Functional annotation of variants in monogenic diabetes genes to support diagnostic interpretation



A thesis submitted to the University of Oxford in candidature for the degree of Doctor of Philosophy

by

Sara Althari

Lincoln College, University of Oxford

and

Oxford Centre for Diabetes, Endocrinology and Metabolism

Radcliffe department of Medicine, University of Oxford

Hilary Term 2018

Acknowledgments

هذا من فضل ربي... اللهم لك الحمد حمداً كثيراً طيباً مباركاً فيه يليق بجلال وجهك وعظيم سلطانك

This thesis is dedicated to my angel mother, Khawla Alkuraya, who is my idol, my best friend, my mentor, and my biggest inspiration. I am the product of her boundless love and unconditional support. I love you, Mama. Thank you.

I am deeply grateful to everyone who has shared this incredible journey with me. My sincerest thanks to my wonderful DPhil supervisors Prof Anna Gloyn and Prof Mark McCarthy for their support, guidance, patience and generosity. Their outstanding mentorship allowed me to grow and flourish both personally and professionally. Thank you for looking out for me, giving me the tools to excel and succeed, sharing your wisdom and brilliance, and for giving me the opportunities to be the intellectual driver of exciting and cutting-edge research.

Special thanks to some special souls: Nic, Nat, Mvdb, and Fernando. Nic, my fellow ENFP, I could not have done it without you. Thank you for the soul food (both literal - in the form of Brent's cooking - and metaphorical – in the form of deep existential spiritual chats). Nat, OPA! Thanks for all the good times, for the care, love and support. I appreciate you and cherish our friendship. Martijn, where do I begin? Thank you for being a brilliant teacher, friend and mentor. You trained me well (check out the contents of this thesis!). I dedicate every statistical and computational analysis in here to you. Fernando, my fellow MAVE-RAVER, my dear friend, words fail to convey my gratitude. I could not have picked a better co-pilot on this flight to functional genomic glory! Thank you for your dedication, sincerity, and for your art, your kindness and generosity.

I would also like to extend my sincere gratitude to the inimitable Prof Jose Florez for being a warm and gracious host during my time as a visiting graduate student at the Broad Institute of Harvard & MIT. I feel deeply honoured and privileged to call Jose a mentor. Many thanks also to the brilliant MODY Team at Exeter – especially Prof Sian Ellard and Kevin Colclough (aka Kev) - for their generosity and for providing an exceptionally warm, engaging, and uplifting collaborative research environment during my brief visit.

To Jen, my best friend, I express my love and gratitude through a series of hashtags that only we understand: #WHOAH #saveme #fluffsandsmokes #liesofahuttublady #lindafromHR #onlyicanbetheduckmeme #delightedtoshare #proudtoannounce #vicious #rugged #rigid #hues #URGENT #thanksfortheaesthetic #jeanbutton #smackonadress #muffins #workfromhome. Our friendship is the dream. To the posse (Maayan, Nora, Almut, Cloe and Kathy) thanks for all the laughs and the wonderful memories. To the original OCDEM gang, Clarissa, Neelam, Soren, Natasha, Hesh, thanks for all the great times. To the entire Gloyn-McCarthy team, thank you for making my experience so special.

Last, but never least, I would like to express my undying gratitude to my family... the ultimate support system... the greatest blessing. Thank you so much.

Declaration

The work in this thesis is the original work of the author. Experiments were conducted in the Oxford Centre for Diabetes Endocrinology and Metabolism, University of Oxford, under the supervision of Professor Anna L. Gloyn and Professor Mark McCarthy. Funding was provided by the i) NIHR Oxford Biomedical Research Centre, ii) Wellcome Trust, iii) the Radcliffe Department of Medicine, University of Oxford, and iv) Saudi Arabian Cultural Bureau in London, Saudi Arabian Ministry of Education, Kingdom of Saudi Arabia.

Multi-dimensional *in vitro* functional follow-up of exome-detected *HNF1A* missense alleles (chapter 3) was conducted by Dr. Amanda Bennett and Dr. Jana Rundle (University of Oxford) and Laeya Najmi, Dr. Lise Bjørkhaug, and Dr. Ingvild Aukcrust (University of Bergen). The alleles were originally identified in a large multi-consortia (GoT2D and T2D-GENES) exome-sequencing study of ~13K Type 2 Diabetes (T2D) cases and controls from five ancestry groups, published in *Nature* (Fuchsberger et al, 2016).

The *HNF1A* saturation mutagenesis library (chapter 4) was synthesised by Twist Biosciences (San Francisco, USA). RNA sequencing (chapter 4) was performed by the Oxford Genomics Centre (Wellcome Centre Human Genetics, University of Oxford). A portion of *HNF1A* target validation experiments described in chapter 4 were conducted by the author at the Broad Institute of MIT and Harvard during her time there as a visiting graduate student (June-September 2016).

This thesis is approximately 50,000 words.

Functional annotation of variants in monogenic diabetes genes to support diagnostic interpretation

Sara Althari, Lincoln College

Submitted in candidature for the degree of Doctor of Philosophy

Hilary term 2018

Abstract

Maturity-Onset Diabetes of the Young (MODY) is a monogenic form of hyperglycaemia that is both genetically and clinically heterogeneous. The majority of MODY cases (80%) are caused by heterozygous damaging alleles in clinically actionable genes (*HNF1A, HNF4A and GCK*). For carriers of these alleles, an accurate molecular diagnosis informs prognosis, management and treatment strategies, and is contingent upon understanding the functional consequences of known and novel variants. High-fidelity molecular assays, which enable comprehensive and contextual functional characterisation, at scale and speed, are needed to address the interpretive challenge. Herein, such efforts, with varying degrees of throughput, are described in the context of missense variants in *HNF1A* and *GCK*.

Using multi-dimensional functional data from 73 *HNF1A* missense alleles identified through exome-sequencing of ~13K T2D cases and controls, and a combination of unsupervised learning methods, I produced high resolution variant clusters along the *HNF1A* phenotypic spectrum. To demonstrate robustness of the approach, I used the functionally annotated allele subgroups to re-evaluate molecular diagnoses in a national MODY diagnostic registry (Exeter, UK). This resulted in confident reclassification of n=7 alleles from pathogenic/likely pathogenic to VUS/likely benign in the clinical database. I also established a systematic pipeline for building a transactivation-centric Multiplexed Assay of Variant Effects (MAVE) to characterize an *HNF1A* saturation mutagenesis library. Using TM4SF4 as an endogenous functional readout in flow cytometric analyses, I was able to segregate a series of *HNF1A* alleles (n=8) based on transactivation-dependent distributions of TM4SF4-positive HepG2 cell populations: 2-8% from MODY-causal loss-of-function alleles, 30% from a low frequency allele associated with T2D risk, and 60-70% from neutral alleles and wild-type. This discriminatory capacity suggests the assay is sufficiently robust to score a pooled *HNF1A* variant library using TM4SF4 expression as a molecular signature. For variants in *GCK*, I characterised the function of two previously unannotated missense variants (hypomorphic homozygous p.E157K/p.E157K and heterozygous p.F133L/N) identified through genetic testing to aid in their clinical interpretation. Both showed catalytic and/or stability defects in *in vitro* kinetic assays consistent with GCK-MODY (p.E157K relative activity index = 0.15; p.F133L relative stability index = 0.53).

Together, these functional genomic annotation efforts help resolve the complex relationship between genotype, molecular dysfunction and clinical phenotype for alleles in *HNF1A* and *GCK* and assist in medical diagnostic interpretation.

List of common abbreviations

ATP	Adenosine Triphosphate
ATP_{km}	half maximal ATP concentration
BMI	Body Mass Index
bp	base pair
cDNA	complementary DNA
DNA	Deoxyribonucleic acid
Dox	Doxycycline
FACS	Fluorescence-activated cell sorting
GCK	Glucokinase
gDNA	genomic DNA
GKA	glucokinase activator
Glucose $S_{0.5}$	half maximal glucose concentration
GoF	gain-of-function
GWAS	Genome-wide association studies
Hb1Ac	Glycated haemoglobin
hclust	Hierarchical Clustering
HGMD	Human Gene Mutation Database
HNF1A	hepatocyte nuclear factor 1-alpha
HNF4A	hepatocyte nuclear factor 4-alpha
Kb	Kilobase
K_{cat}	maximal specific enzyme activity
LoF	loss-of-function
MAF	Minor Allele Frequency
MODY	Maturity-Onset Diabetes of the Young
NADPH	Nicotinamide adenine dinucleotide phosphate
NGS	Next generation sequencing
n_H	hill number
OHA	Oral Hypoglycaemic Agent
OR	Odds Ratio
PCA	Principal Component Analysis
PCR	polymerase chain reaction
PNDM	Permanent Neonatal Diabetes
PPARG	Peroxisome Proliferator-Activated Receptor Gamma
PTV	Protein truncating variant
qPCR	quantitative PCR
RAI	Relative Activity Index
RNA	Ribonucleic Acid
RNA-seq	messenger Ribonucleic Acid sequencing
RSI	Relative Stability Index
SDM	site-directed mutagenesis
SNP	Single Nucleotide polymorphism
SU	Sulphonylurea
T1D	Type 1 Diabetes
T2D	Type 2 Diabetes
TM4SF4	Transmembrane 4 L six family member 4
WT	wild-type

Contents

Chapter 1. Introduction

1.1 Classic definition of Maturity-Onset Diabetes of the Young (MODY).....2
1.2 Genetic culprits of MODY and estimations of prevalence.....5
1.3 Efforts to find the missed and mis-diagnosed MODY cases.....11
1.3.1 Clinical significance of an accurate molecular diagnosis of MODY..... 12
1.3.2 Current MODY detection methods.....14
1.3.3 Biomarkers for MODY detection and classification.....16
1.4 The challenge in interpretation of sequence variants in MODY genes...18
1.4.1 Variants in monogenic diabetes genes: a continuum of effects……..19
1.4.2 Assessing variant causality: contextualize with caution.....29
1.5 Establishing the functional consequences of variants in known MODY genes……30
1.5.1 *in vitro* assays to study the functional impact of MODY gene alleles……...31
1.5.2 Advances in the molecular characterization of sequence variants……..…41
1.6 Thesis Aims ……..…49

Chapter 2. General Methods

2.1 Cell culturing conditions.....51
2.2 Fluorescence *in situ* hybridisation (FISH) to characterise HNF1A copy number in HepG2 cells.....52
2.3 DNA quantification and visualisation……...53
2.4 Molecular cloning and mutagenesis……..54
2.4.1 Site-Directed Mutagenesis (SDM)……..……..54
2.4.2 Sanger sequencing.....55
2.4.3 Restriction digest……..……...55
2.4.4 Purification of digested vector by gel extraction…....……56
2.4.5 Ligation.....57
2.4.6 Polymerase Chain Reaction (PCR).....57
2.5 Transformation of competent bacteria and plasmid DNA isolation.....59
2.5.1 Transformation of competent Escherichia coli.....59
2.5.2 Transformation of competent cells for SDM (XL-Blue).....59
2.5.3 DNA extraction by miniprep.....60
2.5.4 DNA extraction by midiprep.....60
2.6 Antibiotic kill curves.....61
2.7 Transient and stable transfections.....63
2.8 Lentiviral experiments..... 63
2.8.1 Lentiviral production..... 64
2.8.2 Lentiviral titration..... 64
2.8.3 Lentiviral infection.....64

2.9 Expression analysis of *HNF1A* and downstream targets in HepG2 cells..... 66
2.9.1 RNA extraction..... 66
2.9.2 cDNA synthesis..... 66
2.9.3 Real time TaqMan® quantitative PCR.....66
2.9.4 Protein analysis..... 68
2.10 Flow cytometry.....71
2.10.1 HeLa cells.....72
2.10.2 HepG2 cells.....72
2.11 Production of recombinant human glucokinase protein.....73
2.12 Glucokinase glucose- and ATP-dependent assays.....75
2.13 *HNF1A* and *GCK* cDNA sequences.....77

Chapter 3. Unsupervised clustering of exome-detected *HNF1A* alleles using multi-dimensional functional data

3.1 Background
3.1.1 HNF-1a *in vitro* function: stratification of alleles along the pathophysiological continuum.....82
3.1.2 Large-scale collaborative functional follow up of *HNF1A* alleles.....84
3.1.3 Chapter preview.....86
3.2 Methods
3.2.1 *HNF1A* sequence.....90
3.2.2 Bioinformatic prediction.....90
3.2.3 Stratification of *HNF1A* alleles with unsupervised learning methods.....91
3.2.4 Variant-phenotype mapping.....95
3.3 Results
3.3.1 Allele frequency and computational predictions of pathogenicity.....100
3.3.2 Data preparation.....108
3.3.3 Analysis optimisation.....113
3.3.4 *HNF1A* variant classification using unsupervised learning tools.....119
3.3.5 Re-evaluating MODY diagnoses based on unsupervised functional classification.....130
3.3.6 Transactivation as the most informative functional parameter.....142
3.4 Discussion
3.4.1 Summary.....145
3.4.2 Utility of functionally and clinically annotated allele clusters.....146

Chapter 4. Building a Multiplexed Assay of Variant Effects (MAVE) to screen all possible *HNF1A* missense alleles

4.1 Background
4.1.1. The challenge in establishing clinical significance of *HNF1A* sequence variants.....148

4.1.2. Towards molecular diagnostic certainty: large-scale prospective functional annotation.....151
4.1.3 Chapter preview.....153
4.2 Methods
4.2.1 *HNF1A* sequences.....159
4.2.2 Plasmids, molecular cloning, and mutagenesis.....160
4.2.3 Doxycycline dose induction.....161
4.2.4 RNA Sequencing.....164
4.2.5 Differential expression analysis.....165
4.2.6 Bioinformatic analysis.....165
4.2.7 Genome editing.....167
4.2.8 Site Saturation Variant Library (SSVL) synthesis, high-throughput cloning and verification.....168
4.3 Results
4.3.1 Artificial reporter.....172
4.3.2 Endogenous reporter.....181
4.3.3 Biological insights from transcriptomic profiling of *HNF1A* MAVE model.....193
4.4 Discussion
4.4.1 Summary and conclusions.....201
4.4.2 Ongoing efforts: application of an anti-TM4SF4 HepG2 MAVE for *HNF1A*.....205

Chapter 5. Comprehensive functional assessment of *GCK* missense alleles identified in clinical genetic testing

5.1 Background
5.1.1 Glucokinase mutational spectrum.....213
5.1.2 Glucokinase and type 2 diabetes (T2D).....215
5.1.3 Chapter preview.....216
5.1.3 Case reports.....219
5.2 Methods
5.2.1 Glucokinase mutagenesis and recombinant protein production.....225
5.2.2 Functional assays.....225
5.2.3 Bioinformatic analysis.....228
5.2.4 Graphical and statistical analysis.....229
5.2.5 Structural modelling.....230
5.3 Results
5.3.1 *GCK* p.F133L (c.397T>C).....231
5.3.2 *GCK* p.E157K (c.469G>A).....236
5.4 Discussion
5.4.1 GCK-p.F133L Interpretation and reporting.....242
5.4.2 GCK-p.E157K Interpretation and reporting.....245
5.4.3 Summary and conclusions.....248

Chapter 6.General Discussion

6.1 Summary of findings.....253
6.2 Context and limitations.....257

6.3 Looking forward.....261
6.4 Conclusion.....264

Appendix

References.....269

1

Introduction[1]

[1] This chapter is based partly on a review article published in the *Review of Diabetic Studies*, authored by Sara Althari and Prof Anna L. Gloyn (Althari and Gloyn, 2015).

In the age of genomic sequencing, we are faced with the challenge of making clinically meaningful extrapolations and extracting translational value from massive datasets that are being generated at an unprecedentedly rapid pace. Mendelian and other presumed monogenic disorders, such as Maturity-Onset Diabetes of the Young (MODY), are no exception to this challenge, despite their seemingly straightforward genetics and pathophysiology. In this chapter, I highlight some of these challenges in the context of MODY genetics, as I review the latest efforts in variant identification, interpretation, and functional annotation, which, in my opinion, warrant a thorough re-evaluation of the existing classification scheme of diabetes.

I begin by presenting the traditional 'pre-genome' view of MODY, which is that it is a dominantly inherited form of young-onset diabetes defined by highly penetrant rare, damaging, segregating alleles in a small subset of genes. The classic understanding of MODY as a disease with strict Mendelian boundaries originated from familial studies and efforts which were based upon phenotype-driven methods of ascertainment. By reviewing genetic and genomic insights obtained from large-scale sequencing efforts, and developments in mapping sequence-function relationships in the laboratory with varying levels of throughput, I highlight the various ways in which the conventional view of MODY as a discrete diabetes subtype needs to be amended in order to adequately assess the clinical impact of alleles in *HNF1A* and *GCK*.

1.1 Classic definition of Maturity-Onset Diabetes of the Young (MODY)

Maturity-onset diabetes of the young (MODY) (MIM606391) is a rare disorder that results in nonketotic hyperglycemia in young adults as a consequence of altered

pancreatic beta-cell function, primarily through reduced glucose-stimulated insulin secretion (Tattersall 1974; Tattersall and Fajans, 1975). It is characterized by a Mendelian autosomal dominant mode of inheritance, early age of onset (typically <25 years), and non-insulin dependency. It is estimated to comprise 1-2% of all diabetes cases (Frayling et al, 2001; Murphy et al, 2008). However, prevalence rates of MODY are inaccurate and underestimated globally as it is frequently undiagnosed or misclassified as type 1 diabetes (T1D) or type 2 diabetes (T2D) because of overlapping clinical features (Ledermann 1995; Shields et al, 2010; Thanabalasingham and Owen, 2011; Colclough et al, 2014).

In terms of molecular etiology, nearly 80% of all MODY cases are attributable to loss-of-function (LoF) heterozygous variants in one of three genes, namely the transcription factors hepatocyte nuclear factor-1 alpha (*HNF1A*) and 4-alpha (*HNF4A*), and the glycolytic enzyme glucokinase (*GCK*) (**Table 1.1**) (Fajans et al, 2001; Murphy et al, 2008; Shields et al, 2010; McDonald and Ellard, 2013; Pihoker et al, 2013; Colclough et al, 2014). Mutations in another transcription factor gene, *HNF1B,* account for a minority of phenotypically distinct MODY cases, characterized by profound renal complications which often progress to complete kidney failure demanding renal replacement (Bingham and Hattersley, 2004) (**Table 1.1**). Although more genes have been implicated in the pathogenesis of MODY, most are either exceptionally rare—reported once and detected only in a handful of families—or are more commonly responsible for neonatal diabetes, such as genes encoding Kir6.2 (ATP-sensitive inwardly rectifying potassium channel) and sulphonylurea 1 (SUR1) subunits of the pancreatic beta-cell ATP-sensitive potassium channel, *KCNJ11* and *ABCC8,*

respectively (Fajans et al, 2001; Ellard et al, 2008; Murphy et al, 2008; Shields et al, 2010; McDonald and Ellard, 2013) (**Table 1.1**).

The quest to find new MODY genes is ongoing, motivated by the ability of rare familial disease genetics to unravel deep biological and mechanistic insights and novel therapeutic targets. Traditionally, MODY genes were discovered by linkage mapping of disease-associated loci within large pedigrees with extended family history of diabetes, followed by sequencing (Froguel et al, 1992; Hattersley et al, 1992; Yamagata et al, 1996; Vaxillaire et al, 1995). Despite advancements in sequencing technologies and resources to facilitate robust identification of candidate variants/genes, the discovery of novel MODY genes has been rather limited in success.

One main reason is that co-segregation testing demands suitably large pedigrees with Mendelian inheritance of hyperglycemia. There is also the challenge of distinguishing MODY from other subtypes of diabetes, particularly using classic diagnostic criteria of young-onset and no obesity as the obesity epidemic grows among young adults (Colclough et al, 2013; Patel et al, 2017). Although, importantly, with increased technological and analytical developments in genetics and genomics, the once thought to be discrete genetic and clinical subdivisions of diabetes mellitus are being viewed more as a continuum with increasingly evident overlap between the pathophysiological and genetic properties of rare (‘monogenic’) and common (‘polygenic’) forms (Flannick et al, 2016).

A novel missense variant in Wolfram syndrome 1 (*WFS1*) and loss-of-function alleles (missense and nonsense) in Adaptor Protein, Phosphotyrosine Interaction, PH domain, and leucine zipper containing 1 (*APPL1*) were identified

as new rare causes of MODY (Prudente et al, 2015; Bonnycastle et al, 2013). A recent study identified novel heterozygous protein truncating variants (PTVs) in Regulatory Factor X6 (*RFX6*) as a new genetic cause of MODY with reduced penetrance compared to MODY-causal alleles in *HNF1A* and *HNF4A* (Patel et al, 2017). Patel and colleagues used a combination of next-generation sequencing in well-phenotyped MODY cases (from clinical discovery and replication cohorts) and large control cohorts from resources like the Exome Aggregation Consortium (ExAC) database, as well as co-segregation analysis, to reliably evaluate causality of the PTVs in *RFX6* (Patel et al, 2017; Lek et al, 2016). From a physiological standpoint, they identified gastric inhibitory polypeptide (GIP) deficiency in *RFX6* PTV heterozygotes, a feature foreign to other forms of human diabetes, and supports physiological data from murine studies on the regulation of GIP expression and secretion by RFX6 (Patel et al, 2017).

MODY is classically defined as a monogenic form of diabetes, although it is clinically, metabolically, and genetically heterogeneous as molecular dysfunction of presumed MODY-causal genes manifests in a variety of clinical phenotypes (Hattersley 1998; Murphy et al, 2008; McDonald and Ellard, 2013). Management and treatment strategies vary with the different genetic subtypes of MODY. Thus, an accurate genetic diagnosis is the key to adequate treatment, and entails proper interpretation and assessment of known and novel variants detected in both the research and clinical sphere.

1.2 Genetic culprits of MODY and estimations of prevalence

Since the first clinical case report that described young-onset non-insulin dependent diabetes in three families (Tattersall 1974), a substantial amount of

progress has been made in defining the genetic architecture of MODY. Early on, this was achieved by linkage analysis in large multigenerational pedigrees. In 1992, *GCK* was identified on chromosome 7p as the first MODY gene via linkage studies in British and French families (Froguel et al, 1992; Hattersley et al, 1992; Velho et al, 1992).

Shortly thereafter, microsatellite marker analysis and positional cloning identified *HNF1A* at chromosome 12q in MODY families (Vaxillaire et al, 1995; Yamagata et al, 1996b). Similar efforts also led to the discovery of *HNF4A* in several MODY pedigrees (Bell et al, 1991; Fajans et al, 1994; Yamagata et al, 1996a). The foregoing findings facilitated variant identification in those genes, which seemed to be responsible for familial MODY phenotypes. Since then, >620 *GCK* mutations, >414 *HNF1A* mutations, and >103 *HNF4A* mutations in over 1440, 1240, and 170 European Caucasian families, respectively, have been reported (Osbak et al, 2009; Colclough et al, 2013; Colclough et al, 2014). These mutations include missense, nonsense, frameshift, in-frame deletions/insertions/duplications, splice site, promoter region, and whole/partial gene deletions (Colclough et al, 2013; Ellard et al, 2008; Gloyn et al 2003; Ellard et al, 2007; Osbak et al, 2009).

Yet, the true population prevalence of MODY remains unknown and geographically variable because of arbitrary and inconsistent methods of clinical ascertainment and patient recruitment worldwide (Shields et al, 2010; Thanabalasingham and Owen, 2011; Chakera et al, 2014; Colclough et al, 2014).

Table 1.1. Maturity-Onset Diabetes of the Young (MODY) genes

MODY SUBTYPES (REF)	DISTINGUISHING FEATURES	OTHER FORMS OF MONOGENIC DIABETES	OTHER DISEASE PHENOTYPES
***GCK*-MODY** (Steele et al, 2014; Chakera et al, 2015; Njolstad et al 2001; Osbak et al 2009; Gloyn et al, 2003)	Heterozygous inactivating mutations result in asymptomatic mild fasting hyperglycaemia from birth (5.5–8.3 mmol/l). Small OGTT increment (< 3 mmol/l). No insulin or pharmacological treatment required. Slight alterations in glycaemic control with age. Risk of complications in line with the general population.	Homozygous or compound heterozygous inactivating mutations cause Permanent Neonatal Diabetes (PNDM). Carries of heterozygous inactivating mutations are at risk for developing gestational diabetes.	Heterozygous activating mutations cause hyperinsulinemic hypoglycaemia (Glaser et al, 1998).
TRANSCRIPTION FACTOR MODY			
***HNF1A*-MODY** (Yamagata et al, 1996; Colclough et al, 2013; Shepherd et al, 2001; Pearson et al 2003; Stride et al, 2002; Murphy et al, 2008)	Typically presents in early adulthood/adolescence, though age of onset dependant on mutation position within the gene. High sensitivity to treatment with low dose sulphonylurea tablets. Progressive beta-cell dysfunction and increasing penetrance with age: 63% by age 25; 93.6% by age 50; 98.7% by age 75. Large glucose increments in OGTT (> 5 mmol/l). Elevated high density lipoprotein (HDL) cholesterol levels.		Hepatocellular adenoma (biallelic inactivating variants) (Zucman-Rossi et al, 2006).
***HNF4A*-MODY** (Yamagata et al, 1996; Colclough et al, 2013; Pearson et al, 2003; Murphy et al 2008; Pearson et al, 2005; Pearson et al, 2007; Flanagan et al, 2010)	Macrosomia (birth weight increase by 800g compared to non-carriers). Reduced HDL-cholesterol, triglycerides, apolipoproteins A1 and A2, and increased low density lipoprotein (LDL) cholesterol concentrations.		~10% heterozygous mutation carries present with diazoxide-responsive congenital hyperinsulinaemic hypoglycaemia in the first week of life (Flanagan et al, 2010).
***HNF1B*-MODY** (Horikawa et al 1997; Bingham and Hattersley, 2004; Shields et al, 2010; Bellanne-Chantelot et al, 2005; Pearson et al, 2004)	Requires early insulin therapy. Nondiabetic cystic renal disease. Abnormalities in renal structure and development and genital tract abnormalities in female carriers. Low birth weight.	Mutations may cause Transient Neonatal Diabetes (TNDM).	Renal Cysts and Diabetes (RCAD) in heterozygous allele carries (Edghill et al, 2006).

***PDX1*-MODY** (Stoffers et al, 1997; De Franco et al, 2013; Vaxillaire and Froguel, 2008)	Very Rare. Heterozygous mutation reported in a single family.	Homozygous hypomorphic variant reported to cause PNDM without exocrine pancreatic insufficiency (De Franco et al, 2013).	
***NEUROD1*-MODY** (Rubio-Cabezas et al, 2010; Kristinsson et al, 2001; Vaxillaire and Froguel, 2008)	Rare. Heterozygous mutations reported in five families to date.	Two homozygous loss-of-function mutations in two patients with PNDM and neurological abnormalities have been reported.	
POTASSIUM CHANNEL (K_{ATP}) MODY			
***KCNJ11*-MODY** (Inagaki et al, 1995; Gloyn et al, 2004; Bonnefond et al, 2012; Sagen et al, 2004; Vaxillaire et al, 2004; Flanagan et al, 2006; Gloyn et al, 2006)	Glycaemic control achieved with high dose sulphonylurea therapy.	Heterozygous activating mutations are the commonest cause of PNDM or TNDM.	Recessive loss-of-function mutations manifest in congenital hyperinsulinism of infancy (CHI) (Thomas et al, 1996; Dunne et al, 2004; Glaser et al, 2000).
***ABCC8*-MODY** (Aguilar-Bryan et al, 1995; Babenko et al, 2006; Ellard et al, 2007; Gloyn et al, 2006; Bowman et al, 2012)	Glycaemic control achieved with high dose sulphonylurea therapy.	Heterozygous activating mutations are the commonest cause of PNDM or TNDM.	Loss-of-function mutations (dominant and recessive) manifest in CHI (Shemer et al, 2012; Chandran et al, 2013; Glaser et al, 2000).
RARER FORMS OF MODY			
***CEL*-MODY** (Raeder et al, 2006; Torsvik et al, 2010; Vesterhus et al, 2008; Raeder et al, 2013)	Rare. Heterozygous single base deletions in two families and a short 3-repeat allele in one.		Faecal elastase deficiency and exocrine pancreatic insufficiency (Raeder et al, 2006; Vesterhus et al, 2008).
***INS*-MODY** (Edghill et al, 2008; Molven et al, 2008; Molven and Njolstad 2011)	Rare. Heterozygous co-segregating missense variants identified in 6 European families.	Heterozygous missense and autosomal dominant mis-folding mutations cause PNDM (Polak et al, 2008; Colombo et al, 2008). Recessive homozygous or compound heterozygous also reported to cause neonatal diabetes (Garin et al, 2010)	

For example, in Italy, France, and Spain, where pediatric testing for fasting blood glucose is routine practice, mild fasting hyperglycemia as a result of GCK-MODY (5.4-8.3 mmol/l) is more frequently detected resulting in higher reported prevalence (Velho et al, 1997; Lehto et al, 1999; Costa et al, 2000; Fajans et al, 2001; Sagen et al, 2008; Lorini et al, 2009; Colclough et al, 2014). These reports are consistent with the observation that GCK-MODY is present from birth, unlike transcription factor-MODY which is characterized by progressive beta-cell dysfunction with onset between the second and fifth decade and increasing penetrance with age (Thanabalasingham and Owen, 2011).

In contrast to Southern Europe, HNF1A-MODY is more commonly diagnosed in the United Kingdom and Scandinavia, where blood glucose tests are mainly reserved for referred patients for whom a diagnosis of diabetes has already been made, but rarely conducted in asymptomatic individuals (Shields et al, 2010; Thanabalasingham and Owen, 2011). Efforts have been made to estimate the minimum population prevalence of MODY among adults and children in European countries, albeit inconclusively, as studies were neither systematic nor up to scale (Ledermann 1995; Pearson et al, 2005; Eide et al, 2008; Schober et al, 2009; Shields et al, 2010; Kropff et al, 2011). A community-based cross-sectional study by Kropff and colleagues estimated the prevalence of HNF1A-MODY in the UK to be 84 cases per million (95% CI 36-131). The study also estimated that nearly 90% of HNF1A-MODY cases in the UK, of which there are at least 5,000-6,000, remain undiagnosed (Shields et al, 2010; Kropff et al, 2011).

In 2010, Shields and colleagues estimated the minimum prevalence of MODY in the UK to be 108 cases per million by examining UK referrals for genetic testing (>2K probands and >1K relatives) between 1996 and 2009 (Shields et al, 2010).

If MODY accounts for ~1% of diabetes, 27,000 MODY cases are predicted in the UK, and ~6,000 based on the minimum prevalence estimated in this study (Shields et al, 2010). Yet, by 2010, only 1,177 cases in the UK had received a molecular diagnosis of MODY, meaning >80% of cases are likely to be either mis- or undiagnosed by genetic testing (Shields et al, 2010).

To determine MODY prevalence accurately, more large-scale population-based assessments and molecular-genetic screens of patients with diabetes are needed. Chakera and colleagues have conducted such a study using a subset of Euro-Caucasian female patients from the Atlantic Diabetes in Pregnancy (Atlantic DIP) cohort (n = 5,500) (O'Sullivan et al, 2011; Chakera et al, 2014). Based on their analysis, it was estimated that the prevalence of GCK-MODY would be 1.1 in 1,000 (95% CI 0.3-2.9) among white Europeans (Chakera et al, 2014).

Other groups have evaluated the frequency of MODY in youth populations by genetically screening individuals from large pediatric diabetes cohorts. Some have avoided classical MODY presentations and selection based on referral patterns to offset ascertainment bias. Based on such studies, the minimum prevalence rates of MODY among children in Norway, Poland, Germany, and the United States have been estimated to be at 3.1/100,000, 4.2-4.6/100,000, 2.39/100,000, and 2.1/100,000, respectively (Neu et al, 2009; Fendler et al, 2012; Irgens et al, 2013; Pihoker et al, 2013). The SEARCH for Diabetes in Youth study estimated the prevalence of *HNF1A*, *HNF4A* and *GCK* MODY in the pediatric diabetes in the population to be 1.2% using a large multicenter pediatric US diabetes cohort where selection criteria were limited to a negative autoantibody test and fasting C-peptide (FCP) level of >265 pmol/L (Pihoker et al, 2013).

Most recently, the Fremantle Diabetes Study Phase II determined the prevalence of monogenic diabetes (MODY and PNDM) amongst Australians (European and non-European ethnic backgrounds) using a cohort of patients under the age of 35 and considered to be at risk for MODY based on UK risk prediction models (Shields et al, 2012; MODY probability calculator http://www.diabetesgenes.org/content/mody-probability-calculator). Genotypes were resolved for those identified to be at risk for monogenic diabetes. They reported 45 and 89 cases per million as the overall population prevalence of PNDM and MODY, respectively (Davis et al, 2017).

A recent population-based assessment of monogenic diabetes prevalence was conducted by the UNITED (Using pharmacogeNetics to Improve Treatment in Early-onset Diabetes) study team, and estimated minimum prevalence to be 3.6% of patients diagnosed under the age of 30y (Shields et al, 2017). Prevalence estimates were determined as a proportion of positive cases out of total recruited for the study (Shields et al, 2017). The study tested the robustness of a biomarker screening pathway (C-peptide and islet autoantibodies) to discriminate monogenic diabetes from T1D for which MODY is frequently misdiagnosed (reported positive predictive value of 20%) (Shields et al, 2017).

1.3 Efforts to find the missed and mis-diagnosed MODY cases

Apart from enhancing epidemiological data, there is great therapeutic and prognostic value in providing an accurate molecular genetic diagnosis of MODY. Genetically discriminating MODY from other types of diabetes, as well as MODY subtypes from each other, has profound implications on treatment and management strategies for allele carriers and can inform genetic counselling and

screening of family members (Murphy et al, 2008; Bellanne-Chantelot et al, 2011; Thanabalasingham et al, 2012).

1.3.1 Clinical significance of an accurate molecular diagnosis of MODY

Patients with undiagnosed MODY often follow glycemic control plans specific to either T1D or T2D, including exogenous insulin, which could be avoided with a proper clinical and genetic diagnosis (Shepherd et al, 2009; Thanabalasingham and Owen, 2011). There are profound negative impacts on quality of life associated with insulin administration, particularly from a psychosocial standpoint (Pouwer and Hermanns, 2009). Individuals prescribed insulin treatment express a number of concerns and fears, which include stress and anxiety related to self-management and self-monitoring (risk of hypoglycemia), weight gain, stigma, and hindrance to normal social functioning (Pouwer and Hermanns, 2009).

Individuals diagnosed with HNF1A-MODY are highly responsive to treatment with a low dose of the oral hypoglycemic agent (OHA) sulphonylurea, and can avoid insulin injections (either entirely or until late in life) (Pearson et al, 2000). In fact, early studies testing the efficacy of sulphonylurea treatment reported a five times greater glycemic control response than metformin and a four times greater response than insulin therapy as measured by lowering fasting glucose in HNF1A-MODY subjects using BMI- and glycaemia-matched T2D controls (Pearson et al, 2000; Pearson et al, 2003). The case, in terms of responsiveness to low-dose sulphonylurea tablets, is similar for patients with MODY as a result of mutations in *HNF4A* (Pearson et al. 2005). Transcription factor MODY patients are typically born normoglycemic, and develop progressive beta-cell dysfunction in adolescence or early adulthood (< 25 years) (Shepherd et al, 2009; Bellanne-

Chantelot et al, 2011). HNF1A-MODY is associated with micro- and macrovascular complications similar to those developed by T1D and T2D patients (Isomaa et al, 1998).

GCK-MODY is another example of how a genetic diagnosis can spare affected individuals ineffective and burdensome treatment. Unlike transcription factor MODY, GCK-MODY patients have lifelong, asymptomatic mild fasting hyperglycemia, present from birth, with only minor deterioration with age in line with the normal population (Prisco et al, 2000; Fajans et a, 2001; Murphy et al, 2008) (**Table 1.1**). Glucose is regulated in affected individuals, and insulin response is maintained, at elevated set-points (Byrne et al, 1994; Guenat et al, 2000). Micro- and macrovascular complications are rare in GCK-MODY patients, and according to recent cross-sectional studies, are less frequent than in T1D and T2D individuals, and comparable to healthy controls (Steele et al, 2014). As a result of this homeostatic control and lack of severe complications associated with the GCK-MODY subtype, patients usually require no treatment (Pearson et al. 2001).

A recent cross-sectional study evaluated the impact of pharmacological treatment, either insulin or oral hypoglycemic agents, on glycemic control in 799 UK GCK-MODY patients, and found no difference in HbA1c between patients on or off treatment (48 vs. 46 mmol/mol; 6.5% vs. 6.4%; $p = 0.11$) (Stride et al, 2014). Stride and colleagues performed a longitudinal study to assess the pharmacotherapeutic impact of stopping treatment in GCK-MODY patients - who had been receiving OHAs or insulin - on glycaemia. They also found no significant alterations in Hb1Ac levels between the different treatments (mean change: -0.68 mmol/mol (95% CI: -2.97, 1.61)) (Stride et al, 2014).

The identification of alleles in *GCK* or *HNF4A* also has important implications for pregnancy management. Inherited mutations in *HNF4A* are associated with an 800 g increase in birth weight of the baby. Therefore, babies of affected parents are 50% more likely to develop macrosomia (Pearson et al, 2007). Additionally, about 10% of *HNF4A* mutation carriers experience diazoxide-responsive hyperinsulinemic hypoglycemia in the first week of life (Flanagan et al. 2010) (**Table 1.1**). With GCK-MODY, babies that do not inherit the heterozygous *GCK* allele from an affected mother are at risk of excessive birth weight because of increased fetal insulin secretion and insulin-mediated growth in response to maternal hyperglycemia (Hattersley et al, 1998; Spyer et al, 2009). Conversely, babies that have either developed a *de novo GCK* mutation or inherited only one allele paternally have a birth weight reduced by nearly 500 g as a direct effect of reduced fetal insulin secretion (Hattersley et al, 1998; Spyer et al, 2009). Babies that do inherit heterozygous inactivating *GCK* mutations from GCK-MODY-affected mothers are born with a normal weight as they have the same elevated set-point for glucose regulation (Spyer et al, 2009).

1.3.2 Current MODY detection methods

Although the cost of high-throughput genomic sequencing is rapidly falling, whole exome/whole genome applications are still too expensive to sequence patients indiscriminately in the clinical diagnostic setting. The gold standard for molecular genetic testing of MODY is direct Sanger sequencing of coding, conserved splice, and minimal promoter regions of known MODY genes (99% sensitivity for heterozygous base alleles) (Colclough et al, 2014). Multiplex Ligation-dependent Probe Amplification (MLPA) assays are used to for gene dosage analysis to identify whole/partial gene deletions which constitute 1-3% of variation in MODY

genes (Colclough et al, 2014). A targeted next-generation sequencing exon-capture assay for a panel of MODY genes has been successfully designed, developed and implemented (confirmation by Sanger sequencing) (Ellard et al, 2013).

Individuals are referred for molecular diagnostic testing of MODY on the basis of clinical criteria, which include early age of diabetes onset, positive family history, endogenous insulin secretion, non-obesity, and absence of pancreatic autoantibodies (Ellard et al, 2008; Murphy et al, 2008; Bellanne-Chantelot et al, 2011; Colclough et al, 2014). Alone, these are neither reliable nor sensitive enough, as only 50% of European cases diagnosed with MODY fit these traditional criteria (~35% in the UK) (Shields et al, 2010).

This has led to the widely-quoted statement that the conventional clinical criteria miss 50% of MODY. However, it is worth pointing out that this statement depends on the assumption that individuals with diabetes who fail to meet the explicit clinical criteria for MODY, but who harbour a variant in a MODY gene thought to be functional, must have MODY rather than T2D. As I will describe later, this seems to invite a rather expansive, flexible view of MODY which extends well beyond the original monogenic confines, and which ignores the range of clinical phenotypes that can be associated with a single MODY allele.

Limitations associated with these selection criteria were highlighted in a recent study in which MODY genes, *HNF1A*, *HNF4A*, and *GCK*, were sequenced in a large group of European patients clinically diagnosed as having either T1D or early-onset T2D (Thanabalasingham et al, 2012). This sequencing effort showed that 25% of the subjects with apparent T1D or T2D before the age of 30 were in

fact MODY cases with pathogenic variants in one of the three MODY genes. Thus, Thanabalasingham and colleagues concluded that broadening the current selection criteria to include all patients diagnosed with C-peptide-positive diabetes for molecular diagnostic sequencing before the age of 30 increases sensitivity, potentially sparing more misdiagnosed individuals unnecessary treatment (Thanabalasingham et al, 2012).

1.3.3 Biomarkers for MODY detection and classification

Efforts centered on biomarker utility have emerged within the last 5 years, with the aim of refining the imperfect MODY selection criteria that qualify patients for genetic screening. Glucose increments in oral glucose tolerance tests can be used to discriminate between HNF1A and GCK-MODY (Colclough et al, 2014). Patients with HNF1A-MODY show larger glucose increments (>5 mmol/l) as a consequence of a more compromised insulin secretory capacity compared to GCK-MODY where glucose increments are modest (< 3 mmol/l) (McDonald and Ellard, 2013). Pal and colleagues evaluated the discriminatory capacity of serum 1,5 anhydroglucitol (1,5 AG) levels in UK subjects with HNF1-MODY, GCK-MODY, T1D, T2D, and latent autoimmune diabetes in adults (Pal et al, 2010). They found that 1,5 AG levels were significantly higher in GCK-MODY than in any other group ($p \leq 0.0003$), with greatest discriminatory accuracy for GCK-MODY from HNF1A-MODY subjects (C-statistic: 0.86; 86% sensitivity and 84% specificity for 1,5 AG >7.5 µg/ml) (Pal et al, 2010). Elevated plasma-HDL-cholesterol levels in HNF1A-MODY patients have also been shown to differentiate from T2D (>1.12 mmol/L, ROC AUC = 0.76) where HDL levels are typically reduced (McDonald et al, 2012).

Genome-wide association study (GWAS) findings have inspired research into the discriminatory capacity of another set of candidate biomarkers, namely highly sensitive C-reactive protein (hsCRP) in serum and plasma glycan profiles, to discriminate HNF1A-MODY from other types of diabetes (Lauc et al, 2010; Owen et al, 2010; McDonald et al, 2011; Thanabalasingham et al, 2011; Thanabalasingham et al, 2013). Levels of hsCRP were initially tested in a small UK cohort, and then replicated in two large European studies, which included about 700 HNF1A-MODY cases (Owen et al, 2010; McDonald et al, 2011; Thanabalasingham et al, 2011). These studies showed that hsCRP levels are significantly reduced in HNF1A-MODY cases - which is consistent with the fact that the *CRP* gene is under the regulation of *HNF1A* - and that HNF1A-MODY cases could be discriminated from other types of diabetes (Owen et al, 2010; McDonald et al, 2011; Thanabalasingham et al, 2011). Levels of hsCRP discriminated HNF1A-MODY from T2D with high sensitivity and specificity: studies have shown that 70-80% of HNF1A-MODY patients have <0.5 mg/l hsCRP, compared to <20% of individuals with type 2 diabetes (McDonald et al, 2011; Thanabalasingham et al, 2011; Owen 2013).

