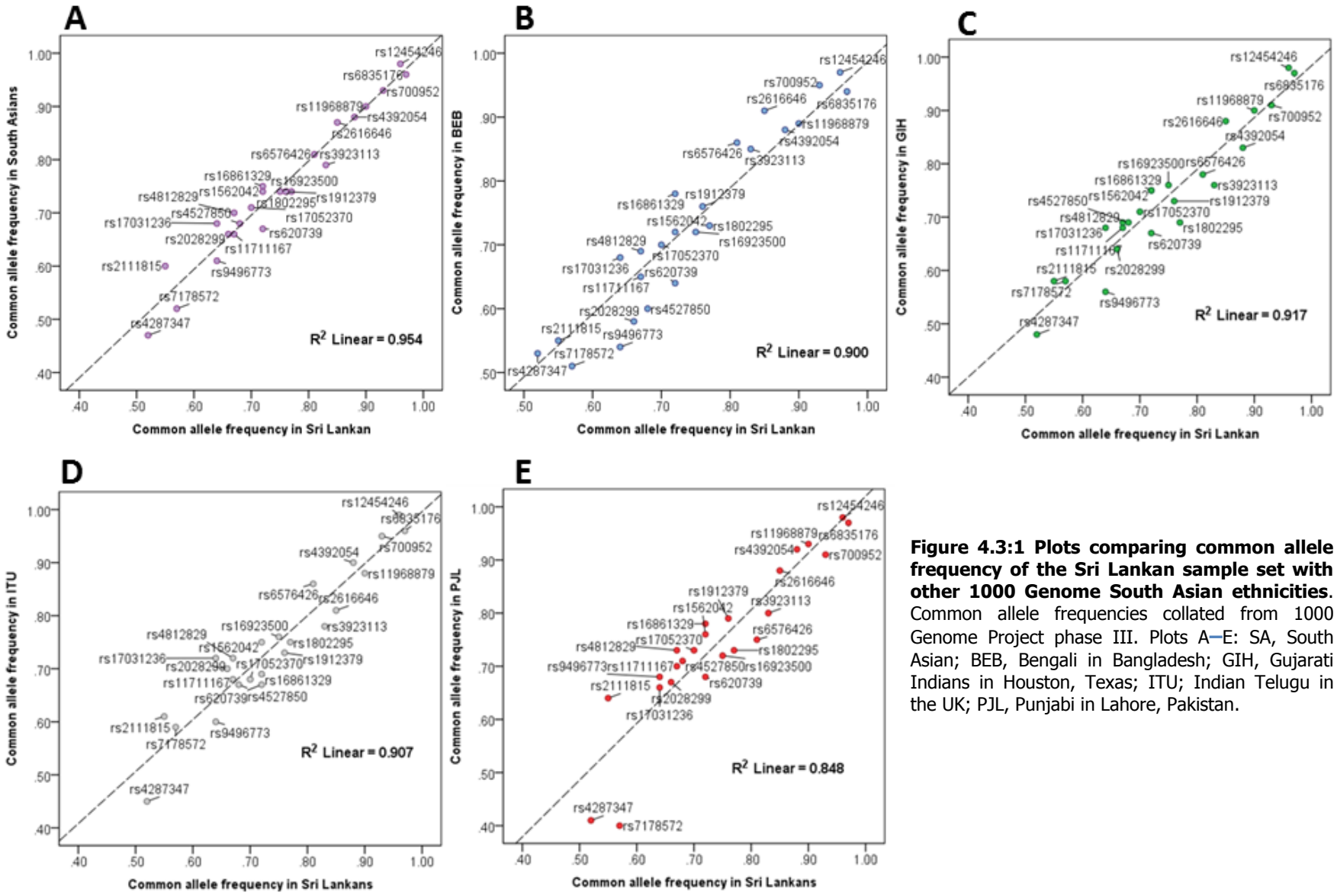


Figure 4.3:1 Plots comparing common allele frequency of the Sri Lankan sample set with other 1000 Genome South Asian ethnicities. Common allele frequencies collated from 1000 Genome Project phase III. Plots A–E: SA, South Asian; BEB, Bengali in Bangladesh; GIH, Gujarati Indians in Houston, Texas; ITU; Indian Telugu in the UK; PJL, Punjabi in Lahore, Pakistan.

Therefore, initially I compared the effect size confidence intervals to determine whether there was overlap between the Sri Lankan samples and other data. This demonstrated that 13 of the 19 (68%) SNPs had overlapping confidence intervals across Sri Lankans and South Asians, and 19/19 (100%) across Sri Lankans and Europeans. The consistency of the effect size estimates is less good, with 9/19 SNPs directionally consistent when comparing Sri Lankans and South Asians and 10/19 compared to Europeans. Between South Asians and Europeans, all SNPs were directionally consistent with the exception of rs4392054 near *ACSM1*, OR 1.17 [1.09–1.26] in South Asians, OR 0.99 [0.92–1.17] in Europeans. Taken together, these data again show good correlation between the results for the Sri Lankan only and combined South Asian T2D GWAS. To assess the impact of T2D risk variants with these small effect sizes in the Sri Lankan population, a much larger sample set will be needed. To evaluate a SNP with OR 1.1 and RAF 0.39 at alpha 0.05 (replication threshold as opposed to genome wide significant) with ~80% power, almost 3000 cases and 3000 controls will be needed (calculated using http://csg.sph.umich.edu//abecasis/CaTS/) (Skol et al. 2006).

By looking at both allele frequencies and odds ratios together, it is possible to determine the relationship between the two. South Asians have a higher prevalence for T2D than Europeans, which might be explained by variants having larger effect sizes at either end of the allele frequency spectrum, thus contributing to higher disease prevalence. Figure 4.3.4 illustrates the relationship between the odds ratio and allele frequency across the three populations. As expected the European data show that for common variants the effect size is mostly unaffected by the RAF. The variants in the Sri Lankan samples show a similar relationship between effect size and allele frequency. However, the South Asian dataset shows a slightly different pattern. These observations are confirmed by analysis of variance, which found no significant difference in the relationship between odds ratio and RAF in Sri Lankans compared to either Europeans (P=0.70) or South Asians (P=0.21), but demonstrate that there is a significant population effect between South Asians and Europeans (P=0.01). It is not possible to explore what contributes to this difference in the South Asian dataset, as it is an external data source. However, the similarity of the Sri Lankan dataset to the European samples suggests that the Sri Lankan component is not driving the difference observed in the South Asian samples.

SNP	CHR	Alternate/risk allele	Sri Lankan RAF	South Asian RAF (GWAS)	European RAF (GWAS)
rs17031236	2	C/A	0.64	0.68	0.73
rs3923113	2	C/A	0.83	0.74	0.64
rs700952	2	T/C	0.07	0.10	0.17
rs11711167	3	A/G	0.33	0.32	0.23
rs16861329	3	A/G	0.72	0.76	0.86
rs6835176	4	G/A	0.03	0.04	0.03
rs9496773	6	G/A	0.64	0.65	0.75
rs11968879	6	G/A	0.10	0.09	0.00
rs2111815	7	G/A	0.45	0.41	0.33
rs16923500	8	G/A	0.25	0.24	0.20
rs4527850	8	C/T	0.68	0.75	0.73
rs1562042	9	A/G	0.72	0.77	0.90
rs1802295	10	G/A	0.23	0.27	0.31
rs2616646	10	A/G	0.15	0.15	0.17
rs4287347	11	A/G	0.52	0.48	0.51
rs6576426	15	A/G	0.19	0.17	0.10
rs620739	15	A/G	0.28	0.36	0.43
rs7178572	15	A/G	0.43	0.52	0.71
rs2028299	15	A/C	0.34	0.31	0.28
rs4392054	16	G/A	0.12	0.12	0.09
rs17052370	17	G/A	0.30	0.26	0.19
rs1912379	18	A/G	0.25	0.25	0.16
rs12454246	18	G/A	0.96	0.98	0.98
rs4812829	20	G/A	0.33	0.29	0.16

Table 4.3:2 Summary of the risk allele frequency for the Sri Lankan sample, South Asian and Europeans datasets. Data for South Asians and Europeans were attained using previously published GWAS data and 1000 Genomes Project (phase III). Risk allele frequency GWAS data were unavailable for rs17052370, rs700952, rs11968879, rs4527850, rs12454246; therefore, HapMap GIH was used for South Asians and HapMap CEU for European data.

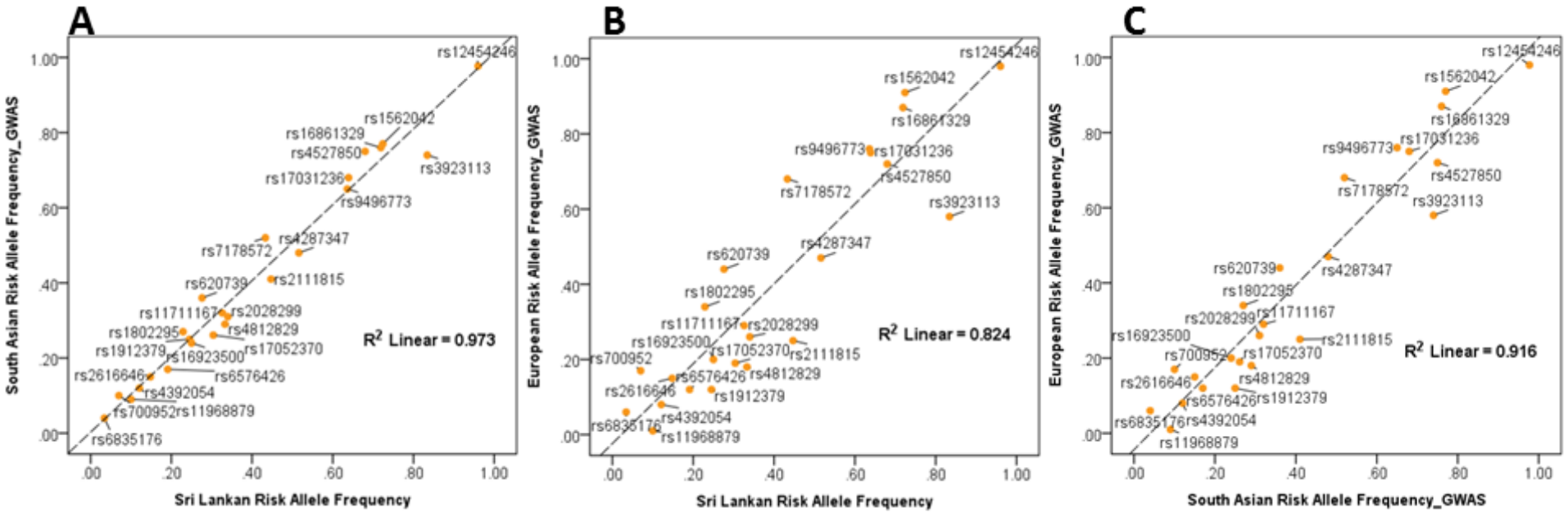


Figure 4.3:2 Plots comparing RAF between Sri Lankans, South Asians and Europeans. Data for SA and Europeans obtained from GWAS data. GWAS risk allele frequency data obtained from the original studies, which included replication in DIAGRAM (individuals of European ancestry). Plots A–C compare Sri Lankan vs South Asians, Sri Lankan vs European and South Asian vs European, respectively (Kooner et al. 2011; Saxena et al. 2013).

SNP	Chr.	Nearest gene	Sri Lankan OR (95% CI) Sample size: 2312	South Asian OR [SAT2D] *Sample size: ~15,000-18,200	European OR [DIGRAM+] *Sample size: ~18,000-22,570
rs17031236	2	*FHL2*	1.04 (0.91–1.19)	1.11 (1.05–1.16)	1.02 (0.98–1.07)
rs3923113	2	*GRB14*	0.92 (0.78–1.08)	1.15 (1.09–1.21)	1.05 (1.01–1.10)
rs11711167	3	*ZNF148*	1.05 (0.92–1.20)	1.10 (1.05–1.16)	1.02 (0.98–1.07)
rs16861329	3	*ST6GAL1*	1.19 (1.04–1.37)	1.12 (1.07–1.19)	1.02 (0.95–1.09)
rs6835176	4	*RBM47*	1.17 (0.83–1.65)	1.27 (1.14–1.42)	1.03 (0.93–1.15)
rs9496773	6	*PHACTR2*	0.97 (0.86–1.10)	1.11 (1.06–1.17)	1.04 (0.98–1.09)
rs2111815	7	*CHCHD3*	0.97 (0.85–1.09)	1.10 (1.05–1.15)	1.05 (1.00–1.09)
rs16923500	8	*CYP7A1*	0.86 (0.75–1.00)	1.12 (1.06–1.18)	1.05 (1.00–1.10)
rs1562042	9	*SMC2*	0.96 (0.84–1.10)	1.12 (1.06–1.19)	1.04 (0.97–1.11)
rs1802295	10	*VPS26A*	1.07 (0.93–1.24)	1.14 (1.08–1.20)	1.04 (1.00–1.09)
rs2616646	10	*KCNMA1*	0.90 (0.75–1.07)	1.16 (1.09–1.23)	1.02 (0.96–1.08)
rs4287347	11	*OPCML*	1.12 (0.99–1.27)	1.10 (1.05–1.15)	1.02 (0.98–1.07)
rs6576426	15	*UBE3A*	0.98 (0.84–1.15)	1.15 (1.08–1.23)	1.03 (0.97–1.10)
rs620739	15	*ATP8B4*	0.91 (0.80–1.05)	1.11 (1.06–1.16)	1.02 (0.98–1.06)
rs7178572	15	*HMG20A*	1.06 (0.94–1.20)	1.10 (1.05–1.15)	1.07 (1.02–1.12)
rs2028299	15	*AP3S2*	1.14 (1.00–1.30)	1.11 (1.05–1.16)	1.05 (1.00–1.09)
rs4392054	16	*ACSM1*	0.87 (0.72–1.06)	1.17 (1.09–1.26)	0.99 (0.92–1.17)
rs1912379	18	*RPS2P6*	0.93 (0.80–1.07)	1.14 (1.08–1.20)	1.00 (0.94–1.06)
rs4812829	20	*HNF4A*	1.07 (0.94–1.21)	1.14 (1.08–1.19)	1.08 (1.02–1.14)

Table 4.3:3 Summary of the effect size estimates that were available for 19 of the 24 SNPs. Odds ratios (OR) for South Asians and Europeans were taken from previously published data from the South Asian GWAS (SAT2D) and the replication in the DIAGRAM+ cohort for the European odds ratio (Kooner et al. 2011). *Sample size for the SAT2D and DIAGRAM+ was variable at each locus; therefore, the range is listed. Chr., chromosome.

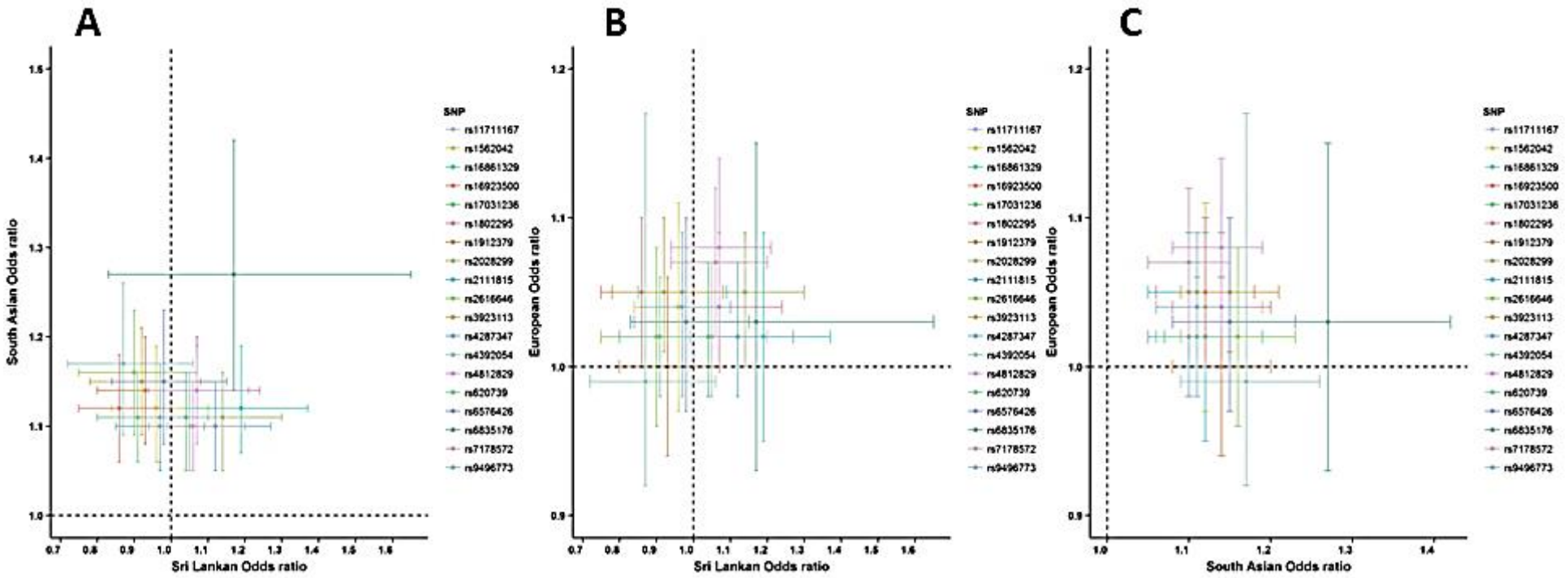


Figure 4.3:3 Plots comparing the effect size estimates for 19 of the 24 SNPs. Comparisons were made using odds ratios calculated for Sri Lankans in this study. Odds ratio for South Asians and Europeans were acquired from previously published data (Kooner et al. 2011). The vertical and horizontal threshold lines represent the 95% Confidence Intervals (CIs) at each SNP. Graphs plotted by Dr Martijn van de Bunt.

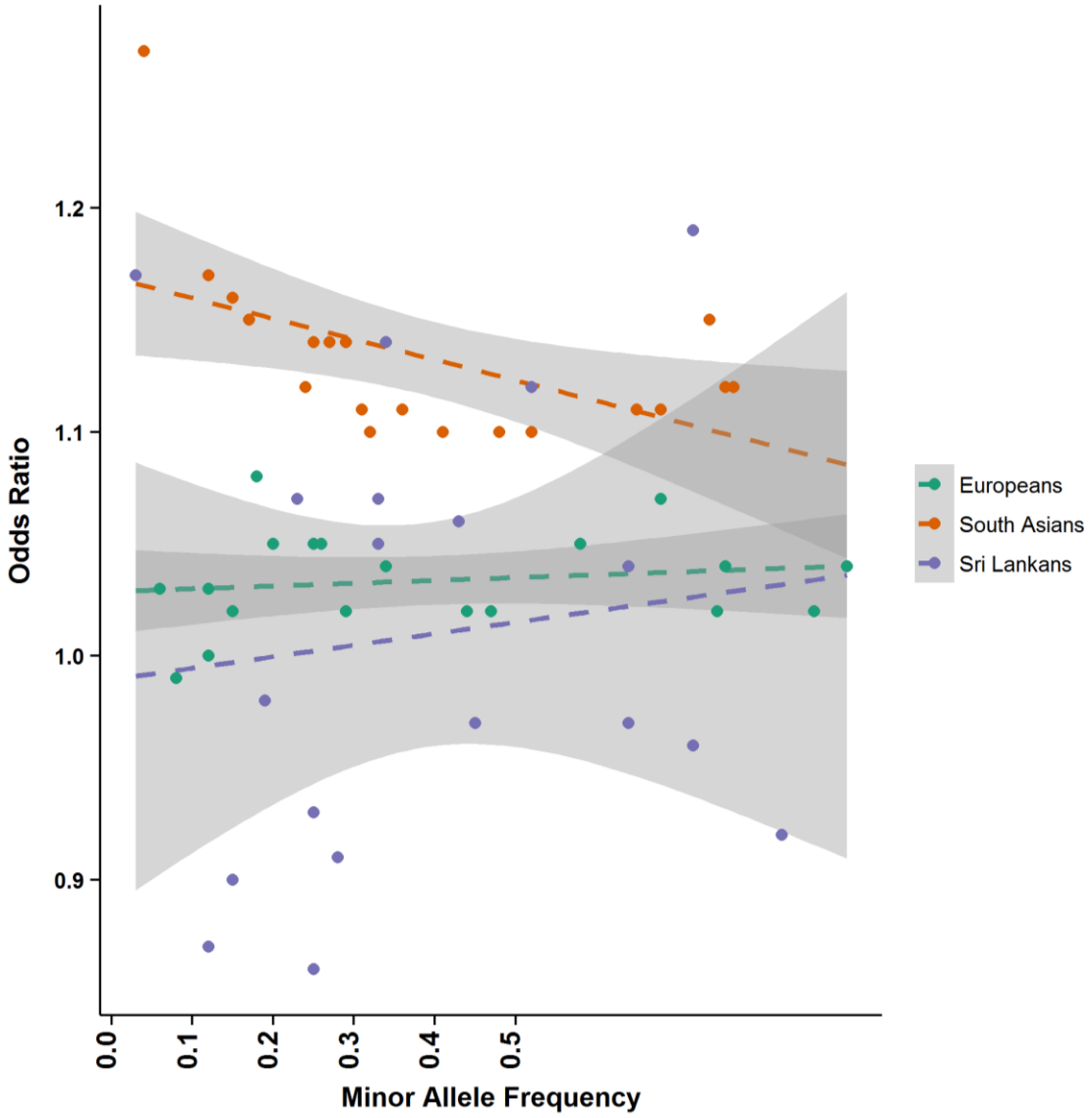


Figure 4.3:4: Plot comparing the relationship between the odds ratio and allele frequency between the three different populations. Grey shaded areas represent 95% Confidence Intervals (95%Cis) at each SNP for each ethnic group. Graphs plotted by Dr Martijn van de Bunt.

4.4 Discussion

Over the past decade large-scale genome-wide association studies have successfully identified common genetic variants (MAF >5%) that contribute to T2D risk (Morris 2014). The vast improvement in genotyping technology has enabled the investigation of the genome in many unrelated individuals with and without disease (Pearson and Manolio 2008). Given the differences in genetic architecture, disease prevalence, presentation, language and sociocultural behaviours between different ethnic groups, investigations are necessary to determine whether these common variants have the same effect across different ancestral groups or whether disease susceptibility is driven by group-specific genetic determinants (Sanghera and Blackett 2012).

The data presented in this chapter were part of two large-scale genome-wide studies focused on individuals for South Asian ancestry. Both studies comprised large sample sizes which was necessary in order to have sufficient power to detect novel variants with modest effect size (Altshuler et al. 2008).The subset of SNPs genotyped in the Sri Lankan samples as part of the replication phase exemplifies this. Irrespective of study, no SNP reached stringent genome-wide significance in the Sri Lankan samples. The sample size for the Sri Lankan sample set was modest compared with the large-scale GWAS. Power calculations recommend that for ~80% power at least 3000 (case and control) Sri Lankan samples will be required to detect SNPs with an OR 1.1 and RAF 0.39 at alpha 0.05, in contrast to the current Sri Lankan sample size, which only had 39% power. In the context of genome-wide association, the Sri Lankan sample set had no power to detect any associations.

The combination of the large population cohorts of South Asian ancestry in Kooner et al. (2011) (study A) identified six novel loci reaching genome-wide significance contributing to T2D risk. These loci were found within or near *GRB14*, *ST6GAL1*, *VPS26A*, *AP3S2*, *HMG20A* and *HNF4A* (Kooner et al. 2011). All were directionally consistent in the Sri Lankan sample set with the exception of *GRB14*. This inconsistency is likely to be attributable to the modest sample size and limited power of the Sri Lankan sample set to detect variants with a modest effect size. Of these newly identified loci, several novel biological pathways have emerged, suggesting a molecular mechanism for T2D pathogenesis.

The SNP rs16861329 is intronic within *ST6GAL1*, and showed nominal significance (P<0.05) in the Sri Lankan sample set. The gene encodes for a membrane protein

predominantly found in the Golgi apparatus and is important in glycosylation. The protein is involved in the post-translational modification of cell surface components by catalysing the transfer of sialic acid to galactose-containing substrates (Dall'Olio 2000). It is suggested that the addition of sialic acid during glycosylation might affect insulin action and cell surface trafficking, influencing glucose metabolism, and/or T2D (Kooner et al. 2011; Woodard-Grice et al. 2008). The SNP (rs16861329) is also near *ADIPOQ*, which encodes for adiponectin, a hormone secreted by adipocytes. Adiponectin mediates a number of metabolic processes including glucose metabolism by enhancing insulin sensitivity. The main mechanism by which adiponectin promotes insulin sensitivity is through increased fatty acid oxidation and inhibition of hepatic glucose production (Lihn et al. 2005).

Despite the lack of association with T2D risk in the Sri Lankan samples, the overall biological support suggests that the locus near *GRB14* is a strong candidate for association. *GRB14* is an adapter protein that binds to insulin receptors and insulin-like factors to inhibit tyrosine kinase. Compared to wild-type mice, studies in *Grb14*$^{-/-}$ mice showed that these mice had higher lean mass and better glucose homeostasis despite lower insulin and improved insulin sensitivity (Holt et al. 2009; Kooner et al. 2011). The risk allele in the (Kooner et al. 2011) study was associated with reduced insulin sensitivity, suggesting a differential mechanism in humans(Kooner et al. 2011). Furthermore, the risk allele is also in strong linkage disequilibrium with a SNP reported to be associated with central adiposity and expression of *GRB14* in adipose tissue (Heid et al. 2010). South Asians tend to have increased central adiposity in comparison to European counterparts (Misra and Shrivastava 2013). Therefore, the adverse effects mediated through *GRB14* may be increased because of the increased adiposity in the South Asians. Another favourable candidate for association was at the *HNF4A* locus. Mutations in *HNF4A* are well reported to cause MODY, resulting in impaired pancreatic beta-cell function and insulin secretion (Hattersley 1998). Here, as anticipated the risk allele in South Asians was associated with reduced pancreatic-beta-cell function (Kooner et al. 2011).

The second GWA study (study B) revealed a novel locus in the *SGCG* gene that was associated with T2D (P=1.82×10^{-8}). This locus was identified through a Sikh-only meta-analysis, combining data from stages 1, 2a and 3a. The decision for which SNPs should be replicated was chosen by the central analysis team and by chance, the associated SNP was not chosen for replication in the Sri Lankan samples; so no direct inferences

can be made as to whether the direction and size of effect is consistent in Sri Lankans. *SGCG* is a member of the sacroglycan complex of transmembrane proteins, which is highly expressed in skeletal muscle. Mice that lack sacroglycan complexes including *SGCG* in adipose and skeletal muscle were shown to be glucose intolerant and exhibited whole-body insulin resistance due to impaired insulin-stimulated glucose uptake in skeletal muscle (Barresi et al. 2000; Saxena et al. 2013). The replication at stage 3b using diverse South Asian cohorts identified four loci that showed suggestive association with T2D risk after meta-analysis (Saxena et al. 2013). These SNPs were in or near loci *HMG1L1* (P=1.55×10^{-6}), *PLXNA4* (P=6.70×10^{-5}), *SCAP* (P=7.40×10^{-5}) and Chr5.p11 (P=7.74×10^{-5}). One of these signals was in the *SCAP* gene, which plays a key role in regulating lipid metabolism (Nakajima et al. 1999). Further studies in larger samples are required to confirm these loci (Saxena et al. 2013).

South Asian populations have a higher prevalence for T2D. This usually occurs at a younger age and at lower BMI thresholds than in European counterparts. Furthermore, South Asians tend to have greater waist-to-hip ratios and larger waist sizes (Ramachandran et al. 1997). These characteristics tend to lead to a negative metabolic profile, predisposing many South Asians to higher levels of circulating insulin and a greater degree of insulin resistance (Unnikrishnan et al. 2014). The second part of the study aimed to determine whether the SNPs genotyped as part of the replication studies were ethnic-specific or shared between South Asians and Europeans. If dissimilar, this might account for some of the differences in genetic predisposition for T2D risk between the two groups. The assessment of the RAF across groups revealed slight variations but no consistent trend with any group having a greater RAF for the genotyped SNPs. The comparison of the effect sizes between the Sri Lankans, South Asians and Europeans showed that that not all SNPs were directionally consistent with effect in the Sri Lankan samples. Between South Asians and Europeans, there was a strong overlap in directional consistency for the effect size estimates. Examining the relationship between the odds ratio and RAF for the genotyped SNPs showed that there was no significant difference between the two in the Sri Lankan samples and Europeans. In the South Asian samples, there seems to be significant population effect and the effect sizes at each loci appeared to be considerably inflated. As the South Asian dataset was external, it was not possible to explore what might be driving the difference observed in the South Asians. One suggestion could be differences in population stratification between the composite South Asian subpopulations used in the South Asian GWAS, with the possibility of one cohort

driving the apparent difference. A recent trans-ethnic meta-analysis demonstrated that many of the common T2D risk variants are shared across multiple ancestries (Mahajan et al. 2014), supporting the suggestion that the causal variants in this study are most likely shared.

The identification of novel common T2D susceptibility loci in South Asians has added to the number of established risk loci whilst elucidating potential novel biological pathways underlying T2D pathogenesis. As more GWA studies with larger sample sizes are performed in South Asians it is likely that further variants with smaller effects and possibly unique to this population will be discovered (Sohani et al. 2014). One recent GWAS described by (Tabassum et al. 2013) using Indian samples identified a novel region at 2q21 harbouring three potential candidates, *TMEM163*, *RAB3GAP1* and *ACMSD*, with the strongest signal mapping to *TMEM163* (Tabassum et al. 2013). All three are involved in neuronal pathways, suggesting alternative pathways contributing to T2D pathogenesis. Furthermore, the study showed a consistent direction of effect with the six loci identified in the Kooner et al. (2011) study, providing additional evidence to suggest that common variants are shared across populations (Tabassum et al. 2013).

Many of the common genetic variants implicated in T2D risk have been found to be shared across multiple ancestral groups, implicating similar biological mechanisms increasing disease risk (Mahajan et al. 2014; Waters et al. 2010). Therefore, understanding the genetic component of why South Asians have an increased predisposition for T2D still remains unanswered. Additional genetic contributions are likely to be from rare or low-frequency variants, which require further investigations in South Asians (Sohani et al. 2014).

The absence of specific information regarding genetic variation in South Asian has been a limiting factor in the design of custom genotyping arrays (Chambers et al. 2014; Rosenberg et al. 2010). A recent study interrogating the South Asian genome using whole-exome and whole-genome sequencing identified 3 million new genetic variants (Chambers et al. 2014). The study demonstrated that common variants (AF>5%) were almost entirely shared amongst three of the populations (European, African and East Asians) in the 1000 Genomes population. Through principal component analysis, (Chambers et al. 2014) was able to show the presence of a north–south gradient between South Asians, consistent with the historic admixture between the ANI and ASI

(Chambers et al. 2014). Many of the common variants identified which were specific to South Asians were found in gene pathways responsible for carbohydrate and lipid metabolism, suggesting an enrichment of genetic variants specific in South Asians related to energy storage and metabolism (Chambers et al. 2014). Most of the novel variants identified were classified as either low-frequency (AF≥1 and <5%) or rare (AF<1%). However, to evaluate their influence on disease status will require direct genotyping as rare and low-frequency variants are poorly captured by common haplotypes and are challenging to impute (Chambers et al. 2014). The identification of population-specific genetic variants in South Asians, particularly in pathways linked to metabolic pathways, such as adipogenesis and insulin signalling, could be explained by historical compensatory mechanisms for unstable food supplies (thrifty-gene hypothesis). Although in the past these genetic variants might have been protective, changes in lifestyle and dietary habits may now cause these genetic variants to be unfavourable to health, thus contributing towards an increased risk for disease such as T2D (Chambers et al. 2014; Neel 1962). It is expected that the dataset of genetic variants specific to South Asians, generated by (Chambers et al. 2014), will support the design of customised microarrays, giving the opportunity for further study to investigate the underlying genetic susceptibility of highly prevalent diseases in South Asians, such as T2D and cardiovascular disease(Chambers et al. 2014).

In conclusion, data presented in this chapter summarise two GWA studies conducted in populations of South Asian ancestry. Both reported the identification of novel T2D loci, adding to the number of confirmed T2D susceptibility loci and highlighting novel biological pathways. Further comparisons between Sri Lankans, South Asians and Europeans demonstrated that these common variants are likely to be shared across populations and not ethnic-specific. Further studies are required in South Asians to investigate whether population-specific, low-frequency and rare variants may contribute to the differences seen in disease prevalence and presentation for T2D in this high-risk group.

Chapter 5

Functional assessment of missense hepatocyte nuclear factor-1-alpha (HNF1A) gene variants identified in subjects with biomarker profiles indicative of HNF1A dysfunction

Below is the reference to the publication that the results presented in this chapter contribute. The full manuscript is provided in the appendix section:

- Thanabalasingham, G., et al. (2013), 'Mutations in HNF1A result in marked alterations of plasma glycan profile', *Diabetes,* 62 (4), 1329-37.

5.1 Introduction

5.1.1 The importance of diagnosing monogenic diabetes in the clinic

The misdiagnosis of maturity onset diabetes of the young (MODY) in the clinic is common as the clinical characteristics of monogenic diabetes are not easily distinguishable from type 1 diabetes (T1D) and particularly type 2 diabetes (T2D). Like T1D patients MODY patients tend to be young and lean, and like T2D patients they do not require insulin treatment and are usually beta-cell antibody negative (Thanabalasingham and Owen 2011). When MODY is correctly diagnosed, it influences treatment and informs on disease prognosis for the patient and their family members; i.e. HNF1A-MODY can be treated successfully with low dose sulphonylurea (Pearson et al. 2003; Shepherd et al. 2009).

HNF1A-MODY is estimated to represent 52% of all MODY cases in the UK with diagnosis and identification occurring via genetic testing (Shields et al. 2010). The frequency at which MODY testing occurs is variable across the United Kingdom and is attributed to differences in physician awareness of MODY, the availability and cost of genetic testing and ethnic differences in diabetes prevalence and referral rates (Shields et al. 2010). Recently, plasma/serum biomarkers have been investigated as a potential method for identifying subjects with the most common subtype of MODY, specifically HNF1A-MODY, in patients with diabetes. If successful, biomarkers might be able to discriminate between HNF1A-MODY and T2D and may offer a cost-effective method for prioritising patients for genetic testing. HNF1A is a widely expressed transcription factor that plays a key role in a number of tissues, including the liver, kidney and the pancreatic beta-cell (Ferrer 2002). Therefore, determining whether allelic variants in HNF1A also result in extra-pancreatic affects might give an alternative way of recognising this subtype of diabetes (Servitja et al. 2009; Shih et al. 2001). Over the past few years, as described in the Chapter 1 (section 1.6.5), a number of candidate biomarkers have been evaluated as a discriminatory tool for HNF1A-MODY. However, many of these have had limited success as they have not been sensitive or specific enough to identify cases of HNF1A-MODY from other subtypes of diabetes.

Findings from genome-wide association studies (GWASs) have discovered novel pathways that are now being investigated which might be used to distinguish between HNF1A-MODY and common forms of diabetes. One of the first candidate biomarkers to be evaluated because of GWAS outcomes was high-sensitivity C-reactive protein (hsCRP)

(McDonald et al. 2011; Owen et al. 2010; Thanabalasingham et al. 2011). Chapter 6 of this thesis will provide more detail on hsCRP and HNF1A-MODY as it explores hsCRP as a potential biomarker in South Asians.

In 2010, Lauc and colleagues demonstrated that HNF1A regulates plasma protein fucosylation and common genetic variation near *HNF1A* influences plasma glycan profiles (Lauc et al. 2010). As with CRP, it was hypothesised that plasma glycan profiles might be a potential biomarker that could identify patients with inactivating mutations in *HNF1A*.

5.1.2 Using glycan profiles as a biomarker for HNF1A-MODY

Glycosylation of proteins is the most common post-translational modification (Marino et al. 2010). It is a complex enzymatic process whereby complex carbohydrates (glycans) are attached to proteins. This influences many functional aspects of proteins, including cellular localisation, protein turnover and quality, and protein folding (Marino et al. 2010). Fucosylation is a form of glycosylation in which a fucose residue is added to a glycan (Ma et al. 2006). The GWAS in 2010 focused on the pathways and genes that regulate one specific type of glycan, N-linked glycans (Lauc et al. 2010). The term N-linked glycans is used to refer to the attachment of the glycan residue to a protein at the nitrogen atom of an asparagine (ASN) residue via recognition of a consensus sequence (Imperiali and O'Connor 1999). The GWAS study identified HNF1A as an important regulator of fucosylation as it regulates the expression of key genes at various stages of the pathway (Lauc et al. 2010). Together with *HNF1A*, they showed that *HNF4A* also plays a role in the complex interaction in regulating fucose synthesis (Lauc et al. 2010). Functional studies demonstrated that HNF1A is involved in both the *de novo* and salvage pathways by regulating the expression of two genes encoding GDP-mannose-4-6 dehydratase (GMDS) and L-fucokinase, respectively (Figure 5.1.1).

The upregulation of both GMDS and L-fucokinase by HNF1A increases the availability of fucose in the glycosylation pathway (Lauc et al. 2010). In addition, HNF1A reciprocally regulates core versus antennary fucosylation by influencing the expression of the fucosyltransferase (*FUT*) genes, which are responsible for adding fucose to proteins or lipids (Lauc et al. 2010). HNF1A promotes the activity of *FUT* genes involved in antennary fucosylation, whilst repressing *FUT8*, which adds fucose to the core-N-acetylglucosamine (GLcNAc). Consequently, HNF1A reduces the amount of GDP-fucose for core-fucosylation and increases the availability for antennary fucosylation (Figure

5.1.1) (Lauc et al. 2010). Fucosylation is the rate-limiting step in plasma protein glycosylation and so it has been hypothesised that patients with sequence alterations in *HNF1A,* such as those found in HNF1A-MODY, may have altered plasma glycan profiles (Lauc et al. 2010).

Glycan profiles have been investigated as a novel approach to distinguish patients with HNF1A-MODY from other forms of diabetes (Thanabalasingham et al. 2013). The initial study compared the plasma glycome profiles between patients with HNF1A-MODY and those with early onset T2D patients, and identified 15 (out of 29) glycan profile measurements that were significantly ($P<0.05$) different between the two groups. As predicted, patients with heterozygous loss of function variants in *HNF1A* had an increased proportion of glycans without antennary fucose (Thanabalasingham et al. 2013). From the 15 profiles that were different between the two groups one of these, the DG9-glycan index was selected for further validation in patients with differing types of diabetes. The DG9-glycan index was based on two glycan profiles that measured triantennary glycans with (DG9) and without (DG8) antennary fucose. The DG9-glycan index is the summary of the proportion [DG9-to-(DG8+DG9) ratio] of triantennary glycans that are fucosylated (Thanabalasingham et al. 2013). The validation process identified that the DG9-glycan index (≤0.16) had 88% sensitivity and 81% specificity to discriminate between HNF1A-MODY and early onset T2D, and 88% sensitivity and 88% specificity for HNF1A-MODY and T1D (Thanabalasingham et al. 2013).

The first part of this chapter focuses on the second part of this collaborative study in which the clinical performance of the DG9-glycan index was evaluated as a potential clinical screening tool. My role was to complete the capillary sequencing of *HNF1A* in patients with young onset diabetes thought to have either type 1 or type 2 diabetes, all of whom have a DG9-glycan index suggestive of *HNF1A* dysfunction (Thanabalasingham et al. 2013)

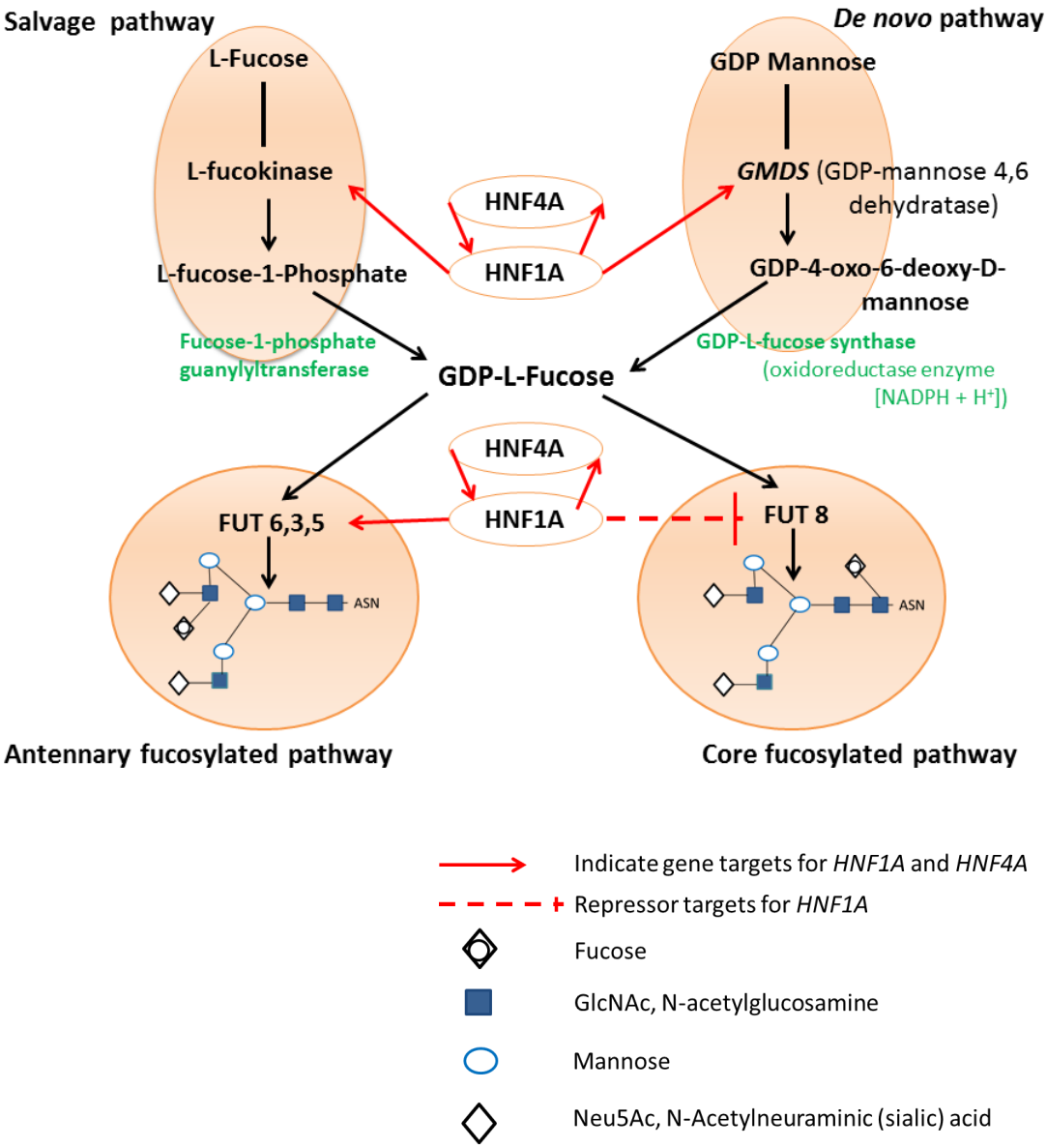


Figure 5.1.1: The role of HNF1A in the fucosylation pathway. The illustration shows that both HNF1A and HNF4A are involved in complex interactions regulating the fucosylation pathway. HNF1A up-regulates both GMDS and L-fucokinase, increasing the pool of fucose available for glycosylation. HNF1A also mediates the expression of the fucosyltransferase (*FUT*) genes, which are involved in adding fucose to proteins via two pathways. Red arrows indicate genes that are targets of HNF1A and HNF4A. Dashed red lines designate repressor targets for HNF1A. ASN; asparagine residue on glycosylated protein. Figure adapted from (Lauc et al. 2010).

5.1.3 Determining whether HNF1A variants identified in patients with biomarker profiles suggestive of HNF1A dysfunction cause HNF1A-MODY

Even though a *HNF1A* variant may be identified in a patient with a biomarker profile indicative of HNF1A dysfunction, determining whether the variant is causal for a diabetic phenotype requires further investigation. In some instances, allelic variants within *HNF1A* are characterised as causing HNF1A-MODY. More recently, studies have indicated that the same *HNF1A* variant is causal for a spectrum of diabetic phenotypes and can be categorised as truly causal for MODY, influencing T2D risk, or neutral (Flannick et al. 2013). For example, studies have proposed that the common coding SNP p.A98V in *HNF1A* might influence T2D risk (Bergmann et al. 2008; Holmkvist et al. 2006; Steinthorsdottir et al. 2014). Another recent study identified a low-frequency variant p.E508K that was found to be associated with a fivefold increase for T2D prevalence in the Latino population (Estrada et al. 2014). The recent study by Flannick et al. (2013) looked at rare variants in the general population, in a subset of genes that cause monogenic diabetes (including *HNF1A*). Crucially, they identified low-frequency variants in MODY genes previously implicated as causal for MODY in individuals who remained euglycemic even at middle age(Flannick et al. 2013). The study highlighted that the classification of variants based on bioinformatics, previous publications, and mutation databases did not provide sufficient evidence to determine an individual's disease risk (Flannick et al. 2013). For example, the cataloguing of human genetic variation via the 1000 Genomes Project, the Exome Sequencing Project (ESP) and the Exome Aggregation Consortium (ExAC) has been extremely useful in providing information about whether the variants are present and at what frequency in a range of populations. However, it does not provide any details regarding the functional contribution of the variant to protein function (Abecasis et al. 2012; ESP ; ExAc). The Flannick et al. (2013) study suggested that in order to gain a comprehensive understanding of which variants have a large effect size increasing disease risk, family studies (even though these may be impractical) and functional characterisation of variants will be required(Flannick et al. 2013).

The second part of this Chapter explores whether variants in *HNF1A*, identified in individuals with glycan profiles indicative of HNF1A dysfunction, affect transcription factor function using a suite of *in vitro* assays. In combination with family studies, I then use these data to determine whether it is possible to ascertain whether these *HNF1A* variants are causal for diabetes.

5.1.4 Experimental aims

The aims of this study were

1. To identify whether glycan profiles can be used as a marker for HNF1A dysfunction by screening patients diagnosed with T1D or young-onset T2D whose glycan profiles are indicative of HNF1A dysfunction; and
2. To characterise functionally *HNF1A* variants *in vitro* to establish whether these variants affect HNF1A function.

5.2 Samples and methods

5.2.1 Subjects studied

This study was part of a large collaborative study that evaluated different glycan indices to find an appropriate marker that might differentiate between HNF1A-MODY and other common forms of diabetes. As part of this study I screened individuals with young onset diabetes (≤ 45 years old) clinically labelled as either T1D or T2D. All participants had a DG9-glycan index ≤0.16, which is defined as the proportion of triantennary glycans that are fucosylated [DG9-to-(DG8+DG9) ratio]. The subjects sequenced had not been previously suspected of having HNF1A-MODY. Capillary sequencing was performed in these subjects to evaluate whether the DG9-glycan index level of ≤0.16 was a useful screening marker to detect variants in *HNF1A* (Thanabalasingham et al. 2013).

All 10 exons of the *HNF1A* gene were screened in 97 individuals from four European centres (detailed below). DNA amplification and subsequent bi-directional sequencing was performed on DNA samples shipped from the relevant centres.

5.2.1.1 Young Diabetes in Oxford (YDX) Study

Twenty-two subjects were sequenced from this sample set. All YDX patients have been ascertained from the South of England. All participants had a provisional clinical diagnosis of T2D at ≤45 years of age. In addition, all T2D patients were positive for C-peptide, had no requirement for permanent insulin within 3 months of diagnosis, and had negative glutamic acid decarboxylase (GAD) antibodies. The criteria for patients labelled with T1D included permanent insulin therapy from diagnosis, a C-peptide <0.1 nmol/L and/or being positive for GAD antibodies. The study was approved by the Oxfordshire Local Research Ethics Committee and all subjects gave informed consent (Thanabalasingham et al. 2013).

5.2.1.2 The Orkney Complex Disease Study (ORCADES)

Twenty-one subjects were from the Orkney Complex Disease Study (ORCADES). This study was performed in the Scottish archipelago of Orkney and data were collected between 2005 and 2011. Plasma samples from 889 participants aged 18 to 100 years from a subgroup of ten islands were available for the study. Diabetic status was determined in one of the following ways: self-reported diabetes, treatment with an anti-diabetic medication regardless of glycaemic status or a record of HbA1c measurement that was greater or equal to 6.5% (47.5 mmol/mol). All subjects signed consent prior to participation and the study followed the ethical guidelines of the 1975 Declaration of Helsinki (Thanabalasingham et al. 2013).

5.2.1.3 Subjects from the CROATIA-Vis and CROATIA-Korcula studies

Thirty-six subjects were from two Croatian cohorts. Members of the CROATIA-Vis study were recruited between the ages of 18 and 93 years from two villages, Vis and Komiza, on the Dalmatian island of Vis during 2003 and 2004. Subjects in the CROATIA-Korcula study were recruited in 2007 and 2008 and were between the ages of 18 and 98 years. Subjects in both cohorts were enlisted irrespective of any specific phenotype. All information was collected in a similar manner across studies, including fasting plasma samples, biochemical and physiological measurements and medical questionnaire data. Diabetic status was determined by reports of diabetes recorded in the medical history, treatment with an anti-diabetic medication regardless of glycaemic status or HbA1c measurement equal to or greater than 6.5% (47.5 mmol/mol). All participants gave signed consent and all studies conformed to the ethical guidelines of the 1975 Declaration of Helsinki (Thanabalasingham et al. 2013).

5.2.1.4 Subjects from the Danish study

Eighteen subjects labelled with T1D or T2D were identified from the outpatient clinic at the Steno Diabetes Centre. The T2D patients were diagnosed with diabetes ≤45 years old, were C-peptide positive, with no requirement for insulin within 12 months of diagnosis, and had negative GAD antibodies. The T1D patients had HbA1c >8% (63.9 mmol/mol), were GAD antibody positive and/or C-peptide <10pmol/L and had diabetes for 3 years or more. All subjects were of North European ethnicity and all subjects provided written consent (Thanabalasingham et al. 2013).

5.2.2 PCR and gel electrophoresis of HNF1A exons

Genomic DNA samples were amplified by PCR using primers for specific exons detailed in Table 5.2.1. All primers were tested to establish optimal Mg^{2+} concentrations and

annealing temperatures by carrying out a magnesium titration and temperature gradient PCR. All samples underwent PCR using the Ampli Taq Gold (Applied Biosytems) method. The thermo-cycling programme used was the same for all exons and is described in Chapter 2 (section 2.3). All PCR reactions were checked on 2% agarose gel (5g agarose (SIGMA), 250µl TrisBorateEDTA (TBE-10×) and 5µl Ethidium bromide). Expected band sizes for each exon were checked against 100bp ladder (New England BioLabs). PCR products were stored at -20°C until required for sequencing.

5.2.3 Capillary sequencing of HNF1A exons

All PCR products underwent bi-directional sequencing using M13 tailed primers. Sequencing was performed as described in Chapter 2 (section 2.4), on an ABI 3700 platform. All analysis was carried out using Mutation Surveyor v3.97 (SoftGenetics, Cambridge, UK). The reference sequence to which all samples were compared (NM_000545.3) was downloaded from the Ensembl website (http://www.ensembl.org).

5.2.4 Assessing the pathogenicity of HNF1A variants identified

To assess the pathogenicity of the variants identified, five main approaches were employed. First, family studies if possible were performed to confirm whether variants co-segregated with diabetes. Second, *in silico* prediction tools were used to evaluate the effect of the amino acid substitution on protein structure as described below in section 5.2.5. Third, the conservation for each variant across nine different species was assessed using ClustalW, since variants affecting amino acid residues, which are highly conserved throughout evolution, are more likely to be pathogenic. Fourth, the presence of each variant was checked in 1000 Genomes (Phase 3 release), the Exome Sequencing Project (ESP) and the Exome Aggregation Consortium (ExAc) browser. The ExAc dataset contains subjects with T2D and possibly MODY; however, as the phenotypes of the participants in these datasets are unknown, limited inferences can be made about the variants and disease status. Finally, *in vitro* functional studies were established to determine whether the variants affected HNF1A function.

5.2.5 In silico prediction tools

As described in Chapter 2 (section 2.18), three *in silico* prediction tools PolyPhen2, SIFT and Condel were used to evaluate whether the identified HNF1A variants affected protein function

Exon	Forward (M13 tailed) (All primers start 5′ TGTAAAACGACGGCCAGT)	Reverse (M13 tailed) (All primers start 5′ CAGGAAACAGCTATGACC)	Size (bp)
1	GCAGTGGGTGCAAGGAGT	GCCCCTCTAGGCTCTCCTG	528
2	CAAGTCTCTGTCCTCATGACC	TCCCACTGACTTCCTTTCC	588
3	GGGCAAGGTCAGGGGAATGGA	CAGCCCAGACCAAACCAGCAC	342
4	CAGATCTGCCAGCCTCAAACC	ACCTTCCCACGTGTCCCTTG	434
5	TGGAGTTTGAAGTGCTGAGG	GCCAAGGAAAGATGAGGTTG	320
6	TGTAAGGAAAACCCAACCTCA	AGTGGCTCTTCCAGCTCCT	430
7	GCTCTGGGAAGGAGAGGTG	CATCCATTGACAGCCAACC	346
8	AGGCTCCCCTGCATCAAG	GGGTGAGGGGCTCTGTCA	398
9	ACCAAGCAGGTAAGGTCCAG	GTGACGGACAGCAACAGAAG	331
10	GTACCCCTAGGGACAGGCAGG	ACCCCCCAAGCAGGCAGTACA	283

Table 5.2.1: Primer sequences of primers used to amplify exons in *HNF1A.*

5.2.6 Genotyping of D526N in the ORCADE population

Genotyping was performed using TaqMan technology to assess whether the p.D526N (c.1576 G>A) variant was enriched in the Orkney population. DNA samples from ORCADES were sent over from the laboratory of Professor James Wilson (Edinburgh University). A customised TaqMan assay was designed with VIC and FAM labelled probes using NCBI and Primer Express v3.0 (Applied Biosystems). The primer sequences generated were D526N_F CGAGGTGGCCCAGTACAC and D526N_R GCCTGGACCTTACCTGCTT.

DNA samples from 2147 individuals were genotyped and genotyping calls determined using the methodologies described in Chapter 2 (sections 2.4.2 and 2.4.3). The original proband with the p.D526N variant was included as a positive control.

5.2.7 Site-directed mutagenesis and construction of HNF1A mutant proteins

In vitro studies were performed using a pRK5 expression construct comprising the full-length cDNA for human *HNF1A* (NM_000545) with an in-frame N-terminal FLAG tag. This construct was generated by Dr Jana Rundle, a post-doctoral researcher in the Gloyn group (OCDEM, University of Oxford). Briefly, the vector was made using an Image Clone as a PCR template for *HNF1A*. An EcoRI and a FLAG-TAG (DYKDDDDK) both 5' were incorporated into the *HNF1A* sequence in addition to a 3' BamHI recognition site. The PCR product was cloned into a TOPO2.1 vector, and sequenced using M13F and M13R sequencing primers. A S1A custom primer was used for internal sequence verification. Once confirmed the EcoRI and BamHI restriction enzymes were used to clone *HNF1A* into the pRK5 vector. The cDNA sequence for *HNF1A* contained four SNPs, three of which caused no change in amino acid sequence. One SNP (rs1169288) at c.79 A>C resulted in an amino acid sequence change p.I27L (isoleucine to leucine). Within the 1000 Genomes Project this SNP has a minor allele frequency (MAF) of 33% in Europeans (Abecasis et al. 2012; Colclough et al. 2013). Previous studies have evaluated this SNP in the context of its contribution to T2D risk and the results have been inconsistent. Therefore, this SNP was mutated back to wild-type (27I C>A) (Holmkvist et al. 2006; Shaat et al. 2006; Winckler et al. 2005). All mutagenesis was carried out using the QuikChange II Site-Directed mutagenesis kit (Agilent Technologies Inc, Santa Clara CA) using primers designed complementary to the mutations listed in Table 5.2.2.

HNF1A variant	Forward Primer	Reverse Primer
L27I	GAGCAAAGAGGCACTG**A**TCCAGGCACTGGGTG	CACCCAGTGCCTGGA**T**CAGTGCCTCTTTGCTC
G47R	CCCCTGGACAAG**C**GGGAGTCCTGCG	CGCAGGACTCCC**G**CTTGTCCAGGGG
R203H	GAAGGGGCGGAGGAACC**A**TTTCAAGTGGGGCCCAG	CTGGGCCCCACTTGAAA**T**GGTTCCTCCGCCCCTTC
A251T	GTCCCCATCACAG**A**CACAGGGGCTGGG	CCCAGCCCCTGTGTC**T**GTGATGGGGAC
D526N	GACTATGCTCATCACC**A**ACACCACCAACCTGAG	CTCAGGTTGGTGGTGT**T**GGTGATGAGCATAGTC
P112L	CCTTCTGCAGGAGGACC**T**GTGGCGTGTGGCG	CGCCACACGCCAC**A**GGTCCTCCTGCAGAAGG
R271W	CAACTGGTTTGCCAAC**T**GGCGCAAAGAAGAAGCC	GGCTTCTTCTTTGCGCC**A**GTTGGCAAACCAGTTG

Table 5.2:2: Primer sequences used for site directed mutagenesis. The purification for each primer was salt free. The letter in red denotes the nucleotide that was incorporated into the sequence during mutagenesis.

Details of the mutagenesis steps are given in Chapter 2 (section 2.6). All generated plasmid constructs were verified by DNA sequencing prior to commencement of expression studies as described in Chapter 2 (section 2.12).

5.2.8 Transactivation assays

The Promega Dual-Luciferase Reporter Assay System was used to complete all transactivation studies. All assays were measured 24 hours post transfection using a Veritas Microplate Luminometer (Promega, Madison, WI, USA). All assays were performed in HeLa and INS-1 cells using Lipofectamine 2000 as the transfection reagent. Details of transient transfections are as described in Chapter 2 (section 2.14). The methodology employed for the transactivation assays was the same for each promoter construct and these have been outlined in Chapter 2 (section 2.15).

5.2.8.1 Generation of the human albumin and HNF4A-P2 promoter constructs

HNF1A regulates a number of genes within different tissues. In the liver, it regulates many liver-specific genes and in the pancreas, HNF1A is involved in regulating genes linked to glucose metabolism and transport (Galan et al. 2011). To investigate if there were tissue-specific differences in transcriptional activity between wild-type and variant HNF1A two tissue-specific promoter constructs were used. The first encompassed the minimal sequence of the human albumin promoter. The albumin promoter contains six *cis* elements A–F that act as binding sites for transcription factors. To gain optimal liver-specific transcriptional activity three of these elements must be bound to proteins, one of which is HNF1A. Elimination of one of these sites or disruption of one of these key proteins causes a considerable loss in promoter activity (Lichtsteiner and Schibler 1989 ; Mendel and Crabtree 1991).

The albumin reporter construct used in the transactivation experiments was cloned into the pGL3 luciferase vector by Dr Jana Rundle (OCDEM, University of Oxford). Briefly, using restriction enzymes KpnI and NcoI the SV40 enhancer and albumin promoter was cut out of the pDRIV5s-SV40-hALB. The pGL3-basic vector (Promega) which contains the luciferase gene was opened up using KpnI and NcoI (Promega) and the promoter fragment was left to ligate in an overnight reaction at 4°C. The next day the ligated fragments were transformed and colonies grown overnight at 37°C as described in chapter 2 (section 2.9). The following day colonies were picked and grown and DNA was prepared from bacterial cultures. The new pGL3-hAlb promoter construct was sequenced to verify that the entire albumin promoter sequencing was correctly inserted into the pGL3 luciferase vector.

The second promoter, HNF4A promoter (P2), was sent from Professor Pal Njølstad's group in Norway (University Hospital, Bergen). In the pancreas, *HNF4A* expression is regulated by *HNF1A*. In the liver, the absence of *HNF1A* does not alter *HNF4A* expression (Hansen et al. 2002). It is thought that this differential regulation observed in the pancreas is due to the presence of a tissue-specific promoter, HNF4A-P2 (Hansen et al. 2002; Thomas et al. 2001). Studies in mouse islets have shown that the P2 promoter has a high affinity-binding site for hnf1a, suggesting that hnf1a directly mediates expression of hnf4a in pancreatic islets (Boj et al. 2001). Prior to using this promoter, the construct was sequenced to confirm the presence of the HNF4A-P2 sequence.