GWA studies evaluating the human plasma N-glycome revealed novel associations with common variants in the *HNF1A* locus (Lauc et al, 2010). Functional analysis inspired by the GWAS findings exposed *HNF1A* as a master regulator of plasma protein fucosylation (Lauc et al 2010). Specifically, *HNF1A* contributes to posttranslational modification of glycoproteins by upregulating antennary fucosylation of glycoproteins modified with *N*-linked glycans (Lauc et al. 2010). This generated the hypothesis that loss-of-function mutations in *HNF1A* might be accompanied by alterations in plasma glycome profiles of valuable

diagnostic utility (Thanabalasingham et al, 2013). Indeed, a pilot study and a large replication, comparing the DG9-glycan index (the ratio of fucosylated to nonfucosylated triantennary glycans) of HNF1A-MODY patients with other diabetes subtypes, reported good discrimination from T1D and T2D (C-statistic ≥0.90) (Thanabalasingham et al, 2013).

Until we are financially, technically, and analytically equipped to interpret both coding and non-coding sequence variants identified in patients unselected for extreme phenotypes, diagnostic clinical biomarkers appear to be the most attractive and cost-effective adjuncts for patient prioritization as indirect readouts of variant function. When high-throughput, next-generation sequencing technologies are affordable and widely applied in clinics, these biomarkers will likely serve more as complementary validation, and aid in sequence interpretation and assessments of variant-level pathogenicity.

1.4 The challenge in interpretation of sequence variants in MODY genes

Focusing on the 'functional genome', the subset of the human genome enriched for rare, highly penetrant, protein-altering variants, has helped to unravel the pathogenesis of numerous monogenic disorders, and to understand the influence of rare alleles on complex disease risk (Stenson et al, 2009; Ponting and Hardison, 2011; Rands et al, 2014). Sufficiently powered positional cloning studies have been highly successful in identifying protein-coding variations in genes underlying inherited monogenic diseases. Nowadays, in the research setting, targeted exome sequencing has become the primary approach for readily and cost-effectively detecting rare coding variants with large effect sizes in the human genome (Bamshad et al, 2011; MacArthur et al, 2014). However, one of

the major challenges associated with exome-based efforts has been to distinguish disease-causing alleles from those that impact protein function, and from the vast majority of neutral variants with no discernable effects on molecular activity (Xue et al, 2012; MacArthur et al, 2014). Another layer of complexity is added when considering that most detected alleles, particularly in the clinical genomic context, are likely to be novel (i.e. neither previously reported nor characterized).

1.4.1 Variants in monogenic diabetes genes: a continuum of effects

Data from whole exome-sequencing studies are increasingly blurring the genetic boundaries previously thought to be discrete between rare, Mendelian, monogenic disorders and common, complex, polygenic ones, as well as between genomes of healthy and diseased individuals (**Figure 1.1**) (MacArthur et al, 2012; Xue et al, 2012; Flannick et al, 2013). Empirical data from large-scale, high throughput resequencing studies have unmasked the presence of nearly 100 classic loss-of-function (LoF) alleles (splice site single nucleotide variants (SNVs), insertions/deletions (indels), nonsense truncating mutations) per typical healthy human genome, mostly in the heterozygous state (MacArthur et al, 2012; MacArthur et al, 2014). Additional resequencing efforts have estimated up to roughly 400 damaging alleles per individual (>2 bona fide disease alleles per individual); a number that is expected to rise as methods of rare variant ascertainment continue to improve (Xue et al, 2012).

(i) Rare protein-coding MODY variants in normoglycemic individuals.

To date, the majority of genetic sequencing studies have drastically overestimated the penetrance of rare protein-coding alleles—especially in the

context of monogenic disorders—by employing inherently biased methods of sample ascertainment (Goring et al, 2001; Flannick et al, 2013). Most variant effect size estimates have been based on analysis of the exomes of either exceptionally high-risk individuals or those who occupy the extreme ends of a phenotypic distribution (**Figure 1.1A**) (Goring et al, 2001; Bamshad et al, 2011; Flannick et al, 2013).

Although such approaches have been instrumental in identifying disease variants, they can be misleading: by default, they dismiss scenarios of reduced allelic penetrance, i.e., they fail to capture events along the phenotypic spectrum (**Figure 1.1B**) (MacArthur et al, 2014). Consequently, erroneous or exaggerated estimates of variant pathogenicity are made, resulting in over-diagnosis or misdiagnosis, and the prescription of treatment and intervention regimens to an increasingly high number of low-risk individuals (Flannick et al, 2013). By studying well-phenotyped, population-based or case-control cohorts, in which cases are not preselected on the basis of satisfying classical MODY diagnostic criteria, for example, we could mitigate the upward bias, and remedy inflated effect size estimations, as has been demonstrated in the cases of *BRCA1* and *BRCA2* in breast cancer studies (Begg 2002; Flannick et al, 2013).

A recent study showed that following the aforementioned design guidelines is helpful in characterizing the spectrum of low-frequency alleles in the genes under investigation in the general population (Flannick et al, 2013). By targeted sequencing of the three MODY-causing genes (*GCK, HNF1A, HNF4A*), in addition to four of the rarer genes implicated in MODY (*PDX1, INS, NEUROD1, and HNF1B*), in well-phenotyped, randomly-ascertained individuals from population-based cohorts, Flannick and colleagues found that some MODY-

causing alleles exhibited incomplete penetrance in carriers who remained euglycemic through middle age. The variants were either (i) annotated as MODY-causal, and classified as pathogenic in the Human Gene Mutation Database (HGMD) (Stenson et al. 2009), or (ii) met strict criteria for ascribing *de novo* pathogenicity: rare, conserved, and predicted to be protein-damaging based on *in silico* tools (Flannick et al. 2013).

Studies like this, which have reliably exposed putative disease-causing or harmful alleles in the genome of normal healthy humans, impose many challenges on variant interpretation and pathogenicity assignments. Most notably, such findings challenge the notion that Mendelian disease variants are simpler to interpret—particularly in the medical sequencing setting—because of their supposedly deterministic nature. As has been shown, rare, protein-altering, and presumably pathogenic variants in *HNF1A,* for instance, do not necessarily give rise to a MODY phenotype in all individuals; they can also either contribute to T2D predisposition or exist innocuously in the genomes of healthy individuals (**Figure 1.1B and Figure 1.2**).

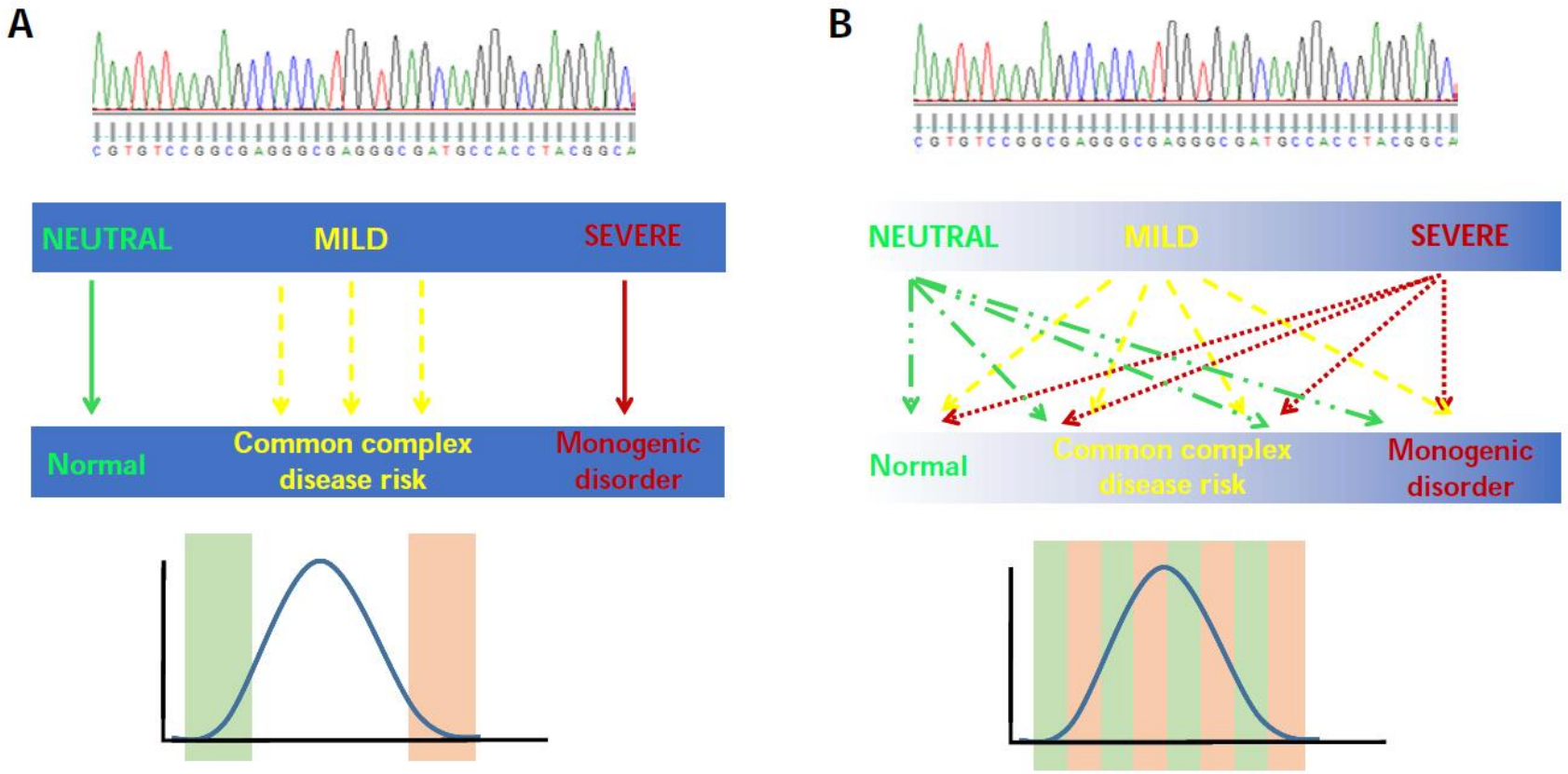


Figure 1.1. Relationship between mutational severity in known Mendelian disease-causing genes and clinical phenotype. (**A**) Traditional, deterministic understanding of Mendelian disease genetics based on biased modes of sample ascertainment. (**B**) Spectrum-based understanding of Mendelian disease genetics, as revealed by large unbiased population-wide resequencing efforts. This approach encompasses scenarios of incompletely penetrant, monogenic, disease-causing alleles in healthy individuals and, for example, monogenic disorders falsely attributed to a neutral variant in a known risk gene, while the disorder is actually caused by an allele in a different gene.

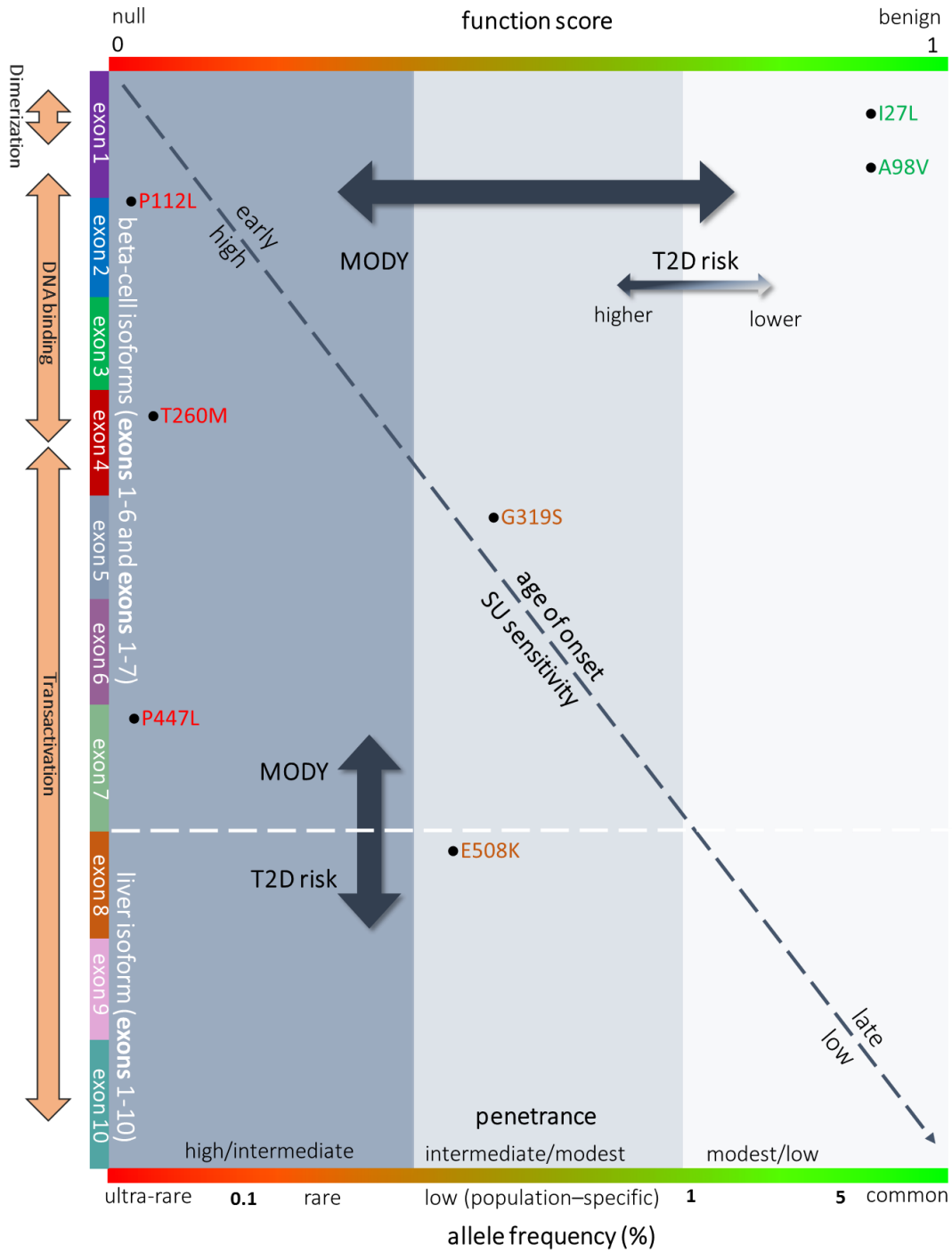


Figure 1.2. The complex, probabilistic relationship between *HNF1A* protein-coding variation and diabetes mellitus: a hypothesis. The *HNF1A* pathophysiological continuum is defined a) horizontally, by rare, familial, functionally disruptive alleles with large effects and minimal environmental contribution on one end (early-onset, SU-repsonsive diabetes), and, on the other end, common, functionally moderate/benign alleles of modest effect which modify risk for developing multifactorial polygenic diabetes (T2D predisposition); b) vertically, by isoform and functional domain expression: disruptive variation (low function score) in the first seven coding exons of the beta-cell isoforms associate with strong, monogenic, familial diabetes, whereas equally damaging alleles in the longer, liver-dominant *HNF1A* isoforms represent a major T2D molecular risk factor. Abbreviations: SU = sulphonylurea; T2D = type 2 diabetes; MODY = maturity onset diabetes of the young.

(ii) Low-frequency protein-coding MODY variants associated with increased T2D risk

A study by the SIGMA consortium also nicely demonstrates the complexities outlined above. This study was among the first systematic, high-resolution screens for sequence variation in protein-coding regions in a large population unselected for an extreme monogenic clinical phenotype. The consortium reported the detection of a low-frequency missense variant in *HNF1A* (p.E508K, c.1522G>A, exon 8) after sequencing whole exomes of Mexican and Latino T2D cases (n = 1794) and diabetes-free controls (n = 1962) (Consortium STD, 2014). It was observed in 2.1% of cases and 0.36% non-diabetic controls (OR 5.48; 95% CI 2.83-10.61; $p = 4.4 \times 10^{-7}$), and found to be associated with a 5-fold increase in T2D risk for carriers (Consortium STD, 2014). The association of the p.E508K variant in *HNF1A* with a T2D phenotype was replicated in the T2D-GENES (Type 2 Diabetes Genetic Exploration by Next-Generation Sequencing in Multi-Ethnic Samples) Hispanic group and in a *de novo* genotyping data set from self-identified indigenous Mexicans (DMS2) (OR 4.16; 95% CI 2.93-8.38). There was no evidence of p.E508K in exomes from large multi-ethnic data sets, other than in Latino samples from these consortia cohorts, confirming its exclusivity to Mexican/Latino American ethnic groups (Consortium STD, 2014).

The p.E508K variant exhibited incomplete penetrance, suggesting it is a moderate diabetes allele (Consortium STD, 2014). This may be linked in part to its intermediate effect on protein function, but could also be related to its location in exon 8 of *HNF1A*, which is only present in the longest of the three *HNF1A* genetic isoforms, termed *HNF1A*(A). Harries and colleagues have shown that the *HNF1A* isoforms exhibit differential expression, with *HNF1A*(A) (exons 1-10)

being the major isoform in liver and kidney (and fetal pancreatic beta-cells), but the minor one in the adult pancreatic islet, where expression of the two shorter isoforms, *HNF1A*(B) (exons 1-7) and *HNF1A*(C) (exons 1-6), dominates (**Figure 1.2**) (Harries et al, 2006). The same study revealed variation in age of onset of HNF1A-MODY as a consequence of mutation position. Individuals with mutations in exons 8-10, present exclusively in *HNF1A*(A), developed MODY about 7 years later than those with mutations harbored by the first 6 exons present in all three isomers (Harries et al, 2006). The penetrance of the p.E508K Latino variant could be influenced by this isomeric expression trend, which poses an additional challenge to the interpretation of the functional relevance and clinical significance of *HNF1A* alleles. For example, a protein-altering variation in *HNF1A*(A) may not necessarily translate to an adult beta-cell effect, yet it may nonetheless show clear effects on liver biomarkers, such as hsCRP, which have a good discriminatory capacity for HNF1A-MODY. In such instances, it is important to consider mutation position and isoform-specific expression patterns alongside biochemical phenotypes for making well-informed evaluations of variant pathogenicity (**Figure 1.2**).

An early candidate gene study of *HNF1A* identified another population-specific risk allele in *HNF1A*, p.G319S, private to the Canadian Oji-Cree population (Hegele et al, 1999). The variant was found to be associated with early-onset T2D (OR 4.0 and 1.97 in homozygous and heterozygous carriers) (**Figure 1.2**) (Hegele et al, 1999). A positive correlation was observed between HNF1A-p.G319S genotype and diabetes age-of-onset (Hegele et al, 1999).

(iii) Common variants of modest effects in MODY gene loci influence T2D risk. Early large-scale case-control efforts exploring the link between common variants

(MAF > 5%) in MODY genes and T2D predisposition reported very modest associations (Weedon et al, 2005; Winckler et al, 2007). In a meta-analysis of GWAS data from n=8,130 T2D cases and n= 38,987 controls, a common variant near *HNF1A* (rs7957197, OR 1.14, P= 4.6 × 10^{-7}) was identified as a new genome-wide T2D susceptibility locus (Voight et al, 2010). The same effort revealed T2D association signals near *HNF1B* and *WFS1* (Voight et al, 2010).

A study in which genetic fine-mapping of 39 GWAS-determined T2D risk loci was performed in nearly 27,000 cases and 57,000 controls of European ancestry reported the nonsynonymous common coding polymorphisms p.Ile27Leu (rs1169288, OR, 1.09; P = 8.1x10^{-15}: MAF = 33%) and p.Ala98Val (rs1800574, OR, 1.22; P = 5.1x10^{-10}; MAF = 2.7 %), and an intergenic SNP (chr12:121440833:D, OR, 1.07, P= 2.9x10^{-10}; MAF= 4.2%) as three index variants driving the T2D association signal mapping to *HNF1A* (**Figure 1.2**) (Gaulton et al, 2015). The variants were found to be in weak linkage disequilibrium ([I27L, P = 4.4 × 10^{-14}; r^2 = 0.09], [A98V, P = 4.2x10^{-10}, r^2 = 0.01], [chr12:121440833:D, P= 2.9x10^{-10}, r^2 = 0.19) with previously reported lead *HNF1A* GWAS SNP rs12427353 (intronic *HNF1A* variant, OR 1.07, P= 6.5x10^{-8}) (Morris et al, 2012; Gaulton et al, 2015). An association between HNF1A-p.I27L and increased T2D risk had been previously reported in elderly (age > 60) and overweight (BMI>25 kg/m^2) individuals (OR 2.3, P = 0.002) (Holmkvist et al, 2006). Gaulton and colleagues also reported coding allele in *HNF4A* p.Thr139Ile (rs1800961, OR= 1.16, MAF= 3.5%, P= 2.3x10^{-9}) as the index variant for a novel T2D association signal mapping to the *HNF4A* locus (P = 1.4 × 10^{-9}) (Gaulton et al, 2015).

A recent effort in which fine-mapping of coding associations was conducted using meta-analysis of GWAS data from 50,160 T2D cases and 465,272 controls of European ancestry, Mahajan and colleagues showed that coding variants accounted for >80% of the posterior probability at 16 out of 38 distinct T2D coding variant association signals (Mahajan et al, 2017). The coding allele in *HNF1A* p.Ala146Val (rs1800574, MAF=2.6%) was identified as one of the lead SNPs in the fine-mapping analysis. The common *HNF1A* coding allele p.Gly226Ala (rs56348580, MAF= 31.1%) has also been implicated as causal for T2D association at the *HNF1A* p.Ile57Leu signal (Mahajan et al, 2017; Mahajan et al, 2018).

Taken together, these findings showcase the genetic complementarity between rare and common disorders. Genome-wide significant common coding alleles in *HNF1A* associated with T2D susceptibility might modulate the impact of rare functional protein coding variation and partly explain variability in phenotype and penetrance between carriers. Indeed, well-studied Mendelian genetics (MODY) can inform and accelerate our understanding of common complex disorders (T2D), for which genome-wide association efforts have provided extensive catalogues of common variants with low to modest effect sizes. Importantly, these studies address some of the past limitations associated with evaluating the implication of variants in disease. For instance, we could not have concluded that the role of the p.E508K *HNF1A* allele in the pathogenesis of MODY is minor (i.e. not sufficient to cause MODY in isolation) without the a) detection of this allele in a state of incomplete penetrance in individuals with elevated T2D risk and b) the biological follow-up provided by the study group showing intermediate function compared to WT and known MODY alleles (Consortium STD, 2014).

These complex genomic insights warrant a more nuanced understanding of the phenotypic manifestation of *HNF1A* gene variants: some alleles are sufficient for early-onset SU-responsive diabetes ('HNF1A-MODY') in some people, some modify susceptibility for developing complex multifactorial hyperglycemia later in life (T2D), and most alleles will likely manifest as benign and neutral **(Figures 1.1B and 1.2)**. **Figure 1.3** illustrates the interpretive challenge in light of this updated view.

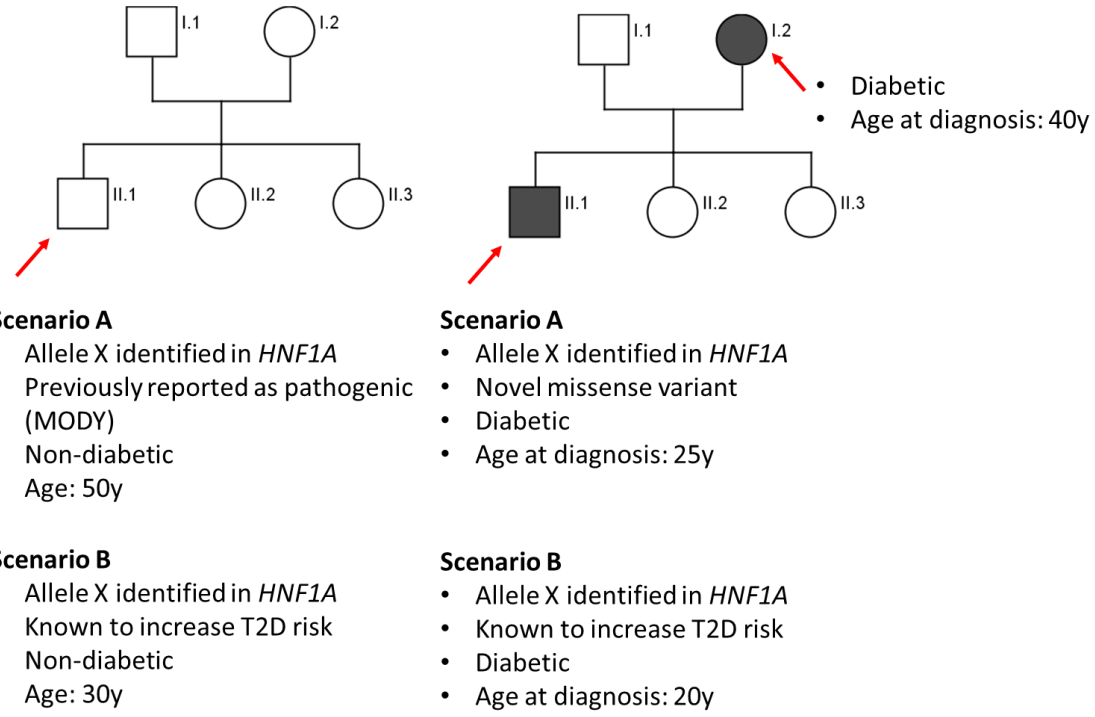


Figure 1.3. Hypothetical pedigrees which illustrate the challenge in the interpretation of known and novel *HNF1A* alleles as more individuals without a classic MODY phenotype are sequenced. As next generation sequencing becomes ubiquitous in the medical genomic setting, an increasing number of variants of uncertain clinical significance in *HNF1A* are bound to be detected in individuals along the phenotypic spectrum. Does the allele cause familial early-onset hyperglycemia? Does it considerably impact T2D risk? What if it is known to impact T2D risk but is detected in a young individual (<25y) with diabetes? If the individual carries an *HNF1A* allele and is diagnosed with diabetes, would they individual benefit from lifestyle modifications, insulin, or treatment with sulphonylurea? Should family members be screened and counselled? What if the allele was erroneously classified as pathogenic (i.e. MODY-causal)? What about the impact of genomic context on allele manifestation (modified expression by regulatory allele or co-expression with other alleles in *cis* or *trans*)? These questions highlight some of the challenges which the era of genomics has exposed, and which had not been previously considered in the context of Mendelian disease gene variants.

1.4.2 Assessing variant causality: contextualize with caution

In the light of these and other emerging challenges, MacArthur and colleagues have recommended a set of guidelines to assist investigators in cautiously and reliably defining the role of sequence variants in human diseases (MacArthur et al, 2014). In essence, these guidelines call for applying more systematic and unbiased approaches to identification, molecular characterization, and definition of the clinical relevance of sequence variants implicated in disease, in particular those of large effects, influencing both severe monogenic disorders and complex disease risk (**Figure 1.4**) (MacArthur et al, 2014). Some recommendations for meticulously designed sequencing study (population-based or case-control) include random recruitment of a large number of unrelated participants from multiple ethnic backgrounds (mitigate the impact of population stratification), with matched controls for whom clinical phenotypic data are available (MacArthur et al, 2014).

Furthermore, a rigorous assessment of conservation, novelty, and frequency should be performed on the variant. Family-based segregation data and detection of the same *de novo* mutation in multiple individuals from the same cohort are necessary criteria for determining causality, but they should not be considered as definitive (MacArthur et al, 2014). The latter are rare in the known MODY-causal genes, *GCK*, *HNF1A*, *and HNF4A*, but expected to increase considerably as more and more individuals unselected for a positive family history and classic MODY phenotype are sequenced in the research and clinical diagnostic settings **(Figures 1.1B and 1.3)** (Stanik et al, 2014). *A priori* functional and clinical information can be inconclusive and often misguiding and unreliable due to prevalence of false positive findings. Therefore, with each investigation,

researchers should conduct a host of complementary functional and statistical analyses, explicitly report results by highlighting limitations and ambiguities, and update relevant mutational catalogues and variant databases (**Figure 1.4**) (MacArthur et al, 2014). Transparency in data sharing, critical evaluation of variant-level implication in disease, confirmatory experimental assays, and an evaluation of pharmacotherapeutic targetability of variant-harboring proteins, will help to accelerate annotation efforts and to realize the translational potential of population-sequencing endeavors (Plenge et al, 2013; MacArthur et al, 2014).

1.5 Establishing the functional consequences of variants in known MODY genes

The ability to identify disease-relevant sequence variants exceeds the ability to interpret their true biological effects by far. This 'interpretative gap' is a major roadblock in the path bridging basic research to the clinical arena (Cutting 2014). Genetic findings are rife with false positives, variants of unknown clinical significance, and poor assessments of disease liability (Cooper 2015; Manrai et al, 2016; MacArthur et al, 2014). Proper annotation and functional follow-up of identified sequence variants, in ways which enable experimental validation of disruption and phenotypic recapitulation, is paramount to closing the gap (**Figure 1.4**) (Cutting 2014; MacArthur et al, 2014). Defining the actionable potential of sequence alleles and making accurate assignments of pathogenicity require employment of comprehensive and scalable molecular characterization pipelines. These must be meticulously designed and customized to the disease genes and disease phenotype under interrogation. For example, it would be futile to assess the function of naturally occurring HNF1A-MODY variants in a mouse model, because heterozygous *HNF1A* mutations have no effect on glycemic

status in mice (in contrast to humans) as their insulin-secretory function and glucose tolerance are normal (Pontoglio et al, 1998). Even within the mouse, there are inconsistencies with phenotypic consequences in humans when comparing global and tissue-specific genetic manipulations (Boj et al, 2012). As such, decisions regarding the most suitable model systems—cell lines or animals—depend primarily on which most effectively recapitulate human disease pathology.

1.5.1 *in vitro* assays to study the functional impact of MODY gene alleles

For MODY genes, *HNF1A* and *GCK*, there are sets of well-established, multi-tiered, reproducible assays to characterise the impact of sequence-detected alleles on protein function *in vitro*. For transcription factor MODY (such as *HNF1A*), experiments designed to interrogate the capacity for DNA binding and transactivation are the most widely applied. In the case of the gene encoding glucokinase enzyme, taking identified variants through a series of kinetic analyses is the most relevant approach for making sound extrapolations regarding function and protein stability (Osbak et al, 2009; Raimondo et al, 2014).

To characterize the molecular function of *HNF1A* variants detected in MODY probands, Galán and colleagues employed a variety of *in vitro* techniques (Galan et al, 2011). They identified n=6 alleles (5 missense and 1 allele in the 5' untranslated region [5'UTR]) from an *HNF1A*-specific mutation screen in patients who were selected for familial early-onset hyperglycemia; the alleles exhibited co-segregation with the disease phenotype, and were absent in healthy controls (Galan et al, 2011). Site-directed mutagenesis was used to synthesise the detected alleles onto *HNF1A* cDNA plasmid constructs, which, along with wild-

type *HNF1A*, were used to transfect cell lines, Min6 (mouse insulinoma-derived pancreatic beta-cells) and Cos7 (monkey kidney-derived). Since *HNF1A* regulates the expression of transcripts in pancreatic islets and hepatocytes, they co-transfected different tissue-specific target promoter sequences in each cell line to test the ability of the variants to drive reporter gene expression (Galan et al. 2011). The promoter sequences used were either expressed exclusively in pancreatic beta cells (HNF-4a P2 promoter and insulin), liver (beta-fibrinogen), or in both tissue types (GLUT-2). The direct impact of the clinical variants on HNF-1a protein expression, stability, and function was thoroughly evaluated using immunoblot (protein expression), luciferase reporter (transactivation) analyses, and electrophoretic mobility shift assays (EMSA) (DNA binding) (Galan et al, 2011). The 5'UTR alleles exhibited impaired molecular function in Min6 cells (~50% luciferase activity compared to WT); two missense alleles demonstrated consistent functional trends across tissues and promoters (p.T196A WT-like; p.R271W dysfunctional), whereas the molecular behavior of the other three alleles (p.E235G, p.V133M, p.P378R) was heterogeneous and variable depending on promoter architecture (degree of affinity) and tissue type (noise and biological relevance) (Galan et al, 2011).

Performing complementary assays and testing multiple variables in parallel is important, particularly when the aim is to acquire an objective and systematic understanding of the functional consequences of sequence variants. For example, although Min6 cells, which express endogenous HNF-1a, are more physiologically similar to pancreatic beta-cells, the endogenous protein may mask the effect of transfected mutants in a way that would not naturally occur in carriers (Galan et al, 2011). To address the limitations associated with the

interpretation of data from a single experimental system, the impact of the alleles on HNF-1a function was also evaluated in Cos7 cells, which lack endogenous *HNF1A*, providing an opportunity for more straightforward interpretations of allele effects in a potentially less noisy system. Galán and colleagues found that DNA binding and transactivation analyses correlated across different promoters and cell types. They speculated that *in vitro* promoter- and tissue-dependent variability in the performance of allele-harbouring HNF-1a might help to explain the complex and progressive nature of the HNF1A-MODY phenotype (Galan et al, 2011).

The study which identified the rare nonsynonymous missense *HNF1A* allele, p.E508K, in Mexicans and US Latinos, exploited similar methods to annotate the variant functionally (Consortium STD, 2014). As p.E508K lies within the transactivation domain of *HNF1A*, investigators from the SIGMA Type 2 Diabetes consortium performed a thorough set of transcriptional activation assays. These assays compared the transfected mutant's ability to induce luciferase reporter gene expression via multiple promoter sites (HNF-4a P2, GLUT-2, rat albumin) and two different cell lines (HeLa and Min6), against wild-type HNF-1A, three known HNF1A-MODY mutations (p.P447L, p.R229Q, p.P379fsdelCT), and a wild-type-like-functioning *HNF1A* mutant reported in a single patient with T2D (p.M490T) (Consortium STD, 2014). Consistently, across all promoters and cell types, they found that the allele was less transcriptionally active than the wild-type protein ($p < 0.0001$) and p.M490T. However, the variant showed much greater transactivation than the three variants previously identified in HNF1A-MODY families (Consortium STD, 2014). In their EMSA DNA-binding analysis, the ability of the p.E508K mutant protein to bind rat albumin promoter was

indistinguishable from that of the wild-type protein. A third complementary approach to characterize the function of the mutant protein was to assess the intracellular localization in the cell lines, using a previously reported cytosol-retained variant, p.Q466X, as a control (*in vivo*, this PTV would be subject to nonsense-mediated decay [NMD] thereby causing haploinsufficiency) (Consortium STD, 2014).

Based on these experimental and statistically evaluated data on a single variant, along with phenotypic characteristics of the p.E508K carriers from a large population, a more confident assessment of pathogenicity is possible. Another benefit of this study is the specification of limitations, including imperfect exome coverage, and avoidance of an ascertainment bias which allowed for a reliable assessment of allelic penetrance (**Figures 1.1 and 1.2**) (Consortium STD, 2014). Overall, the authors were able to conclude that p.E508K impacts protein function, albeit not as severely as other *bona fide* HNF1A-MODY mutations. This is in line with the association of the variant with T2D risk elevation in the studied population, rather than MODY causation (**Figure 1.2**).

The private Oji-Cree variant associated with early-onset T2D, p.G319S, was also functionally investigated (Triggs-Raine et al, 2002). Triggs-Raine et al evaluated the impact of the variant on transactivation (luciferase reporter experiments in HeLa cells), protein stability (Western blot analysis) and DNA binding (EMSAs). The ability of the allele to activate transcription was roughly 54% compared to WT, with no discernable impact on protein stability or DNA binding properties (Triggs-Raine et al, 2002). Again, this is consistent with milder *in vitro* functional effects of *HNF1A* T2D-associated risk alleles compared to those reported to be MODY-causal (**Figure 1.2).**

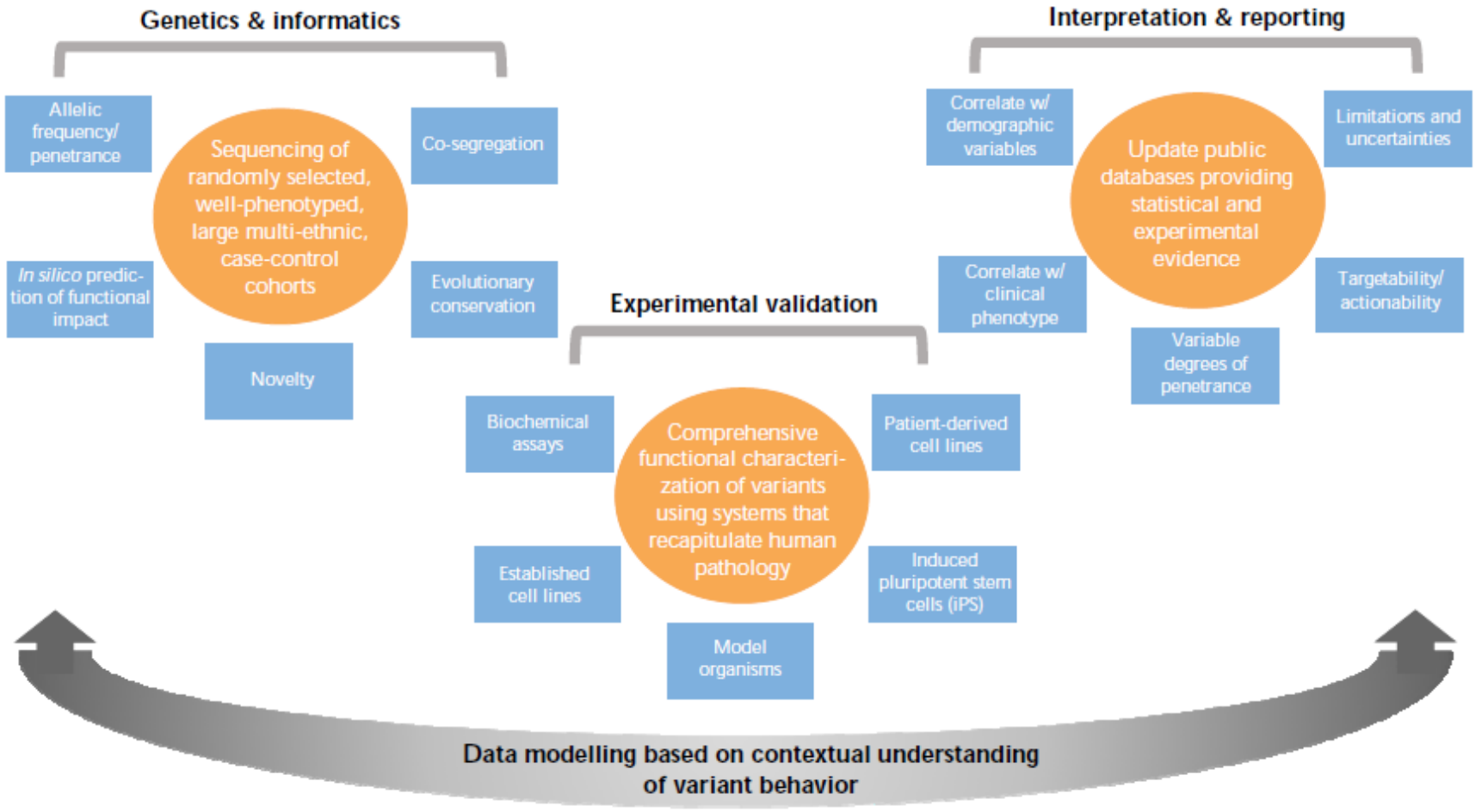


Figure 1.4. Recommended guidelines for a systematic and objective approach to identifying, characterizing, and assessing the pathogenicity of rare protein-coding genetic variants. (MacArthur et al, 2014)

The impact of p.G319S on *HNF1A* mRNA stability was investigated in another study (Harries et al, 2008). Harries and colleagues hypothesized that the variant might exert its effect through splice site alteration as it sits in the 3' end of exon 4. To test this hypothesis, they characterised *HNF1A* transcripts by RT-PCR in Epstein Barr virus (EBV)-transformed lymphoblastoid cell lines derived from an *HNF1A* p.G319S homozygote and two unrelated controls (Harries et al, 2008). They discovered that p.G319S produces two abnormal transcript isoforms with premature termination codons. Full length p.G319S represented only 24% of *HNF1A* mRNA transcripts (~11% of active HNF-1a protein in p.G319S homozygotes with residual activity of 66%) (Harries et al, 2008).

The nonsynonymous common alleles in *HNF1A* associated with increased T2D risk, L27 and V98, were interrogated both *in vitro* and *in vivo* for effects on protein function (Holmkvist et al, 2006). When assessed in combination, the alleles exhibited moderately reduced transcriptional activity (50-70%) in reporter gene assays in both HeLa and INS-1 cells using GLUT2 promoter (Holmkvist et al, 2006). When evaluated independently, significant results were only reported for L27 in HeLa cell experiments (~70% activity) (Holmkvist et al, 2006). These polymorphisms also demonstrated a modest yet significant reduction in insulin response to glucose *in vivo* (Holmkvist et al, 2006).

A very recent study assessed the impact of 27 rare missense variants (MAF < 1%) on HNF-1a transcriptional activity (luciferase reporter assay using rat albumin promoter in HeLa cells), DNA binding ability (EMSA) and nuclear localization (immunofluorescence in HeLa cells) (Najmi et al, 2017). These variants were originally identified through exome sequencing of population-based and T2D case-control cohorts as part of the aforementioned effort by Flannick et

al to evaluate the impact of rare monogenic diabetes variants on phenotypic outcome (Flannick et al, 2013; Najmi et al, 2017). Eleven of the 27 tested alleles exhibited impaired transactivation, DNA binding or nuclear localization compared to WT, though they were easily distinguishable in behavior from MODY-causal alleles used as positive controls in the study (p.P112L, p.P447L, p.Q466X [nuclear localization control], p.R263C [DNA binding control]). A T2D risk prediction model in which only alleles that demonstrated <60% transcriptional activity compared to WT were included showed robust association with T2D using counts from cases and controls in Framingham Heart Study (FHS) Offspring cohort (n = 1653), the Jackson Heart Study (JHS) cohort (n = 1691), and the Extreme T2D cohort (n = 771), in which the variants were originally detected (OR 5.04; P = 0.0007) (Najmi et al, 2017).

There is also a well-established series of biochemical assays for kinetically characterizing variants identified in GCK-MODY cases. Of the 620 MODY mutations which have been identified and reported in *GCK*, about 80 of them have been functionally characterized (based on the last mutation update by Osbak et al, published in 2009) (Osbak et al, 2009). For MODY-specific heterozygous inactivating mutations in GCK, the enzyme's maximal specific activity (K_{cat}) is usually decreased accompanied by various trends in S0.5 (glucose affinity at half maximal activity), ATP Km (ATP concentration at half maximal activity), and Hill coefficient values (Osbak et al, 2009) **(Figure 1.5).** The phosphorylating capacity of the glucokinase enzyme is normally diminished in proteins harboring MODY-associated variants. This is a functional defect reflected by the mild hyperglycemia in GCK-MODY patients, which is due to reduced glucose processing (Davis et al, 1999). However, kinetic dysfunction *in*

vitro does not directly correlate with the clinical phenotype, as the wild-type *GCK* allele in mutation-carrying GCK-MODY patients plays a compensatory role, and is upregulated by glucose at a posttranslational level (Liang et al, 1992). Other enzyme characteristics such as protein stability and regulation by glucokinase regulatory protein (GKRP) often explain mutational mechanisms in cases where kinetic data are inconsistent (Burke et al, 1999; Gloyn et al, 2005; Arden et al, 2007; Raimondo et al, 2014). GCK-MODY mutations have been catalytically assessed in physiologically relevant cell lines such as Min6 (Arden et al, 2007), and the enzyme's kinetics have been characterized in pancreatic islets as a function of glucose concentration (Liang et al, 1992). Moreover, unlike HNF1A-MODY mutations, global and beta-cell-specific heterozygous knockouts of *GCK* in mice recapitulate the moderate early-onset diabetic phenotype and reduced glucose-stimulated insulin secretory response in GCK-MODY patients (Grupe et al, 1995; Terauchi et al, 1995; Postic et al, 1999).

Transparent reporting of findings is equally pivotal, as this allows clinicians and scientists to make independent interpretations and assessments of the published data and the way in which they were generated. Combined family and multi-tiered functional studies of rare nonsynonymous glucokinase alleles have shown that analyzing variants, previously reported and characterized as pathogenic, sometimes showed the opposite effect (i.e. benign impact on function) (Steele et al, 2011; Beer et al, 2012). Steele and colleagues found that a heterozygous *GCK* missense variant, p.T342P, which was previously reported as pathogenic loss-of-function, neither demonstrated impaired catalytic activity nor co-segregation with the disease phenotype in the two families in which it was identified (Steele et al, 2011). The variant was also classified as benign by three different *in silico*

prediction tools, SIFT, PolyPhen, and Mutation Taster (Steele et al, 2011). The p.T342P substitution was not inherited by relatives with an overtly diabetic or elevated fasting glucose phenotype, and other relatives in whom the allele was identified were euglycemic (<5.5 mmol/l fasting glucose) (Steele et al, 2011).

Another functional effort, which aimed to characterize 6 heterozygous missense variants identified in combination in 3 individuals referred for GCK-MODY genetic testing, also refuted some classically held notions regarding interpretation and pathogenicity assignment of heterozygous inactivating missense *GCK* variants (Beer et al, 2012). Beer and colleagues found that two mutations, p.R43H and p.G68H, present in *trans* and independently reported to be causal for GCK-MODY, were associated with a moderate hyperglycemic phenotype rather than permanent neonatal diabetes mellitus (PNDM), the disease phenotype typically associated with *trans* GCK-inactivating mutations **(Figure 1.5)** (Beer et al, 2012).

A range of functional assessments, including kinetic and thermostability tests as well as evaluations of the behavior of mutation-harboring GCK in the presence of glucokinase activators and GKRP, revealed that R43H-GCK protein instability explained pathogenicity of the allele, despite wild-type-like catalytic function. GCK-p.G68D, which was inherited from an unaffected mother, was neither unstable nor abnormal kinetically, besides very mild activation (Beer et al, 2012). The familial and functional evidence of p.G68D neutrality challenges previous reports of p.G68D pathogenicity. In sum, these studies emphasize the need for thorough and critical evaluation of identified mutations before defining and reporting these as clinically significant and likely to be disease-relevant.

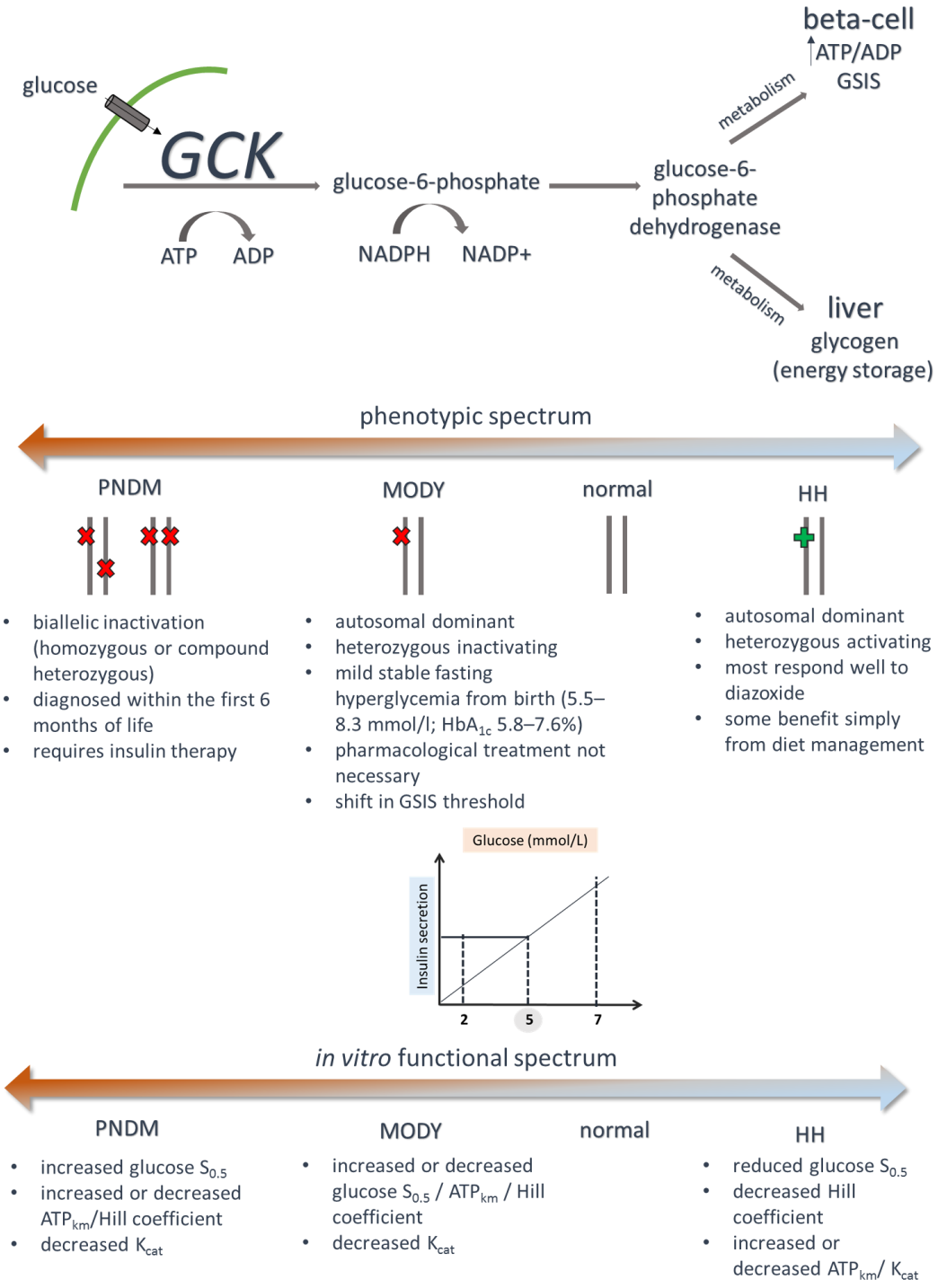


Figure 1.4. Genetic variants in glucokinase (GCK) give rise to a spectrum of glycemic disorders and display variable functional kinetic trends *in vitro*. Abbreviations: PNDM = permanent neonatal diabetes mellitus; MODY= maturity onset diabetes of the young; HH = hyperinsulinemic hypoglycemia. (Osbak et al, 2009; Chakera et al, 2015; Stride et al, 2013; Matschinsky et al, 2002).

1.5.2 Advances in the molecular characterization of sequence variants

(i) Human-derived cellular systems

The human pancreas remains largely inaccessible for experimental manipulation, and researchers continue to rely primarily on organ donation and transplantation systems. Therefore, human pancreatic beta-cell lines (Ravassard et al, 2011; Scharfmann et al, 2014) and induced pluripotent stem cells (iPS) (Hua et al, 2013; Shang et al, 2014; Stepniewski et al, 2015), are by far the most informative models for studying MODY and the human-specific phenotypic manifestations. Cellular phenotyping in patient-derived iPS cells provides the necessary and highly sought after contextual information (e.g. haplotypes and genomic background) which are otherwise absent from other systems or difficult to access from a technical and analytical point of view. Patient-derived stem cells also allow for reliable assessments of pharmacological drug targeting and response, in addition to the opportunity to perform genetic rescue experiments which provide unparalleled insight into molecular mechanism and causality. In the context of MODY, most disease-causing genes play critical regulatory and developmental roles in more than one tissue type, such as *HNF1B* in the kidney and *GCK*, *HNF1A*, and *HNF4A* in the hepatocyte. Therefore, patient-derived iPS cells provide an opportunity for simultaneous investigations of tissue-specific effects via unique maturation switches (Teo et al, 2013). The insulin-producing pancreatic beta-cells, recently developed *in vitro* from human pluripotent stem cells (hPSCs), also provide an excellent genomic and cellular context for such interrogations, not to mention their immense potential in cell replacement therapy (Pagliuca et al, 2014; Rezania et al, 2014).

Teo and colleagues describe the *in vitro* development of human iPS cells derived from skin biopsies of MODY patients, each with a mutation in *HNF1A*, *HNF1B*, *HNF4A*, *GCK*, or *CEL*. These were morphologically and functionally identical to human embryonic stem cells and other human pluripotent stem cells, expressing pluripotency markers OCT4, SOX2, NANOG, SSEA-4, and TRA-1-60 (Teo et al, 2013). Since then, a study was published in which investigators performed a thorough characterization of reprogrammed human iPS cells derived from primary fibroblasts of patients with HNF1A-MODY (Stepniewski et al, 2015). As more findings emerge from such pre-clinical disease-modeling systems, our knowledge of the pathophysiological basis of MODY, and of the precise biological mechanisms underlying beta-cell dysfunction, will be enhanced significantly.

Another successful patient-derived stem cell endeavor was described, although it recapitulated specifically the phenotype of a rare syndromic form of diabetes known as Wolfram syndrome (MIM222300) (Shang et al, 2014; Urano 2014). Autosomal recessive or compound heterozygous mutations *WFS1*, which encodes an ER transmembrane protein, give rise to Wolfram syndrome (Urano 2014). The most prominent clinical features of Wolfram syndrome include juvenile insulin-dependent diabetes, optic atrophy, and deafness (Barrett et al, 1995; Marshall et al, 2013). Shang and colleagues created insulin-producing C-peptide-positive beta-cells from skin fibroblasts of patients with Wolfram syndrome and control subjects (Shang et al, 2014). After differentiation, the mutations in *WFS1* inducing the insulin-dependent diabetic phenotype in the Wolfram patients yielded characteristically low insulin content and increased activity of the unfolded protein response (UPR) (Shang et al, 2014). Upon exposing the cells to the chemical chaperone, 4-phenyl butyric acid (4PBA), UPR pathway activity was

reduced and insulin content was adjusted to levels similar to those in control cells, suggesting 4PBA as a novel therapeutic candidate (Shang et al, 2014). *WFS1* locus rescue restored insulin levels and maintained normal insulin response to secretagogues, even under conditions of endoplasmic reticulum stress, in which cells otherwise demonstrated impaired insulin processing (Shang et al, 2014).

(ii) Medium-throughput functional annotation, high-throughput variant library scanning, and genome-editing applications

The patient-tissue-based techniques outlined above are excellent experimental models for candidate gene and variant analyses. However, these remain in early stages of technical optimisation and implementation, and are thus rather low in throughput and limited in scale. Until they are optimised, scalable, and better understood, medium-to-high-throughput variant screens in *in vitro* cell systems amenable to multiplexing currently represent the most attractive approach for functional interrogation of a large numbers of alleles in parallel (MacArthur et al, 2014).

To evaluate the pathogenicity of 18 rare nonsynonymous alleles in *GCKR*, the gene encoding glucokinase regulatory protein (GKRP), identified in n=458 hypertriglyceridemia cases and n=333 controls, Rees and colleagues performed a series of complementary functional assessments (Rees et al, 2014). GKRP interacts with and regulates the function of hepatic GCK, and it has been reported that rare nonsynonymous variants in *GCKR* are enriched in cases of hypertriglyceridemia (Rees et al, 2014; Johansen et al, 2010). Using an array of techniques with varied levels of throughput, they interrogated the biochemical

(cell-free) and cell biological (primary mouse hepatocytes, HeLa cells and Cos7 cells) effects of the *GCKR* alleles (Rees et al, 2014).

The ability of allele-harbouring GKRP to interact with GCK protein and inhibit its function was investigated spectrophotometrically using glucose-6-phosphate dehydrogenase-linked kinetic assays, fluorescence microscopy, and homogenous time-resolved fluorescence (HTRF) which enables cell free assessment of protein-protein interactions of multiple variants in parallel (Rees et al, 2014). Ultimately, the rare alleles which demonstrated impaired function across the various assays did not co-segregate with triglyceride levels in the re-evaluated pedigrees, highlighting the complexity and multifactorial nature of lipid traits (Rees et al, 2014).

A study on rare nonsynonymous missense variants identified in *PPARG* from exome sequencing of a large multi-ethnic case-control type 2 diabetes cohort implemented a novel quantitative medium-throughput differentiation assay (Majithia et al, 2014). Majithia and colleagues combined bioinformatic prediction, systematic molecular characterization, drug response, and clinical phenotypic data to aid the interpretation of variants in *PPARG*, and accurately determine their association with type 2 diabetes risk. The function of 49 variants was evaluated on the basis of their ability to rescue differentiation of human preadipocytes *in vitro* (Majithia et al, 2014). The investigators evaluated the impact of the variants on differentiation by combining high-content microscopy and customized image analysis (Majithia et al, 2014). Interestingly, they found that individuals from their cohort who were heterozygous carriers of variants in *PPARG* with significantly reduced function in their characterization pipeline, showed a sevenfold increase in T2D risk. In contrast, when the investigators analyzed all of the rare variants

identified in *PPARG*, without filtering for *in vitro* dysfunction, they did not observe a significant association with T2D risk (Majithia et al, 2014).

Another study employed similar techniques for characterizing variants in low-density lipoprotein receptor (*LDLR*) (Thormaehlen et al. 2015). Thormaehlan and colleagues leveraged the exomes of a large European cohort of early-onset myocardial infarction (MI) cases and MI-free controls to identify rare missense variants in *LDLR* associated with premature MI. Their aim was to discriminate disruptive, functional, and potentially pathogenic alleles from the majority of neutral missense variants (Thormaehlen et al, 2015). Their approach involved consolidation of multiple complementary datasets: *in silico* prediction, locus-specific *a priori* clinical information, systematic cell-based functional profiling of identified alleles, and available phenotypic information. The impact of naturally occurring missense alleles on LDLR protein function was assessed *in vitro* via uptake of fluorescently labeled LDL using microscopy coupled with a customized quantitative multi-parametric image analysis pipeline (Thormaehlen et al, 2015).

New technologies have enabled massively parallel mutational scanning and synthesis of site-saturation libraries of tens of thousands of regulatory (MPRAs) and protein-coding variants (MAVEs) (Melnikov et al, 2012; Melnikov et al, 2014) (Majithia et al, 2016; Gasperini et al, 2016). Indeed, the advent of Multiplexed (functional) Assays of Variant Effects (MAVEs) has made it possible to, in principle, map sequence-function relationships at once in a single experiment for nearly all possible variants in disease-relevant genes (Gasperini et al, 2016; Starita et al, 2017). By prospectively scoring the function of alleles based on various cellular phenotypes, pooled molecular characterization assays will assist tremendously in overcoming the variant-interpretation crisis.

In fact, these functional genomic innovations have already been implemented for variants in two medically relevant genes, *PPARG* and *BRCA1* (Majithia et al, 2016; Starita et al, 2015). Starita and colleagues performed a MAVE in which they interrogated the function of 2,000 missense alleles in the RING domain of BRCA1 - the most medically relevant element of the gene where more than half of the pathogenic variants reside (Starita et al, 2015). Each allele in the library was assigned two function scores, as well as a combined score, on the basis of two distinct molecular mechanisms in multiplexed biochemical assays: the impact of RING domain variants on E3 ubiquitin ligase activity and ability to bind to the RING domain of BARD1. A supporter vector regression (SVL) model trained using combined function scores was used to generate a high confidence look up table of predicted ability of RING domain variants to rescue homology-directed DNA repair (HDR) – the essential role of *BRCA1* in tumor suppression (Starita et al, 2016).

Majithia and colleagues used a single functional signature – namely, transactivation – to annotate a site saturation library of all possible single-amino acid substitutions in *PPARG* (n=9,595 missense variants). Using flow cytometry, they were able to discriminate benign *PPARG* alleles from damaging and pathogenic lipodystrophy variants based on *PPARG*-dependent CD36 fluorescent signal on the surface of THP1 macrophages. Function scores that were assigned to each allele based on frequency in each of the two activity bins (CD36+ or CD36-) were used to build and train a variant classifier by supervised machine learning. The classifier, which has been tested against nearly 60 novel missense alleles ascertained in clinical and population-based sequencing, is publically available as a web application for the community to utilise and update.

Genome-scale nucleic editing and functional screening via RNA-guided CRISPR (clustered regularly interspaced short palindromic repeat)-associated Cas9 nuclease technology represent another promising trajectory for scalable variant annotation (Mali et al. 2013; Hsu et al. 2014; Shalem et al, 2015). CRISPR-Cas9-mediated screens allow for a wide variety of targeted, sequence-specific perturbations to achieve unbiased complementary cellular phenotypic readouts with high efficiency, including knockdown, knockout, and activation (Shalem et al. 2015). Genomic editing by Cas9 provides an opportunity to answer context-dependent experimental and clinical questions, such as whether cellular phenotype from a patient-derived sample is directly attributable to the effect of a sequence variant, or is influenced by regulatory modifiers and/or genetic background. Moreover, with CRISPR-Cas9 technology, *in vivo* levels of gene dosage and expression can be replicated, unlike the majority of artificial cell-based manipulations, which are typically limited to overexpression or incomplete silencing of the gene(s) of interest.

Indeed, understanding the functional consequences of known and novel sequence variants is crucial for making accurate estimates of variant pathogenicity (rare disease causality, common disease risk modification, or benign impact). Multi-tiered low- and medium- throughput functional investigations have aided immensely in the interpretation of variants and will no doubt be utilised to calibrate high-throughput-derived annotations. Though, the ‘interpretive void’ is likely to expand as an ever increasing number of variants of uncertain clinical significance (VUS) are detected with increasing sequencing applications in research and clinical settings (Cooper 2015). Post hoc functional interrogations are too slow to benefit the carrier of a VUS. Emerging technologies

such as MAVEs are making it possible to prospectively annotate the function of genetic perturbations massively in parallel (Starita et al, 2017).

High-throughput functional datasets will grow in size with increased application of MAVEs and MPRAs - statistical models trained with these data will (a) enable generation of pre-computed function scores for clinically relevant loci, and (b) improve existing prediction algorithms which operate on the basis of multi-data integration (functional data, protein sequence homology, conservation, structure, variant frequency, amino acid substitutions) (Starita et al, 2017; Eilbeck et al, 2017).

1.6 Thesis Aims

My doctoral research projects share the common goal of mapping sequence-function relationships for variants in *HNF1A* and *GCK* to maximize their diagnostic, prognostic and therapeutic utility.

Aim 1 (Chapter 3): My objective was to develop an analytical framework for robust and unbiased variant classification using multi-dimensional functional follow-up data from exome-identified missense alleles in *HNF1A*. The method should facilitate a) informative spatial organization of variants along the *HNF1A* pathophysiological continuum (**Figure 1.2)**, b) an exploration of the contributions of each *in vitro* molecular mechanism on functional stratification, and c) mapping of variant effects on molecular function (defined by multiple mechanistic angles) with phenotypic outcome using clinical and genetic data from a national MODY diagnostic registry (UK).

Aim 2 (Chapter 4): My aim was to build and optimise a high-throughput *in vitro* assay that would enable functional screening of a saturation mutagenesis library harboring all possible missense substitutions in *HNF1A*. More specifically, I sought to develop a MAVE for *HNF1A* with transcriptional activity as a functional signature in a cell-based assay, and to interrogate the system's discriminatory capacity using a well-characterized *HNF1A* allelic series with robust clinical, functional, and genetic evidence.

Aim 3 (Chapter 5): My goal was to assist in the clinical interpretation of two novel *GCK* missense variants of uncertain clinical significance (VUS), identified through genetic testing, by performing comprehensive *in vitro* functional characterisation experiments.

2

General Methods

This chapter provides detailed descriptions of the more common and well-established *in vitro* techniques I have applied throughout my doctoral research. Methods which I have applied or developed for specific chapter aims are described in the methods section of each results chapter. All procedures were carried out in accordance with institutional biosafety requirements.

2.1 Cell culturing conditions

Human hepatocellular carcinoma HepG2 (endogenous reporter screen system for *HNF1A*; **Chapter 4**), human cervical cancer HeLa (artificial reporter screen system for *HNF1A*; **Chapter 4**), and transformed human embryonic kidney HEK 293T (for producing and packaging lentiviral particles; **subsection 2.8**) cells were cultured at 37°C as detailed in **Table 2.1**.

All cell lines were tested negative for mycoplasma contamination using the MycoAlert Assay kit (Lonza). For cell maintenance, briefly, cells were washed with 1X phosphate buffered saline (PBS) (Sigma Aldrich). In a T75 flask, 1 ml of TrypLE Select (Gibco 12563–011) was added to each flask for 3-5 min at 37°C to dislodge cells. To neutralise the trypsin solution, 9 ml of culture medium was added to the flask. Depending on cell line, cells were spun down at 200 g for 5 min post-neutralisation in a centrifuge (Beckman GS-15R, Beckman Coulter), then resuspended in 10 ml of fresh culture medium. Cells were counted using the Cellometer Auto T4 cell counter (Nexcelom Bioscience).

Table 2.1. Cell line culture medium and cell-specific culturing procedures.

HepG2	Dulbecco's Modified Eagle's Medium (DMEM) (D6429, Sigma Aldrich), 10% (v/v) foetal bovine serum (FBS) (10500-064, Gibco by Life Technologies), 100 U/ml penicillin and 100 µg/ml streptomycin (15140122, Gibco by Life Technologies)	Cells were split 1:10 twice weekly. Re-suspended cells were filtered through a 40 µm cell strainer (Corning) before seeding.
HEK293T	DMEM (D6429, Sigma Aldrich), 10% (v/v) FBS	Cells were split in a 1:20 ratio twice weekly.
HeLa	DMEM (D6429, Sigma Aldrich), 10% (v/v) FBS, 100 U/ml penicillin, 100 µg/ml streptomycin	Cells were split 1:10 twice weekly.

2.2 Fluorescence *in situ* hybridisation (FISH) to characterise HNF1A copy number in HepG2 cells

Fluorescent in situ hybridization of *HNF1A* at chr12:121,416,371-121,440,314 was performed by the chromosome dynamics team at the Wellcome Centre for Human Genetics (Elsie Hodimont and Catharine Green). The Bacterial Artificial Chromosome (BAC) probe CTD-3028F22 at chr12:121,404,747-121,553,715 was used as well as BAC RP11-148L11 from their in-house library which covers the chromosome 12 telomere. As a control for the FISH experiment, the BACs were labelled and hybridized to metaphase spreads from peripheral blood of a healthy donor. The CTD-3028F22 mapped to the correct genomic location in the control cells The two BACs were hybridized to metaphase spreads from the HepG2, a minimum of 200 cells (metaphase and interphase) were analysed. The percentage of cells showing one, two or more signals by probe per cell is summarized in **Table 2.2**.

Table 2.2 *HNF1A* copy number analysis in HepG2 cells

	1 signal by Chr 12 (3 copies)	1 signal by Chr 12 (4 copies)	Extra copy not on Chr12
HepG2	8.5%	89.2%	2.3%

2.3 DNA quantification and visualisation

DNA quantity (ng/µl) and quality assessments were conducted by measuring sample absorbance at 280nm using Nanodrop 2000 (Thermo Scientific). Quality was evaluated by confirming 260/280nm absorbance ratio between 1.8 and 2.0 (minimum protein contamination) and a 260/230nm absorbance ratio between 2.0-2.2 (minimum contamination with phenol/other organic compounds). Concentration and purity evaluation were undertaken with 2µl of the DNA sample using the nucleic acids and DNA50 application programmes following blanking and calibration of the Nanodrop instrument with distilled water (or appropriate elution buffer if used to collect DNA sample).

One percent agarose gels were made from 1g agarose (Sigma Aldrich Company Ltd, Gillingham, UK) dissolved in 100ml Tris/Borate/Ethylenediaminetetraacetic acid (TBE) buffer (NBS Biologicals, Huntingdon, UK); the solution was then heated in an 800W microwave for 2 minutes to aid dissolution. Upon cool-down to temperature below 50°C, I added 2.5µl ethidium bromide (EtBr), an intercalator of DNA, to the solution which was then poured into a MultiSUB gel electrophoresis tank with well combs (Appleton Woods, Ltd, Birmingham, UK). The gel was then allowed to solidify at room temperature for 30 minutes, at which point the tank would be filled with 500ml of 1X TBE buffer. A minimum of 20ng of DNA would be mixed with orange-G loading buffer (Sigma Aldrich Company Ltd)

at 2% DNA volume by a brief vortex, and loaded onto the gel alongside 5-10 ul of a 100bp or 1Kb DNA ladder (depending on size of product) (New England Biolabs UK, Ltd, Hitchin, UK and Promega UK Ltd, Southampton, UK, respectively). Electrophoresis was carried out by passing a 120V through the gel for 45-60 min (depending on size of product to be resolved). The gel was then viewed with a ChemiDoc XRS+ system and Image Lab ™ software (Bio-Rad Laboratories, Hemel Hempstead, UK).

2.4 Molecular cloning and mutagenesis

2.4.1 Site-Directed Mutagenesis (SDM)

Site-directed mutagenesis (SDM) was carried out on the relevant *HNF1A* or *GCK* cDNA sequences (**subsection 2.13**) to study the genetic variants described in Chapters 4 and 5. Primers were designed using the Quikchange Primer Design (Agilent Technologies) software, which is optimised to design primers that are 25 to 45 bases in length, of melting temperature ≥78°C, with minimum GC content of 40%, that terminate in one or more C or G bases, and that contain base mismatches located close to the middle of the primer. Primer sequences are found in the relevant Chapters 3, 4 and 5 and were all purchased as salt-free oligonucleotides from Eurofins Genomics (Germany).

SDM reactions were performed using the Quikchange II Site-Directed Mutagenesis (Agilent Technologies) kit and the protocol adapted from the manufacturer's guidelines. Each reaction was set up with 50 ng of template DNA and 50 ng/µl each of forward and reverse primers and performed on a tetrad thermal cycler (DNA Engine, MJ Research). See **Table 2.3** for the composition of the reaction mix and **Table 2.4** for the thermal cycler conditions for the

polymerase chain reaction (PCR). Upon completion of template amplification, reactions were cooled to 4°C for at least 2 min. Samples were then subjected to Dpn I-mediated endonuclease digestion for 1h at 37°C to remove template DNA which had not been mutagenised.

Table 2.3. Site-Directed Mutagenesis (SDM) reaction mix for one reaction.

Reagent	Volume (µl)
Reaction buffer (10X)	5
Double-stranded DNA template (10 ng/µl)	2.5
Forward primer (50 ng/µl)	5
Reverse primer (50 ng/µl)	5
dNTP mix	1
Nuclease-free H_2O	31.5
*Pfu*Ultra High Fidelity DNA polymerase (2.5U/µl)	1 (add last)

Primers were purchased from Eurofins Genomics. All other reagents were obtained from the Quikchange II Site-Directed Mutagenesis kit (Agilent Technologies).

Table 2.4. SDM thermal cycler conditions.

Segment	Cycles	Temperature	Time
DNA denaturation	1	95ºC	2 min
Primer annealing Extension	18	95ºC	1 min
		55ºC	1 min
		68ºC	min=amplicon kb
End	-	4ºC	∞

2.4.2 Sanger sequencing

All sequences introduced synthetically by Site-Directed Mutagenesis (SDM) and all ligation reactions were verified by Sanger sequencing. Sequencing primers were designed to either capture the ligation sites (integration) or to span the introduced nucleotide change by SDM. Samples were sent to Source BioScience (UK) or Eurofins Genomics (Germany).

2.4.3 Restriction digest

To create an empty vector or excise an open reading frame (ORF) of interest from plasmid DNA, the digestion reactions detailed in **Table 2.5** were set up.

Reactions were incubated at 37°C for 2-3 hours. An input DNA quantity of 1-5 µg was used for each reactions. The enzymes used for each unique restriction digest application are specified in the relevant chapters. The digested vector was further treated with Thermosensitive Alkaline Phosphatase (TSAP) (Promega) to remove phosphate groups from the 5' sticky ends to prevent self-ligation of the vector.

Table 2.5. DNA restriction digest protocol for one reaction

Reagents	**Quantity**
DNA	1 µg
10X buffer (multicore) (Promega)	2 µl
Enzyme 1 (Promega)	0.5 µl
Enzyme 2 (Promega)	0.5 µl
BSA (10 ug/ul) (Promega)	0.2 µl
Nuclease-free H_2O	up to 20 µl

Reactions were carried out in 1.5 ml microcentrifuge tubes. Reagents were obtained from Promega.

2.4.4 Purification of digested vector by gel extraction

Restriction digest products were size separated and visualised on agarose gels as per DNA visualisation method described previously (**subsection 2.3**). DNA bands corresponding to size separated linearised vector or insert were excised from the rest of the gel using a scalpel under ultraviolet (UV) light and placed in a 1.5 ml microcentrifuge tube. The gel may be stored at -20°C until extraction. The QIAquick Gel Extraction kit (Qiagen) was used to extract the DNA from the agarose gel according to manufacturer's guidelines. Sodium acetate at 3 M, pH 5 (BDH Chemicals) was added at 10 µl per sample. The DNA was eluted in 30 µl of nuclease-free water (Ambion, ThermoFisher Scientific) and quantified before use (Nanodrop 2000).

2.4.5 Ligation

Ligation reactions were set up using the T4 DNA ligase and ligase buffer (**Table 2.6**) (Promega). The insert was added in excess in the ligation reaction typically using a vector to insert molar ratio of 1 to 3 (ligation calculator http://www.insilico.uni-duesseldorf.de/Lig_Input.html). Samples were incubated overnight at 4°C and collected the next day for transformation into E. coli cells as described **in subsection 2.5.1**.

Table 2.6. DNA ligation protocol using T4 DNA polymerase.

Reagents	Quantity
Vector DNA	X (3:1 molar ratio insert:vector)
Insert DNA	X (3:1 molar ratio insert:vector)
10X T4 DNA ligase buffer	2 µl
T4 DNA ligase	1 µl
Nuclease-free H_2O	up to reaction volume (10-20 µl)

Reagents were obtained from Promega.

2.4.6 Polymerase Chain Reaction (PCR)

To isolate/linearise mCherry red fluorescent reporter sequence for molecular cloning downstream human albumin promoter, it was PCR amplified from pCW57d-PPARGopt-cherrypk-pgk-rtta (mCherry in-frame with *PPARG*; vector supplied by Dr. Amit Majithia, Broad Institute of MIT and Harvard) using the primers, reaction and amplification conditions detailed in **Tables 2.7, 2.8 and 2.9**. The reason the sequence was amplified rather than excised by restriction digest was that enzyme recognition sites were incompatible with the destination vector (pGL3-Basic vector backbone [Promega]) into which human albumin promoter sequence had been cloned by Dr. Jana Rundle for *HNF1A* luciferase reporter gene assays. PCR products were purified using QIAquick PCR purification Kit (Qiagen cat no.28104) and analysed on a 1% agarose gel (mCherry amplicon size 941 bp).

The mCherry PCR amplification primer sequences were designed to include 3 bases flanking the restriction enzyme (RE) recognition sites in the destination vector (pGL3-albumin-luciferase) as in **Table 2.7.**

Table 2.7 Primers for mCherry amplification with integrated restriction sites.

	Primer Sequence (5' to 3')
mCherry_Forward NcoI RE site	5'-GCACCATGGATGATGGTGGACGGGGATAACTCC-3'
mCherry_Reverse XbaI RE site	5'-GACTCTAGATCACTTATACAGCTCGTCCATCCCG-3'

GCA and GAC chosen as anchoring bases by virtue of being bases immediately upstream NcoI and XbaI, respectively, in the pGL3-albumin-luciferase destination vector. Primers were purchased from Eurofins Genomics.

Table 2.8. Phusion PCR mix for one reaction.

Reagents	**µl**
5X High-Fidelity buffer	4
dNTP (10mM)	0.4
Forward Primer 10uM	0.5
Reverse Primer 10uM	0.5
DNA (25 ng/ul)	2
Phusion High-Fidelity DNA Polymerase (Thermo Scientific)	0.2
H_2O	up to 20

Table 2.9. Phusion PCR thermal cycler conditions.

		2-step protocol
Temperature	**Time**	**Cycles**
98°C	30 sec	1
98°C	10 sec	34X
72°C	1 min	1
72°C	10 min	1
4°C	End	1

2.5 Transformation of competent bacteria and plasmid DNA isolation

2.5.1 Transformation of competent Escherichia coli

Plasmid DNA was introduced into competent E. coli cells by transformation. One to two microlitres of ligation reaction, or 2-5 ul 1:10 dilution of ligation reaction in nuclease-free H2O (Ambion, ThermoFisher) or up to 100 ng of plasmid DNA was added to 50 µl of DH5α competent cells (Life Technologies) in 1.5 ml tubes on ice. Transformation reactions were mixed gently by flicking, followed by incubation on ice for 20 min. Reactions were subjected to heat shock at 42°C for 45 seconds (heat block), then immediately placed on ice for 2 min. For plasmids with kanamycin resistance cassette 500 µl of super optimal broth (SOC) media (Invitrogen) was added to each reaction and the reaction incubated for 60 min at 37°C in an orbital shaking incubator set at 225 rpm (Stuart Orbital Incubator S150, Jencons Scientific Ltd). Total reaction volume for amp-resistant bacteria (~50 µl), or 200 µl of kan-resistant bacteria pre-incubated in SOC, was spread evenly across LB (Lysogeny broth) agar plates with the relevant antibiotic – 25 µg/ml of kanamycin (Sigma Aldrich) or 50 µg/ml of ampicillin (Sigma Aldrich). Plates were incubated for 18 hours at 37°C.

2.5.2 Transformation of competent cells for SDM (XL-Blue)

Plasmid DNA was introduced into competent E. coli cells by transformation, in a protocol roughly identical to that described in **subsection 2.5.1**. Two microlitres of SDM reaction was added to 30 µl of XL1-Blue supercompetent cells (Agilent Technologies). After transformation and overnight incubation at 37°C, at least 4 isolated colonies were picked per plate for DNA extraction by miniprep (**subsection 2.5.3**) and the DNA outsourced for Sanger sequencing. Plasmid

samples with the desired alleles were selected and expanded by midiprep as described in **subection 2.5.4**. All midiprep samples generated were verified by sequencing as described in **subsections 2.5.4 and 2.4.2.**

2.5.3 DNA extraction by miniprep

Bacterial colonies were picked from LB agar culture plates and each single colony was used to inoculate 10 ml LB medium (Sigma Alrich) with antibiotic (50-100 µg/mL) and cultured for 16h (overnight) at 37°C in the orbital shaking incubator at 250 rpm (Stuart Orbital Incubator S150, Jencons Scientific Ltd). After incubation, 500 µl of bacterial sample was first collected and added to an equal volume of 37% DMSO (Sigma Aldrich) made up in distilled H_2O in a 1.5 ml tube. DMSO stocks were stored at -80°C and used for further plasmid preparation. The remaining cultured bacteria was pelleted at 13,000 rpm for 10 min in a Force 1618 benchtop microcentrifuge (Labnet International). The supernatant was discarded and the DNA from the pellet extracted using the QIAprep Spin Miniprep kit (Qiagen) or PureYield Plasmid Miniprep system (Promega), according to manufacturers' instructions. The DNA was eluted in 40 µl of RNase-free water and stored at -20°C.

2.5.4 DNA extraction by midiprep

The DMSO stock for the plasmid DNA to be extracted was scraped with a plastic pipette tip and used to inoculate 10 ml LB medium with antibiotic (50-100 µg/ml) and cultured for at least 5 hours at 37°C in the incubator shaking at 250 rpm (Stuart Orbital Incubator S150, Jencons Scientific Ltd). This starter culture was then added to 100 ml LB medium with antibiotic in 500 ml flasks and placed in the shaking incubator (Labnet 311DS, Labnet International) for 16 hours

(overnight) at 37°C. After incubation, the culture was transferred to 50 ml Falcon tubes (Corning) and pelleted at 3,500 rpm for 10 min (Heraeus Multifuge 3SR, Thermo Scientific). The supernatant was discarded and the DNA from the pellet extracted using the QIAprep Spin Midiprep kit (Qiagen) or PureYield Plasmid Midiprep system (Promega), according to manufacturers' instructions. The DNA was eluted in at least 450 µl of RNase-free water and stored at -20°C.

2.6 Antibiotic kill curves

Cells under interrogation were seeded at a density of 1.0-5.0x10^4 cells in 100 µl culture media in individual wells of 96-well flat clear bottom black polystyrene TC-treated microplates (Corning, product #3904). A range of puromycin (0, 0.5, 1, 1.5, 2, 2.5, 3 µg/ml) or hygromycin (for HeLa: 200-600 µg/ml; for HepG2: 200-1000 µg/ml) concentrations were added to the wells in duplicate. The assay scheme also included a negative control of DMEM culture media. After incubation at 37°C in a 5% CO2 atmosphere for 48hrs (puro), or 5 days (HeLa hygro; pulse every 48hrs), or 7 days (HepG2 hygro; pulse every 48hrs 48 hours) 10 µl alamarBlue™ Reagent (ThermoFisher #88951) were added to each well. Following a 4-hr incubation period, the change in solution colour (metabolic activity-indicator) was measured at A570 nm and A600 nm with the Molecular Devices M3 plate reader running SoftMax Pro (Software version 2.05.5). The percentage reduction of AlamarBlue™ Reagent was calculated according to the formula provided by the manufacturer (ThermoFisher Scientific #88951).

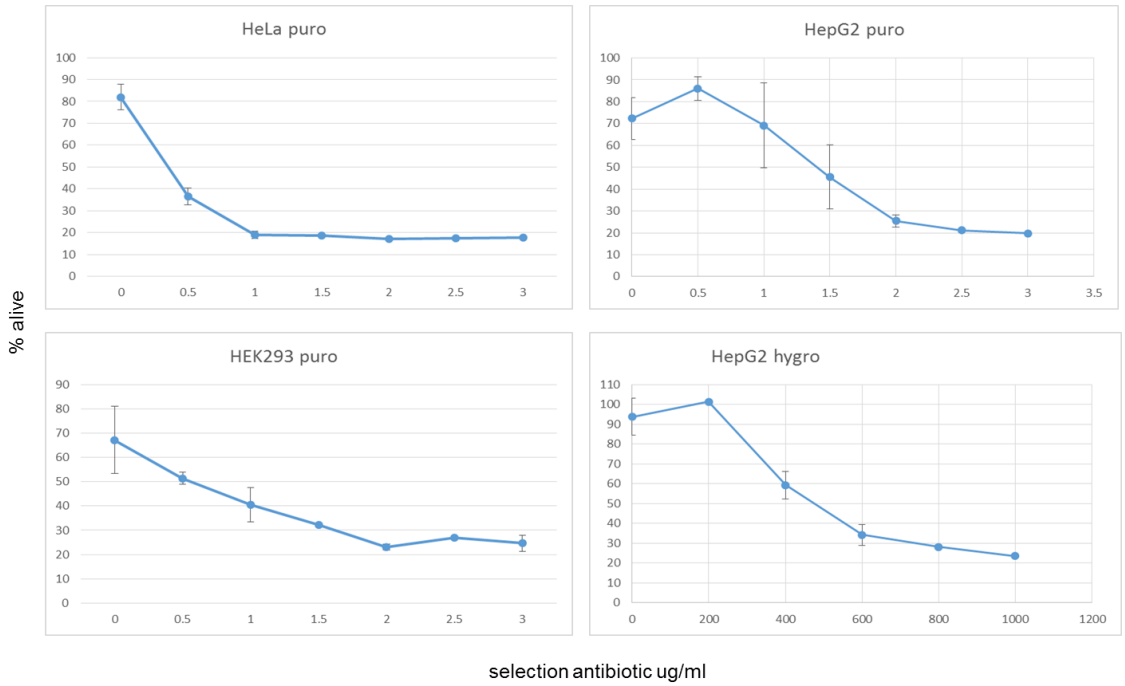


Figure 2.1. Viability of HepG2, HEK293T and HeLa cells following exposure to a range of antibiotic concentrations (puromycin 0-3 µ/ml; hygromycin (0-1000 µg/ml). Kill curves were generated over a time course of 48 hours and 7 days for puromycin and hygromycin, respectively. Viability was determined using alamarBlue™ Cell Viability Reagent.

2.7 Transient and stable transfections

HeLa cells were seeded 18-24 hours prior to transfection. Prior to transfection, the culture medium over cells was replaced with transfection medium (culture medium without FCS and antibiotics). Cells were transiently transfected with *HNF1A* expression and reporter gene vectors using the Lipofectamine® 2000 transfection reagent (Life Technologies) following manufacturer's guidelines. Transfection medium was removed and replaced with fresh complete medium 4-6 hours after transfection. The plating density ($6.5x10^5$ in 6-well plate format), amount of DNA (2.5 µg) and ratio of DNA to transfection reagent (5 ul Lipofectamine to 1 ug DNA) were determined after optimisation experiments. DNA and transfection reagent mix was brought up to 125 µl with Opti-MEM reduced serum medium (31985, Gibco, Life Technologies). To generate a stable cell line expressing mCherry reporter, albumin promoter and hygromycin resistance, reporter plasmid pgl4.26_hALB_mCherry **(see Chapter 4 subsection 4.2.2 and Figure 4.4 for plasmid details)** was used to transfect HeLa cells. Hygromycin B (ThermoFisher cat no. 10687010) was added to the cells 48 hrs post-transfection at a dose-response determined concentration of 550 µg/ml. The surviving hygromycin resistant colony was cultured under selection by hygromycin for a 3-week period, until cells demonstrated complete resistance to the antibiotic.