5.2.8.2 Optimisation of experimental conditions for each promoter construct

To determine optimal experimental conditions for the transactivation assay, titrations of the experimental construct and the *HNF1A* expression construct were performed in both HeLa and INS-1 cell lines. Firstly, increasing amounts of the albumin and HNF4A-P2 promoter constructs alone were transfected, to identify a DNA concentration that produced minimal luciferase activity in the basal state. Increasing amounts of wild-type *HNF1A* expression construct were then co-transfected with the decided amount of promoter construct to identify a DNA concentration that provided maximum fold induction of luciferase activity above background.

With the exception of the *HNF4A-P2* promoter construct in INS-1 cells, all cells were co-transfected with 0.25µg of either wild-type or mutant pRK5-HNF1A, 0.5µg of human albumin/HNF4A-P2 luciferase reporter construct (Promega) and 0.01µg of pRL-SV40 (Promega). Empty pGL3 vector was used to make the final DNA concentration to 1µg. The outcomes of the optimisation studies for the HNF4A-P2 promoter are provided in section 5.3, below.

5.2.9 Protein expression and subcellular localisation

5.2.9.1 Transactivation assay lysate western blot analysis

Protein expression studies were performed using the HeLa cell lysates collected from the transactivation assays using the albumin reporter construct. As described in chapter 2 (section 2.16.1), representative western blots were completed using one triplicate well from two separate luciferase assays. The Bio-Rad gel and imaging system (Bio-Rad, Hertfordshire, UK) were used to assess HNF1A protein levels.

5.2.9.2 Cytosolic and nuclear fractionation protein extracts

Cytosolic and nuclear fractions were prepared using whole cell extracts. Western blot analysis was then performed to evaluate whether the *HNF1A v*ariants altered correct nuclear localisation of the HNF1A protein. All relevant details of this are provided in Chapter 2 (sections 2.16). All membranes were blocked using 20ml 5% Marvel non-fat milk powder and Tris buffered Saline and Tween 20 (TBST). Using the N-terminal FLAG tag appended to the HNF1A cDNA sequence in the pRK5 vector, a primary FLAG antibody was used to visualise HNF1A protein. The monoclonal anti-FLAG (Sigma-Aldrich Company Ltd, Gillingham, UK) was used at 1:50,000 concentration and incubated overnight at 4^{o}C. The following day membranes were blocked for 1 hour using rabbit anti-mouse IgG (H+L) secondary antibody horseradish peroxidase (HRP) conjugate (Thermo Scientific, Rockford, IL, USA) at a 1:2,500 concentration.

To assess whether there was any contamination between the cytosolic and nuclear fractions, two antibodies were chosen to act as quality controls. The first, for cytosolic fractions, was anti-GAPDH (Santa Cruz, CA, USA), which was used at a concentration of 1:4,000 incubated overnight at 4^{o}C. The following day the membrane was blocked with anti-goat (Santa Cruz, CA, USA) using a 1:2,500 dilution. The second primary, for the nuclear fraction, was Topoisomerase IIα antibody (Cell Signalling Technology, Danvers, MA, USA) used at 1:2,500. Following an overnight incubation at 4^{o}C the membrane was blocked for 1 hour using goat anti-rabbit IgG (H+L) secondary antibody, HRP conjugate (Thermo Scientific, Rockford, IL, USA) at 1:2,500 concentration. These antibodies were probed on subsequent days and all images were captured using the Chemi-doc imager (Bio-Rad, Hertfordshire, UK).

5.2.9.3 Immunofluorescence

The methodology used for immunostaining is described in Chapter 2 (section 2.17). All images were viewed using the confocal microscope (Nikon model, Eclipse, E600FN) using Bio-Rad coloured filters (red, green, and blue). Software available online, Zeiss LSM Image browser, was used to superimpose all images (http://www.zeiss.co.uk/microscopy/en_de/downloads/lsm-5-series.html).

5.2.10 Statistical analysis

To compare the mean fold activation between wild-type HNF1A and each mutant, an analysis of variance (ANOVA) was performed. Average luminescence values for each experiment were log transformed to normalise the data prior to running a univariate analysis. An ANOVA was used as it allowed one single test to compare multiple *HNF1A*

variants relative to wild-type. IBM SPSS v20.0 was used to perform all analyses. An unpaired t-test was performed to compare differences in protein levels for wild-type and variant HNF1A observed in the western blot analysis. A *P*-value <0.05 was considered significant.

5.3 Results

5.3.1 Identification of HNF1A variants, follow-up family studies and in silico predictions

DNA from all 97 subjects was amplified for all 10 exons of *HNF1A* via PCR. All PCR products were visualised using standard gel electrophoresis. This confirmed that the PCR reaction had worked for each exon in all individuals prior to commencement of DNA sequencing. Completion of capillary sequencing across all 10 *HNF1A* exons identified 4 individuals with heterozygous missense variants (Figures 5.3.1–5.3.4). Table 5.3.1 provides a summary of the clinical characteristics of the patients and the *in silico* evaluation for each variant.

The first variant p.G47R (c.139 G>C p.Gly47Arg) in exon one was identified in an individual from the YDX cohort. The proband was diagnosed with presumed T2D at age 37 years. Additional capillary sequencing identified the same variant in an affected parent who was diagnosed with diabetes at 70 years (Figure 5.3.1B). Although the proband was diagnosed with diabetes at a young age (37 years), which is consistent with a diagnosis of MODY, the affected parent's diabetes was diagnosed at a later age. Conservation data showed that this residue is not conserved and *in silico*, prediction suggests that this variant is unlikely to affect protein function (Table 5.3.1). The p.G47R variant was reported in a single European American individual from the NHLBI Exome Sequencing Project and was identified once in the ExAC dataset (Table 5.3.1). Similar to that of the 1000 genomes the phenotypic characteristics of subjects in ESP are not declared; therefore, the p.G47R variant may be pathogenic.

The p.R203H (c.608 G>A p.Arg203His) was identified in a patient from the YDX cohort. This variant in exon 3 has been previously described in 19 other families, suggesting that this variant is casual for a MODY phenotype (Colclough et al. 2013). Further capillary sequencing of DNA from two affected siblings confirmed they also carried the p.R203H variant. Clinical data revealed a two-generation history of young onset diabetes (Figure 5.3.2B). The proband was treated with insulin and thought to have T1D when diagnosed at age 31 years. Further investigation showed the proband to have residual

endogenous insulin secretion 17 years after diagnosis (C-peptide, 0.27 nmol/L). Since the patient has been identified as carrying the p.R203H variant, it has been possible to alter diabetes treatment for one sibling by successfully stopping basal-bolus insulin. This patient's diabetes is now well controlled by taking the sulphonylurea gilclazide. It is hoped that that this therapy change will also be considered by the proband. Conservation analysis and *in silico* data both support this variant as being pathogenic (Table 5.3.1), and the p.R203H variant was not seen in the 1000 Genomes Project, ESP or ExAC. Overall, the family data and the successful modification of treatment to sulphonylurea suggest that p.R203H is a genuine MODY mutation and the cause of diabetes in this pedigree.

Variant p.A251T (c.751 G>A p.Ala251Thr) in exon four was identified in an individual from the CROATIA-Korcula cohort. This proband was diagnosed with diabetes at 43 years and is being treated with sulphonylurea and metformin therapy 25 years later. No follow-up genetic studies were possible for this proband. Family details indicate that the proband's mother and sister had diabetes in old age (Figure 5.3.3B). There have been no observations of p.A251T in the 1000 Genomes Project, ESP or ExAC. Conservation analysis indicated that this variant is not highly conserved across species and there were inconsistences amongst the three *in silico* prediction programmes.

The p.D526N (c.1576 G>A Asp526Asn) variant located in exon 8 has been previously described as a MODY-causing mutation (Bellanne-Chantelot et al. 2008). This missense variant was seen in an individual from the ORCADES cohort. The proband was diagnosed with diabetes at 63 years, which is inconsistent with a diagnosis of MODY. Further capillary sequencing of family members failed to demonstrate co-segregation with a clinical phenotype consistent with MODY (Figure 5.3.4B). The variant was present in an unaffected sibling who at the age of 80 years, despite having an elevated HbA1c (6% or 42.1 mmol/mol), has not yet been clinically diagnosed with diabetes. There was no evidence of p.D526N in the 1000 Genomes, ESP or ExAC projects. Although the conservation and *in silico* data (Table 5.3.1) suggest that this variant may have a deleterious effect on protein function, family studies indicate that this variant is unlikely to be a fully penetrant MODY mutation.

Exon	Nucleotide position & substitution	Protein Change	Previously reported	PolyPhen 2 prediction	SIFT prediction	Condel prediction	Highly conserved across species*	Present in 1000 Genomes (Phase 3 release)	Present in the ESP data	Present in ExAC (Version 0.3)	Evidence for co-segregation
1	c.139 G>C	Gly47Arg	No	Benign	Tolerated	Neutral	No	No	Yes (one EA individual)	Yes	Inconclusive
3	c.608 G>A	Arg203His	Yes	Probably damaging	Affect protein function	Deleterious	Yes	No	No	No	Yes
4	c.751 G>A	Ala251Thr	No	Probably damaging	Tolerated	Deleterious	No	No	No	No	Family studies not completed
8	c.1576 G>A	Asp526Asn	Yes	Probably damaging	Affect protein function	Deleterious	Yes	No	No	No	No

Table 5.3.1 *HNF1A* missense variants identified through capillary sequencing. *Highly conserved is defined as no variation observed across seven orthologous genes (includes at least five mammalian orthologues plus a species of frog and fish) (Flanagan et al. 2010). EA, European American; ESP, Exome Sequencing Project; ExAc, Exome Aggregation Consortium.

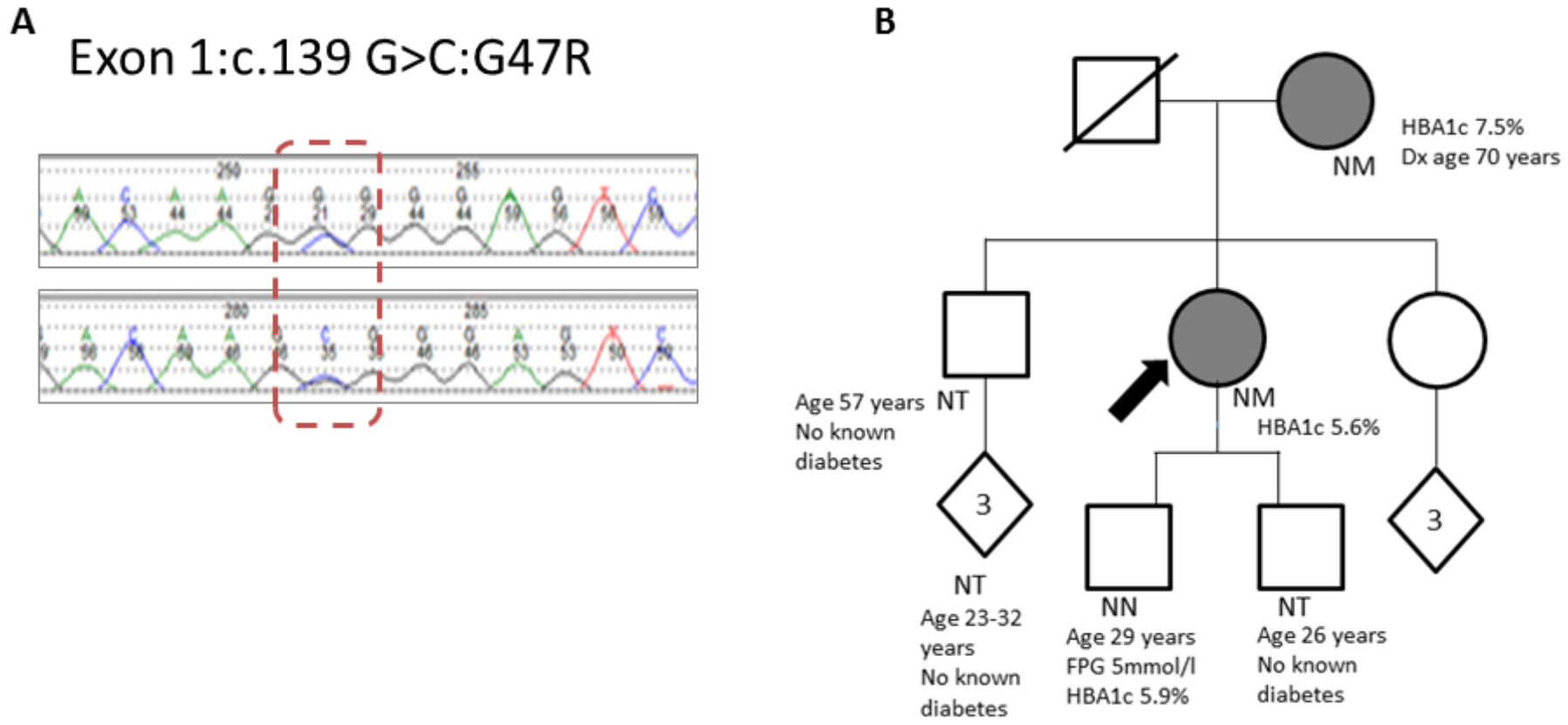


Figure 5.3.1: Electropherogram from capillary sequencing identifying the p.G47R variant and the family pedigree of the proband's. Squares represent males and circles females. Filled symbols indicate diabetic subjects. Diamonds represent number of offspring, sex unknown. NM, heterozygous mutation carrier; NN, no mutation; NT, not tested; Dx, age of diagnosis; FPG, fasting plasma glucose. The black arrow designates the proband.

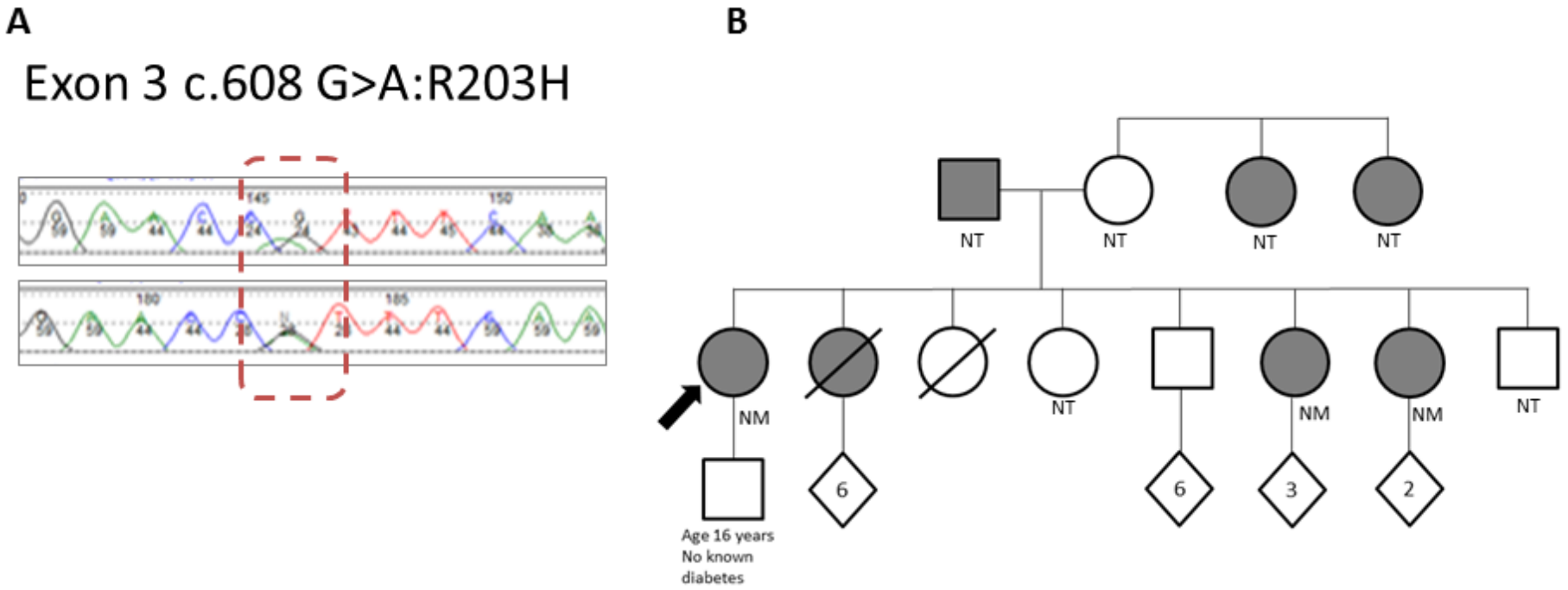


Figure 5.3.2: Electropherogram from capillary sequencing identifying the p.R203H variant and the family pedigree of the proband's. Squares represent males and circles females. Filled symbols indicate diabetic subjects. Diamonds represent number of offspring, sex unknown. NM, heterozygous mutation carrier; NN, no mutation; NT, not tested. The black arrow designates the proband.

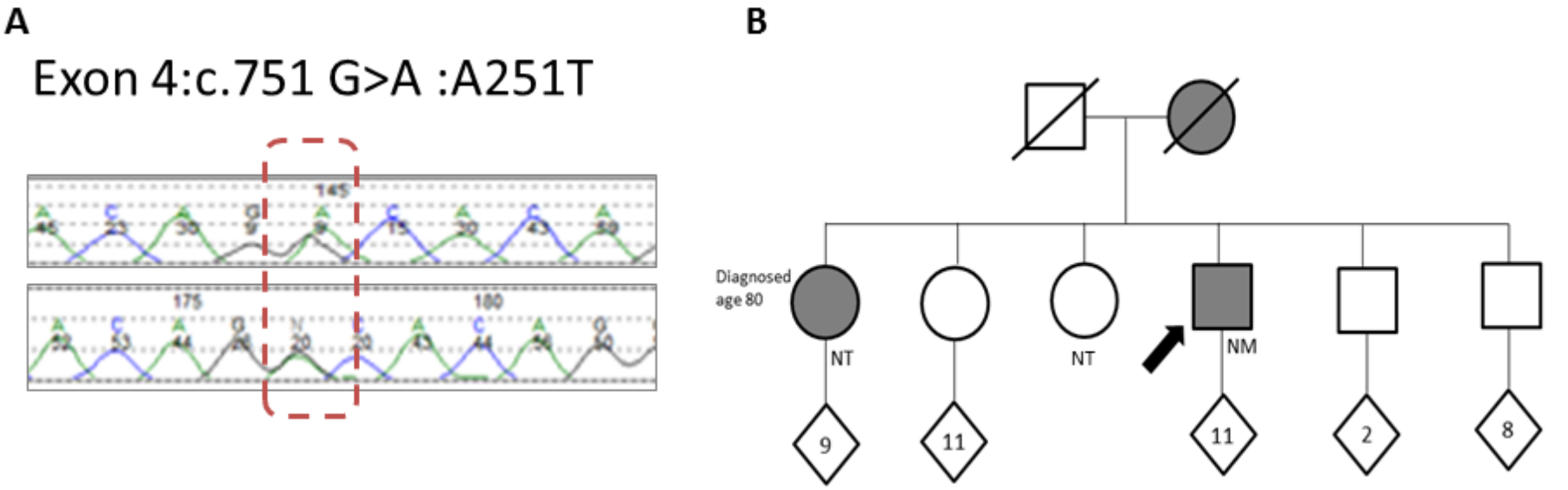


Figure 5.3.3: Electropherogram from capillary sequencing identifying the p.A251T variant and the family pedigree of the proband's. Squares represent males and circles females. Filled symbols indicate diabetic subjects. Diamonds represent number of offspring, sex unknown. NM, heterozygous mutation carrier; NN, no mutation; NT, not tested. The black arrow designates the proband.

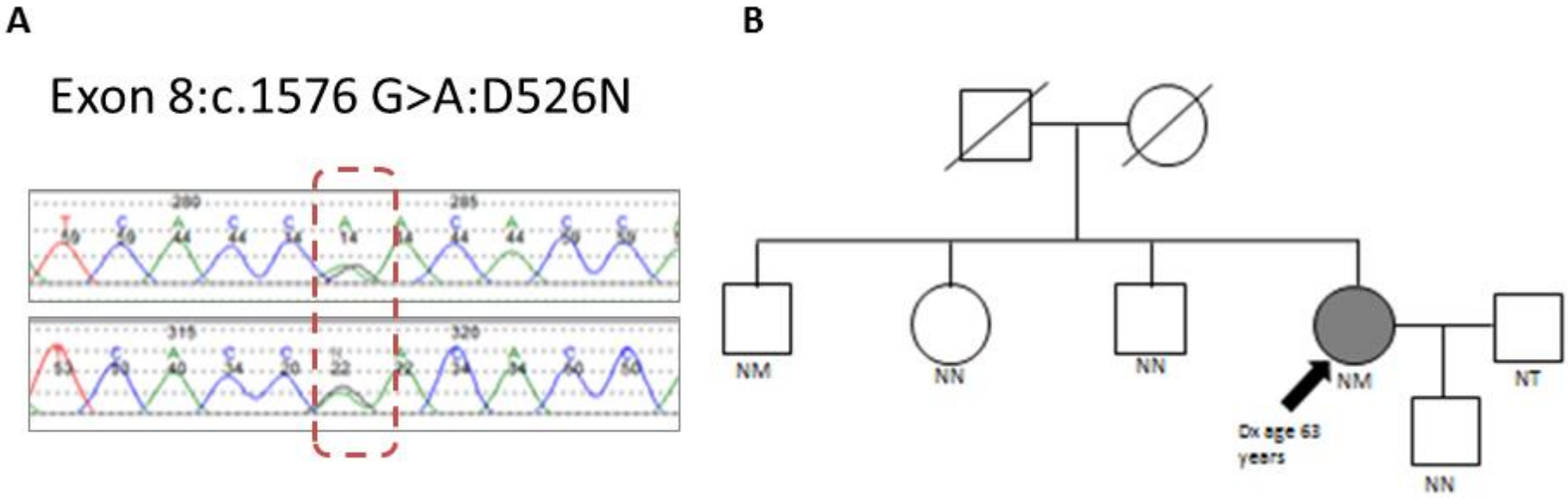


Figure 5.3.4: Electropherogram from capillary sequencing identifying the p.D526N variant and the family pedigree of the proband's. Squares represent males and circles females. Filled symbols indicate diabetic subjects. Dx, age of diagnosis; NM, heterozygous mutation carrier; NN, no mutation; NT, not tested. The black arrow designates the proband.

5.3.2 Genotyping of the p.D526N variant in the ORCADES cohort

Given the support for pathogenicity from the *in silico* evaluation of the p.D526N variant I was keen to explore a potential role for this variant in T2D pathogenesis. Prior to genotyping, three individuals were identified as carriers of the p.D526N variant. First the proband, second the sibling of the proband and then a third individual identified through exome sequencing performed in the laboratory of Professor James Wilson (Edinburgh University), which I confirmed by capillary sequencing. The Orkney Isles are considered a population isolate, leading us to hypothesize that the p.D526N variant may be enriched in this population. This finding presented an opportunity to study the p.D526N variant in the context of T2D risk and see whether this variant had an increased occurrence in patients with T2D.

Out of 2147 individuals genotyped from the ORCADE cohort a further 14 (~1%) subjects were identified as being heterozygous for the p.D526N variant. Overall, the genotyping call rate was 98%. Figure 5.3.5 displays a representative genotyping cluster plot. Relatedness between variant carriers was checked and confirmed with Professor James Wilson and was consistent with no carriers being related. Available clinical data showed that one individual had a phenotype consistent with a diabetes diagnosis (HbA1c 7.8% or 61.7 mmol/mol) and FPG levels (7.7mmol/L). All the rest (13/14) had HbA1c and FPG levels that supported a non-diabetic phenotype. It was noted that, at the time of genotyping, half of these individuals were under the age of 50 and therefore could develop diabetes later on in life. Therefore, whether the p.D526N variant may contribute to increased T2D susceptibility is undetermined.

5.3.3 Functional characterisation of HNF1A variant proteins: transactivation studies

HNF1A regulates gene expression in a number of tissues involved in glucose homeostasis including the liver and β-cell. To assess the effect of the identified missense variants on HNF1A function we used two tissue-specific promoters, one for the liver (albumin) and the second for the β-cell (HNF4A-P2). The ability of wild-type and variant HNF1A to activate the human albumin and HNF4A-P2 promoter was measured via luciferase assays.

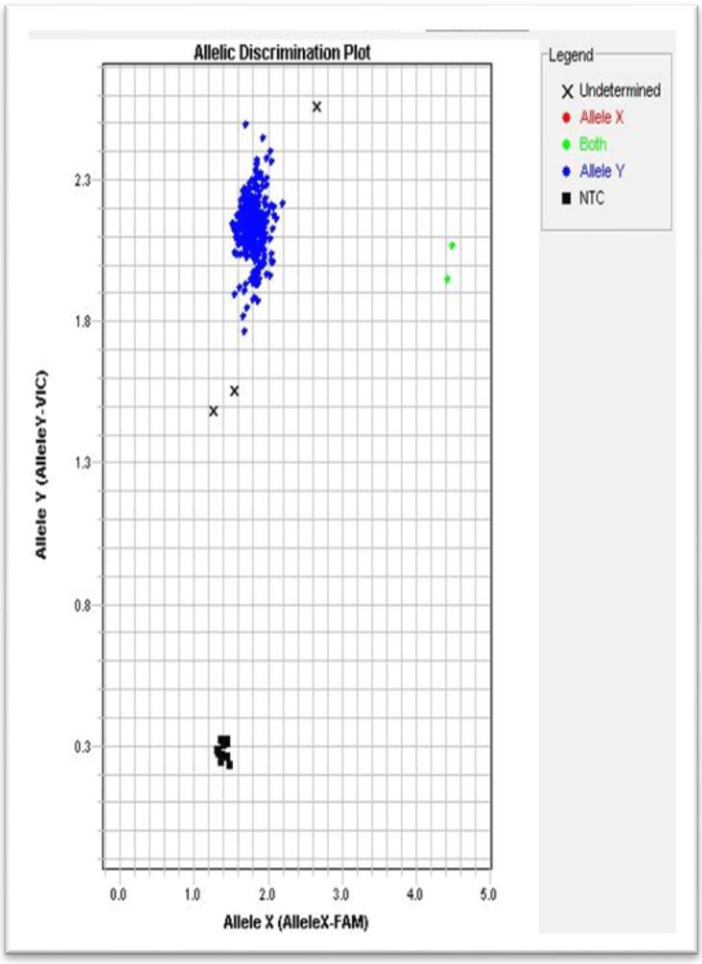


Figure 5.3:5 Representative genotyping cluster plot as produced by the SDSv2.3 software. FAM (A allele) fluorescence (*x*-axis) is plotted against VIC (G allele) fluorescence (*y*-axis). **Blue points** denote homozygotes GG and **green points** denote heterozygotes (GA). Black crosses (X) indicate samples of undetermined genotype and black squares represent non-template controls (dH_20) which should not amplify.

All experiments were performed in two different cell lines, one lacking endogenous *HNF1A* (HeLa cells) and one in the presence of endogenous *HNF1A* (INS-1 cells). INS-1 cells are derived from a rat insulinoma and are used as a model for a rodent β-cell (Asfari et al. 1992). The presence of endogenous *HNF1A* in the INS-1 cells gave the opportunity for evaluating the variants in a heterozygous state, analogous to mutations carried, and for considering the effects of HNF1A dimerization.

5.3.3.1 Optimisation of promoter constructs

Prior to beginning the transactivation assays, each promoter construct was optimised in each cell type to find ideal conditions for each experiment. As detailed above in section 5.2.8.2, only the HNF4A-P2 reporter constructs in INS-1 cells required altered conditions likely due to the presence of endogenous *HNF1A* in INS-1 cells, which gave a higher basal activity of the HNF4A-P2 promoter.

Figure 5.3.6 shows the results of the titration experiments performed with only the HNF4A-P2 promoter construct in INS-1 cells. Figure 5.3.7 shows the results of the subsequent experiments using two different concentrations of the reporter construct co-transfected with varying amounts of wild-type HNF1A. From the results of these experiments, it was decided that 0.05µg of wild-type or variant pRK5-HNF1A will be co-transfected with 0.05µg HNF4A-P2 as this provided the maximal response of luciferase activity above background in INS-1 cells.

5.3.3.2 Transcriptional activity of wild-type and variant HNF1A proteins

Figure 5.3.8 and Figure 5.3.9 display the results of the transactivation assays comparing wild-type and variant HNF1A using the albumin and HNF4A-P2 constructs respectively. All assays contained two positive controls, HNF1A-P112L and HNF1A-R271W. Both p.P112L and p.R271W lie in the DNA binding domain of HNF1A and have both been previously shown to modify HNF1A activity (Bjorkhaug et al. 2000; Bjorkhaug et al. 2003; Galan et al. 2011; Xu et al. 2002). As anticipated both HNF1A-P112L and HNF1A-R271W showed significant reduction in activity across all cell lines and promoter constructs tested ($P<0.05$). Both HNF1A-P112L and HNF1A-R271W are well characterised and have been reported in 14 and 13 different families, respectively, as casual for MODY (Colclough et al. 2013).

In HeLa cells using the albumin reporter construct only HNF1A-G47R behaved like wild-type HNF1A (Figure 5.3.8a). The rest HNF1A-R203H (43% of WT), HNF1A-A251T (21% of WT) and HNF1A-D526N (50% of WT) was significantly ($P<0.05$) decreased compared

to wild-type HNF1A. In INS-1 cells with the albumin reporter construct, a nominal difference was seen between p.G47R (P=0.013) and wild-type HNF1A. Alleles HNF1A-R203H (40% of WT) and HNF1A-D526N (53% of WT) had reduced activity relative to wild-type HNF1A and that was comparable to promoter activity in the HeLa cells (Figure 5.3.8b). No significant difference in activity was seen between wild-type and HNF1A-A251T in INS-1 cells.

When transcriptional activity was evaluated using the HNF4A-P2 promoter, only HNF1A-R203H and HNF1A-D526N showed significantly reduced activity relative to wild-type in both cell lines. Although both HNF1A-R203H and HNF1A-D526N behaved in a similar way, the reduction in promoter activity was far less (30–40% of wild-type) in INS-1 cells compared to that observed in HeLa cells (~60–80% of wild-type) (Figures 5.3.9 a & b). This may be due to a cell-specific effect due to the presence of endogenous HNF1A. The endogenous HNF1A is able to form dimers with the mutated HNF1A, thereby possibly compensating for the effect of the variant in a loss-of-function manner. In contrast, the lack of coactivators in HeLa cells maybe causing the HNF4A-P2 promoter to be less active.

5.3.4 Protein expression and localisation of HNF1A variant proteins

Using HeLa cell lysates collected from the transactivation studies (albumin promoter); western blots were performed from two independent experiments to confirm whether the reduction in transcriptional activity was due to reduced protein expression. Figure 5.3.10 shows the result of a representative blot and the results of the combined densitometry for the two independent experiments. Visualisation of the membrane and densitometry demonstrated a consistent reduction in protein levels when normalised to wild-type protein for HNF1A-P112L, HNF1A-R203H, HNF1A-R271W and HNF1A-D526N. The results of an unpaired t-test comparing wild-type and variant HNF1A showed no statistical difference (P>0.05). The reduced transactivation activities observed for HNF1A-P112L, HNF1A-R203H, HNF1A-R271W and HNF1A-D526N are likely to be influenced by the reduction in protein levels for these variants. Three of the alleles, p.P112L, p.R203H and p.R271W, are found in the DNA binding domain of HNF1A; therefore, the reduced transactivation potential may also be caused by weakened DNA binding affinity. Reduced DNA binding affinity for HNF1A-P112L and HNF1A-R271W has been previously reported, supporting the hypothesis (Bjorkhaug et al. 2000; Bjorkhaug et al. 2003).

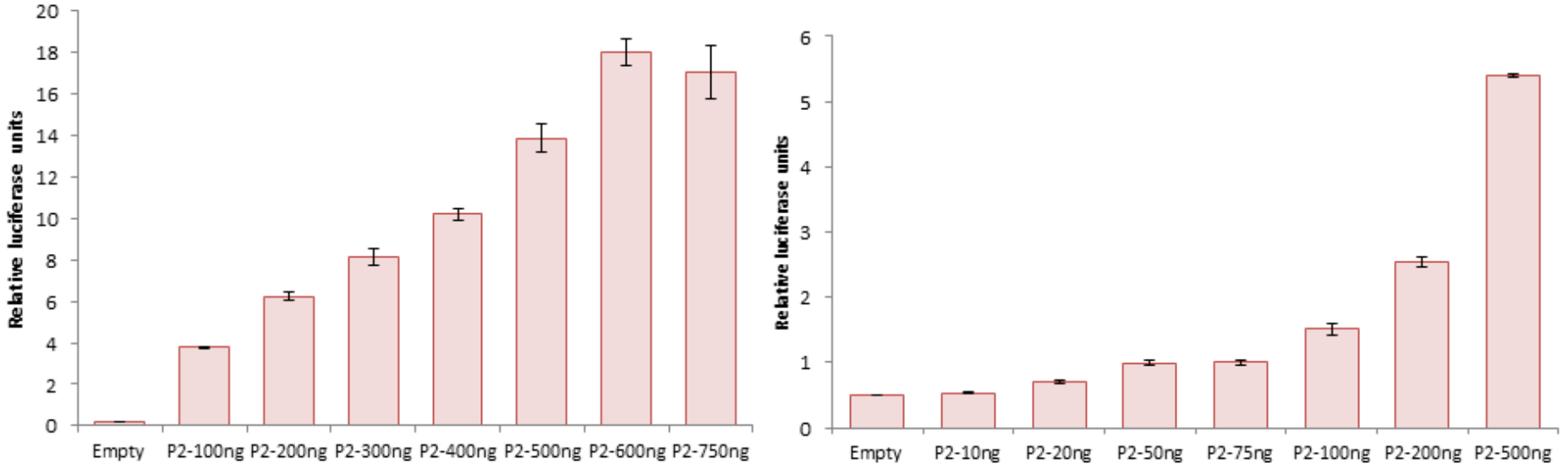


Figure 5.3.6: Transcriptional activity of the HNF4A-P2 promoter construct at various DNA concentrations in INS-1 cells. Two separate experiments were performed to find the optimal DNA concentrations required of the promoter, which would have minimal activity with the endogenous factors in the INS-1 cell line. Each bar represents the average of a triplicate and the error bars represent the standard error of the mean (SEM).

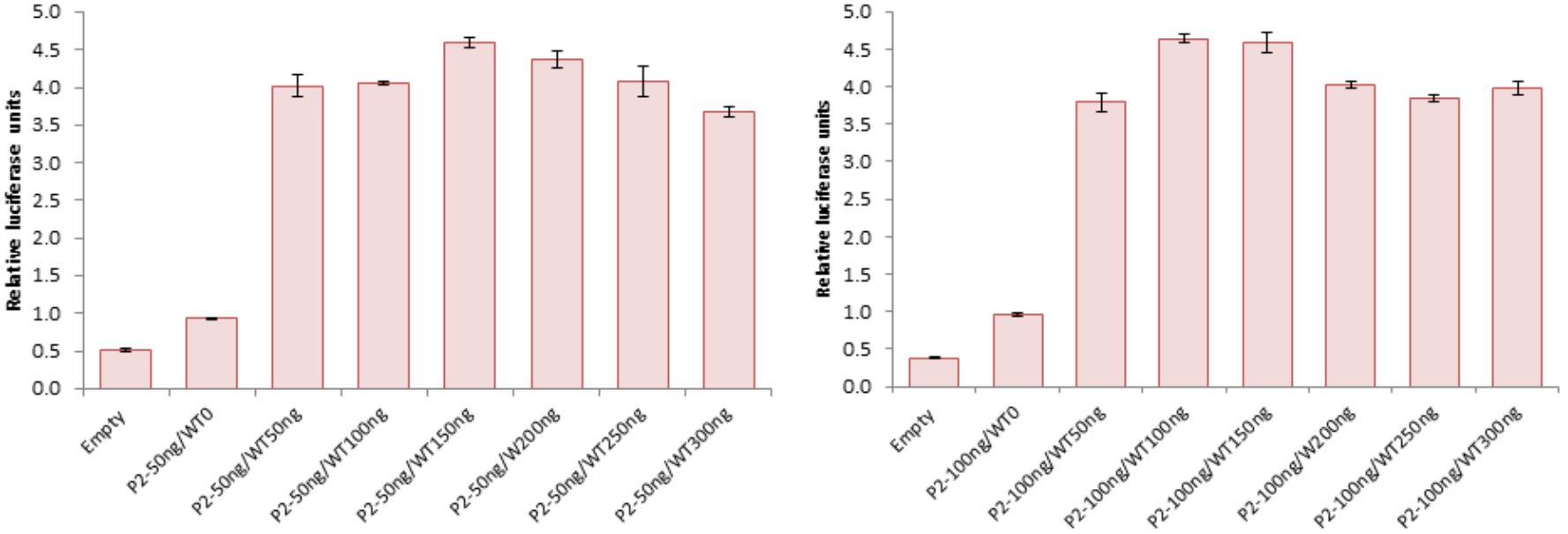


Figure 5.3.7: Transcriptional activity of the HNF4A-P2 promoter with varying concentrations of wild-type HNF1A in INS-1 cells. Graph 1 denotes the results of when 50ng of HNF4A-P2 promoter is co-transfected with varying concentrations of wild-type HNF1A (0–300ng). Graph 2 represents the transcriptional activity of the HNF4A-P2 promoter when using 100ng with varying concentrations of wild-type HNF1A. Each bar represents the average of a triplicate. Error bars represent the standard error of the mean (SEM).

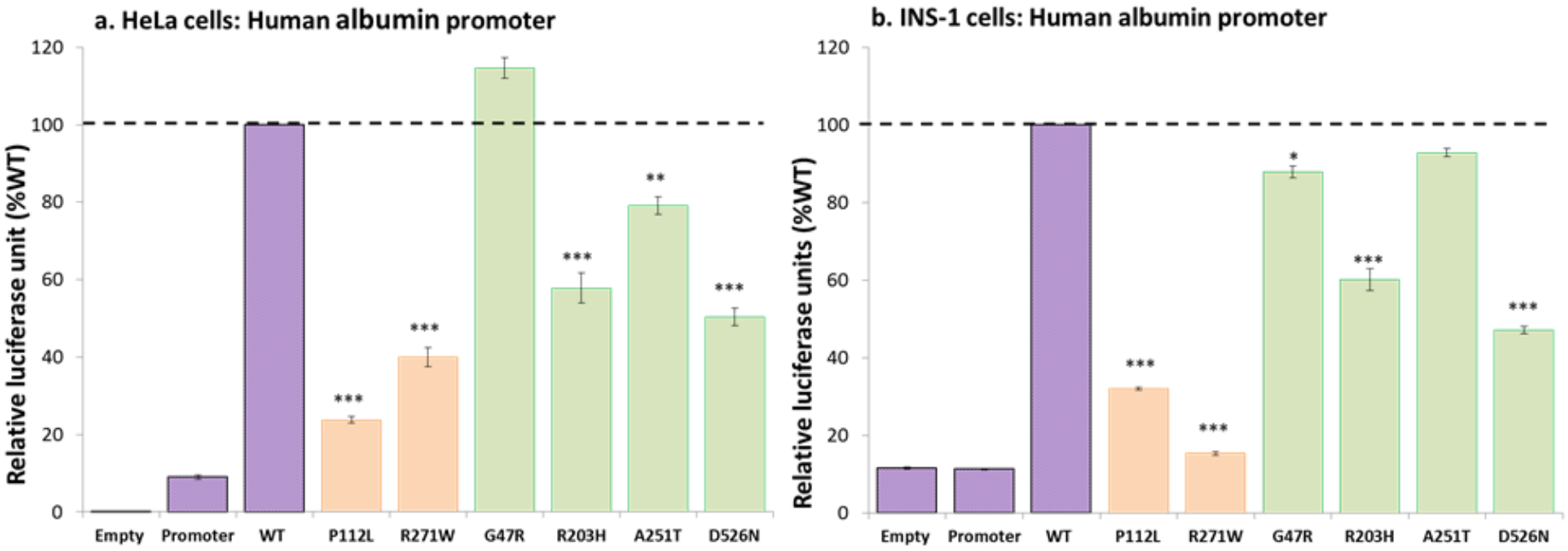


Figure 5.3.8: Transcriptional activity of wild-type (WT) and mutant HNF1A in HeLa and INS1 cells using the albumin reporter construct (*n*=4). P112L and R271W (orange bars) were used as positive controls as they have been previously shown to alter HNF1A activity. Cells were harvested 24 hours after transfection and transcriptional activity was measured using a dual luciferase (firefly) reporter system. Luciferase values were normalised to HNF1A-WT and are presented as percentage WT (100%). All experiments were completed in triplicate and repeated four times with different DNA preparations. An ANOVA was carried out to compare the means of the different groups. *≤0.05–0.01, **<0.01–0.001 and ***<0.001. Error bars represent the standard error of the mean (SEM).

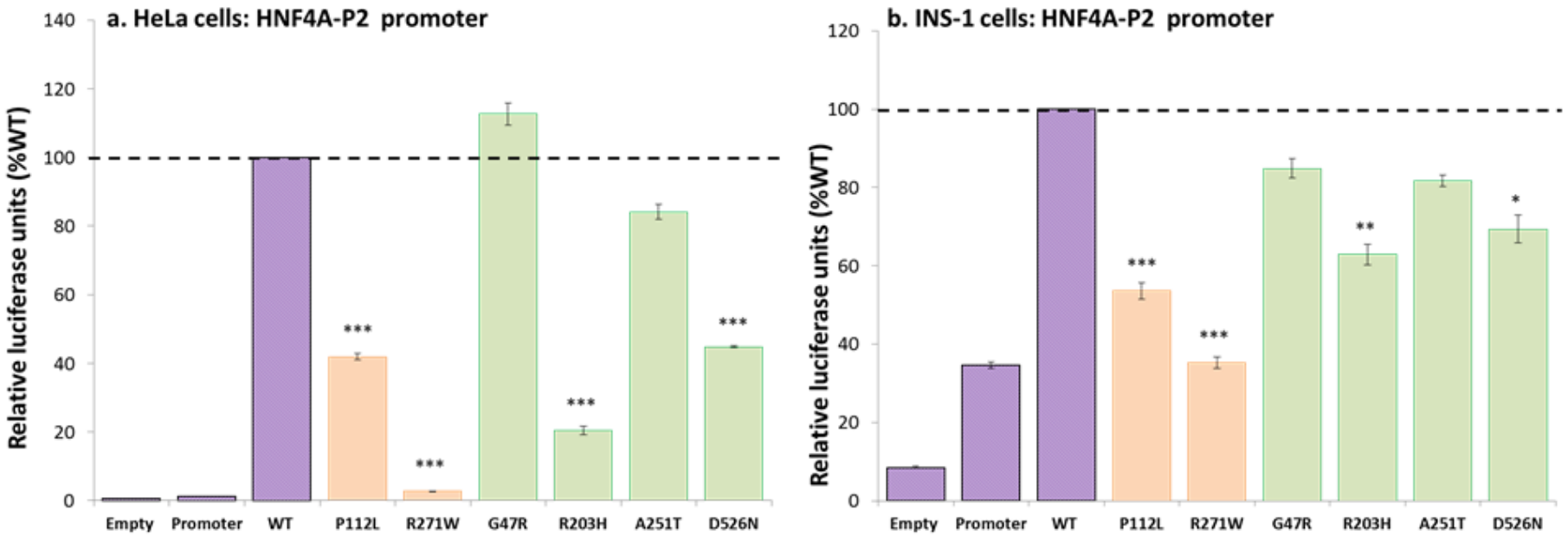


Figure 5.3.9: Transcriptional activity of wild-type (WT) and mutant HNF1A in HeLa and INS1 cells using the HNF4A-P2 promoter (*n*=4). P112L and R271W (orange bars) were used as positive controls as they have been previously shown to alter HNF1A activity. Cells were harvested 24 hours after transfection and transcriptional activity was measured using a dual luciferase (firefly) reporter system. Luciferase values were normalised to HNF1A-WT and are presented as percentage WT (100%). All experiments were completed in triplicate and repeated four times with different DNA preparations. An ANOVA was carried out to compare the means of the different groups. *≤0.05–0.01, **<0.01–0.001 and ***<0.001. Error bars represent the standard error of the mean (SEM).

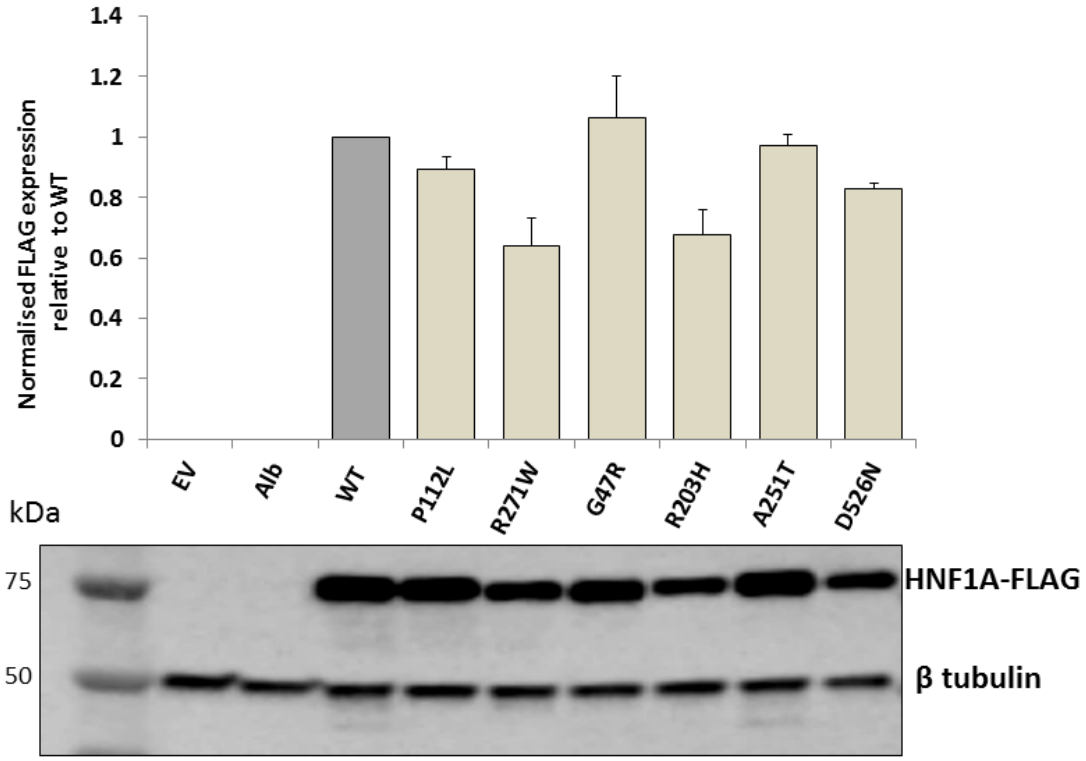


Figure 5.3.10: Expression levels of wild-type and variant HNF1A. Expression levels in HeLa cells were determined by western blot and densitometry using transactivation cell lysates. The in-frame N-terminal FLAG tag was used to detect protein expression of wild-type and mutant HNF1A. Tubulin was used as a control as protein expression is expected to be equal across all samples. Lysates used were those from the luciferase assays. Data are presented as a mean ± standard error of the mean for two independent experiments. An unpaired t-test was performed to compare the means between wild-type and variant HNF1A, no significant differences were observed (P>0.05). EV, empty vector; Alb, albumin; WT, wild-type. The western blot was performed by Dr Jana Rundle.

Due to the length of time taken to optimise the cytosolic and nuclear fractionation technique, only representative western blot analyses were possible. Figure 5.3.11 shows an example of western blot using fractionated INS-1 cell lysates. Two control antibodies anti-GAPDH and Topoisomerase IIα were used to assess cross-contamination between the cytosolic and nuclear fractions, respectively. Visualisation of the membrane displayed that the fractionation protocol was producing reasonably pure cell lysates. It was noted that non-specific bands were seen when samples were blotted with the Topoisomerase IIα antibody, which were present in the cytoplasmic fraction. Western blot analysis in HeLa cells would demonstrate whether this was a cell-specific effect. Further western blots would allow for quantification between the cytoplasmic and nuclear fractions using densitometry. This could then be used in addition to the immunostaining to determine whether the HNF1A variants alter nuclear localisation.

5.3.5 Assessing nuclear localisation of **HNF1A** *variant proteins*

Immunofluorescence was used to look at the subcellular localisation of wild-type and mutant HNF1A to explore whether the variants would affect nuclear localisation in both cell lines. An in-frame N-terminal HNF1β construct was used as an immunostaining positive control as this has previously demonstrated to localise to the nucleus (Barbacci et al. 2004). Figures 5.3.12(A–H) and 5.3.13(A–H) illustrate examples of representative cells viewed for each variant in HeLa and INS-1 cells, respectively. Due to the length of time taken to optimise experimental conditions it was only possible to view each variant once under the confocal microscope. As expected, cells transfected with HNF1β all localised to the nucleus (Figures 5.3.12A and 5.3.13A). Wild-type HNF1A also displayed nuclear localisation in almost all cells viewed for both cell lines. Overall, all variants investigated showed a pattern consistent with normal nuclear protein staining. In HeLa cells HNF1A-R271W (Figure 5.3.12D) was seen to have unusual punctate staining. This would need to be confirmed by further staining to assess whether this was genuine or due to experimental artefact. Although it was clear that the majority of the protein staining was confined to the nucleus for all variants, some staining was seen in the cytoplasm. Previous reports of HNF1A-R271W targeting both the cytoplasm and nucleus are consistent with what has been observed in this study (Bjorkhaug et al. 2003). Consequently, it is possible that the staining patterns observed for these HNF1A variants are genuine and may influence the ability of HNF1A to target the nucleus completely. Further staining would be required to either support or refute this observation.

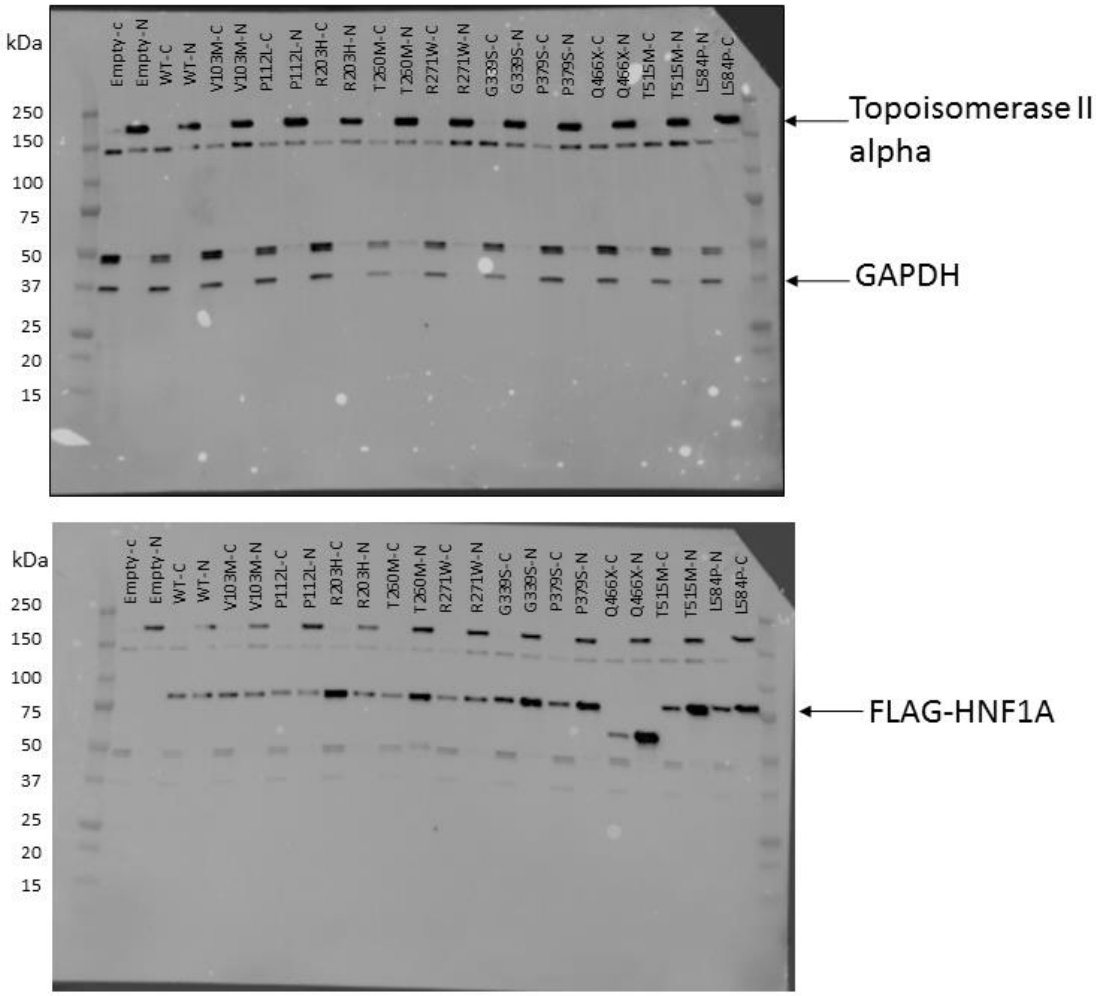


Figure 5.3.11: Western blot using fractionated cytoplasmic and nuclear fractionated nuclear cell lysates to illustrate whether the variants altered HNF1A cellular localisation. The in-frame N-terminal FLAG tag was used to detect protein expression of wild-type and mutant HNF1A. Antibodies, anti-GAPDH and Topoisomerase IIα were used as a control, as protein expression is expected to be equal across all samples, and to detect cross-contamination between fractionated samples. WT, wild-type; C, cytoplasmic; N, nuclear.

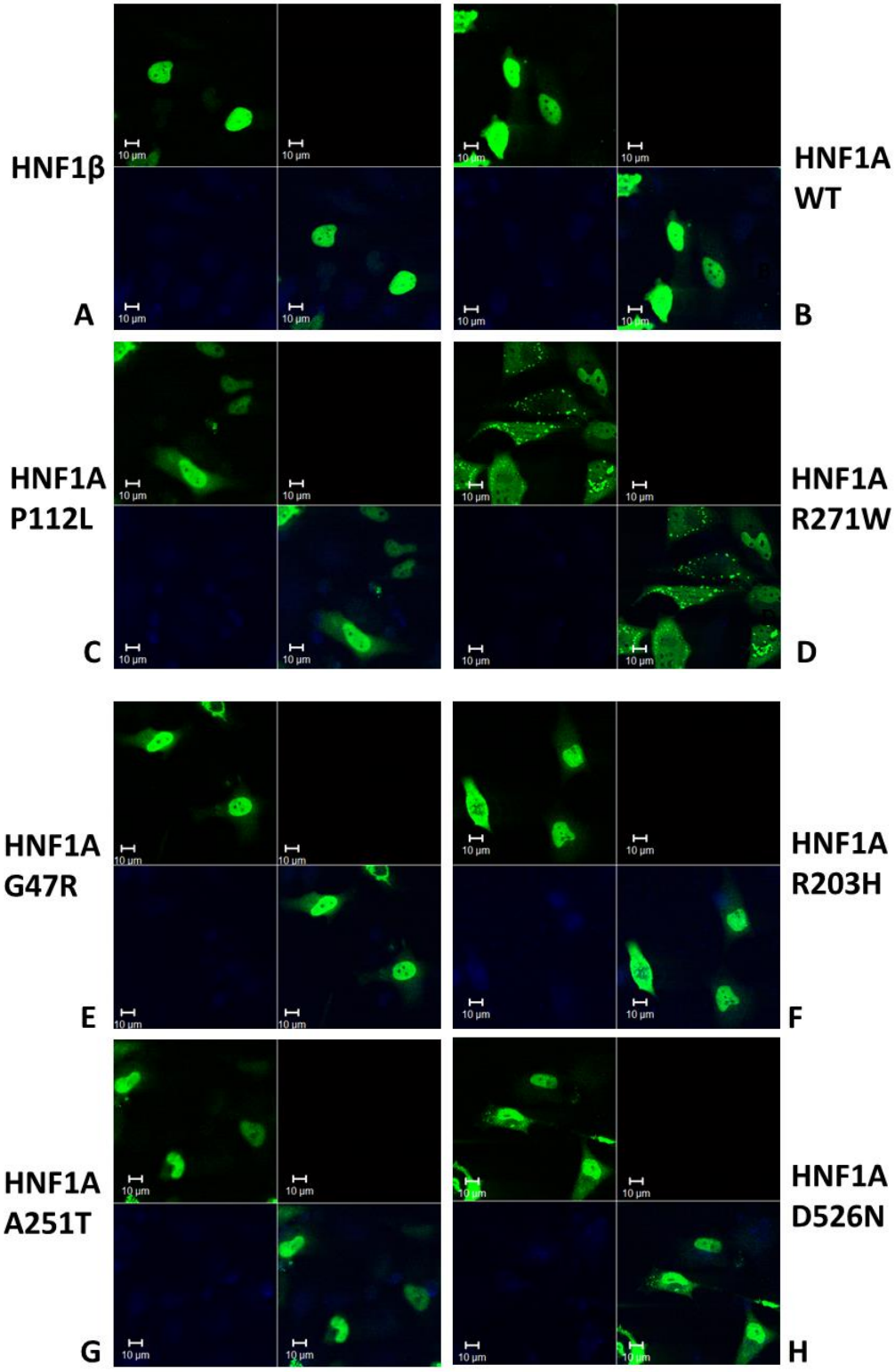


Figure 5.3.12: Subcellular localisation of wild-type and variant HNF1A in HeLa cells. Representative images were captured in HeLa cells that were transiently transfected for 24 hours prior to immunostaining. Cells were stained with Alexa Fluor-488 (Green) and far-red nuclear stain (NucBlue Live 647 Ready probes) and viewed using the confocal microscope using Bio-Rad filters. Images were superimposed using Zeiss LSM Image browser (http://www.zeiss.co.uk/microscopy/en_de/downloads/lsm-5-series.html).

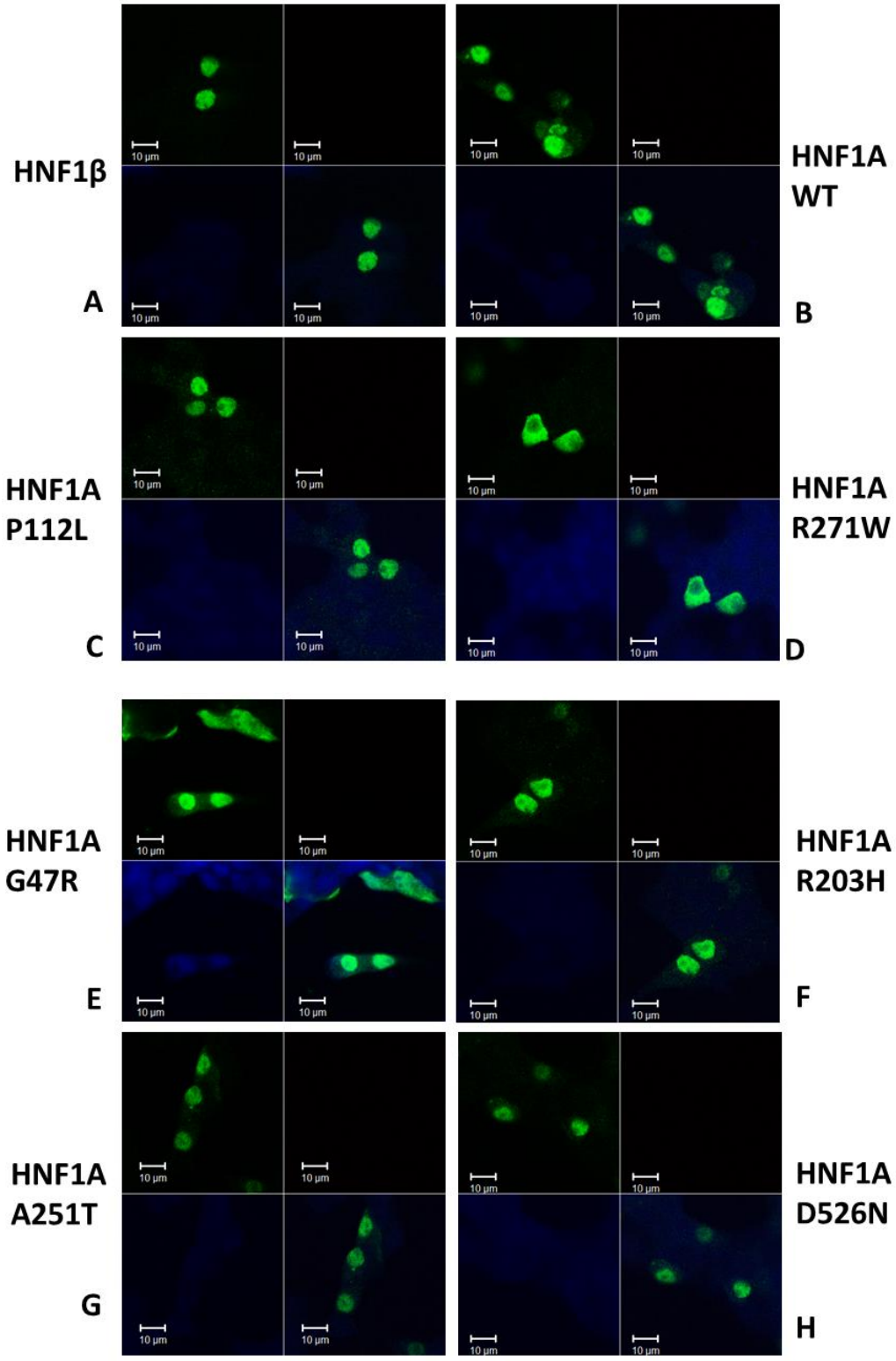


Figure 5.3.13: Subcellular localisation of wild-type and variant HNF1A in INS-1 cells. Representative images were captured in INS-1 cells that were transiently transfected for 24 hours prior to immunostaining. Cells were stained with Alexa Fluor-488 (Green) and far-red nuclear stain (NucBlue Live 647 Ready probes) and viewed using the confocal microscope using Bio-Rad filters. Images were superimposed using Zeiss LSM Image browser (http://www.zeiss.co.uk/microscopy/en_de/downloads/lsm-5-series.html).

5.4 Discussion

Efforts to improve diagnostic testing for HNF1A-MODY, by exploring the potential of biomarkers as a clinical diagnostic tool, have been the focus of recent research (Bonner et al. 2013; McDonald et al. 2011; Owen et al. 2010; Thanabalasingham et al. 2011; Thanabalasingham et al. 2013). Differentiating between HNF1A-MODY and other common subtypes of diabetes will improve the prioritisation of patients for *HNF1A* genetic screening (Thanabalasingham et al. 2012). An accurate diagnosis of HNF1A-MODY has important implications for the patients' treatment and disease prognosis as cases of HNF1A-MODY can be successfully treated with low-dose sulphonylurea (Pearson et al. 2003).

To assess the potential of the DG9-glycan index as a marker to discriminate between patients with HNF1A-MODY from those with T1D and T2D, I screened all exons of *HNF1A* in patients with a DG9-glycan index <0.16. Of the 97 subjects screened, only four individuals (4%) were identified as carriers of heterozygous missense mutations in *HNF1A*. Preliminary validation of the DG9-glycan index indicated that a diagnostic threshold of ≤0.16 had 88% sensitivity and 81% specificity to discriminate between HNF1A-MODY and early onset T2D (C-statistic 0.91), and 88% sensitivity and 88% specificity for T1D (C-statistic 0.94) (Thanabalasingham et al. 2013). The sensitivity of the assay represents the percentage of individuals with a disease correctly identified as having the disease (true positive rate). The specificity of an assay denotes the percentage of individuals without the disease correctly recognised by the assay as not having the condition (false positive rate) (Saah and Hoover 1997). Consequently, given the high sensitivity and specificity of the assay, it was expected that a higher proportion of patients with mutations in *HNF1A* would be found. There are several possible reasons for why this proportion was lower than expected. First, the discriminatory power calculated based on the C-statistic used in the validation study may have been overestimated (Thanabalasingham et al. 2013). Second, the number of novel HNF1A-MODY cases may have been filtered out because of existing clinical investigations and previous *HNF1A* sequencing (Thanabalasingham et al. 2013). Finally, many other factors influence glycan profiles including genetic factors. Common variants in the fucosyltransferase genes *FUT6* and *FUT8* were also found to influence N-glycan plasma levels (Lauc et al. 2010). Therefore, other genetic influences may be driving the changes in N-plasma glycan levels. As the cohort of patients tested in this study was small, further validation studies are required in larger cohorts to assess the value of the DG9-

glycan index as a potential discriminatory biomarker for HNF1A-MODY (Thanabalasingham et al. 2013).

The discriminatory effect of the DG9-index between HNF1A-MODY and T2D is comparable with that of hsCRP, but hsCRP is less able than glycan profiles to distinguish between HNF1A-MODY and T1D (Owen et al. 2010; Thanabalasingham et al. 2011; Thanabalasingham et al. 2013). In contrast to hsCRP, glycan profiles have the added benefit of being less affected by inflammation (Thanabalasingham et al. 2013). The lack of high-throughput techniques for glycan analysis and the associated costs makes this a less amenable diagnostic tool than hsCRP, which is widely available and inexpensive to perform (Owen et al. 2010; Thanabalasingham et al. 2011; Thanabalasingham et al. 2013). In the future, it is likely that a combination of biomarker profiles will be used as they all vary in their ability to discriminate between different subtypes of diabetes. For example, it is anticipated that combined biomarkers and clinical information, which includes age of onset and family history, will improve our ability to differentiate between diabetes subtypes.

Next, I functionally evaluated the four HNF1A variants identified through capillary sequencing. All variants were identified in patients with a glycan profile suggestive of HNF1A dysfunction. Of the four variants assessed, only two alleles, HNF1A-R203H and HNF1A-D526N, were seen to affect HNF1A function. Of these, only p.R203H had sufficient evidence to support that this was a genuine pathogenic MODY mutation. This heterozygous missense mutation has been previously reported in 19 other families as co-segregating as a MODY mutation (Colclough et al. 2013). The amino acid residue is located in exon 3 within the DNA binding domain of HNF1A. This residue is highly conserved amongst different species and therefore it is plausible that the change in amino acid at this position causes DNA binding to be weakened. An approach that could be used to determine whether DNA binding affinity was affected is an electrophoretic mobility shift assay (EMSA). An EMSA is a technique that can be used to study protein–DNA interactions. For example, HNF1A variant protein can be combined with DNA containing the sequence of the HNF1A binding site for the human albumin promoter. The subsequent protein–DNA mixtures are separated usually on a polyacrylamide gel and the ratio of bound to unbound DNA and protein can be calculated to determine the affinity of the protein and nucleic acid.

The later age of onset of diabetes and variable penetrance associated with variants in exon 8–10 of *HNF1A* makes the interpretation of the previously described MODY variant p.D526N, located within exon 8, challenging (Bellanne-Chantelot et al. 2008; Harries et al. 2006). The p.D526N variant is located in exon 8, present in only isoform A of HNF1A and expressed at low levels in beta cells (Harries et al. 2006).

It has been suggested that the type of variant (missense variants), and those within the later exons (8–10) that are not represented in all three isoforms, may manifest in a milder diabetic phenotype or later on in life due to compensation from the other isoforms (Bellanne-Chantelot et al. 2008; Harries et al. 2006). Isoform A is the predominant transcript in the liver, whereas in the pancreas it is HNF1A (B) (Harries et al. 2006). Therefore, reduced transcriptional activity due to p.D526N is likely to influence liver-specific genes that maybe involved in hepatic glucose metabolism as opposed to beta-cell relevant genes (Bach and Yaniv 1993).

In the case of p.D526N, the DG9-glycan biomarker profile was able to identify a patient, with a variant that *in vitro* assays clearly showed had an effect on HNF1A function. In my study, transient transfections exhibited reduced transcriptional activity of p.D526N with both promoter constructs evaluated. This is likely due to reduced protein levels and slightly altered nuclear localisation of the variant protein. Although DNA binding studies were not undertaken, it is unlikely that DNA binding affinity would be compromised, as p.D526N is located in the transactivation domain of *HNF1A.* However, further studies would be required to confirm this. Recently, another rare *HNF1A* variant, p.E508K, also in exon 8 and the transactivation domain of HNF1A, was associated with a fivefold increase of T2D prevalence but incomplete penetrance in the Latino population (Estrada et al. 2014). Of the participants studied, the p.E508K variant was seen in 2.1% of T2D cases (Estrada et al. 2014). Functional studies in HeLa and Min6 cells using three promoter constructs demonstrated reduced transcriptional activity for p.E508K variant due to reduced protein expression and irregular subcellular localisation; however, DNA binding was comparable to wild-type HNF1A (Estrada et al. 2014).

Several other population-specific variants have been related to *HNF1A.* The private p.G319S variant resulting in abnormal mRNA processing is associated with early-onset T2D in the Canadian Oji-Cree population (Harries et al. 2008; Hegele et al. 1999b). Additionally, a rare mutation 290fsdelC is associated with MODY and T2D in the Icelandic population (Kristinsson et al. 2001; Steinthorsdottir et al. 2014). Genotyping of

the p.D526N variant to identify whether the variant was enriched in the ORCADES population detected only one carrier with a diabetic phenotype. Half of the individuals identified as carriers of the p.D526N variant in the ORCADES population were under 50 years old, so the development of T2D later in life cannot be ruled out. Functional data for the p.D526N show that the variant impacts HNF1A function comparable to that observed for the p.E508K variant. Therefore, it is plausible that p.D526N is a rare variant with variable penetrance and is associated with T2D risk in certain individuals.

The data collected for the final two variants, p.G47R, and p.A251T, displayed limited evidence to suggest that either of these variants cause MODY. Family data for the p.G47R variant were inconsistent with the variant co-segregating as a fully penetrant MODY mutation. Only the transactivation assay in INS-1 cells with the albumin promoter showed a nominal difference in promoter activity with p.G47R relative to wild-type HNF1A (P<0.013). For all other studies, p.G47R was seen to act like wild-type protein. A study by Bjorkhaug et al. (2003) found, in a single individual, a variant at the same amino acid residue but with a different amino acid substitution (p.G47E) (Bjorkhaug et al. 2003). The results of their functional investigations showed close to normal levels of activity of the mutant p.G47E protein (Bjorkhaug et al. 2003). The location of the G47 amino acid is just outside the DNA dimerization domain (1–32) and before the DNA binding domain of HNF1A (Bellanne-Chantelot et al. 2008; Bjorkhaug et al. 2003; Galan et al. 2011; Vaxillaire et al. 1999). It is therefore possible that amino acid substitutions at this position do not affect protein function or structure, as important domains are not interrupted. Like the suggestion for the p.G47E, the p.G47R variant could be a rare variant that is not fully penetrant and increases risk for developing diabetes in certain individuals (Bjorkhaug et al. 2003).

The p.A251T is a novel variant located in the homeodomain region of the DNA binding domain. Family follow-up was not possible so it is not known whether the family members with diabetes were also carriers of the same amino acid substitution. Again *in vitro* evaluation displayed p.A251T to function similar to that of wild-type HNF1A. Only one transactivation assay in HeLa cells using the albumin reporter construct showed a significant reduction in activity of p.A251T compared to wild-type protein. The p.A251T sits in the DNA binding domain; therefore, impaired DNA binding for the albumin promoter may have caused the reduced transcriptional activity observed. Additional DNA binding studies would be required to confirm this hypothesis. Furthermore, the inconsistencies between the *in silico* programmes and lack of conservation across

species all support the idea that it is unlikely that this variant causes MODY. Comparable to p.G47R, this may be a rare variant, which is not fully penetrant or contributes to diabetes risk in certain individuals.

This study has shown that glycan profiles are able to identify patients with HNF1A variants. However, *in vitro* assays demonstrated that not all variants affect HNF1A function. The results presented in this chapter show that, despite having a suite of information that includes family studies and functional characterisation of variants, it is still difficult to ascribe pathogenicity to some variants in *HNF1A*, limiting the clinical interpretations that can be made for at risk individuals.

Future work is now being directed by advances in high-throughput sequencing capture methods and next-generation sequencing technologies, giving the opportunity to comprehensively interrogate genes already implicated in Mendelian disorders (Ku et al. 2011). It is anticipated that this will identify new genes linked to complex diseases, uncover novel biological pathways, and elucidate novel pharmacotherapy targets (Ku et al. 2011). Exome-sequencing is now a cost-effective method that can systematically identify both common and rare alleles in protein-coding sequences (Kiezun et al. 2012). Understanding which of these variants impact human health and disease risk is the next challenge (Cutting 2014). Current work taking place in the laboratories of supervisors Professor Gloyn and Professor McCarthy (University of Oxford, OCDEM) in collaboration with Professor Njølstad's group in Norway (University Hospital, Bergen) as part of the T2DGENES/GoT2D consortia is the functional characterisation of almost 80 *HNF1A* missense variants identified through exome-sequencing projects. Cases (6500) and controls (6500) were ascertained from five multiethnic ancestries (African American, East Asian, European, Hispanic and South Asian). Many of the *in vitro* assays described in the chapter are now routinely used for the characterisation of *HNF1A* variants as part of this large collaborative effort. Complete functional evaluation of the *HNF1A* missense variants will provide a detailed catalogue of how these variants impact gene function *in vitro*, helping to interpret medical and re-sequencing data.

Recent studies focused on the Peroxisome proliferator-activated receptor gamma (*PPARG*) and the low-density lipoprotein receptor (*LDLR*) genes have provide good examples of how high-throughput experimental assays *in vitro* can provide comprehensive characterisation of missense rare-coding variants identified in the general population (Majithia et al. 2014; Thormaehlen et al. 2015). PPARG is a nuclear

hormone receptor essential for adipogenesis (Tzameli et al. 2004). Majithia et al. (2014) developed a novel *in vitro* assay, evaluating the *PPARG* variants identified in a multiethnic population composed of T2D cases and controls(Majithia et al. 2014). The assay measured the ability of the *PPARG* variants to stimulate differentiation of human preadipocytes. Using this assay Majithia et al. (2014) identified nine novel variants in *PPARG* that had reduced ability to stimulate adipocyte differentiation (Majithia et al. 2014). All nine variants were associated with an increased risk for T2D in carriers (Majithia et al. 2014). Leading on from this example, the laboratories of supervisors Professor Gloyn and Professor McCarthy (University of Oxford, OCDEM), in collaboration with the Broad Institute, are developing a functional high-throughput assay for *HNF1A* for all possible missense variants. The assay will determine the functional consequence of missense *HNF1A* variants found in genomic sequencing. The aim is to create a clinical prediction tool that correlates mutation severity with clinical features, including biomarker profiles and sulphonylurea sensitivity, which will help guide pharmacotherapy for at-risk individuals.

Finally, induced pluripotent stem cells (iPSCs) represent a novel approach that can used to model diabetes (Takahashi and Yamanaka 2006). Two recent studies have reported iPSCs that have been derived from HNF1A-MODY patients (Stepniewski et al. 2015; Teo et al. 2013). The accessibility of MODY-iPSCs *in vitro* provides the opportunity to generate cells of pancreatic lineage and study the developmental and molecular mechanisms that cause impaired insulin production in HNF1A-MODY patients whilst at the same time elucidating the role of *HNF1A* (Stepniewski et al. 2015).

Chapter 6

Functional assessment of missense hepatocyte nuclear factor-1-alpha (*HNF1A*) gene variants identified in subjects of South Asian ancestry

6.1 Introduction

6.1.1 Maturity Onset Diabetes of the Young (MODY) in South Asians

Our understanding of MODY in the South Asian population is limited, and this is reflected in the clinic where MODY is frequently overlooked in South Asian populations with a diagnosis of type 2 diabetes (T2D) usually favoured. This occurs despite several clinical features of T2D onset in South Asians being similar to the clinical characteristics of MODY, including the young age of onset of diabetes and lower BMI thresholds. The discrepancies in testing for MODY between Europeans and South Asians have been attributed to several issues. First, the high prevalence of young onset T2D in Asians: if young South Asians are found to have non-insulin-dependent diabetes and antibody-negative diabetes they are assumed to have T2D and not referred for genetic testing. Second, it is possible that although South Asian families are being identified for further genetic investigations, such referrals are in decline, possibly due to a lack of willingness of the patient to take up further investigations. Studies focusing on prenatal testing have highlighted that there was a lower uptake of genetic testing by South Asians, due to fewer South Asian families being offered testing and also fewer women requesting to be tested (Jayawardena et al. 2012; Porter et al. 2006). Finally, most studies looking at MODY have been in Europeans, with very few studies in other ethnic groups.

One of the first studies investigating MODY in South Asians was reported by Porter et al. in 2006, which identified mutations in two known MODY genes (*HNF1A* and *GCK*) in three South Asian probands. Porter et al. (2006) suggested that the occurrence of MODY was the same in UK South Asians when compared to European counterparts even though a lower referral rate for MODY testing in South Asians was observed. Clinical presentation of MODY and response to appropriate treatment was also comparable between both ancestry groups (Porter et al. 2006). Subsequently, several investigations examining MODY in India have been reported (Anuradha et al. 2011; Radha et al. 2009; Sahu et al. 2007). Together, these studies sequenced *HNF1A*, *HNF4A* and *GCK* in a subset of patients clinically classified as having MODY. Genetic testing revealed variants in both *HNF1A* and *HNF4A*, giving an overall molecular diagnosis of 3–9% across all studies in patients strongly suspected clinically of having a MODY diagnosis (Anuradha et al. 2011; Radha et al. 2009; Sahu et al. 2007). The study by Radha et al. sequenced *HNF1A* in 96 patients clinically suspected of having MODY. Of these, 9 probands (out of 96) were recorded as carrying 9 different HNF1A variants. For one of these, p.Arg263His, additional family studies were possible and revealed 7 additional carriers of

p.Arg263His, all with diabetes on the maternal lineage, suggestive of this variant co-segregating with HNF1A-MODY (Radha et al. 2009). In addition, Radha et al. functionally characterised two novel *HNF1A* promoter variants. Although statistically non-significant both promoter variants were reported to have reduced transcriptional activity compared to wild-type HNF1A and may increase risk for diabetes susceptibility (Radha et al. 2009).

Given the limited number of studies examining MODY in South Asians, our understanding of MODY in the South Asian population is incomplete. Consequently, physicians may not consider investigating MODY in non-European populations due to the lack of awareness of the clinical features of MODY and the considerable overlap with T2D when considering clinical criteria only.

In South Asians, the onset of T2D occurs almost a decade earlier and at a lower BMI and results in a potential overlap with MODY. As a result, it is possible that cases of MODY are being missed in South Asians. One approach currently being investigated that may help find the 'missing MODY cases' is the use of biomarkers that are moderately non-invasive which might be used routinely in the clinic to help select and prioritise patients for genetic screening. To date, almost all evaluation of biomarkers has been performed in European cohorts and it is not known whether these studies can be translated to other populations. This chapter focuses on one particular biomarker, the highly sensitive C-reactive protein (hsCRP), and evaluates whether hsCRP can be used as a discriminatory tool for HNF1A-MODY in South Asians.

6.1.2 Using highly sensitive C-reactive protein (hsCRP) as a biomarker for HNF1A-MODY

Sequence alterations in *HNF1A* are the most common cause of MODY in the UK (Shields et al. 2010). Correctly classifying patients with HNF1A-MODY has important clinical value, as patients can effectively maintain glycaemic control when treated with low-dose sulphonylurea, thus improving quality of life and disease prognosis for the patient and their families. Over the past few years, independent genome-wide association studies (GWAS) focusing on C-reactive protein (CRP) have shown that common variants mapping close to *HNF1A* (and considered likely to be acting through altered regulation of *HNF1A*) may alter CRP levels (Okada et al. 2011; Reiner et al. 2008; Ridker et al. 2008). This prompted investigations to determine whether CRP levels could be used as a biomarker to screen for HNF1A-MODY. Studies in European patients have now demonstrated that patients with HNF1A-MODY have lower levels of hsCRP compared with other forms of diabetes (Owen et al. 2010). The studies demonstrated that by

using a combined approach, which included low hsCRP (≤0.2mg/L) and current MODY diagnostic criteria (age of diabetes onset ≤25 years and first-degree family history of diabetes), it is possible to discriminate between HNF1A-MODY and T2D, with a sensitivity and specificity approaching ~80% (McDonald et al. 2011; Owen et al. 2010; Thanabalasingham et al. 2011). It is not known whether these same criteria of low hsCRP levels can achieve similar levels of sensitivity and specificity to differentiate between HNF1A-MODY and T2D in other ethnic groups so this requires further evaluation.

Measuring hsCRP levels in patients is advantageous as the test is routinely available, relatively non-invasive, and inexpensive. However, hsCRP is commonly used as a marker for inflammation and is consequently raised with other diseases, including atherosclerosis, cancer, and coronary heart disease, so low levels could be masked in the presence of concurrent diseases (Black et al. 2004; Hansson 2005; Toniatti et al. 1990b).

CRP is an inducible liver-specific gene. It primarily functions as an acute phase protein and levels increase in response to acute and chronic inflammatory conditions. Both IL-6 and IL-1β alongside other liver-specific transcription factors mediate gene expression of CRP (Black et al. 2004; Nishikawa et al. 2008; Toniatti et al. 1990a). Studies examining the interaction between HNF1A and CRP have shown that the *CRP* promoter has two binding sites for HNF1A. The two binding sites on the *CRP* promoter have several alterations compared to the consensus sequence but are still able to bind to *HNF1A* albeit with lower binding affinity. Mutations in either of these two sites have been shown to prevent HNF1A binding, leading to a loss of CRP expression (Toniatti et al. 1990a). HNF1A is responsible for conferring both the liver specificity of CRP and for forming transcriptional complexes with other proteins that are essential for the expression of *CRP* (Nishikawa et al. 2008; Toniatti et al. 1990a). The biological interaction between HNF1A and CRP makes it a plausible link to investigate as a potential biomarker for HNF1A-MODY.

6.1.3 Evaluating hsCRP as a biomarker for HNF1A-MODY in South Asians

Studies assessing CRP levels across different populations have shown that basal levels of CRP differ across ethnic groups (Shah and Kanaya 2014). Reports directly comparing healthy South Asians and Europeans have verified that basal levels of CRP are elevated in Asian Indians (Chambers et al. 2001; Forouhi et al. 2001). Additionally, elevated CRP levels have been correlated with increased BMI in South Asians and Europeans (Forouhi et al. 2001; Visser et al. 1999). The processes underlying systemic low-grade inflammation altering CRP levels have been attributed to several biological mechanisms, including central adiposity and markers of insulin resistance (Pannacciulli et al. 2001; Plomgaard et al. 2005), both of which have an increased propensity in South Asians.

The evaluation of whether low levels of hsCRP can identify South Asian subjects at high risk of having HNF1A-MODY was performed by a member of my supervisor's team, Dr Saima Aitzaz. In Dr Saima Aitzaz's study, 206 South Asian individuals were screened using capillary sequencing for mutations in *HNF1A*. Of these, 188 subjects had an hsCRP ≤1.0 mg/L and 18 had CRP levels ≥1.0 mg/L but were diagnosed with diabetes below the age of 30 years. Details of the samples chosen by Dr Saima Aitzaz that underwent capillary sequencing will be provided below in section 6.2.1 as context. The cutoff ≤1.0 mg/L was chosen based on the previously published evaluations of hsCRP in Europeans and that 80% of HNF1A-MODY patients had CRP ≤1.0 mg/L (Owen et al. 2010). Seven *HNF1A* sequence alterations were identified in patients with hsCRP levels ≤1.0mg/L; these included two heterozygous frameshift mutations and five heterozygous missense variants. My role was to functionally characterise the five-missense variants using *in vitro* assays. It is not known whether these *HNF1A* variants, identified in patients diagnosed with T2D and having low hsCRP levels, influence HNF1A function, and are causal for the diabetic phenotype.

6.1.4 Experimental aims

The aims of this study were:

1. To functionally characterise *HNF1A* variants identified in individuals of South Asian ancestry with hsCRP levels indicative of HNF1A dysfunction; and
2. To evaluate whether hsCRP is a good marker for HNF1A dysfunction in the South Asian population.

6.2 Samples and methods

6.2.1 Subjects

In total, 206 South Asian participants clinically diagnosed as having early-onset T2D from three independent studies were selected for *HNF1A* capillary sequencing. Participants were selected based on two criteria: (a) early-onset T2D <50 years old and hsCRP ≤1.0 mg/L or (b) age of onset for T2D ≤30 years irrespective of hsCRP levels. Capillary sequencing and identification of *HNF1A* variants was completed by Dr Saima Aitzaz. The clinical characteristics of the selected patients are detailed in Table 6.2.1.

6.2.1.1 London Life Sciences Prospective Population (LOLIPOP) study

The LOLIPOP study is a population-based cohort comprising multiethnic samples ascertained from GP surgeries in West London. The cohort includes both healthy controls and diabetes patients. South Asian patients with T2D who were <50 years at sampling (*n*=749) were identified. Of these, 159 subjects with hsCRP <1.0 mg/L were selected for *HNF1A* sequencing. The age at which diabetes was diagnosed was unavailable for the subjects in this cohort. The LOLIPOP study is ethically approved by the West London Ethics Committee and all subjects gave informed consent.

6.2.1.2 Young Diabetes in Oxford (YDX) study

Subjects in the YDX cohort are from Oxfordshire and were all diagnosed with diabetes at ≤45 years of age. Forty South Asian patients labelled with T2D were identified. All subjects were C-peptide positive, had no requirement for permanent insulin within 3 months of diagnosis, and were negative for GAD antibodies. From the 40 patients, 17 were selected for *HNF1A* sequencing. Out of these, 16 patients had hsCRP <1.0 mg/L and one subject had a CRP >1.0 mg/L but was included because they had an age at diagnosis of diabetes ≤30 years. All subjects gave informed consent and the study was approved by the Oxfordshire Local Research Ethics Committee.

6.2.1.3 South Asian Birmingham study

The last cohort included South Asians diagnosed with T2D at ≤40 years of age (*n*=72) from Birmingham, UK. Thirty subjects (13 subjects with hsCRP ≤1.0 mg/L and 17 subjects with hsCRP >1.0 mg/L but an age at diagnosis of diabetes of ≤30 years) were taken forward for capillary sequencing. The study was approved by the Local Research Ethics Committee and all subjects gave informed consent.

	Subjects with hsCRP ≤1.0 mg/L (*n*=188)			Subjects with age at diagnosis of diabetes ≤30 years (irrespective of hsCRP) *n*=18	
	LOLIPOP *n*=159	**Birmingham *n*=13**	**YDX *n*=16**	**Birmingham *n*=17**	**YDX *n*=1**
Age at sampling (years)*	45.3 (35.0–49.9)	39.0 (24.0–69.0)	49.4 (32.1–70.3)	41.0 (21.0–57.0)	25
Age of onset (years)*	N/A	31.0 (20.0–39.0)	36.0 (26.0–45.0)	26.0 (19.0–30.0)	21
BMI (kg/m^2)	26.0 (23.9–28.7)	N/A	24.9 (22.3–26.0)	N/A	30.7
HbA1c (%)	7.2 (6.2–8.4)	N/A	7.4 (6.6–8.3)	NA	6.0
Glucose (mmol/L)	7.6 (6.5–9.3)	N/A	7.9 (5.5–8.8)	N/A	4.5
HDL-cholesterol (mmol/L)	1.2 (1.1–1.3)	N/A	1.3 (1.0–1.5)	N/A	1.30
Total cholesterol (mmol/L)	4.9 (4.2–5.6)	N/A	4.2 (3.8–4.7)	N/A	4.7
Triglycerides (mmol/L)	1.5 (1.1–2.3)	N/A	1.3 (0.7–1.6)	N/A	1.40
HsCRP (mg/L)*	0.4 (0.0–1.0)	0.25 (0.02–0.87)	0.14 (0.01–0.84)	6.20 (1.80–28.5)	13.20

Table 6.2.1 Clinical and biochemical characteristics of subjects selected for sequencing. Data are presented as median (interquartile range). NA, data not available. * Criteria used for selection are shown as median range (hsCRP ≤1.0 mg/L or age of diagnosis <30 years irrespective of hsCRP levels). Data provided by Dr Saima Aitzaz.

6.2.2 Mutational screening of HNF1A

Genetic testing and analysis was performed in the laboratory of collaborator Professor Sian Ellard in a CPA-accredited Molecular Genetics Laboratory at the Royal Devon and Exeter National Health Service (NHS) Foundation Trust. Semi-automated unidirectional sequencing of *HNF1A* exons 1–10 was performed on an ABI 3730 capillary sequencer (Carlsbad, CA, USA) and analysed using Mutation Surveyor v3.24 (SoftGenetics, State College, PA, USA).

6.2.3 Assessing the pathogenicity of HNF1A variants identified

The pathogenicity of the identified *HNF1A* variants was evaluated using a pipeline similar to that described in Chapter 5. All variants were evaluated using *in silico* prediction tools. The presence of each variant was checked in 1000 Genomes Project (phase 3 release), the Exome sequencing project (ESP) and the Exome Aggregation Consortium (ExAC, version 0.3). In the 1000 Genomes Project and ESP the disease status of participants is unknown. In the ExAC database, it is known that some participants have T2D and some may have MODY but it is not known who these people are. All variants were then functionally characterised by assessing HNF1A transcriptional activity using two distinct reporter constructs (albumin and CRP) in HeLa and INS-1 cells. In addition, protein expression studies via western blot analysis and subcellular localisation of the variants via immunofluorescence was completed.

6.2.4 In silico assessment of the HNF1A variants

As described in section 2.18, three bioinformatics software packages were used to evaluate whether the variants would affect protein function. These included PolyPhen2, SIFT and Condel.

6.2.5 Site-directed mutagenesis and construction of HNF1A mutants

In vitro studies were performed with a pRK5 expression vector comprising full-length human *HNF1A* cDNA containing an in-frame N-terminal FLAG tag as described in Chapter 5, (section 5.2.7). Mutagenesis was carried out using the QuikChange II Site-Directed mutagenesis kit (Agilent Technologies Inc, Santa Clara, CA, USA) as described previously in Chapter 2, (section 2.5). Primers were designed complementary to the mutations and are listed in Table 6.2.2. All generated constructs were verified by DNA sequencing prior to commencement of expression studies as described in Chapter 2, (section 2.12).

***HNF1A* variant**	**Primer sequence**
V103M_F	5'-cccaccagaaagccAtggtggagaccctt-3'
V103M_R	5'-aagggtctccaccaTggctttctggtggg-3'
G339S_F	5'-cctcaagcagcggcAgtcccttagtgaca-3'
G339S_R	5'-tgtcactaagggacTgccgctgcttgagg-3'
P379S_F	5'-gccccctcccTctgtcagcacc-3'
P379S_R	5'-ggtgctgacagAggggaggggc-3'
T515M_F	5'-ggcccagtacacccacaTaggcctgct-3'
T515M_R	5'-agcaggcctAtgtgggtgtactgggcc-3'
L584P_F	5'-ggcccaccggcCcagcgccagcc-3'
L584P_R	5'-ggctggcgctgGgccggtgggcc-3'
P112L_F	5'-ccttctgcaggaggaccTgtggcgtgtggcg-3'
P112L_R	5'-cgccacacgccacAggtcctcctgcagaagg-3'
R271W_F	5'-caactggtttgccaacTggcgcaaagaagaagcc-3'
R271W_R	5'-ggcttcttctttgcgccAgttggcaaaccagttg-3'

Table 6.2:2: Primer sequences used for site-directed mutagenesis. The purification for each primer was salt free. The letter in red denotes the nucleotide that was incorporated into the sequence during mutagenesis.

6.2.6 Generation of pGL3-CRP reporter construct

A CRP expression construct was designed to evaluate whether the *HNF1A* variants identified in patients with reduced CRP levels affected HNF1A function. The CRP sequence was retrieved using UCSC and based on the previous publications by Dr Jana Rundle, a postdoc in the Gloyn lab, and was ordered from GeneArt (Life Technologies), arriving in a pMA vector (Toniatti et al. 1990a). Additional 5'Mlu and 3'XhoI sites were added to either end of the sequence for ease of cloning into pGL3-basic vector, which contained the luciferase (firefly) gene. DNA for the pMA-CRP and the pGL3-basic vector was transformed as described in Chapter 2, (section 2.6) and left overnight to grow at 37°C. The following day a single colony was picked and grown and DNA was subsequently prepared using the QIAprep Spin Miniprep kit as described in Chapter 2, (section 2.9). The DNA for both pMA-CRP and pGL3-basic was digested as described in Chapter 2, section (2.11) using two restriction enzymes, MLU I and Xho 1 (Promega UK Ltd, Southampton, UK), for 2.5 hours on a heat block at 37°C. After digestion, DNA was run on a 1.5% (3g of agarose and 200ml of 1xTAE) gel at 140V for approximately 45 minutes. Care was taken to ensure there was sufficient space between each sample when running the gel. Next, the DNA bands were cut out carefully using ultra violet (UV) light for the CRP (~600bp) and the pGL3-basic (~4.8kb). DNA bands were stored in 1.5-ml microcentrifuge tubes at -20°C. Following manufacturer guidelines, the Qiagen gel extraction kit was used to extract the DNA from the agarose gel (QIAgen Ltd, Crawley, UK). An overnight ligation process was then performed, which required the insert (CRP) to be in vast excess of the vector. The appropriate volumes of DNA were added to a 1.5-ml microcentrifuge tube and samples were incubated overnight at 4°C. The next day, the ligated DNA was transformed using DH5α competent cells as described in Chapter 2, (section 2.6) and grown overnight on ampicillin resistant plates at 37°C. Single colonies were then picked the following day, bacteria were grown and DNA was prepared using the Qiagen mini prep kit as per manufacturer guidelines. The DNA was sequenced to confirm that the CRP promoter was correctly inserted in the correct orientation into the open pGL3-vector.

6.2.7 Transactivation assays

Transactivation assays were performed in HeLa and INS-1 cells using the dual reporter luciferase system as described earlier in Chapters 2 and 5 (sections 2.15 and 5.8), respectively. The transactivation potential of the *HNF1A* variants was assessed using two liver-specific reporter constructs, albumin and CRP. The promoters of both these genes contain part of the consensus binding sequence for *HNF1A,* which is required to facilitate

gene expression (Lichtsteiner and Schibler 1989; Mendel and Crabtree 1991; Toniatti et al. 1990a). The albumin construct was evaluated in both HeLa and INS-1 cells and the CRP promoter in HeLa cells only. Optimisation experiments were completed to determine the ideal conditions under which to use the *CRP* promoter. The results of these experiments indicated that the DNA concentrations used for the albumin construct and wild-type or variant HNF1A were also appropriate for the *CRP* promoter construct. Two positive controls, HNF1A-P112L and HNF1A-R271W (described in Chapter 5), were included in all transactivation assays, as they have both been previously reported as MODY mutations and have been demonstrated to alter *HNF1A* function (Bjorkhaug et al. 2000; Bjorkhaug et al. 2003; Colclough et al. 2013).

Transcriptional activity was measured 24 hours post transfection using the Dual Luciferase Assay kit (Promega, Madison, WI, USA) and the Veritas Microplate Luminometer (Promega). Three independent experiments on consecutive days were performed in triplicate for each cell line. Percentage transcriptional activity was calculated by normalising to wild-type HNF1A.

6.2.8 Protein expression studies—western blots

Western blot studies were performed using the HeLa cell lysates collected from the transactivation assays using the albumin reporter construct. As described in Chapter 2, (section 2.16.1), two representative western blots were completed using one triplicate well from two separate luciferase assays. Protein expression levels were evaluated using the Bio-Rad gel and imaging system (Bio-Rad, Hertfordshire, UK). Densitometry was performed by combining the data from the two experiments.

6.2.9 Cellular localisation—immunofluorescence

Immunostaining was performed in both HeLa and INS-1 cells as described in Chapter 2, section 2.17. Images were viewed using the confocal microscope (Nikon model, Eclipse, E600FN) using the Bio-Rad coloured filters. All images were superimposed using software available online (http://www.zeiss.co.uk/microscopy/en_de/downloads/lsm-5-series.html).

6.2.10 Statistical Analysis

An ANOVA was carried out to compare the mean fold activation between wild-type HNF1A and each variant as described previously in Chapter 2 (section 2.19). An unpaired t-test was used to compare differences in protein levels for wild-type and

variant HNF1A observed in the western blot analysis. A *P*-value of <0.05 was considered significant.

6.3 Results

6.3.1 Identification of HNF1A variants in South Asian Subjects

Capillary sequencing of *HNF1A* in 206 South Asian participants in the study completed by Dr Saima Aitzaz revealed seven *HNF1A* sequence alterations. All variants were identified in patients with hsCRP ≤1.0 mg/L. Table 6.3.1 summarises the details of the seven variants including the hsCRP levels of each patient and the *in silico* evaluation for each variant.

Of the five missense variants identified, two have been previously published (c.307 G>A.pV103M and c.1135 C>T.p.P379S) whilst three have not been previously described in the literature (c.1015 G>A.p.G339S, c.1544 C>T.p.T515M, c.1751 T>C.p.L584P). Both frameshift mutations had not been previously reported. Four of the missense variants, p.V103M, p.P379S, p.T515M and p.L584P, were identified in subjects from the LOLIPOP cohort. Participants from the LOLIPOP cohort were part of a prospective population study and therefore individual family information was unavailable for these patients. One variant, p.G339S, was found in a patient from the YDX cohort. Members of the YDX study were recruited from the Oxfordshire area and so it was possible to recruit family members for further genetic studies.

In vitro assays were carried out to evaluate the five missense variants in line with the functional pipeline described in Chapter 5.

- Figures 6.3.1 (a & b) display the results of the transactivation assays performed in HeLa and INS-1 cells using the albumin reporter construct.
- Figure 6.3.2 shows the results of the transactivation assay completed in HeLa cells using the CRP promoter construct.
- Figure 6.3.3 displays the western blot analysis and densitometry comparing protein expression levels of HNF1A protein variant.
- Figures 6.3.4 and 6.3.5 show illustrative images captured in HeLa and INS-1 cells, respectively, to determine subcellular localisation of variant HNF1A proteins.

Below are described the results of the clinical, genetic (where available) and functional assessment of each missense variant.

6.3.2 c.307 G>A.p.V103M

The p.V103M variant was identified in a subject from the LOLIPOP cohort with the patient having an hsCRP of 0.4 mg/L. The p.V103M variant has been previously reported as a neutral polymorphism in a study that examined HNF1A-MODY in South Asian patients. The variant was identified in six unrelated subjects with diabetes but no information was provided to suggest that this variant did or did not co-segregate with diabetes (Radha et al. 2009). In the present study, PolyPhen 2 predicted the variant to influence protein function whilst both SIFT and Condel predicted the variant to be tolerated (Table 6.3.1). The p.V103M variant was identified in one European in the Exome Sequencing Project whilst in the ExAC dataset, the p.V103M variant was identified twice in Europeans and once in South Asians. The variant has not been reported in the 1000 Genomes Project (phase 3).

The transactivation potential of HNF1A-V103M was assessed using both the albumin and the CRP promoter constructs (Figures 6.3.1 and 6.3.2). With the albumin promoter, HNF1A-V103M showed a statistically significant (P<0.05) reduction in activity compared with wild-type HNF1A across both HeLa and INS-1 cells. Evaluation with the *CRP* construct in HeLa cells showed a greater reduction in activity, almost 30% (P=0.00004) relative to wild-type protein, the highest compared with the other alleles investigated. Western blot analysis and densitometry revealed no difference in protein levels relative to wild-type protein (Figure 6.3.3). Subcellular localisation studies in both HeLa and INS-1 cells showed that HNF1A-V103M exhibited nuclear localisation similar to that of wild-type HNF1A and staining was seen mainly in the nucleus (Figures 6.3.4 and 6.3.5). The transactivation assays with both the albumin and CRP promoters suggest that HNF1A-V103M affects HNF1A activity. As both protein expression and localisation studies seemed to behave similar to wild-type, it is postulated that this variant, located in the DNA binding domain of HNF1A, may cause weakened DNA binding affinity, resulting in reduced activity, and this requires further investigation.

Although the study by Radha et al. considered p.V103M to be a 'polymorphism' the variant was identified in six subjects with diabetes. Furthermore, the functional work in the current study suggests that HNF1A-V103M affects HNF1A function. Therefore, p.V103M is considered to increase risk for T2D in the present study.

Exon	Nucleotide position & substitution	Amino acid change	HNF1A domain	hsCRP mg/L	PolyPhen 2	SIFT	Condel	Previously published
1	c.14delT	p.L5Rfs17 The deletion results in a frameshift at codon Leu5. The new reading frame creates a STOP codon downstream at position 16.	Dimerization	0.4	N/A	N/A	N/A	Not published
1	c.307 G>A	Val103Met	DNA binding	0.4	Probably damaging	Tolerated	Neutral	Radha et al. (2009)
4	c.872dupG	p.292Rfs25 The duplication at position Gly292 causes a frameshift. The new reading frame results on a stop codon 24 positions downstream.	Transactivation	0.5	N/A	N/A	N/A	Not published
5	c.1015 G>A	Gly339Ser	Transactivation	0.8	Benign	Tolerated	Neutral	Not published
6	c.1135 C>T	Pro379Ser	Transactivation	0.8	Probably damaging	Tolerated	Deleterious	Bellanne-Chantelot et al. (2008)
8	c.1544 C>T	Thr515Met	Transactivation	ND	Probably damaging	Affect protein function	Deleterious	Not published
9	c.1751 T>C	Leu584Pro	Transactivation	0.1	Possibly damaging	Affect protein function	Neutral	Not published

Table 6.3.1: Summary of *HNF1A* variants identified in South Asians by capillary sequencing. ND, Not detected. Clinical and biochemical data provided by Dr Saima Atizaz.

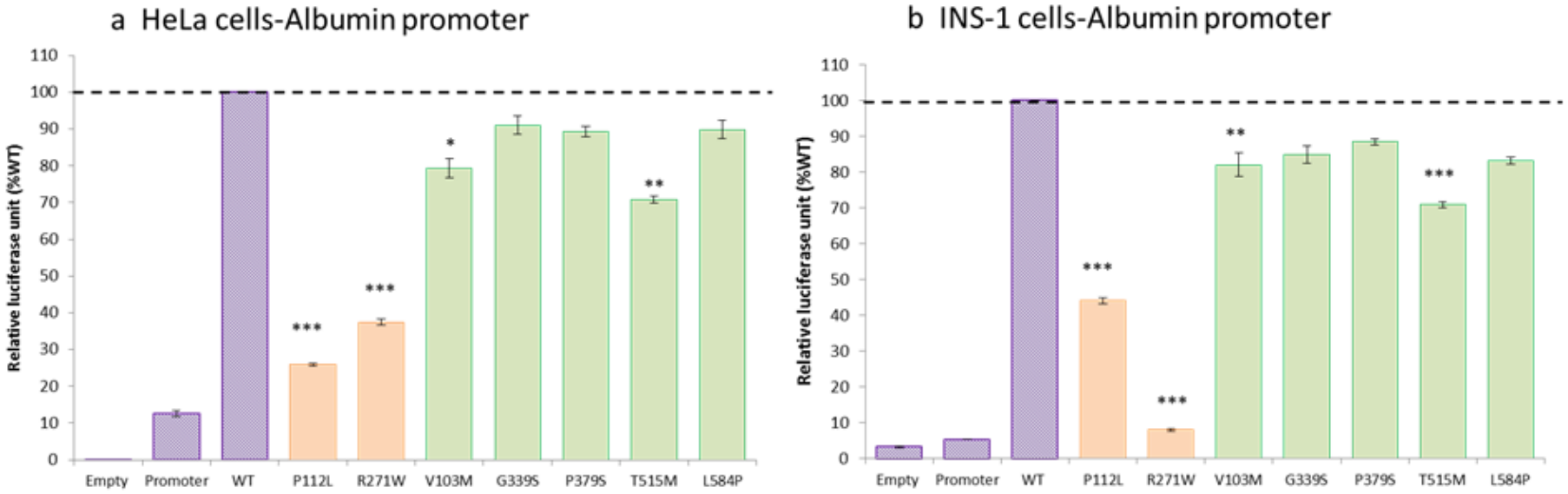


Figure 6.3.1: Transcriptional activity of wild-type (WT) and mutant HNF1A in HeLa and INS-1 cells using the albumin reporter construct (*n*=3). P112L and R271W (orange bars) were used as positive controls as they have been previously shown to alter HNF1A activity. Cells were harvested 24 hours after transfection and transcriptional activity was measured using a dual luciferase (firefly) reporter system. Luciferase values were normalised to HNF1A-WT and are presented as percentage WT (100%). All experiments were completed in triplicate and repeated four times with different DNA preparations. An ANOVA was carried out to compare the means of the different groups, *≤0.05–0.01, **<0.01–0.001 and ***<0.001. Error bars represent the standard error of the mean (SEM).

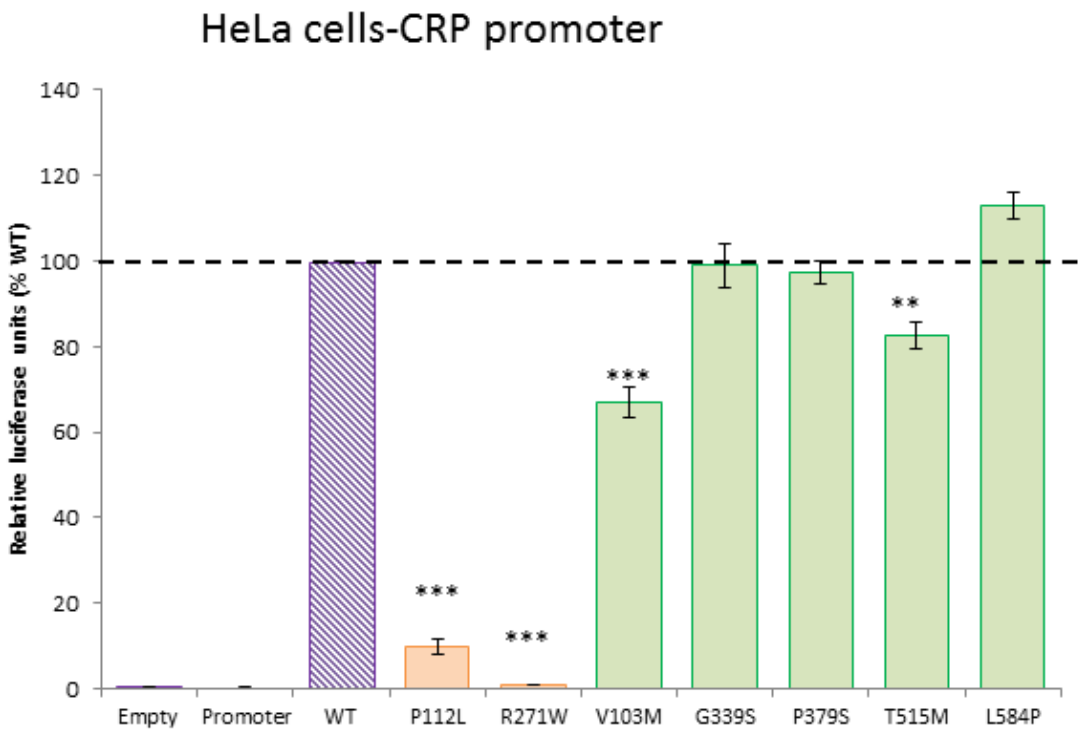


Figure 6.3.2: Transcriptional activity of wild-type (WT) and mutant HNF1A in HeLa cells using the CRP reporter construct (*n*=4). P112L and R271W (orange bars) were used as positive controls as they have been previously shown to alter HNF1A activity. Cells were harvested 24 hours after transfection and transcriptional activity was measured using a dual luciferase (firefly) reporter system. Luciferase values were normalised to HNF1A-WT and are presented as percentage WT (100%). All experiments were completed in triplicate and repeated four times with different DNA preparations. An ANOVA was carried out to compare the means of the different groups, *≤0.05–0.01, **≤0.01–0.001 and ***<0.001. Error bars represent the standard error of the mean (SEM).

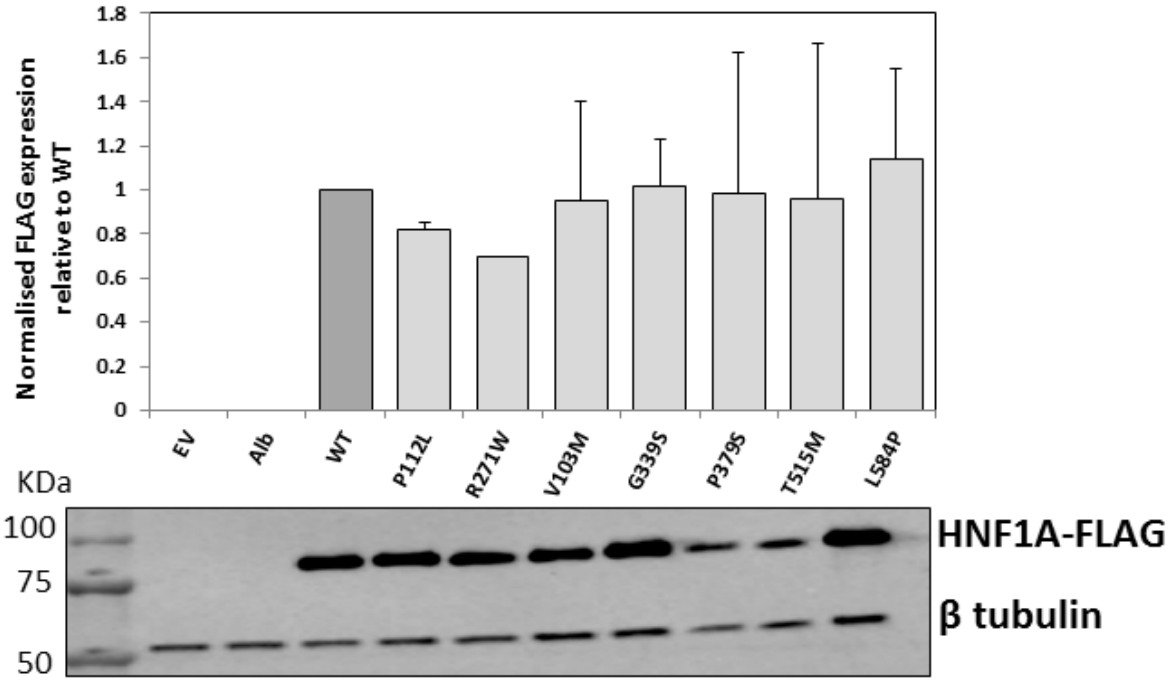


Figure 6.3.3: Expression levels of wild-type and variant HNF1A. Expression levels in HeLa cells were determined by western blot and densitometry using transactivation cell lysates. The in-frame N-terminal FLAG tag was used to detect protein expression of wild-type and mutant HNF1A. Tubulin was used as a control as protein expression was expected to be equal across all samples. Lysates used were those from the luciferase assays. Data are presented as a mean ± standard error of the mean for 2 independent experiments. An unpaired t-test was performed to compare the means between wild-type and variant HNF1A; no significant differences were observed ($P>0.05$). EV, empty vector; Alb, albumin; WT, wild-type. The western blot was run by Dr Jana Rundle using the cell lysates that I collected under experimental conditions that I optimised.

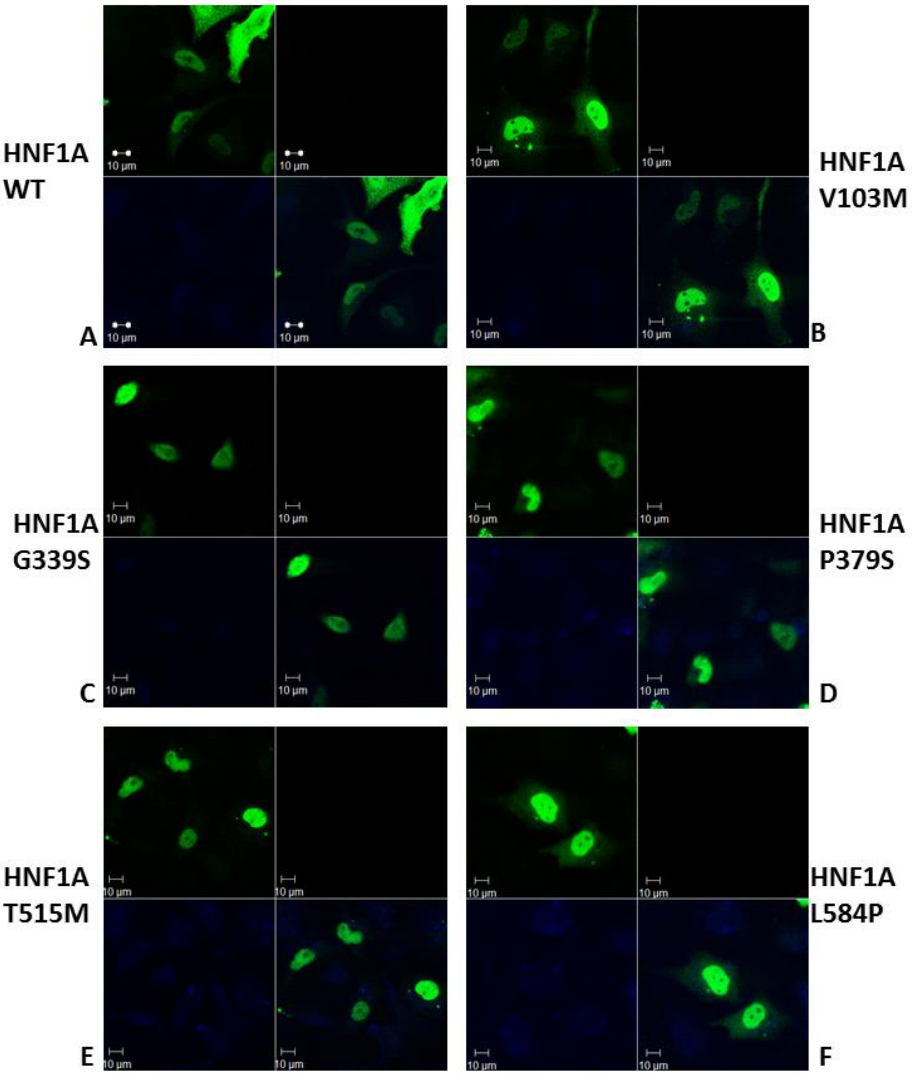


Figure 6.3:4: Subcellular localisation of wild-type and variant HNF1A in HeLa cells. Representative images were captured in INS-1 cells, which were transiently transfected for 24 hours prior to immunostaining. Cells were stained with Alexa Fluor-488 (Green) and far-red nuclear stain (NucBlue Live 647 Ready probes) and viewed using the confocal microscope using Bio-Rad filters. Images were superimposed using Zeiss LSM Image browser (http://www.zeiss.co.uk/microscopy/en_de/downloads/lsm-5-series.html).

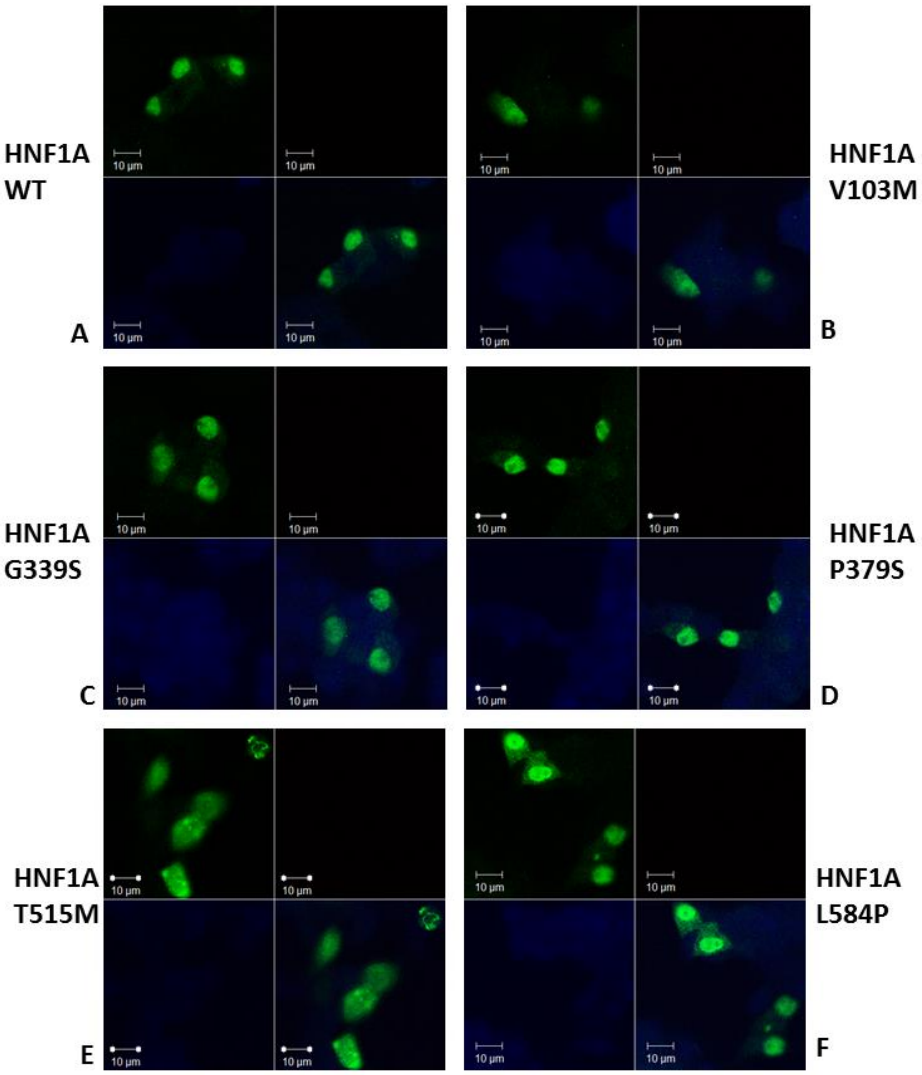


Figure 6.3:5: Subcellular localisation of wild-type and variant HNF1A in INS-1 cells. Representative images were captured in INS-1 cells, which were transiently transfected for 24 hours prior to immunostaining. Cells were stained with Alexa Fluor-488 (Green) and far-red nuclear stain (NucBlue Live 647 Ready probes) and viewed using the confocal microscope using Bio-Rad filters. Images were superimposed using Zeiss LSM Image browser (http://www.zeiss.co.uk/microscopy/en_de/downloads/lsm-5-series.html).

6.3.3 c.1015 G>A.p.G339S

As the p.G339S variant was identified in a participant from the YDX cohort with an hsCRP of 0.8 mg/L, it was possible to recruit family members for further genetic studies. Figure 6.3.6 illustrates the family pedigree of the proband.

Of the family members screened, three were found to be carriers of the p.G339S variant. The sister of the proband, who was found to carry the p.G339S allele, is currently asymptomatic aged 29 years, though the development of diabetes later on in life cannot be ruled out. Furthermore, there is no history of the affected members showing sensitivity to sulphonylurea treatment. The evidence as to whether the p.G339S allele co-segregates with a MODY phenotype in this family is inconclusive. Although the young age at which diabetes was diagnosed, for the proband and mother is consistent with a MODY phenotype, the presence of an unaffected relative with the allele and the apparent lack of sensitivity to sulphonylurea treatment by affected carriers are not supportive of a MODY phenotype. *In silico* evaluation by all three programmes agreed that this variant would not affect protein function (Table 6.3.1). The p.G339S variant was not observed in the ESP or 1000 Genomes Project. The ExAC database reports 19 heterozygous observations of the p.G339S variant, 13 in Latino individuals and 6 in South Asians; however, the phenotype of the participants is unknown so we do not know whether these carriers have diabetes.

All *in vitro* assays evaluating HNF1A-G339S demonstrated that this allele had no impact on HNF1A function supporting the *in silico* predictions. Taking together inconclusive family data and functional evaluation that suggests the allele behaves similar to that of wild-type HNF1A it is likely that p.G339S is a neutral variant.