2.8 Lentiviral experiments

Viral production and infection was performed according to the Broad Institute's Gene Perturbation Platform (GPP) protocols available through their web portal: https://portals.broadinstitute.org/gpp/public/resources/protocols.

2.8.1 Lentiviral production

Lentiviral expression plasmid PCW-*HNF1A*-puroR [pCW57.1 Dest (pLXI_TRC401)] was packaged into viral particles by co-transfecting HEK293T cells with the 2nd generation packaging plasmid PAX (pCMV-dR8.74psPAX2) and envelope VSV-G-expressing plasmid pMD2.G using FuGENE® 6 Transfection Reagent (Promega) (**Table 2.10**).

Table 2.10. Transfection mixture for lentiviral packaging and production.

Reagent	Quantity in 10cm plate format
packaging plasmid PAX (pCMV-dR8.74psPAX2)	9 µg
envelope VSV-G-expressing plasmid pMD2.G	0.9 µg
HNF1A lentiviral transgene expression vector pCW57.1 Dest (pLXI_TRC401)	9 µg
Opti-MEM reduced serum medium (31985, Gibco, Life Technologies) to total volume	225 µl

All vectors supplied by Dr. Amit Majithia, collaborator at the Broad Institute at MIT and Harvard.

2.8.2 Lentiviral titration

To determine quality and concentration of produced viral particles, and to calculate Multiplicity of Infection (MOI) (amount of viral serum to achieve a desired number of viral particles/integration events per cell), lentiviral titres were measured as per the protocol detailed in **Table 2.11**.

2.8.3 Lentiviral infection

High quality (i.e. high titre) lentivirus was used to infect HepG2 cells at high MOIs (50% virus in culture media) for initial assessments and low MOI (0.8-1.2) (3.125-12.5% virus in culture media depending on titre) to replicate single variant entry per cell high-throughput conditions **(Table 2.12).**

Table 2.11. Lentiviral titration protocol.

Day 1	HEK293T are seeded at a density of $2.5x10^4$ in wells of 96-well flat clear bottom black polystyrene TC-treated microplates (Corning, product #3904).
Day 2	A 2-fold serial dilution in DMEM supplemented with 8 ug/ml polybrene (Sigma cat no. 107689) of each viral harvest is used to infect HEK293T cells; the serial dilution consists of 6 points whereby viral harvest makes up either 50, 25, 12.5, 6.25, 3.125, 1.5625% of the infection media requiring 12 wells per harvest for duplicate samples; control samples include a column of cells to which puro is added on Day 3 for selection and a column of cells to which polybrene supplemented media is added on Day 2 and are not subjected to puro selection on Day 3.
Day 3	After 16-18 hours (overnight infection), infection media is replaced with puromycin (2 ug/ml) supplemented media for a 48-hour selection of transduced cells.
Day 5	A cell viability titre is established using AlamarBlue™ reagent by replacing selection media with 1:10 dilution of reagent in media, and incubating the samples for 2hrs at 37°C in a 5% CO2 atmosphere; viability is analysed via fluorescent endpoint reads (either Synergy H4 reader 1404211C software version 2.05.5, or Molecular Devices M3 plate reader running SoftMax Pro software version 2.05.5) using excitation and emission wavelengths 550-555nm and 585-590nm, respectively, at 25.8°C. Multiplicity of Infection (MOI) is calculated using fluorescence values the natural logarithm (LN) of (1-fraction of alive cells). The proportion of alive cells is calculated per well using the following formula ({fluorescence – [avg fluorescence of +puro control column]}/{[avg fluorescence of -puro/+polybrene control column]-[avg fluorescence of +puro control column]}).

Table 2.12. Lentiviral infection protocol.

Day 1	Seed cells at appropriate density (For HepG2 cells this was $3.8\text{-}4.0x10^6$ in 10cm dish or $2.5x10^5$ cells in 6-well plate)
Day 2	Infect cells with lentivirus diluted in culture media supplemented with 8 µg/ml transduction agent polybrene (Sigma cat no. 107689)
Day 3	Remove infection media and replace with selection media (culture media supplemented with puromycin (2.5-3 µg/ml depending on cell density)
Day 5 (48 hrs puro selection)	Assay infected cells (i.e. re-plate for doxycycline dose induction; details in **Chapter 4 subsection 4.2.3**)

Protocol is optimised for HepG2 cells.

2.9 Expression analysis of *HNF1A* and downstream targets in HepG2 cells

2.9.1 RNA extraction

RNA was extracted from HepG2 cells using the TRIzol® reagent (Life Technologies) according to manufacturer's guidelines. After phase separation and isolation, RNA precipitation from the aqueous layer with isopropanol was carried out with overnight incubation at 4°C. Glycogen (R0551, Life Technologies) was also added during this incubation step to increase the efficiency of recovering RNA from small amounts of starting material. RNA pellets were resuspended in RNase-free H2O (Ambion, Thermo Fisher Scientific) and quantified using a NanoDrop 8000 spectrophotometer (Thermo Scientific). RNA samples were stored at -20°C.

2.9.2 cDNA synthesis

cDNA was synthesised using the SuperScript III First-Strand Synthesis kit (18080-051, Life Technologies) according to manufacturer's guidelines, from RNA samples obtained from HepG2 cells. Reactions were carried out in 96-well PCR plates (4titude Ltd) in the tetrad thermal cycler (DNA Engine, MJ Research). Fifty nanograms to 1 µg of input RNA was used and primed with oligo(dT) or a mixture of oligo(dT) and random hexamers where stated. cDNA synthesis reactions were stored at 4°C for short-term or at -20°C for long-term storage.

2.9.3 Real time TaqMan® quantitative PCR

Quantitative PCR was performed on the Applied Biosystems 7900HT Fast Real-Time PCR system on the SDS v2.3 software using TaqMan® gene expression mastermix (Applied Biosystems) and commercially-available TaqMan® gene expression assays (Applied Biosystems). All samples were analysed in triplicate

wells in 384-well plates (4titude Ltd) for the gene(s) of interest (*HNF1A*, *CUBN, CLDN2, TM4SF4, TMEM27*) and a housekeeping gene (*PPIA*) (**Table 2.13**). To quantify expression of codon-optimised (recoded) *HNF1A* **(see Chapter 4 subsection 4.2.1 for details)**, I designed a custom Taqman® assay (probe/primer pair) using Primer Express™ Software v3.0.1 which is detailed in **Table 2.14**. Reactions were comprised of 5 µl TaqMan® Gene Expression Master Mix (Applied Biosystems), 2µl cDNA (1:10 dilution in 0.01M TRIS HCL), 0.5 µl assay (see **Table 2.14**) and 2.5 µl H2O to a total volume of 10 µl per well. Where multiplex reactions were carried out, a FAM- and a VIC-labelled assay were combined. Non-template control (NTC) wells containing H_2O were also included on each plate. All amplification curves were visually checked. Analysis was performed using the comparative Ct ($2^{-\Delta\Delta Ct}$) method on Microsoft Excel and graphs were plotted using *'ggplot2'* package in *R* v3.3.1.

Table 2.13. TaqMan® gene expression assays

Gene	**Probe ID**
CLDN2	Hs00252666_s1
CUBN	Hs00153607_m1
HNF1A	Hs00167041_m1
TMEM27	Hs00252907_m1
TM4SF4	Hs00270335_m1
PPIA	Hs04194521_s1

Reagents were obtained from Applied Biosystems.

Table 2.14 Custom Taqman® assay for recoded (codon-optimised) *HNF1A*

Forward Primer	CCACTGGTGACCGTGTCTACA
Reverse Primer	TTGGCTCCAGGCCAGTTG
Probe	CTCTGCACCAGGTCT

Probe/primer pair designed using Primer Express™ Software v3.0.1 (Applied Biosystems).

2.9.4 Protein analysis

(i) Cell lysis

Total protein lysates were collected from HepG2 cells transduced with PCW-*HNF1A*-puroR and exposed to doxycycline for 48 hrs (to induce expression of *HNF1A*) (**see chapter 3 subsection for more details**). The protocol was performed on ice to minimise protein degradation. First, cells (in a 6-well plate format) were washed in cold PBS (2ml per well). Two hundred µl of lysis buffer solution (10mL Roche Complete Lysis-M, 1 tablet Roche PhosStop, 1 tablet Roche Complete Protease Inhibitor; vortex to dissolve) was then added to each well. Cells were lysed for 10 min on an orbital shaker at room temperature, and then samples were transferred from the 6-well plate to 1.5 ml Eppendorf tubes. Lysates were centrifuged at 14,000 rpm for 20 min at 4°C (Allegra 64R, Beckman Coulter). The supernatant (post-nuclear fraction) was transferred into pre-chilled 1.5 ml microcentrifuge tubes and snap frozen on dry ice before long-term storage at -80°C.

(ii) Protein quantification

Cell lysates were quantified using the Bradford assay reagent (Bio-Rad Laboratories), which detects changes in absorbance of the Coomassie Brilliant Blue dye. When bound to protein, the blue form of the dye is stabilised and can be measured at 595nm, yielding a measure of protein concentration. Two microlitres of each sample were aliquoted into glass assay tubes (Fisher Scientific) containing 38 µl 0.1 M NaOH (Sigma Aldrich). Bradford reagent (1 ml of 1X) was made up in distilled H_2O, added to each tube, mixed by brief vortexing, and incubated for 10 min. Each sample was measured in duplicate on a 96-well

clear plate. Absorbance was read at 595nm on the Synergy H4 reader 1404211C software version 2.05.5. Protein concentrations were calculated based on a standard curve generated with bovine serum albumin (BSA, Sigma Aldrich).

(iii) Western blot analysis

Protein lysates were separated by sodium dodecyl sulphate polyacrylamide gel electrophoresis (SDS-PAGE) on a Bio-Rad Criterion TGX stain-free 4-12% pre-cast gel (1.0 mm, 18-well). The 1X Tris/Glycine/SDS (TGS) gel running buffer consists of 25 mM Tris, 192 mM glycine and 0.1% SDS pH 8.3 (Bio-Rad Laboratories).

Loading samples were prepared by adding 32.5 µl lysate to 5 µl 10X NuPAGE™ Sample Reducing Agent and 12.5 µl 6X NuPAGE™ Loading Buffer (both ThermoFisher Scientific). Samples were subject to mild heating at 70°C for 5 min in a thermocycler. A constant amount of protein (25-50 µg) was loaded in each lane for every gel run alongside a molecular weight marker (Precision Plus All Blue standard, Bio-Rad Laboratories). Gels were run at 200V for 1h.

Proteins were transferred to a 0.2 µm polyvinylidene difluoride (PVDF) membrane using the Trans-Blot Turbo transfer system (Bio-Rad Laboratories). Following transfer, the membrane was imaged to confirm protein transfer. Membranes were washed briefly in wash buffer made up of 1 mM Tris-HCl pH 7.4, 5 M NaCl (both Sigma Aldrich), 0.05% Tween-20 (Fisher Scientific) (TBST), then blocked with 5% non-fat milk powder in TBST for 1h before primary antibody incubation. Primary antibodies used for determining protein expression and recommended dilutions are shown in **Table 2.15**. Dilutions were made in 5% milk in TBST and added to membranes for incubation overnight at 4°C with gentle shaking (Stuart

see-saw rocker, Bibby Scientific Ltd). The following day, membranes were washed 3 times for 10 min each in TBST with agitation, followed by secondary antibody incubation for 1h with gentle shaking at room temperature. The secondary antibodies used and dilutions are shown in **Table 2.16**. After incubation membranes were washed 5 times for 5 min each in TBST with agitation. Membranes were then submerged in 2mL 1:1 SuperSignal West Pico Chemiluminescent Substrate (for HRP) for 3-5 minutes in the dark and then visualised on the Amersham™ Imager 600 (GE Healthcare).

Table 2.15. Primary antibodies used for western blot.

Antibody	Recommended dilutions	Source
Anti-HNF-1a (rabbit monoclonal)	1:1000	Cell Signalling Technology (D7Z2Q)
Anti-TM4SF4 (rabbit polyclonal)	1:500	Abcam ab84300
Anti-Cubilin (rabbit monoclonal to [EPR16259-27] antibody - C-terminal)	1:10,000	Abcam ab191073
Anti-Claudin-2 (rabbit polyclonal)	1:500	Abcam ab53032
Anti-TMEM27 (mouse polyclonal)	1:500	Abcam ab89058
Anti-Actin (goat monoclonal)	1:200	Santa Cruz sc-1615

All antibodies were stored at -20°C. Recommended dilutions were optimised.

Table 2.16. Secondary antibodies used for western blot.

Antibody	Recommended dilutions	Source
Anti-Rabbit IgG HRP	1:1000	R&D Systems HAF008
Anti-Mouse IgG HRP	1:1000	R&D Systems HAF018
Anti-Goat IgG HRP	1:5000	Abcam ab6741

All antibodies were stored at 4°C. Recommended dilutions by the manufacturer.

2.10 Flow cytometry

Several molecular signatures downstream HNF-1a were tested in flow cytometric analyses for robustness of *HNF1A* allelic segregation on the basis of emitted fluorescent signals which correspond to degree of molecular dysfunction (**see Chapter 4 for details**). Flow cytometric cellular analysis was performed on *in vitro* cellular models tested as high-throughput screening systems for *HNF1A* (HepG2 and HeLa).

Live cells were analysed on several different platforms:

- BD LSRFortessa™ cell analyser at the Flow Cytometry facility at the Kennedy Institute of Rheumatology, Oxford, UK (operated by technician, Jonathan Webber). All of the cytometry results in **Chapter 4 subsection 4.3.1** were generated using this platform, as well as a small number of confirmatory experiments in **Chapter 4 subsection 4.3.2**.

- BD Accuri C6™ cytometer at the Broad Institute of MIT and Harvard, Boston, MA (operated by technician, Robert Rice). This instrument was used for preliminary *TM4SF4* signal testing during my 3-month period as a visiting graduate student at the Broad Institute (June 2016-September 2016).

- SH800S Sony Cell Sorter at the Wellcome Centre for Human Genetics, Oxford, UK (operated by myself after induction and training). The majority of results shown in **Chapter 4 subsection 4.3.2**. were generated by me using this machine.

2.10.1 HeLa cells

Cells were harvested with TrypLE Select (Gibco 12563–011), resuspended in cold magnesium and calcium-free phosphate buffer saline (PBS) (Gibco cat no. 10010023) supplemented with supplemented with 2% FCS and 5mM Ethylenediaminetetraacetic acid (EDTA) to a final cell concentration of 2-5x10^6 cells per ml. Samples were filtered using BD Falcon® 100 µm filter caps (Corning Product #352360) into either sterile 1.5 ml microcentrifuge tubes or Falcon® 5mL Round Bottom Polystyrene Test Tube (Product #352235), depending scale and analysis platform. Data were analysed with FlowJo (version 10.0.8).

2.10.2 HepG2 cells

After a 7-day dox dose induction, cells from 6 well plates were harvested with TrypLE Select (Gibco 12563–011) and washed twice with cold magnesium and calcium-free PBS (Gibco cat no. 10010023), supplemented with 0.5% Bovine Serum Albumin (BSA) (Sigma) and 2mM EDTA, first with 5 mls and then with 2mls, spinning at 400xg for 5 minutes in 4°C. After the second wash, cells are resuspended in 100 ul cold Ca++ and Mg++ free PBS supplemented with 3% BSA and 2 mM EDTA (staining buffer) and stained with 10-12ul of Phycoerythrin (PE)-conjugated mouse monoclonal anti-human TM4SF4 (anti-human TM4SF4-PE; R&D Systems cat no. FAB7998P) for 30-40 min the dark at room temperature. Stained samples are then diluted by adding wash buffer up to 1 ml per sample and transferring to tubes containing 4 mls of additional wash buffer. Following a spin at 400xg for 5 minutes in 4°C, the stained cells are washed twice. Between 0.5-2x10^6 cells are finally resuspended in 200ul wash buffer and filtered using BD Falcon® 40 µm filter caps (Corning Product #352340) into either sterile

1.5 ml microcentrifuge tubes or Falcon® 5mL Round Bottom Polystyrene Test Tube (Product #352235), depending scale and analysis platform. Data were analysed with FlowJo (version 10.0.8).

An identical workflow was used to prepare samples assayed for *CLDN2* fluorescent signal with minor adjustments for primary and secondary staining. After trypsinisation and the first two washes, samples were stained with 10 ug/ml anti-Claudin 2 antibody (Rabbit polyclonal to Claudin-2; Abcam, ab53032), as recommended by the manufacturer, for 30 minutes in the dark at room temperature. A secondary stain was performed following two washes by resuspending cells in a 100ul solution of Goat Anti-Rabbit IgG H&L (Alexa Fluor® 488) (Abcam, ab150077).

2.11 Production of recombinant human glucokinase protein

The restriction sites BamHI and EcoRI were used to clone human pancreatic beta-call glucokinase cDNA sequence NM_000162.3 into pGEX-3X vector (Sigma Aldrich Company Ltd) (**see Chapter 5 Figure 5.2 for plasmid map**). Recombinant human GCK in the form of a fusion protein with Glutathione S-Transferase (GST) was purified following heterologous expression in BL21 supercompetent E.coli (Stratgene Europe) and affinity chromatography through a glutathione-agarose column (Sigma).

More specifically, a single transformed BL21 colony (cultured on agar-ampicillin) was used to inoculate 5ml LB-ampicillin medium. This was cultured overnight at 37°C, shaking at 250 rpm. The next day, the 5ml starter culture was added to 500ml LB-ampicillin medium, which was in turn cultured overnight (37°C, 250 rpm). Protein production was induced by adding 0.2M isopropyl β-D-1-

thiogalactopyranoside (IPTG) once the culture yielded an optical density (OD) of 0.6 at 600nm (indication that cells have reached exponential growth). IPTG is a structural mimic of lactose which binds to the lac repressor sequence (lacI) making the lac operon promoter site accessible to T7 RNA polymerase upstream human GCK cDNA, thereby enabling transcription (Daber et al, 2007). Following addition of IPTG, the culture was incubated overnight at 22°C (high yield GCK protein induction temperature), shaking at 150 rpm. IPTG induced cells were pelleted via centrifugation at 3500g, 4°C, for 10 minutes, after which the pellet was resuspended in ice-cold PBS-DTT (Sigma) and lysed via sonication (VibraCell sonicator, amplitude 20, 40% max power). Ten percent Triton-X (Sigma) was then added to the lysed cells, which were mixed gently by inversion and centrifuged at 5000g, 4°C, for 8 minutes to pellet cell debris and retain cytoplasmic proteins in suspension. Glutathione agarose dissolved in PBS-DTT was then added to the cytoplasmic-protein supernatant, which was then mixed gently by inversion for 2 minutes to facilitate GST-tag-glutathione binding.

Purification of GCK protein from the solution was conducted in a 4°C cold room where it was poured into a PolyPrep chromatography column (Bio-Rad Laboratories Ltd). This allowed for slow elution of buffer/undesired cellular proteins whilst retaining recombinant glucokinase bound to glutathione-agarose in the column. Upon adequate washing of column contents, bound recombinant GCK was eluted by the addition of glutathione elution buffer (**Table 2.17**). Eluent fractions were collected in 1.5 ml Eppendorf tubes on ice. Protein yield in each eluent fraction was determined by measuring absorbance at 595nm using 5-fold-diluted Bradford reagent (Bio-Rad). The fractions with the highest protein yield

were pooled and storage buffer (**Table 2.17**) was added to the pooled eluents in a 2:3 ratio before snap-freezing on dry ice and storage at -80°C.

Table 2.17. Components of buffers used in recombinant GCK production.

Buffers used in recombinant protein production	Recipe
Glutathione elution buffer (solution mixed by vortexing and stored at 4°C in the dark)	6 µM glutathione (GE Healthcare Biosciences) 0.8M potassium chloride (KCl) 0.2M pH 6.8 Tris HCl ([hydroxymethyl] aminomethane-hydrochloric acid) 0.02M DTT (Dithiothreitol)
GCK storage buffer (before addition to purified and eluted GCK, a working stock of 10ml with 5mM DTT was made)	5.5mM glucose 0.25M KCl 0.06M pH 8.0 Tris-HCl 60ml glycerol

2.12 Glucokinase glucose- and ATP-dependent assays

GCK activity was measured spectrophotometrically as a rate of NADPH formation glucose 6-phosphate dehydrogenase-coupled assays. To interrogate wild-type and allele-harbouring GCK affinity for glucose, the following reaction mix was prepared on ice: 7.5mM $MgCl_2$, 187.5mM KCl, 125mM HEPES buffer (4-[2-hydroxyethyl]-1-piperazineethanesulfonic acid), 1.25mM $NADP^+$, 6.25mM ATP, 2.5mM DTT, 0.125% BSA (bovine serum albumin), and 6.45ml distilled water. Sixty six units of glucose-6-phosphate dehydrogenase (Sigma) were added to the reaction mix upon confirming a pH level of 7.4. The reaction mix was used to make 1ml of 0.2M glucose solution (Sigma) before the remainder was diluted using 3.5ml distilled water. The diluted mix was used to perform a serial dilution of the 0.2M glucose stock solution to yield the following concentration range (mM): 100, 50, 25, 12.5, 6.25, 3.31, 1.56, 0.78, 0.39, 0.20, 0.10, 0. GCK, at a concentration of 10 mol/L, was then added to 8 glass test tubes containing 1ml diluted reaction mix, which was then mixed by vortexing. Two 96-well

spectrophotometer plates were then prepared as follows: one plate contained the diluted glucose samples and the other contained the GCK enzyme dilution to be tested. The contents of the latter were transferred to the corresponding wells of the former, and the reaction rate was measured using Molecular Devices M3 plate reader running SoftMax Pro (Software version 2.05.5), which recorder the rate of NADPH formation based on absorbance at 340nm. Kaleidagraph v3.5 (Synergy Software) was used to plot assay data fitted to the Hill equation (Hill 1910; Kesavan et al, 1997), which enabled accurate quantification of the half maximal glucose concentration ($S_{0.5}$) and Hill number (n_H).

(2.1) $$v = (S^{nH} \times V_{max}) / (S^{nH} \times S_{0.5}{}^{nH})$$

Equation 2.1. The Hill equation (Hill 1910; Kesavan et al, 1997). 'v' is reaction rate, 'S' represents glucose concentration, and '$S_{0.5}$' the half maximal glucose concentration. To fit the assay reaction data (v) over the tested glucose concentration range (S), Kaleidagraph was provided with prior estimates (based on wild-type GCK activity *in vitro*) of each of the equation parameters: V_{max} =10; $S_{0.5}$= 7.5; n_H= 1.5.

The maximal reaction rate (V_{max}) and total enzyme concentration (E_t) were used to calculate the overall turnover rate of the enzyme (K_{cat}; equation 2.4), which was derived from a rearrangement of the Michaelis-Menton equation (equation 2.2) (Michaelis and Menton, 1913).

(2.2) $$v = (V_{max} \times S) / (K_m + S)$$

(2.3) $$V_{max} = K_{cat} \times E_t$$

(2.4) $$K_{cat} = [V_{max} \times (K_m + S)] / (E_t \times S)$$

Equations 2.2, 2.3 and 2.4. Michaelis-Menton equation (Michaelis and Menton, 1913) is shown in 2.2, where K_m corresponds to glucose $S_{0.5}$ and ATP_{km} in the glucose- and ATP- dependent assays, respectively.

A similar procedure was followed to evaluate GCK affinity for ATP (second substrate). The reaction mix was prepared on ice and was identical to that prepared for glucose-dependent assays with the exception of 18.75mM glucose and no ATP. The reaction mix was used to make 100mM ATP-$MgCl_2$ solution, which was serially diluted to yield the following ATP-$MgCl_2$ concentration range (mM): 5, 4, 3, 2.5, 2, 1.5, 1, 0.75, 0.5, 0.25, 0.1, 0. One millilitre of reaction mix was then added to 8 glass test tubes on ice, to which then 10mol/L of GCK enzyme were also added. Two 96-well spectrophotometer plates were then prepared as follows: one plate contained the diluted ATP-$MgCl_2$ samples and the other contained the GCK enzyme dilution to be tested. A multi-channel pipette was then used to transfer the contents from the latter to the former, and reaction rate was recorded as previously described. The Michaelis-Menton equation (equation 2.2) was used to fit assay data and determine the half maximal ATP concentration (ATP_{km}) for each tested enzyme.

2.13 *HNF1A* and *GCK* cDNA sequences

Table 2.18. *HNF1A* cDNA sequences used in chapters 3 and 4

HNF1A reference cDNA template (27Leu) NM_000545.6 (liver isoform)	ATGGTTTCTAAACTGAGCCAGCTGCAGACGGAGCTCCTGGCGGC CCTGCTGGAGTCAGGGCTGAGCAAAGAGGCACTGCTCCAGGCAC TGGGTGAGCCGGGGCCCTACCTCCTGGCTGGAGAAGGCCCCCT GGACAAGGGGGAGTCCTGCGGCGGCGGTCGAGGGGAGCTGGCT GAGCTGCCCAATGGGCTGGGGGAGACTCGGGGCTCCGAGGACG AGACGGACGACGATGGGGAAGACTTCACGCCACCCATCCTCAAA GAGCTGGAGAACCTCAGCCCTGAGGAGGCGGCCCACCAGAAAG CCGTGGTGGAGACCCTTCTGCAGGAGGACCCGTGGCGTGTGGC GAAGATGGTCAAGTCCTACCTGCAGCAGCACAACATCCCACAGC GGGAGGTGGTCGATACCACTGGCCTCAACCAGTCCCACCTGTCC CAACACCTCAACAAGGGCACTCCCATGAAGACGCAGAAGCGGGC CGCCCTGTACACCTGGTACGTCCGCAAGCAGCGAGAGGTGGCG CAGCAGTTCACCCATGCAGGGCAGGGAGGGCTGATTGAAGAGCC CACAGGTGATGAGCTACCAACCAAGAAGGGGCGGAGGAACCGTT TCAAGTGGGGCCCAGCATCCCAGCAGATCCTGTTCCAGGCCTAT GAGAGGCAGAAGAACCCTAGCAAGGAGGAGCGAGAGACGCTAG TGGAGGAGTGCAATAGGGCGGAATGCATCCAGAGAGGGGTGTCC CCATCACAGGCACAGGGGCTGGGCTCCAACCTCGTCACGGAGGT GCGTGTCTACAACTGGTTTGCCAACCGGCGCAAAGAAGAAGCCT TCCGGCACAAGCTGGCCATGGACACGTACAGCGGGCCCCCCCC AGGGCCAGGCCCGGGACCTGCGCTGCCCGCTCACAGCTCCCCT GGCCTGCCTCCACCTGCCCTCTCCCCCAGTAAGGTCCACGGTGT

	GCGCTATGGACAGCCTGCGACCAGTGAGACTGCAGAAGTACCCT CAAGCAGCGGCGGTCCCTTAGTGACAGTGTCTACACCCCTCCAC CAAGTGTCCCCCACGGGCCTGGAGCCCAGCCACAGCCTGCTGA GTACAGAAGCCAAGCTGGTCTCAGCAGCTGGGGGCCCCCTCCCC CCTGTCAGCACCCTGACAGCACTGCACAGCTTGGAGCAGACATC CCCAGGCCTCAACCAGCAGCCCCAGAACCTCATCATGGCCTCAC TTCCTGGGGTCATGACCATCGGGCCTGGTGAGCCTGCCTCCCTG GGTCCTACGTTCACCAACACAGGTGCCTCCACCCTGGTCATCGG CCTGGCCTCCACGCAGGCACAGAGTGTGCCGGTCATCAACAGCA TGGGCAGCAGCCTGACCACCCTGCAGCCCGTCCAGTTCTCCCAG CCGCTGCACCCCTCCTACCAGCAGCCGCTCATGCCACCTGTGCA GAGCCATGTGACCCAGAGCCCCTTCATGGCCACCATGGCTCAGC TGCAGAGCCCCCACGCCCTCTACAGCCACAAGCCCGAGGTGGCC CAGTACACCCACACGGGCCTGCTCCCGCAGACTATGCTCATCAC CGACACCACCAACCTGAGCGCCCTGGCCAGCCTCACGCCCACCA AGCAGGTCTTCACCTCAGACACTGAGGCCTCCAGTGAGTCCGGG CTTCACACGCCGGCATCTCAGGCCACCACCCTCCACGTCCCCAG CCAGGACCCTGCCGGCATCCAGCACCTGCAGCCGGCCCACCGG CTCAGCGCCAGCCCCACAGTGTCCTCCAGCAGCCTGGTGCTGTA CCAGAGCTCAGACTCCAGCAATGGCCAGAGCCACCTGCTGCCAT CCAACCACAGCGTCATCGAGACCTTCATCTCCACCCAGATGGCCT CTTCCTCCCAGTAA
HNF1A recoded cDNA template (27Leu)	ATGGTGTCAAAACTGTCACAGCTGCAGACTGAACTGCTGGCTGCT CTGCTGGAAAGCGGACTGAGTAAGGAGGCTCTGATCCAGGCTCT GGGAGAACCAGGACCATATCTGCTGGCAGGAGAAGGCCCTCTGG ATAAGGGAGAGTCTTGCGGCGGAGGAAGAGGAGAGCTGGCAGA ACTGCCCAATGGCCTGGGAGAGACCCGCGGGTCCGAGGACGAA ACTGACGACGATGGAGAAGACTTTACACCACCCATTCTGAAGGAA CTGGAAAACCTGTCACCTGAAGAGGCCGCTCATCAGAAGGCTGT CGTCGAAACACTGCTGCAGGAGGACCCATGGCGAGTGGCTAAGA TGGTCAAGTCATATCTGCAGCAGCATAACATTCCACAGAGAGAAG TCGTCGATACCACCGGACTGAACCAGAGCCATCTGTCACAGCATC TGAATAAGGGCACACCAATGAAGACTCAGAAACGCGCTGCCCTG TACACTTGGTACGTCAGAAAACAGAGGGAAGTCGCTCAGCAGTTT ACTCATGCTGGCCAGGGAGGGCTGATTGAGGAACCAACCGGAGA TGAACTGCCAACTAAGAAAGGAAGGCGAAATAGGTTTAAATGGGG ACCAGCCTCACAGCAGATTCTGTTCCAGGCTTATGAACGCCAGAA AAACCCTAGCAAGGAAGAGAGAGAGACCCTGGTGGAAGAGTGTA ATCGCGCCGAGTGTATCCAGCGCGGCGTGAGTCCTAGCCAGGCC CAGGGACTGGGCTCAAATCTGGTCACAGAGGTGCGAGTGTACAA TTGGTTTGCAAATCGCAGGAAAGAGGAGGCATTCCGGCACAAGC TGGCAATGGATACATACTCCGGGCCTCCACCTGGCCCTGGACCA GGACCTGCACTGCCCGCTCACAGTTCTCCAGGACTGCCTCCACC TGCACTGTCTCCTAGCAAAGTCCATGGAGTGAGATACGGCCAGC CTGCAACCTCCGAAACTGCTGAGGTGCCATCCAGCTCTGGAGGA CCACTGGTGACCGTGTCTACACCTCTGCACCAGGTCTCTCCAACT GGCCTGGAGCCAAGTCACAGTCTGCTGAGTACCGAAGCAAAGCT GGTCTCTGCTGCCGGCGGCCCTCTGCCACCCGTGTCAACACTGA CCGCTCTGCATTCCCTGGAACAGACCAGTCCCGGGCTGAATCAG CAGCCTCAGAACCTGATCATGGCCAGCCTGCCTGGCGTGATGAC AATTGGCCCTGGCGAGCCCGCTTCCCTGGGCCCAACATTCACAA ATACTGGCGCTTCTACACTGGTCATCGGCCTGGCTTCAACTCAGG CCCAGTCCGTGCCTGTGATTAACAGTATGGGATCTTCACTGACCA CTCTGCAGCCCGTCCAGTTCTCCCAGCCACTGCATCCCTCATATC AGCAGCCACTGATGCCACCAGTCCAGTCACATGTGACCCAGTCT CCTTTCATGGCAACCATGGCTCAGCTGCAGTCACCTCATGCTCTG TACAGCCACAAGCCTGAGGTGGCACAGTACACTCATACCGGGCT GCTGCCACAGACAATGCTGATTACTGATACCACTAATCTGAGCGC CCTGGCTAGTCTGACACCAACCAAGCAGGTGTTTACCTCAGACAC CGAGGCATCTTCCGAGTCAGGACTGCACACTCCCGCCTCCCAGG

	CAACCACACTGCACGTGCCTTCTCAGGACCCAGCTGGCATCCAG CATCTGCAGCCTGCCCACCGGCTGAGCGCATCTCCCACCGTGTC ATCCTCTAGTCTGGTGCTGTACCAGTCAAGCGACTCCTCAAACGG ACAGTCCCATCTGCTGCCTAGCAATCATAGTGTCATTGAGACTTTT ATTTCTACTCAGATGGCATCATCATCACAGTAG

Table 2.19. *GCK* cDNA sequence used in chapter 5	
GCK reference cDNA template NM_000162.3 hGCK (beta cell isoform)	ATGCTGGACGACAGAGCCAGGATGGAGGCCGCCAAGAAGGAGA AGGTAGAGCAGATCCTGGCAGAGTTCCAGCTGCAGGAGGAGGAC CTGAAGAAGGTGATGAGACGGATGCAGAAGGAGATGGACCGCG GCCTGAGGCTGGAGACCCATGAAGAGGCCAGTGTGAAGATGCTG CCCACCTACGTGCGCTCCACCCCAGAAGGCTCAGAAGTCGGGGA CTTCCTCTCCCTGGACCTGGGTGGCACTAACTTCAGGGTGATGCT GGTGAAGGTGGGAGAAGGTGAGGAGGGGCAGTGGAGCGTGAAG ACCAAACACCAGATGTACTCCATCCCCGAGGACGCCATGACCGG CACTGCTGAGATGCTCTTCGACTACATCTCTGAGTGCATCTCCGA CTTCCTGGACAAGCATCAGATGAAACACAAGAAGCTGCCCCTGG GCTTCACCTTCTCCTTTCCTGTGAGGCACGAAGACATCGATAAGG GCATCCTTCTCAACTGGACCAAGGGCTTCAAGGCCTCAGGAGCA GAAGGGAACAATGTCGTGGGGCTTCTGCGAGACGCTATCAAACG GAGAGGGGACTTTGAAATGGATGTGGTGGCAATGGTGAATGACA CGGTGGCCACGATGATCTCCTGCTACTACGAAGACCATCAGTGC GAGGTCGGCATGATCGTGGGCACGGGCTGCAATGCCTGCTACAT GGAGGAGATGCAGAATGTGGAGCTGGTGGAGGGGGACGAGGGC CGCATGTGCGTCAATACCGAGTGGGGCGCCTTCGGGGACTCCG GCGAGCTGGACGAGTTCCTGCTGGAGTATGACCGCCTGGTGGAC GAGAGCTCTGCAAACCCCGGTCAGCAGCTGTATGAGAAGCTCAT AGGTGGCAAGTACATGGGCGAGCTGGTGCGGCTTGTGCTGCTCA GGCTCGTGGACGAAAACCTGCTCTTCCACGGGGAGGCCTCCGAG CAGCTGCGCACACGCGGAGCCTTCGAGACGCGCTTCGTGTCGCA GGTGGAGAGCGACACGGGCGACCGCAAGCAGATCTACAACATCC TGAGCACGCTGGGGCTGCGACCCTCGACCACCGACTGCGACATC GTGCGCCGCGCCTGCGAGAGCGTGTCTACGCGCGCTGCGCACA TGTGCTCGGCGGGGCTGGCGGGCGTCATCAACCGCATGCGCGA GAGCCGCAGCGAGGACGTAATGCGCATCACTGTGGGCGTGGAT GGCTCCGTGTACAAGCTGCACCCCAGCTTCAAGGAGCGGTTCCA TGCCAGCGTGCGCAGGCTGACGCCCAGCTGCGAGATCACCTTCA TCGAGTCGGAGGAGGGCAGTGGCCGGGGCGCGGCCCTGGTCTC GGCGGTGGCCTGTAAGAAGGCCTGTATGCTGGGCCAGTGA

3

Unsupervised clustering of exome-detected *HNF1A* alleles using multi-dimensional functional data

3.1 Background

The hepatocyte nuclear factor 1-alpha gene (*HNF1A*) is roughly 23Kb in length, comprised of 10 exons and encodes a 631 amino-acid atypical homeodomain-containing transcription factor (HNF-1a). Though initially considered a hepatocyte-specific transcriptional regulator, *HNF1A* is also expressed in the kidney, intestine and pancreas. It is a key member of an autoregulatory transcription factor circuit which plays a critical role in pancreas maturation and differentiation (Boj et al, 2001).

HNF-1a consists of 3 functional domains: 1) a myosin-like dimerization domain at the N(amino)-terminus (amino acids 1-33), 2) a bipartite (POU-homeodomain) DNA binding domain (amino acids 100-184 and 198-281), and 3) a tripartite transactivation region at the C(carboxyl)-terminus (amino acid 282-631) (Vaxillaire et al 1999). It is also associated with temporal and spatial variation in isoform expression (Bach and Yaniv 1993; Harries et al, 2006).

HNF1A produces three alternatively processed transcripts [*HNF1A*(A), *HNF1A*(B), *HNF1A*(C)], which are identical at the 5' end, and differ in length as a consequence of differential splicing and polyadenylation (Bach and Yaniv 1993). *HNF1A*(A) is the longest transcript (exons 1-10) and is the predominant one in the hepatocyte; whereas *HNF1A(*B) and *HNF1A*(C) terminate at exon 6 and 7, respectively, and represent the predominant isoforms in kidney and adult pancreas (Bach and Yaniv 1993; Harries et al, 2006). In the fetal pancreas, *HNF1A*(A) is the abundant transcript (Bach and Yaniv 1993; Harries et al, 2006).

3.1.1 HNF-1a *in vitro* function: stratification of alleles along the pathophysiological continuum

HNF1A is a clinically actionable gene where accurate variant interpretation has tangible impact on prognostic evaluation, treatment and management. Individuals with a deleterious *HNF1A* allele and young-onset diabetes are sensitive to treatment with sulphonylurea (oral hypoglycaemic agent) and can often avoid insulin injections until late in life (Yamagata et al, 1996; Shepherd et al, 2001; Stride et al, 2002; Pearson et al, 2000). Lifestyle interventions can be used preventatively by individuals in whom an *HNF1A* variant known to substantively influence T2D risk is identified early in life (Pearson et al, 2003).

Large-scale multi-ethnic genetic sequencing studies have been instrumental in resolving the complexity of genotype-phenotype relationships in the context of Mendelian disease genes such as *HNF1A* (Flannick et al, 2013; Alkuraya 2015; Chen et al, 2016). The understanding of *HNF1A* variant manifestation was traditionally limited to the notion that heterozygous highly penetrant loss-of-function alleles in *HNF1A* give rise to a clinically distinct diabetes subtype characterised by non-obesity, young age of onset (<25y), and dominant inheritance (Maturity onset diabetes of the young, MODY) (Stride et al, 2002; Murphy et al, 2008). While this genotype-phenotype correlation is true for a subset of *HNF1A* allele carriers, it represents one end of a broad spectrum of *HNF1A* variant effects. Genome-wide association and next-generation sequencing studies of randomly ascertained individuals have challenged binary assumptions and overinflated pathogenicity estimates regarding variants in *HNF1A* (and other Mendelian disease genes) (Flannick et al, 2013).

Such efforts have shown that, in the general population, bona fide loss-of-function alleles in *HNF1A* exhibit incomplete penetrance (Flannick et al, 2013). Whole-exome sequencing in T2D cases and controls of Mexican and Mexican American ancestry revealed a low frequency missense allele in *HNF1A* p.E508K associated with a five-fold increase in T2D prevalence (Consortium STD, 2014). Another population-specific *HNF1A* missense allele p.G391S was discovered in Canadian Oji-Cree and was also show to elevate T2D risk (Hegele et al, 1999; Triggs-Raine et al, 2002). Common missense alleles in *HNF1A* of low effect have also been associated with T2D predisposition in populations of European ancestry (p.A98V and p.I27L) (Gaulton et al, 2015) (**see Chapter 1 section 1.4.1 for more detail**).

The identification of *HNF1A* alleles along the clinical phenotypic spectrum is no longer a formidable task. The challenge today is in the ability to map *HNF1A* sequence-function relationships at high fidelity, using molecular characterisation and analytical pipelines with sensitivity to capture the subtleties along the pathophysiological continuum. Rigorous functional follow-up of sequence-identified alleles in *HNF1A* is crucial to making correct assignments of pathogenicity. Indeed, functional data are considered a strong line of evidence for accurate clinical diagnostic classification of variants (Richards et al, 2015).

Gold standard molecular characterisation pipelines for studying the function of transcription factors consist primarily of *in vitro* cellular assays which interrogate transcriptional activity, protein abundance, localisation of the transcription factor to the nucleus, and binding to relevant DNA recognition sequences. The experimental techniques used to interrogate these mechanisms *in vitro* vary in robustness and reliability hence the wide variability of reported functional consequences of *HNF1A* variants in the literature. Functional and clinical

phenotypic severity are also influenced by allele position in *HNF1A*. The transactivation domain in HNF-1a is more tolerant to genetic variation (<25% alleles associated with MODY) compared to the DNA binding and dimerization domains (>60%, >45-55%, respectively) (Consortium STD, 2014; Flannick et al, 2016). With respect to isoform variation, Harries and colleagues showed that diabetes was diagnosed later in life (25.5y) when individuals carried a damaging *HNF1A* allele expressed exclusively in the hepatocyte-dominant isoform *HNF1A*(A) compared to one expressed in all three isoforms (exons 1-7) (18y) (Harries et al, 2006; Bellanne-Chantelot et al, 2008).

3.1.2 Large-scale collaborative functional follow up of *HNF1A* alleles

A large-scale exome sequencing study was conducted on 12,940 T2D cases and controls from five different ancestry groups with the aim of understanding the impact of low frequency (0.5% < MAF < 5%) and rare (MAF < 0.5%) protein coding alleles on T2D susceptibility (GoT2D/T2D-GENES consortia; Fuchsberger et al, 2016) **(Figure 3.1A)**. A total of 84 rare protein-altering variants in *HNF1A* (i.e. missense, protein-truncating) were identified in the study at various frequencies in cases and controls (some present exclusively in either cases or controls). The alleles did not on aggregate enrich for a T2D phenotype under the several filters used in the study (rare MAF <0.1%, conserved and predicted damaging [PolyPhen: SKAT p=0.30, BURDEN p=0.37]) (Fuchsberger et al, 2016).