6.3.4 c.1135 C>T.p.P379S

The p.P379S variant, identified in a subject from the LOLIPOP cohort, has been previously reported in two unrelated individuals. Bellanne-Chantelot et al. (2008) examined 356 unrelated HNF1A-MODY patients (87% were Europeans). No additional family data were provided in this report for the p.P379S variant, so it is unknown whether this variant co-segregates with diabetes. In the present study, the hsCRP of the proband was 0.8 mg/L. The variant was not present in the 1000 Genomes Project or in the ESP. The p.P379S variant was observed in the ExAC data set in seven South Asians but their diabetic status is unknown, so limited inferences can be made about whether this variant might increase for

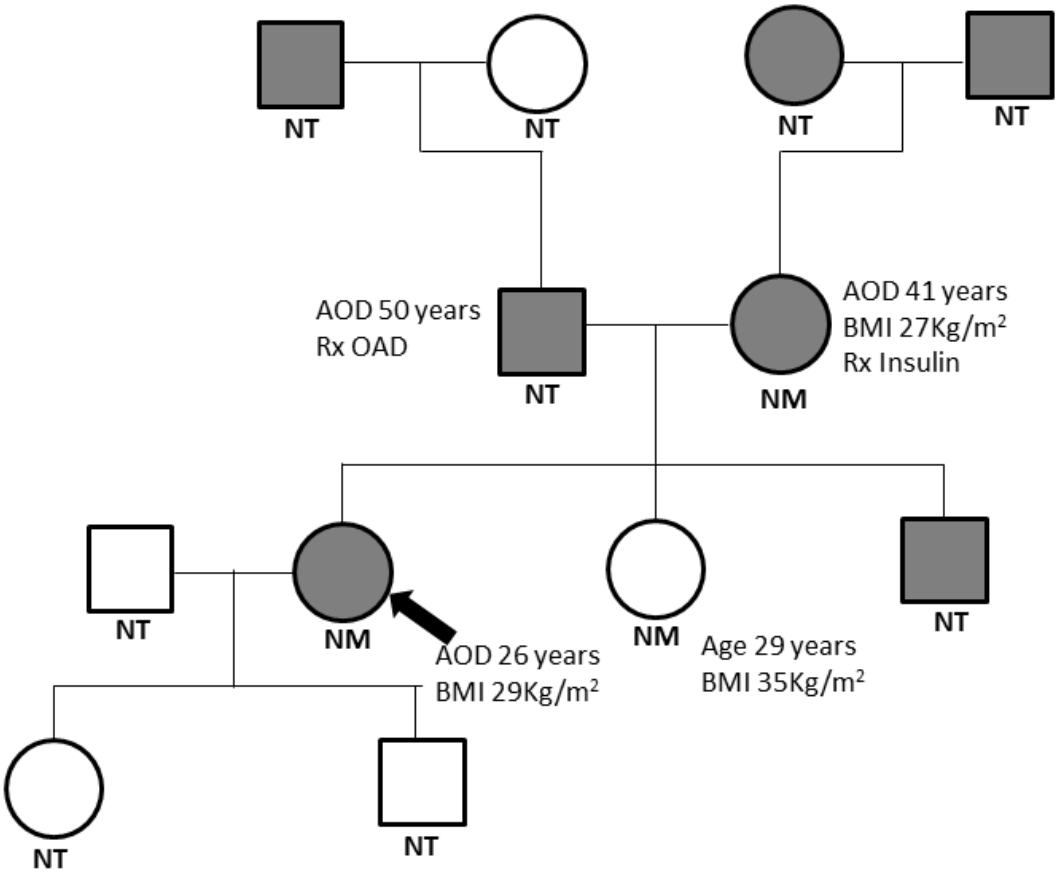


Figure 6.3:6: Family pedigree of the proband identified with p.G339S variant. Squares represent males and circles females. Filled symbols indicate diabetic subjects. NM, heterozygous mutation carrier; NN, no mutation; NT, not-tested; Rx, diabetes treatment; OAD, oral antidiabetic agent; BMI, body mass index; AOD, age at diagnosis. The black arrow designates the proband.

T2D. Two out of the three bioinformatics tools, PolyPhen2 and Condel, predicted that this variant might affect protein function (Table 6.3.1). Functional characterisation of this allele showed no difference compared with wild-type HNF1A protein across all three assays. This suggests that p.P379S is neutral in the current study.

6.3.5 c.1544 C>T.pT515M

The p.T515M variant was identified in a subject from the LOLIPOP cohort with the patient having undetectable hsCRP levels. *In silico* predictions, all agreed that this variant would have a deleterious effect on protein function. The variant was not present in either the ESP or the 1000 Genomes Project but was present in the ExAC database in two European individuals. Transactivation assays showed a statistically significant reduction in transcriptional activity with both the albumin ($P<0.05$) and the *CRP* ($P<0.01$) promoter constructs for HNF1A-T515M (Figures 6.3.1 and 6.3.2). Protein expression studies and nuclear localisation studies appeared to be similar to wild-type HNF1A. The results from the transactivation assays showed an impact on HNF1A function which was consistent with the *in silico* predictions. In the current study, it is possible that the p.T515M allele increases risk for T2D.

6.3.6 c.1751 T>C.p.L584P

The p.L584P variant was identified in a proband from the LOLIPOP cohort with an hsCRP of 0.1 mg/L. There were inconsistencies between the *in silico* programmes with PolyPhen 2 and SIFT predicted that p.L584P would affect protein function, whereas Condel predicted a neutral outcome (Table 6.3.1). The p.L584P variant was reported in one South Asian in the ExAC database and once in the 1000 Genomes Project; however, the variant was not present in the ESP. Functional assessment of this allele across all three functional assays demonstrated that HNF1A-L584P behaved similar to wild-type HNF1A. In the current study functional evaluation supports p.L584P as being a neutral variant.

6.3.7 c.14delT and c.872dupG

Probands detected with these single-base frameshift mutations were identified in the LOLIPOP cohort in participants with hsCRP 0.4 and 0.5 mg/L, respectively. The deletion results in a premature stop codon at residue 16, whilst the duplication causes a stop codon at position 24. Although not functionally characterised it is highly probable that both these frameshift mutations will have a deleterious effect on HNF1A protein synthesis and will have a causal role for diabetes in the patients.

6.4 Discussion

The frequency at which HNF1A-MODY is diagnosed in South Asians is much lower than in European counterparts (Porter et al. 2006; Shields et al. 2010). As very few studies have examined HNF1A-MODY in the South Asian population, additional studies are needed to broaden our understanding of this subtype of diabetes in this ethnic group. It is not known whether HNF1A-MODY has a lower occurrence in South Asians or whether cases of HNF1A-MODY are being missed or misdiagnosed in this ethnic group. In this study I functionally characterised *HNF1A* missense variants, detected in South Asian individuals with low hsCRP (<1.0 mg/L), giving the opportunity to investigate whether hsCRP could be used to detect MODY in high-risk South Asians.

The high prevalence of T2D, which tends to be diagnosed at a younger age and at lower BMI thresholds, favours the automatic diagnosis of T2D in most South Asian patients despite the overlap of clinical characteristics with HNF1A-MODY. Biomarkers, particularly low levels of plasma hsCRP, have been shown to be useful in discriminating between HNF1A-MODY and young onset T2D in European samples (McDonald et al. 2011; Owen et al. 2010; Thanabalasingham et al. 2011). Determining whether low levels of hsCRP is a good predictor of HNF1A-MODY in South Asians would offer a quick routine test that could be used to prioritise patients for *HNF1A* screening. Dr Saima Atizaz, a member of my supervisor's group, in her studies undertook capillary sequencing of *HNF1A* in 206 (out of a possible 861) South Asian patients diagnosed with apparent T2D. All selected participants had an hsCRP <1.0 mg/L or had been diagnosed with diabetes ≤30 years of age (irrespective of hsCRP levels). Sequencing detected seven sequence alterations within *HNF1A*, and by using *in silico* bioinformatics tools, it was predicted that six of the seven variants would be likely pathogenic. This gave an overall prevalence of 0.7% of HNF1A-MODY amongst a group of young onset T2D patients (unpublished, Dr Atizaz).

Of the five-missense variants that I functionally characterised, only two alleles, HNF1A-V103M and HNF1A-T515M, affected HNF1A function across both promoter constructs tested. It is unclear why the previous report of p.V103M identified in six diabetic patients was termed a 'polymorphism' (as opposed to 'mutation'—hence implying neutrality) (Radha et al. 2009). In the same study by Radha et al. they functionally characterised two variants found in the HNF1A promoter region, one of which (c.-373C>T) was labelled a 'polymorphism'. Although not statistically significant, transactivation assays evaluating the two HNF1A promoter regions revealed reduced activity relative to wild-

type for the promoter region that included the variant (c.-373C>T) labelled as a polymorphism (Radha et al. 2009). In the current study, the p.V103M allele clearly demonstrated reduced activity relative to wild-type HNF1A across both reporter constructs studied, similar to that seen in the Radha et al. study for the c.-373C>T allele located in the HNF1A promoter region. This interpretation of both p.V103M and c.-373C>T as (neutral) polymorphisms highlights the current challenge of assigning pathogenicity to variants (Radha et al. 2009). For the HNF1A-V103M allele, protein expression studies and nuclear localisation studies appeared to be similar to that of wild-type HNF1A. It is postulated that the reduced transactivation potential could be due to weakened DNA binding affinity. Further studies such as an EMSA—that enables you to measure the affinity between a protein and a particular nucleic acid sequence—could be performed to determine the impact of the variant proteins on DNA binding affinity.

The HNF1A-T515M allele, located in exon 8 and in the transactivation domain of HNF1A, was predicted to impact protein function by all three *in silico* tools and this was confirmed by the transactivation assays across both promoter constructs analysed. Missense variants in the later exons of *HNF1A* (8–10) are associated with a later age of onset for diabetes as only one isoform of *HNF1A* is affected. Consequently, clinical expression may be associated with a milder phenotype (Bellanne-Chantelot et al. 2008; Harries et al. 2006). Similar to the p.D526N discussed in Chapter 5 and the p.E508K variant identified in the Latino population (Estrada et al. 2014) it is thought that p.T515M may increase risk for T2D in the current study.

In the present study based on functional work only, the other three alleles assessed, HNF1A-G339S, HNF1A-P379S and HNF1A-L584P, are proposed to be neutral variants as they behaved similar to wild-type HNF1A in all three *in vitro* assays. Although a three-generation family history was available for the proband that carried the p.G339S allele, there was insufficient evidence to suggest that this variant co-segregated with MODY. Furthermore, in the ExAC sequence variant database the p.G339S was seen in 19 individuals, the highest of all the variants examined. One possibility for this high number could be that this nucleotide is weakly conserved across species and so susceptible to substitution. However, as the diabetic status of each carrier is unknown in the ExAc database, only suggestive inferences can be made about this variant.

All probands identified with *HNF1A* sequence alterations had hsCRP <1.0 mg/L. Assessment of HNF1A function with the CRP promoter showed that only two of the five

alleles influenced promoter activity and even then activity was only reduced to 70% for the HNF1A-V103M in comparisons with wild-type HNF1A. The two notable exceptions were the two positive controls HNF1A-P112L and HNF1A-R271W, both of which had nearly ~90% reduction in promoter activity with the CRP promoter. Both p.P112L and p.R271W and have been previously reported to cause MODY and have been shown to alter HNF1A function (Bjorkhaug et al. 2000; Bjorkhaug et al. 2003; Colclough et al. 2013). Like HNF1A-V103M, both HNF1A-P112L and HNF1A-R271W are located in the DNA binding domain of HNF1A. Subsequently, instead of weakening DNA binding as suspected with p.V103M, both the p.P112L and the p.R271W alleles may disrupt the consensus binding sequence that the CRP promoter recognises, consequently, abolishing DNA binding.

Although functionally not characterised, it is highly probable that the two frameshift mutations resulting in premature stop codons would affect HNF1A protein synthesis. The probands detected with the premature truncating variants (PTV) both had hsCRP levels of 0.4 and 0.5 mg/L, whilst the proband detected with p.L584P variant, a suspected neutral variant, had an hsCRP of 0.1mg/L. It is predicted that those with deleterious mutations such as PTVs would have lower hsCRP than those patients harbouring neutral variants. Therefore, it is possible that genetic variation in genes other than *HNF1A* can alter plasma hsCRP levels. For example, common variation in or near *CRP*, *LEPR*, *APOE, IL-6* and *GCKR* have all been associated with influencing hsCRP levels in Europeans (Reiner et al. 2008; Ridker et al. 2008). Investigations in North Indians have confirmed that common variants in *CRP* and *LEPR* are linked with changes in hsCRP levels (Mahajan et al. 2011). The current study only focused on *HNF1A*; therefore, further studies are warranted to determine whether other genetic mechanisms involved in regulating plasma hsCRP levels might be driving the lower levels of CRP detected in these patients.

A limitation of this study is the absence of family and clinical data for four of the variants investigated. It is not known whether the subjects carrying the variants fit into the classical criteria for HNF1A-MODY or if the patients respond to low-dose sulphonylurea. Therefore, the interpretations made on whether the alleles influence diabetes risk are based on the results of the functional assays only. One possibility for why three of the five alleles appear to be neutral is that the *in vitro* assays performed may not have been sensitive enough to detect small changes that might be relevant for a role in T2D. It is likely that the assays used would have sufficient sensitivity to detect those variants

causal for MODY as the anticipated effect on the promoter constructs would be greater. Furthermore, both reporter constructs (albumin and CRP) used were liver specific and displayed similar outcomes in activity for the five variants assessed. Additional studies using a beta cell relevant reporter construct such as HNF4A-P2 (as described in the results chapter 5) could be performed to gain a further understanding of whether these variants might be associated with diabetes risk. Further transactivation studies in a liver cell line expressing endogenous HNF1A and CRP might also be informative. These experiments would allow for the evaluation of the variants in the heterozygote state and in the presence of HNF1A dimerization.

Protein expression studies (Figure 6.3.3) were performed using the cell lysates collected after the transactivation assays, with no statistical difference observed between wild-type and variant HNF1A proteins. Although this was useful for determining directly whether the changes observed in activity were because of reduced protein levels, it was not possible to quantify the amount of protein loaded due to buffer incompatibility. In addition, the sample size for the western blots was small and therefore further studies are warranted to confirm the findings. Immunostaining of the HNF1A variant proteins in HeLa and INS-1 cells (Figures 6.3.4 and 6.3.5, respectively) to determine whether the alleles altered nuclear localisation demonstrated that all the variant proteins exhibited nuclear localisation similar to wild-type HNF1A. Studies examining the nuclear localisation signal of *HNF1A* have demonstrated that amino acid residues 197–205 are critical for correct targeting of HNF1A protein to the nucleus (Bjorkhaug et al. 2005). None of the variant proteins investigated in the current study fell in the region thought to be essential for nuclear signalling, consistent with the pattern of staining observed. For all variants including wild-type HNF1A, it was clear that some protein staining was observed in the cytoplasm. The overexpression of HNF1A in this system may have over-saturated the cell, causing HNF1A to leak out into the cytoplasm, thus the staining being detected in both cellular compartments.

In conclusion, this study has shown that although low hsCRP (<1.0mg/L) is able to detect HNF1A variants in South Asians they are not necessary causal for HNF1A-MODY. *In vitro* assays established that only two alleles might contribute to lower CRP levels. Therefore, based on the current studies of functional evaluation the readjusted prevalence of HNF1A-MODY is 0.5% (including p.V103M, p.T515M and the two PTVs) in a group of South Asian young onset T2D patients. However, the unavailability of family data and detailed clinical phenotype makes it difficult to ascertain completely whether

p.V103M and p.T515M are causal for MODY or increase risk for T2D. Further evaluation is required to determine whether HNF1A alleles would be found in those South Asian patients with normal/or high hsCRP levels or whether 'low hsCRP' detects South Asian patients with the highest risk for HNF1A-MODY. In addition, as basal levels of CRP are elevated in South Asians (Chambers et al. 2001) a higher cutoff for hsCRP may be necessary to have the same sensitivity and specificity as that reported in Europeans for hsCRP to discriminate between HNF1A-MODY and T2D.

Current work in the laboratories of supervisors Professor Anna Gloyn and Professor Mark McCarthy is now being directed by the functional work described in this chapter. As part of the T2DGENES/GoT2D consortia, HNF1A variants particularly identified in South Asian and East Asian subjects are currently being functionally characterised, as additional phenotypic data are available, which includes information on hsCRP levels. This will provide a much larger dataset for evaluating the performance of hsCRP as a potential diagnostic tool in non-European cohorts. If successful, hsCRP has the potential to be a relatively non-invasive, non-genetic marker that might be used diagnostically to prioritise individuals for genetic screening to check for mutations in *HNF1A.*

Chapter 7

General Discussion

The advent of genome-wide association studies saw the start of a transformative period in type 2 diabetes (T2D) research broadening our understanding of the biological mechanisms linked to T2D susceptibility. These studies have identified common genetic variants throughout the genome that are associated with T2D, all with modest effect sizes (reviewed in (Prasad and Groop 2015). Many of the initial studies were performed in individuals of European ancestry, requiring similar studies to be performed in non-European populations. Studying ethnically diverse populations has given the opportunity to investigate whether the associated loci are: (i) shared (supporting the view that the identified T2D signals are common), (ii) important in disease susceptibility and likely harbour the casual variant or (iii) novel signals proposing new biology.

Work presented in Chapters 3 and 4 is focused around individuals of South Asian origin. Chapter 3 complements the growing number of studies, which aim to replicate the association findings from Europeans studies, improving our understanding about whether these common variants can be generalised across populations. The examination of a subset of robustly associated single nucleotide polymorphisms (SNPs) with T2D in South Asians of Sri Lankan descent revealed no systematic difference in effect size or risk allele frequency for these SNPs when compared with the initial observations in Europeans. This is exemplified in studies focusing on other South Asian subgroups, which have also shown considerable overlap with the common variants evaluated in Chapter 3 and European studies (Chauhan et al. 2010; Chidambaram et al. 2010; Rees et al. 2011; Sanghera et al. 2008). The directional consistency (OR>1 for the risk allele) for many of the associated T2D loci has been demonstrated in several studies comparing multiple ethnic populations (Haiman et al. 2012; Waters et al. 2010). This consistent pattern of associations suggests that the established loci are important in disease risk, and it is likely that the functional common variant is likely to be shared across populations.

Chapter 4 describes that through collaborative large-scale genome-wide association studies (GWAS) it has been possible to identify novel common genetic variants associated with T2D in individuals of South Asian ancestry. The identification of these variants has added to the growing number of loci associated with T2D, revealed novel biological pathways and suggested that population-specific variants may be important. By investigating the genetic component of T2D predisposition in South Asians, the novel genetic variants might explain differences in T2D prevalence and presentation in this high-risk group.

Studies investigating the physiological indices of a subset of the associated T2D loci have shown that the majority of them influence beta cell function, insulin secretion or processing, with very few implicated with insulin sensitivity (Dimas et al. 2014). The South Asian GWAS described in Chapter 4 discovered six novel T2D susceptibility loci, including a variant near the gene encoding for Growth Factor Receptor-Bound Protein 14 (*GRB14*) (Kooner et al. 2011). GRB14 is an adaptor protein, which plays a role in signalling pathways involved in growth and metabolism. It has been suggested that the protein has an inhibitory function on insulin receptor signalling (Depetris et al. 2005). Studies in knockout mice for $Grb14^{-/-}$ exhibit better glucose tolerance even though they have lower levels of circulating insulin (Cooney et al. 2004). Furthermore, the overexpression of Grb14 has been shown to lead to increased insulin resistance (Cariou et al. 2004). Therefore, the link between *GRB14* and insulin sensitivity suggests a strong candidate for association (Kooner et al. 2011). In addition, other investigations have implicated *GRB14* as candidate for metabolic syndrome (Kraja et al. 2014). More recently, in a study exploring the relationship between endometriosis and fat distribution, variants near/in *GRB14* were found to be involved (Rahmioglu et al. 2015). Taken together, these findings indicate that *GRB14* is a plausible biological candidate for influencing T2D susceptibility.

Genome-wide association studies in non-European populations similar to those described in Chapter 4 have facilitated large-scale trans-ethnic meta-analysis and assisted in improving the resolution for fine mapping of causal variants (Morris 2011). The increase in sample size by combining multiple multiethnic GWAS studies has increased power to detect novel loci. A recent trans-ethnic meta-analysis study, which included individuals of South Asian ancestry from the GWAS described in Chapter 4, discovered seven novel T2D risk loci and confirmed that the majority of the common variants are shared across ancestral groups. The study also highlighted numerous common variants, which did not reach genome-wide significance but are homogenous across all ancestral groups and could contribute to T2D heritability (Mahajan et al. 2014). Furthermore, by exploiting differences in LD structure between populations and assuming the causal variant is the same across populations, comprehensive fine-mapping studies were undertaken to help refine T2D association signals. Mahajan et al. (2014) showed that at the *JAZF1* locus it was possible to reduce the genomic interval to include four possible SNPs compared to the previous nine, since these four SNPs were in LD with the lead SNP across all ancestral groups included in the meta-analysis.

Currently, it seems unlikely that common genetic variants can be attributed to the differences in presentation and prevalence seen in South Asians. Low- and rare-frequency variants yet to be detected could account for some of the variation. As described in Chapter 1, several recent examples applying next generation sequencing technologies in isolated populations have successfully identified novel low- and rare-frequency variants, which influence T2D susceptibility (Estrada et al. 2014; Steinthorsdottir et al. 2014). These studies highlight that further studies utilising whole exome and genome sequencing of South Asian individuals may reveal novel variants that might be responsible for the increased prevalence for T2D.

Changes in lifestyle, dietary behaviours, and reduced physical inactivity are undoubtedly driving the rise in T2D prevalence, especially in South Asians. Strategies to prevent and slow the progression of the disease in this high-risk group are required (Nolan et al. 2011). Lifestyle interventions involving nutritional alternations (reduced intake of saturated fat and increased dietary fibre) alongside regular physical activity have shown to reduce the relative risk by 36% (Lindstrom et al. 2006). For the South Asian population, the implementation of a lifestyle change can be met with reservations especially in native South Asian countries as religious and cultural beliefs play a key role in daily life. The current public health challenge is incorporating and then encouraging a lifestyle change to reduce the risk for diabetes (Walker et al. 2010).

In addition to identifying a substantial number of T2D susceptibility loci, GWAS have identified potential pathways that might have useful clinical applications as candidate biomarkers. The outcomes of several GWAS studies have highlighted that *HNF1A* may be important in regulating the levels of plasma glycans and hsCRP (Lauc et al. 2010; Okada et al. 2011; Reiner et al. 2008; Ridker 2001). This has led to the examination that glycan profiles and hsCRP could be used as non-genetic markers to detect HNF1A dysfunction as that caused in HNF1A-MODY. If successful, these biomarkers could be translated into the clinic as a diagnostic tool for discriminating between different subtypes of diabetes and HNF1A-MODY. The definitive molecular diagnosis of this subtype of MODY has important implications for the patient's treatment and disease management, as many patients can be managed successfully using oral sulphonylureas rather than insulin therapy (Pearson et al. 2003). HNF1A-MODY usually presents with an autosomal dominant inheritance pattern, and so for family members follow-up and genetic screening is important to assess risk for diabetes.

HNF1A is a particularly useful functional model because the physiological function of HNF1A is well established. This means that understanding physiological changes that may occur when there are variants in the *HNF1A* gene is more tractable than in cases when the function of the genes at associated loci are unknown. As the function and tissue expression patterns of *HNF1A* are known, searching for a measurable biomarker profile might be a convenient method for discriminating between subtypes of diabetes in the clinic. At present, the similarity in clinical presentation between HNF1A-MODY and other common subtypes of diabetes, particularly T2D, results in many cases being missed or misdiagnosed. Given the improved disease prognosis and quality of life for patients accurately diagnosed with HNF1A-MODY, work presented in Chapters 5 and 6 investigated whether non-genetic biomarkers could be used as an adjunct to clinical characteristics to distinguish those patients at highest-risk for HNF1A-MODY to guide appropriate genetic testing in the clinic.

Chapter 5 formed part of a study that evaluated whether plasma glycan profiles might be altered in patients with HNF1A-MODY. Capillary sequencing of patients clinically diagnosed with presumed T1D or T2D, all with a glycan profile (DG9-index) suggestive of *HNF1A* dysfunction, revealed 4 (out of 97) patients carrying variants in *HNF1A*, demonstrating that it was possible to identify patients with variants in *HNF1A* using the DG9-glycan index. However, given the sensitivity and specificity of the assay the number of cases detected was lower than expected (Thanabalasingham et al. 2013). The current study evaluated the DG9-glycan index only as it showed the most potential; however, in the initial validation study other glycan measurements differed significantly ($P<0.05$) between HNF1A-MODY and T2D patients and these might also warrant further investigation (Thanabalasingham et al. 2013). The functional *in vitro* assays were performed to determine whether the DG9-index was a good measure of *HNF1A* function, and whether the HNF1A variant was a functional and causal disease phenotype. Taken together, a combination of family studies (where possible), *in silico* evaluation and functional work showed that of the four variants identified, only one (p.R203H) is likely to be a fully penetrant HNF1A-MODY mutation. The other three *HNF1A* variants are suggested to increase risk for T2D in certain individuals in the current study.

The observation that *HNF1A* variants are associated with a spectrum of diabetic phenotypes adds to the complexity of determining whether a particular variant is causal for MODY, increasing risk for T2D or in some cases neutral, resulting in no adverse clinical effects. The sequencing of several MODY genes, including *HNF1A* in the study by

Flannick et al. (2013), discovered that the same rare variants previously implicated in HNF1A-MODY are found in the general population who remain euglycemic even at middle age. This highlighted that ascribing pathogenicity to rare variants remains a challenge, and the penetrance of a variant cannot be generalised in the population and that interpretation of diabetes risk may have to be assessed on an individual basis. This is exemplified by the p.D526N allele described in Chapter 5, which was discovered in a proband diagnosed with presumed T2D aged 63 years old with a DG9-glycan index indicative of HNF1A dysfunction. Subsequent genetic screening in family members revealed a sibling who carried the p.D526N allele but remained asymptomatic at aged 80 years. Functional evaluation of HNF1A-D526N suggested that the allele affects HNF1A function and could contribute to T2D susceptibility.

Several studies have evaluated hsCRP in individuals of European origin and have shown that it can be used to discriminate between HNF1A-MODY and T2D (McDonald et al. 2011; Owen et al. 2010; Thanabalasingham et al. 2011). Progress into translating these findings into non-European cohorts has been limited. Therefore, to address this Chapter 6 aimed to investigate whether hsCRP could be used as a biomarker in South Asians for detecting patients with the highest risk of HNF1A-MODY. Despite the overlapping clinical characteristics between HNF1A-MODY and T2D, HNF1A-MODY is largely overlooked in the clinic due to the high prevalence of T2D in South Asians and limited investigation of HNF1A-MODY in this ethnic group. Functional evaluation of five *HNF1A* variants, all identified in South Asians patients with young onset diabetes and with hsCRP levels <1.0mg/ml demonstrated that only two alleles (p.V103M and p.T515M) affected HNF1A function and may lead to increased T2D susceptibility. Three alleles (p.G339S, p.P379S and p.L584P) exhibited no effect on HNF1A function despite being identified in patients with low levels of hsCRP, and are suspected of being neutral variants in the present study and are unlikely to be causal for a diabetes phenotype. South Asians have increased prevalence of T2D driven by the increased tendency for insulin resistance, which is exacerbated by a sedentary lifestyle and dietary choices. Therefore, these features in addition to other genetic modifiers are likely to contribute to early onset diabetes predisposition.

Chapters 5 and 6 confirmed that glycan profiles (DG9-glycan index) and hsCRP are able to detect patients with variants in *HNF1A*. Whether these biomarker profiles can be used in the clinic as a measure of HNF1A dysfunction requires further investigation. Additional validation for the DG9-glycan index is necessary in an unselected group of patients with

young onset diabetes to assess the performance of the glycan index in a scenario that represents a typical diabetes clinic. Furthermore, routine clinical practise is limited due to the cost and lack of standardised glycan profiling (Thanabalasingham et al. 2013). In addition, investigations in non-European samples are still required. Similarly, the evaluation of hsCRP requires further assessment in South Asians. Whether HNF1A alleles can be detected in patients with normal or high hsCRP levels remains to be determined.

Whilst there are instances in which hsCRP has been used as a diagnostic tool for HNF1A-MODY (Besser et al. 2012), genetic testing is still the predominant method for screening. Although, the cost of diagnostic sequencing is decreasing, biomarkers will be beneficial, as they will enable targeted screening of *HNF1A*. Overall, in Chapters 5 and 6, nine *HNF1A* variants were identified in patients with diabetes due to biomarker profiles suggestive of HNF1A function. Of the nine variants, functional evaluation suggests that six of them may affect protein activity influencing diabetes susceptibility, proposing that neither biomarker outperformed the other as a measure of HNF1A dysfunction. Therefore, compared with existing diagnostic criteria (age of onset and family history), biomarkers might be able to give an indication of HNF1A function. Therefore, clinical application of biomarkers will be a useful adjunct to the clinical characteristics currently used in the diagnosis of HNF1A-MODY. Furthermore, it is proposed that in the future clinical diagnostics will require a composite of biomarkers to detect those patients with the highest risk of carrying a variant that affects HNF1A activity. Therefore, my work demonstrates that, as additional novel variants in *HNF1A* are uncovered, through large-scale sequencing projects, biomarker profiles could provide some insight into which variants might be implicated in diabetes risk.

The current challenge of interpreting the clinical significance of variants identified in *HNF1A* has been discussed in Chapters 5 and 6. Even with a suite of available tools (family studies, *in silico* information and functional characterisation); the data presented here demonstrate that it can be difficult to elucidate whether variants will cause a disease phenotype. Out of the nine HNF1A variant proteins assessed, only one (p.R203H) had sufficient evidence to suggest it was a genuine HNF1A-MODY mutation. For the other alleles investigated, the data indicated an increase risk for T2D in certain individuals, or no effect on HNF1A function, implying neutrality. This raises questions about whether MODY is a discrete entity or is it more complicated than previously presumed. Evidence now suggests that the lines between those variants causing MODY, those increasing susceptibility for T2D and those with no effect are blurred, and the risk

for diabetes needs to be determined on an individual basis, considering an array of factors that include, family history, ethnicity, impact of variants on *HNF1A* activity alongside lifestyle factors and environmental exposures.

Future work is now directed by the development of high-throughput functional assays, which will enable the cataloguing of all missense variants in *HNF1A*. Along with clinical information and bioformatics, it is anticipated this will assist in clinical translation to determine an individual's risk, which will help direct appropriate treatment and disease management. Induced pluripotent stem cells (iPSCs) are now being used to model different subtypes of diabetes including HNF1A-MODY. These studies will provide a valuable tool for studying the exact molecular mechanisms responsible for HNF1A-MODY (Stepniewski et al. 2015). In addition, the CRISPR-Cas9 system is proving to be innovative, by providing targeted methodology to edit the genome and look at the effects of specifically introduced variants. For example, it may be possible to look at *HNF1A* variants to elucidate the effects of the genetic variants in different cell types.

The data presented in this thesis have explored the role of common and rare genetic variants in T2D predisposition. Genome-wide association studies have been extended to South Asian populations (Chapter 4), an ethnic group with the highest prevalence for T2D. These studies have identified novel T2D risk variants, highlighting novel biology and demonstrating the importance of studying non-European populations. My work on the Sri Lankan samples (Chapter 3) was the first systematic review of known common T2D risk variants in this ethnic subgroup and showed that the variants are shared and not ethnic specific. In addition, this thesis has examined whether different biomarkers profiles are able to measure HNF1A dysfunction and so detect patients at highest risk for HNF1A-MODY (Chapter 5 and 6). My experimental evaluation of *HNF1A* alleles has demonstrated that not all variants have a detrimental impact on protein activity—in keeping with the suggestion that genetic alterations in *HNF1A* are associated with a spectrum of phenotypes. Together, my work has started to explore the current challenges faced with interpreting the clinical significance of coding variants implicated in diabetes susceptibly. Continued research is essential to enable successful clinical translation to help determine an individual's risk for diabetes to ensure that appropriate treatment and preventative care is provided.

Appendices

References and Manuscripts

Abecasis, G. R., et al. (2012), 'An integrated map of genetic variation from 1,092 human genomes', *Nature,* 491 (7422), 56-65.
ADA (2012), 'Diagnosis and classification of diabetes mellitus', *Diabetes Care,* 35 Suppl 1, S64-71.
Adzhubei, I. A., et al. (2010), 'A method and server for predicting damaging missense mutations', *Nat Methods,* 7 (4), 248-9.
Alberti, K. G. and Zimmet, P. Z. (1998), 'Definition, diagnosis and classification of diabetes mellitus and its complications. Part 1: diagnosis and classification of diabetes mellitus provisional report of a WHO consultation', *Diabet Med,* 15 (7), 539-53.
Altshuler, D., Daly, M. J., and Lander, E. S. (2008), 'Genetic mapping in human disease', *Science,* 322 (5903), 881-8.
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.
Amarasekara, A. A., et al. (2014), 'An ethnographic study of diabetes health beliefs and practices in Sri Lankan adults', *Int Nurs Rev,* 61 (4), 507-14.
Anderson, C. A., et al. (2011), 'Synthetic associations are unlikely to account for many common disease genome-wide association signals', *PLoS Biol,* 9 (1), e1000580.
Antuna-Puente, B., et al. (2008), 'Adipokines: the missing link between insulin resistance and obesity', *Diabetes Metab,* 34 (1), 2-11.
Anuradha, S., Radha, V., and Mohan, V. (2011), 'Association of novel variants in the hepatocyte nuclear factor 4A gene with maturity onset diabetes of the young and early onset type 2 diabetes', *Clin Genet,* 80 (6), 541-9.
Asfari, M., et al. (1992), 'Establishment of 2-mercaptoethanol-dependent differentiated insulin-secreting cell lines', *Endocrinology,* 130 (1), 167-78.
Ashcroft, F. M. and Rorsman, P. (2012), 'Diabetes mellitus and the beta cell: the last ten years', *Cell,* 148 (6), 1160-71.
Ayub, Q., et al. (2014), 'Revisiting the thrifty gene hypothesis via 65 loci associated with susceptibility to type 2 diabetes', *Am J Hum Genet,* 94 (2), 176-85.
Babenko, A. P., et al. (2006), 'Activating mutations in the ABCC8 gene in neonatal diabetes mellitus', *N Engl J Med,* 355 (5), 456-66.
Bach, I. and Yaniv, M. (1993), 'More potent transcriptional activators or a transdominant inhibitor of the HNF1 homeoprotein family are generated by alternative RNA processing', *Embo j,* 12 (11), 4229-42.
Bach, I., et al. (1990), 'Cloning of human hepatic nuclear factor 1 (HNF1) and chromosomal localization of its gene in man and mouse', *Genomics,* 8 (1), 155-64.
Bakker, L. E., et al. (2013), 'Pathogenesis of type 2 diabetes in South Asians', *Eur J Endocrinol,* 169 (5), R99-r114.
Bamshad, M. J., et al. (2011), 'Exome sequencing as a tool for Mendelian disease gene discovery', *Nat Rev Genet,* 12 (11), 745-55.
Banerji, M. A., et al. (1999), 'Body composition, visceral fat, leptin, and insulin resistance in Asian Indian men', *J Clin Endocrinol Metab,* 84 (1), 137-44.
Barbacci, E., et al. (2004), 'HNF1beta/TCF2 mutations impair transactivation potential through altered co-regulator recruitment', *Hum Mol Genet,* 13 (24), 3139-49.
Barresi, R., et al. (2000), 'Expression of gamma -sarcoglycan in smooth muscle and its interaction with the smooth muscle sarcoglycan-sarcospan complex', *J Biol Chem,* 275 (49), 38554-60.
Barroso, I. (2005), 'Genetics of Type 2 diabetes', *Diabet Med,* 22 (5), 517-35.
Bartel, D. P. (2004), 'MicroRNAs: genomics, biogenesis, mechanism, and function', *Cell,* 116 (2), 281-97.

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.
Begg, C. B. (2002), 'On the use of familial aggregation in population-based case probands for calculating penetrance', *J Natl Cancer Inst,* 94 (16), 1221-6.
Bellanne-Chantelot, C., et al. (2008), 'The type and the position of HNF1A mutation modulate age at diagnosis of diabetes in patients with maturity-onset diabetes of the young (MODY)-3', *Diabetes,* 57 (2), 503-8.
Bentley, A. R., et al. (2014), 'Gene-based sequencing identifies lipid-influencing variants with ethnicity-specific effects in African Americans', *PLoS Genet,* 10 (3), e1004190.
Bergmann, A., et al. (2008), 'The A98V single nucleotide polymorphism (SNP) in hepatic nuclear factor 1 alpha (HNF-1alpha) is associated with insulin sensitivity and beta-cell function', *Exp Clin Endocrinol Diabetes,* 116 Suppl 1, S50-5.
Besser, R. E., et al. (2012), 'Using highly sensitive C-reactive protein measurement to diagnose MODY in a family with suspected type 2 diabetes', *BMJ Case Rep,* 2012.
Bjorkhaug, L., et al. (2005), 'Functional dissection of the HNF-1alpha transcription factor: a study on nuclear localization and transcriptional activation', *DNA Cell Biol,* 24 (11), 661-9.
Bjorkhaug, L., et al. (2000), 'MODY associated with two novel hepatocyte nuclear factor-1alpha loss-of-function mutations (P112L and Q466X)', *Biochem Biophys Res Commun,* 279 (3), 792-8.
Bjorkhaug, L., et al. (2003), 'Hepatocyte nuclear factor-1 alpha gene mutations and diabetes in Norway', *J Clin Endocrinol Metab,* 88 (2), 920-31.
Black, S., Kushner, I., and Samols, D. (2004), 'C-reactive Protein', *J Biol Chem,* 279 (47), 48487-90.
Blumenfeld, M., et al. (1991), 'Hepatic nuclear factor 1 (HNF1) shows a wider distribution than products of its known target genes in developing mouse', *Development,* 113 (2), 589-99.
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.
Boj, S. F., et al. (2001), 'A transcription factor regulatory circuit in differentiated pancreatic cells', *Proc Natl Acad Sci U S A,* 98 (25), 14481-6.
Bonnefond, A., et al. (2012a), 'Whole-exome sequencing and high throughput genotyping identified KCNJ11 as the thirteenth MODY gene', *PLoS One,* 7 (6), e37423.
Bonnefond, A., et al. (2012b), 'Rare MTNR1B variants impairing melatonin receptor 1B function contribute to type 2 diabetes', *Nat Genet,* 44 (3), 297-301.
Bonner, C., et al. (2013), 'Identification of circulating microRNAs in HNF1A-MODY carriers', *Diabetologia,* 56 (8), 1743-51.
Borowiec, M., et al. (2009), 'Mutations at the BLK locus linked to maturity onset diabetes of the young and beta-cell dysfunction', *Proc Natl Acad Sci U S A,* 106 (34), 14460-5.
Bruun, C., et al. (2013), 'Amputations and foot ulcers in patients newly diagnosed with type 2 diabetes mellitus and observed for 19 years. The role of age, gender and co-morbidity', *Diabet Med,* 30 (8), 964-72.
Cariou, B., et al. (2004), 'Increased adipose tissue expression of Grb14 in several models of insulin resistance', *FASEB J,* 18 (9), 965-7.
Catli, G., et al. (2013), 'A novel GATA6 mutation leading to congenital heart defects and permanent neonatal diabetes: a case report', *Diabetes Metab,* 39 (4), 370-4.

Cervin, C., et al. (2010), 'An investigation of serum concentration of apoM as a potential MODY3 marker using a novel ELISA', *J Intern Med,* 267 (3), 316-21.
Chambers, J. C., et al. (2001), 'C-reactive protein, insulin resistance, central obesity, and coronary heart disease risk in Indian Asians from the United Kingdom compared with European whites', *Circulation,* 104 (2), 145-50.
Chambers, J. C., et al. (2014), 'The South asian genome', *PLoS One,* 9 (8), e102645.
Chandalia, M., et al. (2007), 'Insulin resistance and body fat distribution in South Asian men compared to Caucasian men', *PLoS One,* 2 (8), e812.
Chaubey, G., et al. (2007), 'Peopling of South Asia: investigating the caste-tribe continuum in India', *Bioessays,* 29 (1), 91-100.
Chauhan, G., et al. (2010), 'Impact of common variants of PPARG, KCNJ11, TCF7L2, SLC30A8, HHEX, CDKN2A, IGF2BP2, and CDKAL1 on the risk of type 2 diabetes in 5,164 Indians', *Diabetes,* 59 (8), 2068-74.
Chidambaram, M., Radha, V., and Mohan, V. (2010), 'Replication of recently described type 2 diabetes gene variants in a South Indian population', *Metabolism,* 59 (12), 1760-6.
Chimienti, F., et al. (2004), 'Identification and cloning of a beta-cell-specific zinc transporter, ZnT-8, localized into insulin secretory granules', *Diabetes,* 53 (9), 2330-7.
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.
Clarke, P., et al. (2003), 'The impact of diabetes-related complications on healthcare costs: results from the United Kingdom Prospective Diabetes Study (UKPDS Study No. 65)', *Diabet Med,* 20 (6), 442-50.
Colclough, K., et al. (2013), 'Mutations in the genes encoding the transcription factors hepatocyte nuclear factor 1 alpha and 4 alpha in maturity-onset diabetes of the young and hyperinsulinemic hypoglycemia', *Hum Mutat,* 34 (5), 669-85.
Consortium, HapMap (2005), 'A haplotype map of the human genome', *Nature,* 437 (7063), 1299-320.
Consortium, The Wellcome Trust Case Control (2007), 'Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls', *Nature,* 447 (7145), 661-78.
Cooney, G. J., et al. (2004), 'Improved glucose homeostasis and enhanced insulin signalling in Grb14-deficient mice', *EMBO J,* 23 (3), 582-93.
Cutting, G. R. (2014), 'Annotating DNA variants is the next major goal for human genetics', *Am J Hum Genet,* 94 (1), 5-10.
Dall'Olio, F. (2000), 'The sialyl-alpha2,6-lactosaminyl-structure: biosynthesis and functional role', *Glycoconj J,* 17 (10), 669-76.
De Franco, E., et al. (2013), 'GATA6 mutations cause a broad phenotypic spectrum of diabetes from pancreatic agenesis to adult-onset diabetes without exocrine insufficiency', *Diabetes,* 62 (3), 993-7.
Depetris, R. S., et al. (2005), 'Structural basis for inhibition of the insulin receptor by the adaptor protein Grb14', *Mol Cell,* 20 (2), 325-33.
Dhawan, J., et al. (1994), 'Insulin resistance, high prevalence of diabetes, and cardiovascular risk in immigrant Asians. Genetic or environmental effect?', *Br Heart J,* 72 (5), 413-21.
Diabetes UK 'DIABETES: FACTS AND STATS', <http://www.diabetes.org.uk/Documents/About%20Us/Statistics/Diabetes-key-stats-guidelines-April2014.pdf>, accessed.
Diamond, J. (2003), 'The double puzzle of diabetes', *Nature,* 423 (6940), 599-602.
Dickson, S. P., et al. (2010), 'Rare variants create synthetic genome-wide associations', *PLoS Biol,* 8 (1), e1000294.

Dimas, A. S., et al. (2014), 'Impact of type 2 diabetes susceptibility variants on quantitative glycemic traits reveals mechanistic heterogeneity', *Diabetes,* 63 (6), 2158-71.
Dinneen, S., Gerich, J., and Rizza, R. (1992), 'Carbohydrate metabolism in non-insulin-dependent diabetes mellitus', *N Engl J Med,* 327 (10), 707-13.
Dukes, I. D., et al. (1998), 'Defective pancreatic beta-cell glycolytic signaling in hepatocyte nuclear factor-1alpha-deficient mice', *J Biol Chem,* 273 (38), 24457-64.
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.
Edghill, E. L., et al. (2008), 'Insulin mutation screening in 1,044 patients with diabetes: mutations in the INS gene are a common cause of neonatal diabetes but a rare cause of diabetes diagnosed in childhood or adulthood', *Diabetes,* 57 (4), 1034-42.
Ellard, S., Bellanne-Chantelot, C., and Hattersley, A. T. (2008), 'Best practice guidelines for the molecular genetic diagnosis of maturity-onset diabetes of the young', *Diabetologia,* 51 (4), 546-53.
ESP 'Exome Variant Server, NHLBI GO Exome Sequencing Project (ESP), Seattle, WA', <http://evs.gs.washington.edu/EVS/>, accessed.
Estrada, K., et al. (2014), 'Association of a low-frequency variant in HNF1A with type 2 diabetes in a Latino population', *Jama,* 311 (22), 2305-14.
Etheridge, A., et al. (2011), 'Extracellular microRNA: a new source of biomarkers', *Mutat Res,* 717 (1-2), 85-90.
ExAc 'Exome Aggregation Consortium, Cambridge, MA', <http://exac.broadinstitute.org/>., accessed.
Fajans, S. S., Bell, G. I., and Polonsky, K. S. (2001), 'Molecular mechanisms and clinical pathophysiology of maturity-onset diabetes of the young', *N Engl J Med,* 345 (13), 971-80.
Ferrer, J. (2002), 'A genetic switch in pancreatic beta-cells: implications for differentiation and haploinsufficiency', *Diabetes,* 51 (8), 2355-62.
Flanagan, S. E., Patch, A. M., and Ellard, S. (2010), 'Using SIFT and PolyPhen to predict loss-of-function and gain-of-function mutations', *Genet Test Mol Biomarkers,* 14 (4), 533-7.
Flannick, J., et al. (2013), 'Assessing the phenotypic effects in the general population of rare variants in genes for a dominant Mendelian form of diabetes', 45 (11), 1380-5.
Forouhi, N. G., Sattar, N., and McKeigue, P. M. (2001), 'Relation of C-reactive protein to body fat distribution and features of the metabolic syndrome in Europeans and South Asians', *Int J Obes Relat Metab Disord,* 25 (9), 1327-31.
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.
Frazer, K. A., et al. (2007), 'A second generation human haplotype map of over 3.1 million SNPs', *Nature,* 449 (7164), 851-61.
Froguel, P., et al. (1993), 'Familial hyperglycemia due to mutations in glucokinase. Definition of a subtype of diabetes mellitus', *N Engl J Med,* 328 (10), 697-702.
Galan, M., et al. (2011), 'Differential effects of HNF-1alpha mutations associated with familial young-onset diabetes on target gene regulation', *Mol Med,* 17 (3-4), 256-65.
Genomes, 1000 <http://browser.1000genomes.org/index.html>, accessed.
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. (2012), 'Metabolic profiling in Maturity-onset diabetes of the young (MODY) and young onset type 2 diabetes fails to detect robust urinary biomarkers', *PLoS One,* 7 (7), e40962.
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.
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.
Goldstein, D. B. (2009), 'Common genetic variation and human traits', *N Engl J Med,* 360 (17), 1696-8.
Gonzalez-Perez, A. and Lopez-Bigas, N. (2011), 'Improving the assessment of the outcome of nonsynonymous SNVs with a consensus deleteriousness score, Condel', *Am J Hum Genet,* 88 (4), 440-9.
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.
Gujral, U. P., et al. (2013), 'Type 2 diabetes in South Asians: similarities and differences with white Caucasian and other populations', *Ann N Y Acad Sci,* 1281, 51-63.
Haiman, C. A., et al. (2012), 'Consistent directions of effect for established type 2 diabetes risk variants across populations: the population architecture using Genomics and Epidemiology (PAGE) Consortium', *Diabetes,* 61 (6), 1642-7.
Hales, C. N. and Barker, D. J. P. (1992), 'Type 2 (non-insulin-dependent) diabetes mellitus: the thrifty phenotype hypothesis', *Diabetologia,* 35 (7), 595-601.
Hansen, S. K., et al. (2002), 'Genetic evidence that HNF-1alpha-dependent transcriptional control of HNF-4alpha is essential for human pancreatic beta cell function', *J Clin Invest,* 110 (6), 827-33.
Hansson, G. K. (2005), 'Inflammation, atherosclerosis, and coronary artery disease', *N Engl J Med,* 352 (16), 1685-95.
Harries, L. W., Hattersley, A. T., and Ellard, S. (2004), 'Messenger RNA transcripts of the hepatocyte nuclear factor-1alpha gene containing premature termination codons are subject to nonsense-mediated decay', *Diabetes,* 53 (2), 500-4.
Harries, L. W., Brown, J. E., and Gloyn, A. L. (2009), 'Species-specific differences in the expression of the HNF1A, HNF1B and HNF4A genes', *PLoS One,* 4 (11), e7855.
Harries, L. W., et al. (2006), 'Isomers of the TCF1 gene encoding hepatocyte nuclear factor-1 alpha show differential expression in the pancreas and define the relationship between mutation position and clinical phenotype in monogenic diabetes', *Hum Mol Genet,* 15 (14), 2216-24.
Harries, L. W., et al. (2008), 'Diabetes susceptibility in the Canadian Oji-Cree population is moderated by abnormal mRNA processing of HNF1A G319S transcripts', *Diabetes,* 57 (7), 1978-82.
Hattersley, A. T. (1998), 'Maturity-onset diabetes of the young: clinical heterogeneity explained by genetic heterogeneity', *Diabet Med,* 15 (1), 15-24.
Hattersley, A. T., et al. (1992), 'Linkage of type 2 diabetes to the glucokinase gene', *Lancet,* 339 (8805), 1307-10.
Hegele, R. A., et al. (1999a), 'Hepatocyte nuclear factor-1 alpha G319S. A private mutation in Oji-Cree associated with type 2 diabetes', *Diabetes Care,* 22 (3), 524.
--- (1999b), 'The hepatic nuclear factor-1alpha G319S variant is associated with early-onset type 2 diabetes in Canadian Oji-Cree', *J Clin Endocrinol Metab,* 84 (3), 1077-82.

Heid, I. M., 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 (11), 949-60.
Holmkvist, J., et al. (2006), 'Common variants in HNF-1 alpha and risk of type 2 diabetes', *Diabetologia,* 49 (12), 2882-91.
Holt, L. J., et al. (2009), 'Dual ablation of Grb10 and Grb14 in mice reveals their combined role in regulation of insulin signaling and glucose homeostasis', *Mol Endocrinol,* 23 (9), 1406-14.
Horikawa, Y., et al. (1997), 'Mutation in hepatocyte nuclear factor-1 beta gene (TCF2) associated with MODY', *Nat Genet,* 17 (4), 384-5.
Hotamisligil, G. S. (2000), 'Molecular mechanisms of insulin resistance and the role of the adipocyte', *Int J Obes Relat Metab Disord,* 24 Suppl 4, S23-7.
http://biostats.usc.edu/Quanto.html 'Quanto'.
http://pngu.mgh.harvard.edu/~purcell/plink/ 'PLINK'.
Huyghe, J. R., et al. (2013), 'Exome array analysis identifies new loci and low-frequency variants influencing insulin processing and secretion', *Nat Genet,* 45 (2), 197-201.
IDF (2013), *International Diabetes Federation. Diabetes Atlas, 6th edn. Brussels, Belgium: International Diabetes Federation*.
Imperiali, B. and O'Connor, S. E. (1999), 'Effect of N-linked glycosylation on glycopeptide and glycoprotein structure', *Curr Opin Chem Biol,* 3 (6), 643-9.
Jayasekara, R. S. and Schultz, T. (2007), 'Health status, trends, and issues in Sri Lanka', *Nurs Health Sci,* 9 (3), 228-33.
Jayawardena, R., et al. (2012), 'Prevalence and trends of the diabetes epidemic in South Asia: a systematic review and meta-analysis', *BMC Public Health,* 12, 380.
Jayawardena, R., et al. (2014), 'Energy and nutrient intakes among Sri Lankan adults', *Int Arch Med,* 7, 34.
Johansen, C. T., et al. (2010), 'Excess of rare variants in genes identified by genome-wide association study of hypertriglyceridemia', *Nat Genet,* 42 (8), 684-7.
Jones, H. W., Jr. (1997), 'Record of the first physician to see Henrietta Lacks at the Johns Hopkins Hospital: history of the beginning of the HeLa cell line', *Am J Obstet Gynecol,* 176 (6), S227-8.
Kahn, S. E., Hull, R. L., and Utzschneider, K. M. (2006), 'Mechanisms linking obesity to insulin resistance and type 2 diabetes', *Nature,* 444 (7121), 840-6.
Kalra, S., Mercuri, M., and Anand, S. S. (2013), 'Measures of body fat in South Asian adults', *Nutr Diabetes,* 3, e69.
Katulanda, P., et al. (2010), 'Prevalence of overweight and obesity in Sri Lankan adults', *Obes Rev,* 11 (11), 751-6.
Katulanda, P., et al. (2012), 'Metabolic syndrome among Sri Lankan adults: prevalence, patterns and correlates', *Diabetol Metab Syndr,* 4 (1), 24.
Katulanda, P., et al. (2014), 'The prevalence, predictors and associations of hypertension in Sri Lanka: a cross-sectional population based national survey', *Clin Exp Hypertens,* 36 (7), 484-91.
Katulanda, P., et al. (2008a), 'Prevalence and projections of diabetes and pre-diabetes in adults in Sri Lanka--Sri Lanka Diabetes, Cardiovascular Study (SLDCS)', *Diabet Med,* 25 (9), 1062-9.
Katulanda, P., et al. (2008b), 'Diabetes mellitus among young adults in Sri Lanka--role of GAD antibodies in classification and treatment: the Sri Lanka Young Diabetes study', *Diabetologia,* 51 (8), 1368-74.
Keinan, A. and Clark, A. G. (2012), 'Recent explosive human population growth has resulted in an excess of rare genetic variants', *Science,* 336 (6082), 740-3.

Kiezun, A., et al. (2012), 'Exome sequencing and the genetic basis of complex traits', *Nat Genet,* 44 (6), 623-30.
Kim, I., et al. (2011), 'A low-risk ZnT-8 allele (W325) for post-transplantation diabetes mellitus is protective against cyclosporin A-induced impairment of insulin secretion', *Pharmacogenomics J,* 11 (3), 191-8.
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.
Kraft, P. (2008), 'Curses--winner's and otherwise--in genetic epidemiology', *Epidemiology,* 19 (5), 649-51; discussion 57-8.
Kraja, A. T., et al. (2014), 'Pleiotropic genes for metabolic syndrome and inflammation', *Mol Genet Metab,* 112 (4), 317-38.
Kristinsson, S. Y., et al. (2001), 'MODY in Iceland is associated with mutations in HNF-1alpha and a novel mutation in NeuroD1', *Diabetologia,* 44 (11), 2098-103.
Kshatriya, G. K. (1995), 'Genetic affinities of Sri Lankan populations', *Hum Biol,* 67 (6), 843-66.
Ku, C. S., Naidoo, N., and Pawitan, Y. (2011), 'Revisiting Mendelian disorders through exome sequencing', *Hum Genet,* 129 (4), 351-70.
Kumar, P., Henikoff, S., and Ng, P. C. (2009), 'Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm', *Nat Protoc,* 4 (7), 1073-81.
Lango Allen, H., et al. (2012), 'GATA6 haploinsufficiency causes pancreatic agenesis in humans', *Nat Genet,* 44 (1), 20-2.
Lauc, G., et al. (2010), 'Genomics meets glycomics-the first GWAS study of human N-Glycome identifies HNF1alpha as a master regulator of plasma protein fucosylation', *PLoS Genet,* 6 (12), e1001256.
Lee, Y. H., Sauer, B., and Gonzalez, F. J. (1998), 'Laron dwarfism and non-insulin-dependent diabetes mellitus in the Hnf-1alpha knockout mouse', *Mol Cell Biol,* 18 (5), 3059-68.
Lees, M. J. and Whitelaw, M. L. (1999), 'Multiple roles of ligand in transforming the dioxin receptor to an active basic helix-loop-helix/PAS transcription factor complex with the nuclear protein Arnt', *Mol Cell Biol,* 19 (8), 5811-22.
Leon, D. A. (2008), 'Cities, urbanization and health', *Int J Epidemiol,* 37 (1), 4-8.
Lichtsteiner, S. and Schibler, U. (1989), 'A glycosylated liver-specific transcription factor stimulates transcription of the albumin gene', *Cell,* 57 (7), 1179-87.
Lihn, A. S., Pedersen, S. B., and Richelsen, B. (2005), 'Adiponectin: action, regulation and association to insulin sensitivity', *Obes Rev,* 6 (1), 13-21.
Lindstrom, J., et al. (2006), 'Sustained reduction in the incidence of type 2 diabetes by lifestyle intervention: follow-up of the Finnish Diabetes Prevention Study', *Lancet,* 368 (9548), 1673-9.
Lucey, B. P., Nelson-Rees, W. A., and Hutchins, G. M. (2009), 'Henrietta Lacks, HeLa cells, and cell culture contamination', *Arch Pathol Lab Med,* 133 (9), 1463-7.
Lupski, J. R., et al. (2011), 'Clan genomics and the complex architecture of human disease', *Cell,* 147 (1), 32-43.
Lyssenko, V., et al. (2009), 'Common variant in MTNR1B associated with increased risk of type 2 diabetes and impaired early insulin secretion', *Nat Genet,* 41 (1), 82-8.
Ma, B., Simala-Grant, J. L., and Taylor, D. E. (2006), 'Fucosylation in prokaryotes and eukaryotes', *Glycobiology,* 16 (12), 158r-84r.
Mahajan, A., et al. (2011), 'Common variants in CRP and LEPR influence high sensitivity C-reactive protein levels in North Indians', *PLoS One,* 6 (9), e24645.

Mahajan, A., et al. (2014), 'Genome-wide trans-ancestry meta-analysis provides insight into the genetic architecture of type 2 diabetes susceptibility', *Nat Genet,* 46 (3), 234-44.
Majithia, A. R., et al. (2014), 'Rare variants in PPARG with decreased activity in adipocyte differentiation are associated with increased risk of type 2 diabetes', *Proc Natl Acad Sci U S A,* 111 (36), 13127-32.
Malecki, M. T., et al. (1999), 'Mutations in NEUROD1 are associated with the development of type 2 diabetes mellitus', *Nat Genet,* 23 (3), 323-8.
Manolio, T. A., et al. (2009), 'Finding the missing heritability of complex diseases', *Nature,* 461 (7265), 747-53.
Marchini, J., et al. (2007), 'A new multipoint method for genome-wide association studies by imputation of genotypes', *Nat Genet,* 39 (7), 906-13.
Marigorta, U. M. and Navarro, A. (2013), 'High trans-ethnic replicability of GWAS results implies common causal variants', *PLoS Genet,* 9 (6), e1003566.
Marino, K., et al. (2010), 'A systematic approach to protein glycosylation analysis: a path through the maze', *Nat Chem Biol,* 6 (10), 713-23.
Matschinsky, F., et al. (1993), 'Glucokinase as pancreatic beta cell glucose sensor and diabetes gene', *J Clin Invest,* 92 (5), 2092-8.
McCarthy, M. I. (2010), 'Genomics, type 2 diabetes, and obesity', *N Engl J Med,* 363 (24), 2339-50.
--- (2011), 'The importance of global studies of the genetics of type 2 diabetes', *Diabetes Metab J,* 35 (2), 91-100.
McCarthy, M. I., et al. (2008), 'Genome-wide association studies for complex traits: consensus, uncertainty and challenges', *Nat Rev Genet,* 9 (5), 356-69.
McDonald, T. J., et al. (2011), 'High-sensitivity CRP discriminates HNF1A-MODY from other subtypes of diabetes', *Diabetes Care,* 34 (8), 1860-2.
McKeigue, P. M., Shah, B., and Marmot, M. G. (1991), 'Relation of central obesity and insulin resistance with high diabetes prevalence and cardiovascular risk in South Asians', *Lancet,* 337 (8738), 382-6.
Medici, F., et al. (1999), 'Concordance rate for type II diabetes mellitus in monozygotic twins: actuarial analysis', *Diabetologia,* 42 (2), 146-50.
Meigs, J. B., Cupples, L. A., and Wilson, P. W. (2000), 'Parental transmission of type 2 diabetes: the Framingham Offspring Study', *Diabetes,* 49 (12), 2201-7.
Mendel, D. B. and Crabtree, G. R. (1991), 'HNF-1, a member of a novel class of dimerizing homeodomain proteins', *J Biol Chem,* 266 (2), 677-80.
Misra, A. and Shrivastava, U. (2013), 'Obesity and dyslipidemia in South Asians', *Nutrients,* 5 (7), 2708-33.
Misra, A., et al. (2014), 'Diabetes in South asians', *Diabet Med,* 31 (10), 1153-62.
Mohan, V., et al. (1986), 'Serum immunoreactive insulin responses to a glucose load in Asian Indian and European type 2 (non-insulin-dependent) diabetic patients and control subjects', *Diabetologia,* 29 (4), 235-7.
Molven, A. and Njolstad, P. R. (2011), 'Role of molecular genetics in transforming diagnosis of diabetes mellitus', *Expert Rev Mol Diagn,* 11 (3), 313-20.
Moorjani, P., et al. (2013), 'Genetic evidence for recent population mixture in India', *Am J Hum Genet,* 93 (3), 422-38.
Morris, A. P. (2011), 'Transethnic meta-analysis of genomewide association studies', *Genet Epidemiol,* 35 (8), 809-22.
--- (2014), 'Fine mapping of type 2 diabetes susceptibility loci', *Curr Diab Rep,* 14 (11), 549.
Morris, A. P., et al. (2012), 'Large-scale association analysis provides insights into the genetic architecture and pathophysiology of type 2 diabetes', *Nat Genet,* 44 (9), 981-90.

Mughal, S. A., et al. (2013), 'Apolipoprotein M can discriminate HNF1A-MODY from Type 1 diabetes', *Diabet Med,* 30 (2), 246-50.
Murphy, R., Ellard, S., and Hattersley, A. T. (2008), 'Clinical implications of a molecular genetic classification of monogenic beta-cell diabetes', *Nat Clin Pract Endocrinol Metab,* 4 (4), 200-13.
Nakajima, T., et al. (1999), 'Genomic structure and chromosomal mapping of the human sterol regulatory element binding protein (SREBP) cleavage-activating protein (SCAP) gene', *J Hum Genet,* 44 (6), 402-7.
Nammo, T., et al. (2008), 'Expression of HNF-4alpha (MODY1), HNF-1beta (MODY5), and HNF-1alpha (MODY3) proteins in the developing mouse pancreas', *Gene Expr Patterns,* 8 (2), 96-106.
Neel, J. V. (1962), 'Diabetes mellitus: a "thrifty" genotype rendered detrimental by "progress"?', *Am J Hum Genet,* 14, 353-62.
Nejentsev, S., et al. (2009), 'Rare variants of IFIH1, a gene implicated in antiviral responses, protect against type 1 diabetes', *Science,* 324 (5925), 387-9.
Nelson, M. R., et al. (2012), 'An abundance of rare functional variants in 202 drug target genes sequenced in 14,002 people', *Science,* 337 (6090), 100-4.
Neve, B., et al. (2005), 'Role of transcription factor KLF11 and its diabetes-associated gene variants in pancreatic beta cell function', *Proc Natl Acad Sci U S A,* 102 (13), 4807-12.
Newman, B., et al. (1987), 'Concordance for type 2 (non-insulin-dependent) diabetes mellitus in male twins', *Diabetologia,* 30 (10), 763-8.
Ng, P. C. and Henikoff, S. (2001), 'Predicting deleterious amino acid substitutions', *Genome Res,* 11 (5), 863-74.
--- (2003), 'SIFT: Predicting amino acid changes that affect protein function', *Nucleic Acids Res,* 31 (13), 3812-4.
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.
Nishikawa, T., et al. (2008), 'Transcriptional complex formation of c-Fos, STAT3, and hepatocyte NF-1 alpha is essential for cytokine-driven C-reactive protein gene expression', *J Immunol,* 180 (5), 3492-501.
Njolstad, P. R., et al. (2001), 'Neonatal diabetes mellitus due to complete glucokinase deficiency', *N Engl J Med,* 344 (21), 1588-92.
Nolan, C. J., Damm, P., and Prentki, M. (2011), 'Type 2 diabetes across generations: from pathophysiology to prevention and management', *Lancet,* 378 (9786), 169-81.
Nothnagel, M., et al. (2009), 'A comprehensive evaluation of SNP genotype imputation', *Hum Genet,* 125 (2), 163-71.
Nowak, N., et al. (2013), 'Cystatin C is not a good candidate biomarker for HNF1A-MODY', *Acta Diabetol,* 50 (5), 815-20.
Okada, Y., et al. (2011), 'Genome-wide association study for C-reactive protein levels identified pleiotropic associations in the IL6 locus', *Hum Mol Genet,* 20 (6), 1224-31.
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.
Ormond, K. E. (2008), 'Medical ethics for the genome world: a paper from the 2007 William Beaumont hospital symposium on molecular pathology', *J Mol Diagn,* 10 (5), 377-82.

Orozco, G., Barrett, J. C., and Zeggini, E. (2010), 'Synthetic associations in the context of genome-wide association scan signals', *Hum Mol Genet,* 19 (R2), R137-44.
Owen, K. and Hattersley, A. T. (2001), 'Maturity-onset diabetes of the young: from clinical description to molecular genetic characterization', *Best Pract Res Clin Endocrinol Metab,* 15 (3), 309-23.
Owen, K. R., et 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 (9), 1919-24.
Palmer, N. D., et al. (2012), 'A genome-wide association search for type 2 diabetes genes in African Americans', *PLoS One,* 7 (1), e29202.
Pannacciulli, N., et al. (2001), 'C-reactive protein is independently associated with total body fat, central fat, and insulin resistance in adult women', *Int J Obes Relat Metab Disord,* 25 (10), 1416-20.
Papiha, S. S., Mastana, S. S., and Jayasekara, R. (1996), 'Genetic variation in Sri Lanka', *Hum Biol,* 68 (5), 707-37.
Pearson, E. R., et al. (2003), 'Genetic cause of hyperglycaemia and response to treatment in diabetes', *Lancet,* 362 (9392), 1275-81.
Pearson, T. A. and Manolio, T. A. (2008), 'How to interpret a genome-wide association study', *Jama,* 299 (11), 1335-44.
Plengvidhya, N., et al. (2007), 'PAX4 mutations in Thais with maturity onset diabetes of the young', *J Clin Endocrinol Metab,* 92 (7), 2821-6.
Plomgaard, P., et al. (2005), 'Tumor necrosis factor-alpha induces skeletal muscle insulin resistance in healthy human subjects via inhibition of Akt substrate 160 phosphorylation', *Diabetes,* 54 (10), 2939-45.
Pontoglio, M., et al. (1996), 'Hepatocyte nuclear factor 1 inactivation results in hepatic dysfunction, phenylketonuria, and renal Fanconi syndrome', *Cell,* 84 (4), 575-85.
Pontoglio, M., et al. (1998), 'Defective insulin secretion in hepatocyte nuclear factor 1alpha-deficient mice', *J Clin Invest,* 101 (10), 2215-22.
Popkin, B. M. (2006), 'Global nutrition dynamics: the world is shifting rapidly toward a diet linked with noncommunicable diseases', *Am J Clin Nutr,* 84 (2), 289-98.
Porter, J. R., et al. (2006), 'Asian MODY: are we missing an important diagnosis?', *Diabet Med,* 23 (11), 1257-60.
Poulsen, P., et al. (1999), 'Heritability of type II (non-insulin-dependent) diabetes mellitus and abnormal glucose tolerance--a population-based twin study', *Diabetologia,* 42 (2), 139-45.
Prasad, R. B. and Groop, L. (2015), 'Genetics of type 2 diabetes-pitfalls and possibilities', *Genes (Basel),* 6 (1), 87-123.
Prokopenko, I., et al. (2009), 'Variants in MTNR1B influence fasting glucose levels', *Nat Genet,* 41 (1), 77-81.
Pubudu De Silva, A., et al. (2012), 'Social, cultural and economical determinants of diabetes mellitus in Kalutara district, Sri Lanka: a cross sectional descriptive study', *Int J Equity Health,* 11, 76.
Purcell, S., et al. (2007), 'PLINK: a tool set for whole-genome association and population-based linkage analyses', *Am J Hum Genet,* 81 (3), 559-75.
Radha, V., et al. (2009), 'Identification of novel variants in the hepatocyte nuclear factor-1alpha gene in South Indian patients with maturity onset diabetes of young', *J Clin Endocrinol Metab,* 94 (6), 1959-65.
Raeder, H., et al. (2006), 'Mutations in the CEL VNTR cause a syndrome of diabetes and pancreatic exocrine dysfunction', *Nat Genet,* 38 (1), 54-62.
Rahmioglu, N., et al. (2015), 'Genome-wide enrichment analysis between endometriosis and obesity-related traits reveals novel susceptibility loci', *Hum Mol Genet,* 24 (4), 1185-99.

Raji, A., et al. (2001), 'Body fat distribution and insulin resistance in healthy Asian Indians and Caucasians', *J Clin Endocrinol Metab,* 86 (11), 5366-71.
Ramachandran, A., Snehalatha, C., and Ma, R. C. (2014), 'Diabetes in South-East Asia: an update', *Diabetes Res Clin Pract,* 103 (2), 231-7.
Ramachandran, A., et al. (1997), 'Risk of noninsulin dependent diabetes mellitus conferred by obesity and central adiposity in different ethnic groups: a comparative analysis between Asian Indians, Mexican Americans and Whites', *Diabetes Res Clin Pract,* 36 (2), 121-5.
Ramachandran, A., et al. (2006), 'The Indian Diabetes Prevention Programme shows that lifestyle modification and metformin prevent type 2 diabetes in Asian Indian subjects with impaired glucose tolerance (IDPP-1)', *Diabetologia,* 49 (2), 289-97.
Ramachandran, A., et al. (2004), 'Temporal changes in prevalence of diabetes and impaired glucose tolerance associated with lifestyle transition occurring in the rural population in India', *Diabetologia,* 47 (5), 860-5.
Ramachandran, Ambady, Wan Ma, Ronald Ching, and Snehalatha, Chamukuttan (2010), 'Diabetes in Asia', *The Lancet,* 375 (9712), 408-18.
Rees, M. G., et al. (2012), 'Cellular characterisation of the GCKR P446L variant associated with type 2 diabetes risk', *Diabetologia,* 55 (1), 114-22.
Rees, S. D., et al. (2011), 'Replication of 13 genome-wide association (GWA)-validated risk variants for type 2 diabetes in Pakistani populations', *Diabetologia,* 54 (6), 1368-74.
Reich, D., et al. (2009), 'Reconstructing Indian population history', *Nature,* 461 (7263), 489-94.
Reiner, A. P., et al. (2008), 'Polymorphisms of the HNF1A gene encoding hepatocyte nuclear factor-1 alpha are associated with C-reactive protein', *Am J Hum Genet,* 82 (5), 1193-201.
Ridker, P. M. (2001), 'High-sensitivity C-reactive protein: potential adjunct for global risk assessment in the primary prevention of cardiovascular disease', *Circulation,* 103 (13), 1813-8.
Ridker, P. M., et al. (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 (5), 1185-92.
Roglic, G. and Unwin, N. (2010), 'Mortality attributable to diabetes: estimates for the year 2010', *Diabetes Res Clin Pract,* 87 (1), 15-9.
Rosenberg, N. A., et al. (2010), 'Genome-wide association studies in diverse populations', *Nat Rev Genet,* 11 (5), 356-66.
Rosenblum, B. B., et al. (1997), 'New dye-labeled terminators for improved DNA sequencing patterns', *Nucleic Acids Res,* 25 (22), 4500-4.
Ryffel, G. U. (2001), 'Mutations in the human genes encoding the transcription factors of the hepatocyte nuclear factor (HNF)1 and HNF4 families: functional and pathological consequences', *J Mol Endocrinol,* 27 (1), 11-29.
Saah, A. J. and Hoover, D. R. (1997), '"Sensitivity" and "specificity" reconsidered: the meaning of these terms in analytical and diagnostic settings', *Ann Intern Med,* 126 (1), 91-4.
Sahu, R. P., et al. (2007), 'Etiology of early-onset type 2 diabetes in Indians: islet autoimmunity and mutations in hepatocyte nuclear factor 1alpha and mitochondrial gene', *J Clin Endocrinol Metab,* 92 (7), 2462-7.
Sandhu, M. S., et al. (2007), 'Common variants in WFS1 confer risk of type 2 diabetes', *Nat Genet,* 39 (8), 951-3.
Sanger, F. and Coulson, A. R. (1975), 'A rapid method for determining sequences in DNA by primed synthesis with DNA polymerase', *J Mol Biol,* 94 (3), 441-8.

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.
Sanghera, D. K. and Blackett, P. R. (2012), 'Type 2 Diabetes Genetics: Beyond GWAS', *J Diabetes Metab,* 3 (198).
Sanghera, D. K., et al. (2008), 'Impact of nine common type 2 diabetes risk polymorphisms in Asian Indian Sikhs: PPARG2 (Pro12Ala), IGF2BP2, TCF7L2 and FTO variants confer a significant risk', *BMC Med Genet,* 9, 59.
Sarwar, N., et al. (2010), 'Diabetes mellitus, fasting blood glucose concentration, and risk of vascular disease: a collaborative meta-analysis of 102 prospective studies', *Lancet,* 375 (9733), 2215-22.
Saxena, R., et al. (2013), 'Genome-wide association study identifies a novel locus contributing to type 2 diabetes susceptibility in Sikhs of Punjabi origin from India', *Diabetes,* 62 (5), 1746-55.
Saxena, R., et al. (2007), 'Genome-wide association analysis identifies loci for type 2 diabetes and triglyceride levels', *Science,* 316 (5829), 1331-6.
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), 410-25.
Schwitzgebel, V. M. (2014), 'Many faces of monogenic diabetes', *J Diabetes Investig,* 5 (2), 121-33.
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.
Servitja, J. M., et al. (2009), 'Hnf1alpha (MODY3) controls tissue-specific transcriptional programs and exerts opposed effects on cell growth in pancreatic islets and liver', *Mol Cell Biol,* 29 (11), 2945-59.
Shaat, N., et al. (2006), 'Common variants in MODY genes increase the risk of gestational diabetes mellitus', *Diabetologia,* 49 (7), 1545-51.
Shah, A. and Kanaya, A. M. (2014), 'Diabetes and associated complications in the South Asian population', *Curr Cardiol Rep,* 16 (5), 476.
Shepherd, M., et al. (2009), 'A genetic diagnosis of HNF1A diabetes alters treatment and improves glycaemic control in the majority of insulin-treated patients', *Diabet Med,* 26 (4), 437-41.
Shields, B. M., et al. (2010), 'Maturity-onset diabetes of the young (MODY): how many cases are we missing?', *Diabetologia,* 53 (12), 2504-8.
Shih, D. Q., et al. (2001), 'Loss of HNF-1alpha function in mice leads to abnormal expression of genes involved in pancreatic islet development and metabolism', *Diabetes,* 50 (11), 2472-80.
Skol, A. D., et al. (2006), 'Joint analysis is more efficient than replication-based analysis for two-stage genome-wide association studies', *Nat Genet,* 38 (2), 209-13.
Sladek, R., et al. (2007), 'A genome-wide association study identifies novel risk loci for type 2 diabetes', *Nature,* 445 (7130), 881-5.
Smith, L. M., et al. (1986), 'Fluorescence detection in automated DNA sequence analysis', *Nature,* 321 (6071), 674-9.
Sohani, Z. N., et al. (2014), 'Does genetic heterogeneity account for the divergent risk of type 2 diabetes in South Asian and white European populations?', *Diabetologia,* 57 (11), 2270-81.
Srinath Reddy, K., et al. (2005), 'Responding to the threat of chronic diseases in India', *Lancet,* 366 (9498), 1744-9.
Steinthorsdottir, V., Thorleifsson, G., and Sulem, P. (2014), 'Identification of low-frequency and rare sequence variants associated with elevated or reduced risk of type 2 diabetes', 46 (3), 294-8.
Steinthorsdottir, V., et al. (2007), 'A variant in CDKAL1 influences insulin response and risk of type 2 diabetes', *Nat Genet,* 39 (6), 770-5.

Stepniewski, J., et al. (2015), 'Induced pluripotent stem cells as a model for diabetes investigation', *Sci Rep,* 5, 8597.
Stoffers, D. A., et al. (1997), 'Early-onset type-II diabetes mellitus (MODY4) linked to IPF1', *Nat Genet,* 17 (2), 138-9.
Stratton, I. M., et al. (2000), 'Association of glycaemia with macrovascular and microvascular complications of type 2 diabetes (UKPDS 35): prospective observational study', *BMJ,* 321 (7258), 405-12.
Stride, A., et al. (2014), 'Cross-sectional and longitudinal studies suggest pharmacological treatment used in patients with glucokinase mutations does not alter glycaemia', *Diabetologia,* 57 (1), 54-6.
Strimbu, Kyle and Tavel, Jorge A. (2010), 'What are Biomarkers?', *Current opinion in HIV and AIDS,* 5 (6), 463-66.
Stumvoll, M., Goldstein, B. J., and van Haeften, T. W. (2005), 'Type 2 diabetes: principles of pathogenesis and therapy', *Lancet,* 365 (9467), 1333-46.
Stumvoll, M., et al. (2003), 'Glucose allostasis', *Diabetes,* 52 (4), 903-9.
Tabassum, R., et al. (2013), 'Genome-wide association study for type 2 diabetes in Indians identifies a new susceptibility locus at 2q21', *Diabetes,* 62 (3), 977-86.
Takahashi, K. and Yamanaka, S. (2006), 'Induction of pluripotent stem cells from mouse embryonic and adult fibroblast cultures by defined factors', *Cell,* 126 (4), 663-76.
Tattersall, R. B. and Fajans, S. S. (1975), 'A difference between the inheritance of classical juvenile-onset and maturity-onset type diabetes of young people', *Diabetes,* 24 (1), 44-53.
Tennessen, J. A., et al. (2012), 'Evolution and functional impact of rare coding variation from deep sequencing of human exomes', *Science,* 337 (6090), 64-9.
Teo, A. K., et al. (2013), 'Derivation of human induced pluripotent stem cells from patients with maturity onset diabetes of the young', *J Biol Chem,* 288 (8), 5353-6.
Thanabalasingham, G. and Owen, K. R. (2011), 'Diagnosis and management of maturity onset diabetes of the young (MODY)', *Bmj,* 343, d6044.
Thanabalasingham, G., et al. (2012), 'Systematic assessment of etiology in adults with a clinical diagnosis of young-onset type 2 diabetes is a successful strategy for identifying maturity-onset diabetes of the young', *Diabetes Care,* 35 (6), 1206-12.
Thanabalasingham, G., et al. (2011), 'A large multi-centre European study validates high-sensitivity C-reactive protein (hsCRP) as a clinical biomarker for the diagnosis of diabetes subtypes', *Diabetologia,* 54 (11), 2801-10.
Thanabalasingham, G., et al. (2013), 'Mutations in HNF1A result in marked alterations of plasma glycan profile', *Diabetes,* 62 (4), 1329-37.
Thomas, H., et al. (2002), 'Evidence for haploinsufficiency of the human HNF1alpha gene revealed by functional characterization of MODY3-associated mutations', *Biol Chem,* 383 (11), 1691-700.
Thomas, H., et al. (2001), 'A distant upstream promoter of the HNF-4alpha gene connects the transcription factors involved in maturity-onset diabetes of the young', *Hum Mol Genet,* 10 (19), 2089-97.
Thormaehlen, A. S., et al. (2015), 'Systematic cell-based phenotyping of missense alleles empowers rare variant association studies: a case for LDLR and myocardial infarction', *PLoS Genet,* 11 (2), e1004855.
Tomei, L., Cortese, R., and De Francesco, R. (1992), 'A POU-A related region dictates DNA binding specificity of LFB1/HNF1 by orienting the two XL-homeodomains in the dimer', *Embo j,* 11 (11), 4119-29.

Toniatti, C., et al. (1990a), 'Synergistic trans-activation of the human C-reactive protein promoter by transcription factor HNF-1 binding at two distinct sites', *Embo j,* 9 (13), 4467-75.
Toniatti, C., et al. (1990b), 'Regulation of the human C-reactive protein gene, a major marker of inflammation and cancer', *Mol Biol Med,* 7 (3), 199-212.
Triggs-Raine, B. L., et al. (2002), 'HNF-1alpha G319S, a transactivation-deficient mutant, is associated with altered dynamics of diabetes onset in an Oji-Cree community', *Proc Natl Acad Sci U S A,* 99 (7), 4614-9.
Tzameli, I., et al. (2004), 'Regulated production of a peroxisome proliferator-activated receptor-gamma ligand during an early phase of adipocyte differentiation in 3T3-L1 adipocytes', *J Biol Chem,* 279 (34), 36093-102.
Unnikrishnan, R., Anjana, R. M., and Mohan, V. (2014), 'Diabetes in South Asians: is the phenotype different?', *Diabetes,* 63 (1), 53-5.
Vaxillaire, M., Bonnefond, A., and Froguel, P. (2012), 'The lessons of early-onset monogenic diabetes for the understanding of diabetes pathogenesis', *Best Pract Res Clin Endocrinol Metab,* 26 (2), 171-87.
Vaxillaire, M., et al. (1999), 'Anatomy of a homeoprotein revealed by the analysis of human MODY3 mutations', *J Biol Chem,* 274 (50), 35639-46.
Velho, G., et al. (1992), 'Primary pancreatic beta-cell secretory defect caused by mutations in glucokinase gene in kindreds of maturity onset diabetes of the young', *Lancet,* 340 (8817), 444-8.
Visser, M., et al. (1999), 'Elevated C-reactive protein levels in overweight and obese adults', *JAMA,* 282 (22), 2131-5.
Voight, B. F., et al. (2010), 'Twelve type 2 diabetes susceptibility loci identified through large-scale association analysis', *Nat Genet,* 42 (7), 579-89.
Walker, K. Z., et al. (2010), 'Diet and exercise in the prevention of diabetes', *J Hum Nutr Diet,* 23 (4), 344-52.
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).
Whiting, D. R., et al. (2011), 'IDF diabetes atlas: global estimates of the prevalence of diabetes for 2011 and 2030', *Diabetes Res Clin Pract,* 94 (3), 311-21.
WHO (2012), 'WHO Guidelines Approved by the Guidelines Review Committee', *Prevention and Control of Noncommunicable Diseases: Guidelines for Primary Health Care in Low Resource Settings* (Geneva: World Health Organization

World Health Organization.).
Winckler, W., et al. (2005), 'Association of common variation in the HNF1alpha gene region with risk of type 2 diabetes', *Diabetes,* 54 (8), 2336-42.
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.
Woodard-Grice, A. V., et al. (2008), 'Proteolytic shedding of ST6Gal-I by BACE1 regulates the glycosylation and function of alpha4beta1 integrins', *J Biol Chem,* 283 (39), 26364-73.
Xing, J., et al. (2010), 'Genetic diversity in India and the inference of Eurasian population expansion', *Genome Biol,* 11 (11), R113.
Xu, J. Y., et al. (2002), 'Mutations in the hepatocyte nuclear factor-1alpha gene in Chinese MODY families: prevalence and functional analysis', *Diabetologia,* 45 (5), 744-6.
Yajnik, C. S., et al. (2003), 'Neonatal anthropometry: the thin-fat Indian baby. The Pune Maternal Nutrition Study', *Int J Obes Relat Metab Disord,* 27 (2), 173-80.

Yamagata, K., et al. (1996a), 'Mutations in the hepatocyte nuclear factor-4alpha gene in maturity-onset diabetes of the young (MODY1)', *Nature,* 384 (6608), 458-60.
Yamagata, K., et al. (1996b), 'Mutations in the hepatocyte nuclear factor-1alpha gene in maturity-onset diabetes of the young (MODY3)', *Nature,* 384 (6608), 455-8.
Yasuda, K., et al. (2008), 'Variants in KCNQ1 are associated with susceptibility to type 2 diabetes mellitus', *Nat Genet,* 40 (9), 1092-7.
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.
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.

OPEN ACCESS Freely available online

PLOS ONE

Evaluation of Common Type 2 Diabetes Risk Variants in a South Asian Population of Sri Lankan Descent

Neelam Hassanali[1], N. Maneka G. De Silva[2], Neil Robertson[1,3], N. William Rayner[1,3], Amy Barrett[1], Amanda J. Bennett[1], Christopher J. Groves[1], David R. Matthews[1,4], Prasad Katulanda[1,5], Timothy M. Frayling[2], Mark I. McCarthy[1,3,6]*

1 Oxford Centre for Diabetes, Endocrinology & Metabolism, University of Oxford, Churchill Hospital, Oxford, United Kingdom, **2** Genetics of Complex Traits, University of Exeter Medical School, Exeter, United Kingdom, **3** Wellcome Trust Centre for Human Genetics, University of Oxford, Oxford, United Kingdom, **4** Harris Manchester College, University of Oxford, Oxford, United Kingdom, **5** Department of Clinical Medicine, Faculty of Medicine, University of Colombo, Colombo, Sri Lanka, **6** Oxford National Institute for Health Research Biomedical Research Centre, Churchill Hospital, Oxford, United Kingdom

Abstract

Introduction: Most studies seeking common variant associations with type 2 diabetes (T2D) have focused on individuals of European ancestry. These discoveries need to be evaluated in other major ancestral groups, to understand ethnic differences in predisposition, and establish whether these contribute to variation in T2D prevalence and presentation. This study aims to establish whether common variants conferring T2D-risk in Europeans contribute to T2D-susceptibility in the South Asian population of Sri Lanka.

Methodology: Lead single nucleotide polymorphism (SNPs) at 37 T2D-risk loci attaining genome-wide significance in Europeans were genotyped in 878 T2D cases and 1523 normoglycaemic controls from Sri Lanka. Association testing was performed by logistic regression adjusting for age and sex and by the Cochran-Mantel-Haenszel test after stratifying according to self-identified ethnolinguistic subgroup. A weighted genetic risk score was generated to examine the combined effect of these SNPs on T2D-risk in the Sri Lankan population.

Results: Of the 36 SNPs passing quality control, sixteen showed nominal ($p<0.05$) association in Sri Lankan samples, fifteen of those directionally-consistent with the original signal. Overall, these association findings were robust to analyses that accounted for membership of ethnolinguistic subgroups. Overall, the odds ratios for 31 of the 36 SNPs were directionally-consistent with those observed in Europeans ($p=3.2\times10^{-6}$). Allelic odds ratios and risk allele frequencies in Sri Lankan subjects were not systematically different to those reported in Europeans. Genetic risk score and risk of T2D were strongly related in Sri Lankans (per allele OR 1.10 [95%CI 1.08–1.13], $p=1.2\times10^{-17}$).

Conclusion: Our data indicate that most T2D-risk variants identified in Europeans have similar effects in South Asians from Sri Lanka, and that systematic difference in common variant associations are unlikely to explain inter-ethnic differences in prevalence or presentation of T2D.

Citation: Hassanali N, De Silva NMG, Robertson N, Rayner NW, Barrett A, et al. (2014) Evaluation of Common Type 2 Diabetes Risk Variants in a South Asian Population of Sri Lankan Descent. PLoS ONE 9(6): e98608. doi:10.1371/journal.pone.0098608

Editor: Balraj Mittal, Sanjay Gandhi Medical Institute, India

Received October 29, 2013; **Accepted** May 5, 2014; **Published** June 13, 2014

Copyright: © 2014 Hassanali et al. This is an open-access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

Funding: The authors acknowledge the support provided by Diabetes UK (09/0003943) and by the European Commission (ENGAGE: HEALTH-F4-2007-201413). NH is funded by a Diabetes UK PhD studentship award. Collection of the SLYDS and SLDCS cohorts were supported by (The National Science Foundation of Sri Lanka, the NIHR Biomedical Research Centre and the Sri Lanka Young Diabetes study group). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

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

* E-mail: mark.mccarthy@drl.ox.ac.uk

Introduction

Type 2 diabetes (T2D) is a major global health concern that is currently estimated to affect 336 million people worldwide [1]. It is widely accepted that T2D is a complex disorder and individual risk reflects the influence of environmental factors on a background of genetic predisposition. Over the past three decades the prevalence of T2D in South Asians has shown a particularly dramatic increase [2,3], prompted by profound changes in socioeconomic factors and lifestyle. Compared to European counterparts, South Asians tend to be diagnosed with diabetes earlier, with a lower BMI and display a more rapid decline in glycaemic control over time [3]. The increased prevalence of T2D extends to South Asian groups living outside their native countries, and this suggests that there may also be an underlying biological predisposition in addition to environmental and lifestyle factors [2,4].

The advent of large scale genome-wide association studies (GWAS) has led to the identification of over 70 genetic loci that contribute to T2D risk [5–17]. Most of the early studies were conducted in Europeans, but increasingly, similar approaches are being deployed in samples of South Asian, East Asian and African origin [11,13,18–22]. These studies have revealed novel signals

Table 1. Individual SNP association results for T2D risk in the Sri Lankan case control samples.

Nearest Gene	CHR	SNP	Alleles	T2D Risk allele in Sri Lankans	T2D Risk allele in Europeans	RAF in Sri Lankans	Odds Ratio (95% CI) adjusted for age and sex	p-value adjusted for age and sex	Odds ratio for CMH test (95% CI) adjusted for age, sex	CMH test p-value adjusted for age, sex
NOTCH2	1	rs10923931	G/T	T	T	0.25	1.04 (0.90–1.20)	0.566	1.04 (0.90–1.20)	0.610
PROX1	1	rs340874	C/T	C	C	0.51	1.13 (1.00–1.29)	0.053	1.16 (1.02–1.31)	0.021
BCL11A	2	rs243021	G/A	A	A	0.49	1.23 (1.09–1.41)	1.1×10^{-3}	1.25 (1.11–1.41)	3.6×10^{-4}
GCKR	2	rs780094	C/T	C	C	0.20	0.87 (0.75–1.03)	0.110	0.91 (0.78–1.06)	0.21
IRS1	2	rs7578326	A/G	A	A	0.81	1.20 (1.03–1.41)	0.023	1.17 (1.00–1.37)	0.055
THADA	2	rs7578597	T/C	T	T	0.89	1.17 (0.96–1.44)	0.118	1.10 (0.91–1.35)	0.321
ADAMTS9	3	rs4607103	C/T	C	C	0.46	1.02 (0.89–1.15)	0.781	0.99 (0.87–1.11)	0.815
ADCY5	3	rs11708067	A/G	A	A	0.81	1.22 (1.04–1.45)	0.016	1.28 (1.09–1.50)	2.7×10^{-3}
IGF2BP2	3	rs4402960	G/T	T	T	0.49	1.10 (0.97–1.26)	0.125	1.10 (0.97–1.25)	0.124
PPARG	3	rs1801282	C/G	C	C	0.91	1.21 (0.97–1.52)	0.093	1.30 (1.04–1.62)	0.021
WFS1	4	rs10010131	G/A	G	G	0.76	1.24 (1.07–1.45)	5.6×10^{-3}	1.25 (1.07–1.45)	4.1×10^{-3}
ZBED3	5	rs4457053	A/G	G	G	0.20	1.25 (1.08–1.46)	3.6×10^{-3}	1.23 (1.06–1.43)	6.7×10^{-3}
CDKAL1	6	rs10946398	T/C	C	C	0.24	1.09 (0.94–1.27)	0.240	1.07 (0.92–1.24)	0.390
DGKB/TMEM195	7	rs2191349	T/G	T	T	0.66	0.99 (0.86–1.12)	0.821	0.99 (0.87–1.13)	0.855
GCK	7	rs4607517	G/A	A	A	0.12	0.92 (0.75–1.13)	0.441	0.93 (0.76–1.13)	0.449
JAZF1	7	rs864745	C/T	T	T	0.79	1.17 (1.00–1.37)	0.049	1.14 (0.98–1.33)	0.082
KLF14	7	rs972283	G/A	G	G	0.63	1.01 (0.89–1.16)	0.834	1.03 (0.91–1.17)	0.633
SLC30A8	8	rs13266634	C/T	C	C	0.77	1.32 (1.14–1.55)	3.9×10^{-4}	1.33 (1.14–1.55)	3.0×10^{-4}
TP53INP1	8	rs896854	C/T	T	T	0.40	1.03 (0.90–1.17)	0.66	1.02 (0.90–1.16)	0.761
CDKN2A/B	9	rs10811661	T/C	T	T	0.85	1.34 (1.12–1.61)	1.2×10^{-3}	1.32 (1.10–1.57)	2.0×10^{-3}
CHCHD9	9	rs13292136	C/T	C	C	0.84	1.29 (1.08–1.54)	4.4×10^{-3}	1.31 (1.10–1.56)	2.4×10^{-3}
CDC123/CAMK1D	10	rs12779790	A/G	G	G	0.14	1.09 (0.91–1.31)	0.327	1.09 (0.91–1.30)	0.346
HHEX/IDE	10	rs1111875	C/T	C	C	0.36	1.14 (1.01–1.30)	0.048	1.17 (1.03–1.32)	0.017
TCF7L2	10	rs7903146	C/T	T	T	0.34	1.38 (1.21–1.59)	2.8×10^{-6}	1.35 (1.19–1.54)	6.1×10^{-6}
CENTD2	11	rs1552224	A/C	A	A	0.82	1.12 (0.95–1.33)	0.174	1.11 (0.95–1.31)	0.194
KCNJ11	11	rs5219	C/T	T	T	0.34	1.15 (1.01–1.31)	0.035	1.16 (1.02–1.31)	0.026
KCNQ1	11	rs231362	A/G	G	G	0.76	1.03 (0.89–1.19)	0.728	1.03 (0.90–1.20)	0.644
MTNR1B	11	rs10830963	C/G	G	G	0.84	1.12 (0.99–1.27)	0.067	1.10 (0.97–1.25)	0.132
HMGA2	12	rs1531343	G/C	C	C	0.19	1.23 (1.05–1.45)	8.8×10^{-3}	1.22 (1.04–1.43)	0.013
HNF1A	12	rs7957197	T/A	T	T	0.96	0.94 (0.67–1.33)	0.735	1.02 (0.73–1.42)	0.907
TSPAN8/LGR5	12	rs7961581	T/C	C	C	0.35	0.87 (0.77–0.99)	0.049	0.84 (0.74–0.96)	0.012
PRC1	15	rs8042680	A/C	A	A	0.74	1.18 (1.03–1.37)	0.020	1.24 (1.08–1.43)	0.003
ZFAND6	15	rs11634397	G/A	G	G	0.53	1.00 (0.89–1.14)	0.933	1.00 (0.88–1.13)	0.996

Table 1. Cont.

Nearest Gene	CHR	SNP	Alleles	T2D Risk allele in Sri Lankans	T2D Risk allele in Europeans	RAF in Sri Lankans	Odds Ratio (95% CI) adjusted for age and sex	p-value adjusted for age and sex	Odds ratio for CMH test (95% CI) adjusted for age, sex	CMH test p-value adjusted for age, sex
FTO	16	rs9939609	A/T	A	A	0.36	1.11 (0.98–1.27)	0.089	1.14 (1.00–1.30)	0.044
HNF1B(TCF2)	17	rs757210	C/T	T	T	0.28	1.01 (0.88–1.16)	0.851	0.99 (0.87–1.14)	0.931
DUSP9	X	rs5945326	A/G	A	A	0.57	1.15 (1.03–1.29)	0.010	1.29 (1.14–1.47)	6.1×10^{-5}

Chr: chromosome.

doi:10.1371/journal.pone.0098608.t001

[4,13,23], but have also shown appreciable overlap with associations first discovered in European groups [19–22].

Here, we extend these studies to South Asians from the island of Sri Lanka. As elsewhere in South Asia, the incidence of T2D is increasing and it is predicted that by 2030 approximately 14% of the adult population will have the condition, many of them undiagnosed [1]. Relatively little is known about the genetic predisposition of T2D in this country. In this study, we determined whether a set of T2D-risk variants reaching genome-wide significance in Europeans carry the same disease risk in South Asians from Sri Lanka.

Materials and Methodology

Study Samples

Cases and controls were ascertained from two independent collections of South Asian subjects from Sri Lanka. T2D cases (n = 1001, 44% male) were recruits to the Sri Lankan Young Diabetes Study (SLYDS), consecutively ascertained from private and government diabetes clinics [24]. Age of diabetes diagnosis was between 16–40 years and all participants were under the age of 45 years at recruitment. Of the 1001, 965 had DNA samples available for genotyping. Within these individuals, T2D status was defined if Glutamic Acid Decarboxylase Autoantibodies (GADA) titre was $\leq$14 units/ml, and if the interval between diagnosis and the initiation of insulin therapy was at least six months [24]. Twenty-nine samples were excluded due to missing GADA data, 48 for having insulin treatment within 6 months of diagnosis and 10 individuals for a positive diagnosis of mitochondrial diabetes (mt3243 A>G;), leaving 878 T2D cases available for inclusion [24].

Control subjects were participants in the Sri Lankan Diabetes Cardiovascular Study (SLDCS), a cross-sectional epidemiological study that used a multi-stage random cluster sampling technique to recruit 4388 subjects across seven Sri Lankan provinces [25]. DNA collection was initiated partway through the study and DNA samples were available for 1769 subjects. Of these, 1523 individuals who were confirmed as normoglycaemic based on oral glucose tolerance data (interpreted according to then-current ADA and WHO criteria), and under the age of 80 years, were included in this study [26].

At the time of recruitment, participants in both studies were classified according to the major ethnolinguistic and religious groups in Sri Lanka, using categories specified, for example, in the national Census, and based on a combination of spoken language, religion/cultural identification and surname [27]. The majority were Sinhalese (86%) or Tamil (5.4%), with the rest categorised as Muslim (8.3%) or Burgher (0.26%) or having other designations (0.04%).

Participant collection was approved by the Ethical Review Committee of the University of Colombo. All participants provided informed written consent [24,25].

Genotyping and quality control

We genotyped the lead single nucleotide polymorphisms (SNPs) at 37 T2D-risk loci that had reached genome-wide significance in Europeans from studies published as of mid-2010 [28,29]. We genotyped 2401 individuals (878 cases, 1523 controls) using Applied Biosystems TaqMan SNP genotyping assays on an Applied Biosystems 7900HT system. Seventy-four samples with a low (<80%) overall call rate were removed from further analysis leaving 2327 individuals (830 cases, 1497 controls) for final analysis. The average genotyping call rate for these 2327 individuals was 97%.

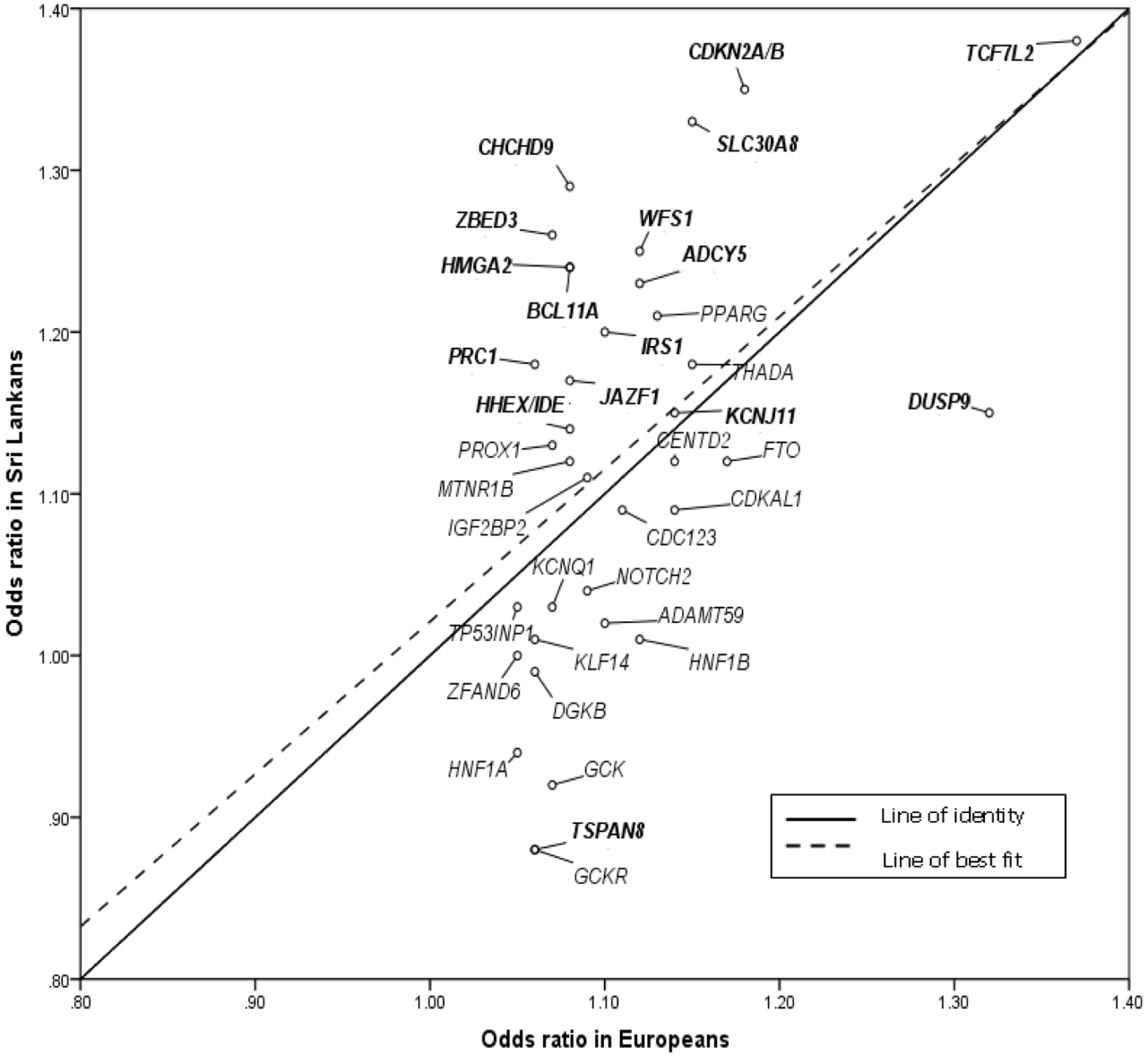


Figure 1. Comparison of allelic ORs between South Asians from Sri Lanka and Europeans. Loci labelled in bold are variants that showed nominal significance ($p<0.05$) in Sri Lankan subjects. European ORs used were derived from previously reported studies (Table S1).
doi:10.1371/journal.pone.0098608.g001

Statistical Analysis

All 37 SNPs were in Hardy Weinberg equilibrium (HWE) except for the SNP rs2237892 in the *KCNQ1* locus ($p<0.001$ in controls) which was removed from subsequent analyses. First, we used logistic regression to assess the association between each individual SNP and T2D status assuming a log additive model. All associations were adjusted for age and sex. In addition we reanalysed the case-control data after stratifying for self-identified ethnic subgroup using the Cochran–Mantel–Haenszel (CMH) test. The individual SNP association analyses were undertaken in PLINK v1.07 [30,31].

Next we tested the significance of genetic risk scores (GRS) that combine information from all 36 T2D associated SNPs using logistic regression. The SNPs were coded as 0,1 or 2 corresponding to the number of T2Ds risk increasing alleles in Europeans, except for the X chromosome SNP rs5945326 at *DUSP*9 where male genotypes were coded as 0 or 1, and female genotypes as 0, 0.5 or 1 (to reflect random X inactivation). These analyses were performed using Stata/SE version 10.1 for Windows (StataCorp, Brownsville, TX). To create the GRS, we used individuals with genotypes available from at least 29 of the 36 type 2 diabetes SNPs (i.e. 80% of the SNPs genotyped), and accounted for the varying effect sizes of each SNP using equation 1, where w is the natural log of the per allele type 2 diabetes odds ratio (OR) reported in Europeans.

$$\text{Weighted score} = (w_1 \times \text{SNP}_1) + (\text{w}_1 \times \text{SNP}_2) + \ldots\ldots\ldots + (\text{w}_\text{n} \times \text{SNP}_\text{n}) \quad (1)$$

We rescaled the weighted score from above to reflect the number of available SNPs using equation (2) as described previously [32].

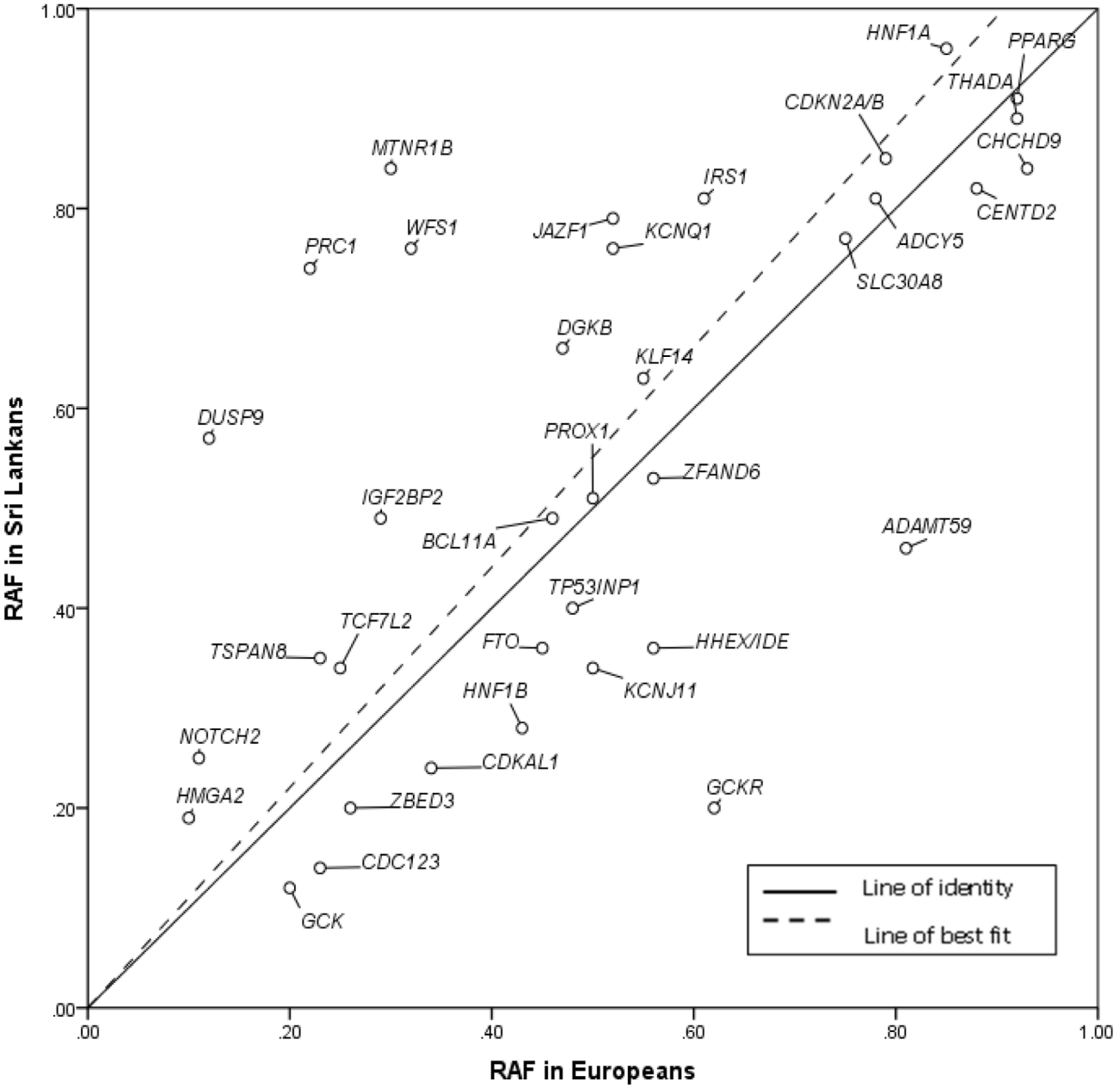


Figure 2. Comparison of risk allele frequencies (RAF) between South Asians from Sri Lanka samples included in this study (cases and controls combined) and previously reported RAF for Europeans (from HapMap).
doi:10.1371/journal.pone.0098608.g002

$$\text{Genetic risk score} = \frac{\text{Weighted score} \times \text{Number of SNPs available}}{\text{Sum weight of the available SNPs}} \quad (2)$$

We used this weighted GRS as the independent variable and T2D status as the dependent variable in logistic regression analyses. We also stratified individuals into quintiles of GRS.

To compare effect size estimates for Sri Lankan case-control samples with those observed in Europeans, we compiled odds ratio (OR) estimates for each locus for European case-control data from the literature [6,7,14–16,29,33–36] (Table S1). To minimise inflation of these estimates in initial genome-wide association discovery samples (the "winner's curse"), we used OR values from replication samples wherever possible.

Power Calculations

Quanto was used to calculate power under assumptions of a log-additive model, a disease prevalence of 10%, and a significance threshold (α) of 0.05. Power was calculated for each individual SNP using allelic odds ratios (for European case-control comparisons) collated from the literature and effect allele frequencies from the CEU component of HapMap (Table S1) [37].

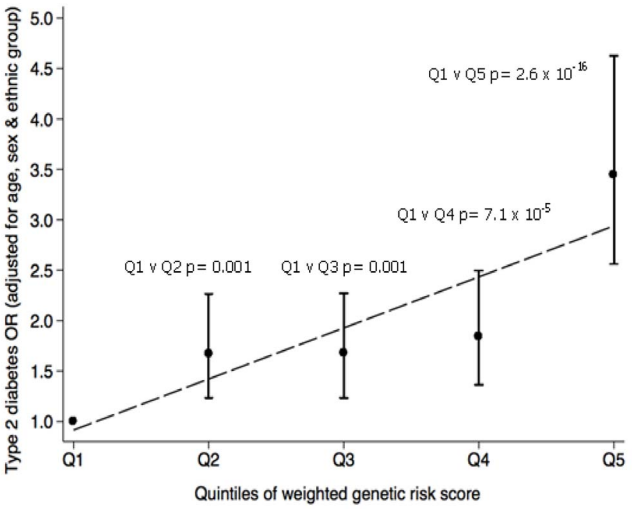


Figure 3. The combined impact of the 36 T2D-associated SNPs on T2D risk in T2D cases and controls of South Asian origin from Sri Lanka. Subjects were grouped into quintiles of the weighted genetic risk score. Circles represent the T2D odds ratio (adjusted for age, sex and ethnic group) when comparing each quintile group to the group in the lowest quintile (Q1). The capped lines represent the 95% CI of the T2D odds ratios.
doi:10.1371/journal.pone.0098608.g003

Results

Individual SNP associations for the 36 SNPs in 830 T2D patients and 1497 controls are summarised in Table 1. Nominal associations with T2D ($p \leq 0.05$) were observed for 16 of the 36 SNPs tested when adjusted for age and sex. Given some case-control imbalance with respect to "self-identified" ethnolinguistic subgroups (e.g. Sinhala, Tamil), case-control analyses were repeated in stratified samples using the CMH test. The results obtained were broadly comparable with highly-correlated odds ratios (Table 1). In the CMH analysis, a total of 17 loci were nominally-associated with T2D, fourteen of them overlapping with the non-stratified analysis. Two loci (*IRS1* and *JAZF1*) were no longer associated ($p < 0.05$) in the CMH test, but three others (*PROX1*, *PPARG* and *FTO*) became significant for the first time. Subsequent analyses were performed (unless otherwise stated) on the combined dataset.

The strongest effects were seen at *TCF7L2* (OR 1.38 [95% CI 1.21–1.58], $p = 2.8 \times 10^{-6}$) and *SLC30A8* (OR 1.32 [95% CI 1.14–1.55], $p = 3.9 \times 10^{-4}$). At 15 of the 16 loci, the direction of effect was consistent with that reported in the original discovery study, the exception being rs7961581 (*TSPAN8/LGR5*: OR 0.87 [95% CI 0.77–0.99], $p = 0.049$). In all, 31 of the 36 SNPs tested showed evidence of association that was directionally consistent with previous studies in Europeans (i.e. the same allele increases T2D risk in the two populations) (binomial $p = 3.2 \times 10^{-6}$).

Given the substantial differences in prevalence and presentation of T2D between European and South Asian populations, we sought evidence for consistent differences in effect size or allele frequency between the present study and previous reports from European studies (see methods). We found significant correlations between Sri Lankan and European samples for both the allelic odds ratio point estimates (Figure 1; $r = 0.50$, $p = 1.8 \times 10^{-3}$) and risk allele frequencies (Figure 2: $r = 0.64$, $p = 2.3 \times 10^{-5}$) but no suggestion of systematic differences in either. Allelic OR point estimates were greater in Sri Lankans than Europeans at 19 of 36 loci, and risk allele frequencies at 20 of 36.

As expected, individuals carrying greater numbers of (weighted) T2D risk increasing alleles had increased T2D risk (Figure 3), with an allelic OR of 1.10 (95%CI: 1.08–1.13, $p = 1.2 \times 10^{-17}$) per unit of the weighted genetic risk score. Individuals in the highest quintile of the genetic risk score had more than three-fold higher odds of T2D (3.44 [95%CI: 2.56–4.63] $p = 2.6 \times 10^{-16}$) when compared to individuals in the lowest quintile.

Discussion

In this study, we have shown that established T2D-risk variants, most of them first identified in European samples, show strong enrichment of association in T2D cases and controls of South Asian origin from Sri Lanka. This pattern of enrichment, along with the absence of any systematic difference in risk-allele frequency or odds ratios between Sri Lankan and European samples has several important corollaries.

Firstly, these data provide further evidence for the transethnic consistency in allelic patterns of association for T2D, building on similar findings seen in a variety of ethnic groups including other samples of South Asian origin [17,19–22,38]. These patterns of transethnic consistency are consistent with a model in which the (often unknown) casual variants driving these association signals are also themselves common, a model which is also increasingly supported by fine mapping data [17,39]. However, definitive confirmation of this model will require comprehensive identification of variants in these regions (e.g. via genome sequencing studies that are ongoing) such that the contribution of variants of all frequencies to disease predisposition can be directly tested.

Secondly, despite differences in both the prevalence and presentation of T2D between Sri Lanka and Europe, we observed no systematic differences in either risk allele frequency or effect size. We conclude therefore, that these ethnic differences in epidemiological and physiological patterns cannot be attributed to differences in common variant predisposition.

Though the general patterns are clear, the relatively modest sample sizes available in this study limit the inferences that can be made at any individual locus. None of the variants tested reached stringent genome-wide significance, and only about half of the loci reached nominal significance (i.e. $p \leq 0.05$). The patterns of association seen even amongst those variants not reaching nominal significance (from the twenty SNP associations with $p > 0.05$, fifteen are directionally consistent with the associations reported in Europeans) indicate that many of these are likely to be false-negatives reflecting the limited power of our study. Many of these loci have very modest odds ratios and would have required much larger sample sizes to be detected than were available to us. Indeed, amongst the 15 loci with no formal evidence ($p > 0.05$) of association in our study, but displaying directional consistency with data from Europeans, are several that show evidence of association in other South Asian case-control studies. For example, variants at the *GCKR* and *CDC123* loci were not associated with T2D in the present study, but have strong associations in far larger meta-analyses of South Asian samples [13].

In summary, we have shown that common T2D risk variants identified in Europeans have a similar genetic risk in Sri Lankans, adding further to the evidence that South Asians and Europeans share many overlapping common variants which contribute to T2D risk.

Supporting Information

Table S1 Summary of the reported allele frequency and odds ratio in Europeans for the T2D SNPs investigated.
(DOCX)

Author Contributions

Conceived and designed the experiments: NH NMDS PK MIM. Performed the experiments: NH. Analyzed the data: NH NMDS MIM. Contributed reagents/materials/analysis tools: PK. Wrote the paper: NH NMDS MIM. Contributed to the acquisition of the data: NH NMDS NR NWR AB AJB CJG DRM PK TMF MIM. Revising the manuscript critically and approved the final version: NH NMDS NR NWR AB AJB CJG DRM PK TMF MIM.

References

1. Whiting DR, Guariguata L, Weil C, Shaw J (2011) IDF diabetes atlas: global estimates of the prevalence of diabetes for 2011 and 2030. Diabetes research and clinical practice 94: 311–321.
2. Gujral UP, Pradeepa R, Weber MB, Narayan KM, Mohan V (2013) Type 2 diabetes in South Asians: similarities and differences with white Caucasian and other populations. Annals of the New York Academy of Sciences 1281: 51–63.
3. Jayawardena R, Ranasinghe P, Byrne NM, Soares MJ, Katulanda P, et al. (2012) Prevalence and trends of the diabetes epidemic in South Asia: a systematic review and meta-analysis. BMC Public Health 12: 380.
4. Saxena R, Saleheen D, Been LF, Garavito ML, Braun T, et al. (2013) Genome-wide association study identifies a novel locus contributing to type 2 diabetes susceptibility in Sikhs of Punjabi origin from India. Diabetes 62: 1746–1755.
5. Morris AP, Voight BF, Teslovich TM, Ferreira T, Segre AV, et al. (2012) Large-scale association analysis provides insights into the genetic architecture and pathophysiology of type 2 diabetes. Nature genetics 44: 981–990.
6. Zeggini E, Weedon MN, Lindgren CM, Frayling TM, Elliott KS, et al. (2007) Replication of genome-wide association signals in UK samples reveals risk loci for type 2 diabetes. Science 316: 1336–1341.
7. Zeggini E, Scott LJ, Saxena R, Voight BF, Marchini JL, 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: 638–645.
8. Johansen CT, Wang J, Lanktree MB, Cao H, McIntyre AD, et al. (2010) Excess of rare variants in genes identified by genome-wide association study of hypertriglyceridemia. Nat Genet 42: 684–687.
9. Steinthorsdottir V, Thorleifsson G, Reynisdottir I, Benediktsson R, Jonsdottir T, et al. (2007) A variant in CDKAL1 influences insulin response and risk of type 2 diabetes. Nat Genet 39: 770–775.
10. Scott LJ, Mohlke KL, Bonnycastle LL, Willer CJ, Li Y, et al. (2007) A genome-wide association study of type 2 diabetes in Finns detects multiple susceptibility variants. Science 316: 1341–1345.
11. Saxena R, Elbers CC, Guo Y, Peter I, Gaunt TR, et al. (2012) Large-scale gene-centric meta-analysis across 39 studies identifies type 2 diabetes loci. Am J Hum Genet 90: 410–425.
12. Lyssenko V, Nagorny CL, Erdos MR, Wierup N, Jonsson A, et al. (2009) Common variant in MTNR1B associated with increased risk of type 2 diabetes and impaired early insulin secretion. Nat Genet 41: 82–88.
13. Kooner JS, Saleheen D, Sim X, Sehmi J, Zhang W, et al. (2011) Genome-wide association study in individuals of South Asian ancestry identifies six new type 2 diabetes susceptibility loci. Nat Genet 43: 984–989.
14. Frayling TM, Timpson NJ, Weedon MN, Zeggini E, Freathy RM, et al. (2007) A common variant in the FTO gene is associated with body mass index and predisposes to childhood and adult obesity. Science 316: 889–894.
15. Dupuis J, Langenberg C, Prokopenko I, Saxena R, Soranzo N, et al. (2010) New genetic loci implicated in fasting glucose homeostasis and their impact on type 2 diabetes risk. Nat Genet 42: 105–116.
16. Sladek R, Rocheleau G, Rung J, Dina C, Shen L, et al. (2007) A genome-wide association study identifies novel risk loci for type 2 diabetes. Nature 445: 881–885.
17. Mahajan A, Jin MG, Zhang W, Below JE, Gaulton KJ, et al. (2014) Genome-wide trans-ancestry meta-analysis provides insight into the genetic architecture of type 2 diabetes susceptibility. Nat Genet 46: 234–244.
18. Palmer ND, McDonough CW, Hicks PJ, Roh BH, Wing MR, et al. (2012) A genome-wide association search for type 2 diabetes genes in African Americans. PLoS One 7: e29202.
19. Chauhan G, Spurgeon CJ, Tabassum R, Bhaskar S, Kulkarni SR, et al. (2010) Impact of common variants of PPARG, KCNJ11, TCF7L2, SLC30A8, HHEX, CDKN2A, IGF2BP2, and CDKAL1 on the risk of type 2 diabetes in 5,164 Indians. Diabetes 59: 2068–2074.
20. Chidambaram M, Radha V, Mohan V (2010) Replication of recently described type 2 diabetes gene variants in a South Indian population. Metabolism 59: 1760–1766.
21. Rees SD, Hydrie MZ, Shera AS, Kumar S, O'Hare JP, et al. (2011) Replication of 13 genome-wide association (GWA)-validated risk variants for type 2 diabetes in Pakistani populations. Diabetologia 54: 1368–1374.
22. Sanghera DK, Ortega L, Han S, Singh J, Ralhan SK, et al. (2008) Impact of nine common type 2 diabetes risk polymorphisms in Asian Indian Sikhs: PPARG2 (Pro12Ala), IGF2BP2, TCF7L2 and FTO variants confer a significant risk. BMC Med Genet 9: 59.
23. Tabassum R, Chauhan G, Dwivedi OP, Mahajan A, Jaiswal A, et al. (2013) Genome-wide association study for type 2 diabetes in Indians identifies a new susceptibility locus at 2q21. Diabetes 62: 977–986.
24. Katulanda P, Shine B, Katulanda GW, Silva A, Asfir EL, et al. (2008) Diabetes mellitus among young adults in Sri Lanka–role of GAD antibodies in classification and treatment: the Sri Lanka Young Diabetes study. Diabetologia 51: 1368–1374.
25. Katulanda P, Constantine GR, Mahesh JG, Sheriff R, Seneviratne RD, et al. (2008) Prevalence and projections of diabetes and pre-diabetes in adults in Sri Lanka–Sri Lanka Diabetes, Cardiovascular Study (SLDCS). Diabet Med 25: 1062–1069.
26. Alberti KG, Zimmet PZ (1998) Definition, diagnosis and classification of diabetes mellitus and its complications. Part 1: diagnosis and classification of diabetes mellitus provisional report of a WHO consultation. Diabet Med 15: 539–553.
27. Population by Ethnicity and District http://www.statistics.gov.lk/PopHouSat/PopulationAtla_2012/03_DistrictMaps/Map%20P2.4.1%20Population%20by%20Ethnicity%20and%20District,%202012.pdf. Accessed 2014 May 12.
28. McCarthy MI (2010) Genomics, type 2 diabetes, and obesity. N Engl J Med 363: 2339–2350.
29. Voight BF, Scott LJ, Steinthorsdottir V, Morris AP, Dina C, et al. (2010) Twelve type 2 diabetes susceptibility loci identified through large-scale association analysis. Nat Genet 42: 579–589.
30. http://pngu.mgh.harvard.edu/~purcell/plink/ v1.07 downloaded in 2010.
31. Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MA, et al. (2007) PLINK: a tool set for whole-genome association and population-based linkage analyses. Am J Hum Genet 81: 559–575.
32. Lin X, Song K, Lim N, Yuan X, Johnson T, et al. (2009) Risk prediction of prevalent diabetes in a Swiss population using a weighted genetic score–the CoLaus Study. Diabetologia 52: 600–608.
33. Altshuler D, Hirschhorn JN, Klannemark M, Lindgren CM, Vohl MC, et al. (2000) The common PPARgamma Pro12Ala polymorphism is associated with decreased risk of type 2 diabetes. Nat Genet 26: 76–80.
34. Gloyn AL, Cummings EA, Edghill EL, Harries LW, Scott R, et al. (2004) Permanent neonatal diabetes due to paternal germline mosaicism for an activating mutation of the KCNJ11 Gene encoding the Kir6.2 subunit of the beta-cell potassium adenosine triphosphate channel. J Clin Endocrinol Metab 89: 3932–3935.
35. Grant SF, Thorleifsson G, Reynisdottir I, Benediktsson R, Manolescu A, et al. (2006) Variant of transcription factor 7-like 2 (TCF7L2) gene confers risk of type 2 diabetes. Nat Genet 38: 320–323.
36. Sandhu MS, Weedon MN, Fawcett KA, Wasson J, Debenham SL, et al. (2007) Common variants in WFS1 confer risk of type 2 diabetes. Nat Genet 39: 951–953.
37. Quanto software downloaded Available: http://hydra.usc.edu/gxe/. Accessed 2011 May 4.
38. Waters KM, Stram DO, Hassanein MT, Le Marchand L, Wilkens LR, et al. (2010) Consistent association of type 2 diabetes risk variants found in europeans in diverse racial and ethnic groups. PLoS Genet 6: e1001078.
39. Maller JB, McVean G, Byrnes J, Vukcevic D, Palin K, et al. (2012) Bayesian refinement of association signals for 14 loci in 3 common diseases. Nat Genet 44: 1294–1301.

nature genetics

Genome-wide association study in individuals of South Asian ancestry identifies six new type 2 diabetes susceptibility loci

Jaspal S Kooner[1–3,46], Danish Saleheen[4,5,46], Xueling Sim[6,46], Joban Sehmi[1,2,46], Weihua Zhang[7,46], Philippe Frossard[4,46], Latonya F Been[8], Kee-Seng Chia[6,9], Antigone S Dimas[10,11], Neelam Hassanali[12], Tazeen Jafar[13,14], Jeremy B M Jowett[15], Xinzhong Li[1], Venkatesan Radha[16], Simon D Rees[17,18], Fumihiko Takeuchi[19], Robin Young[5], Tin Aung[20,21], Abdul Basit[22], Manickam Chidambaram[16], Debashish Das[2], Elin Grundberg[23], Åsa K Hedman[11], Zafar I Hydrie[22], Muhammed Islam[13], Chiea-Chuen Khor[6,21,24], Sudhir Kowlessur[25], Malene M Kristensen[15], Samuel Liju[16], Wei-Yen Lim[6], David R Matthews[12], Jianjun Liu[24], Andrew P Morris[11], Alexandra C Nica[10], Janani M Pinidiyapathirage[26], Inga Prokopenko[11], Asif Rasheed[4], Maria Samuel[4], Nabi Shah[4], A Samad Shera[27], Kerrin S Small[23,28], Chen Suo[6], Ananda R Wickremasinghe[26], Tien Yin Wong[20,21,29], Mingyu Yang[30], Fan Zhang[30], DIAGRAM[31], MuTHER[31], Goncalo R Abecasis[32], Anthony H Barnett[17,18], Mark Caulfield[33], Panos Deloukas[34], Timothy M Frayling[35], Philippe Froguel[36], Norihiro Kato[19], Prasad Katulanda[12,37], M Ann Kelly[17,18], Junbin Liang[30], Viswanathan Mohan[16,38], Dharambir K Sanghera[8], James Scott[1], Mark Seielstad[39], Paul Z Zimmet[15], Paul Elliott[7,40,46], Yik Ying Teo[6,9,24,41,42,46], Mark I McCarthy[11,12,43,46], John Danesh[5,46], E Shyong Tai[9,44–46] & John C Chambers[2,3,7,46]

© 2011 Nature America, Inc. All rights reserved.