To understand and characterise the relationship between *HNF1A* sequence variation, molecular dysfunction and clinical phenotype, a total of 73 *HNF1A* missense alleles detected in the exomes of ~13K multi-ethnic T2D cases and

controls were chosen for functional follow-up by research groups at Oxford (Prof Anna Gloyn and Prof Mark McCarthy) and Bergen (Prof Pal Njølstad) **(Figure 3.1)** (Fuchsberger et al 2016). The variants were divided between teams at Oxford and Bergen such that each was responsible for molecular characterisation of a unique subset of exome-detected alleles (n >30), a shared subset of exome-detected alleles (n=5), shared exome-detected known T2D risk control alleles (n=2) and a shared subset of MODY references alleles (n=6) **(Figure 3.2)**.

The teams at Oxford and Bergen adopted an *in vitro* molecular characterisation pipeline that interrogates the function of *HNF1A* alleles from multiple mechanistic angles using gold standard functional assays: transactivation (luciferase reporter gene assays) in multiple cell lines, DNA binding (electrophoretic mobility shift EMSA assays), protein expression (western blot), and nuclear localisation (cell fractionation followed by western blot) **(Figure 3.1B).** The Oxford variant set was functionally characterised by Dr. Amanda Bennett and Dr. Jana Rundle, and the Bergen variant set by Laeya Najmi, Dr. Lise Bjørkhaug and Dr. Ingvild Aukrust, per the pipeline depicted in **Figure 3.1B**.

To artificially express *HNF1A in vitro*, *HNF1A* cDNA was cloned into an appropriate expression plasmid **(see subsection 3.2.1)**, and the alleles were introduced to the sequence via site-directed mutagenesis. The cell systems were transiently transfected with *HNF1A* cDNA sequences and assayed 48 hours post-transfection for functional performance. By direct introduction of cDNA mRNA surveillance mechanisms are evaded and thus reduced protein expression in these investigations can be primarily interpreted as reduction in protein stability.

Dr. Amanda Bennett and Dr. Jana Rundle (Oxford) and Laeya Najmi, Dr. Lise Bjørkhaug and Dr. Ingvild Aukrust (Bergen) interrogated transcriptional activity in two different cell systems each with a unique promoter architecture: 1) Rat insulin secreting beta cell derived line INS-1 (which endogenously expresses *HNF1A*), and HNF-4a-P2 promoter, for which *HNF1A* has a low binding affinity, and 2) Human cervical cancer HeLa cell line (in which *HNF1A* is not endogenously expressed), and albumin promoter (rat), for which *HNF1A* has highest consensus binding affinity. They evaluated HNF-1a protein abundance and nuclear translocation in HeLa cells. Additionally, they conducted DNA binding assays for the subset of alleles which exhibited impaired transcriptional activity and were located between amino acids 1-280 (dimerization and DNA binding domains). EMSAs were performed by measuring the level of binding of HNF-1a proteins (exogenously expressed in and extracted from HeLa cells) to fluorescently labelled rat albumin oligonucleotide.

3.1.3 Chapter preview

The *HNF1A* functional dataset generated by the research teams at the Universities of Oxford and Bergen represents the largest multi-parametric *in vitro* functional annotation effort to date. From an analytical and data interpretation standpoint, this presented some key challenges. It was unclear how best to process a multidimensional functional dataset of this scale (n=79 functionally characterised alleles) generated by multiple research groups using slightly variable experimental protocols for each assay. It was also unclear whether all parameters should be weighted equally (i.e. an integrative analysis using data from multiple functional dimensions) or whether the standard approach of evaluating each functional parameter independently would be best.

In this chapter, I address these challenges as I describe my efforts towards building and validating an analytical strategy to map multi-dimensional *in vitro* function to *HNF1A* glycaemic phenotypes *in vivo*. My hypothesis was that severity of molecular dysfunction *in vitro* (WT/WT-like, moderate/intermediate impact, loss-of-function/deleterious) would correlate positively with the severity of clinical phenotype (benign, increased T2D risk, young-onset SU responsive hyperglycaemia).

The pipeline I developed exploits unsupervised learning techniques which enable informative exploration of the structure of large, complex datasets and the identification of subgroups within them. Using principal component analysis (PCA) for dimension reduction and hierarchical clustering for useful visualisation of structural patterns in multivariate space, I generated allele clusters which I was able to annotate broadly across the *HNF1A* phenotypic continuum (benign, T2D risk-modifying, and MODY-casual). To demonstrate the clinical utility of unsupervised functional stratification of *HNF1A* alleles, I used the data to re-evaluate molecular diagnoses in the UK MODY diagnostic registry at Exeter. This effort resulted in reclassification of n=7 *HNF1A* alleles from pathogenic/likely pathogenic to VUS (variant of uncertain clinical significance)/likely benign.

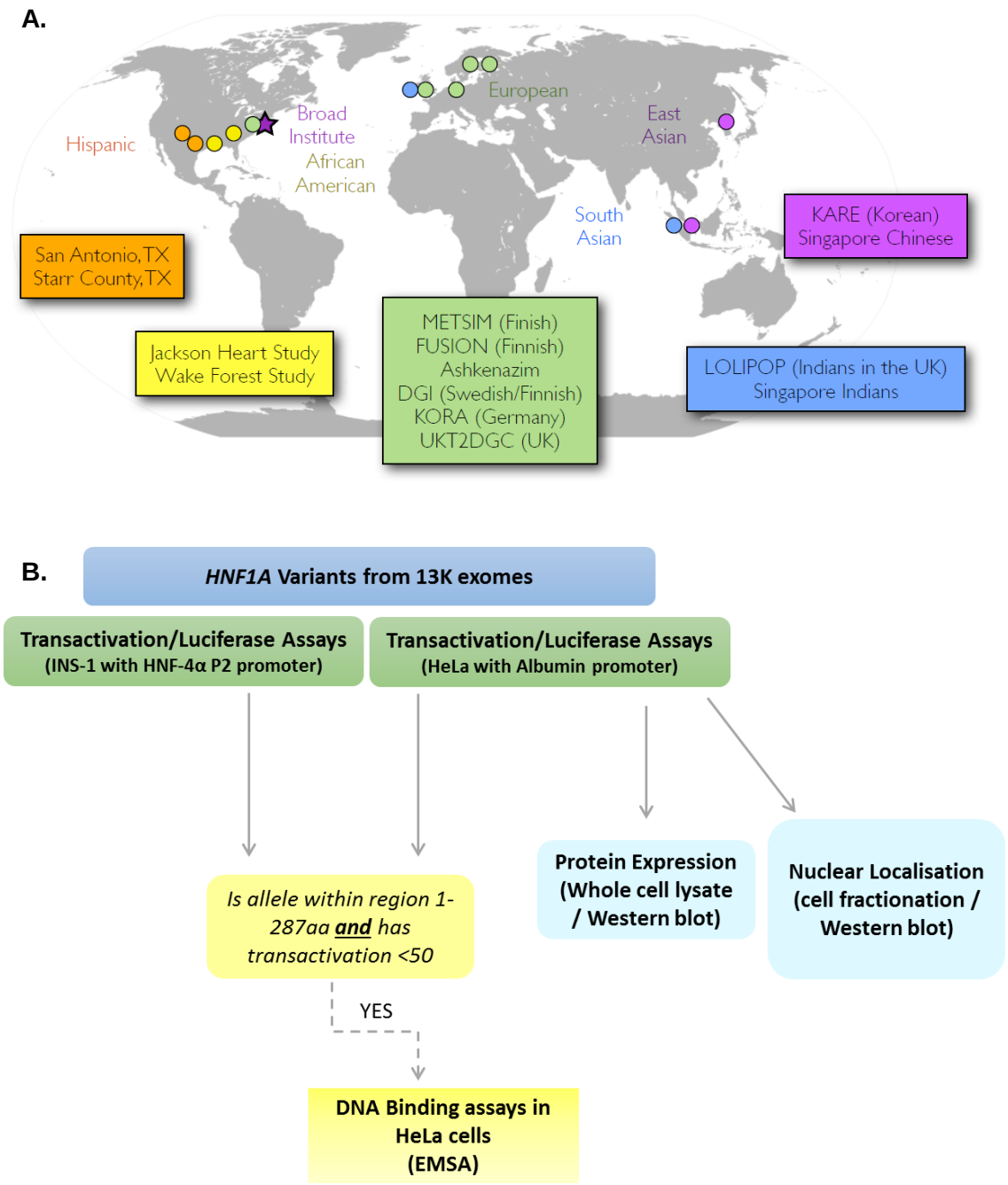


Figure 3.1. Ethnic composition of a multi-consortia led exome sequencing study of n=12,940 T2D cases and controls (A). Molecular characterisation pipeline adopted by research groups at Oxford and Bergen to annotate the function of exome-detected *HNF1A* missense alleles (B). Exome sequence data was generated from ~13K individuals (6,504 T2D cases and 6,436 controls) from five ancestry groups with 82x mean coverage across the protein coding sequence of 18,281 genes, identifying 3.04 million alleles (1.19 million protein-altering) (Fuchsberger et al, 2016). Figures courtesy of Prof Mark McCarthy (A) and Prof Anna L Gloyn (B; modified with permission).

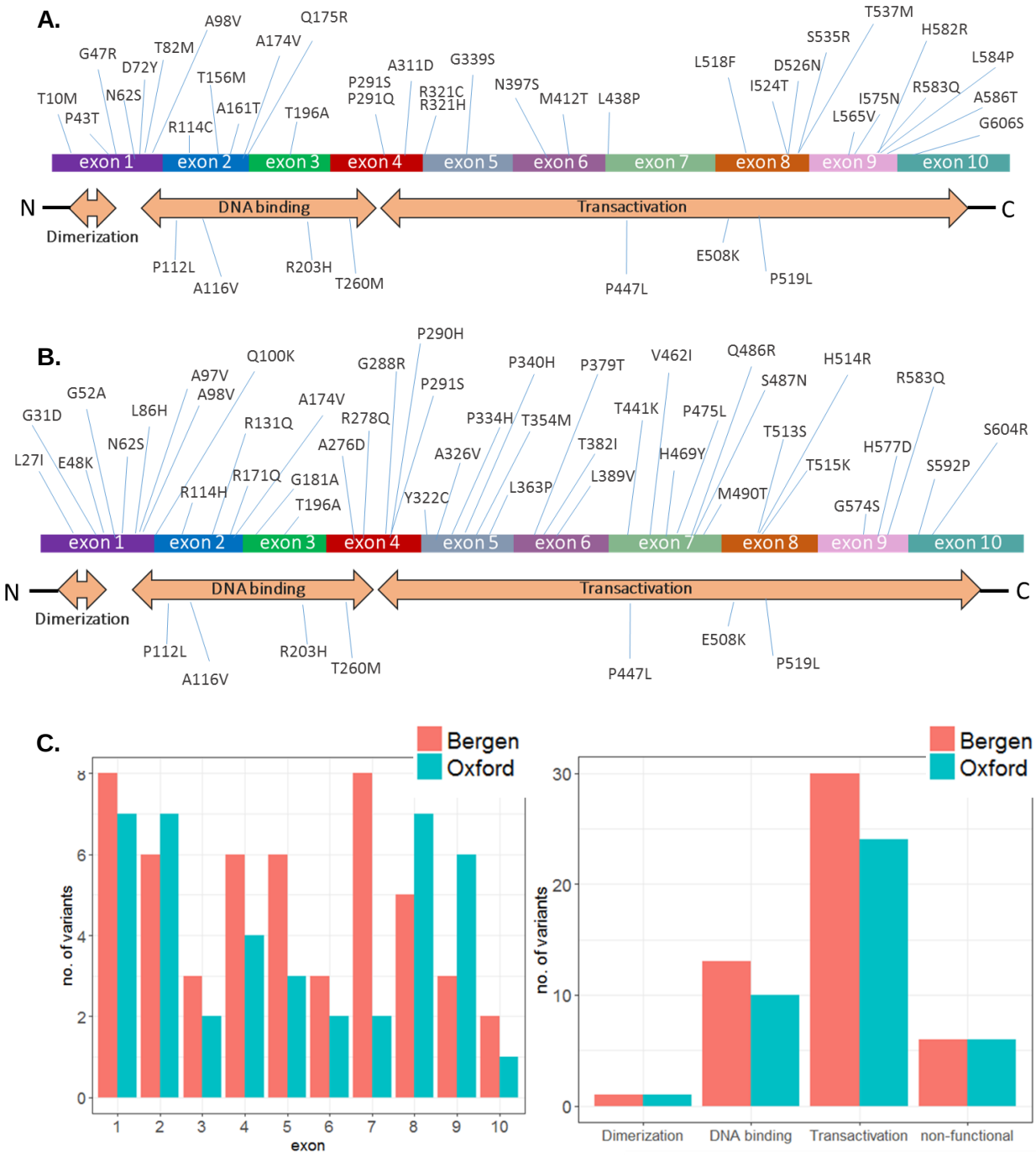


Figure 3.2. Division of exome-identified non-synonymous missense alleles in *HNF1A* between Oxford and Bergen for large-scale collaborative functional follow-up. (A) Oxford variant set **(B)** Bergen variant set **(C)** Distribution of exome-detected *HNF1A* alleles by centre based on exon position and functional domain.

3.2 Methods

3.2.1 *HNF1A* sequence

Human liver *HNF1A* cDNA in expression vector pcDNA3.1/HisC (NM_000545.6; Ensembl gene ID ENSG00000135100, transcript ID ENST00000257555) was used for all functional assessments. *HNF1A* NM_000545 is the longest transcript and is the one used by diagnostic genetic testing laboratories in the UK (NHS) (Colclough et al, 2014). The reference ('wild-type') sequence used in this study harbours the common coding allele p.Ile27Leu (p.Leu27Ile also tested as part of the Bergen variant set) (Gaulton et al, 2015).

3.2.2 Bioinformatic prediction

Since different computational prediction algorithms rely on different methods, criteria and levels of data integration for assigning pathogenicity (conservation, frequency, function, structure), consensus across multiple tools on the deleteriousness of an allele is considered a supporting line of evidence for variant classification (Richards et al, 2015). I therefore used the following four *in silico* tools to evaluate the pathogenicity of the exome-identified *HNF1A* missense alleles and the MODY reference alleles **(see subsection 3.3.1)**:

- SIFT (Sorts Intolerant From Tolerant): missense variant prediction based on protein sequence conservation (Kumar et al, 2009)
- PolyPhen-2 (polymorphism phenotyping version 2): missense variant pathogenicity scores based predicted effect of amino acid substitution on the structure and function of the resulting protein (Adzhubei et al 2013)

- MutationTaster: integrative tool which predicts on basis of sequence conservation (PhyloP and PhastCons multiple sequence alignment scores) and data from dbSNP, ClinVar, ENCODE, 1000 Genomes Project (Schwarz et al, 2014; Eilbeck et al, 2017)

- Combined Annotation Dependent Depletion (CADD): scores based on multi-data integration (evolutionary conservation metrics, SIFT and PolyPhen-2 missense pathogenicity scores, and functional genomic data such as transcription factor binding, DNase I hypersensitivity) (Kircher et al 2014; Eilbeck et al, 2017)

3.2.3 Stratification of *HNF1A* alleles with unsupervised learning methods

The rationale behind the methodological choices I made in this chapter were driven by the hypothesis that unsupervised learning techniques would enable unbiased spatial organisation of *HNF1A* alleles along the spectrum of *in vivo* glycaemic phenotypes on the basis of multi-dimensional *in vitro* function.

Clustering and principal component analysis (PCA) are the most widely used unsupervised learning techniques to facilitate exploration and examination of the structure of large multi-layered datasets which have not been category-annotated (i.e. not supervised) (Becker et al, 1988; Mardia et al, 1979; Venables and Ripley, 2002; Jain and Dubes, 1988). By reducing dimensionality and revealing maximal uncorrelated variance, these tools allow for informative visualisation and discovery of subgroups. I employed these methods to uncover potentially interesting patterns within the complex multidimensional *HNF1A* functional datasets.

(i) PCA

I used the *prcomp()* function in the R stats package v3.5.0 to perform principal component analysis of multi-parametric *HNF1A* functional data generated by both research groups at Oxford and Bergen. Each dataset was prepared independently for PCA **(see subsection 3.3.2)**. Polished data matrices were scaled and centred within the *prcomp()*. Eigenvalues (explained variance), eigenvectors (contributions to variance or to the orientation of data transformation in multivariate space), PC coordinates (scores), and standard deviation (square root of eigenvalues) were extracted from the analysis and plotted with ggplot2 and the factoextra R packages.

(ii) K-means partitioning

I used 4 different algorithms to systematically and empirically determine the most suitable number cluster to divide the orthogonally transformed data in PC space **(see subsection 3.3.4):**

1. Elbow method: the number of k cluster is varied from 1 to 10 and the within-cluster sum of square (WSS) is computed for each k. The point at which the WSS curve bends corresponds to the optimal number of clusters.

2. Silhouette method (Kaufman and Rousseeuw 1990): computes an average silhouette values (quality of cluster compactness) for different k cluster 1 to 10. A higher average silhouette values indicates higher quality clustering.

3. Gap statistic method (Tibshirani et al, 2002): computes total within-cluster variation for k_1 to k_{max} and compares against null reference data distribution. The gap statistic is estimated as deviation of the observed

intra-cluster variation from expected. The optimal number of k clusters is chosen as that which maximises the gap statistic (greatest deviation from random uniform distribution).

4. The *NbClust()* function (Charrad et al, 2014): in a single function call, this method provides 30 indices for determining the optimal number of clusters and provides the best one based on the majority rule (distance set as Euclidean).

(iii) Hierarchical clustering

I used the hclust() function in the R stats package to perform hierarchical clustering using a data matrix comprised of alleles and their PC coordinates from the PCs which contributed to >85% of data variance (Kaiser's criterion). The hclust() function computes the Euclidean distance (measure of the dissimilarity between each pair of observations by computing linear distance) between data points which it then organises hierarchically based on one of 4 hclust() methods (see comparative analysis using Oxford variant set in **subsection 3.3.3(ii)**).

Based on a literature review of the most common methods for hierarchical agglomerative clustering, I selected the WARD ('ward.d2') minimum variance method. The WARD minimum variance method ('ward.d2') (Ward, 1963) is an agglomerative bottom-up hierarchical clustering method based on sum-of-squares criterion: subjects begin as independent clusters (n clusters of size 1) in multivariate Euclidean space, and then all possible pairs of clusters are combined and the sum of the squared distances (Euclidean) within each cluster is calculated (Murtagh and Legendre 2014). This is then summed over all clusters; the combination that yields the lowest sum of squared distances is selected. This

method produces minimal within-group dispersion/variance at each fusion (Murtagh and Legendre 2014).

In comparative analyses, the WARD method has been reported as demonstrating high robustness, high sensitivity to data standardization, and high definition of clusters (Singh 2008). After testing all possible cluster pairs, the WARD method produces a combination of clusters with the lowest sum of squared distances (low levels of dissimilarity). Therefore, applying the WARD method to data analysed by PCA (which reveals maximum uncorrelated variance) minimises further interference with the structure and variance revealed by PCA whilst allowing for visualisation based on multiple PC dimensions (>85% variance).

Other common hclust methodologies include complete, average, and single linkage which vary in robustness depending on the complexity of the dataset. Average linkage typically results in a loss of within-cluster information with the iterative and progressive reduction of dissimilarity at each fusion (Legendre, 1998). In complete linkage, the merging of clusters is based on the smallest maximum pairwise distance (i.e. fusion dictated by the similarity of the most dissimilar members of two clusters at each step), and it is, as such, more sensitive to outliers (Jain et al, 1999). Conversely, single linkage clustering is based on the smallest minimum pairwise distance (i.e. fusion dictated by the similarity of most similar/closest members of two different clusters) (Legendre, 1998; Jain et al, 1999). This method typically yields unintuitive patterns arranged in a long chain dendrogram (Jain et al, 1999).

3.2.4 Variant-phenotype mapping

This effort was a collaboration with Mr. Kevin Colclough and Professor Sian Ellard, Royal Devon and Exeter NHS Foundation Trust, Molecular Genetics Laboratory, Exeter, UK. I was hosted by Mr. Colclough and Prof Ellard in Exeter to survey their MODY diagnostic registry for overlapping *HNF1A* missense alleles with the Oxford-Bergen functional dataset.

The MODY database is divided into the classification categories pathogenic/likely pathogenic and variant of uncertain clinical significance (VUS)/ likely benign. It contains information on each MODY gene allele that is identified in individuals who undergo genetic testing for monogenic diabetes in the diagnostic clinic at Exeter. Before the advent of next-generation sequencing (NGS), genetic testing was done by Sanger sequencing of specific MODY genes which the clinician would indicate and request based on examination of the patient. Since 2013, the clinical diagnostic lab at Exeter have used a next generation sequencing exon-capture based panel of 29 monogenic diabetes genes (encompassing promoter regions and splice sites), which they have developed for more comprehensive genetic testing (Ellard et al, 2013).

A total of n=155 *HNF1A* missense alleles were documented in the diagnostic registry under both classifications. Since only missense alleles were characterised in the Oxford-Bergen effort, I only surveyed and extracted those from the clinical database. Those categorised as VUS/likely benign were primarily *HNF1A* variants which had co-occurred with a pathogenic allele in another known MODY gene (e.g. *HNF4A*). **Table 3.1** shows the list of clinical features that were available and which I extracted from the database for alleles which overlapped

with the Oxford-Bergen dataset (not all features available for each allele). We also revisited the original clinical reports of carriers for more details - particularly where clinical features were sparse - such as extra pancreatic features, vascular complications, additional family history data, and whether, for example, other MODY genes were sequenced as part of an NGS MODY panel (Ellard et al, 2013).

Table 3.1. Features of *HNF1A* allele carriers documented in UK MODY registry

Features	Measure
Age at diagnosis	age (years)
Initial Treatment	Diet/OHA/Insulin
Current Treatment	Diet/OHA/Insulin
Sulphonylurea sensitivity	Yes/No
BMI	BMI (kg/m^2)
GAD -ve	Yes/No
GAD +ve	Yes/No
ICA +ve	Yes/No
ICA -ve	Yes/No
ZnT8 -ve	Yes/No
ZnT8 +ve	Yes/No
IA2 -ve	Yes/No
IA2 +ve	Yes/No
UCPCR Value	nmol/mmol
C-Peptide	pmol/L
No of generations DM	number of generations with diabetes
Mother DM	Yes/No
Father DM	Yes/No
No of Children DM	number of children with diabetes
Ethnic Origin	ethnicity

GAD, ICA, IA2, ZnT8 = islet autoantibodies; UCPCR = urinary C-peptide creatinine ratio; OHA = oral hypoglycaemic agent; BMI = body mass index; DM= diabetes mellitus.

3.3 Results

The molecular function of 73 missense alleles in *HNF1A*, detected by exome sequencing of a multi-ethnic T2D case-control cohort, was investigated (by Dr. Jana Rundle and Dr. Amanda Bennett at the University of Oxford, and Dr. Lise Bjørkhaug, Dr. Ingvild Aukrust and Laeya Najmi at the University of Bergen) from 4 different mechanistic angles: transcriptional activity, DNA binding, protein abundance, and nuclear localisation **(Figures 3.1 and 3.2)**.

To maximise the utility of the multiparametric *HNF1A* functional datasets generated, I designed an unsupervised learning pipeline **(Figure 3.3)** that would enable the following:

1. A holistic assessment of the functional data by using information from as many mechanistic dimensions as possible to gain a comprehensive idea of variant behaviour.
2. Unbiased stratification of alleles along the spectrum of glycaemic phenotypes and T2D risk associated with *HNF1A* based on multi-parametric data to avoid arbitrary functional cut-offs based on 1-D functional data.
3. An assessment of the relative contributions of each functional dimension to molecular dysfunction/allele pathogenicity.
4. Map allelic effects on molecular function with phenotypic outcome using clinical and genetic data from the UK MODY diagnostic registry (Exeter) to annotate functional clusters with clinical knowledge and identify potentially discordant classifications between functional genotype and clinical phenotype.

Briefly, the pipeline I developed begins with preparation of the dataset for unsupervised analysis. This is followed by the addition of 2 scores to each allele, one to account for functional domain and the other for spatial variation in *HNF1A* isoform expression (exon position), as there are well-established correlations between variant position in *HNF1A* and clinical phenotype (Harries et al, 2006; Bellanne-Chantelot et al, 2008). The polished dataset is processed using commonly applied unsupervised statistical analyses: principal component analysis and hierarchical clustering. I chose principal component analysis for its dimension reduction and data structure-revealing properties. I applied a minimum variance distance-based hierarchical clustering method following dimension reduction to enable 2D visualisation of subgroups (dendrogram) that are hierarchically organised on the basis of >2 dimensions of information. This allows for the retention of assigned PC scores/coordinates from the total number of informative principal components (those which contribute >85% of data variance).

I begin this section by providing details on the exome-detected alleles from the study in which they were identified in addition to results from *in silico* predictions of pathogenicity and allele counts in the general population (gnomAD). I then present results from the various stages of analysis development, optimisation and validation which have culminated in the pipeline depicted in **Figure 3.3**.

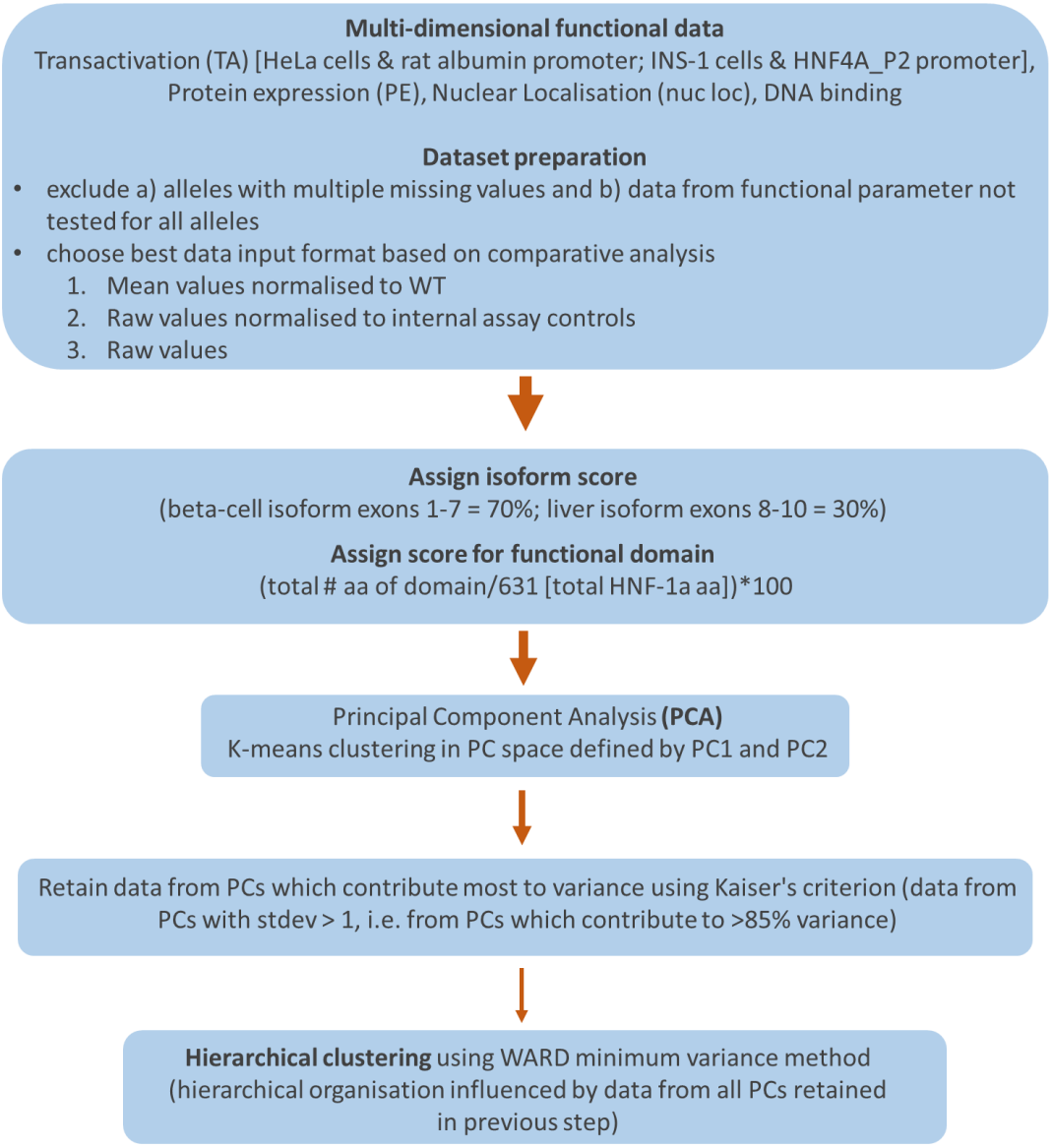


Figure 3.3. Analytical pipeline for unsupervised stratification of *HNF1A* missense alleles along the *in vivo* glycaemic spectrum using multi-dimensional *in vitro* functional data. The pipeline I developed to perform unbiased classification of functionally characterised missense alleles in *HNF1A* is as follows: 1) dataset preparation (omitting missing values and unevenly represented functional parameters, and selecting the optimal input data format), 2) assignment of weights for tissue-specific isoform expression (liver v pancreatic beta-cell dominant isoforms) and functional domain (dimerization, undefined, DNA binding, transactivation), 3) Dimension reduction by principal component analysis, 4) retention of values from principal components which contribute the most to data variance according to Kaiser's criterion (>85% explained variance by eigenvalues i.e. all PCs with standard deviation > 1) to avoid fitting noise, 5) hierarchical clustering analysis of data points from PCs retained in step 4 using WARD minimum variance method, 6) visualisation of clusters plotted in PC space (k-means-derived) and as dendrograms. Abbreviations: aa= amino acids; stdev = standard deviation; PCs = principal components.

3.3.1 Allele frequency and computational predictions of pathogenicity

To enable a comprehensive assessment of *HNF1A* variant-effects, I compiled case-control and population-level allele frequency data and pathogenicity predictions from multiple *in silico* tools for the missense alleles identified in the exomes of 13K T2D cases and controls from five ancestry groups **(Table 3.2).** There was no evidence for *HNF1A* association with T2D susceptibility in the study neither at the single variant nor at the gene level, only a marginal aggregate association with T2D risk in a multigene test (*P*=0.023) which included rare coding alleles in a set of genes implicated in monogenic/syndromic diabetes or related glycaemic traits (Fuchsberger et al, 2016).

The vast majority of alleles were identified in both T2D cases and controls, and for the few observed exclusively in T2D cases, they were identified with a frequency of either 1 or 2 cases per allele **(Table 3.2)**. Consensus was rare among the multiple bioinformatic prediction tools I used to assess missense variant pathogenicity (SIFT, PolyPhen-2, MutationTaster and CADD scores). With respect to CADD scores, pathogenicity cut-offs are rather arbitrary and relative, though a score >15 is suggested (by virtue of being the median value for all possible nonsynonymous and canonical splice site variants) (Kircher et al, 2014). Using a CADD cut-off score of 15, roughly 70% of the missense alleles would be bioinformatically classified as disease-causing, which is a drastic overestimation.

To further aid in a detailed investigation of variant pathogenicity, I extracted allele frequency values from the genome Aggregation Database gnomAD (http://gnomad.broadinstitute.org/) which is comprised of sequence data from

>120K exomes and >15K genomes as part of large-scale case-control cohort or population sequencing studies. **Figure 3.4** shows allele counts in gnomAD of all the functionally evaluated *HNF1A* missense alleles in this study. The red and orange dashed lines in the allele count plots represent frequency thresholds dictated by MODY reference alleles and the population-specific T2D risk-elevating allele (p.E508K), respectively **(Figure 3.4)**. Contrasting the frequency of alleles with unannotated function against those with known functional consequences and variant-level implication in disease provides an additional layer of context required for accurate and reliable allele classification.

In **Table 3.3**, I present the various layers of evidence for pathogenicity of the known MODY reference alleles used in this study (n=6). There is strong functional (damaging in luciferase reporter gene assays), clinical (SU response) and/or genetic (co-segregation) evidence to support the role of this allelic series in MODY pathogenesis **(references in Table 3.3)**. Previous functional studies have shown that alleles, p.R203H, p.P112L, and p.T260M, are deleterious *in vitro* based on luciferase transactivation assays. For the alleles, p.R203H, p.P519L, p.T260M, and p.A116V, there is evidence for co-segregation with classic HNF1A-MODY pathology (young-onset, dominantly inherited, non-insulin dependent diabetes), and SU sensitivity in carriers **(Table 3.3)**.

As expected, there is strong consensus among *in silico* prediction algorithms on the pathogenicity of the MODY reference alleles. They have been reported in a total of 57 families (p.P519L in 5, p.T260M in 13, p.A116V in 10, p.R203H in 4, p.P112L in 14, and p.P447L in 11) with diabetes and strong evidence of co-segregation and/or sulphonylurea sensitivity (see references in table; Colclough et al, 2014). All of the alleles are present in the Human Gene Mutation Database

(HGMD) and 4 out of 6 have been annotated as pathogenic (MODY) or likely pathogenic in ClinVar **(Table 3.3)**. The alleles are rare: only half of them are present in gnomAD, with an allele count of either or 1 or 2 (AF < 0.0008%) **(Table 3.3; Figure 3.4)**.

Table 3.2. Exome-detected *HNF1A* missense alleles selected for functional follow-up (n=73)

Variant	CDS position NM_000545		African American	East Asian	European	Hispanic	South Asian	Case	Ctrl	SIFT	PolyPhen-2	Mutation Taster	CADD
p.T10M	29	aCg/aTg	1	1	0	1	0	0	3	tolerated	probably damaging 0.965	polymorphism	23.7
p.I27L	79	Atc/Ctc	484	1483	2829	1118	1433	3746	3601	tolerated	benign 0.025	polymorphism	25.6
p.G31D	92	gGt/gAt	1	0	2	6	0	5	4	deleterious	benign 0.021	disease causing	22.8
p.P43T	127	Ccc/Acc	0	0	0	1	0	0	1	tolerated	benign 0.018	polymorphism	8.259
p.G47R	139	Ggg/Cgg	0	0	1	0	0	1	0	deleterious	benign 0.001	polymorphism	10.71
p.E48K	142	Gag/Aag	0	0	2	0	0	1	1	deleterious	benign 0.322	disease causing	18.32
p.G52A	155	gGc/gCc	17	0	0	0	0	6	11	deleterious	benign 0.001	polymorphism	0.001
p.N62S	185	aAt/aGt	0	1	0	0	1	1	1	deleterious	probably damaging 0.975	disease causing	22.1
p.D72Y	214	Gac/Tac	1	0	0	0	0	1	0	deleterious	probably damaging 0.956	disease causing	26.7
p.T82M	245	aCg/aTg	0	1	0	0	0	0	1	deleterious	benign 0.025	polymorphism	27
p.L86H	257	cTc/cAc	0	0	1	0	0	0	1	deleterious	probably damaging 1.000	disease causing	26.3
p.A97V	290	gCg/gTg	1	0	0	1	0	1	1	deleterious	probably damaging 0.966	disease causing	25.2
p.A98V	293	gCc/gTc	26	0	357	54	281	396	322	deleterious	benign	polymorphism	23.4
p.Q100K	298	Cag/Aag	1	0	0	0	0	1	0	deleterious	benign 0.397	disease causing	20.9
p.R114C	340	Cgt/Tgt	0	0	0	0	1	1	0	deleterious	probably damaging 1.000	disease causing	34
p.R114H	341	cGt/cAt	0	0	1	0	0	0	1	tolerated	benign 0.378	disease causing	22.9

Variant	CDS position NM_000545		African American	East Asian	European	Hispanic	South Asian	Case	Ctrl	SIFT	PolyPhen-2	Mutation Taster	CADD
p.R131Q	392	cGg/cAg	0	0	1	0	0	1	0	deleterious	probably damaging 1.000	disease causing	34
p.T156M	467	aCg/aTg	0	0	0	0	2	0	2	deleterious	probably damaging 1.000	disease causing	23.6
p.A161T	481	Gcc/Acc	0	0	0	0	1	0	1	deleterious	probably damaging 1.000	disease causing	31
p.R171Q	512	cGa/cAa	0	1	0	0	0	1	0	deleterious	probably damaging 1.000	disease causing	24.1
p.A174V	521	gCg/gTg	0	0	0	3	0	1	2	deleterious	possibly damaging 0.845	disease causing	23
p.Q175R	524	cAg/cGg	1	0	0	0	0	0	1	tolerated	benign 0.000	disease causing	6.627
p.G181A	542	gGg/gCg	0	0	0	3	0	2	1	deleterious	benign 0.011	disease causing	10.44
p.T196A	586	Acc/Gcc	0	0	5	0	0	1	4	deleterious	benign 0.255	disease causing	14.81
p.A276D	827	gCc/gAc	1	0	0	0	0	1	0	deleterious	probably damaging 0.999	disease causing	34
p.R278Q	833	cGg/cAg	0	1	0	0	0	1	0	deleterious	probably damaging 0.998	disease causing	34
p.G288R	862	Ggg/Agg	0	0	1	4	0	3	2	deleterious	probably damaging 0.999	disease causing	24.4
p.P290H	869	cCc/cAc	0	0	1	2	0	2	0	tolerated	possibly damaging 0.857	polymorphism	20.7
p.P291S	871	Cca/Tca	0	0	7	1	0	4	4	deleterious	benign 0.001	polymorphism	0.019
p.P291Q	872	cCa/cAa	5	0	0	0	2	4	3	deleterious	benign 0.174	polymorphism	7.117
p.A311D	932	gCc/gAc	0	1	0	0	0	0	1	deleterious	possibly damaging 0.650	polymorphism	16.63
p.R321C	961	Cgc/Tgc	0	0	0	0	1	0	1	deleterious	probably damaging 1.000	disease causing	26.5
p.R321H	962	cGc/cAc	0	1	0	0	0	1	0	deleterious	probably damaging 1.000	disease causing	25.5
p.Y322C	965	tAt/tGt	10	0	0	0	0	6	4	deleterious	probably damaging 1.000	disease causing	24.7
p.A326V	977	gCg/gTg	0	3	0	0	0	1	2	tolerated	benign 0.153	polymorphism	11.69
p.P334H	1001	cCc/cAc	0	1	0	0	0	1	0	deleterious	probably damaging 0.913	disease causing	23.1

Variant	CDS position NM_000545		African American	East Asian	European	Hispanic	South Asian	Case	Ctrl	SIFT	PolyPhen-2	Mutation Taster	CADD
p.G339S	1015	Ggt/Agt	0	0	0	4	1	3	2	tolerated	benign 0.046	polymorphism	9.323
p.P340A	1018	Ccc/Gcc	0	0	2	0	0	1	1	tolerated	benign 0.002	polymorphism	0.161
p.T354M	1061	aCg/aTg	0	0	1	0	0	0	1	deleterious	probably damaging 0.990	polymorphism	23
p.L363P	1088	cTg/cCg	0	0	1	0	0	0	1	deleterious	probably damaging 1.000	disease causing	NA
p.P379T	1135	Cct/Act	0	0	3	1	7	8	3	deleterious	probably damaging 1.000	disease causing	26.5
p.T382I	1145	aCc/aTc	0	0	1	0	0	1	0	deleterious	probably damaging 1.000	disease causing	29.3
p.L389V	1165	Ttg/Gtg	29	0	0	0	0	15	14	tolerated	possibly damaging 0.763	disease causing	12.49
p.N397S	1190	aAc/aGc	0	0	0	0	1	0	1	tolerated	benign 0.139	polymorphism	14.24
p.M412T	1235	aTg/aCg	0	0	0	0	1	1	0	deleterious	probably damaging 0.999	disease causing	25.6
p.L438P	1313	cTg/cCg	0	0	0	0	1	1	0	deleterious	probably damaging 0.999	disease causing	28.1
p.T441K	1322	aCg/aAg	0	0	1	0	1	1	1	deleterious	benign 0.064	disease causing	25.7
p.V462I	1384	Gtc/Atc	1	0	0	0	0	0	1	deleterious	probably damaging 0.959	disease causing	19.24
p.H469Y	1405	Cac/Tac	0	0	1	0	0	1	0	deleterious	probably damaging 0.998	disease causing	23.5
p.P475L	1424	cCg/cTg	0	0	1	0	0	1	0	deleterious	probably damaging 0.999	disease causing	23.8
p.Q486R	1457	cAg/cGg	0	0	0	1	0	0	1	deleterious	possibly damaging 0.701	disease causing	25
p.S487N	1460	aGc/aAc	463	1630	2556	1144	1466	3672	3587	deleterious	benign 0.014	polymorphism	15.32
p.M490T	1469	aTg/aCg	0	0	1	0	0	0	1	deleterious	probably damaging 0.989	disease causing	25.8
p.E508K	1522	Gag/Aag	0	0	0	7	0	6	1	deleterious	probably damaging 0.991	disease causing	32
p.T513S	1537	Acc/Tcc	0	0	1	0	0	1	0	tolerated	benign 0.139	polymorphism	9.125
p.H514R	1541	cAc/cGc	0	0	19	0	0	14	5	deleterious	probably damaging 0.996	disease causing	22.5

Variant	CDS position NM_000545		African American	East Asian	European	Hispanic	South Asian	Case	Ctrl	SIFT	PolyPhen-2	Mutation Taster	CADD
p.T515K	1544	aCg/aAg	4	2	0	0	0	2	4	deleterious	probably damaging 0.962	disease causing	35
p.L518F	1552	Ctc/Ttc	0	0	0	0	1	0	1	tolerated	benign 0.004	disease causing	17.58
p.I524T	1571	aTc/aCc	0	0	0	0	1	0	1	tolerated	probably damaging 0.989	disease causing	25.4
p.D526N	1576	Gac/Aac	0	0	1	0	0	1	0	deleterious	probably damaging 1.000	disease causing	NA
p.S535R	1605	agC/agA	0	0	0	0	1	1	0	deleterious	possibly damaging 0.928	disease causing	33
p.T537M	1610	aCg/aTg	0	0	0	0	1	0	1	deleterious	probably damaging 1.000	disease causing	35
p.L565V	1693	Ctc/Gtc	0	0	0	0	1	0	1	tolerated	benign 0.002	polymorphism	17.81
p.S574G	1720	Agc/Ggc	2075	2144	4561	1878	0	5373	5285	deleterious	benign 0.002	polymorphism	5.414
p.I575N	1724	aTc/aAc	0	0	0	0	3	1	2	deleterious	probably damaging 0.993	disease causing	25.4
p.H577D	1729	Cac/Gac	0	0	5	0	0	4	1	deleterious	benign 0.000	disease causing	24.2
p.H582R	1745	cAc/cGc	1	0	0	0	0	0	1	tolerated	possibly damaging 0.761	disease causing	13.64
p.R583Q	1748	cGg/cAg	2	0	3	0	1	1	5	deleterious	benign 0.007	disease causing	23.6
p.L584P	1751	cTc/cCc	0	0	0	0	1	1	0	deleterious	possibly damaging 0.956	disease causing	25
p.A586T	1756	Gcc/Acc	1	0	0	0	0	1	0	tolerated	benign 0.001	polymorphism	8.411
p.S592P	1774	Tcc/Ccc	0	0	1	0	0	0	1	deleterious	possibly damaging 0.949	disease causing	NA
p.S604R	1812	agC/agG	1	0	0	0	0	0	1	tolerated	benign 0.055	polymorphism	23.9
p.G606S	1816	Ggc/Agc	0	0	0	0	1	1	0	tolerated	benign 0.003	disease causing	12.26

Colour coding of first two columns (Variant and CDS position NM_0005454): blue= Oxford variant set; red= Bergen variant set; gray = alleles shared and characterised by both research groups at Oxford and Bergen.

Table 3.3. MODY reference alleles (n=6)

Amino acid substitution (refs)	CDS position NM_000545	RefSNP	ClinVar (condition; clinical significance)	gnomAD allele frequency	Human Gene Mutation Database HGMD	SIFT	PolyPhen-2	mutationTaster	CADD
p.P112L (Bjørkhaug et al, 2000; Xu et al, 2002; Bjørkhaug et al 2003; Bellanne-Chantelot,et al 2008)	c.335C>T	rs137853243	MODY3; pathogenic	4.064×10^{-6}	present (HGMD ID CM004772)	**deleterious**	**probably damaging 1.000**	**disease causing**	**34**
p.P447L (Hansen et al, 1997; Yamagata et al, 1996; Bjørkhaug et al 2003; Bellanne-Chantelot,et al 2008; Vaxillaire et al, 1999; Estalella et al, 2007; Johansen et al, 2005)	c.1340C>T	rs137853236	MODY3; pathogenic	not present	present (HGMD ID CM961362)	**deleterious**	**probably damaging 1.000**	**disease causing**	**34**
p.A116V (Bellanne-Chantelot,et al 2008; Klupa et al, 2002; Colclough et al, 2013)	c.347C>T	rs752886203	not present	4.063×10^{-6}	present (HGMD ID CM015315)	**deleterious**	**probably damaging 1.000**	**disease causing**	**34**
p.R203H (Awata et al 1998; Bellanne -Chantelot, et al 2008; Johansen et al 2005; Ng et al 1999; Pruhova et al 2003; Tonooka et al 2002; Weng et al 2002; Xu et al 2005)	c.608G>A	rs587780357	T1D; likely pathogenic	8.125×10^{-6}	present (HGMD ID CM993816)	**deleterious**	**probably damaging 1.000**	**disease causing**	**29**
p.P519L (Bellanne-Chantelot, et al 2008; Frayling et al 1997)	c.1556C>T	rs749673816	not present	not present	present (HGMD ID CM971460)	**deleterious**	**probably damaging 1.000**	**disease causing**	**31**
p.T260M (Bellanne-Chantelot, et al 2008; Johansen et al, 2005; Yorifuji et al, 2012; Bazalova et al, 2010; Glucksmann et al, 1997)	c.779C>T	rs886039544	not provided; pathogenic	not present	present HGMD ID CM971457)	**deleterious**	**probably damaging 1.000**	**disease causing**	**NA**

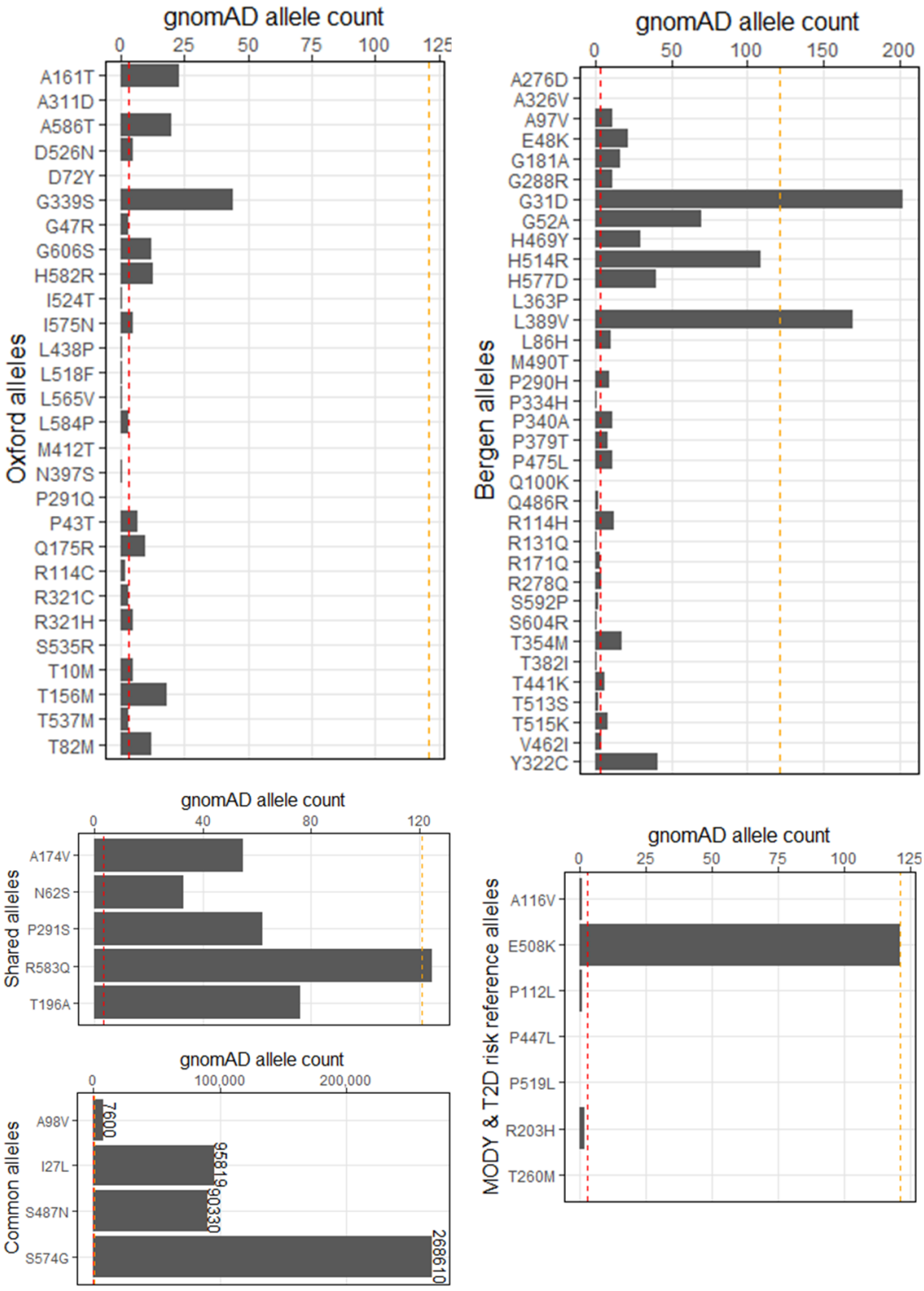


Figure 3.4. Frequency of n=79 functionally characterised *HNF1A* missense alleles in gnomAD (n=73 exome-detected; n=6 MODY reference). The red and orange dashed lines mark known ultra-rare, MODY pathogenic (allele count ≤ 2, AF < 0.0008%) and low frequency T2D predisposing allele frequencies (allele count ≤ 121, AF <0.04%) respectively.

3.3.2 Data preparation

Here, I describe the ways in which I prepared the *HNF1A* functional datasets, generated by teams at Oxford and Bergen, for PCA and clustering. This required a) polishing and trimming the datasets to ensure that all alleles are represented by an equal number of data points (number of replicates per functional parameter per allele as well as number of tested parameters per allele) and b) choosing the optimal format of input data for the analysis between unprocessed, semi-processed and fully processed/normalised functional data.

(i) Missing values

An important consideration for dimension reduction and clustering analysis is evenness in the number data points across variables or levels of information in the input data matrix. Since DNA binding ability was only tested for a small subset of alleles in each centre (Oxford and Bergen), I excluded EMSA data from the analysis. From the Bergen dataset, I also excluded three alleles from the analysis: p.A116V (missing data on protein expression by western blot), p.P447L (missing nuclear localisation data) and p.P475L (missing a biological replicate for protein expression by western blot). Additionally, the number of replicates for transactivation (both in HeLa and INS-1 cells) was reduced to n=8 in the Bergen dataset (versus n=9 TA replicates in Oxford dataset) due to high frequency of a single missing replicate per allele.

(ii) Input data format

For each of the 79 molecularly characterised alleles, and for the 4 complete dimensions of functional data (DNA binding excluded), there were three different data formats per allele per parameter to test:

1. Raw unprocessed values

2. Semi-processed values which are raw data normalised per assay to an internal housekeeping control to minimize technical variability between samples (e.g. renilla in luciferase reporter gene assays, actin/tubulin in protein expression analysis)

3. Fully processed values which are semi-processed data averaged and normalised to wild-type values to reflect difference in function relative to WT (summary data)

There are no recommendations or templates in the literature to guide the formatting of multi-layered experimental data for analysis using unsupervised statistical methods. Therefore, I had to independently determine which of the three formats of input *in vitro* functional data would be most appropriate and would reveal the most accurate and biologically meaningful variant clustering patterns.

Using the Oxford variant set, I analysed each of the three functional data formats by principal component analysis and k-means partitioning (randomly assigned k=4 clusters for simplicity) **(Figure 3.5)**, and assessed format suitability on the basis of the following criteria:

- Distinct clustering and distance between WT, T2D risk reference allele of intermediate function (p.E508K), and damaging MODY reference alleles **(Table 3.3)**.

- Quality and resolution of k-means-derived clusters in multivariate Euclidean space

- informed, semi-supervised assessment of cluster composition (i.e. examination of allele clustering trends using knowledge of *in vitro* molecular function)

- Distribution quality of orthogonally transformed alleles along PC axes
  - visual assessment of the impact of the different levels of functional data normalisation, or lack thereof, on the spatial organisation (distribution) of alleles and on intra-/inter-cluster variability

Of the three data formats, the semi-processed values (i.e. raw data normalised to internal assay-specific controls) yielded the highest resolution clusters and distribution of alleles in multivariate space. With semi-processed data, WT, p.E508K, and the majority of MODY reference alleles occupied separate distinct clusters that were spatially distributed along PC1 (which explains half of the linearly uncorrelated variance) in a manner which matched a predicted sequence-function continuum **(Figure 3.5B)**. Very few alleles from the Oxford variant set exhibited deleterious effects in functional assays; this trend was most accurately represented in the analysis of semi-processed data where only two exome-detected variants occupied the same cluster as 4/6 MODY control alleles. The cluster defined by the MODY reference alleles (p.P112L, p.P519L, p.P447L, and p.T260M) and the cluster defined by WT were separated by an intermediate cluster defined by known T2D risk alleles (p.E508K and p.A98V) **(Figure 3.5B)**.

PCA of mean values normalised to WT function (i.e. fully processed data) yielded low complexity distribution which appeared to resolve only the most extreme differences. Variance was noticeably diminished by over-normalisation, as is evident from concentrated clustering of alleles around WT. While WT, p.E508K

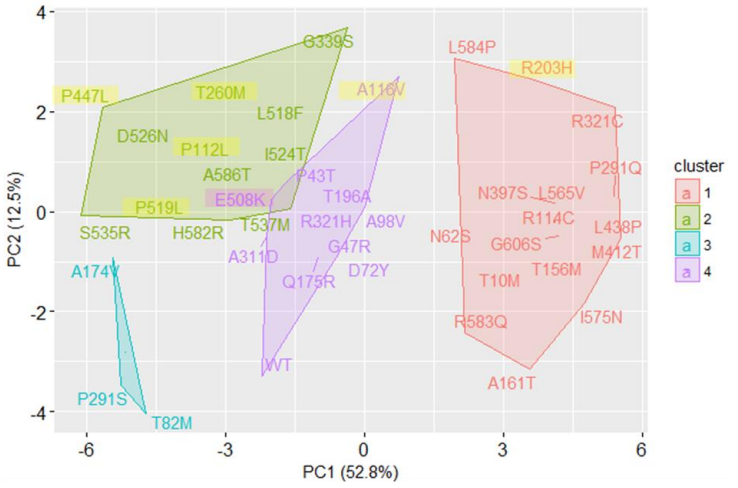
and 3 MODY reference alleles occupied different clusters, the sequence-function gradient was not as clearly resolved in PC space. Most coordinates were within a 2-point distance from WT, with the exception of a small cluster of 3 MODY control alleles and p.D526N **(Figure 3.5C)**. There was no obvious intermediate cluster defined by T2D risk alleles: p.E508K sat in a cluster with 3 MODY alleles and p.A98V in an independent cluster that was shared by neither WT nor p.E508K **(Figure 3.5C)**.

Clusters derived from raw data were similar to those generated using semi-processed data, though not as well defined, which is likely due to the retention of a high level of technical noise and variability. Using unprocessed data, WT and p.E508K occupied the same intermediate cluster which overlaps with a cluster characterised by 4/6 MODY reference alleles **(Figure 3.5A)**.

In conclusion, functional data normalised to internal assay controls (i.e. semi-processed) yielded the most robust clustering trends in PC space based on distribution quality and known and expected sequence-function relationships.

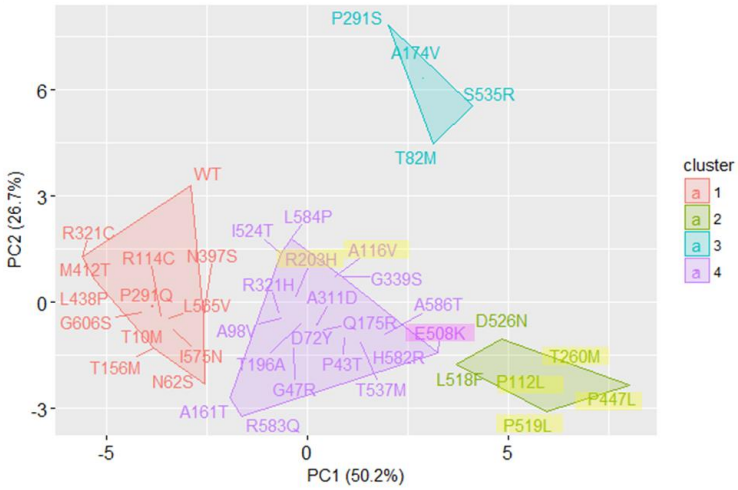
A.

Raw functional data (unprocessed)



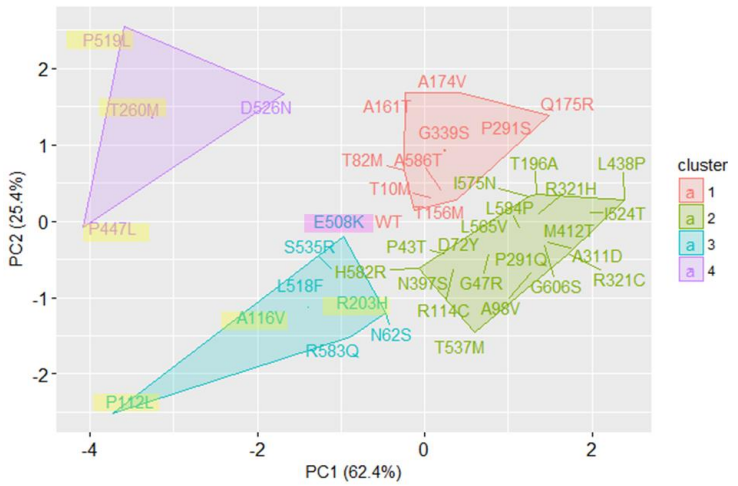
B.

Semi-processed functional data (normalised to internal assay controls to minimise technical variability)



C.

Fully processed functional data (average semi-processed data normalised to WT activity)



MODY reference

T2D risk reference

Figure 3.5. Assessment of optimal input format of multi-dimensional functional data (luciferase transactivation in HeLa and INS-1 cells, protein expression, and nuclear translocation) from the Oxford variant set using principal component analysis (PCA) and k-means clustering. (A) K-means-derived clusters from raw functional data **(B)** K-means-derived clusters from raw data normalised to internal assay controls (firefly renilla in luciferase assays, tubulin in protein expression analysis by Western blot, and the cytoplasmic ratio of HNF-1a protein expression in nuclear localisation assays). **(C)** K-means-derived clusters from averaged, doubly normalised functional data (raw normalised to internal assay controls, averaged across technical replicates, and normalised to WT activity).

3.3.3 Analysis optimisation

To refine the quality and enhance the clinical and biological relevance of allele clusters driven by semi-processed functional characterisation data, I assigned weights to each allele for tissue-specific expression of *HNF1A* isoforms (implications for clinical phenotypic manifestation) and functional domain (varied levels of mutation tolerance) **3.3.3(i)**. Additionally, I compared the performance of my hierarchical clustering method of choice (WARD) against the other most commonly applied hierarchical clustering techniques by examining cluster resolution on the basis of multi-dimensional functional data and added scores for isoform expression and functional domain **3.3.3(ii).**

(i) Scores for tissue-dependant transcript isoform expression and functional domain.

The relationship between allele position in *HNF1A* (spatial variation in isoform expression) and clinical phenotype (diabetes age of onset) is well established (Harries et al, 2006; **see subsection 3.1.1**). To factor this biological information into the model, I assigned weights to each allele of the 79 functionally characterised *HNF1A* alleles for functional domain and tissue-specific isoform expression (via exon position) **(Figure 3.3)**. I then performed a comparative analysis between the functional dataset from the Oxford variant set with and without annotations for exon position and functional domain using PCA and hclust to assess impact of the weights on cluster resolution **(Figure 3.6)**.

In addition to assessment criteria delineated in subsection **3.3.2(ii)** – particularly clustering patterns of WT and MODY/T2D risk reference alleles and subgroup composition that is consistent with *in vitro* function - I specifically examined

reorganisation or restructuring influenced by assigned weights for exon position/functional domain.

I expected the added scores to have subtle yet physiologically relevant effects on the organisation of alleles in multivariate space. Therefore, while not an empirically determined value, I divided the alleles into k=6 clusters in PC space to enable a detailed visual examination of finer differences in grouping patterns. Data from the PCs which were driving the majority of variance were retained in hclust analysis using the WARD minimum variance method (PC 1-3 for **Figure 3.6A** and PC 1-5 for **Figure 3.6B**). The dendrograms were divided into 6 subgroups (gray boxes) to a) facilitate comparison with k-means clusters and b) resolve subgroups at a height between 5 and 10 where the level of information per cluster is neither too broad (higher fusions) nor too detailed (lower fusions).

Broadly, and as expected, clustering trends were almost identical with and without scores for exon position and functional domain – that is, the two largest separations (highest fusions at 15-20 on the height scale) were comprised of the same alleles **(Figure 3.6)**. The main difference worth noting is the allele composition of the clusters defined by the majority of MODY reference alleles on the far right in the dendrograms.

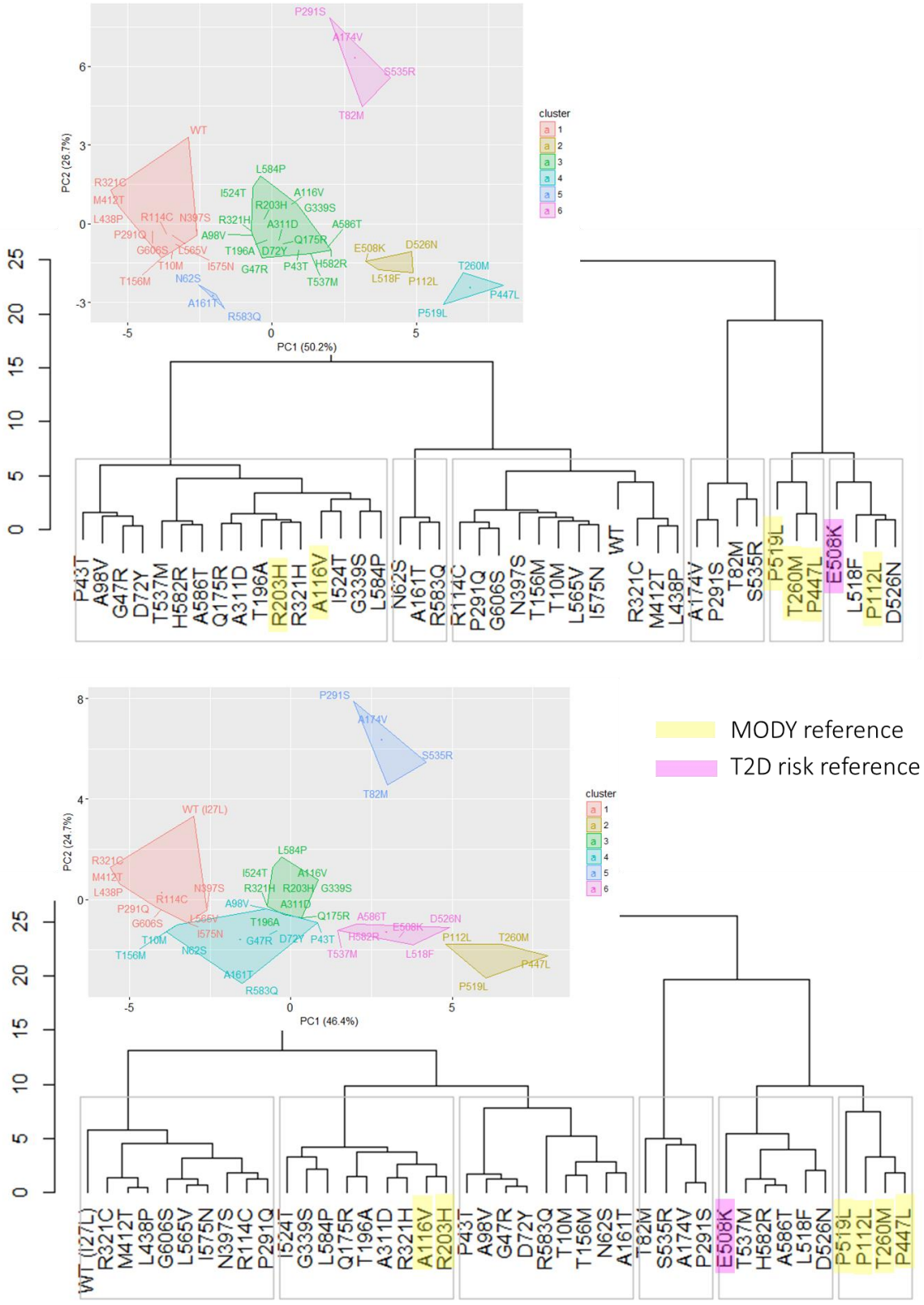


Figure 3.6. Principal component analysis (PCA) and hierarchical clustering (hclust) using multi-dimensional functional data from Oxford variant set without (top) and with (bottom) assigned weights for isoform expression and functional domain. The clusters in PC space were defined by the cutting the hclust dendrograms into k=6 subgroups. For the functional dataset without assigned weights **(top)** values from the first 3 PCs (>85%) were included in hclust analysis; whereas values from the first 5 PCs (>85%) variance were retained in hclust analysis of the functional dataset with assigned isoform and domain weights **(bottom).**

Scores for differential isoform expression in the hepatocyte/kidney (*HNF1A*(A); exons 1-10; aa 1-631) and pancreatic beta cells (*HNF1A*(B); exons 1-7; aa 1-500 and *HNF1A*(C); exons 1-6; aa 1-436), resulted in a clearer separation between 'functional' liver isoform alleles (e.g. p.D526N, p.L518F) from the pathogenic and functionally deleterious MODY reference alleles (**Figure 3.6B**). Moving up, the liver isoform alleles fuse with the moderately impaired T2D risk allele p.E508K, and then ultimately with the exclusive MODY allele cluster at height=10 **(Figure 3.6B)**.

With respect to function domain scores, subtle yet noteworthy differences were observed, particularly in the order of the alleles which constitute the three leftmost clusters in each dendrogram **(Figure 3.6)**. For example, the MODY reference variants, p.R203H and p.A116V, both located in the *HNF1A* DNA binding domain and severely impair DNA binding *in vitro*, are directly linked in the weighted dendrogram **(Figure 3.6B)** whereas they occupy separate subgroups without function domain annotations **(Figure 3.6A)**. The resolution and quality of those clusters was also enhanced, as the number of alleles per cluster is more evenly distributed after weight assignment **(Figure 3.6)**.

(ii) Comparison of hierarchical clustering methods

The rationale behind incorporating hierarchical clustering to the analytical pipeline is that it would permit 2D visualisation of distances or (dis)similarities between *HNF1A* alleles as dictated by coordinates or PC scores from the most informative principal components (n=PCs contributing >85% variance).

I compared the performance of the WARD method against complete, single and average linkage methods using the functional data from the Oxford variant set,

with assigned weights for function domain and exon position, using their coordinates from the first 5 PCs (>85% variance) **(Figure 3.7)**. Again, I assessed performance based on a) predicted grouping patterns of WT, and MODY/T2D risk reference alleles and b) known molecular function of alleles which co-occupied clusters defined by WT, T2D risk and MODY reference alleles.

Based on these metrics, single linkage clustering was poorest in performance and yielded the lowest resolution clusters **(Figure 3.7A)**. Incidentally, single linkage did not correlate well with the WARD method, which was most well correlated with complete linkage followed by average linkage **(Figure 3.7B)**. At the final fusion, the WARD method neatly separated all alleles into two large distinct groups **(Figure 3.7A)**:

1. One is clearly defined by WT, common T2D risk-modifying (p.A98V) and low frequency (p.N62S, p.T196A) alleles **(Figure 3.4),** as well as the alleles that behaved WT-like in functional assays (the vast majority of the Oxford variant set).

2. The other is defined by a cluster of MODY reference alleles (p.P519L, p.P112L, p.T260M, and p.P447L), linked to a cluster of functional liver alleles, linked to the low frequency T2D risk increasing allele p.E508K. This subgroup was linked to a cluster of 4 alleles which demonstrated variable cell-type dependant transcriptional activity (2 of which are known low frequency alleles, p.P291S and p.A174V) **(Figure 3.4)**.

With the exception of the cluster defined by the 4 alleles with inconsistent functional trends, the complete linkage method yields roughly similar subgrouping of alleles, though not as well resolved, as the WARD method **(Figure 3.7)**.

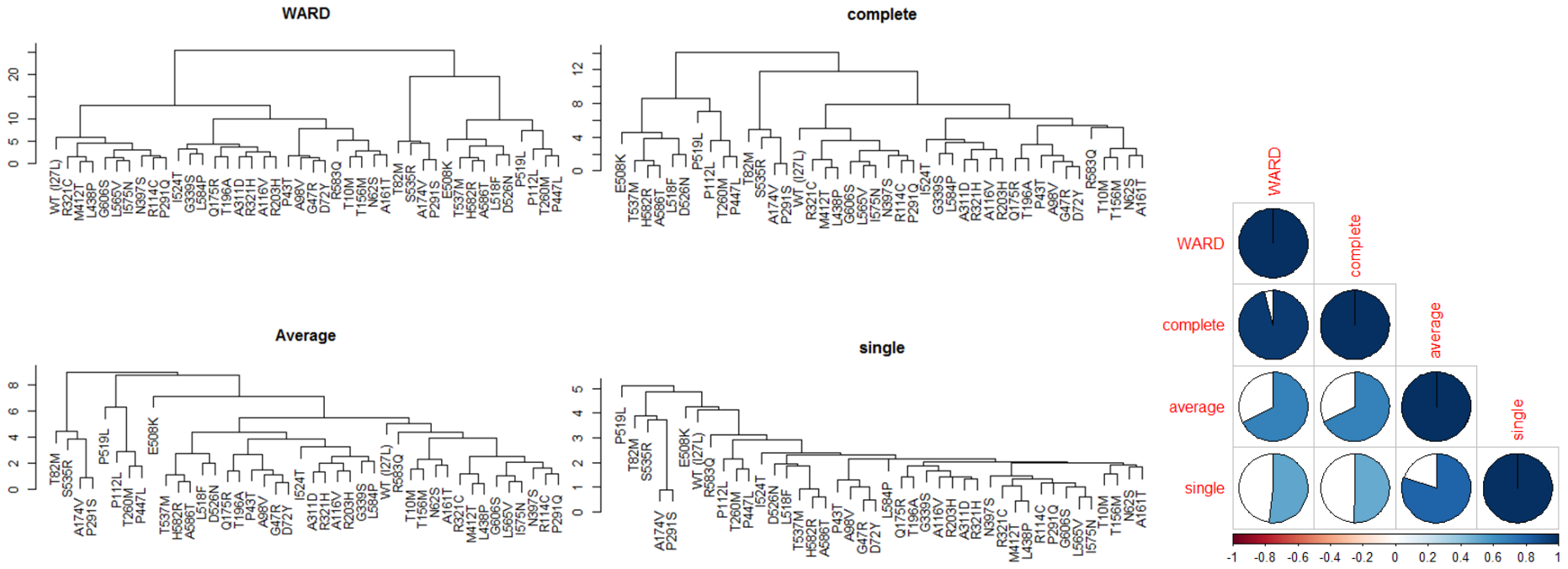


Figure 3.7. Comparative analysis of hierarchical clustering (hclust) methods using the multi-dimensional functional data from the Oxford variant set, with assigned weights for isoform expression and functional domain. (A) Dendrograms obtained from the various hclust methods (B) Cophenetic correlation between the different tested hclust methods.

3.3.4 *HNF1A* variant classification using unsupervised learning tools

In this section, I describe the results of large-scale unsupervised stratification of exome-detected *HNF1A* alleles using multi-tiered *in vitro* functional data, generated by teams at Oxford and Bergen, with the aim of mapping molecular dysfunction to clinical phenotype.

Prior to implementing the analysis, I trimmed and polished the respective datasets for unsupervised processing. Using data from the Oxford variant set, I determined that semi-processed data (i.e. raw values normalised to technical assay controls) was the most appropriate input data format by comparing allele cluster resolution in PC space with other formats of data **(Figure 3.5).** I also validated my hierarchical clustering methodology and demonstrated that the incorporation of scores for exon position and isoform expression refines cluster definition across the *HNF1A* spectrum of effects **(Figures 3.6 and 3.7).**

(i) Oxford variant set

I applied principal component analysis to the Oxford variant set **(Figure 3.8 and 3.10A).** To understand the structure of the orthogonally transformed multidimensional functional data, I dissected the principal components (n=25 which is equivalent to number of dimensions per allele). **Figure 3.8A** shows the number of PCs (n dimensions) against the percent explained variance by each. The first 5 PCs explain >85% of variance in the Oxford dataset. **Figure 3.8B** shows the eigenvectors or PC loadings, which reflect the contributions of each functional parameter to the variance explained by each PC. Transcriptional activity was the greatest contributor to data variance or the parameter which most heavily impacted the orientation of data transformation in multivariate space. This

is evident from the fact that PC1 and PC2, which together explain about 70% of data variance, are driven largely by transactivation data **(Figure 3.8B).**

Next, I systematically diagnosed the number of clusters that would suit the data using 4 different statistical methods (**Figure 3.9; see subsection 3.2.3(ii)**). The majority of tested algorithms agreed on 4 as the optimal number of k clusters for data partitioning **(Figure 3.9)**. I visualised the transformed data via 1) a 2-dimensional PC plot (PC1 v PC2) in which allele clusters were defined by k=4 groups **(Figure 3.10A),** and 2) a dendrogram produced by the WARD minimum variance hierarchical clustering method using the PC scores of alleles from the first 5 PCs to avoid fitting noise **(Figure 3.10B)**.

In previous subsections, I have provided detailed descriptions of the clustering patterns of Oxford alleles based on the WARD method (**3.3.3**). In **Figure 3.10B** I highlighted shared alleles between the centres (green), alleles shared on the basis of being T2D risk controls (pink) and alleles shared on the basis of being deleterious controls associated with a classic MODY phenotype (yellow). I used the *in vitro* functional spectrum (refined by exon position and isoform expression weights) to map the *HNF1A in vivo* spectrum onto the dendrogram, from benign/healthy to T2D risk-modifying to MODY (inherited early-onset hyperglycaemia, likely to be SU-responsive).

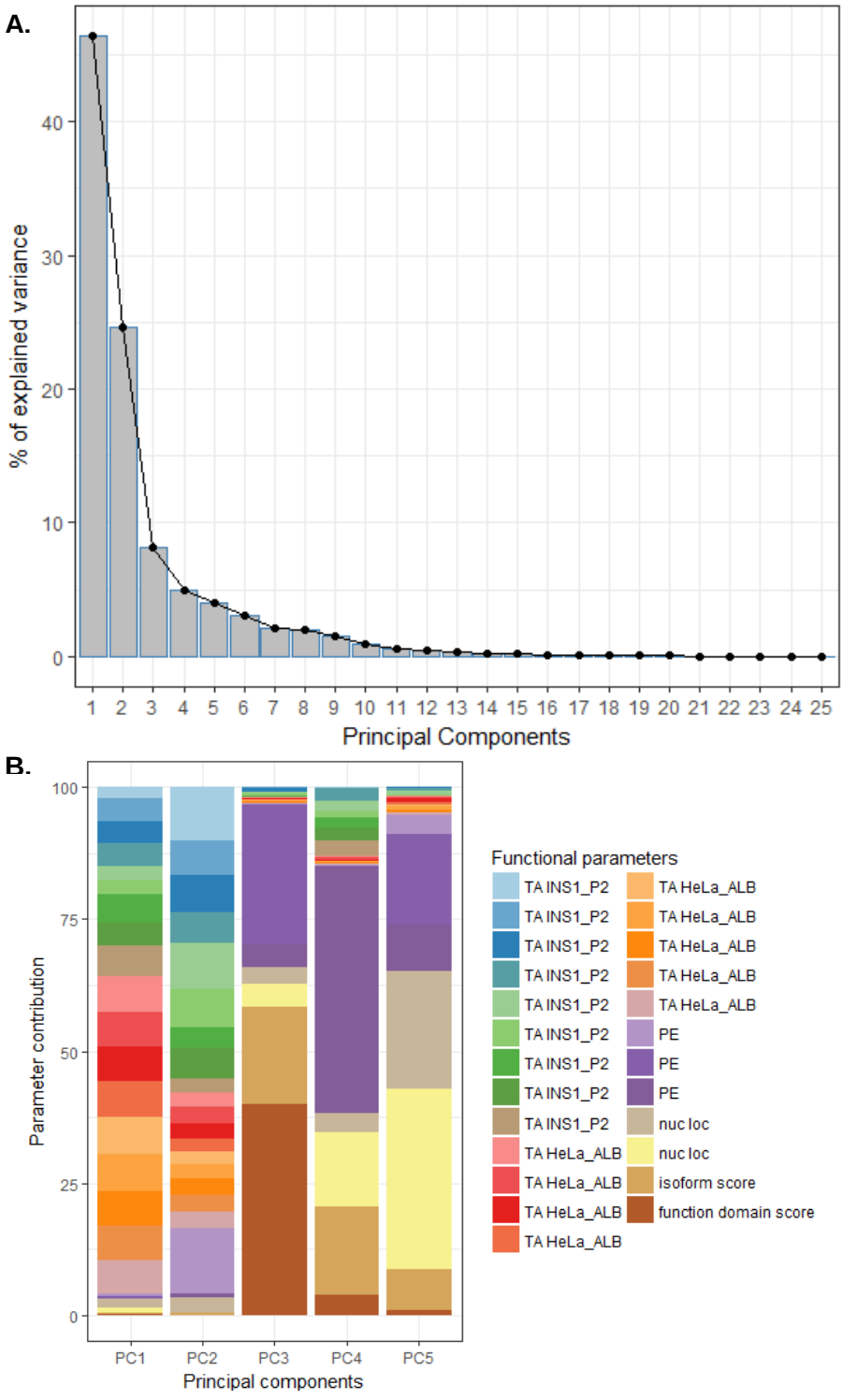


Figure 3.8. Dissection of Oxford variant set PCA. (A) Variance explained by each principal component (n= number of analysed dimensions =23). **(B)** PC loadings (eigenvectors) from PCA of Oxford alleles. TA INS1_P2 and TA HeLa_ALB are transcriptional activity data from INS-1 cells using HNF-4a-P2 promoter and from HeLa cells using rat albumin promoter, respectively (n=9 replicates per assay). PE = protein expression (n=3 replicates); nuc loc = nuclear localisation data (n=2 replicates).

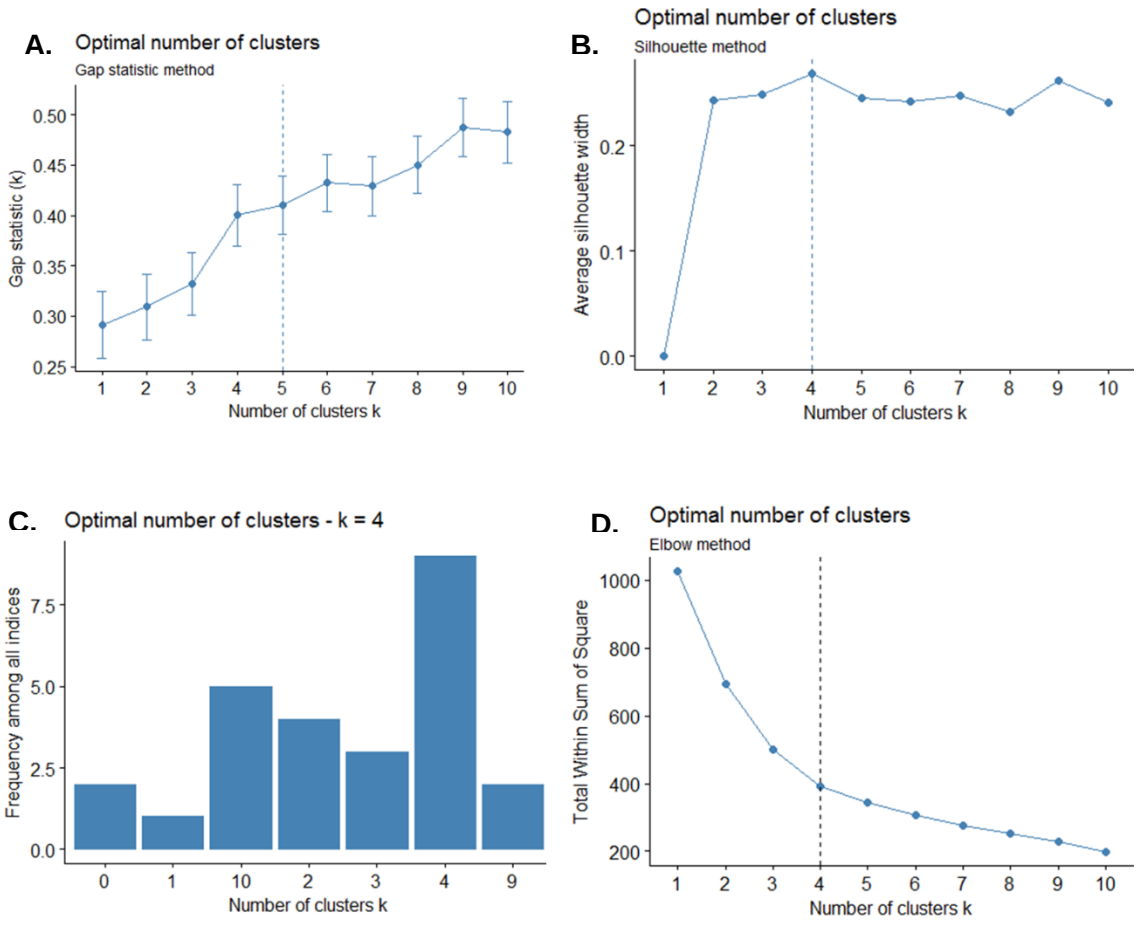


Figure 3.9. Assessment of the optimal number of clusters to use for partitioning the multi-dimensional dataset from the Oxford variant set. Four different statistical methods were used to evaluate the optimal number of k clusters into the functional data should be divided in PC space [method details provided in subsection **3.2.3(ii)**] **(A)** gap statistic method (recommends k=5 clusters as marked by the dashed line) **(B)** average silhouette method (recommends k=4 clusters as marked by the dashed line) **(C)** NbClust majority rule out of 30 indices using euclidean distance (k=4 clusters most frequently selected among tested indices) **(D)** elbow method (recommends k=4 clusters as marked by the dashed line).

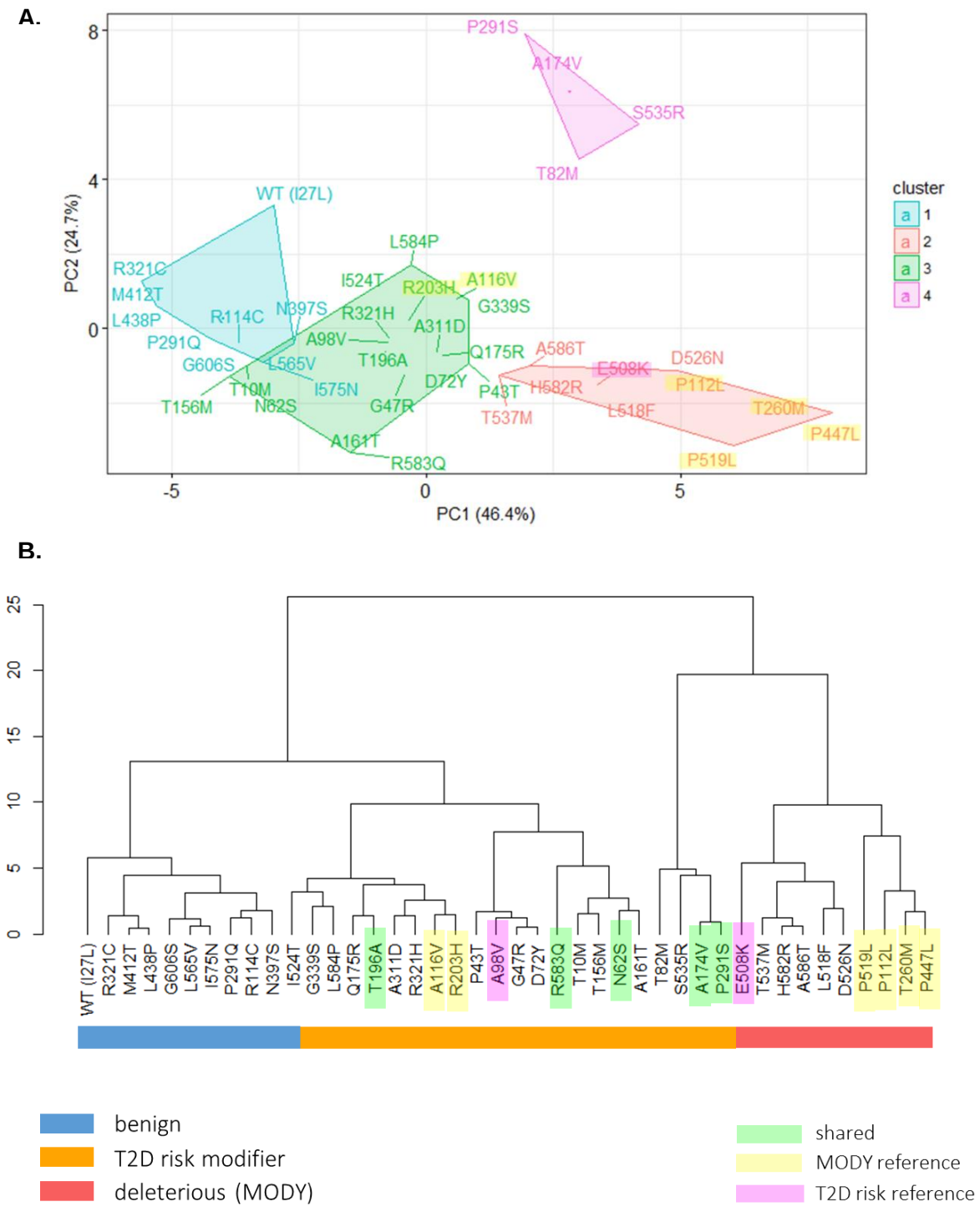


Figure 3.10. Unsupervised clustering of *HNF1A* missense alleles characterised at Oxford. **(A)** k-means clustering (k=4) of alleles in principal component (PC) space defined by PC1 and PC2 **(B)** Hierarchical clustering analysis (WARD minimum variance method) using orthogonally transformed functional data from PC1-PC5 (>85% explained variance). To optimise visualisation of the function-phenotype gradient, some branches were rotated (e.g. WT rotated from right of p.N397S to left of p.R321C; cluster defined by MODY reference alleles and adjacent cluster with p.T82M, p.S535R, p.A174V, p.P291S flipped so that the former subgroups resides at the very right end of the spectrum).