We carried out a genome-wide association study of type-2 diabetes (T2D) in individuals of South Asian ancestry. Our discovery set included 5,561 individuals with T2D (cases) and 14,458 controls drawn from studies in London, Pakistan and Singapore. We identified 20 independent SNPs associated with T2D at $P < 10^{-4}$ for testing in a replication sample of 13,170 cases and 25,398 controls, also all of South Asian ancestry. In the combined analysis, we identified common genetic variants at six loci (*GRB14*, *ST6GAL1*, *VPS26A*, *HMG20A*, *AP3S2* and *HNF4A*) newly associated with T2D ($P = 4.1 \times 10^{-8}$ to $P = 1.9 \times 10^{-11}$). SNPs at *GRB14* were also associated with insulin sensitivity ($P = 5.0 \times 10^{-4}$), and SNPs at *ST6GAL1* and *HNF4A* were also associated with pancreatic beta-cell function ($P = 0.02$ and $P = 0.001$, respectively). Our findings provide additional insight into mechanisms underlying T2D and show the potential for new discovery from genetic association studies in South Asians, a population with increased susceptibility to T2D.

Individuals with South Asian ancestry are at an up to fourfold higher risk of T2D compared to European populations[1,2]. T2D currently affects ~55 million South Asians worldwide and is projected to affect ~80 million South Asians by 2030 (ref. 3), which will comprise one-quarter of all people with T2D worldwide. Though unhealthy diet, obesity and physical inactivity contribute to T2D, it also has an important genetic contribution[4]. T2D is heritable in South Asians, as it is in other populations[5].

To date, genome-wide association studies (GWAS) have identified variants in and around 42 genes as associated with susceptibility to T2D[4]. However, these studies have been predominantly performed in populations of European ancestry. We therefore used genome-wide association to identify common genetic variants underlying risk of T2D in South Asians. The study design is shown in **Figure 1**.

In stage one, we analyzed data from GWAS performed in 5,561 South Asian cases and 14,458 South Asian controls from the London Life Sciences Population (LOLIPOP) study[6], the Pakistan Risk of Myocardial Infarction Study (PROMIS)[7] and the Singapore Indian Eye (SINDI) Study[8] using Illumina genotyping arrays. We identified South Asians as people with ancestry from the Indian subcontinent (India, Pakistan, Sri Lanka and Bangladesh). We summarized the characteristics of the participants and genotyping arrays used (**Supplementary Note** and **Supplementary Table 1**). We excluded samples with <95% call rate as well as SNPs with call rate <97%, Hardy-Weinberg equilibrium $P < 10^{-6}$ or minor allele frequency <1%. We used principal components analysis to assess for population substructure[9]; the results of this analysis confirmed that the GWAS samples were representative of South Asian ancestry with no evidence for stratification between cases and controls or between the three studies (**Supplementary Figs. 1** and **2**).

*A full list of authors' affiliations appears at the end of this paper.

Received 4 January; accepted 3 August; published online 28 August; corrected online (details online) 16 September 2011; doi:10.1038/ng.921

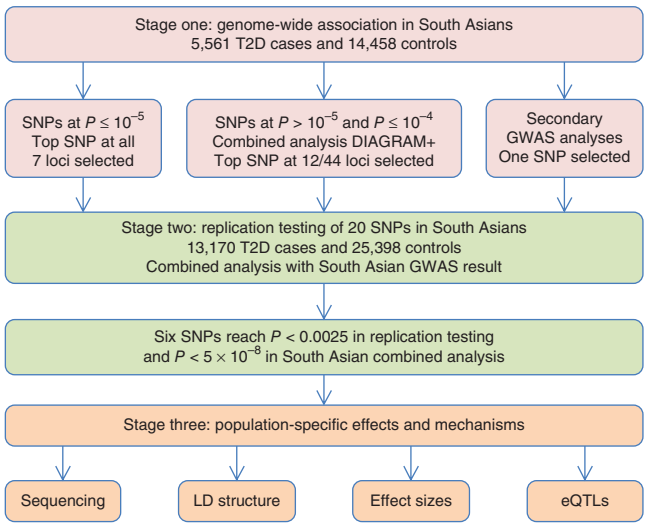


Figure 1 Summary of study design.

The primary analysis tested the association with T2D of the 568,976 autosomal SNPs that had been directly genotyped and passed quality control testing among the South Asian GWAS participants. We chose to limit the primary analysis to directly genotyped rather than imputed SNPs, as a South-Asian–specific dense haplotype map has not been described. This approach was enabled by the use of Illumina microarrays for all samples (17,880 samples on the Illumina 610 or 660 and 2,139 samples on the Illumina 317). We tested SNP associations with T2D separately among men and women in each study using logistic regression and an additive genetic model. We included principal components as covariates to adjust for population substructure[9]; there were no other covariates used in the regression analyses. We combined the results from the separate studies using a fixed effects inverse variance meta-analysis implemented in METAL (see URLs) and output association test results adjusted for the genomic control inflation factor. There was no evidence for inflation of the test statistics ($\lambda = 1.00$ to $\lambda = 1.03$; **Supplementary Table 2** and **Supplementary Fig. 3**).

In the primary analysis, one locus reached genome-wide significance ($P < 5 \times 10^{-8}$); the lead SNP was rs7903146 in *TCF7L2*, which is a well described T2D risk variant (**Fig. 2**)[10]. There were a further 59 independent loci associated with T2D at $P \leq 1 \times 10^{-4}$, of which 50 have not previously been described in GWAS of T2D (**Supplementary Table 3**). To help prioritize these loci for replication testing in stage two, we carried out a combined analysis of the South Asian discovery data with results from the DIAGRAM+ GWAS meta-analysis (8,130 cases with T2D and 38,987 controls of European ancestry)[11]. We then took forward for replication testing (i) all loci associated with T2D in South Asians alone at $P \leq 1 \times 10^{-5}$ ($N = 7$) and (ii) among the loci associated with T2D in South Asians at $P > 1 \times 10^{-5}$ and $P \leq 1 \times 10^{-4}$, the 12 independent SNPs with the lowest P values in the combined analysis with results from DIAGRAM+ (corresponding to ~$P < 10^{-3}$). This strategy was designed to maximize discovery of both genetic loci specific to South Asians as well as loci shared with European populations.

We carried out replication testing in 13,170 cases and 25,398 controls of South Asian ancestry (**Supplementary Note** and **Supplementary Table 4**). We tested SNP associations with T2D separately in each cohort and then combined the results by inverse variance meta-analysis. Six SNPs were associated with T2D in the South Asian replication samples at $P < 2.5 \times 10^{-3}$ ($P < 0.05$ after correction for multiple testing): rs3923113 near *GRB14*, rs16861329 in *ST6GAL1*, rs1802295 in *VPS26A*, rs2028299 near *AP3S2*, rs7178572 in *HMG20A* and rs4812829 in *HNF4A* (**Table 1**, **Supplementary Table 5** and **Supplementary Fig. 4**). These six SNPs reached genome-wide significance ($P < 5 \times 10^{-8}$) in a combined analysis of GWAS data from South Asians and the results for South Asians from the replication samples (**Table 1**).

Of the 12 SNPs with modest statistical evidence in the South Asian GWAS ($P > 1 \times 10^{-5}$ and $P \leq 1 \times 10^{-4}$) that we carried forward for replication testing based on the combined analysis with the DIAGRAM+ data, 3 showed replication and reached genome-wide significance in South Asians. This lends support to the view that selection of SNPs based on a combined analysis with Europeans is likely to enrich for true associations.

As secondary analyses, we carried out the following five GWAS meta-analyses: (i) male-gender specific; (ii) female-gender specific; (iii) body mass index (BMI) adjusted; (iv) lean T2D cases (BMI < 25 kg/m^2) compared to overweight controls (BMI > 25 kg/m^2); and (v) analysis of GWAS data imputed with untyped SNPs from the HapMap2 reference panel (for a total of 2,646,472 SNPs). One locus reached $P < 5 \times 10^{-8}$ for association with T2D among women (**Supplementary Fig. 5**). However, the sentinel SNP (rs17052370, near *UBBP4*, odds ratio (OR) = 1.37, 95% confidence interval (CI) 1.23–1.52, $P = 8.9 \times 10^{-9}$) did not replicate among South Asian women either in the replication samples (OR = 1.03, 95% CI 0.97–1.09, $P = 0.32$) or in the combined analysis of the GWAS and replication data (OR = 1.10, 95% CI 1.04–1.15, $P = 3.4 \times 10^{-4}$). There were no additional loci identified from the other secondary analysis (**Supplementary Fig. 5**).

Because overweight status and obesity are major risk factors for T2D[12], we investigated whether the relationship of the sentinel SNPs with T2D might be mediated through adiposity. We found that the sentinel SNPs are not associated with BMI or waist-hip ratio in South Asians (**Supplementary Table 6**). Furthermore, the association of the sentinel SNPs with T2D is not materially influenced by additional adjustment for measures of obesity among South Asians in the replication samples (**Supplementary Table 7**). Adiposity therefore does not mediate the relationships between these SNPs and T2D.

To provide insight into the genetic mechanisms underlying the observed associations with T2D, we sequenced the six associated regions (1 Mb on either side of sentinel SNPs) in 109 South Asians. We identified 49,145 genetic variants at these loci in South Asians, of which 24,902 are not present in the HapMap or 1000 Genomes Project[13] datasets (see URLs; **Supplementary Table 8**). However, there were no pathogenic SNPs (coding, splice site or regulatory) in linkage disequilibrium (LD) at $r^2 > 0.5$ with the sentinel SNPs; we obtained the same results in searches of the HapMap and 1000 Genomes Project datasets.

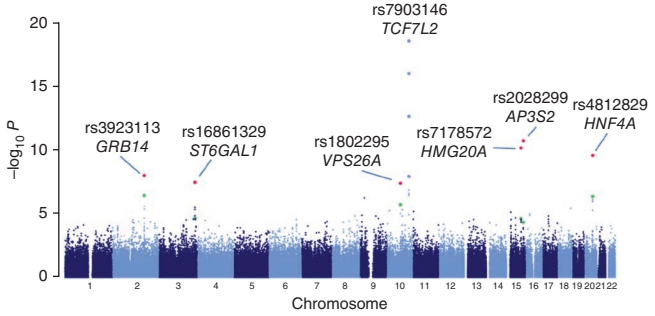


Figure 2 Manhattan plot for the primary South Asian genome-wide association analysis of men and women using directly genotyped SNPs. At the six new loci reaching genome wide-significance, the sentinel SNPs are indicated with a green dot for the GWAS result and a red dot for the combined analysis of the GWAS and replication data in South Asians.

© 2011 Nature America, Inc. All rights reserved.

© 2011 Nature America, Inc. All rights reserved.

Table 1 Genomic location and association test results for the sentinel SNPs from the six loci reaching genome-wide significance

SNP	Chr.	Position	Nearest Gene	Alleles[a] (A/R)	RAF		GWAS South Asians			Replication South Asians		Combined analysis South Asians		Europeans (DIAGRAM+)			Global analysis	
					SA	EW	*N*	OR[b]	*P*	OR[b]	*P*	OR[b]	*P*	OR[b]	*P*	P_{hetero}	OR[b]	*P*
rs3923113	2	165,210,095	*GRB14*	C/A	0.74	0.64	18,174	1.15 1.09–1.21	3.7×10^{-7}	1.07 1.03–1.11	6.7×10^{-4}	1.09 1.06–1.13	1.0×10^{-8}	1.05 1.01–1.10	2.0×10^{-2}	0.64	1.08 1.05–1.11	1.6×10^{-9}
rs16861329	3	188,149,155	*ST6GAL1*	A/G	0.75	0.86	18,186	1.12 1.07–1.19	2.3×10^{-5}	1.07 1.03–1.11	1.6×10^{-4}	1.09 1.06–1.12	3.4×10^{-8}	1.02 0.95–1.09	0.62	0.21	1.08 1.05–1.11	1.3×10^{-7}
rs1802295[c]	10	70,601,480	*VPS26A*	G/A	0.26	0.31	15,506*	1.14 1.08–1.20	1.9×10^{-6}	1.06 1.03–1.10	6.6×10^{-4}	1.08 1.05–1.12	4.1×10^{-8}	1.04 1.00–1.09	6.0×10^{-2}	0.50	1.07 1.05–1.10	2.1×10^{-8}
rs7178572	15	75,534,245	*HMG20A*	A/G	0.52	0.71	18,193	1.10 1.05–1.15	2.4×10^{-5}	1.08 1.05–1.12	7.0×10^{-7}	1.09 1.06–1.12	7.1×10^{-11}	1.07 1.02–1.12	2.6×10^{-3}	0.59	1.08 1.06–1.11	9.2×10^{-13}
rs2028299	15	88,175,261	*AP3S2*	A/C	0.31	0.31	18,076	1.11 1.05–1.16	4.8×10^{-5}	1.09 1.06–1.13	1.1×10^{-7}	1.10 1.07–1.13	1.9×10^{-11}	1.05 1.00–1.09	4.0×10^{-2}	0.12	1.08 1.06–1.11	1.2×10^{-11}
rs4812829	20	42,422,681	*HNF4A*	G/A	0.29	0.19	18,186	1.14 1.08–1.19	4.5×10^{-7}	1.07 1.04–1.11	2.8×10^{-5}	1.09 1.06–1.12	2.6×10^{-10}	1.08 1.02–1.14	1.0×10^{-2}	0.90	1.09 1.06–1.12	8.2×10^{-12}

Shown are the genomic location and association test results for the sentinel SNPs from the six loci reaching genome-wide significance ($P < 5 \times 10^{-8}$) among South Asians. Chr., chromosome; RAF, risk allele frequency; SA, South Asian; EW, European; P_{hetero}, the comparison between the odds ratios for type 2 diabetes among South Asians in the replication sample and among Europeans in DIAGRAM+.
[a]For the alleles, R indicates the risk allele and A indicates the alternate allele. [b]Results are provided as odds ratio (OR) (95% confidence interval) per copy of the risk allele. [c]rs1802295 is not available on the Illumina 317K array. Results for Europeans are from the GWAS stage of DIAGRAM+.

To investigate possible transcriptional mechanisms underlying our associations with T2D, we investigated the relationships of sentinel SNPs with *cis*–expression quantitative trait loci (eQTLs) in adipose tissue, liver, peripheral blood leucocytes and other tissues (**Supplementary Table 9**). At 2q24, rs3923113 is in high LD ($r^2 = 0.85$) with rs10195252, which was recently reported to be associated with central adiposity and increased expression of *GRB14* in adipose tissue ($P < 10^{-10}$)[14]. At 15q26, rs2028299 is closely associated with expression of *C15orf38* in skin and fat ($P < 10^{-16}$). At 15q24, rs7178572 is weakly associated with expression of *PSTPIP1* but is in low LD ($r^2 < 0.3$) with the peak SNP associated with this eQTL; *PSTPIP1* is unlikely to be the causal gene at this locus.

To test for heterogeneity between the participating South Asian GWAS cohorts, we compared effect sizes at (i) the loci identified in the present study and (ii) the 42 loci previously reported to be associated with T2D in GWAS. After correction for multiple testing, there was no evidence for heterogeneity of effect at any of these loci (**Supplementary Table 10**). Findings were similar among South Asians in the replication samples, although there was some evidence for heterogeneity at the *GRB14* locus (**Supplementary Fig. 4**). There was also no evidence for heterogeneity of effect between South Asian men and women except at rs17052370, which was suggested in the women-only gender-specific secondary analysis but which did not replicate (**Supplementary Table 11**).

We then investigated whether there was heterogeneity of effect between South Asians and Europeans. Among the 42 loci reported to be associated with T2D, 37 showed consistent direction of effect in South Asians and Europeans ($P = 2.8 \times 10^{-8}$) and 27 were associated with T2D at $P < 0.05$ in South Asians only (**Supplementary Table 12**). Only three loci showed statistical evidence for heterogeneity of effect at $P < 0.001$ ($P < 0.05$ after correction for multiple testing). These findings support the view that the effects of common variants are largely shared between populations[6,15–17] and lend further support to the strategy of prioritizing SNPs from the South Asian GWAS through a combined analysis with results from Europeans. There was also no evidence for heterogeneity of effect between South Asians and Europeans at the six new loci discovered in the present study, including those carried forward for further testing based on GWAS results from South Asians only (**Table 1**).

Finally, we compared the LD structure of the six loci in South Asians and Europeans (**Supplementary Table 13**). There was some evidence for different pairwise LD between the two populations at the *VPS26A* locus, however, these differences were small (**Supplementary Fig. 6**). At the *GRB14*, *ST6GAL1*, *AP3S2*, *HMG20A* and *HNF4A* loci, the haplotype structures were similar and there was no evidence for differences in pairwise LD between South Asians and Europeans.

We identify common variants at the six loci newly associated with T2D in individuals of South Asian ancestry. Regional plots of directly genotyped SNPs are shown in **Figure 3**. Results for genotyped and imputed SNPs are shown in **Supplementary Figures 7** and **8**, and the top ten SNPs (genotyped or imputed) at each locus are listed in **Supplementary Table 14**. At the *GRB14*, *VPS26A* and *HMG20A* loci, the genotyped SNP is that most strongly associated with T2D. At the *ST6GAL1*, *AP3S2* and *HNF4A* loci, there were imputed SNPs which had stronger *P* values than those for the directly genotyped SNPs. For the *STGAL1* and *AP3S2* loci, the sentinel genotyped SNP is in perfect LD with the lead imputed SNP. At *HNF4A*, the sentinel genotyped SNP (rs4812829) and the lead imputed SNP (rs4812831) are in partial LD ($r^2 = 0.43$). However, these LD estimates are calculated from HapMap European CEU data, so they may not be applicable to South Asians.

At 2q24, rs3923113 is closest to *GRB14*, which is a strong candidate for the observed association. GRB14 is an adaptor protein that binds

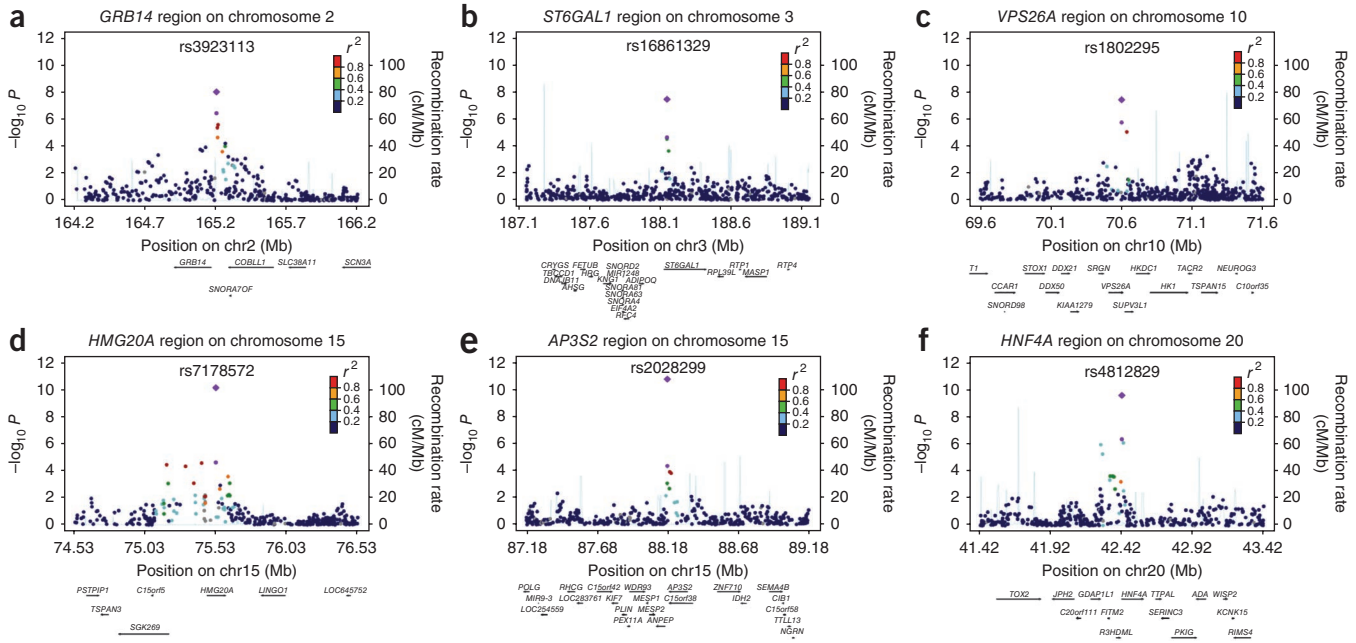


Figure 3 Regional plots for the six loci associated with type 2 diabetes in South Asians. Shown are the *GRB14* region (**a**), *ST6GAL1* region (**b**), *VPS26A* region (**c**), *HMG20A* region (**d**), *AP3S2* region (**e**) and *HNF4A* region (**f**). For the sentinel SNP, a purple circle represents the genome-wide result, and a purple diamond represents the combined analysis with the replication studies. Other genotyped SNPs are color coded according to their pairwise LD with the sentinel SNP calculated in a representative sample of 83 South Asians from the Singapore Genome Variation Project[36]. The recombination rates estimated from the HapMap2 combined panels are shown. Regional plots also incorporating SNPs imputed from HapMap2 are shown in **Supplementary Figures 7** and **8**.

to insulin receptors and insulin-like growth-factor receptors to inhibit tyrosine kinase signaling[18,19]. *Grb14*$^{-/-}$ mice have higher lean mass, better glucose homeostasis despite lower insulin and improved insulin sensitivity[20] compared to wild type. The risk allele of rs3923113 is associated with T2D and with reduced insulin sensitivity in South Asians, suggesting a gain of function (**Supplementary Table 6**). rs3923113 is also in high LD ($r^2 = 0.85$) with rs10195252 (**Supplementary Table 9**), which was recently reported to be associated with central adiposity and expression of *GRB14* in adipose tissue[14].

At 3q27, rs16861329 is intronic in *ST6GAL1*, which encodes an enzyme predominantly located in the Golgi apparatus. ST6GAL1 is involved in post-translational modification of cell-surface components by glycosylation. Although *ST6GAL1* has not previously been linked with glucose metabolism or T2D, glycosylation through addition of sialic acid residues is reported to influence both insulin action and cell surface trafficking[21]. rs16861329 is also near *ADIPOQ*, which encodes adiponectin, a hormone secreted by adipocytes that promotes insulin sensitivity. Adiponectin knockout mice show severe insulin resistance[22], and previous candidate gene studies have suggested an association of genetic variation at *ADIPOQ* with adiponectin levels, obesity and T2D[23]. However, these associations have not been consistently found, and rs16861329 (identified in the present study) is not in LD ($r^2 < 0.1$) with reported *ADIPOQ* variants.

At 10q22, rs1802295 is in *VPS26A*, which encodes a component of the retromer complex, a multimeric protein involved in transport of proteins from endosomes to the trans-Golgi network[24,25]. *VPS26A* is expressed in pancreatic, adipose and other tissues[26], but a relationship with glucose metabolism or T2D has not been described. At 15q26, rs2028299 is closest to *AP3S2*, which encodes a clathrin-associated adaptor complex expressed in adipocytes, pancreatic islets and other tissues that may be involved in vesicle transport and sorting[27]. rs2028299 is associated with expression of *C15orf38*, which encodes a member of an uncharacterized family of proteins. Among the other genes at this locus, *PLIN1* is also a possible candidate for the association with T2D. *PLIN1* encodes Perilipin-1, a phosphoprotein that coats fat droplets in adipocytes and regulates lipolysis by hormone-sensitive lipase[28]. Genetic variation at *PLIN1* has been associated with obesity in humans and in experimental animal models[29,30]. rs2028299 is 1.2 Mb from rs8042680 in *PRC1*, which is associated with T2D in Europeans[11]. Although these two SNPs are not in LD ($r^2 = 0$), we cannot exclude the possibility of a shared mechanism between the two through remote regulatory effects. At 15q24, known biology does not identify any compelling candidates. rs7178572 is intronic in *HMG20A*, which encodes a non-histone chromosomal protein that is widely expressed and which may influence histone methylation and be involved in neuronal development[31,32].

At 20q13, the lead genotyped SNP, rs4812829, is intronic in *HNF4A*, a strong candidate for the observed association. HNF4A is a nuclear transcription factor strongly expressed in liver[33] that regulates transcription of a number of genes including *HNF1A*[34]. Mutations in *HNF4A* are known to cause maturity-onset diabetes of the young type 1, which is characterized by defective pancreatic beta-cell function and impaired insulin secretion[35]. In keeping with this, the risk allele of rs4812829 is associated with reduced pancreatic beta-cell function in South Asians (**Supplementary Table 9**). At the *HNF4A* locus, there were four genotyped SNPs associated with T2D at $P < 10^{-5}$; in a conditional analysis, the effect sizes of these SNPs were substantially reduced, indicating that they are unlikely to be independent signals (**Supplementary Table 15**). An analysis of imputed data identified rs4812831 as the strongest signal at this locus (**Supplementary Table 14**), suggesting that rs4812831 may either be the causal variant or be in high LD with it. The regional plots also show a separate cluster of SNPs associated with T2D that are not in LD with rs4812829 (with the lead SNP being rs12625067), raising the possibility of two separate causal variants at this locus.

This is the first GWAS to investigate genetic factors underlying T2D in individuals of South Asian ancestry, a population with increased

© 2011 Nature America, Inc. All rights reserved.

susceptibility to T2D. We identify common genetic variants at six newly associated loci (*GRB14*, *ST6GAL1*, *VPS26A*, *AP3S2*, *HMG20A* and *HNF4A*) associated with T2D. Our findings provide insight into the genetic mechanisms underlying T2D and show the potential for new discovery from genetic association studies in populations of non-European ancestry.

URLs. METAL, http://www.sph.umich.edu/csg/abecasis/Metal/download/; PLINK, http://pngu.mgh.harvard.edu/~purcell/plink/; MuTHER resource, http://www.muther.ac.uk/; HapMap (release #28), http://hapmap.ncbi.nlm.nih.gov/; 1000 Genomes (July 2010 dataset), http://www.1000genomes.org/.

METHODS

Methods and any associated references are available in the online version of the paper at http://www.nature.com/naturegenetics/.

Note: Supplementary information is available on the Nature Genetics website.

ACKNOWLEDGMENTS

We would like to thank the many colleagues who contributed to collection and phenotypic characterization of the clinical samples as well as genotyping and analysis of the GWAS data. We would also like to acknowledge those individuals who agreed to participate in these studies. Major funding for the work described in this paper comes from Wellcome Trust awards (070854/Z/03/Z, 080747/Z/06/Z, 083270/Z/07/Z, 084723/Z/08/Z); Chennai Wellingdon Corporate Foundation; Diabetes UK (07/0003512); National Institute for Health Research Comprehensive Biomedical Research Centre at Imperial College Healthcare NHS Trust; British Heart Foundation (SP/04/002); Medical Research Council (G0700931); National Institute for Health Research (RP-PG-0407-10371); US National Institutes of Health (DK-25446); KAKENHI (Grant-in-Aid for Scientific Research from the Ministry of Education, Culture, Sports, Science and Technology of Japan; National Center for Global Health and Medicine; US National Institutes of Health (KO1TW006087); National Institute of Diabetes and Digestive and Kidney Diseases (R01DK082766); A*STAR Biomedical Research Council (05/1/21/19/425); Biomedical Research Council Singapore (09/1/35/19/616, 08/1/35/19/550); National Medical Research Council Singapore (NMRC/STaR/0003/2008, 1174/2008); National Science Foundation of Sri Lanka; and Oxford NIHR Biomedical Research Centre. A full list of acknowledgments is provided in the **Supplementary Note**.

AUTHOR CONTRIBUTIONS

Manuscript preparation: J.S.K., D.S., X.S., J. Sehmi, W.Z., P.E., Y.Y.T., M.I.M., J.D., E.S.T. and J.C.C. wrote the manuscript. All authors read and provided critical comment on the manuscript. **Data collection and analysis in the participating studies:** COBRA study: T.J., M.I. and T.M.F. Chennai Urban Rural Epidemiology Study: V.R., M. Chidambaram, S.L. and V.M. Diabetes Genetics in Pakistan and UK Asian Diabetes Studies: S.D.R., A.B., Z.I.H., A.S.S., A.H.B. and M.A.K. London Life Sciences Population Study: J.S.K., W.Z., J. Sehmi, X.L., D.D., G.R.A., J. Scott, M. Caulfield, P. Froguel, P.E., M.I.M. and J.C.C. Mauritius study: J.B.M.J., S.K., M.M.K. and P.Z.Z. Pakistan Risk of Myocardial Infarction Study: D.S., P. Frossard, R.Y., A.R., M. Samuel, N.S., P.D. and J.D. Ragama Health Study: N.K., F.T., A.R.W. and J.M.P. Singapore Consortium of Cohort Studies: K.-S.C., W.-Y.L., C.-C.K., J. Liu and E.S.T. Sikh Diabetes Study: L.F.B. and D.K.S. Singapore Indian Eye Study: X.S., C.S., T.A., T.Y.W., M. Seielstad, Y.Y.T. and E.S.T. Sri Lankan Diabetes Study: N.H., I.P., D.R.M., P.K. and M.I.M. **Sequencing of T2D loci:** M.Y., F.Z., J. Liang, X.L., J.S.K. and J.C.C. **Association results among Europeans in DIAGRAM:** A.P.M. and M.I.M. **eQTL analyses in MuTHER:** A.S.D., E.G., Å.K.H., A.C.N., K.S.S. and M.I.M.

COMPETING FINANCIAL INTERESTS

The authors declare no competing financial interests.

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

1. Chambers, J.C. *et al.* Plasma homocysteine concentrations and risk of coronary heart disease in UK Indian Asian and European men. *Lancet* **355**, 523–527 (2000).
2. Ramachandran, A., Ma, R.C. & Snehalatha, C. Diabetes in Asia. *Lancet* **375**, 408–418 (2010).
3. Shaw, J.E., Sicree, R.A. & Zimmet, P.Z. Global estimates of the prevalence of diabetes for 2010 and 2030. *Diabetes Res. Clin. Pract.* **87**, 4–14 (2010).
4. McCarthy, M.I. Genomics, type 2 diabetes, and obesity. *N. Engl. J. Med.* **363**, 2339–2350 (2010).
5. Jowett, J.B. *et al.* Genetic influences on type 2 diabetes and metabolic syndrome related quantitative traits in Mauritius. *Twin Res. Hum. Genet.* **12**, 44–52 (2009).
6. Chambers, J.C. *et al.* Genetic variation in SCN10A influences cardiac conduction. *Nat. Genet.* **42**, 149–152 (2010).
7. Saleheen, D. *et al.* The Pakistan Risk of Myocardial Infarction Study: a resource for the study of genetic, lifestyle and other determinants of myocardial infarction in South Asia. *Eur. J. Epidemiol.* **24**, 329–338 (2009).
8. Lavanya, R. *et al.* Methodology of the Singapore Indian Chinese Cohort (SICC) eye study: quantifying ethnic variations in the epidemiology of eye diseases in Asians. *Ophthalmic Epidemiol.* **16**, 325–336 (2009).
9. Price, A.L. *et al.* Principal components analysis corrects for stratification in genome-wide association studies. *Nat. Genet.* **38**, 904–909 (2006).
10. Sladek, R. *et al.* A genome-wide association study identifies novel risk loci for type 2 diabetes. *Nature* **445**, 881–885 (2007).
11. Voight, B.F. *et al.* Twelve type 2 diabetes susceptibility loci identified through large-scale association analysis. *Nat. Genet.* **42**, 579–589 (2010).
12. Kahn, S.E., Hull, R.L. & Utzschneider, K.M. Mechanisms linking obesity to insulin resistance and type 2 diabetes. *Nature* **444**, 840–846 (2006).
13. 1000 Genomes Project Consortium. *et al.* A map of human genome variation from population-scale sequencing. *Nature* **467**, 1061–1073 (2010).
14. Heid, I.M. *et al.* Meta-analysis identifies 13 new loci associated with waist-hip ratio and reveals sexual dimorphism in the genetic basis of fat distribution. *Nat. Genet.* **42**, 949–960 (2010).
15. Chambers, J.C. *et al.* Common genetic variation near MC4R is associated with waist circumference and insulin resistance. *Nat. Genet.* **40**, 716–718 (2008).
16. Kooner, J.S. *et al.* Genome-wide scan identifies variation in MLXIPL associated with plasma triglycerides. *Nat. Genet.* **40**, 149–151 (2008).
17. 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.* **43**, 339–344 (2011).
18. Dufresne, A.M. & Smith, R.J. The adapter protein GRB10 is an endogenous negative regulator of insulin-like growth factor signaling. *Endocrinology* **146**, 4399–4409 (2005).
19. Depetris, R.S., Wu, J. & Hubbard, S.R. Structural and functional studies of the Ras-associating and pleckstrin-homology domains of Grb10 and Grb14. *Nat. Struct. Mol. Biol.* **16**, 833–839 (2009).
20. Holt, L.J. *et al.* Dual ablation of Grb10 and Grb14 in mice reveals their combined role in regulation of insulin signaling and glucose homeostasis. *Mol. Endocrinol.* **23**, 1406–1414 (2009).
21. Woodard-Grice, A.V., McBrayer, A.C., Wakefield, J.K., Zhuo, Y. & Bellis, S.L. Proteolytic shedding of ST6Gal-I by BACE1 regulates the glycosylation and function of α4β1 integrins. *J. Biol. Chem.* **283**, 26364–26373 (2008).
22. Maeda, N. *et al.* Diet-induced insulin resistance in mice lacking adiponectin/ACRP30. *Nat. Med.* **8**, 731–737 (2002).
23. Siitonen, N. *et al.* Association of ADIPOQ gene variants with body weight, type 2 diabetes and serum adiponectin concentrations: the Finnish Diabetes Prevention Study. *BMC Med. Genet.* **12**, 5 (2011).
24. Seaman, M.N., Marcusson, E.G., Cereghino, J.L. & Emr, S.D. Endosome to Golgi retrieval of the vacuolar protein sorting receptor, Vps10p, requires the function of the VPS29, VPS30, and VPS35 gene products. *J. Cell Biol.* **137**, 79–92 (1997).
25. Seaman, M.N., Harbour, M.E., Tattersall, D., Read, E. & Bright, N. Membrane recruitment of the cargo-selective retromer subcomplex is catalysed by the small GTPase Rab7 and inhibited by the Rab-GAP TBC1D5. *J. Cell Sci.* **122**, 2371–2382 (2009).
26. Kim, E. *et al.* Identification of novel retromer complexes in the mouse testis. *Biochem. Biophys. Res. Commun.* **375**, 16–21 (2008).
27. Dell'Angelica, E.C. *et al.* AP-3: an adaptor-like protein complex with ubiquitous expression. *EMBO J.* **16**, 917–928 (1997).
28. Brasaemle, D.L. *et al.* Perilipin A increases triacylglycerol storage by decreasing the rate of triacylglycerol hydrolysis. *J. Biol. Chem.* **275**, 38486–38493 (2000).
29. Qi, L. *et al.* Genetic variation at the perilipin (PLIN) locus is associated with obesity-related phenotypes in White women. *Clin. Genet.* **66**, 299–310 (2004).
30. Beller, M. *et al.* PERILIPIN-dependent control of lipid droplet structure and fat storage in *Drosophila*. *Cell Metab.* **12**, 521–532 (2010).
31. Sumoy, L. *et al.* HMG20A and HMG20B map to human chromosomes 15q24 and 19p13.3 and constitute a distinct class of HMG-box genes with ubiquitous expression. *Cytogenet. Cell Genet.* **88**, 62–67 (2000).
32. Artegiani, B. *et al.* The interaction with HMG20a/b proteins suggests a potential role for β-dystrobrevin in neuronal differentiation. *J. Biol. Chem.* **285**, 24740–24750 (2010).
33. Battle, M.A. *et al.* Hepatocyte nuclear factor 4α orchestrates expression of cell adhesion proteins during the epithelial transformation of the developing liver. *Proc. Natl. Acad. Sci. USA* **103**, 8419–8424 (2006).
34. Harries, L.W., Brown, J.E. & Gloyn, A.L. Species-specific differences in the expression of the HNF1A, HNF1B and HNF4A genes. *PLoS ONE* **4**, e7855 (2009).
35. Yamagata, K. *et al.* Mutations in the hepatocyte nuclear factor-4α gene in maturity-onset diabetes of the young (MODY1). *Nature* **384**, 458–460 (1996).
36. Teo, Y.Y. *et al.* Singapore Genome Variation Project: a haplotype map of three Southeast Asian populations. *Genome Res.* **19**, 2154–2162 (2009).

© 2011 Nature America, Inc. All rights reserved.

[1]National Heart and Lung Institute (NHLI), Imperial College London, Hammersmith Hospital, London, UK. [2]Ealing Hospital National Health Service (NHS) Trust, Middlesex, UK. [3]Imperial College Healthcare NHS Trust, London, UK. [4]Center for Non-Communicable Diseases Pakistan, Karachi, Pakistan. [5]Department of Public Health and Primary Care, University of Cambridge, Cambridge, UK. [6]Centre for Molecular Epidemiology, National University of Singapore, Singapore. [7]Epidemiology and Biostatistics, Imperial College London, London, UK. [8]Department of Pediatrics, Section of Genetics, College of Medicine, University of Oklahoma Health Sciences Center, Oklahoma City, Oklahoma, USA. [9]Department of Epidemiology and Public Health, National University of Singapore, Singapore. [10]Department of Genetic Medicine and Development, University of Geneva Medical School, Geneva, Switzerland. [11]Wellcome Trust Centre for Human Genetics, University of Oxford, Oxford, UK. [12]Oxford Centre for Diabetes, Endocrinology & Metabolism, University of Oxford, Churchill Hospital, Oxford, UK. [13]Department of Community Health Sciences, Aga Khan University, Karachi, Pakistan. [14]Department of Medicine, Aga Khan University, Karachi, Pakistan. [15]Baker IDI Heart and Diabetes Institute, Melbourne, Victoria, Australia. [16]Department of Molecular Genetics, Madras Diabetes Research Foundation–Indian Council of Medical Research (ICMR) Advanced Centre for Genomics of Diabetes, Chennai, India. [17]College of Medical and Dental Sciences, University of Birmingham, Birmingham, UK. [18]BioMedical Research Centre, Heart of England NHS Foundation Trust, Birmingham, UK. [19]Department of Gene Diagnostics and Therapeutics, Research Institute, National Center for Global Health and Medicine, Tokyo, Japan. [20]Department of Ophthalmology, National University of Singapore, Singapore. [21]Singapore Eye Research Institute, Singapore National Eye Centre, Singapore. [22]Baqai Institute of Diabetology and Endocrinology, Karachi, Pakistan. [23]Department of Twin Research and Genetic Epidemiology, King's College London, London, UK. [24]Genome Institute of Singapore, Agency for Science, Technology and Research, Singapore. [25]Ministry of Health, Port Louis, Mauritius. [26]Department of Public Health, Faculty of Medicine, University of Kelaniya, Ragama, Sri Lanka. [27]Diabetic Association Pakistan, Karachi, Pakistan. [28]Wellcome Trust Sanger Institute, Wellcome Trust Genome Campus, Hinxton, UK. [29]Center for Eye Research Australia, University of Melbourne, Melbourne, Victoria, Australia. [30]Beijing Genomics Institute, Shenzhen, China. [31]A list of members is provided in the **Supplementary Note**. [32]Center for Statistical Genetics, Department of Biostatistics, University of Michigan School of Public Health, Ann Arbor, Michigan, USA. [33]Clinical Pharmacology and Barts and the London Genome Centre, William Harvey Research Institute, Barts and the London School of Medicine, Queen Mary University of London, London, UK. [34]Human Genetics, Wellcome Trust Sanger Institute, Wellcome Trust Genome Campus, Hinxton, UK. [35]Genetics of Complex Traits, Institute of Biomedical and Clinical Science, Peninsula Medical School, University of Exeter, Exeter, UK. [36]Genomics of Common Diseases, School of Public Health, Imperial College London, Hammersmith Hospital, London, UK. [37]Diabetes Research Unit, Department of Clinical Medicine, University of Colombo, Colombo, Sri Lanka. [38]Dr Mohan's Diabetes Specialties Centre, Chennai, India. [39]Institute of Human Genetics, University of California, San Francisco, California, USA. [40]Medical Research Council (MRC)-Health Protection Agency (HPA) Centre for Environment and Health, Imperial College London, London, UK. [41]Department of Statistics and Applied Probability, National University of Singapore, Singapore. [42]National University of Singapore Graduate School for Integrative Science and Engineering, National University of Singapore, Singapore. [43]Oxford National Institute for Health Research (NIHR) Biomedical Research Centre, Churchill Hospital, Oxford, UK. [44]Department of Medicine, National University of Singapore, Singapore. [45]Duke-National University of Singapore Graduate Medical School, Singapore. [46]These authors contributed equally to this work. Correspondence should be addressed to J.C.C. (john.chambers@ic.ac.uk) or J.S.K. (j.kooner@imperial.ac.uk).

© 2011 Nature America, Inc. All rights reserved.

ONLINE METHODS

Participants. *South Asian T2D cases and controls*. Genome-wide association was carried out among 5,561 South Asian cases with T2D and 14,458 South Asian controls from the London Life Sciences Population (LOLIPOP) study, the Pakistan Risk of Myocardial Infarction Study (PROMIS) and the Singapore Indian Eye (SINDI) study. Replication testing among South Asians was carried out in 13,170 cases and 25,398 control participants from the following studies: the Chennai Urban Rural Epidemiology Study (CURES)[37], the COBRA study[38], the Diabetes Genetics in Pakistan (DGP) and UK Asian Diabetes Study (UKADS)[39,40], the Mauritius study[41], the Ragama Health Study (RHS)[42], the Sikh Diabetes Study (SDS)[43], the Singapore Consortium of Cohort Studies (SCCS), the Sri Lankan Diabetes Study (SLDS)[44], and the LOLIPOP and PROMIS studies. For LOLIPOP and PROMIS, there was no overlap of participants between the genome-wide association and replication testing stages. Full details of the contributing cohorts are provided in the **Supplementary Note** along with the characteristics of the participants (**Supplementary Tables 1** and **4**).

European T2D cases and controls. Associations of SNPs with T2D in Europeans were tested *in silico* using results from the GWAS phase of the DIAGRAM+ study, which comprises 8,130 cases with T2D and 38,987 controls of European ancestry[11]. T2D case or control status was defined using study-specific criteria. SNP associations were tested using an additive genetic model and combined across studies by inverse variance meta-analysis using a fixed effects model.

Genotyping, quality control and statistical methods. Genome-wide association scans were performed using Illumina Infinium BeadChips, and genotypes were called using GenCall or Illuminus algorithms (**Supplementary Table 1**). Samples with a SNP call rate <95% were removed, as were SNPs with call rate <97%, minor allele frequency <1%, or Hardy-Weinberg equilibrium $P < 1.0 \times 10^{-6}$. Hidden relatedness or duplicate samples were sought using identity-by-descent methods implemented in PLINK, and individuals with evidence for relatedness were excluded (pi_hat ≥ 0.5 in LOLIPOP and SINDI or pi_hat ≥ 0.37 in PROMIS to allow for the higher prevalence of consanguinuity in Pakistanis). In the absence of a South Asian–specific haplotype map, the primary analysis tested the association with T2D of the 568,976 autosomal SNPs that had been directly genotyped and passed quality control testing among the South Asian GWAS participants. As secondary analyses and to help fine map the loci identified, we repeated the GWAS analyses after imputation of untyped SNPs from HapMap2. Imputation of genotypes not directly measured was performed using pooled haplotypes from the CEU, YRI and CHB/JPT HapMap2 reference panels and the IMPUTE2 software package as previously described[45].

Principal components analysis was used to identify population outliers by comparison to reference samples from the HapMap YRI, CHB, JPT and CEU panels and the Indian samples collected by Reich and colleagues[46]; samples with eigenvalues inconsistent with South Asian ancestry were removed. Principal components analysis was performed in Eigensoft v3.0 (ref. 9) using a set of 100,864 SNPs common to all three studies and pruned to reduce pairwise LD—this SNP set was selected using the 'indep' option of PLINK (see URLs) with window size 50, step 3 and VIF = 1.3.

Associations of SNPs with T2D were tested using logistic regression and an additive genetic model. Analyses were carried out in men and women separately. Principal components were included as covariates to adjust for residual population stratification; no other covariate adjustments were made. The number of principal components included was study specific: for LOLIPOP we included ten, for PROMIS we included five and for SINDI we included three principal components. A fixed effects inverse variance meta-analysis was used to combine the results for the individual studies. The *P* values were adjusted for the study-level genomic control inflation factor before the meta-analysis and then again for the meta-analysis genomic control inflation factor (that is, a double genomic control correction)[47].

The primary analysis comprised a meta-analysis of the association results for men and women combined at the 568,976 SNPs that had been directly genotyped. Secondary analyses included the following five GWAS meta-analyses: (i) male-gender specific; (ii) female-gender specific; (iii) BMI adjusted; (iv) lean cases (BMI < 25 kg/m^2) compared to overweight controls (BMI > 25 kg/m^2); and (v) analysis of GWAS data imputed with missing genotypes from the HapMap2 reference panel (for a total of 2,646,472 SNPs).

Selection for replication testing. SNPs that were located within a locus that had been previously reported to be associated with T2D were excluded, as were SNPs in LD ($r^2 > 0.5$) with another SNP that had a more significant association with T2D.

Twenty SNPs were carried forward for replication testing, comprising: (i) all SNPs associated with T2D at $P \leq 10^{-5}$ in the primary South Asian–specific genome-wide association analysis of directly genotyped SNPs ($N = 7$); (ii) all SNPs associated with T2D at $P < 5 \times 10^{-8}$ in one of the five secondary GWAS analyses ($N = 1$, identified from the female-gender–specific analysis); and (iii) 12 SNPs prioritized from among the 43 SNPs associated with T2D at $P > 10^{-5}$ and $P \leq 10^{-4}$ in the primary analysis based on the lowest *P* value in a fixed effects inverse variance meta-analysis with results from the GWAS stage of the DIAGRAM+ study (corresponding to ~$P < 10^{-3}$ in the combined analysis)[11].

Replication testing. Genotyping of the replication samples was performed by KASPAR (K-Bioscience Ltd), Sequenom MassArray or TaqMan assays (**Supplementary Table 4**). Samples with <90% call rate were excluded, as were SNPs with call rate <95% or that deviated from Hardy-Weinberg equilibrium at $P < 2.5 \times 10^{-3}$. The associations of SNPs with T2D were tested in each cohort separately; heterogeneity between studies was assessed using a Cochran's *Q* statistic. A fixed effects meta-analysis was then used to combine the results for each SNP across all replication studies with available data and then in a combined analysis with results from the genome-wide association stage. Statistical significance was inferred at $P < 2.5 \times 10^{-3}$ in the replication stage ($P < 0.05$ after Bonferroni correction for 20 SNPs). For the combined analysis of genome-wide and replication data, genome-wide significance was inferred at $P < 5 \times 10^{-8}$.

Power. For SNPs with minor allele frequency >20%, the study had >80% power to identify SNPs with odds ratios for T2D of OR > 1.11 per allele copy at $P < 10^{-4}$ in the genome-wide association phase, OR > 1.06 per allele copy at $P < 2.5 \times 10^{-3}$ in the replication testing stage and OR > 1.08 per allele copy at $P < 5 \times 10^{-8}$ in combined analysis.

Sequencing. Sequencing was performed using a Genome Analyzer II platform (Illumina) at the Beijing Genomics Institute with library preparation and a 91-bp paired-end sequencing strategy according to manufacturer's instructions. Reads were aligned to the human reference genome (NCBI Build 36) using Burrows-Wheeler alignment (BWA)[48], duplicates were removed using SAMtools and the genotype likelihood was calculated by SOAPsnp[49]. We sequenced 109 South Asian samples (including 34 samples with T2D), and the average sequencing depth was ~4×.

Comparison of regional LD patterns. We used the varLD algorithm to compare the regional pattern of LD surrounding the index SNPs for each of the six regions[50]. The use of the targeted varLD algorithm tests the null hypothesis that the regional pattern of correlation between every pair of SNPs in the window is identical across two populations and yields a Monte Carlo *P* value that effectively quantifies the statistical evidence of a deviation from this identity. In our comparisons, we implemented 1,000 iterations for the Monte Carlo procedure across a 300-kb window around the associated SNPs in each region. Each analysis compares two of the following populations: (i) the 60 Europeans in phase 2 of HapMap (CEU); (ii) the 83 South Asians from the Singapore Genome Variation Project[36]; and (iii) 60 randomly chosen control samples from the LOLIPOP study.

Expression QTLs. To determine whether the T2D risk variants detected in this study influenced expression of nearby genes, we accessed a variety of sources, including: (i) publicly available *cis*-eQTL data for brain[51,52], lymphoblastoid cell lines (LCLs)[53–55], fibroblasts[55], liver[56] and T cells[55]; (ii) expression data from the MuTHER consortium[57]; and (iii) *cis*-eQTL data for *GRB14* reported in a recent GWAS of fat distribution[14].

The MuTHER resource (see URLs) includes LCLs, skin and adipose tissue derived simultaneously from a subset of well phenotyped healthy female twins[57]. Whole-genome expression profiling of the samples, each with either two or three technical replicates, was performed using the Illumina Human HT-12 V3 BeadChips (Illumina Inc) according to the protocol supplied by

© 2011 Nature America, Inc. All rights reserved.

doi:10.1038/ng.921

the manufacturer. Log2-transformed expression signals were normalized separately per tissue as follows: quantile normalization was performed across technical replicates of each individual followed by quantile normalization across all individuals. Genotyping was done with a combination of Illumina arrays (HumanHap300, HumanHap610Q, 1MDuo and 1.2MDuo). Untyped HapMap2 SNPs were imputed using the IMPUTE software package (v2). The number of samples with genotypes and expression values per tissue was 778 in LCLs, 667 in skin and 776 in adipose. Association between all SNPs (minor allele frequency >5% and IMPUTE info >0.8) within a gene or within 1 Mb of the gene transcription start or end site and normalized expression values were performed with the GenABEL and ProbABEL packages using the polygenic linear model incorporating a kinship matrix in GenABEL followed by the ProbABEL mmscore score test with imputed genotypes. Age and experimental batch were included as cofactors in the adipose and LCL analysis, and age, experimental batch and concentration were included as cofactors in the skin analysis.

37. Chidambaram, M., Radha, V. & Mohan, V. Replication of recently described type 2 diabetes gene variants in a South Indian population. *Metabolism* **59**, 1760–1766 (2010).
38. Jafar, T.H. *et al.* Community-based interventions to promote blood pressure control in a developing country: a cluster randomized trial. *Ann. Intern. Med.* **151**, 593–601 (2009).
39. Bellary, S. *et al.* Enhanced diabetes care to patients of South Asian ethnic origin (the United Kingdom Asian Diabetes Study): a cluster randomised controlled trial. *Lancet* **371**, 1769–1776 (2008).
40. Rees, S.D. *et al.* An FTO variant is associated with type 2 diabetes in South Asian populations after accounting for body mass index and waist circumference. *Diabet. Med.* **28**, 673–680 (2011).
41. Söderberg, S. *et al.* High incidence of type 2 diabetes and increasing conversion rates from impaired fasting glucose and impaired glucose tolerance to diabetes in Mauritius. *J. Intern. Med.* **256**, 37–47 (2004).
42. Takeuchi, F. *et al.* Common variants at the GCK, GCKR, G6PC2-ABCB11 and MTNR1B loci are associated with fasting glucose in two Asian populations. *Diabetologia* **53**, 299–308 (2010).
43. Sanghera, D.K. *et al.* The Khatri Sikh Diabetes Study (SDS): study design, methodology, sample collection, and initial results. *Hum. Biol.* **78**, 43–63 (2006).
44. Katulanda, P. *et al.* Prevalence and projections of diabetes and pre-diabetes in adults in Sri Lanka–Sri Lanka Diabetes, Cardiovascular Study (SLDCS). *Diabet. Med.* **25**, 1062–1069 (2008).
45. 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.* **39**, 906–913 (2007).
46. Reich, D., Thangaraj, K., Patterson, N., Price, A.L. & Singh, L. Reconstructing Indian population history. *Nature* **461**, 489–494 (2009).
47. Bacanu, S.A., Devlin, B. & Roeder, K. The power of genomic control. *Am. J. Hum. Genet.* **66**, 1933–1944 (2000).
48. Li, H. & Durbin, R. Fast and accurate long-read alignment with Burrows-Wheeler transform. *Bioinformatics* **26**, 589–595 (2010).
49. Li, R. *et al.* SNP detection for massively parallel whole-genome resequencing. *Genome Res.* **19**, 1124–1132 (2009).
50. Ong, R.T. & Teo, Y.Y. varLD: a program for quantifying variation in linkage disequilibrium patterns between populations. *Bioinformatics* **26**, 1269–1270 (2010).
51. Myers, A.J. *et al.* A survey of genetic human cortical gene expression. *Nat. Genet.* **39**, 1494–1499 (2007).
52. Webster, J.A. *et al.* Genetic control of human brain transcript expression in Alzheimer disease. *Am. J. Hum. Genet.* **84**, 445–458 (2009).
53. Dixon, A.L. *et al.* A genome-wide association study of global gene expression. *Nat. Genet.* **39**, 1202–1207 (2007).
54. Ge, B. *et al.* Global patterns of cis variation in human cells revealed by high-density allelic expression analysis. *Nat. Genet.* **41**, 1216–1222 (2009).
55. Dimas, A.S. *et al.* Common regulatory variation impacts gene expression in a cell type-dependent manner. *Science* **325**, 1246–1250 (2009).
56. Schadt, E.E. *et al.* Mapping the genetic architecture of gene expression in human liver. *PLoS Biol.* **6**, e107 (2008).
57. Nica, A.C. *et al.* The architecture of gene regulatory variation across multiple human tissues: the MuTHER study. *PLoS Genet.* **7**, e1002003 (2011).

© 2011 Nature America, Inc. All rights reserved.

Genome-Wide Association Study Identifies a Novel Locus Contributing to Type 2 Diabetes Susceptibility in Sikhs of Punjabi Origin From India

Richa Saxena,[1] Danish Saleheen,[2,3,4] Latonya F. Been,[5] Martha L. Garavito,[5] Timothy Braun,[5] Andrew Bjonnes,[1] Robin Young,[3] Weang Kee Ho,[3] Asif Rasheed,[2] Philippe Frossard,[2] Xueling Sim,[6,7] Neelam Hassanali,[8] Venkatesan Radha,[9] Manickam Chidambaram,[9] Samuel Liju,[9] Simon D. Rees,[10,11] Daniel Peng-Keat Ng,[12] Tien-Yin Wong,[13,14,15] Toshimasa Yamauchi,[16] Kazuo Hara,[16,17] Yasushi Tanaka,[18] Hiroshi Hirose,[19] Mark I. McCarthy,[8,20,21] Andrew P. Morris,[20] DIAGRAM,* MuTHER,* AGEN,* Abdul Basit,[22] Anthony H. Barnett,[11] Prasad Katulanda,[8,23] David Matthews,[8] Viswanathan Mohan,[9,24] Gurpreet S. Wander,[25] Jai Rup Singh,[26] Narinder K. Mehra,[27] Sarju Ralhan,[25] M. Ilyas Kamboh,[28] John J. Mulvihill,[5] Hiroshi Maegawa,[29] Kazuyuki Tobe,[30] Shiro Maeda,[31] Yoon S. Cho,[32] E. Shyong Tai,[12,33,34] M. Ann Kelly,[10,11] John C. Chambers,[35,36,37] Jaspal S. Kooner,[35,36,38] Takashi Kadowaki,[16] Panos Deloukas,[39] Daniel J. Rader,[40] John Danesh,[3] and Dharambir K. Sanghera[5]

We performed a genome-wide association study (GWAS) and a multistage meta-analysis of type 2 diabetes (T2D) in Punjabi Sikhs from India. Our discovery GWAS in 1,616 individuals (842 case subjects) was followed by in silico replication of the top 513 independent single nucleotide polymorphisms (SNPs) ($P < 10^{-3}$) in Punjabi Sikhs (n = 2,819; 801 case subjects). We further replicated 66 SNPs ($P < 10^{-4}$) through genotyping in a Punjabi Sikh sample (n = 2,894; 1,711 case subjects). On combined meta-analysis in Sikh populations (n = 7,329; 3,354 case subjects), we identified a novel locus in association with T2D at 13q12 represented by a directly genotyped intronic SNP (rs9552911, $P = 1.82 \times 10^{-8}$) in the *SGCG* gene. Next, we undertook in silico replication (stage 2b) of the top 513 signals ($P < 10^{-3}$) in 29,157 non-Sikh South Asians (10,971 case subjects) and de novo genotyping of up to 31 top signals ($P < 10^{-4}$) in 10,817 South Asians (5,157 case subjects) (stage 3b). In combined South Asian meta-analysis, we observed six suggestive associations ($P < 10^{-5}$ to $< 10^{-7}$), including SNPs at *HMG1L1/CTCFL*, *PLXNA4*, *SCAP*, and chr5p11. Further evaluation of 31 top SNPs in 33,707 East Asians (16,746 case subjects) (stage 3c) and 47,117 Europeans (8,130 case subjects) (stage 3d), and joint meta-analysis of 128,127 individuals (44,358 case subjects) from 27 multiethnic studies, did not reveal any additional loci nor was there any evidence of replication for the new variant. Our findings provide new evidence on the presence of a population-specific signal in relation to T2D, which may provide additional insights into T2D pathogenesis. ***Diabetes* 62:1746–1755, 2013**

South Asians (people originating from the Indian subcontinent) comprise more than a quarter of the global population and contribute the highest number of patients with type 2 diabetes (T2D) (1). According to latest estimates, ~61 million people in India alone are currently afflicted with T2D, and their number is projected to increase to ~101 million by 2030 (2). Consequently, ~60% of the world's coronary artery disease

From the [1]Center for Human Genetic Research and Department of Anesthesia, Critical Care and Pain Medicine, Massachusetts General Hospital, Harvard Medical School, Boston, Massachusetts; the [2]Center for Non-Communicable Diseases, Karachi, Pakistan; the [3]Department of Public Health and Primary Care, University of Cambridge, Cambridge, U.K.; the [4]Department of Biostatistics and Department of Epidemiology and Medicine, University of Pennsylvania, Philadelphia, Pennsylvania; the [5]Department of Pediatrics, College of Medicine, University of Oklahoma Health Sciences Center, Oklahoma City, Oklahoma; the [6]Center for Statistical Genetics and Department of Statistics, University of Michigan, Ann Arbor, Michigan; the [7]Centre for Molecular Epidemiology, National University of Singapore, Singapore; the [8]Oxford Centre for Diabetes, Endocrinology and Metabolism, University of Oxford, Churchill Hospital, Oxford, U.K.; the [9]Madras Diabetes Research Foundation, Chennai, India; the [10]College of Medical and Dental Sciences, University of Birmingham, Birmingham, U.K.; the [11]Diabetes Centre, Heart of England National Health Service Foundation Trust, Birmingham, U.K.; the [12]Saw Swee Hock School of Public Health, National University of Singapore, Singapore; the [13]Singapore Eye Research Institute, Singapore National Eye Centre, Singapore; the [14]Department of Ophthalmology, National University of Singapore, Singapore; the [15]Centre for Eye Research Australia, University of Melbourne, Melbourne, Australia; the [16]Department of Diabetes and Metabolic Diseases, Graduate School of Medicine, The University of Tokyo, Tokyo, Japan; the [17]Department of Integrated Molecular Science on Metabolic Diseases, 22nd Century Medical and Research Center, The University of Tokyo, Tokyo, Japan; the [18]Department of Internal Medicine, Division of Metabolism and Endocrinology, St. Marianna University School of Medicine, Kanagawa, Japan; the [19]Health Center, Keio University School of Medicine, Tokyo, Japan; the [20]Wellcome Trust Centre for Human Genetics, University of Oxford, Oxford, U.K.; the [21]Oxford National Institute for Health Research Biomedical Research Centre, Churchill Hospital, Oxford, U.K.; the [22]Baqai Institute of Diabetology and Endocrinology, Karachi, Pakistan; the [23]Diabetes Research Unit, Department of Clinical Medicine, University of Colombo, Colombo, Sri Lanka; [24]Dr. Mohan's Diabetes Specialities Centre, Chennai, India; the [25]Hero Dayanand Medical College and Heart Institute, Ludhiana, Punjab, India; the [26]Central University of Punjab, Bathinda, Punjab, India; the [27]All India Institute of Medical Sciences and Research, New Delhi, India; the [28]Department of Human Genetics, University of Pittsburgh, Pittsburgh, Pennsylvania; the [29]Department of Medicine, Shiga University of Medical Science, Shiga, Japan; the [30]First Department of Internal Medicine, University of Toyama, Toyama, Japan; the [31]Laboratory for Endocrinology and Metabolism, RIKEN Center for Genomic Medicine, Kanagawa, Japan; the [32]Department of Biomedical Science, Hallym University, Chuncheon, Gangwondo 200-702, Republic of Korea; the [33]Department of Medicine, Yong Loo Lin School of Medicine, National University of Singapore, Singapore; the [34]Duke-NUS Graduate Medical School Singapore, Singapore; the [35]Ealing Hospital National Health Service Trust, Middlesex, U.K.; the [36]Imperial College Healthcare National Health Service Trust, London, U.K.; [37]Epidemiology and Biostatistics, Imperial College London, London, U.K.; the [38]National Heart and Lung Institute, Imperial College London, Hammersmith Hospital, London, U.K.; the [39]Wellcome Trust Sanger Institute, Hinxton, U.K.; and the [40]Department of Medicine, University of Pennsylvania, Philadelphia, Pennsylvania.

Corresponding author: Dharambir K. Sanghera, dharambir-sanghera@ouhsc.edu.

Received 13 August 2012 and accepted 22 November 2012.

DOI: 10.2337/db12-1077

This article contains Supplementary Data online at http://diabetes.diabetesjournals.org/lookup/suppl/doi:10.2337/db12-1077/-/DC1.

*Complete lists of the members of the DIAGRAM, MuTHER, and AGEN study groups can be found in the Supplementary Data online.

© 2013 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.

See accompanying commentary, p. 1369.

(CAD), a principal cause of mortality in individuals with T2D, is expected to occur in India (3). There is considerable ethnic difference in the prevalence and progression of T2D and CAD. In addition to environmental factors, genetic factors influence disease susceptibility (4). The incidence of T2D and CAD is about three to five times higher in immigrant South Asians compared with Euro-Caucasians, and the age of onset of T2D is roughly a decade earlier in South Asians than in Europeans (5–7). The higher prevalence of T2D among South Asians settled in developed countries compared with the host population reflects the genetic and ethnic predisposition to cardiometabolic disease under an adverse environment and the joint effects of genes and environment in the predisposition to T2D (8). For these reasons, we conducted ethnic-specific genetic studies in a Sikh population to dissect genetic pathways that may contribute to T2D etiology in different ethnic groups.

The vast majority of genome-wide association studies (GWAS) on T2D so far have been performed on Europeans. Studies on non-European populations, especially those with unique demographic and cultural histories, are important for identifying population-specific linkage disequilibrium (LD) patterns and environmental factors that may modulate disease risk or protection (9). Interestingly, many, but not all, of the common loci originally identified in Europeans have been replicated in non-European groups (10–18). Recent GWAS in non-European populations have yielded intriguing new variants (19–21), including six novel signals in South Asians represented by single nucleotide polymorphisms (SNPs) near *GRB14*, *ST6GAL1*, *VPS26A*, *HMG20A*, *AP3S2*, and *HNF4A* in our recent meta-analysis of GWAS (22). Given the existence of marked genetic variability among South Asian communities, in addition to diversity in culture, language, caste system, physical appearance, and diet, they do not constitute a single homogeneous community (23). Therefore, screening populations with a different genetic and racial background or environmental exposures may improve insights about the disease and genetic risk factors (24).

People from India have a complex racial history complicated by the presence of a caste system that has prohibited interbreeding to a great extent and consequently separated people into numerous endogamous groups (25). The Sikhs, a relatively young, inbred population of ~26 million (2% of the Indian population), are from the northwestern province of India and follow a distinct and unique religion born ~500 years ago in Punjab. They have an interesting background for "nontraditional" disease enrichment in the absence of conventional risk factors such as smoking, obesity, and a diet rich in meats (26). Sikhs do not smoke or chew tobacco because of religious and cultural compulsions, and ~50% of them are lifelong vegetarians. Despite the absence of these lifestyle-related risk factors, T2D and CAD have reached epidemic proportions in Sikhs. Our initial genetic studies in a Sikh cohort as part of the Asian Indian Diabetic Heart Study (AIDHS) or the Sikh Diabetes Study (SDS) revealed an association of *FTO* and *MTNR1B*, *ADIPOQ*, and *PPARG* polymorphisms with T2D and risk factors in the absence of obesity (11,27,28). In this investigation, we conducted a GWAS in a relatively homogenous Punjabi Sikh population of 1,850 individuals and performed multistage replication in up to 27 case-control studies of Punjabi, other South Asian, East Asian, and Caucasian ancestries (total n = 128,127; 44,358 T2D case and 83,769 control subjects) (Supplementary Tables 1 and 2). Study design of the discovery, replication, and meta-analysis phases was optimized to detect new population-specific and multiethnic T2D loci (Fig. 1). One important difference in the current study from our previous South Asian GWAS (22) is that in the previous study, the SNPs that were common between South Asians and Europeans were selected for replication based on the European Diabetes Genetics Replication and Meta-analysis (DIAGRAM) sample. However, in this study, the SNP selection was prioritized based on the top signals ($P < 10^{-3}$) from our discovery Sikh cohort.

RESEARCH DESIGN AND METHODS

Participants. Participants were part of the Punjabi Sikh GWAS.

Study sample and characteristics. Our primary Sikh GWAS (discovery) cohort used in this investigation is comprised of 1,616 individuals from the Punjabi Sikh population that was a part of the AIDHS (also named the SDS). The AIDHS/SDS has unique characteristics that are ideal for genetic studies. Sikhs are strictly a nonsmoking population, and ~50% of participants are teetotalers and life-long vegetarians. All individuals for the GWAS discovery cohort were recruited from one geographical location. Diagnosis of T2D was confirmed by scrutinizing medical records for symptoms and use of medication and measuring fasting glucose levels according to the guidelines of the American Diabetes Association (29), as described previously (11). Data on lipids, insulin, glucose, anthropometric measurements, education, socioeconomic status, job grade, diet, and physical activity were available on >95% of the AIDHS/SDS individuals selected for this study. Dietary questions involving alcohol consumption were scored using a scale from 0 to 5; details are described elsewhere (26). T2D is often asymptomatic and remains undiagnosed for many years, especially in people from the developing world due to poor healthcare provisions. Therefore, it is reasonable to assume that the actual age of onset of T2D in Sikhs may range from 39 to 42 years of age compared with the observed age at diagnosis (46 years). This age is in sharp contrast to the mean age at onset of 60 years or above in developed countries (5,26,30). A medical record indicating either *1*) a fasting plasma glucose level ≥7.0 mmol/L (≥126 mg/dL) after a minimum 12-h fast or *2*) a 2-h postglucose level of ≥11.1 mmol/L (≥200 mg/dL) estimated during a 2-h oral glucose tolerance test on more than one occasion, combined with symptoms of diabetes, confirmed the diagnosis. Impaired fasting glucose is defined as a fasting blood glucose level ≥5.6 mmol/L (≥100 mg/dL) but ≤7.0 mmol/L (≤126 mg/dL). Impaired glucose tolerance is defined as a 2-h OGTT >7.8 mmol/L (>140 mg/dL) but <11.1 mmol/L (<200 mg/dL). The 2-h OGTTs were performed according to the criteria of the World Health Organization (75-g oral load of glucose). BMI was calculated as weight (kg)/height $(m)^2$, and waist-to-hip ratio was calculated as the ratio of abdomen or waist circumference to hip circumference. Subjects with type 1 diabetes, or those with a family member with type 1 diabetes, or rare forms of T2D subtypes (maturity-onset diabetes of the young) or secondary diabetes (from, e.g., hemochromatosis or pancreatitis) were excluded from the study. The selection of control subjects was based on a fasting glucose <100.8 mg/dL or a 2-h glucose <141.0 mg/dL. Subjects with impaired fasting glucose or impaired glucose tolerance were excluded when data were analyzed for association of the variants with T2D. All blood samples were obtained at the baseline visits. All participants signed a written informed consent for the investigations. The study was reviewed and approved by the University of Oklahoma Health Sciences Center Institutional Review Board, as well as the Human Subject Protection Committee at the participating hospitals and institutes in India.

South Asian cohorts. For stage 2a replication, the Sikh component of the London Life Sciences Population (LOLIPOP) study (22) comprised 2,919 individuals (801 T2D case and 2,018 control subjects). For stage 2b, the non-Sikh South Asian components of the LOLIPOP and the Pakistan Risk of Myocardial Infarction Study (PROMIS; and the Risk Assessment of Cerebrovascular Events [RACE] study) GWAS (22) comprised 29,157 individuals (10,971 case and 18,186 control subjects) (22). Stage 3a Punjabi-specific replication was carried out on 2,894 individuals (1,711 case and 1,183 control subjects) of Punjabi ancestry from India as part of AIDHS/SDS, and replication testing among South Asians for stage 3b was carried out among 10,817 participants (5,157 case and 5,660 control subjects), which were part of the following studies: Asian Indians from the Singapore Indian Eye (SINDI) study (31), the Chennai Urban Rural Epidemiology Study (CURES) (32), the Diabetes Genetics in Pakistan (DGP) study, the UK Asian Diabetes Study (UKADS) (33), and the Sri Lankan Diabetes Study (SLDS) (34). Details of the contributing cohorts are provided in the Supplementary Data.

East Asian cohorts. Replication testing for stage 3c was carried out on a total of 33,707 East Asians, comprising 14,890 Japanese from RIKEN (n = 7,480

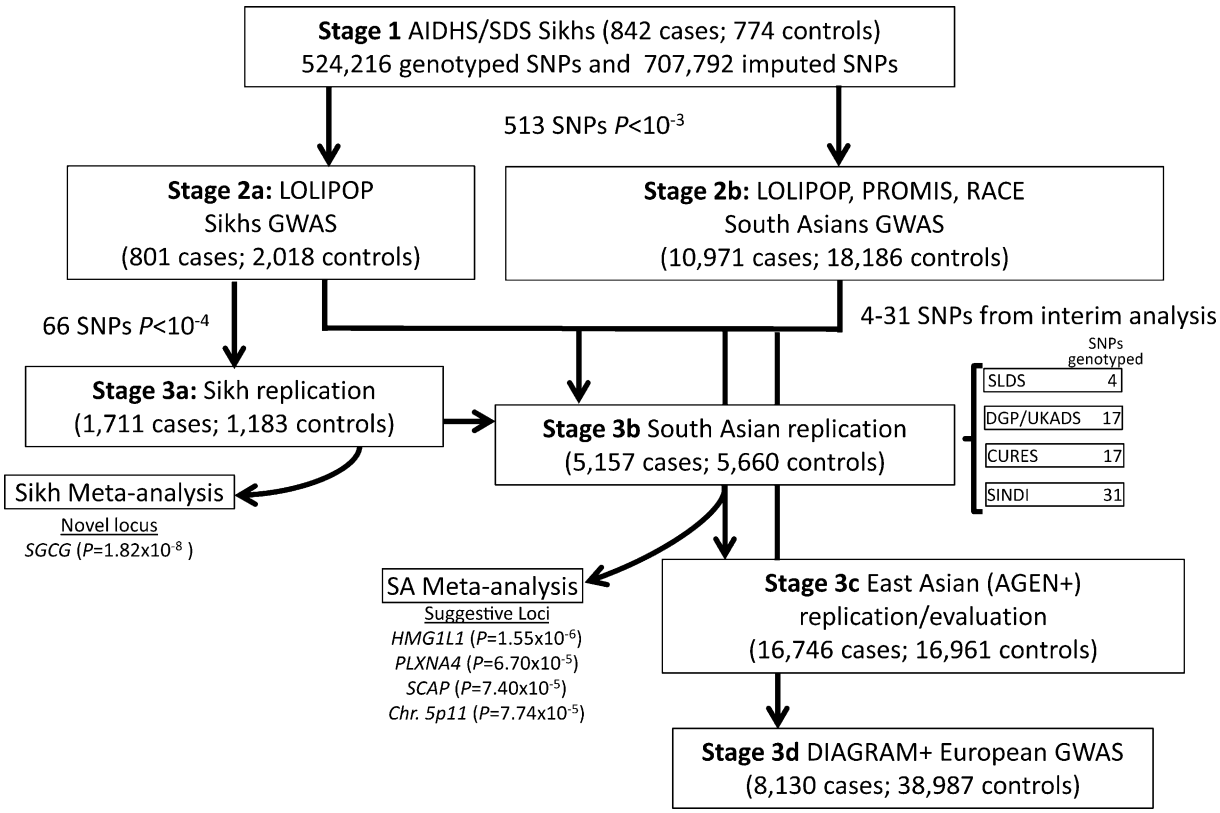


FIG. 1. Summary of study design and outcome of key findings.

genotyped) and BioBank Japan (n = 7,410 GWAS) (19,35) and 18,817 individuals of East Asian ancestry as part of the Asian Genetic Epidemiology Network (AGEN) with genotype data available from eight GWAS (21).

DIAGRAM (Euro-Caucasians). Associations of SNPs with T2D among Europeans were tested in silico using results from the genome-wide association phase of the DIAGRAM study comprising 47,117 subjects (36).

Genotyping and quality control. Genomic DNA was extracted from buffy coats using QiaAmp blood kits (Qiagen, Chatsworth, CA) or by the salting-out procedure (37). Stage 1 genome-wide genotyping was performed using a Human 660W-Quad BeadChip panel (Illumina, Inc., San Diego, CA). We performed pairwise identity-by-state clustering in PLINK across all individuals to assess population stratification; no population outliers were detected. Related individuals with pi-hat >0.3 and samples with <93% call rate were excluded, as were SNPs with call rate <95%. Also excluded were SNPs with Hardy-Weinberg equilibrium (HWE) $P < 10^{-6}$ or minor allele frequency (MAF) <1%. After quality control, 524,216 directly genotyped SNPs in 1,616 subjects (842 case and 774 control subjects) were available for association testing.

Genotyping for de novo SNPs in the replication samples was performed by Sequenom MassArray (BioMark HD MX/HX Genetic Analysis System; Fluidigm) or KASPAR (LGC Genomics KBioscience, London, U.K.). Samples and SNPs with <95% call rate were excluded, as were those that deviated from HWE at $P < 10^{-3}$. The associations of SNPs with T2D were tested in each cohort separately.

Statistical analyses

Association testing. Associations of SNPs with T2D were tested using logistic regression and an additive genetic model. Age, sex, BMI, and 5 or 10 principal components to adjust for residual population stratification were included as covariates. As the existing HapMap2 or HapMap3 and 1000 Genomes data do not include Sikhs, the 5 or 10 principal components used for this correction were estimated using our Sikh population sample and not the HapMap populations. After association analyses, the genomic control inflation factor (λ) was 1.0, so no adjustments were made (Supplementary Fig. 2*A* and *B*).

In addition to the analysis of directly genotyped SNPs, we performed imputation using the Impute 2 program (38–40), which determines the probability distribution of missing genotypes based on a set of known haplotypes and an estimated fine-scale recombination map. Imputation was based on the entire multiethnic HapMap3 reference panel of ~1.5 million autosomal SNPs with MAF >1% in 1,011 individuals from Africa, Asia, Europe, and the Americas (including 1,362,138 SNPs from the Indian population of 100 Gujaratis from Houston [GIH]). Imputation yielded a total of 1,232,008 passing SNPs with MAF >1% in the Sikh GWAS. Imputed SNPs were analyzed using SNPTEST (38,40), adjusted for the covariates age, sex, BMI, and five principal components, which implements frequentist tests that calculate P values and parameter estimates and their standard errors that account for the uncertainty due to the probability distributions of the imputed genotypes, and included only those SNPs with an information score ≥0.5 in the discovery sample as well as in all GWAS used for replication, a measure of the relative statistical information about the additive genetic effect being estimated. The genomic control value for imputed SNPs was 1.02. The inbreeding coefficient and measures of autozygosity were determined using the program PLINK. We identified runs of homozygosity using the metrics defined in Nalls et al. (41), evaluating 1-Mb autosomal regions with at least 50 adjacent SNPs, with a sliding window of 50 SNPs including no more than 2 SNPs with missing genotypes and one possible heterozygous genotype.

Stage 2 replication. We selected all independent association signals (r^2 <0.25) with $P < 10^{-3}$ for lookup in GWAS of *1*) the Sikh component of the LOLIPOP GWAS (22) and *2*) the non-Sikh South Asian components of the LOLIPOP and PROMIS GWAS (22). A fixed-effect, inverse-variance meta-analysis (as implemented in METAL) (42) was used to combine the results for individual studies.

Stage 3 replication. Significant association results with $P < 10^{-3}$ based on meta-analysis of stages 1, 2a, and 2b were selected for de novo or in silico replication in Sikh, South Asian, other Asian, and European populations. In addition, we selected SNPs from a Sikh-only meta-analysis of stages 1a and 2a for genotyping in an in-house Punjabi Sikh T2D case-control population. In our previous South Asian GWAS by Kooner et al. (22), 300 of the 3,200 samples of the AIDHS/SDS (used in replication) were genotyped using Illumina 660 Quad arrays, and the remaining samples (from 1,187 case and 1,632 control subjects) were genotyped using Sequenom MassARRAYs. However, in this study, in addition to GWAS set (n = 1,616), SNPs were genotyped de novo on our remaining replication set (n = 2,894). Signals with $P < 10^{-4}$ after meta-analysis of stages 1, 2a, and 3a were also genotyped in the South Asian, other Asian, and European populations to test if they were specific to the Sikh ethnic group or spanned ethnicities. All meta-analyses were performed using a fixed-effects, inverse-variance meta-analysis implemented in METAL.

MuTHER Consortium. The Multiple Tissue Human Expression Resource (MuTHER; www.muther.ac.uk) includes lymphoblastic cell lines and skin and adipose tissue derived simultaneously from a subset of well-phenotyped

healthy female twins from the Twins UK adult registry. Whole-genome expression profiling of the samples, each with either two or three technical replicates, was performed using the Illumina HumanHT-12 v3 BeadChips according to the protocol supplied by the manufacturer. Log_2-transformed expression signals were normalized separately per tissue as follows. Quantile normalization was performed across technical replicates of each individual followed by quantile normalization across all individuals. Genotyping was performed with a combination of Illumina arrays HumanHap300, HumanHap610Q, 1M-Duo, and 1.2MDuo 1M. Untyped HapMap2 SNPs were imputed using the IMPUTE software package (v2). The number of adipose samples with genotypes and expression values is 776. Association between all SNPs (MAF >5%; IMPUTE info >0.8) within a gene or within 1 Mb of the gene transcription start or end site and normalized expression values was performed with the GenABEL/ProbABEL packages using the polygenic linear model incorporating a kinship matrix in GenABEL followed by the ProbABEL mmscore score test with imputed genotypes. Age and experimental batch were included as cofactors.

RESULTS

Punjabi Sikh discovery GWAS. Clinical characteristics of the stage 1 Punjabi Sikh T2D GWAS cohort and stage 2a and 2b (replication) cohorts are described in Supplementary Table 3. Principal components analysis revealed little population structure (Supplementary Fig. 1). After quality control, 524,216 directly genotyped SNPs in 1,616 subjects (842 case and 774 control subjects) from 1,850 total subjects were available for association testing after removing samples showing cryptic relatedness through identity-by-descent sharing. To increase genome coverage, genotypes were imputed for untyped SNPs using the HapMap3 multiethnic reference panel (see RESEARCH DESIGN AND METHODS), yielding a total of 1,232,008 SNPs for association analyses. The reason for choosing a more cosmopolitan panel and not restricting to the GIH was based on our own data showing equal diversity of the Sikhs from GIH and CEU, and based on previously described advantages of using a worldwide reference panel (39). We performed a GWAS for T2D adjusted for covariates age, sex, BMI, and five principal components (Supplementary Fig. 1); no evidence of inflation was observed (Supplementary Fig. 2*A* and *B*) (see RESEARCH DESIGN AND METHODS).

Replication and meta-analyses in Punjabi Sikh participants. We undertook a two-stage replication in T2D case-control samples of Punjabi Sikh ancestry (stages 2a and 3a in Fig. 1). Lead SNPs representing 513 novel, independent (r^2 <0.25) association signals with $P < 10^{-3}$ in the discovery GWAS (including only two previously known GWAS SNPs from *TCF7L2* and *IGF2BP2* and excluding 62 SNPs with $P < 10^{-3}$ from other known T2D loci) were tested for in silico replication in the Punjabi Sikh subcomponent of the LOLIPOP GWAS comprising 801 T2D case and 2,018 control subjects (Supplementary Table 1). Top SNPs representing 66 putatively novel signals with $P < 10^{-4}$ after stage 1 and 2a meta-analysis using a fixed effects, inverse-variance approach were directly genotyped in the stage 3a sample of 2,894 Punjabi Sikh individuals (1,711 T2D case and 1,183 control subjects) (Fig. 1 and Supplementary Table 2).

In a combined meta-analysis of the three Punjabi studies (n = 7,329), we identified one new locus reaching genome-wide significance ($P < 5 \times 10^{-8}$) along with robust replication of the established SNP rs7903146 in *TCF7L2* ($P = 3.32 \times 10^{-19}$) in Sikhs (Figs. 2, 3, and 4). This novel association signal lies in a 164-kb region of strong LD at 13q12 (harboring genes gamma-sarcoglycan [*SGCG*] and sacsin [*SACS*]) and is represented by a directly genotyped intronic SNP, rs9552911 in *SGCG* (odds ratio [OR] 0.67 [95% CI 0.58–0.77], $P = 1.82 \times 10^{-8}$ for the minor "A" allele) (Table 1, Fig. 4, and Supplementary Table 5). Excluding BMI from the logistic regression model did not affect the association (Supplementary Table 6). Furthermore, including five additional principal components in the model did not attenuate the signal; indeed, the effect and significance were slightly improved (Supplementary Table 6). The genetic variance (R^2) explained by this variant for the T2D phenotype in Punjabi Sikh discovery and replication sets was 1.57 and 1.34%, respectively. There were 15 additional independent loci with suggestive evidence ($P < 10^{-5}$ to $< 10^{-7}$) of association, including six unknown regions along with *IGF2BP2*, originally identified in Caucasians (43) (Supplementary Table 5). Meta-analysis results including non-Sikh Punjabis from PROMIS (Pakistan) revealed suggestive association ($P < 10^{-5}$ to $< 10^{-7}$) at SNPs from three new regions: chromosome 18q21 *ZBTB7C* (rs1893835), 20q13, near *HMG1L1/CTCFL/RBM38/PCK1* (rs328506), and 5q33 (rs17053082) (Supplementary Table 7). Association results for 42 previously reported T2D loci in the Punjabi cohort are summarized in Supplementary Table 14. Most loci showed consistent effect in the same direction and 33 out of 42 were associated with T2D at $P < 0.05$ in Sikhs.

Replication/evaluation and meta-analysis in other South Asians. In order to identify T2D association signals common to Punjabi and other South Asian populations, we tested the association of the 513 top independent signals ($P < 10^{-3}$) derived from the discovery cohort in GWAS from the LOLIPOP, PROMIS, and RACE studies as part of stage 2b replication (10,971 T2D case and 18,186 control subjects) (Fig. 1 and Supplementary Table 1). Thirty-one signals ($P < 10^{-4}$ from an interim analysis with stage 2b) were further genotyped in 10,817 South Asians (5,157 T2D and 5,660 control subjects) (Fig. 1) as part of stage 3b replication. Clinical characteristics of the stage 3 replication cohorts are described in Supplementary Table 4. Combined South Asian meta-analysis revealed nominally significant association in six SNPs with MAF >5% ($P \leq 10^{-4}$), but only the two previously known SNPs in *TCF7L2* and *IGF2BP2* reached genome-wide significance (Table 1 and Supplementary Table 8). Suggestive novel signals included SNPs at chromosome 20q13, near *HMG1L1/CTCFL/RBM38/PCK1* (rs328506), 7q32 near *PLXNA4* (rs1593304), 3p21 in *SCAP* (rs4858889), and 5p11 (rs13155082) (Supplementary Table 8). Further studies and replication in a larger sample will be required to validate these results and identify causal variants at these loci.

Multiethnic replication and meta-analysis. To identify T2D signals spanning ethnicities, we extended the replication of 31 SNPs with $P < 10^{-4}$ in Punjabis and South Asians (stage 3b) to East Asians (AGEN+) and Europeans (DIAGRAM+) in stages 3c and 3d, respectively (Fig. 1). Upon meta-analysis of 31 loci in Asians (South Asians and AGEN +), genome-wide associations were only seen in *TCF7L2* (rs7903146, $P = 1.93 \times 10^{-38}$) and *IGF2BP2* (rs1470579, $P = 1.54 \times 10^{-13}$) (Supplementary Table 9). In joint multiethnic meta-analysis on 128,127 individuals from 27 studies, only two previously known loci, *TCF7L2* (rs7903146, $P = 8.53 \times 10^{-75}$) and *IGF2BP2* (rs1470579, $P = 1.81 \times 10^{-19}$), showed robust associations. Interestingly, none of the Punjabi hits could be independently confirmed in AGEN+ or DIAGRAM+ (notably, the lead rs9552911 variant from *SGCG* was monomorphic in DIAGRAM+) (Table 1 and Supplementary Table 10). Lookup of 50 kb upstream and downstream of SNPs within the *SGCG* locus in the publicly available data of

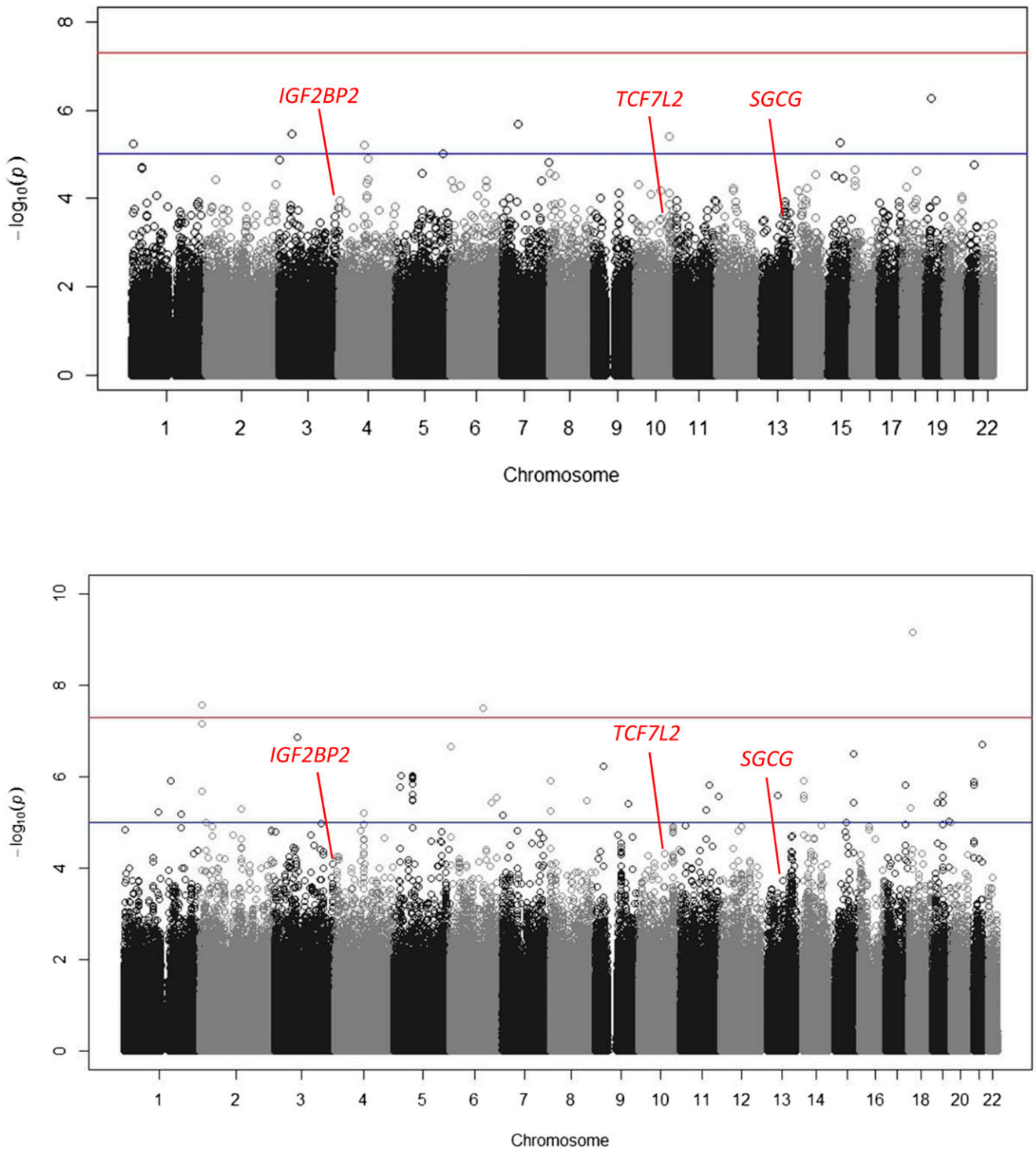


FIG. 2. *A*: Manhattan plot showing primary genome-wide association analysis of the Punjabi Sikh discovery cohort using directly genotyped (524,216) SNPs. *B*: Manhattan plot shows imputed 1,232,008 SNPs on the *x*-axis and $-\log_{10} P$ value of association on the *y*-axis. Locations of the three loci (including one novel locus at SGCG) reached genome-wide significance after combined analysis of the GWAS and replication data in Punjabi Sikhs.

the Meta-Analyses of Glucose and Insulin-Related Traits Consortium (MAGIC) study on glycemic trait GWAS (44,45) revealed several nominal associations of SNPs with fasting blood glucose and 2-h glucose levels (Supplementary Fig. 3). Some of these SNPs also showed an association with fasting blood glucose and waist or waist-to-hip ratio in Sikhs (Supplementary Table 11), but none of these were in LD ($r^2 > 0.20$) with our lead SNP.

Gene expression studies. We examined the expression of *SGCG* and neighboring genes *(FLJ46358*, *MIPEP*, *SACS*, and s*TNFRSF19)* within 1 Mb of the index SNP by *cis* expression quantitative trait locus (eQTL) analysis using adipose tissue, skin, and lymphoblastic cell line gene expression data from the MuTHER Consortium, comprising healthy female twins of European ancestry from Britain. Several SNPs in the *SGCG* region were associated with

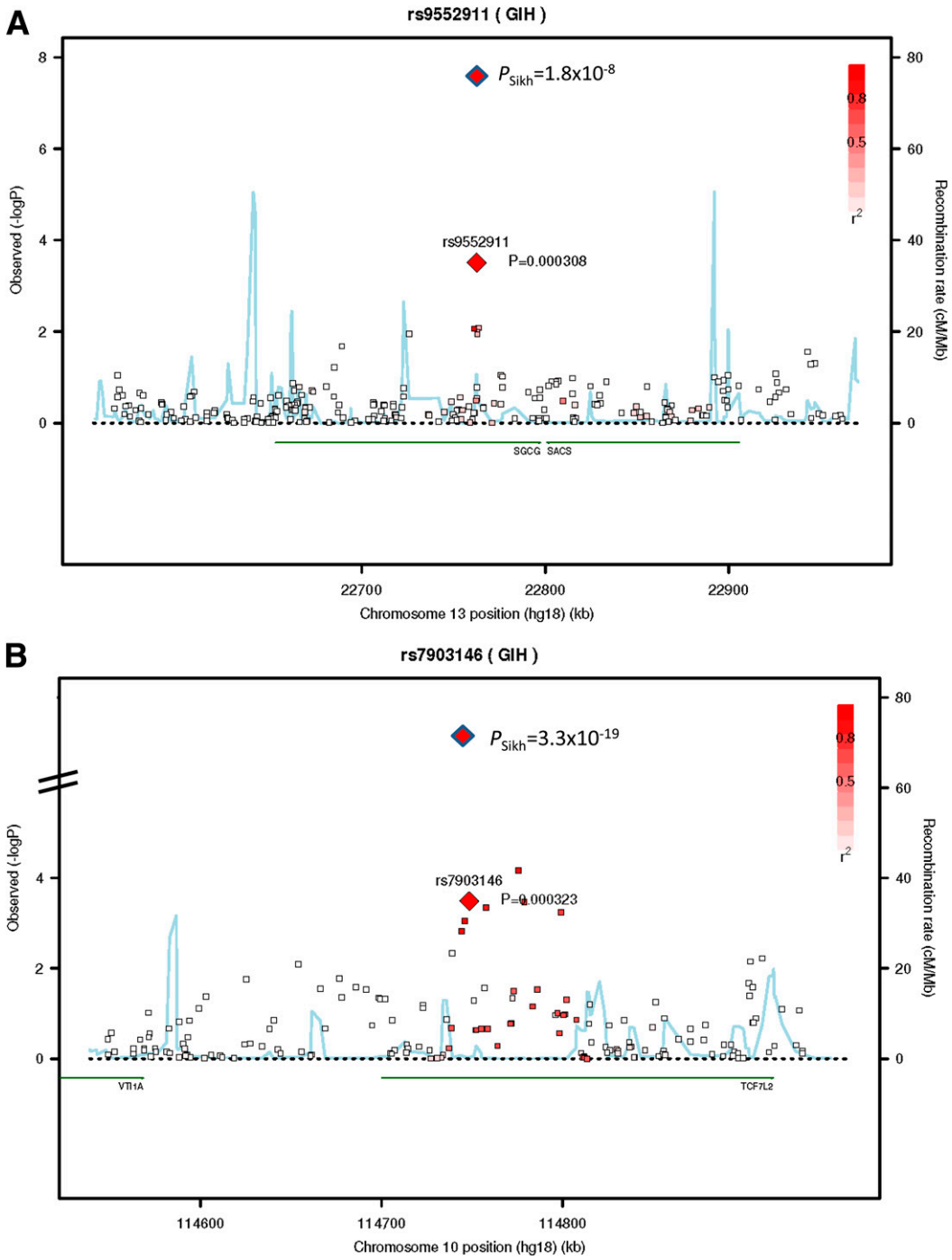


FIG. 3. *A*: Regional association plot for a new T2D locus detected at 13q12 in the *SGCG* gene from the genome-wide meta-analysis in Sikhs. *B*: A strong confirmation of SNPs in the *TCF7L2* gene in Sikh meta-analysis. In these plots, the SNPs showing the most strongly associated signal are depicted as a red diamond with blue border for the combined stage 1, 2a, and 3a results for meta-analysis, and the red diamond with black border shows evidence of association for the stage 1 results. Each square in color shows a SNP with the color scale relating the r^2 value for that SNP and the top SNP taken from the HapMap 3 GIH panel. We present LD using the GIH panel, the closest HapMap population to the Sikhs; however, we note that there could still be differential LD between the reference panel and the Sikh population. At the bottom of the plot, the locations of known genes in the region are shown.

significantly elevated ($P_{eQTL} < 10^{-4}$ to $< 10^{-9}$) expression of SGCG mRNA in adipose tissues (Supplementary Table 12 and Supplementary Fig. 4). One adipose eQTL from MuTHER (rs572303, $P_{eQTL} = 5.47 \times 10^{-4}$) located within *SGCG* showed a nominally significant association with increased waist circumference in Sikhs ($\beta = 0.67$, $P = 5.2 \times 10^{-2}$) (Supplementary Table 11). As shown in Supplementary Fig. 4, the LD patterns in the region (~1.46 Mb) surrounding the *SGCG* variant (rs9552911) varied in East Asians (JPT), Africans (YRI), Caucasians (CEU), Gujarati Indians (GIH), and Sikhs. Interestingly, in Caucasians and Yorubians, this variant was monomorphic. However, several alternative SNPs from this region in Europeans were nominally associated with fasting blood glucose (MAGIC study, r^2 ranging from 0.10 to 0.20 with the index SNP [rs9552911]) and mRNA expression of adipose cells in

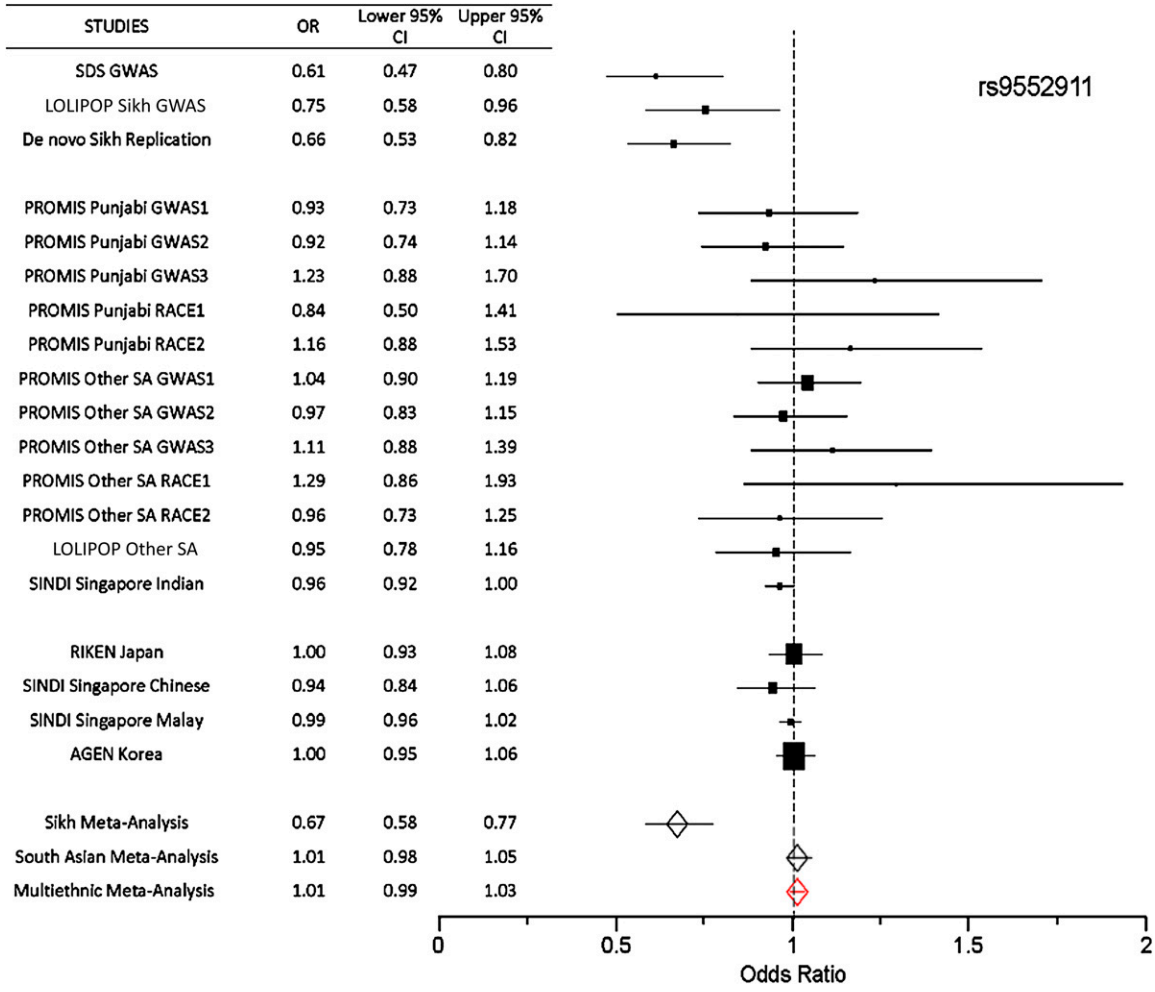


FIG. 4. Forest plot showing the association of lead SNP in the *SGCG* (rs9552911) gene with T2D. For each study, the estimates of the ORs with 95% CI are shown. In addition, meta-analyses of Sikhs and South Asians and multiethnic studies are shown at the bottom. Meta-analysis in Sikhs shows a significant association of rs9552911 with T2D (OR 0.67 [0.58–0.77], $P = 1.82 \times 10^{-8}$). The red diamond signifies the OR and CI across all multiethnic samples.

the MuTHER study (r^2 ranging from 0.14 to 0.26 with the index SNP). These data suggest that population differences may underlie the weak LD. It is possible that a single causal variant may be responsible for these associations, but LD may differ between Sikhs, Europeans, and other populations.

Comparative analysis of autozygosity. We further looked to compare the distributions of inbreeding coefficients and autozygosity as described by Nalls et al. (41). As expected, the inbreeding coefficients in our sample were higher compared with two outbred populations of European Americans, Coriell, and Baltimore Longitudinal Study of Aging (BLSA) ($F = 0.041 \pm 0.018$ in Sikhs vs. $F = 0.007 \pm 0.019$ in Coriell and $F = -0.3 \pm 0.012$ in BLSA), as assessed by Nalls et al. (41). However, these results were similar to other Indian populations previously reported by Reich et al. (46). No significant difference in inbreeding was observed between case and control subjects ($P = 0.59$). Autozygosity analysis determined that there were 19 ± 7 homozygous segments >1 Mb in length, with an average length of 2.0 ± 0.95 Mb. Hence, fewer but longer autozygous segments were found in our population than in outbred populations. No correlation of measures of autozygosity to age was observed ($P > 0.05$) across decades of age.

DISCUSSION

In this GWAS and multistage meta-analysis, a novel locus at 13q12 in the *SGCG* gene (rs9552911, $P = 1.82 \times 10^{-8}$) was identified as associated with T2D susceptibility in Punjabi Sikhs from Northern India.

SGCG is a member of the sarcoglycan complex of transmembrane glycoproteins mutated in autosomal recessive muscular dystrophy, in particular limb-girdle muscular dystrophy type 2C (LGMD2C). *SGCG* is expressed in skeletal muscle, and its high expression is also seen in vascular smooth muscle cells as well as in breast cancer cell lines (47,48). Founder mutations in *SGCG* that cause LGMD2C predate migration of the Romani gypsies of Europe out of India around 1100 AD (49). Due to complete endogamy, this genetically isolated community had an increased incidence of autosomal recessive LGMD2C. *SGCG*-targeted knockout mice displayed a variety of phenotypes, including dystrophic cardiomyopathy and defects in skeletal muscle, metabolism, homeostasis, growth, apoptosis, aging, and behavior (50–53). Mice lacking the sarcoglycan

TABLE 1
GWAS, replication, and meta-analysis results of T2D loci identified in Punjabi Sikhs

SNP	Chromosome position	Nearest gene	Effect/ other allele	EAF		Punjabi Sikh Discovery GWAS (stage 1)	Punjabi Sikh Replication (stages 2a, 3a)	Other South Asian Replication (stages 2b, 3b)	East Asian Evaluation (stage 3c)	European Evaluation (stage 3d)	Punjabi meta-analysis	South Asian meta-analysis	Multiethnic meta-analysis
				Sikh	CEU	OR (95% CI) *P* value	OR (95% CI) *P* value	OR (95% CI) *P* value	OR (95% CI) *P* value	OR (95% CI) *P* value	OR (95% CI) *P* value	OR (95% CI) *P* value	OR (95% CI) *P* value
Novel T2D locus													
rs9552911	13 23864657	*SGCG*	A/G	0.08	0	0.61 (0.47–0.80), $P = 3.08 \times 10^{-4}$	0.69 (0.59–0.82), $P = 1.14 \times 10^{-5}$	1.04 (1.00–1.07), $P = 0.037$	1.00 (0.98–1.03), $P = 0.77$	—	0.67 (0.58–0.77), $P = 1.82 \times 10^{-8}$	1.01 (0.98–1.05), $P = 0.47$	1.01 (0.99–1.03), $P = 0.49$
Previously reported T2D loci													
rs1470579	3 185511687	*IGF2BP2*	A/C	0.59	0.5	0.76 (0.66–0.88), $P = 2.53 \times 10^{-4}$	0.87 (0.8–0.95), $P = 1.87 \times 10^{-3}$	0.95 (0.93–0.97), $P = 1.79 \times 10^{-6}$	0.94 (0.92–0.97), $P = 7.21 \times 10^{-6}$	0.88 (0.92–0.84), $P = 2.17 \times 10^{-9}$	0.84 (0.78–0.9), $P = 5.02 \times 10^{-6}$	0.94 (0.92–0.96), $P = 4.19 \times 10^{-9}$	0.94 (0.92–0.95), $P = 1.81 \times 10^{-19}$
rs7903146	10 114758349	*TCF7L2*	T/C	0.35	0.28	1.31 (1.13–1.52), $P = 3.23 \times 10^{-4}$	1.50 (1.36–1.65), $P = 7.83 \times 10^{-17}$	1.13 (1.10–1.16), $P = 6.12 \times 10^{-25}$	1.12 (1.06–1.19), $P = 1.06 \times 10^{-4}$	1.40 (1.34–1.46), $P = 2.21 \times 10^{-51}$	1.44 (1.33–1.56), $P = 3.32 \times 10^{-19}$	1.15 (1.13–1.18), $P = 2.71 \times 10^{-35}$	1.19 (1.16–1.21), $P = 8.53 \times 10^{-75}$

Novel T2D locus rs9552911											
Punjabi Sikh only	Studies	*n* case subjects	*n* control subjects	Genotype call rate	Controls HWE *P* value	MAF case subjects	MAF control subjects	OR (95% CI)	*P* value	Meta-analysis OR (95% CI) and *P* value	Het *P* value
Punjabi Sikhs	SDS/AIDHS discovery GWAS	842	774	1	0.39	0.06	0.09	0.61 (0.47–0.80)	0.00031	0.67 (0.58–0.77) $P = 1.82 \times 10^{-8}$	0.56
Punjabi Sikhs	LOLIPOP Sikh GWAS	801	2,017	1	0.88	0.07	0.08	0.75 (0.58–0.96)	0.021		
Punjabi Sikhs	De novo Sikh Replication	977	1,493	0.97	0.049	0.06	0.09	0.66 (0.53–0.82)	0.00014		

All *P* values are two sided. CEU, Euro-Caucasians; EAF, effect allele frequency; Het, heterozygous.

complex including SGCG in adipose and skeletal muscle were shown to be glucose intolerant and exhibited whole-body insulin resistance due to impaired insulin-stimulated glucose uptake in skeletal muscle (54).

The allelic distribution of the less common "A" (protective) allele of rs9552911 ranged from 0.06 to 0.15 in South Asians and differed between other South Asians (0.11) and Punjabi Sikhs (0.08) (see details in Supplementary Table 13). Further replication in large independent datasets of South Asians and Punjabi Sikhs would be needed to confirm the pattern of observed association. In view of the complex racial history complicated by a well-defined caste system, Indian populations display a great deal of genetic and cultural diversity (55). Studies suggest that genetic affinity among endogamous communities in India is inversely correlated with geographic distance between them (23). Therefore, it is possible that undetected causal variant(s) or multiple rare variants in LD with this marker arose on a haplotype tagged by rs9552911 in Punjabi Sikhs after divergence from other South and West Indian populations. This variation in the index SNP rs9552911 does not appear to be of recent origin, as suggested by comparative genomic analysis (Supplementary Fig. 5). Two important nuclear hormone receptors and transcription factors (peroxisome proliferator–activated receptor-γ [PPAR-γ] [1 and 2] and PPAR-α) bind to the promoter and intron 1 of the *SGCG* gene. Further, the maturity-onset diabetes of young 4 (MODY4) locus at chromosome 13q12, represented by insulin promoter factor 1 or *PDX-1*, lies next to the *SGCG* locus. Therefore, further in-depth examination by targeted resequencing in the extended region and functional studies may reveal putative causative variants in this extended region and provide insight into the physiological relevance of the observed association.

In summary, our study identified a novel locus associated with T2D in a population of Punjabi Sikh ancestry from Northern India. These findings not only provide new information on previously unknown regions associated with T2D but demonstrate a putative population-specific association that could lead to additional biological insights into T2D pathogenesis.

ACKNOWLEDGMENTS

This work was supported by National Institutes of Health/National Institute of Diabetes and Digestive and Kidney Diseases Grant R01-DK-082766 and a seed grant from the University of Oklahoma Health Sciences Center.

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

Manuscript preparation: R.S. and L.F.B. performed the analysis. R.S., D.S., and D.K.S. wrote the manuscript. Data collection and analysis in the participating studies: AIDHS/SDS: R.S., L.F.B., M.L.G., T.B., A.Bj., G.S.W., J.R.S., N.K.M., S.R., M.I.K., J.J.M., and D.K.S.; PROMIS and RACE study: D.S., R.Y., W.K.H., A.R., P.F., P.D., D.J.R., and J.D.; Singapore Indian Eye Study: X.S., D.P.-K.N., T.-Y.W., and E.S.T.; SLDS: N.H., M.I.M., P.K., and D.M.; CURES: V.R., M.C., S.L., and V.M.; DGP and UKADS: S.D.R., A.Ba., A.H.B., and M.A.K.; BioBank Japan and RIKEN T2D study: T.Y., K.H., Y.T., H.H., H.M., K.T., S.M., and T.K.; AGEN Consortium: Y.S.C.; LOLIPOP: J.C.C. and J.S.K. Association results among Europeans in DIAGRAM: M.I.M. and A.P.M. eQTL analyses in MuTHER: M.I.M. All authors read and provided critical comments on the manuscript. D.K.S. 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.

The authors thank the research participants for their contribution and support for making this study possible. The authors also thank Bansari Mehta and Braden Juengel (University of Oklahoma Health Sciences Center) and Jackie Lane (Harvard Medical School) for their superb technical support in this study. The authors thank the collaborating groups of all the contributing sites including Dr. Nadereh Jafari and staff at the Northwestern University Genomics Core (Evanston, IL) for performing genome-wide genotyping and Dr. Andrew Brooks and staff at Rutgers University (New Brunswick, NJ) for replication genotyping. Study-specific acknowledgments are provided in the Supplementary Data.

REFERENCES

1. Basnyat B, Rajapaksa LC. Cardiovascular and infectious diseases in South Asia: the double whammy. BMJ 2004;328:781
2. International Diabetes Federation. *Diabetes Atlas*. Belgium, International Diabetes Federation, 2011
3. Ghaffar A, Reddy KS, Singhi M. Burden of non-communicable diseases in South Asia. BMJ 2004;328:807–810
4. Zimmet P. Type 2 (non-insulin-dependent) diabetes—an epidemiological overview. Diabetologia 1982;22:399–411
5. McKeigue PM, Pierpoint T, Ferrie JE, Marmot MG. Relationship of glucose intolerance and hyperinsulinaemia to body fat pattern in south Asians and Europeans. Diabetologia 1992;35:785–791
6. Oldroyd J, Banerjee M, Heald A, Cruickshank K. Diabetes and ethnic minorities. Postgrad Med J 2005;81:486–490
7. Nakagami T, Qiao Q, Carstensen B, et al.; DECODE-DECODA Study Group. Age, body mass index and type 2 diabetes-associations modified by ethnicity. Diabetologia 2003;46:1063–1070
8. Zimmet PZ. Kelly West Lecture 1991. Challenges in diabetes epidemiology—from West to the rest. Diabetes Care 1992;15:232–252
9. McCarthy MI, Abecasis GR, Cardon LR, et al. Genome-wide association studies for complex traits: consensus, uncertainty and challenges. Nat Rev Genet 2008;9:356–369
10. Sanghera DK, Nath SK, Ortega L, et al. TCF7L2 polymorphisms are associated with type 2 diabetes in Khatri Sikhs from North India: genetic variation affects lipid levels. Ann Hum Genet 2008;72:499–509
11. Sanghera DK, Ortega L, Han S, et al. Impact of nine common type 2 diabetes risk polymorphisms in Asian Indian Sikhs: PPARG2 (Pro12Ala), IGF2BP2, TCF7L2 and FTO variants confer a significant risk. BMC Med Genet 2008;9:59
12. Hu C, Zhang R, Wang C, et al. A genetic variant of G6PC2 is associated with type 2 diabetes and fasting plasma glucose level in the Chinese population. Diabetologia 2009;52:451–456
13. Hu C, Zhang R, Wang C, et al. PPARG, KCNJ11, CDKAL1, CDKN2A-CDKN2B, IDE-KIF11-HHEX, IGF2BP2 and SLC30A8 are associated with type 2 diabetes in a Chinese population. PLoS ONE 2009;4:e7643
14. Ng MC, Park KS, Oh B, et al. Implication of genetic variants near TCF7L2, SLC30A8, HHEX, CDKAL1, CDKN2A/B, IGF2BP2, and FTO in type 2 diabetes and obesity in 6,719 Asians. Diabetes 2008;57:2226–2233
15. Sanghera DK, Been L, Ortega L, et al. Testing the association of novel meta-analysis-derived diabetes risk genes with type II diabetes and related metabolic traits in Asian Indian Sikhs. J Hum Genet 2009;54:162–168
16. Omori S, Tanaka Y, Horikoshi M, et al. Replication study for the association of new meta-analysis-derived risk loci with susceptibility to type 2 diabetes in 6,244 Japanese individuals. Diabetologia 2009;52:1554–1560
17. Schleinitz D, Tönjes A, Böttcher Y, et al. Lack of significant effects of the type 2 diabetes susceptibility loci JAZF1, CDC123/CAMK1D, NOTCH2, ADAMTS9, THADA, and TSPAN8/LGR5 on diabetes and quantitative metabolic traits. Horm Metab Res 2010;42:14–22
18. Rees SD, Hydrie MZ, Shera AS, et al. Replication of 13 genome-wide association (GWA)-validated risk variants for type 2 diabetes in Pakistani populations. Diabetologia 2011;54:1368–1374
19. 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
20. 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

21. Cho YS, Chen CH, Hu C, et al.; DIAGRAM Consortium; MuTHER Consortium. Meta-analysis of genome-wide association studies identifies eight new loci for type 2 diabetes in east Asians. Nat Genet 2012;44:67–72
22. Kooner JS, Saleheen D, Sim X, et al.; DIAGRAM; MuTHER. Genome-wide association study in individuals of South Asian ancestry identifies six new type 2 diabetes susceptibility loci. Nat Genet 2011;43:984–989
23. Kivisild T, Bamshad MJ, Kaldma K, et al. Deep common ancestry of Indian and Western-Eurasian mitochondrial DNA lineages. Curr Biol 1999;9:1331–1334
24. Mathew CG. New links to the pathogenesis of Crohn disease provided by genome-wide association scans. Nat Rev Genet 2008;9:9–14
25. Singh K, Bhalla V, Kaul L. *People of India.* Vol. X. Oxford, Oxford University Press, 1994, p. 1–3
26. Sanghera DK, Bhatti JS, Bhatti GK, et al. The Khatri Sikh Diabetes Study (SDS): study design, methodology, sample collection, and initial results. Hum Biol 2006;78:43–63
27. Sanghera DK, Demirci FY, Been L, et al. PPARG and ADIPOQ gene polymorphisms increase type 2 diabetes mellitus risk in Asian Indian Sikhs: Pro12Ala still remains as the strongest predictor. Metabolism 2010;59:492–501
28. Been LF, Hatfield JL, Shankar A, et al. A low frequency variant within the GWAS locus of MTNR1B affects fasting glucose concentrations: Genetic risk is modulated by obesity. Nutr Metab Cardiovasc Dis 2012; 22:944–951
29. American Diabetes Association. Diagnosis and classification of diabetes mellitus. Diabetes Care 2004;27(Suppl. 1):S5–S10
30. McKeigue PM, Marmot MG, Syndercombe Court YD, Cottier DE, Rahman S, Riemersma RA. Diabetes, hyperinsulinaemia, and coronary risk factors in Bangladeshis in east London. Br Heart J 1988;60:390–396
31. Lavanya R, Jeganathan VS, Zheng Y, et al. Methodology of the Singapore Indian Chinese Cohort (SICC) eye study: quantifying ethnic variations in the epidemiology of eye diseases in Asians. Ophthalmic Epidemiol 2009;16:325–336
32. Chidambaram M, Radha V, Mohan V. Replication of recently described type 2 diabetes gene variants in a South Indian population. Metabolism 2010;59:1760–1766
33. Rees SD, Islam M, Hydrie MZ, et al. An FTO variant is associated with type 2 diabetes in South Asian populations after accounting for body mass index and waist circumference. Diabet Med 2011;28:673–680
34. Katulanda P, Constantine GR, Mahesh JG, et al. Prevalence and projections of diabetes and pre-diabetes in adults in Sri Lanka—Sri Lanka Diabetes, Cardiovascular Study (SLDCS). Diabet Med 2008;25:1062–1069
35. Gille H, Strahl T, Shaw PE. Activation of ternary complex factor Elk-1 by stress-activated protein kinases. Curr Biol 1995;5:1191–1200
36. 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
37. Miller SA, Dykes DD, Polesky HF. A simple salting out procedure for extracting DNA from human nucleated cells. Nucleic Acids Res 1988;16:1215
38. 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
39. Howie B, Marchini J, Stephens M. Genotype imputation with thousands of genomes. G3 (Bethesda) 2011;1:457–470
40. Marchini J, Howie B. Genotype imputation for genome-wide association studies. Nat Rev Genet 2010;11:499–511
41. Nalls MA, Simon-Sanchez J, Gibbs JR, et al. Measures of autozygosity in decline: globalization, urbanization, and its implications for medical genetics. PLoS Genet 2009;5:e1000415
42. Willer CJ, Li Y, Abecasis GR. METAL: fast and efficient meta-analysis of genomewide association scans. Bioinformatics 2010;26:2190–2191
43. 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
44. Dupuis J, Langenberg C, Prokopenko I, et al.; DIAGRAM Consortium; GIANT Consortium; Global BPgen Consortium; Anders 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
45. 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
46. Reich D, Thangaraj K, Patterson N, Price AL, Singh L. Reconstructing Indian population history. Nature 2009;461:489–494
47. Avondo F, Roncaglia P, Crescenzio N, et al. Fibroblasts from patients with Diamond-Blackfan anaemia show abnormal expression of genes involved in protein synthesis, amino acid metabolism and cancer. BMC Genomics 2009;10:442
48. Barresi R, Moore SA, Stolle CA, Mendell JR, Campbell KP. Expression of gamma-sarcoglycan in smooth muscle and its interaction with the smooth muscle sarcoglycan-sarcospan complex. J Biol Chem 2000;275:38554–38560
49. Piccolo F, Jeanpierre M, Leturcq F, et al. A founder mutation in the gamma-sarcoglycan gene of gypsies possibly predating their migration out of India. Hum Mol Genet 1996;5:2019–2022
50. Goldstein JA, Kelly SM, LoPresti PP, et al. SMAD signaling drives heart and muscle dysfunction in a *Drosophila* model of muscular dystrophy. Hum Mol Genet 2011;20:894–904
51. Townsend D, Yasuda S, McNally E, Metzger JM. Distinct pathophysiological mechanisms of cardiomyopathy in hearts lacking dystrophin or the sarcoglycan complex. FASEB J 2011;25:3106–3114
52. Wheeler MT, Korcarz CE, Collins KA, et al. Secondary coronary artery vasospasm promotes cardiomyopathy progression. Am J Pathol 2004;164:1063–1071
53. Hack AA, Ly CT, Jiang F, et al. Gamma-sarcoglycan deficiency leads to muscle membrane defects and apoptosis independent of dystrophin. J Cell Biol 1998;142:1279–1287
54. Groh S, Zong H, Goddeeris MM, et al. Sarcoglycan complex: implications for metabolic defects in muscular dystrophies. J Biol Chem 2009;284:19178–19182
55. Bamshad M, Kivisild T, Watkins WS, et al. Genetic evidence on the origins of Indian caste populations. Genome Res 2001;11:994–1004

Mutations in *HNF1A* Result in Marked Alterations of Plasma Glycan Profile

Gaya Thanabalasingham,[1,2] Jennifer E. Huffman,[3] Jayesh J. Kattla,[4] Mislav Novokmet,[5] Igor Rudan,[6,7] Anna L. Gloyn,[1,2] Caroline Hayward,[3] Barbara Adamczyk,[4] Rebecca M. Reynolds,[8] Ana Muzinic,[5] Neelam Hassanali,[1] Maja Pucic,[5] Amanda J. Bennett,[1] Abdelkader Essafi,[3] Ozren Polasek,[7] Saima A. Mughal,[1,2] Irma Redzic,[9] Dragan Primorac,[7,10] Lina Zgaga,[6] Ivana Kolcic,[7] Torben Hansen,[11,12,13] Daniela Gasperikova,[14] Erling Tjora,[15,16] Mark W.J. Strachan,[17] Trine Nielsen,[11] Juraj Stanik,[14,18] Iwar Klimes,[14] Oluf B. Pedersen,[11,19,20] Pål R. Njølstad,[15,16] Sarah H. Wild,[6] Ulf Gyllensten,[21] Olga Gornik,[9] James F. Wilson,[6] Nicholas D. Hastie,[3] Harry Campbell,[6] Mark I. McCarthy,[1,2,22] Pauline M. Rudd,[4] Katharine R. Owen,[1,2] Gordan Lauc,[5,9] and Alan F. Wright[3]

A recent genome-wide association study identified hepatocyte nuclear factor 1-α (*HNF1A*) as a key regulator of fucosylation. We hypothesized that loss-of-function *HNF1A* mutations causal for maturity-onset diabetes of the young (MODY) would display altered fucosylation of *N*-linked glycans on plasma proteins and that glycan biomarkers could improve the efficiency of a diagnosis of *HNF1A*-MODY. In a pilot comparison of 33 subjects with *HNF1A*-MODY and 41 subjects with type 2 diabetes, 15 of 29 glycan measurements differed between the two groups. The DG9-glycan index, which is the ratio of fucosylated to nonfucosylated triantennary glycans, provided optimum discrimination in the pilot study and was examined further among additional subjects with *HNF1A*-MODY (n = 188), glucokinase (*GCK*)-MODY (n = 118), hepatocyte nuclear factor 4-α (*HNF4A*)-MODY (n = 40), type 1 diabetes (n = 98), type 2 diabetes (n = 167), and nondiabetic controls (n = 98). The DG9-glycan index was markedly lower in *HNF1A*-MODY than in controls or other diabetes subtypes, offered good discrimination between *HNF1A*-MODY and both type 1 and type 2 diabetes (C statistic ≥0.90), and enabled us to detect three previously undetected *HNF1A* mutations in patients with diabetes. In conclusion, glycan profiles are altered substantially in *HNF1A*-MODY, and the DG9-glycan index has potential clinical value as a diagnostic biomarker of *HNF1A* dysfunction. ***Diabetes*** **62:1329–1337, 2013**

Genome-wide association studies are providing novel insights into the genetic architecture and biological basis of many diseases, but immediate translation into clinical practice has been limited. We recently performed a genome-wide association study of the human plasma *N*-glycome and found evidence of association involving common variants near the hepatocyte nuclear factor 1-α (*HNF1A*) gene; follow-up functional experiments established HNF1A as a master regulator of plasma protein fucosylation (1). Fucosylation, a specific type of glycosylation, comprises the addition of fucose residues to glycans. Here we evaluate the hypothesis that mutations causing a more severe deficit in HNF1A function (resulting in the monogenic subtype of diabetes known as *HNF1A* maturity-onset diabetes of the young [MODY; *HNF1A*-MODY]) are associated with marked alterations of plasma glycome composition, and we assess the value of glycan profiles as a diagnostic biomarker for *HNF1A*-MODY.