(ii) Bergen variant set

I performed the same analysis with the functional dataset generated using the set of variants characterised at Bergen **(Figure 3.3)**. The Bergen dataset is slightly more complex in that it is comprised of more alleles in total, and a wider variety of functional trends compared to the Oxford variant set.

A dissection of the principal components revealed structural information similar to that of the Oxford dataset: a) the first 5 PCs explain >85% of variance among Bergen alleles **(Figure 3.11A)**, and b) Transactivation drives the majority of variance explained by PC1 and PC2, which together explain ~57% of total variance **(Figure 3.11B)**. While trends in dimension reduction were nearly identical between the two datasets, they varied quite substantially in the eigenvector compositions of PC3, which explains ~12% of Bergen data variance and ~8% Oxford data variance. The third principal component from the Oxford dataset is driven almost exclusively by protein expression and localisation data and the added weights for exon position and functional domain **(Figure 3.8B)**. Transcriptional activity remains the dominant contributor to PC3 orientation from the analysis of Bergen data **(Figure 3.11B).**

The dataset was partitioned into 4 clusters, which was determined as the most suitable number clusters based on 3 out of 4 of the statistical tests I used **(Figure 3.12).** Their distribution along PC1 and PC2 is shown in **Figure 3.13A**. As with the Oxford dataset, PC scores of alleles from the first 5 PCs were used for hierarchical clustering (WARD) **(Figure 3.13B)**.

While only n=6 exome-detected alleles clustered with MODY reference alleles in the Oxford dataset, n=16 alleles clustered with the 4 MODY reference alleles in

the dendrogram from the Bergen dataset (2 reference alleles were excluded due to missing data points) **(Figure 3.13B)**. The majority of alleles shared between the two centres (Oxford and Bergen) displayed similar clustering patterns, perhaps with the exception of p.R203H (deleterious in Bergen TA assays, moderately impacted in Oxford TA assays) and p.E508K (in Oxford dendrogram, sits at intersection of risk-modifying and MODY clusters, linked to functional liver-dominant alleles) **(Figures 3.10B and 3.13B)**. I juxtaposed dendrograms comprised solely of the shared alleles between the two centres to better illustrate (mis)alignment in clustering trends **(Figure 3.14)**. The distinct separation at the highest fusion between WT-like/T2D risk-modifying clusters and MODY/deleterious ones which was observed in the Oxford dataset was replicated with the Bergen dataset **(Figures 3.10B and 3.13B)**.

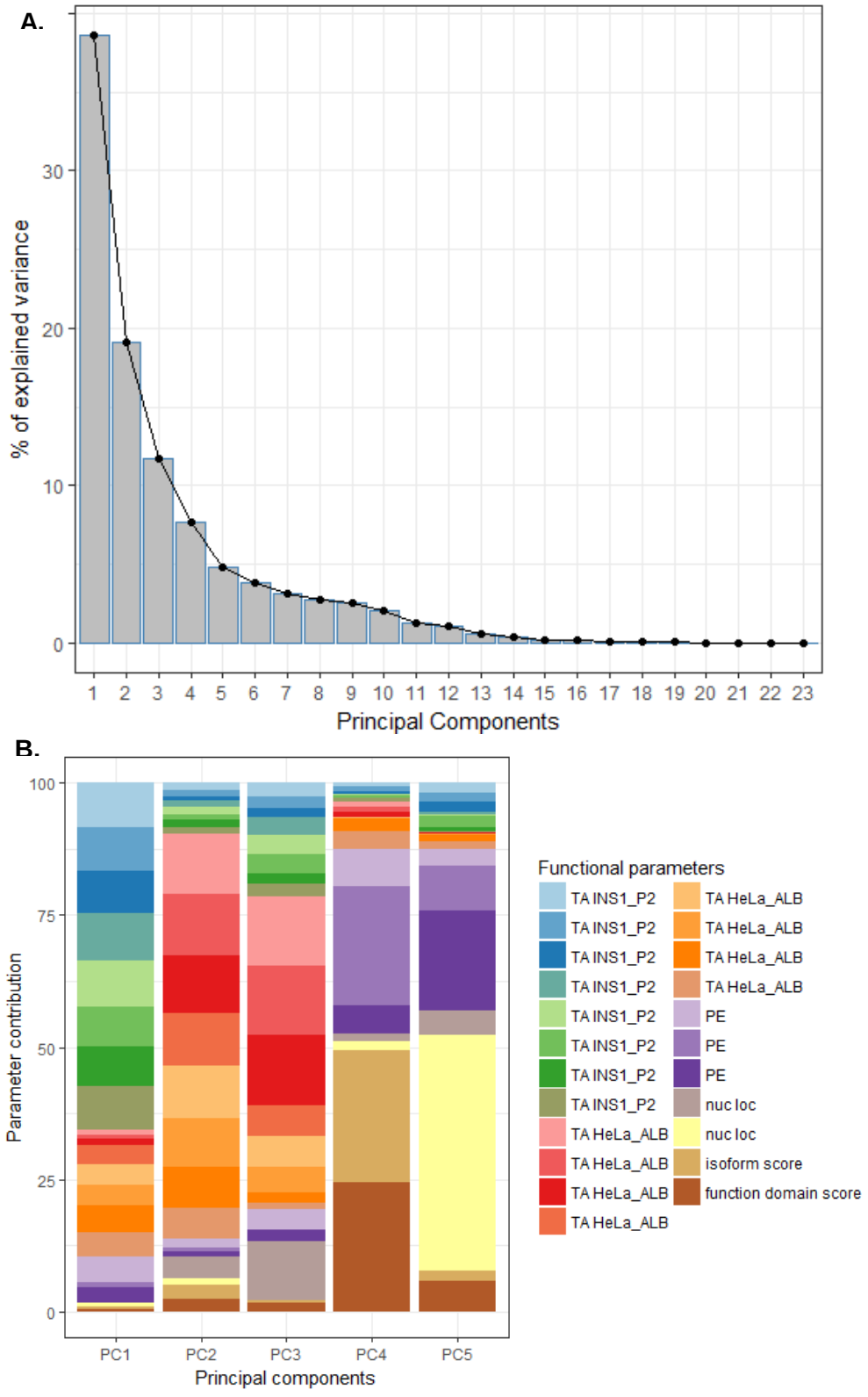


Figure 3.11. Dissection of Bergen variant set PCA. (A) Variance explained by each principal component (n= number of analysed dimensions =23). **(B)** PC loadings (eigenvectors) from PCA of Bergen alleles. TA INS1_P2 and TA HeLa_ALB are transcriptional activity data from INS-1 cells using HNF-4a-P2 promoter and from HeLa cells using rat albumin promoter, respectively (n=8 replicates). PE = protein expression (n=3 replicates); nuc loc = nuclear localisation data (n=2 replicates).

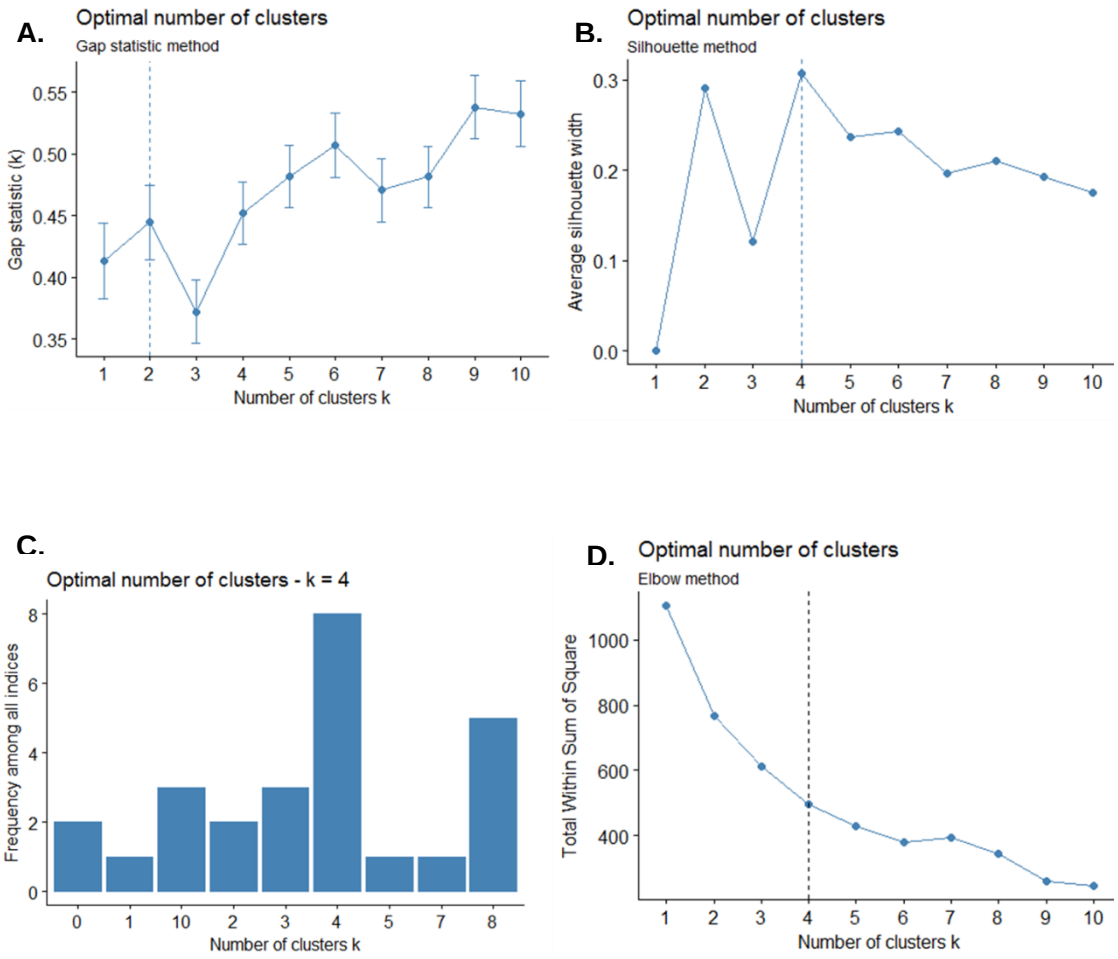


Figure 3.12. Assessment of the optimal number of clusters to use for partitioning the multi-dimensional dataset from the Bergen variant set. Four different statistical methods were used to evaluate the optimal number of k clusters into the functional data should be divided in PC space [method details provided in subsection **3.2.3(ii)**] **(A)** gap statistic method (recommends k=2 clusters as marked by the dashed line) **(B)** average silhouette method (recommends k=4 clusters as marked by the dashed line) **(C)** *NbClust* majority rule across 30 indices using euclidean distance (k=4 clusters most frequently selected among tested indices) **(D)** elbow method (recommends k=4 clusters as marked by the dashed line).

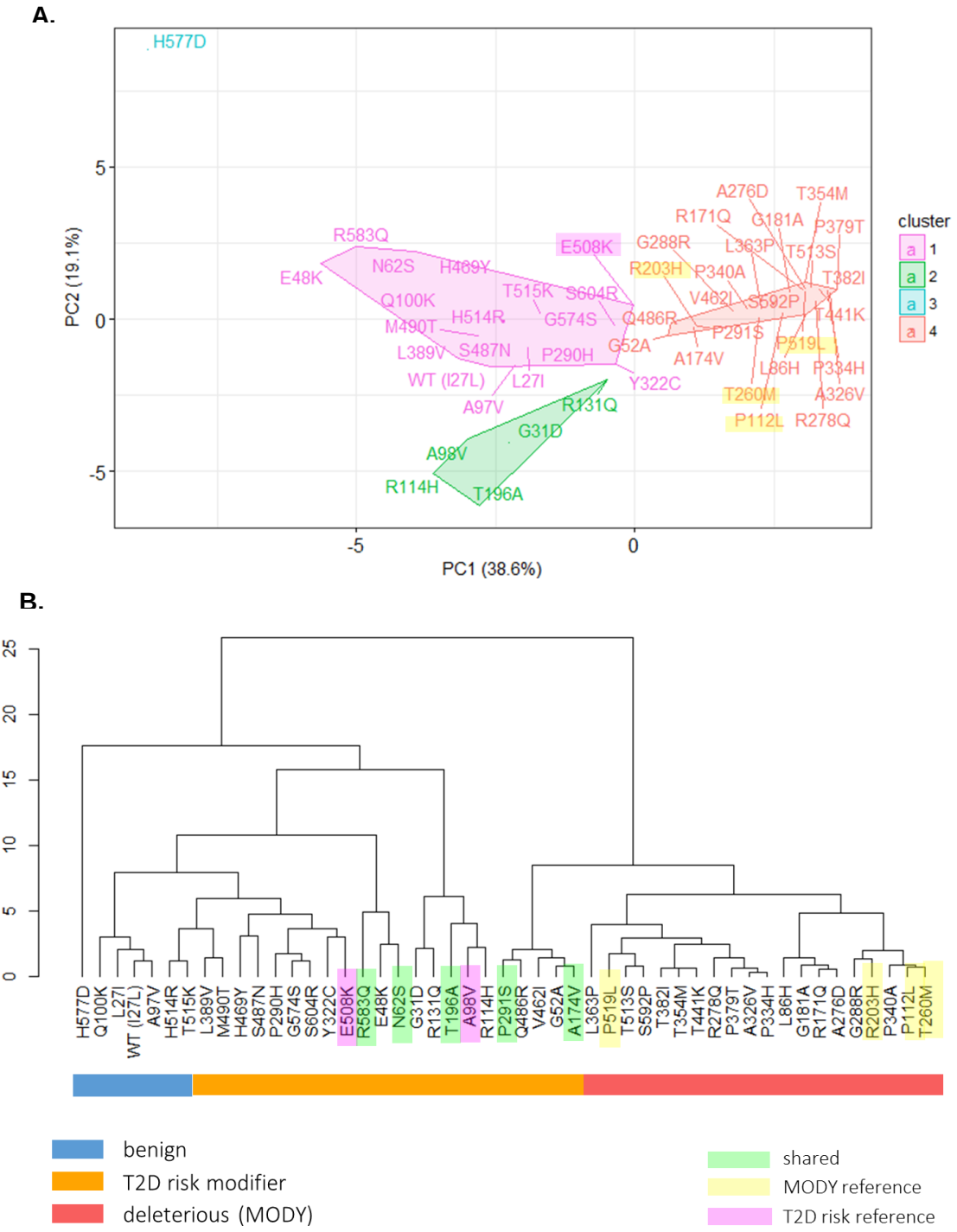


Figure 3.13. Unsupervised clustering of *HNF1A* missense allele set characterised at Bergen. (A) k-means clustering (k=4) of alleles in principal component (PC) space defined by PC1 and PC2 **(B)** Hierarchical clustering analysis (WARD minimum variance method) using orthogonally transformed functional data from PC1-PC5 (>85% explained variance). To optimise visualisation of the function-phenotype gradient, some branches were rotated (e.g. clusters defined by p.A98V and p.N62S rotated to the right of p.E508K rather than default position to the left of p.Q100K. p.L27I and WT; cluster defined by MODY reference alleles and adjacent cluster with p.Q488R, p.G52A, p.V462I, p.A174V, p.P291S flipped so that the former represents the very end of the spectrum on the right). MODY reference alleles, p.A116V and p.P447L, excluded from the analysis due to missing data points.

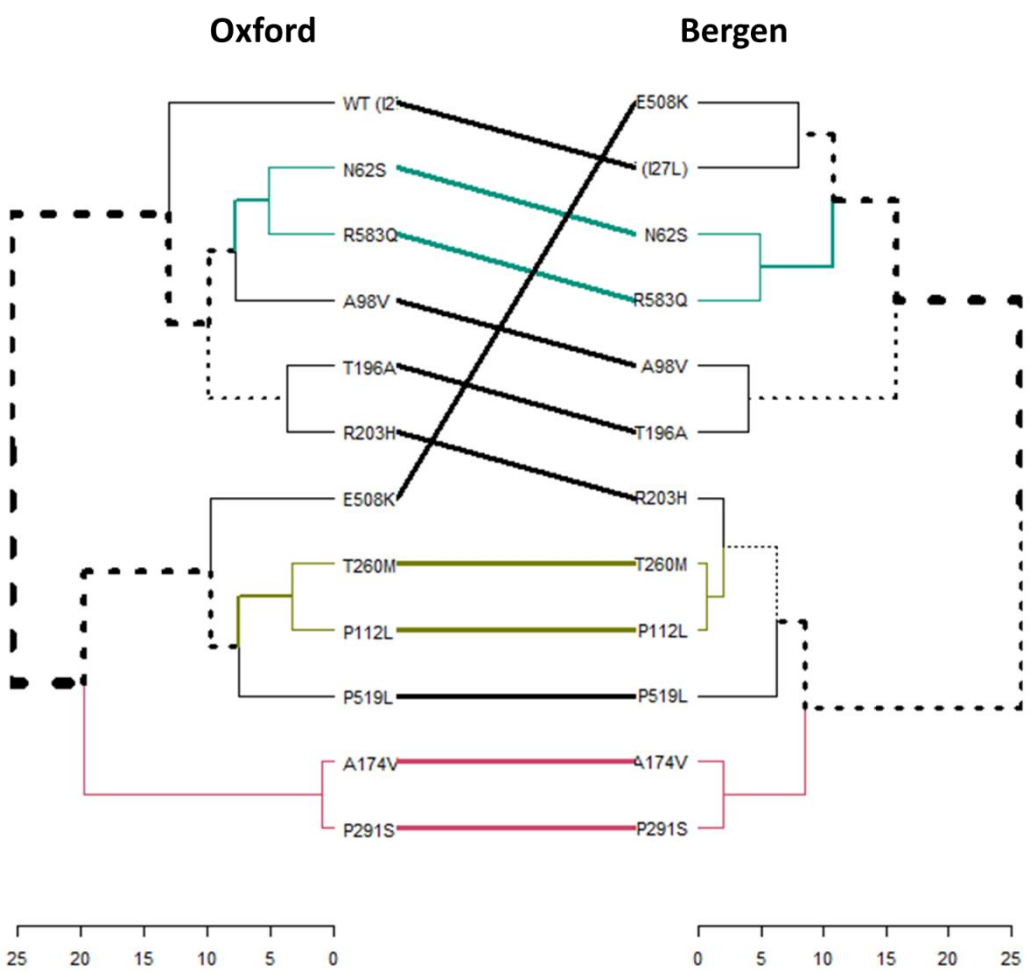


Figure 3.14. Alignment of *HNF1A* missense alleles functionally investigated by both Oxford and Bergen research groups. Dashed lines highlight nodes which contain a combination of alleles not present in the other tree (thickness corresponds to height). The colours and connecting lines highlight subgroups and/or cluster members that are present in both dendrograms. Rotational properties maintained from full variant set dendrogams shown in Figures 10B and 13B. MODY reference alleles, p.A116V and p.P447L, excluded from the analysis of the Bergen variant set due to missing data points.

3.3.5 Re-evaluating MODY diagnoses based on unsupervised functional classification

To assess the medical diagnostic utility of comprehensive and unbiased *HNF1A* sequence-function annotations, I examined their mappability to *HNF1A* clinical phenotype using clinical data from overlapping *HNF1A* missense alleles in the UK MODY diagnostic registry (Exeter).

(i) (Mis)alignment between clinical and functional classifications

At the time I conducted variant-phenotype mapping, there were n=155 *HNF1A* missense alleles identified in genetic testing of individuals with a diabetes phenotype and documented in the MODY diagnostic database at Exeter. A total of n=31 alleles overlapped with the exome-detected alleles and MODY reference variants interrogated at Oxford and Bergen **(Figure 3.15).** Of the 31 shared alleles, 19 were classified as pathogenic/likely pathogenic and 15 were classified as VUS/likely benign in the clinical registry at Exeter **(Figure 3.15).** Three of the 31 shared alleles (p.G606S, p.H469Y, p.P291S) were present under both pathogenic/likely pathogenic and VUS/likely benign classifications. All 15 missense alleles categorised as VUS/likely benign in the Exeter MODY registry demonstrated benign clustering patterns in unsupervised analysis (i.e. did not form subgroups with alleles which exhibited impaired function).

With respect to Oxford-Bergen alleles classified as pathogenic/likely pathogenic in the Exeter database (n=19), there was a mix of consistent and conflicting observations. Concordance between unsupervised functional stratification and clinical variant classification was observed for 9 out of the 19 alleles: six of them were MODY reference alleles, two from the Bergen variant set which clustered

with MODY alleles (p.A276D, p.P379T) **(Figure 3.13B),** and one allele from the Oxford variant set (p.D526N) which also formed a larger subgroup, along with other functional liver alleles, with the MODY reference alleles **(Figure 3.10B)**.

For 10 of the 19 alleles categorised as pathogenic/likely pathogenic at Exeter, sequence-function annotations based on unsupervised clustering patterns did not match clinical diagnostic variant interpretation **(Figure 3.15).** The alleles either co-occupied clusters with known T2D risk modifiers (some moderately impacted in functional assays) or with WT/neutral alleles **(Figures 3.10 and 3.13)**. Discordance between functional genotype and clinical variant interpretation prompted a thorough reassessment of allele pathogenicity on behalf of myself and Mr. Colclough at Exeter. We used information on allele frequency, both in the general population and in the MODY registry **(Figure 3.16)**, as well as clinical phenotypic data associated with the discordant alleles (n=10) to guide and inform a potential variant reclassification effort **(Table 3.3).** We changed the classification of n=7 out of the 10 alleles from pathogenic/likely pathogenic to VUS/likely benign **(Figures 3.16 and 3.17B)**. We also revisited, in depth, the concordant allele sets in both classifications **(Figure 3.17A and 3.18).** One out of 9 concordant alleles, p.D526N, originally annotated as pathogenic/likely pathogenic, was reclassified to VUS/likely benign based on comprehensive variant assessment **(Figure 3.17A).**

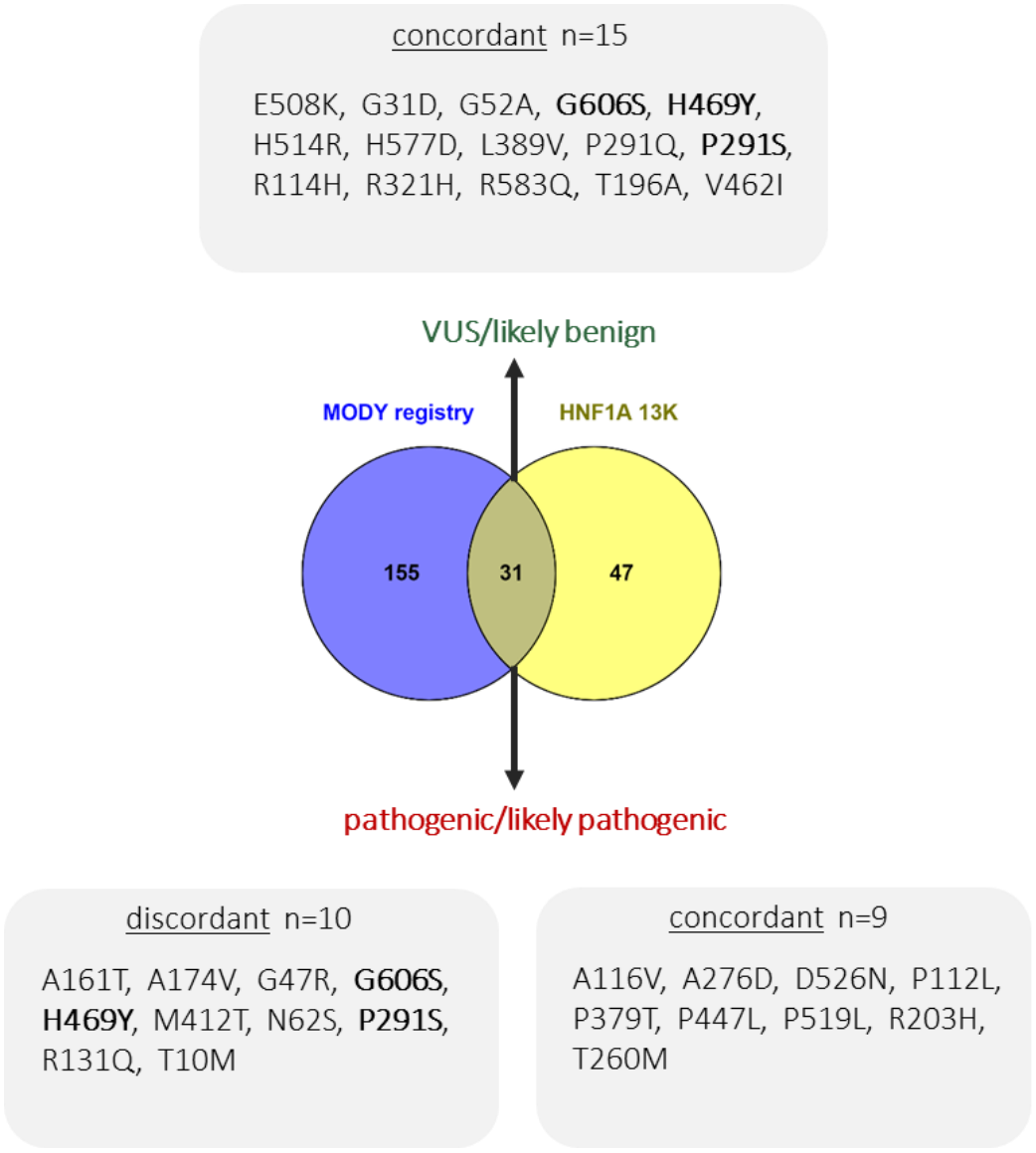


Figure 3.15. Details on the overlapping *HNF1A* missense alleles between the UK MODY diagnostic registry (n=155 total missense alleles) and the Oxford-Bergen functional dataset (n=78 alleles; common missense allele p.I27L [WT] excluded). A total of n=31 alleles from the Oxford-Bergen functional characterisation effort had been detected in the clinical diagnostic laboratory at Exeter and classified in their MODY registry. Three of the 31 alleles were found under both VUS/likely benign and pathogenic/likely pathogenic classifications in the MODY registry (highlighted in bold). All overlapping alleles under the VUS/likely benign classification (n=15) matched unsupervised functional stratification patterns (i.e. did not cluster with deleterious/pathogenic alleles). These include the three alleles (p.G606S, p.H469Y, p.P291S), which were not functionally damaging *in vitro*. Of n=31 total overlapping alleles, n=19 were categorised as pathogenic/likely pathogenic in the UK MODY registry: 9 out of 19 matched clustering patterns, whereas clinical classification for remaining 10 appeared to be discordant with *in vitro* function (3 of the 10 are the alleles which were found under both classifications).

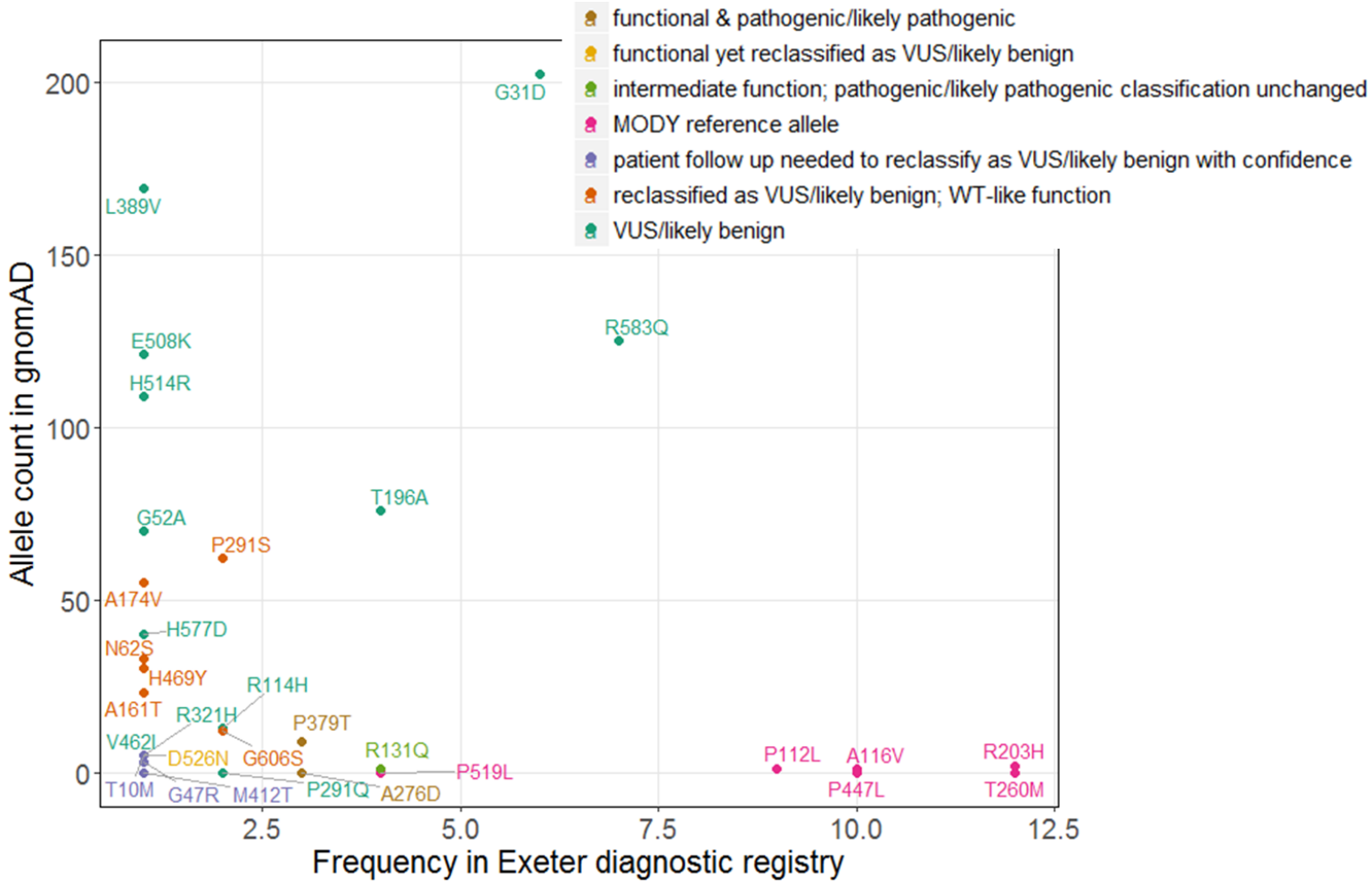


Figure 3.16. Distribution of functionally annotated *HNF1A* missense alleles as a function of frequency in the UK MODY diagnostic registry on the x axis and reported frequency in the genome aggregation database (gnomAD) on the y axis. Alleles are coloured on the basis of the (re)classification scheme on the top right.

Table 3.4. Clinical features associated with discordant *HNF1A* alleles (n=10)

Allele	Age at dx	BMI (kg/m²)	HbA1c (%)	Treatment	Ethnicity	FH	Other	Reclassified as VUS/likely benign
p.A161T	N/A	N/A	N/A	N/A	Arab	N/A	-	Yes
p.A174V	25	N/A	12	initially insulin now diet controlled	White	Mother with diabetes	Pre-NGS panel of MODY genes (only *HNF1A* sequenced)	Yes
p.G47R	38	27	N/A	SU sensitive	White	Child inherited allele has diabetes, islet auto-antibody negative	Pre- NGS panel of MODY genes (only *HNF1A* sequenced)	No
p.G606S	32	24	8.6	OHA and insulin	Middle Eastern	Father with diabetes	-	Yes
p.H469Y	13	21	N/A	Diet and insulin	White	3 generations with diabetes (mother with diabetes)	co-expressed with *HNF1A* p.G207D	Yes
p.M412T	12	36	6	Diet controlled and OHA	N/A	3 generations with diabetes, (mother with diabetes)	Pre-NGS panel of MODY genes (only HNF1A sequenced); twin without variant: BMI 23, OGTT consistent with GCK-MODY	No
p.N62S	36	33	9.6	OHA	White (British)	Mother and 4 siblings with diabetes	retinopathy, nephropathy; Pre-NGS panel of MODY genes (only *HNF1A* sequenced)	Yes
p.P291S	42	29	7.5	initially insulin now OHA	White	Father with diabetes	-	Yes
p.R131Q (4 cases)	13-21	15-20	5.5-7.5	Some diet controlled, some insulin	White (British)	> 2 generations with diabetes	-	No
p.T10M	27	22	N/A	insulin	Chinese	N/A	Pre-NGS panel of MODY genes (only *HNF1A* sequenced); alpha thalassemia	No

Clinical features and additional information available on each the listed alleles from the UK MODY diagnostic registry. With the exception of p.R131Q (n=4 documented cases; information displayed as range across the 4), details are associated with a single case. Abbreviations: NGS = next generation sequencing; dx = diagnosis; SU = sulphonylurea; FH = family history; BMI = body mass index; HbA1c = glycated haemoglobin.

(ii) Clinical reclassification of HNF1A missense alleles

Decisions on allele reclassification were primarily motivated by the juxtaposition of allele frequency values in the general population (based on gnomAD allele counts) against their frequency in Exeter MODY diagnostic registry (highest frequency values belonged to bona fide loss-of-function alleles used as MODY reference controls in this study) **(Figure 3.16)**. Information from other layers of variant annotation such as *in vitro* function (in the tested assays), clinical features, family history, ethnicity, *in silico* prediction all helped to support reclassification decisions. It is important to note that stratification of the functionally characterised *HNF1A* missense alleles by unsupervised clustering analysis is what ultimately enabled this rigorous mapping effort **(Figures 3.15, 3.17 and 3.18)**.

The missense allele p.N62S (gnomAD allele count =33) was consistently dissimilar to dysfunctional alleles in dendrograms and k-means derived clusters from both datasets. At the highest cluster fusion (between height= 15-20), where all alleles are merged into two large yet distinct groups, p.N62S shared a cluster with WT, the known T2D risk modifying allele p.A98V, and another allele shared between the two centres and classified as VUS/likely benign in the registry, p.R583Q (gnomAD allele count = 125) **(Figures 3.10B, 3.13B, and 3.16)**.

Clinical phenotypic data from the MODY registry shows that p.N62S was identified in an obese individual who was diagnosed with diabetes at age 36y **(Table 3.4)**. The patient suffered from microvascular complications (nephropathy and retinopathy) **(Table 3.4)**. These features are not inconsistent with HNF1A-MODY nor with a T2D phenotype (perhaps familial considering number of affected individuals in the carrier's pedigree). *HNF1A* was the only MODY gene

sequenced in this individual as genetic testing was performed before the advent of the NGS MODY gene panel which is the current diagnostic procedure. Although p.N62S allele frequency values alone (gnomAD allele count=33 and detected once in Exeter MODY diagnostic clinic) are sufficient to confidently re-categorise the variant as VUS/likely benign in the context of MODY.

The variants p.A174V and p.P291S, characterised by teams at both Oxford and Bergen, were more difficult to interpret. In the Oxford dataset, these alleles formed a separate outlying k-means-derived cluster **(Figure 3.10A)**. They also occupied an independent subgroup in hierarchical clustering which was more similar to the MODY allele-defined cluster than the one defined by WT and other neutral alleles **(Figure 3.10B)**. They displayed WT-like behaviour across all functional parameters (in some, performance was greater than WT) except for TA in HeLa cells in the Oxford dataset. Unusually high luciferase renilla values (internal luciferase reporter gene assay control used for normalisation) were reported for these alleles as well as p.S535R and p.T82M, which have resulted in an exaggerated reduction in TA values for these variants upon normalisation.

Similar clustering patterns were observed for these alleles using Bergen data, with the exception of k-means partitioning trends in PC space (when k=4 these alleles share the same larger group occupied by impaired and MODY reference alleles) **(Figure 3.13)**. Since two MODY reference alleles were excluded from the analysis of the Bergen dataset due to lack of data points (p.A116V and p.P447L), the resolution of the cluster defined by deleterious alleles may be slightly compromised. Although, in HeLa cells (albumin promoter) transcriptional activity was somewhat impaired by p.A174V and p.P291S (~50-60%). In INS-1 cells

(HNF4A-P2 promoter), p.P291S TA was also reduced (46% compared to WT). Otherwise, the alleles performed similar to WT in assays.

Not only are the activity profiles of p.A174V and p.P291S unlike those of pathogenic dysfunctional alleles, but they also occur at much higher frequency in the general population **(Figures 3.4 and 3.16).** In terms of clinical profiles, p.A174V was detected in an individual with diet-controlled diabetes, which was diagnosed at age 25y **(Table 3.4)**. Again, this was prior to the advent of the NGS MODY gene panel, so only *HNF1A* was sequenced, and it was assumed that the identified variant was causal of the phenotype. The p.P291S variant was detected in an overweight individual who was diagnosed with diabetes in at age 42y **(Table 3.4)**. In another unrelated individual p.P291S was co-expressed with p.G31D and both were annotated as VUS/likely benign in the diagnostic database.

The clinical diagnostic classifications of p.G606S and p.A161T did not match clustering patterns in multivariate space. The alleles did not impact HNF-1a function in the *in vitro* assays tested. In clustering analysis, p.A161T was closest and most similar to p.N62S which was also reclassified as VUS/likely benign, and p.G606S occupied the closest linked subgroup to WT **(Figure 3.10)**. Ultimately, the frequency values associated with these alleles are inconsistent (n=12 alleles in gnomAD and n=2 in the Exeter diagnostic clinic) with those of rare disease-causing (MODY) alleles. The same allele (p.G606S) was detected in *HNF1A* and classified as VUS/likely benign in a single case of hyperinsulinemic hypoglycaemia (on diazoxide treatment) **(Figure 3.18)**.

Despite alignment between clustering pattern and clinical diagnostic interpretation, p.D526N was reclassified from pathogenic/likely pathogenic to

VUS/likely benign. In Oxford TA assays HNF1A-p.D526N was the most impaired of all tested exome-identified alleles (~50% in HeLa and ~80% in INS-1 cells). For perspective, MODY alleles in Oxford assays, exhibited TA range of 20-40% in HeLa cells and 30-50% in INS-1 cells (with the exception of p.R203H and p.A116V which yielded TA values of ~60% and ~100% in HeLa and INS-1 assays, respectively). The allele count of p.D526N in gnomAD is low (observed 5 times) **(Figures 3.4 and 3.16)**, which is consistent with poor *in vitro* function, though plenty of the alleles which were neutral/benign in the tested series of assays have either lower allele counts than p.D526N or are not present at all in gnomAD (e.g. p.L565V, p.R321C, p.L438P, p.P291Q).

Upon revisiting clinical reports associated with the p.D526N case, the clinical picture did not appear to be consistent with HNF1A-MODY, besides presence of diabetes in 3 generations of the carrier's family. The allele carrier was obese (BMI 32.4) and diagnosed with diabetes at age 33y. Other clinical features included dyslipidaemia, polycystic ovary syndrome (PCOS), insulin resistance, and hypertension. Altogether, this suggests that the carrier might have a combination of T2D and HNF1A-MODY. Further, the variant is expressed in the hepatocyte-dominant isoform and is thus unlikely to manifest in a strong beta-cell phenotype despite functionality. For this reason, I integrated isoform weights into the unsupervised clustering model, which separated bona fide loss-of-function MODY alleles from functional alleles expressed in exons 8-10 at the lowest level of hierarchical clustering **(Figure 3.6).**

In k-means-derived clusters and hierarchical organisation, p.H469Y occupied either the same or highly similar (near) subgroups as WT (L27 and I27) as well as known T2D risk modifiers (p.E508K) **(Figure 3.13B).** The clinical features of

the allele carrier described in **Table 3.4** are consistent with severe young-onset familial diabetes. Although it is worth noting that in the same individual, another allele in *HNF1A*, p.G207D, was detected. This allele was identified in a total of 3 cases in the Exeter MODY clinic (including co-occurrence with p.H469Y) and was classified as pathogenic/likely pathogenic each time. Thus, it is likely that p.G207D is the MODY-causal allele and that p.H469Y is either neutral or potentially T2D risk-modifying.

The reported frequency of p.H469Y in the gnomAD database is lower than that of known risk modifiers but also much higher than that of ultra-rare MODY causal alleles **(Figures 3.4 and 3.16).** It was also reported twice VUS/likely benign in the diagnostic registry at Exeter (no information on the presumed deleterious MODY allele with which it was co-expressed in these two cases).

It is possible that some of the noise in functional-clinical mapping **(Figures 3.15 and 3.17)** is a reflection of the heterogeneity in the clinical phenotypic manifestation of *HNF1A* alleles (Flannick et al, 2013). Indeed, the same variant in a MODY gene can give rise to a spectrum of phenotypes depending on genomic and environmental context. It would thus be naive to attempt to draw conclusions regarding variant effects *in vivo* from, for example, a single registry observation. The degree of clinical phenotypic severity or the clinical expression of alleles may be influenced by the impact of regulatory non-coding alleles known to influence *HNF1A* expression and T2D risk, and whether they are in *cis* or *trans* relative to these expression modifying alleles.