Most human proteins are posttranslationally modified by the addition of complex oligosaccharide structures (glycans) (2). Despite the impact on protein structure and function, the clinical consequences of changes in the human glycome remain largely unexplored, primarily because reliable analytical techniques have been developed only recently (3). In recent studies, HNF1A was shown to promote both the de novo and salvage pathways for the synthesis of guanosine diphosphate–fucose (1) and to regulate fucosyltransferase VI (1,4). HNF1A thereby controls the outer-arm (antennary) fucosylation of proteins with *N*-linked glycans through effects on both the supply of activated precursors and the incorporation of fucose (1,4).

Mutations disrupting *HNF1A* are responsible for the most common subtype of monogenic diabetes, *HNF1A*-MODY (5). Like other forms of MODY, *HNF1A*-MODY is characterized

From the [1]Oxford Centre for Diabetes, Endocrinology and Metabolism, University of Oxford, Oxford, U.K.; the [2]Oxford NIHR Biomedical Research Centre, Churchill Hospital, Oxford, U.K.; the [3]MRC Human Genetics Unit, Institute of Genetics and Molecular Medicine, University of Edinburgh, Edinburgh, U.K.; the [4]Dublin-Oxford Glycobiology Laboratory, National Institute for Bioprocessing Research and Training (NIBRT), Dublin, Ireland; [5]Genos Ltd., Glycobiology Division, Zagreb, Croatia; the [6]Centre for Population Health Sciences, University of Edinburgh Medical School, Edinburgh, U.K.; the [7]University of Split School of Medicine, Split, Croatia; the [8]Centre for Cardiovascular Science, Queen's Medical Research Institute, University of Edinburgh, Edinburgh, U.K.; the [9]Faculty of Pharmacy and Biochemistry, University of Zagreb, Zagreb, Croatia; the [10]University of Osijek School of Medicine, Osijek, Croatia; the [11]Novo Nordisk Foundation Center for Basic Metabolic Research, Faculty of Health and Medical Sciences, University of Copenhagen, Copenhagen, Denmark; the [12]Faculty of Health Sciences, University of Southern Denmark, Odense, Denmark; the [13]Steno Diabetes Center, Gentofte, Denmark; the [14]DIABGENE and Diabetes Laboratory, Institute of Experimental Endocrinology, Slovak Academy of Sciences, Bratislava, Slovakia; the [15]Department of Clinical Medicine, University of Bergen, Bergen, Norway; the [16]Department of Pediatrics, Haukeland University Hospital, Bergen, Norway; the [17]Metabolic Unit, Western General Hospital, Edinburgh, U.K.; the [18]Children Diabetes Centre at the First Department of Paediatrics, Faculty of Medicine at the Comenius University, Bratislava, Slovakia; the [19]Faculty of Health Sciences, University of Aarhus, Aarhus, Denmark; the [20]Hagedorn Research Institute, Copenhagen, Denmark; the [21]Department of Genetics and Pathology, Rudbeck Laboratory, Uppsala University, Uppsala, Sweden; and the [22]Wellcome Trust for Human Genetics, University of Oxford, Oxford, U.K.

Corresponding author: Alan F. Wright, alan.wright@igmm.ed.ac.uk.

Received 30 June 2012 and accepted 17 October 2012.

DOI: 10.2337/db12-0880

This article contains Supplementary Data online at http://diabetes.diabetesjournals.org/lookup/suppl/doi:10.2337/db12-0880/-/DC1.

G.T., J.E.H., J.J.K., M.N., I.Ru., and A.L.G. contributed equally to this study.

H.C., M.I.M., P.M.R., K.R.O., G.L., and A.F.W. contributed equally to this study.

© 2013 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.

by autosomal dominant inheritance and pancreatic β-cell dysfunction. This typically leads to diabetes diagnosed in the second to fourth decade of life in the absence of β-cell autoimmunity and insulin resistance (5,6). In clinical practice, diagnostic differentiation between *HNF1A*-MODY and other causes of early-onset diabetes (including type 1 and type 2 diabetes and other forms of MODY) is complicated by the overlap of phenotypic features. In most countries only a minority of cases of *HNF1A*-MODY are referred for definitive molecular testing (i.e., *HNF1A* sequencing), and many patients with *HNF1A*-MODY are misdiagnosed with type 1 or type 2 diabetes (7). Failures or delays in accurate molecular diagnosis can have clinical repercussions because, unlike other forms of diabetes, the optimal treatment for *HNF1A*-MODY is low-dose sulfonylureas (8). Patients with undiagnosed *HNF1A*-MODY may spend many years receiving inappropriate treatment (including exogenous insulin) and experiencing suboptimal glycemic control (9).

We aimed to test the hypothesis that inactivating mutations in *HNF1A*, such as those found in *HNF1A*-MODY, are associated with decreased antennary fucosylation of circulating proteins and to evaluate the clinical translational potential of measuring glycan profiles in diabetes.

RESEARCH DESIGN AND METHODS

Subjects for initial and validation studies. Full details of the subjects are available in Table 1 and the Supplementary Appendix. Subjects carrying mutations in *HNF1A* (n = 221), glucokinase (*GCK*; n = 118), and hepatocyte nuclear factor 4-α (*HNF4A*; n = 40) were recruited from five European centers. Subjects with MODY had an established heterozygous loss-of-function mutation confirmed by sequencing in a certified diagnostic center. All MODY mutations were considered pathogenic if they met one or more of the following criteria: mentioned in previously published reports, presence of a truncating mutation, cosegregation of the mutation with a MODY phenotype within the family, and absence of the variant in normal chromosomes. In addition, we recruited 208 subjects with clinically labeled type 2 diabetes who were diagnosed at an age younger than 45 years, 98 subjects with clinically labeled type 1 diabetes, and 98 subjects who acted as nondiabetic controls. Most samples were collected when the subject was in a fasting state, although fasting status does not influence glycan levels (10).

Glycan release, labeling, and analysis. All samples were stored at −80°C before analysis. Glycan release, labeling, and analysis using hydrophilic interaction high-performance liquid chromatography and sialidase digestion was performed as previously reported (11,12). Chromatograms from fluorescently labeled glycans were separated into 16 glycan groups (GP series) and 13 desialylated glycan groups (DG series), composing a total of 29 peaks (Supplementary Table 1). The amounts of glycans present in each peak were expressed as the percentage of the total plasma glycome. Glycan analysis was performed in two centers: the National Institute for Bioprocessing Research and Training (Dublin, Ireland) and the Glycobiology laboratory of Genos Ltd. (Zagreb, Croatia). Both laboratories used the same columns and separation conditions and previously have demonstrated reproducibility of analytic results within and between laboratories (1,10).

Study design and analysis. All glycan traits were compared using Mann-Whitney U tests in an initial study of 33 subjects with *HNF1A*-MODY and 41 subjects with type 2 diabetes. Based on these results (Supplementary Table 1), the glycan ratio of DG9 to [DG8+DG9], hereafter referred to as the DG9-glycan index, was chosen for further follow-up.

The validation study was performed using 188 subjects with *HNF1A*-MODY, 118 subjects with *GCK*-MODY, 40 subjects with *HNF4A*-MODY, 167 subjects with type 2 diabetes, and 98 subjects with type 1 diabetes cases, plus 98 nondiabetic controls. There was no overlap between the initial and validation studies. We first sought evidence for important covariates through analysis of parameters including age, sex, BMI, HDL, triglycerides, sample origin, processing laboratory, and sample type (plasma vs. serum). An adjusted model incorporated significant covariates (age, BMI, sample origin, processing laboratory) as well as those already known to affect specific glycan traits (sex). An additional analysis was performed without covariates. The use of the DG9-glycan index as a discriminator of diabetes subtypes was analyzed using receiver operator characteristic (ROC) curves from which the C statistic was obtained.

TABLE 1
Clinical characteristics of subjects included in the initial and validation studies

	Initial study		Validation study					
Characteristic	*HNF1A*-MODY (n = 33)	Type 2 diabetes (n = 41)	*HNF1A*-MODY (n = 188)	*GCK*-MODY (n = 118)	*HNF4A*-MODY (n = 40)	Type 1 diabetes (n = 98)	Type 2 diabetes (n = 167)	Nondiabetic controls (n = 98)
Sex (% male)	33	46	45	43	43	51	60	51
BMI (kg/m^2)	25.4 (22.3–28.1)	34.1 (29.7–39.5)	23.3 (21.5–26.1)	22.9 (19.4–26.0)	24.6 (22.7–27.3)	26.5 (23.9–29.5)	31.0 (27.8–35.6)	27.1 (22.7–31.4)
Age at diabetes diagnosis (years)*	21.2 (15.6–29.0)	32.0 (27.0–36.0)	18.0 (15.0–27.0)	26.0 (12.0–37.0)	29.5 (20.0–39.0)	25.5 (14.0–32.0)	40.0 (37.0–43.0)	N/A
Duration of diabetes (years)*	12.0 (9.3–30.0)	6.8 (1.0–12.3)	16.0 (4.0–28.0)	3.0 (0.0–10.0)	17.7 (6.0–34.0)	14.1 (9.7–20.0)	12.0 (6.0–22.0)	N/A
Age at sampling (years)	38.9 (30.0–56.1)	40.2 (34.7–48.2)	40.0 (24.5–50.2)	32.0 (17.0–51.0)	52.0 (31.8–58.2)	42.3 (32.6–50.1)	51.7 (46.2–60.9)	53.7 (49.3–59.1)
HbA1c (%)	7.4 ± 1.8	9.0 ± 2.1	7.1 ± 1.2	6.5 ± 0.7	7.5 ± 1.7	8.8 ± 0.8	7.8 ± 1.4	5.5 ± 0.3
FPG (mmol/L)†	6.6 ± 1.8	9.7 ± 3.6	7.6 ± 2.7	7.1 ± 1.0	7.6 ± 2.6	9.7 ± 4.4	8.4 ± 2.5	5.7 ± 0.5
Total cholesterol (mmol/L)	4.5 ± 0.8	4.7 ± 1.0	4.9 ± 1.4	4.5 ± 0.9	4.9 ± 1.2	4.8 ± 1.3	4.7 ± 1.3	5.3 ± 1.0
HDL-cholesterol (mmol/L)	1.6 ± 0.4	1.1 ± 0.3	1.5 ± 0.5	1.3 ± 0.4	1.5 ± 0.4	1.5 ± 0.4	1.2 ± 0.4	1.6 ± 0.5
Triglycerides (mmol/L)	N/A	2.0 (1.2–2.6)	1.0 (0.7–1.5)	0.8 (0.6–1.3)	0.8 (0.7–1.2)	1.0 (0.8–1.4)	1.8 (1.3–2.4)	1.1 (0.8–1.5)

Normally distributed variables reported as mean ± SD; all others reported as median (25th–75th centiles). FPG, fasting plasma glucose; N/A, not available. *Subjects with diabetes only (i.e., excluded carriers of a nondiabetic mutation). †Unavailable for subjects from Edinburgh.

Performance of the DG9-glycan index as a clinical tool was evaluated by calculating sensitivity and specificity and other measures for the detection of *HNF1A*-MODY at various thresholds. In particular, we estimated posttest diagnostic probabilities based on data from an etiological investigation of young adults with diabetes from the U.K. that indicated pretest probabilities of 4% for *HNF1A*-MODY in young-onset type 2 diabetes and 1% in type 1 diabetes (13).

The effect of the type of *HNF1A* mutation, the HNF1A isoform, and the mutated functional domain of HNF1A on DG9-glycan index levels was assessed. *HNF1A* mutations were classified as either protein-changing mutations (missense mutations resulting in a change of amino acid) or truncating mutations (which generate a premature stop codon). In addition, protein-changing mutations were grouped as exons 1–6 [affecting isoforms *HNF1A* (*A*), (*B*) and (*C*)], exon 7 [isoforms *HNF1A* (*A*) and (*B*)] and exons 8–10 [isoform *HNF1A* (*A*) only] (14,15). Protein-changing mutations also were grouped according to the affected functional domain: dimerization, DNA binding, or transactivation (14,15).

We also assessed evidence for pathogenicity of the *HNF1A* missense mutations included in the validation study and examined whether the DG9-glycan index correlated with other indicators of pathogenicity. These included cosegregation within families, functional characterization of mutant proteins, and in silico prediction of the effect of the amino acid substitution on protein function.

***HNF1A* sequencing.** In the subsequent case-finding study, we evaluated the value of the DG9-glycan index as a screening test for identifying *HNF1A*-MODY in a set of individuals with young-onset diabetes (diagnosed up to age 45 years) who had not previously been suspected of having an *HNF1A* mutation. We tested subjects with a DG9-glycan index <0.16 from the initial or validation studies with clinical labels of type 1 (n = 7) and type 2 diabetes (n = 41), as well as subjects with diabetes of any type diagnosed up to age 45 years from general population cohorts from Croatia (n = 6) and Scotland (n = 3) in whom glycan profiles had been measured previously (1). The 10 exons of *HNF1A* were amplified by PCR and bidirectional sequencing performed using M13 primers and a Big Dye Terminator Cycler Sequencing kit v1.1 (Applied Biosystems, Warrington, U.K.). Reactions were analyzed on an ABI 3730 capillary sequencer (Applied Biosystems), and results were compared with the reference sequence (NM_000545.3) using Mutation Surveyor v3.97 (SoftGenetics, Cambridge, U.K.). Mutation testing was undertaken in family members when available to establish cosegregation. *In silico* analysis of missense mutations was performed using the software program Condel (CONsensus DELeterious score of missense single nucleotide polymorphisms) (16). Condel produces a weighted average of scores from three computational tools [SIFT, Polyphen2, and Mutation-Assessor (17–19)] and classifies missense single nucleotide polymorphisms as probably "deleterious" (i.e., pathogenic) or probably "neutral" (i.e., benign).

All analyses were performed using SPSS version 17.0. The study was performed in accordance with the latest version of the Declaration of Helsinki.

RESULTS

***HNF1A*-MODY and measures of antennary fucosylation.** In the initial study, we found marked differences in the plasma glycome profiles between 33 subjects with *HNF1A*-MODY and 41 subjects with early-onset type 2 diabetes. Fifteen of 29 glycan measures differed significantly between the two groups ($P < 0.05$) (Supplementary Table 1). Patterns were consistent with the known effects of HNF1A on fucosylation (1), in that subjects with loss-of-function mutations in *HNF1A* were characterized by an increase in the proportion of glycans without antennary fucose.

Validation study. For the validation study, we focused on DG9 and DG8 as measures of triantennary glycans with and without antennary fucose, respectively (Fig. 1). Therefore, the DG9-glycan index [DG9-to-(DG8+DG9) ratio] summarizes the proportion of triantennary glycans that are fucosylated. As well as consistency with the existing data on HNF1A effects on fucosylation (1) and strong evidence from the initial study (Supplementary Table 1), triantennary glycans are not affected by the removal of fibrinogen during coagulation (G. Lauc, unpublished observations), allowing our validation studies to include both serum and plasma samples.

The distributions of DG9-glycan index measures for the 709 individuals in the validation study differed significantly between the subject groups (Fig. 2; Supplementary Table 2). Median (interquartile range) DG9-glycan index levels were substantially lower in subjects with *HNF1A*-MODY [0.09 (0.06–0.13)] than in those with young-onset type 2 diabetes [0.25 (0.18–0.33); $P = 1 \times 10^{-39}$ vs. *HNF1A*-MODY], type 1 diabetes [0.28 (0.20–0.34);, $P = 1 \times 10^{-34}$ vs. *HNF1A*-MODY], or *GCK*-MODY [0.25 (0.18–0.31); $P = 5 \times 10^{-32}$ vs. *HNF1A*-MODY]. DG9-glycan index levels in subjects with *HNF1A*-MODY were also lower when compared with controls [0.24 (0.19–0.29); $P = 1 \times 10^{-32}$ vs. *HNF1A*-MODY] and against all other diabetic patients combined [0.25 (0.18–0.31); $P = 5 \times 10^{-55}$ vs. *HNF1A*-MODY]. Consistent with evidence that *HNF4A* also regulates fucosylation (1), cases of *HNF4A*-MODY showed DG9-glycan index levels between those in *HNF1A*-MODY and other forms of diabetes [0.18 (0.09–0.24); $P = 2 \times 10^{-7}$ vs. *HNF1A*-MODY].

Adjustment for significant covariates had no appreciable impact on the magnitude or significance of differences in DG9-glycan index values between groups (Supplementary Table 2).

Receiver operating characteristic (ROC) curve analyses. To test whether glycan profiling had potential as a clinically

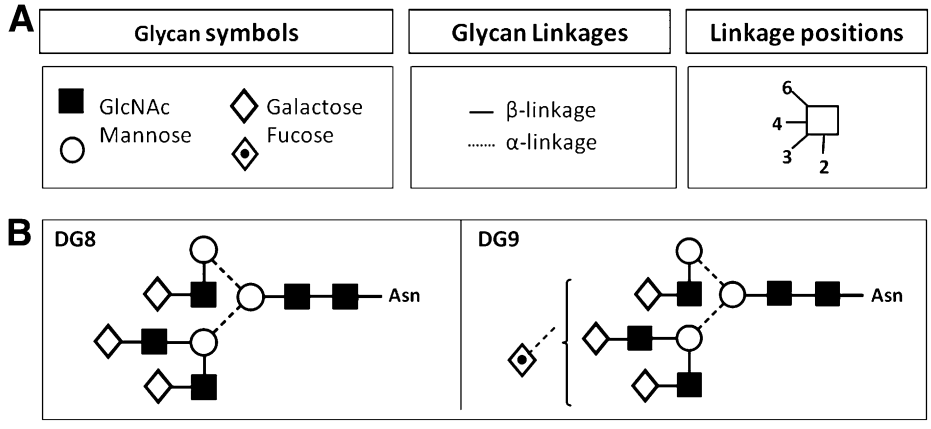


FIG. 1. *A*: Structural symbols for *N*-glycans, their linkages, and the abbreviations used. The Oxford nomenclature has been used to represent *N*-linked glycan composition and structure (20). *B*: Representations of the structures of the major glycans in DG8 and DG9 high-performance liquid chromatography peaks. Both are triantennary glycans. The only difference is the presence of a terminal fucose residue (◈) attached to one of the antennae of the glycans in DG9.

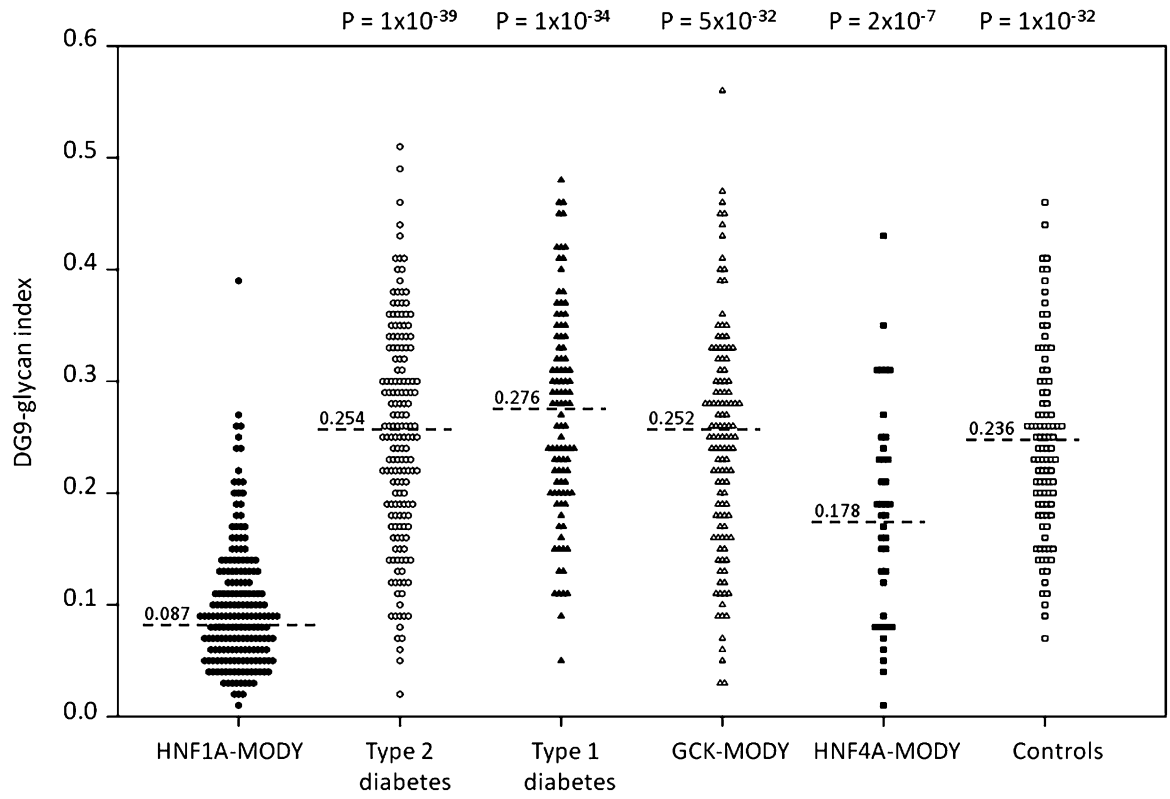


FIG. 2. Dot histograms illustrating the DG9-glycan index in different diabetes subtypes and nondiabetic control subjects. Subjects are represented by the following symbols: ● = *HNF1A*-MODY; ○ = type 2 diabetes; ▲ = type 1 diabetes; △ = *GCK*-MODY (*GCK*-MODY); ■ = *HNF4A*-MODY (*HNF4A*-MODY); □ = nondiabetic controls. *P* values are calculated by Mann-Whitney *U* tests in comparison with subjects with *HNF1A*-MODY. The median value of the DG9-glycan index for each diabetes subtype is highlighted adjacent to a *black dashed line*.

valid screening test, C statistic measures of discriminative accuracy were derived from ROC curve analyses (Fig. 3; Supplementary Table 3). The C statistic was 0.94 for *HNF1A*-MODY against type 1 diabetes and 0.91 for *HNF1A*-MODY against early-onset type 2 diabetes. Similar discrimination was observed for the comparison of *HNF1A*-MODY and *GCK*-MODY (C statistic 0.90), but the DG9-glycan index performed less well in differentiating *HNF4A*-MODY and *HNF1A*-MODY (C statistic 0.76). These measures were not affected by sample type (serum vs. plasma; Supplementary Table 3).

We have reported previously that high-sensitivity C-reactive protein (hs-CRP) is a sensitive and specific biomarker for *HNF1A*-MODY (21,22). In the validation dataset, the ability of the DG9-glycan index to discriminate between *HNF1A*-MODY and type 2 diabetes was comparable with hs-CRP (C statistic 0.91 and 0.94, respectively). However, the DG9-glycan index provided superior discrimination of *HNF1A*-MODY from type 1 diabetes compared with hs-CRP (C statistic 0.94 vs. 0.83 for hs-CRP).

Other glycan measures. As described earlier, our primary validation analyses focused on the DG9-glycan index. However, the availability of full glycome profiles for the validation samples allowed us to explore the relative performance of other measures highlighted in the initial study. Other glycan ratios, such as DG7 to (DG5+DG6), representing the proportion of biantennary fucosylated glycans, offered good discrimination between diabetes subtypes (Supplementary Table 3). The DG7-to-(DG5+DG6) ratio provided near perfect discrimination between *HNF1A*-MODY and type 2 diabetes (C statistic >0.99) in plasma samples, but it performed less well in analyses of serum samples (C statistic 0.78).

Correlates of the DG9-glycan index within the *HNF1A*-MODY group. Eighteen subjects within the *HNF1A*-MODY group did not have diabetes when glycans were sampled. DG9-glycan index levels were not different in the patients with *HNF1A*-MODY with and without diabetes ($P = 0.45$). This further confirms that these changes in antennary fucosylation are specific to loss-of-function *HNF1A* mutations and are unrelated to dysglycemia. DG9-glycan index levels were not correlated with the age at diabetes diagnosis ($P = 0.38$).

Analysis by mutation type showed median DG9-glycan index was lower in subjects with *HNF1A*-MODY with protein-changing mutations than those with truncating mutations (0.08 vs. 0.10; $P = 0.003$) (Table 2). This difference in DG9-glycan index levels did not remain significant when analysis was restricted to one individual per family (the proband), a maneuver that renders the observations independent at the expense of a reduced sample size ($P = 0.18$). Analysis by functional domain indicated lower median DG9-glycan index in subjects with *HNF1A*-MODY with missense mutations affecting the dimerization/DNA-binding domains compared with those with missense mutations in the transactivation domain (0.08 vs. 0.13; $P = 0.04$). The latter relationship was consistent whether we considered all subjects with *HNF1A*-MODY or only the probands ($P = 0.04$ and 0.002, respectively). The *HNF1A* isoform affected had no effect on DG9-glycan index levels ($P = 0.30$).

We also examined whether the DG9-glycan index could be used as a marker of *HNF1A* function to provide additional

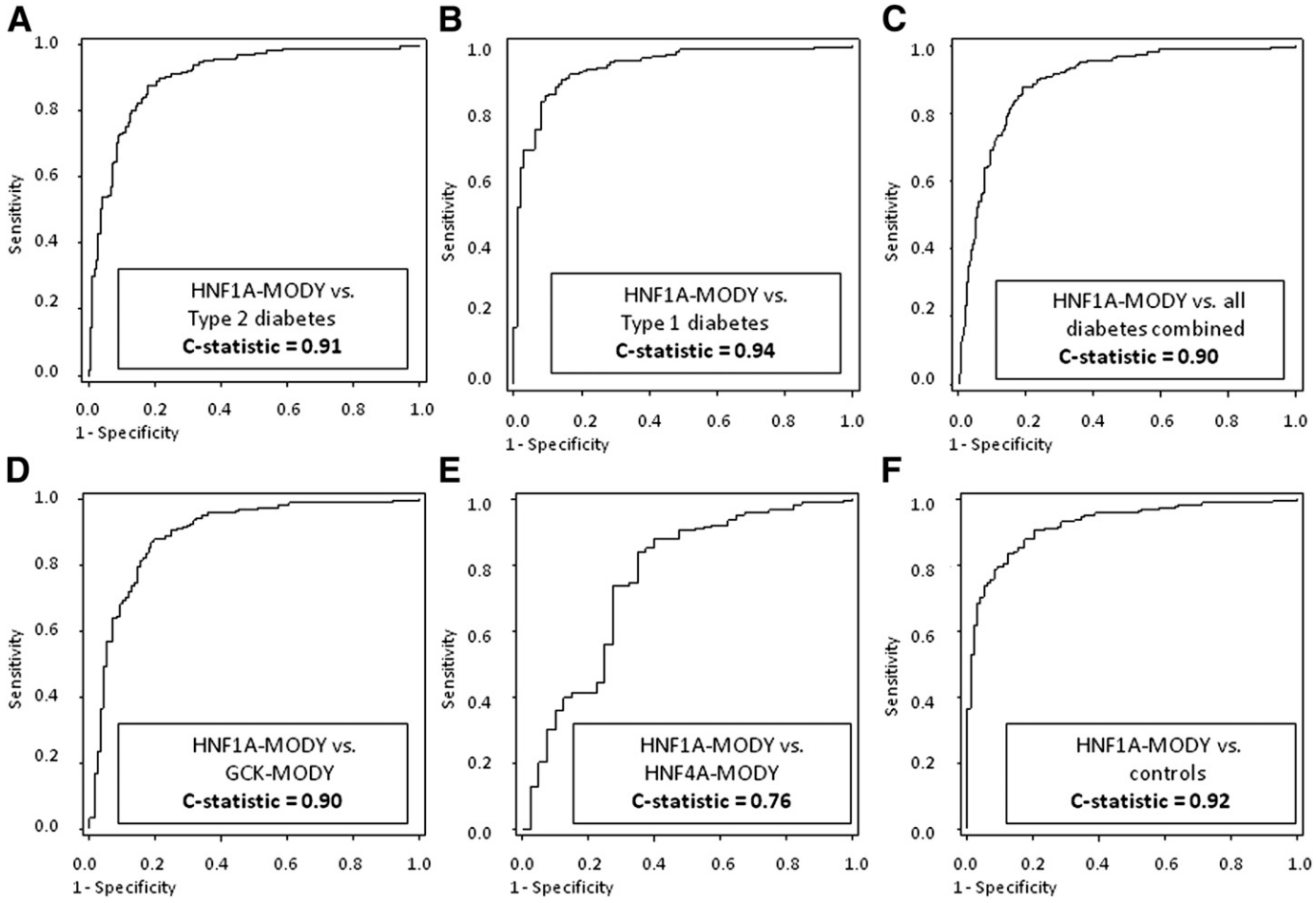


FIG. 3. ROC curves illustrating the performance of the DG9-glycan index to discriminate *HNF1A*-MODY and type 2 diabetes (*A*); *HNF1A*-MODY and type 1 diabetes (*B*); *HNF1A*-MODY and other diabetes subtypes combined (*C*); *HNF1A*-MODY and *GCK*-induced MODY (*GCK*-MODY) (*D*); *HNF1A*-MODY and *HNF4A*-MODY (*E*); and *HNF1A*-MODY and nondiabetic control subjects (*F*).

evidence regarding pathogenicity. All missense *HNF1A* mutations in this study were classified as probably "deleterious" (i.e., pathogenic) or probably "neutral" (i.e., benign) using the software program Condel. Four missense *HNF1A* mutations (in five subjects) had Condel scores within the benign range: c.92G>A p.Gly31Asp, c.142G>A p.Glu48Lys, c.871C>A p.Pro291Thr, and c.1816G>A p.Gly606Ser. Family data for these mutations were generally less supportive of pathogenicity because of incomplete cosegregation, later age at diagnosis of diabetes, or unavailability of family

TABLE 2
DG9-glycan index levels in subjects with *HNF1A*-MODY according to type and position of *HNF1A* mutation

	Subjects with *HNF1A*-MODY (*n*)	DG9-glycan index*	*P*	Unrelated subjects with *HNF1A*-MODY (*n*)	DG9-glycan index*	*P*
Truncating mutations	83	0.10 (0.07–0.21)	0.003	39	0.10 (0.07–0.14)	0.180
Protein-changing mutations	105	0.08 (0.05–0.21)		44	0.08 (0.05–0.14)	
Classified by isoform affected						
Exons 1–6 [isoforms *HNF1A*(*A*), (*B*) and (*C*)]	93	0.08 (0.05–0.21)	0.300	42	0.08 (0.05–0.11)	N/A
Exon 7 [isoforms *HNF1A*(*A*) and (*B*)]	11	0.09 (0.05–0.18)		1	0.13	
Exons 8–10 [isoform *HNF1A*(*A*) only]	1	0.21		1	0.21	
Classified by affected functional domain						
Dimerization/DNA-binding domains†	62	0.08 (0.05–0.17)	0.043	26	0.08 (0.05–0.11)	0.002
Transactivation domain‡	17	0.13 (0.08–0.25)		4	0.21 (0.17–0.24)	

P values calculated using Mann-Whitney *U* and Kruskal-Wallis tests. The Mann-Whitney *U* test was used to test truncating versus protein-changing mutations and dimerization/DNA-binding versus transactivation domains, while the Kruskal-Wallis test was used to compare the isoforms. *Data expressed as median (25th–75th centiles). †Dimerization domain (amino acids 1–32) and DNA-binding domain (amino acids 91–281). ‡Transactivation domain (amino acids 282–631).

members for testing. The median (interquartile range) DG9-glycan index levels in these five subjects were significantly higher compared with the remainder of the group with *HNF1A* missense mutations [0.16 (0.11–0.21) vs. 0.08 (0.05–0.11); $P = 0.01$].

Clinical potential of the DG9-glycan index. To examine the performance of the DG9-glycan index as a diagnostic screen in clinical practice, we based analyses on U.K. data showing that subjects with unrecognized *HNF1A*-MODY account for approximately 4% of young-onset type 2 diabetes (diagnosed ≤45 years) and 1% of type 1 diabetes (13). On the basis of the validation study, we estimate that a diagnostic threshold for the DG9-glycan index of 0.16 confers 88% sensitivity and 81% specificity for the discrimination of *HNF1A*-MODY from young-onset type 2 diabetes and 88% sensitivity and 88% specificity for equivalent comparisons with type 1 diabetes. In contrast, an age younger than 25 years at diagnosis of diabetes, which is the most widely used diagnostic feature for MODY, has lower sensitivity (64%) but higher specificity (99%) for the discrimination of *HNF1A*-MODY from young-onset type 2 diabetes. We calculate that a patient with diabetes who is diagnosed before or at age 45 years, who has an existing clinical label of type 2 diabetes, and who is found to have a DG9-glycan index ≤0.16 has a posttest probability of harboring an underlying *HNF1A* mutation of 16%, whereas the same patient with a DG9-glycan index >0.16 has a 1% posttest probability of having unrecognized *HNF1A*-MODY.

***HNF1A*-MODY case-finding in diabetic subjects.** The results of the validation study indicated that a diagnostic threshold of 0.16 for the DG9-glycan index provided optimum discrimination from both type 1 and type 2 diabetes. *HNF1A* sequencing was performed in 57 subjects with young-onset diabetes and who were clinically labeled as having type 1 or type 2 diabetes, all of whom had a DG9-glycan index ≤0.16. These subjects were either from the initial or validation studies or recruited from general population cohorts in Croatia and Scotland. Three of these 57 individuals were found to have *HNF1A* mutations.

The first proband (Supplementary Fig. 1*A*) was heterozygous for the missense mutation c.608G>A p.Arg203His in exon 3, which previously has been shown to be causal for MODY (23). The phenotype is consistent with *HNF1A*-MODY with a two-generation history of young-adult onset diabetes, and, although treated with insulin and labeled with type 1 diabetes since diagnosis at age 31, this patient was found to have residual endogenous insulin secretion 17 years after diagnosis (C-peptide, 0.27 nmol/L). Two sisters with insulin-treated diabetes also carry the Arg203His mutation. One sister has stopped basal-bolus insulin successfully after the diagnosis of MODY and her diabetes is well controlled with gliclazide; the proband is considering a similar change in therapy. In the second proband (Supplementary Fig. 1*B*), a novel missense variant c.139G>C, p.Gly47Arg was identified in exon 1. This subject, now aged 53 years, was diagnosed with presumed type 2 diabetes at age 37. Her mother, diagnosed with type 2 diabetes at age 70, also carries the mutation. A normoglycemic son, currently 29 years old, does not carry the mutation. In the third proband (Supplementary Fig. 1*C*), a novel missense mutation c.751G>A, p.Ala251Thr was found in exon 4. This proband was diagnosed with diabetes at age 43 and is being managed with sulfonylurea (and metformin) therapy 25 years later. Sensitivity to sulfonylureas and maintenance of diabetes control for many years while taking these agents is typical of subjects with *HNF1A*-MODY. The proband's mother and a sister had diabetes in old age.

To assess pathogenicity for the two novel variants (Gly47Arg and Ala251Thr), we first established that both variants were absent from 400 normal chromosomes and the October 2011 release of the consensus calls for the 1000 Genomes Project (24). The Ala251Thr variant is not reported in the National Heart, Lung, and Blood Institute's Exome Sequencing Project ($N = 6503$; accessed June 2012 via the Exome Variant Server) (25), whereas the Gly47Arg variant is reported in a single European American individual. The phenotypic characteristics of this subject are not declared: because the Exome Sequencing Project includes cases of metabolic and cardiovascular disease, this finding does not exclude a pathogenic role for this variant. Second, we assessed whether the mutated residues were conserved across species. Both Gly47 and Ala251 are highly conserved: Ala251 is conserved in eight of the nine orthologs tested, including frog, chicken, and six mammalian orthologs, whereas Gly47 is conserved in seven of the nine orthologs tested, including *Xenopus*, chicken, and five mammalian orthologs. *In silico* prediction software was more ambiguous: Condel predicted that the variant Gly47Arg is "neutral" and that Ala251Thr is "deleterious." Overall, we regard the clinical and bioinformatic data for these two variants as supportive, but not conclusive, evidence in favor of pathogenicity. The low DG9-glycan index levels associated with Gly47Arg and Ala251Thr (0.12 and 0.15, respectively) provide additional support for pathogenicity. Definitive evidence of a causal role would require more extensive clinical (including the detection of these mutations in additional MODY families) and functional studies.

DISCUSSION

The study confirms the hypothesis that the glycan profile of plasma proteins is altered substantially in those with *HNF1A* mutations. We also demonstrated that these differences could be exploited as biomarkers in diabetes diagnostics and showed that the DG9-glycan index can discriminate *HNF1A*-MODY from both type 1 and type 2 diabetes.

Recent efforts to improve diagnostic performance by identifying biochemical markers specific for MODY subtypes have met with varying success (21,26–32). We recently demonstrated that individuals with *HNF1A*-MODY have low levels of C-reactive protein and that hs-CRP assays can discriminate well between *HNF1A*-MODY and both type 2 diabetes and *HNF4A*-MODY (21,22,33). However, hs-CRP does not provide good discrimination between *HNF1A*-MODY and type 1 diabetes. Furthermore, hs-CRP is an acute inflammatory marker, and diagnostic discrimination can be disturbed by intercurrent infection. Potential advantages of the DG9-glycan index in this context include stability over time (10) and differentiation of *HNF1A*-MODY from both common types of diabetes. Although there is some indication that glycan profiles are affected by acute inflammation (34), all four subjects with *HNF1A*-MODY and elevated levels of hs-CRP (>10 mg/L) in the current study had DG9-glycan indices below 0.16. This suggests that the DG9-glycan index is less prone to spurious elevation from intercurrent infection than hs-CRP, although this will require confirmation in larger numbers.

The ability to discriminate between *HNF1A*-MODY and type 1 diabetes in subjects with recently diagnosed diabetes is particularly important because diagnostic misclassification can lead to the unwarranted decision to recommend lifelong therapy with exogenous insulin. Detectable C-peptide can indicate *HNF1A*-MODY rather than type 1 diabetes of long duration, but it is not helpful close to a diagnosis of diabetes because a substantial proportion of type 1 diabetes patients retain some production of endogenous insulin (35). In this study we did not explicitly examine type 1 diabetes during the honeymoon period; however, glycan profiles are stable within an individual over time, which suggests these measures will continue to provide useful discriminative power from type 1 diabetes close to diagnosis (36). In principle, therefore, the addition of the DG9-glycan index to existing biomarkers such as hs-CRP (21), 1,5-anhydroglucitol (29), pancreatic autoantibodies (37), and C-peptide (13,35) should improve the capacity for clinical discrimination of all major diabetes subtypes.

With respect to clinical utility, our study showed that it was possible to identify subjects with *HNF1A* mutations using the DG9-glycan index. Given the high sensitivity and specificity of the DG9-glycan index, the proportion of cases found was lower than might have been expected (3 of 57; 5%). There are several possible explanations. The first is that the estimate of discriminatory power based on the C statistics calculated during the validation study is inflated, perhaps because of overfitting (38). An alternative explanation is that discovery of novel cases of *HNF1A*-MODY might have been compromised in our study samples by the extent of existing clinical investigation (e.g., approximately 30% of the Oxford samples had undergone *HNF1A* sequencing). Another intriguing possibility is that a small proportion of individuals in large, population-based cohorts have low DG9-glycan index levels due to low-frequency alleles in genes other than *HNF1A*, which play an important role in protein fucosylation, such as those encoding the fucosyltransferases *FUT6* and other genes (1). Further validation of the DG9-glycan index in unselected groups of subjects with young-onset diabetes will be required to assess performance in a more typical clinical scenario in which extensive prior screening for monogenic disease has not been undertaken.

The use of the DG9-glycan index in clinical practice currently is restricted by the cost and limited availability of accurate glycan profiling. Clinical translation is, therefore, dependent on the implementation of a focused assay for specific glycan moieties rather than the global chromatographic profiling used in this study. The DG9-glycan index seems to be the most promising candidate from this study, although the validation study suggests that other measures [such as the DG7-to-(DG5+DG6) ratio] might have superior performance when plasma (rather than serum) samples are available—a finding that requires further confirmation.

There is some evidence for a genotype-phenotype relationship of glycan levels within subjects with *HNF1A*-MODY, including that DG9-glycan index levels are lower in those with protein-changing rather than truncating *HNF1A* mutations. We reported a similar relationship in our studies of hs-CRP; in this case, restricting the analysis to independent probands did not abolish the significance (22). These observations would be consistent with a dominant-negative effect, which has been reported for some *HNF1A* mutations (39,40). Missense mutations disrupting the DNA-binding and dimerization domains have significantly lower DG9-glycan index levels than those affecting the transactivation domain. This indicates a more severe functional outcome of *HNF1A* mutations that disturb DNA binding. This finding is consistent with previous reports of decreased age at diabetes diagnosis in subjects with *HNF1A*-MODY who have missense mutations within the DNA-binding/dimerization domains compared with those in the transactivation domain (14,15).

This study primarily aimed to define the role of glycans as a potential adjunct in the diagnosis of *HNF1A*-MODY and, in particular, the capacity of this measure to improve the targeting of (relatively expensive and therefore restricted) diagnostic sequencing. However, anticipated reductions in the cost of diagnostic sequencing are likely to encourage more liberal access to *HNF1A* sequencing, including many individuals with a relatively low risk of MODY. In this setting, the problem shifts toward correctly interpreting the clinical significance of the many novel *HNF1A* variants that will be uncovered, and we provide preliminary evidence that biomarkers of *HNF1A*, such as the DG9-glycan index, may prove useful in this context.

In summary, the use of the DG9-glycan index both as a biomarker for *HNF1A*-MODY and as a promising biochemical measure of *HNF1A* function represents a compelling example of the potential for rapid clinical translation of a genetic discovery originating from a genome-wide association analysis.

ACKNOWLEDGMENTS

Recruitment in Oxford was supported by the National Institute for Health Research (NIHR) Thames Valley Diabetes Local Research Network, part of the U.K. Clinical Research Network. The study was funded by the Oxford (NIHR) Biomedical Research Centre, the European Community FP7 programs CEED3 (HEALTH-F2-2008-223211), ENGAGE (HEALTHF4-2007-201413), and the Medical Research Council (81696). Glycan analysis was supported by the Croatian Ministry of Science, Education and Sport (Grant 309-0061194-2023), the Croatian Science Foundation (Grant 04-47), and grants from the European Commission GlycoBioM (contract 259869) and HighGlycan (contract 278535). The Danish portion of the study was supported by grants from the Lundbeck Foundation Centre for Applied Medical Genomics in Personalized Disease Prediction, Prevention and Care (LuCAMP), the Danish Diabetes Association, and the Danish Research Council. The CROATIA-Vis study was supported by grants from the Medical Research Council U.K. and Ministry of Science, Education and Sport of the Republic of Croatia (108-1080315-0302) and the European Union framework program 6 European Special Populations Research Network project (contract LSHG-CT-2006-018947). ORCADES was supported by the Chief Scientist Office of the Scottish Government, the Royal Society, and the European Union framework program 6 European Special Populations Research Network project (contract LSHG-CT-2006-018947).

A.L.G. is a Wellcome Trust Senior Fellow in Basic Biomedical Science (095101/Z/10/Z). B.A. and I.Ru. received funding from the European Union FP7 EuroGlycoArrays ITN (contract no. 215536). I.Kl., D.G., and J.S. received grant support from ERDF (Transendogen/26240220051), VEGA 2/0151/11, and APVV-148-10. E.T. and P.R.N. received funding from the University of Bergen, Helse Vest, Innovest, and the Research Council of Norway. N.D.H. and A.F.W. received financial support from the U.K. Medical

Research Council. K.R.O. is a Clinician Scientist funded by the NIHR. H.C., J.K., G.L., I.Ru., P.M.R., and A.F.W. are listed as inventors on a patent application PCT/EP2011/067112, which covers the use of the analysis of antennary fucose for the diagnosis of *HNF1A*-MODY. No other potential conflicts of interest relevant to this article were reported.

No funding bodies played any role in the study design, data collection, and analysis or the preparation of or decision to publish the manuscript.

G.T., J.E.H., J.J.K., M.N., I.Ru., A.L.G., C.H., B.A., O.P., J.F.W., N.D.H., H.C., M.I.M., P.M.R., K.R.O., G.L., and A.F.W. wrote the manuscript. G.T., R.M.R., O.P., S.A.M., D.P., L.Z., I.Ko., T.H., D.G., E.T., M.W.J.S., T.N., J.S., I.Kl., O.B.P., P.R.N., S.H.W., U.G., J.F.W., M.I.M., and K.R.O. enrolled patients. G.T., J.E.H., A.L.G., C.H., A.E., N.H., A.J.B., S.A.M., K.R.O., and G.L. analyzed and interpreted the data. J.J.K., M.N., B.A., A.M., N.H., M.P., A.J.B., D.G., E.T., I.Re., and O.G. acquired the data. I.Ru., A.L.G., C.H., J.F.W., N.D.H., H.C., M.I.M., P.M.R., K.R.O., G.L., and A.F.W. conceptualized and designed the study. All authors read and approved the final manuscript.

H.C., M.I.M., P.M.R., K.R.O., G.L., and A.F.W. are the guarantors of this work and, as such, had full access to all the data in the study and take responsibility for the integrity of the data and the accuracy of the data analysis.

This study was presented in abstract form at the 71st Scientific Sessions of the American Diabetes Association, San Diego, California, 24–28 June 2011, and the 47th Annual Meeting of the European Association for the Study of Diabetes, Lisbon, Portugal, 12–16 September 2011.

ORCADES acknowledges the invaluable contributions of Lorraine Anderson (Centre for Population Health Sciences, University of Edinburgh, Edinburgh, U.K.); the research nurses in Orkney; and the administrative team in Edinburgh. The CROATIA-Vis and CROATIA-Korcula studies acknowledge the invaluable contributions of the recruitment team (including those from the Institute of Anthropological Research in Zagreb) in Vis and Korcula, the administrative teams in Croatia and Edinburgh, and the people of Vis and Korcula. DNA extraction for ORCADES was performed by the Genetics Core Laboratory at the Wellcome Trust Clinical Research Facility, WGH, Edinburgh, Scotland. The Slovak investigators acknowledge the expert technical assistance of Dr. Miroslava Huckova (Diabgene, Slovak Academy of Sciences, Bratislava). M.S. acknowledges Jill Little, Royal Infirmary of Edinburgh, for her work in the MODY clinic. The authors thank the National Heart, Lung, and Blood Institute GO Exome Sequencing Project and its ongoing studies, which produced and provided exome variant calls for comparison: the Lung GO Sequencing Project (HL-102923), the WHI Sequencing Project (HL-102924), the Broad GO Sequencing Project (HL-102925), the Seattle GO Sequencing Project (HL-102926), and the Heart GO Sequencing Project (HL-103010). The authors thank the patients and families participating in the study and are grateful for the help of the research nurses in subject recruitment.

REFERENCES

1. Lauc G, Essafi A, Huffman JE, et al. Genomics meets glycomics-the first GWAS study of human N-Glycome identifies HNF1α as a master regulator of plasma protein fucosylation. PLoS Genet 2010;6:e1001256
2. Apweiler R, Hermjakob H, Sharon N. On the frequency of protein glycosylation, as deduced from analysis of the SWISS-PROT database. Biochim Biophys Acta 1999;1473:4–8
3. Cummings RD. The repertoire of glycan determinants in the human glycome. Mol Biosyst 2009;5:1087–1104
4. Trinchera M, Malagolini N, Chiricolo M, et al. The biosynthesis of the selectin-ligand sialyl Lewis x in colorectal cancer tissues is regulated by fucosyltransferase VI and can be inhibited by an RNA interference-based approach. Int J Biochem Cell Biol 2011;43:130–139
5. Thanabalasingham G, Owen KR. Diagnosis and management of maturity onset diabetes of the young (MODY). BMJ 2011;343:d6044
6. Ellard S, Bellanné-Chantelot C, Hattersley AT; European Molecular Genetics Quality Network (EMQN) MODY group. Best practice guidelines for the molecular genetic diagnosis of maturity-onset diabetes of the young. Diabetologia 2008;51:546–553
7. Shields BM, Hicks S, Shepherd MH, Colclough K, Hattersley AT, Ellard S. Maturity-onset diabetes of the young (MODY): how many cases are we missing? Diabetologia 2010;53:2504–2508
8. Pearson ER, Starkey BJ, Powell RJ, Gribble FM, Clark PM, Hattersley AT. Genetic cause of hyperglycaemia and response to treatment in diabetes. Lancet 2003;362:1275–1281
9. Shepherd M, Shields B, Ellard S, Rubio-Cabezas O, Hattersley AT. A genetic diagnosis of HNF1A diabetes alters treatment and improves glycaemic control in the majority of insulin-treated patients. Diabet Med 2009;26:437–441
10. Gornik O, Wagner J, Pucić M, Knezević A, Redzic I, Lauc G. Stability of N-glycan profiles in human plasma. Glycobiology 2009;19:1547–1553
11. Knezević A, Polasek O, Gornik O, et al. Variability, heritability and environmental determinants of human plasma N-glycome. J Proteome Res 2009;8:694–701
12. Royle L, Campbell MP, Radcliffe CM, et al. HPLC-based analysis of serum N-glycans on a 96-well plate platform with dedicated database software. Anal Biochem 2008;376:1–12
13. Thanabalasingham G, Pal A, Selwood MP, et al. Systematic assessment of etiology in adults with a clinical diagnosis of young-onset type 2 diabetes is a successful strategy for identifying maturity-onset diabetes of the young. Diabetes Care 2012;35:1206–1212
14. Harries LW, Ellard S, Stride A, Morgan NG, Hattersley AT. Isomers of the TCF1 gene encoding hepatocyte nuclear factor-1 alpha show differential expression in the pancreas and define the relationship between mutation position and clinical phenotype in monogenic diabetes. Hum Mol Genet 2006;15:2216–2224
15. Bellanné-Chantelot C, Carette C, Riveline JP, et al. The type and the position of HNF1A mutation modulate age at diagnosis of diabetes in patients with maturity-onset diabetes of the young (MODY)-3. Diabetes 2008;57:503–508
16. González-Pérez A, López-Bigas N. Improving the assessment of the outcome of nonsynonymous SNVs with a consensus deleteriousness score, Condel. Am J Hum Genet 2011;88:440–449
17. Kumar P, Henikoff S, Ng PC. Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm. Nat Protoc 2009;4:1073–1081
18. Adzhubei IA, Schmidt S, Peshkin L, et al. A method and server for predicting damaging missense mutations. Nat Methods 2010;7:248–249
19. Reva B, Antipin Y, Sander C. Predicting the functional impact of protein mutations: application to cancer genomics. Nucleic Acids Res 2011;39:e118
20. Harvey DJ, Merry AH, Royle L, Campbell MP, Dwek RA, Rudd PM. Proposal for a standard system for drawing structural diagrams of N- and O-linked carbohydrates and related compounds. Proteomics 2009;9:3796–3801
21. Owen KR, Thanabalasingham G, James TJ, et al. Assessment of high-sensitivity C-reactive protein levels as diagnostic discriminator of maturity-onset diabetes of the young due to HNF1A mutations. Diabetes Care 2010;33:1919–1924
22. Thanabalasingham G, Shah N, Vaxillaire M, et al. A large multi-centre European study validates high-sensitivity C-reactive protein (hsCRP) as a clinical biomarker for the diagnosis of diabetes subtypes. Diabetologia 2011;54:2801–2810
23. Ellard S, Colclough K. Mutations in the genes encoding the transcription factors hepatocyte nuclear factor 1 alpha (HNF1A) and 4 alpha (HNF4A) in maturity-onset diabetes of the young. Hum Mutat 2006;27:854–869
24. 1000 Genomes Project Consortium. A map of human genome variation from population-scale sequencing. Nature 2010;467:1061–1073
25. National Heart, Lung, and Blood Institute. NHLBI Exome Sequencing Project (ESP. Exome Variant Server [Internet]. Seattle, WA, NESPE. Available from http://evs.gs.washington.edu/EVS/. Accessed June 2012
26. Bingham C, Ellard S, Nicholls AJ, et al. The generalized aminoaciduria seen in patients with hepatocyte nuclear factor-1alpha mutations is a feature of all patients with diabetes and is associated with glucosuria. Diabetes 2001;50:2047–2052

27. Stride A, Pearson ER, Brown A, Gooding K, Castleden HA, Hattersley AT. Serum amino acids in patients with mutations in the hepatocyte nuclear factor-1 alpha gene. Diabet Med 2004;21:928–930
28. Cervin C, Axler O, Holmkvist J, et al. An investigation of serum concentration of apoM as a potential MODY3 marker using a novel ELISA. J Intern Med 2010;267:316–321
29. Pal A, Farmer AJ, Dudley C, et al. Evaluation of serum 1,5 anhydroglucitol levels as a clinical test to differentiate subtypes of diabetes. Diabetes Care 2010;33:252–257
30. Richter S, Shih DQ, Pearson ER, et al. Regulation of apolipoprotein M gene expression by MODY3 gene hepatocyte nuclear factor-1alpha: haploinsufficiency is associated with reduced serum apolipoprotein M levels. Diabetes 2003;52:2989–2995
31. Skupien J, Gorczynska-Kosiorz S, Klupa T, et al. Clinical application of 1,5-anhydroglucitol measurements in patients with hepatocyte nuclear factor-1alpha maturity-onset diabetes of the young. Diabetes Care 2008;31:1496–1501
32. Skupien J, Kepka G, Gorczynska-Kosiorz S, et al. Evaluation of Apolipoprotein M Serum Concentration as a Biomarker of HNF-1alpha MODY. Rev Diabet Stud 2007;4:231–235
33. McDonald TJ, Shields BM, Lawry J, et al. High-sensitivity CRP discriminates HNF1A-MODY from other subtypes of diabetes. Diabetes Care 2011; 34:1860–1862
34. Gornik O, Royle L, Harvey DJ, et al. Changes of serum glycans during sepsis and acute pancreatitis. Glycobiology 2007;17:1321–1332
35. Besser RE, Shepherd MH, McDonald TJ, et al. Urinary C-peptide creatinine ratio is a practical outpatient tool for identifying hepatocyte nuclear factor 1-alpha/hepatocyte nuclear factor 4-alpha maturity-onset diabetes of the young from long-duration type 1 diabetes. Diabetes Care 2011;34:286–291
36. Knezevic A, Gornik O, Polasek O, et al. Effects of aging, body mass index, plasma lipid profiles, and smoking on human plasma N-glycans. Glycobiology 2010;20:959–969
37. McDonald TJ, Colclough K, Brown R, et al. Islet autoantibodies can discriminate maturity-onset diabetes of the young (MODY) from Type 1 diabetes. Diabet Med 2011;28:1028–1033
38. Copas JB. Overestimation of the receiver operating characteristic curve for logistic regression. Biometrika 2002;89:315–331
39. Vaxillaire M, Abderrahmani A, Boutin P, et al. Anatomy of a homeoprotein revealed by the analysis of human MODY3 mutations. J Biol Chem 1999; 274:35639–35646
40. Yoshiuchi I, Yamagata K, Yang Q, et al. Three new mutations in the hepatocyte nuclear factor-1alpha gene in Japanese subjects with diabetes mellitus: clinical features and functional characterization. Diabetologia 1999;42:621–626