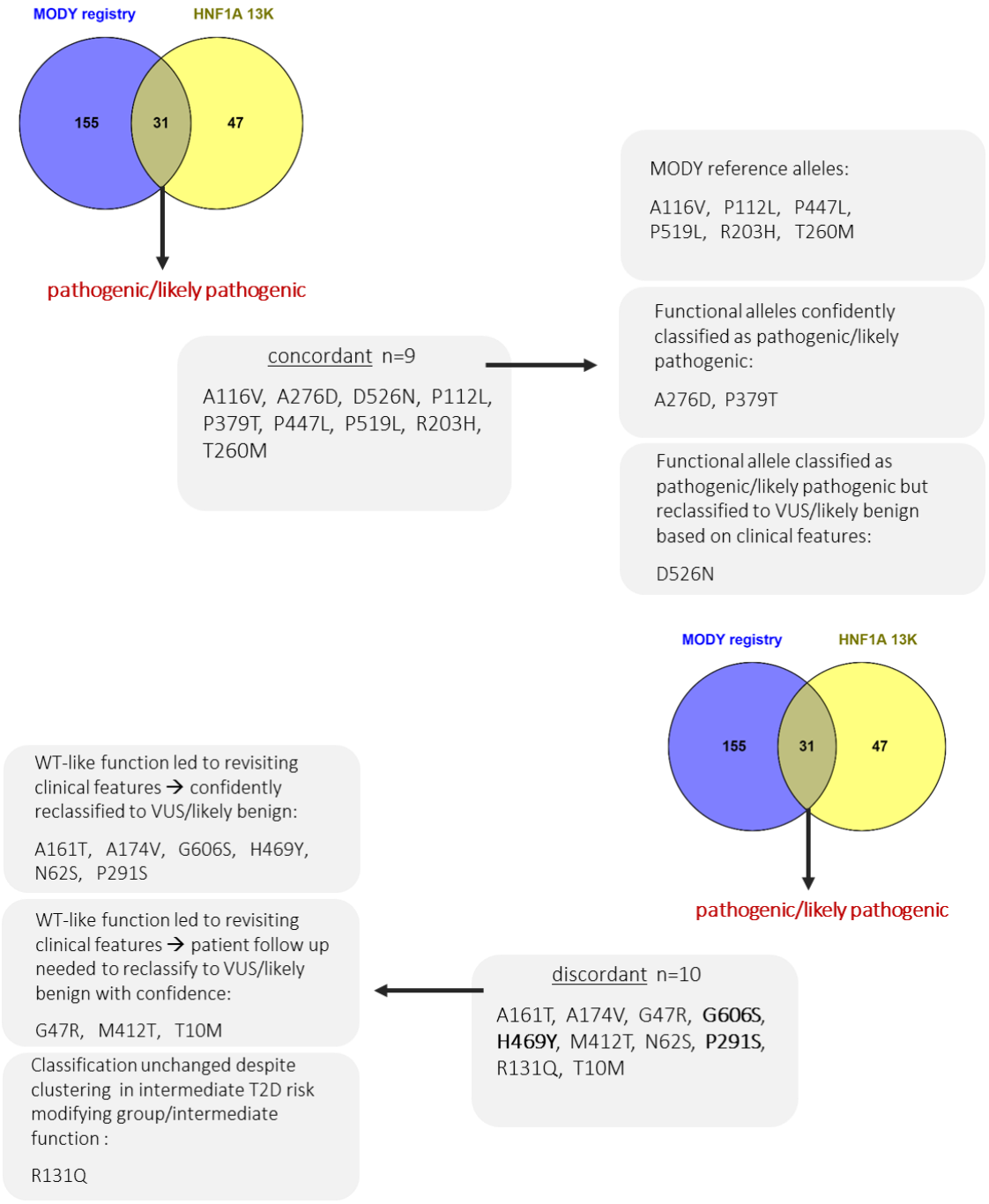


Figure 3.17. (Re)classification informed by unsupervised organisation of functionally characterised *HNF1A* missense from Oxford-Bergen annotated as pathogenic/likely pathogenic in the UK MODY diagnostic registry. Concordant **(top)** and discordant **(bottom)** function-clinical variant classifications.

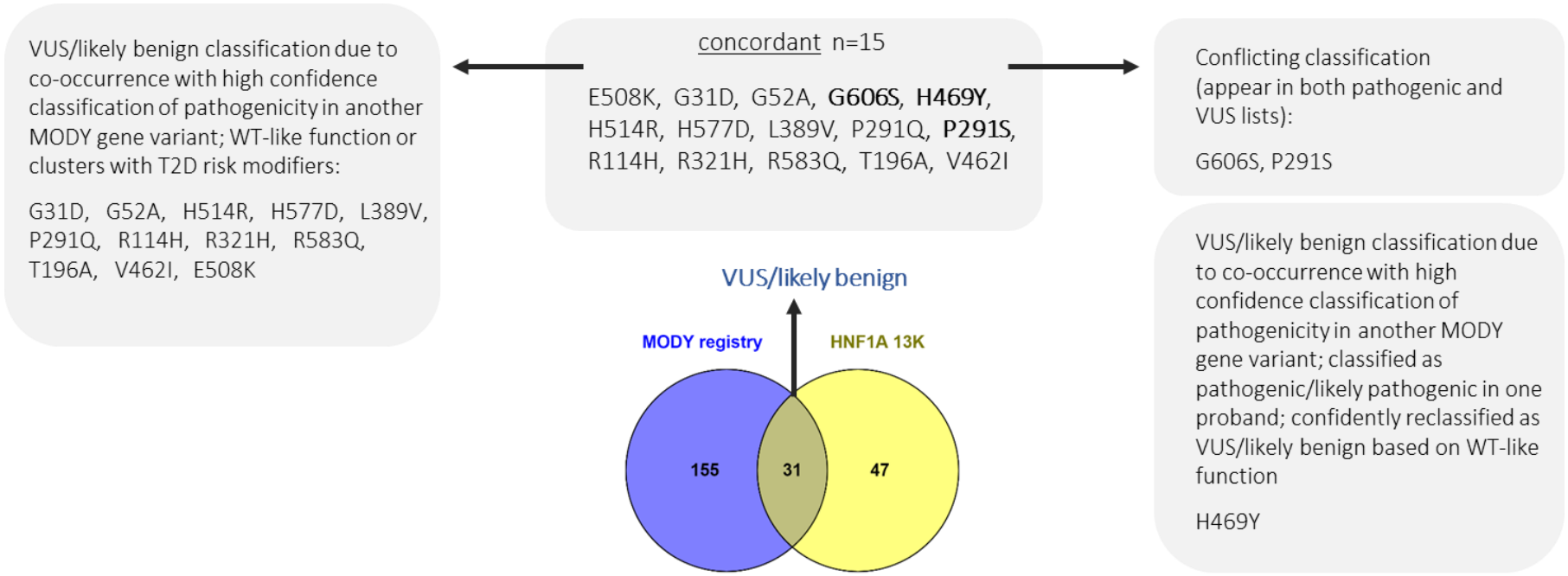


Figure 3.18. *HNF1A* missense alleles which demonstrated concordance between VUS/likely benign classification in the UK MODY diagnostic registry and non-deleterious *in vitro* function in Oxford-Bergen molecular characterisation assessments.

3.3.6 Transactivation as the most informative functional parameter

One observation which remained highly consistent across analyses was the contribution of transcriptional activity to data structure and variance. Upon dissection of the individual principal components, it became evident that transactivation was by far the most impactful functional parameter on the orthogonal transformation of data points in multivariate space. It was the primary contributor to the structural organisation of multi-parametric data. This suggests that it might be a superior functional readout and potentially more informative than other molecular mechanisms for assessing *HNF1A* variant pathogenicity.

In the Oxford-Bergen dataset, it was represented by a greater number of data points (replicates) per allele, which raises the concern of potentially overfitting the model to transcriptional activity data. To address this, I replicated the analysis with a reduced number of data points per TA assay (n=1 biological replicate comprised of n=3 technical replicates; 3 data points in total) using the Oxford variant set. So, overall, each allele had 6 data points for TA and 7 other layers of data (3 for PE, 2 for NucLoc, and 2 assigned weights). I repeated the analysis 3 times, each with a single biological replicate per TA assay. Despite reduction in levels of TA data, transcriptional activity still emerged as the key determinant of data structure **(Figure 3.19).**

Transactivation is a relatively all-encompassing measure of transcription factor protein function. Activating transcription requires transcription factor protein expression, transport into the nucleus, and recognition and binding to DNA sequences. Therefore, defects in transcriptional activity presumably capture defects in any of these biochemical prerequisites. While that may have been the

case for the majority of functionally interrogated *HNF1A* alleles, p.R203H and p.A116V from the Oxford dataset highlight the limitations of this generalisation.

For these alleles, DNA binding ability, a functional parameter excluded from the analysis since it was only tested for a small fraction of alleles, was not adequately captured by transactivation. In the Oxford dataset, p.R203H and p.A116V exhibited WT-like levels of transcriptional activity (up to 100%) in INS-1 cells (HNF-4a-P2 promoter) and 50-60% activity in HeLa cells (rat-albumin promoter). However, in DNA binding assays, these alleles were the most deleterious of all tested MODY controls (~10% compared to WT). The same alleles were consistently deleterious in both luciferase reporter gene assays and interrogations of DNA binding ability in the Bergen dataset.

The consistency and reliability of transactivation data as measured by luciferase reporter gene assays is largely influenced by tissue type and promoter architecture. Further, the protocols for functional assays used to characterise *HNF1A* alleles at Oxford and Bergen varied slightly, and were implemented by several different individuals over the course of the project, which might explain the discrepancy in TA data in the case of p.R203H and p.A116V, and also underlie some of the more subtle differences in *in vitro* function of shared alleles.

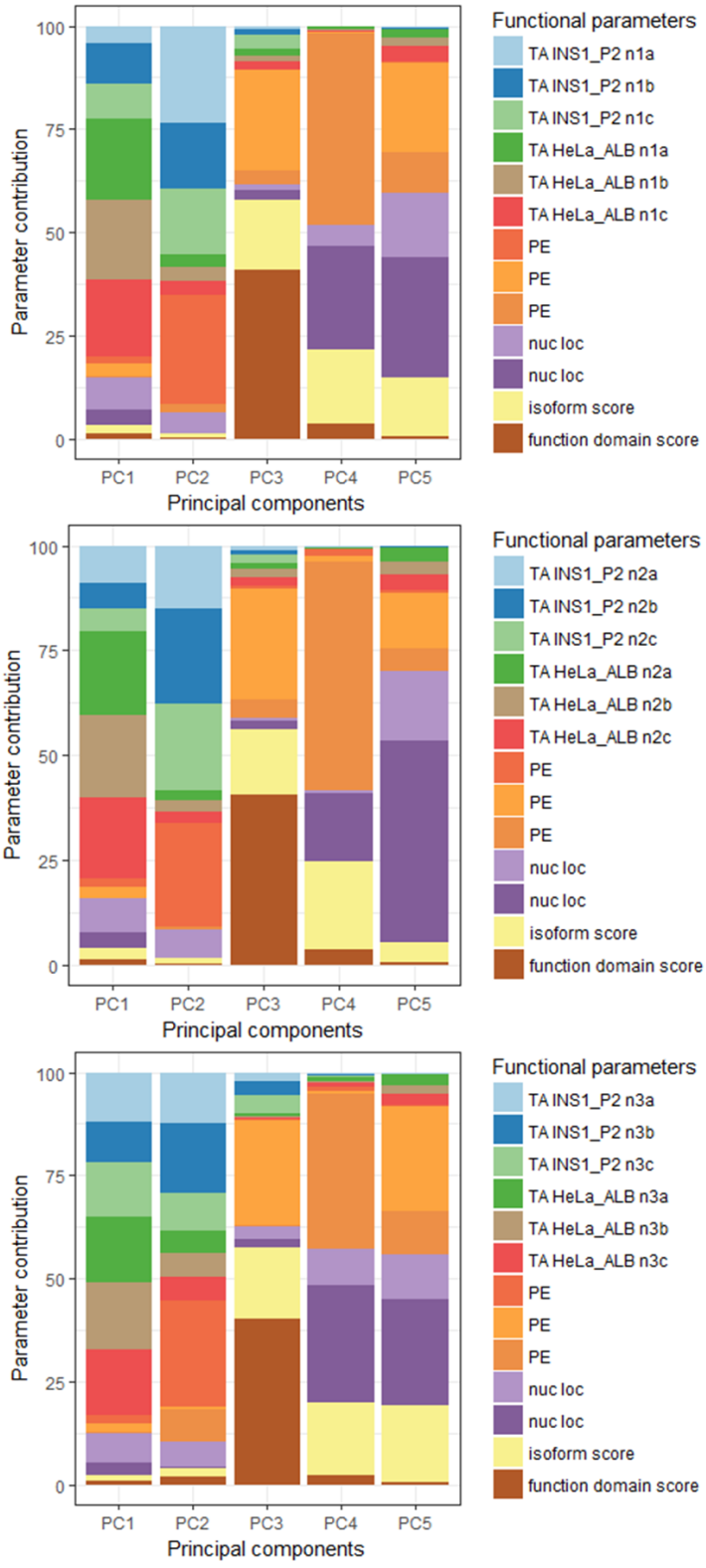


Figure 3.19. PC loadings (eigenvectors) from PCA of Oxford allele set with reduced number of transactivation assay replicates (n1-n3) per iteration. TA INS1_P2 and TA HeLa_ALB are transcriptional activity data from INS-1 cells using HNF4A P2 promoter and from HeLa cells using rat albumin promoter, respectively. PE = protein expression; nuc loc = nuclear localisation data.

3.4 Discussion

3.4.1 Summary

My objective was to process a large multi-parametric *in vitro* functional dataset from 79 *HNF1A* missense alleles (n=73 detected in the exomes of ~13K T2D cases and controls and n=6 which served as deleterious MODY reference alleles). More broadly, the goal was to develop an analytical strategy which effectively utilises multiple layers of functional data per allele for unbiased stratification *HNF1A* missense alleles along the spectrum of *in vivo* glycaemic phenotypes. I have developed, validated and demonstrated the clinical diagnostic utility of such a model for *HNF1A* **(Figure 3.3).**

A combination of dimension reduction and clustering techniques enabled robust organisation and visualisation of exome-identified *HNF1A* alleles in multivariate space. The allele clusters were refined by the integration of scores for tissue-dependant isoform expression and functional domain. For both the Oxford and Bergen variant set, the pipeline captured the spectrum of *HNF1A* molecular dysfunction ranging from neutral/benign effects to modification of multifactorial polygenic diabetes risk to deleterious and causal for early-onset SU-responsive diabetes **(Figures 3.10 and 3.13).**

To supplement the functional clusters with clinical knowledge and to examine their medical diagnostic utility, I explored the UK MODY diagnostic registry (Exeter) for overlapping alleles. Of the 31 alleles present both in the diagnostic database and in this study, myself and collaborator Mr. Kevin Colclough changed the classification of a total of 7 alleles from pathogenic/likely pathogenic to VUS/likely benign **(Figure 3.17).** The reclassification efforts were informed

primarily by allele frequency in the general population **(Figures 3.4 and 3.16)** and clinical features from MODY registry **(Table 3.4)**.

3.4.2 Utility of functionally and clinically annotated allele clusters

Gold standard functional assays are laborious, slow and low in throughout: of all the variant reported in the three major MODY genes, *HNF1A*, *GCK* and *HNF4A*, only ~12% have been functionally annotated (Osbak et al, 2009; Colclough et al, 2013). The recent advent of Multiplexed Assays of Variant Effects (MAVEs) has made it possible to interrogate the function of every possible nucleotide substitution in a single experimental system (Starita et al, 2017).

A meticulously designed MAVE for *HNF1A* would enable functional annotation of all possible missense alleles (>12K) in single *in vitro* assay, which corresponds to roughly 150-fold increase scale compared to the Oxford-Bergen effort. Data from the unsupervised clustering analysis support an *HNF1A* MAVE on multiple levels:

1. As a technology, MAVEs remain in early stages of development and optimisation, and are thus limited to a single functional readout/dimension per assay. PCA of multi-parametric functional data from ~80 exome-detected has revealed transactivation as the most informative mechanistic dimension. As such, it would seem sensible to prioritise building a transactivation-centric MAVE to screen an *HNF1A* saturation mutagenesis library.

2. The comprehensively annotated Oxford-Bergen allele clusters can serve as a powerful training set for supervised classification models built with high-throughput experimental data.

4

Building a Multiplexed Assay of Variant Effects (MAVE) to screen all possible *HNF1A* missense alleles

4.1 Background

4.1.1. The challenge in establishing clinical significance of *HNF1A* sequence variants

The accurate interpretation of the phenotypic consequences of genetic variation remains a central challenge in medical genomics. Understanding the functional impact of variants in known disease genes can significantly enhance interpretive resolution in the diagnostic sequencing setting. For variants in actionable disease genes, where a molecular diagnosis immediately impacts treatment, management and/or prognosis, a catalogue of annotated sequence-function relationships is especially valuable.

Hepatocyte nuclear factor 1-alpha (*HNF1A*), a clinically actionable gene involved in the pathogenesis of diabetes mellitus, exemplifies the challenge in diagnostic interpretation (Tattersall 1974; Yamagata et al, 1996; Ellard et al, 2008; Colclough et al, 2014). Haploinsufficiency (loss-of-function, most commonly through damaging missense alterations) in *HNF1A* can give rise to early-onset monogenic diabetes (MODY-HNF1A; MIM #600496) characterised by Mendelian autosomal dominant inheritance and progressive beta-cell dysfunction (Yamagata et al, 1996; Colclough et al 2013; Tattersall 1974; Frayling et al, 2001). Individuals with HNF1A-MODY are highly responsive to treatment with oral hypoglycaemic agent sulphonylurea (SU) and can avoid insulin injections until late in life (Pearson et al, 2000; Shepherd et al, 2009).

Large-scale genetic sequencing studies without ascertainment bias for extreme phenotypes (monogenic diabetes) have exposed:

- rare, protein-altering *HNF1A* alleles (MAF < 1%) interpreted as MODY-causal in individuals who remained normoglycaemic through middle age (p.H469Y, p.V103M, p.R131Q, p.T515K, p.H505N, p.E48K, p.R583Q, p.R114H, p.T196A, p.A276D) (Flannick et al, 2013; Najmi et al, 2017).
- common (p.A98V, p.I27L) and low frequency (p.E508K, p.G391S) missense alleles that elevate T2D risk or associate with early-onset T2D in population-specific contexts (Consortium STD, 2014; Yang et al 2016; Gaulton et al 2015; Mahajan et al, 2016).
- common intergenic (chr12:121440833:D, OR 1.06, P= 2.9×10^{-10}) and noncoding (intronic lead GWAS SNP rs1242733, OR 1.07, P= $6.5x10^{-8}$) GWAS SNPs near and at the *HNF1A* locus robustly associated with T2D in European ancestry populations (Voight et al, 2010; Morris et al, 2012).

Early identification of T2D risk alleles can help better understand individual T2D predisposition and in turn prompt lifestyle interventions which can reduce risk of developing the disease later in life (McCarthy 2015).

Over 400 mutations (missense, frameshift, splicing, nonsense, and promoter mutations, and in-frame insertions, deletions and duplications) have been reported in *HNF1A*, with highest mutation rates in dimerization (>60% MODY-associated alleles) and DNA binding (45-55%) domains (Consortium STD, 2014; Colclough et al, 2013). Reliable cellular and molecular phenotypes are only available for a small subset of these variants as gold standard functional characterisation methods remain low in throughput, and are both resource and time intensive.

Data from functional characterisation efforts support that the degree of HNF-1a disruption tends to correlate positively with phenotypic severity: MODY-causing protein-altering alleles impair HNF-1a transcriptional activity more severely (≤30% compared to wild-type) than *HNF1A* variants associated with increased risk for developing T2D in population-specific contexts (40-60% that of WT transcriptional activity) (Najmi et al, 2017; Consortium STD, 2014; Balamurugan et al, 2016; **results presented in Chapter 3**). Rare variants which demonstrated intermediate function (between MODY-casual and WT) in transactivation and nuclear localisation assays were associated with a 6-fold increase in T2D predisposition (Flannick et al, 2016; Najmi et al, 2017).

A couple of studies have reported testing the effects of *HNF1A* sequence variants in the context of T2D risk-modifying SNPs or other common alleles. While these assessments have enabled an understanding of the functional consequences of allele co-expression, the *in vitro* experimental design is limited to an interrogation of in *cis* effects. One showed that *HNF1A* harbouring a combination of the common variant p.S487N and the loss-of-function allele p.R171Q was more deleterious in transactivation and nuclear localisation assays compared to p.R171Q alone (Balamurugan et al, 2016). Another study showed that a combination of L27 and V98 showed mildly reduced transcriptional activity (~50% combined vs >60% individually) as measured by GLUT-2 promoter in HeLa and INS-1 cells (Holmkvist et al, 2006).

The ubiquity of genetic sequencing means that more novel and incidentally detected variants of uncertain clinical significance (VUS) are bound to be detected in individuals with less extreme phenotypes (Cooper 2015). Retrospective functional investigations are too slow compared to the pace of

sequence-data generation; at-scale classifications are currently limited to those provided by computational tools, which perform poorly and often conflict (Miosge et al, 2015; Grimm et al, 2015; Flannick et al, 2013).

So, what does it mean when an individual without a classic MODY phenotype and no family history, for example, presents with a novel *HNF1A* variant, or a known allele without an annotated sequence-function relationship or one that is perhaps misclassified? How can we optimise variant-interpretation with the speed, scale, and accuracy required to evaluate clinical significance as next-generation sequencing becomes routine?

4.1.2. Towards molecular diagnostic certainty: large-scale prospective functional annotation

The recent advent of Multiplexed (functional) Assays of Variant Effects (MAVEs) has made it possible to interrogate the biochemical effects of 10^4-10^6 single nucleotide variants (SNVs) at once in a single experimental system using saturation mutagenesis oligonucleotide libraries (Starita et al, 2017). MAVEs provide a unique opportunity to prospectively distinguish disruptive, functional, and potentially pathogenic alleles from the vast majority of neutral ones rapidly and at-scale (Gasperini et al, 2016; Starita et al, 2017).

The general workflow involves tile array-based library synthesis, delivery of the library to an experimental system or model, stratification of alleles on the basis of one or more functional signatures, next generation sequencing to count the occurrence of alleles in each functional category or ‘bin’ and generation of calibrated function scores from read counts using robust statistical frameworks **(Figure 4.1)** (Gasperini et al 2016; Starita et al, 2017). The true clinical diagnostic

utility of MAVEs rests upon the quality and accuracy of experimentally-derived MAVE-based functional annotations. These molecular measurements are used to build and train predictive models which translate the *in vitro* data into likelihood estimates of variant pathogenicity (Gasperini et al, 2016; Starita et al, 2017).

The utility of MAVEs has been recently demonstrated with medically actionable disease genes, *BRCA1* (MIM #113705) and *PPARG* (MIM #601487). All possible missense alleles in the RING domain of *BRCA1* (2K), where most pathogenic mutations occur (58%), were assayed in parallel and scored on the basis of two distinct functional parameters (Starita et al, 2015). A saturation mutagenesis library encoding all possible single amino acid substitutions in *PPARG* (~10K) was interrogated on the basis of transcriptional activity in a high resolution *in vitro* endogenous reporter gene assay (PPARG-responsive CD36 expression on the surface of the human monocytic cell line THP-1) (Majithia et al, 2016). *PPARG* MAVE-derived function scores were used to build a *PPARG* variant classifier by supervised machine learning (Majithia et al, 2016). Majithia and colleagues have made their *PPARG* MAVE-based pathogenicity assignments publically available through a user-friendly web interface (MITER, Missense InTerpretation by Experimental Response http://miter.broadinstitute.org/).

A MAVE designed to stratify all possible single amino acid substitutions in *HNF1A* on the basis of molecular function would provide tremendous diagnostic clarity. The amount and diversity of existing *HNF1A* genomic, functional and clinical datasets can be efficiently utilised to calibrate MAVE-derived function scores, and in turn yield accurate, high performance classification models, presented in the form of an *HNF1A* 'lookup table' (**Figure 4.2**). Upon the identification of an allele in *HNF1A*, clinicians and researchers can consult the annotated lookup table to

make initial judgements of variant effects, informed by the context of the variant carrier(s) (**Figure 4.2**).

4.1.3 Chapter preview

Proper assay design is crucial to obtaining clinically informative and actionable MAVE-based functional annotations. The MAVE community would benefit from a repository of rigorous and well-validated methodological pipelines that are generalizable on the basis of desired functional signature, rather than highly customised to the gene of interest. I aimed to establish a systematic and generalizable pipeline for building a high fidelity MAVE to screen an *HNF1A* variant library harbouring all possible missense alleles using transactivation as a functional readout.

I employed two distinct strategies for building a MAVE for *HNF1A*, both of which utilise transactivation as a molecular signature **(Figure 4.3).** One strategy relies on the fluorescent signal of an artificial reporter gene **(subsection 4.3.1)** and the other on fluorescent signal from a stained cell surface protein encoded by an endogenous *HNF1A* target transcript **(subsection 4.3.2)**. The former seemed an attractive approach to begin with because artificial reporter assays are considered the gold standard for functional investigations of transcription factors *in vitro*. The latter more closely resembles the successfully implemented *PPARG* MAVE (described in **Chapter 1 and subsection 4.1.2**) though is much more technically complex because it requires identification and multi-step validation of robust endogenous target transcripts in a biologically relevant cellular model **(Figure 4.3)**. The key challenge with both strategies is achieving sufficiently high

signal-to-noise and sufficiently wide dynamic range to allow for pooled assessments of a multiplexed oligonucleotide variant library **(Figure 4.1)**.

For this project, I collaborated with Dr. Amit Majithia at the Broad Institute at MIT and Harvard. Majithia and colleagues at the Broad Institute published the *PPARG* MAVE in October 2016. The transactivation-centric MAVE schematic portrayed in **Figure 4.1** is inspired by the *PPARG* MAVE design. A year into the project, Laeya Najmi (a PhD student supervised by Dr. Lise Bjørkhaug and Professor Pal Njølstad at the University of Bergen) joined the *HNF1A* MAVE-building collaboration. To ensure productive synergy of efforts and avoid replication, as a team, we decided to:

1. Invest in my endogenous reporter gene strategy for an *HNF1A* MAVE **(Figure 4.3)** based on the results I gathered from a year of optimising an artificial reporter gene system, which ultimately failed to yield the signal intensity required for high-throughput multiplexed applications **(subsection 4.3.1)**.

2. Work with different cell lines. Laeya Najmi (based between Broad and Bergen) was to adopt the experimental pipeline I designed for identifying MAVE-compatible endogenous target transcripts (**Figure 4.3**), but would do so in a different cell line than the one I had already selected and characterised by the time she joined the collaboration.

3. Utilise different *HNF1A* target-staining techniques. I relied on fluorescent antibody staining of cell surface-expressed *HNF1A* targets **(subsection 4.3.2)**. Najmi and Majithia decided they would pursue an *in situ* RNA-FISH method to enable multiplexed transcript-level readout (not yet tested).

4. Screen two different *HNF1A* saturation mutagenesis libraries. We agreed that two independent libraries harbouring all possible missense alleles would be synthesised on two distinct genetic haplotypes: a) I27L (rs1169288) common polymorphism shown to moderately elevate T2D risk (OR, 1.09; $P = 8.1\times10^{-15}$: MAF = 33%), which I would screen in my cell system using my validated target(s) and staining approach, and b) L27I native wild-type (NM_000545.6) for Najmi and Majithia to screen in their system. Interrogating these libraries in parallel allows us to interrogate independent effects of genetic background on protein function: important contextual information that is largely absent from *in vitro* functional analyses.

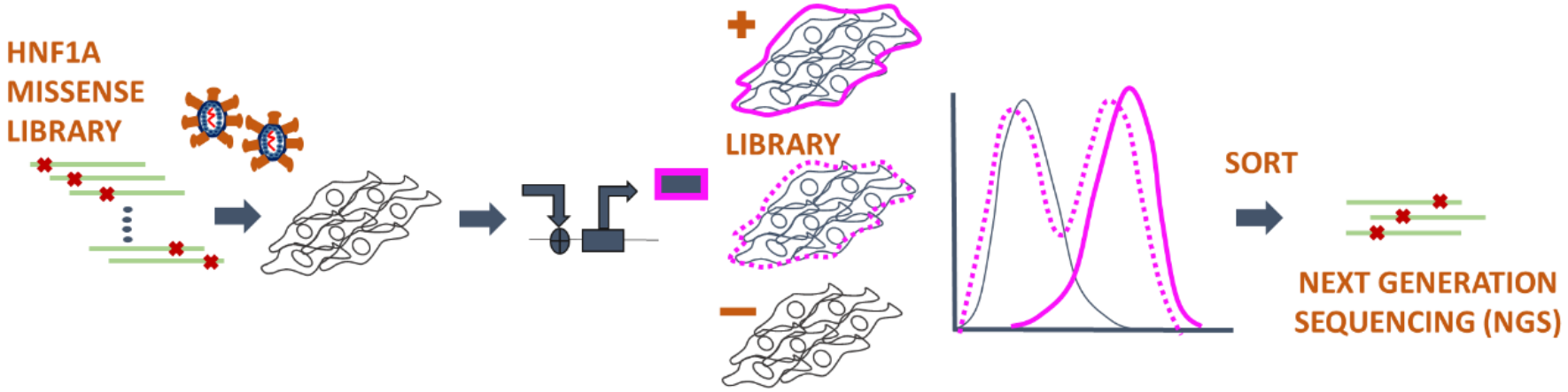


Figure 4.1. Experimental pipeline for high-throughput functional screen of every possible single amino acid substitution in the human *HNF1A* sequence. A multiplexed oligonucleotide library harbouring all possible nonsynonymous missense variants in *HNF1A* is packaged into lentiviral particles which are used to stably integrate the sequences into a relevant cellular system at a rate of a single variant entry per cell. Transduced cells are then sorted on the basis of HNF-1a transcriptional activity, directly reflected by variable fluorescence intensities of induced fluorescent reporters and/or fluorophore conjugated endogenous target transcripts. Genetic material from the sorted cell populations is resolved to nucleotide level using next-generation sequencing technology and computationally processed for variant-phenotype mapping.

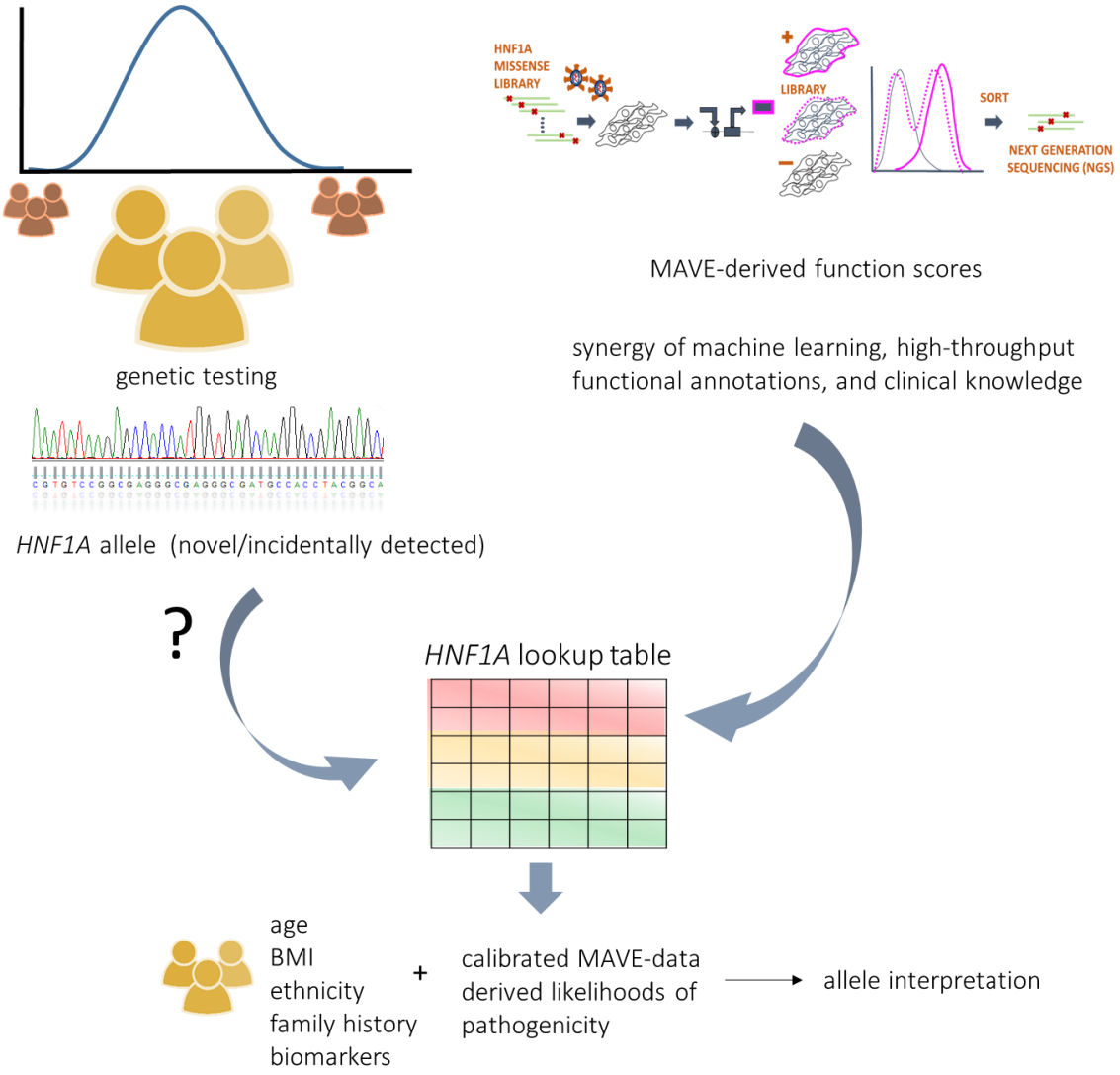


Figure 4.2. The utility of high-throughput experimental data in facilitating clinical diagnostic variant interpretation. With increased next-generation sequencing applications in the clinical setting, known and novel alleles will be ascertained in more individuals without a clear MODY phenotype (i.e. not extremes of a phenotypic distribution). An *HNF1A* MAVE would enable prospective functional annotation of every possible nucleotide substitution based on one or more cellular phenotypes. A high performance variant classifier built using MAVE-based data can then be used to generate an exhaustive catalogue of variant effects which researchers and clinicians can consult upon sequence-identification of an *HNF1A* allele. Multiple layers of clinical, biochemical, bioinformatic and low-throughput *in vitro* molecular data associated with known or previously studied *HNF1A* alleles can be used to refine high-throughput function scores. The lookup table would be comprised of calibrated log-odds likelihoods of allele pathogenicity, which, alongside clinical features, ethnicity, family history, can facilitate accurate molecular diagnosis.

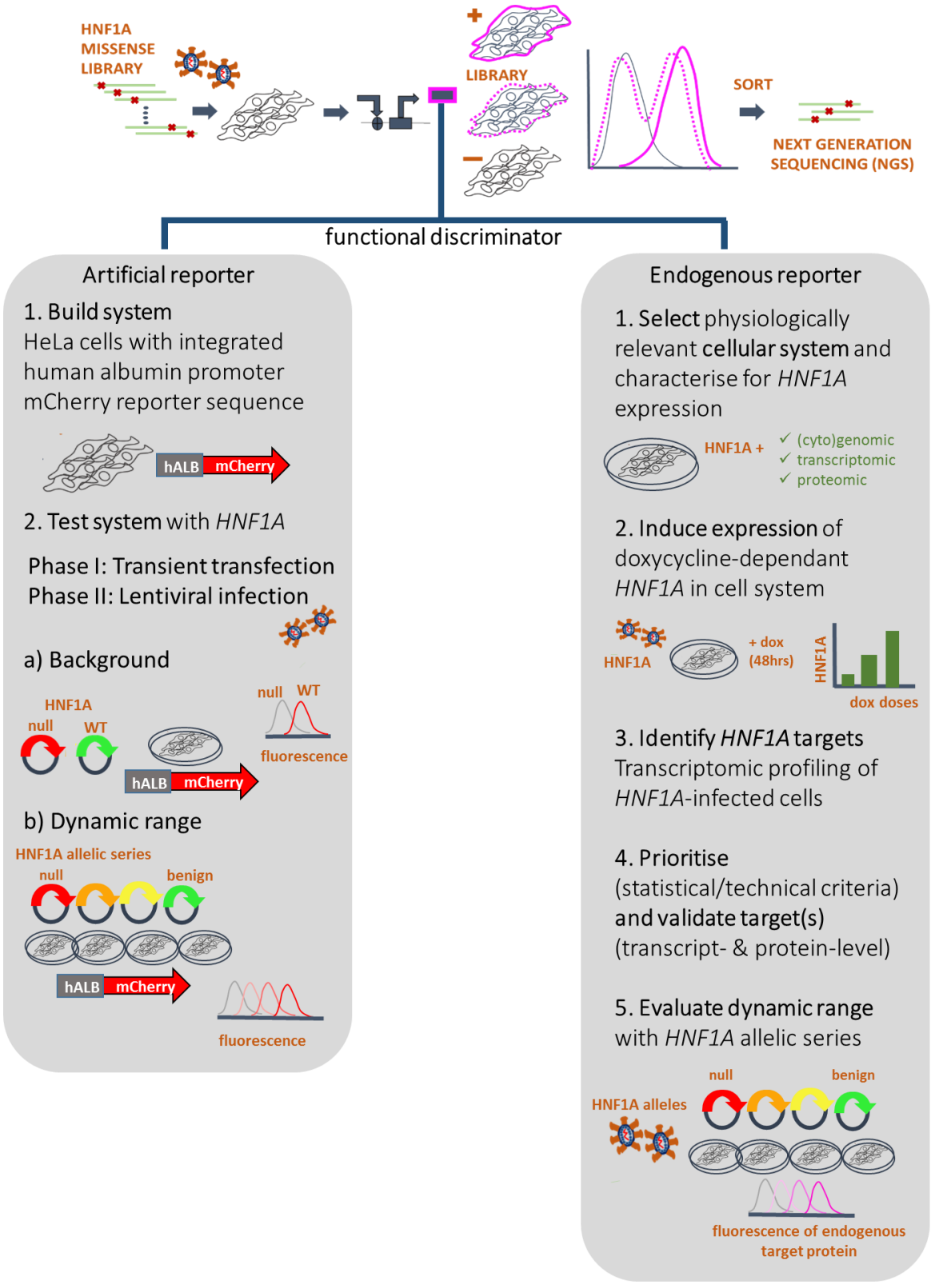


Figure 4.3. Methodological pipeline for building a transactivation-centric MAVE to annotate all possible missense alleles in *HNF1A*. The schematic on top is a blueprint for an *HNF1A* MAVE (see Figure 1 for more details). The assay discriminates multiplexed *HNF1A* alleles based on signal from an HNF-1a transcriptional target, which can either be an artificial reporter gene downstream a relevant promoter sequence in an engineered cell line (left) or an endogenous reporter gene in a physiologically relevant cell system (right).

4.2 Methods

In this section, I describe the specific methods and techniques that I used to build a MAVE for *HNF1A* with transcriptional activity as a functional signature following the pipeline depicted in **Figure 4.3**. I have placed greater emphasis in this section on delineating the technical intricacies of the endogenous reporter strategy **(Figure 4.3)**, library fabrication, and *HNF1A* sequences and expression systems. Details on some of the more general, well-established methods used in this chapter such as cell culture, western blot analysis, PCR, FACS, transfections and lentiviral production/infection/titration can be found in Chapter 2.

4.2.1 *HNF1A* sequences

I used the following *HNF1A* cDNA sequences in independent experiments: a) a sequence homologous to the *HNF1A* reference (NM_000545.6) hereafter referred to as 'native' *HNF1A*, and, b) a sequence which encodes wild-type *HNF1A* protein (same amino acid sequence as native) but is, on the nucleotide level, non-identical (hereafter referred to as 'recoded' *HNF1A*); it is codon-optimised to **(i)** harbour endonuclease recognition sites to facilitate tile array-based synthesis of an *HNF1A* saturation mutagenesis library (Melnikov et al, 2014) **(ii)** confer resistance to *HNF1A* gene editing to enable a delete-and-replace expression strategy if needed in a cell system that endogenously expresses *HNF1A* **(subsection 4.2.6)**, and **(iii)** ensure most efficient mammalian translation, which is partly influenced by codon bias (i.e. codon usage frequency https://www.genscript.com/tools/codon-frequency-table) (Angov, 2011; Ikemura, 1982; Ingolia et al, 2009). Both cDNA sequences were expressed on a genetic background harbouring the common *HNF1A* coding polymorphism p.Ile27Leu.

4.2.2 Plasmids, molecular cloning, and mutagenesis

For *in vitro* mammalian expression of *HNF1A*, I used a couple of plasmid DNA constructs with different expression-control mechanisms:

- For preliminary proof-of-concept experiments in the artificial reporter system **4.3.1(i)**, I used a vector in which native *HNF1A* cDNA was cloned downstream the human cytomegalovirus (CMV) promoter sequence. This promoter architecture enables high-level constitutive expression of the transgene and is commonly used for transient transfections (Xia et al, 2006). See **Figure 4.4A**.

- To achieve titrated delivery, genomic integration, tight control, and high-level inducible expression of *HNF1A* sequences, I cloned native and recoded *HNF1A* cDNA into a tet-ON 2nd generation lentiviral transgene vector via BamHI/EcoRI (NEB). In this construct, *HNF1A* expression is induced by doxycycline: doxycycline binds rtTA (reverse tetracycline-controlled transactivator) which in turn activates the tet response element (TRE) upstream *HNF1A* inducing its expression in a dox dose-responsive manner - in the absence of dox, *HNF1A* is not expressed (Gossen et al, 1995). Second generation lentiviral packaging plasmids were used to generate lentiviral particles using this construct (**see Chapter 2 for details**). **See Figure 4.4B.**

For artificial reporter gene expression:

- I cloned the red fluorescent reporter gene mCherry downstream human albumin promoter sequence (hALB) – a promoter widely used for interrogating HNF-1a transcriptional activity by virtue of its highly

complementary recognition consensus sequence – into a vector which confers hygromycin resistance to enable stable cell line generation **(Figure 4.4C)**.

To evaluate the dynamic range of each reporter gene system, I synthesised *HNF1A* variants that cover the spectrum of functional and clinical phenotypes using the QuikChange Site-Directed XL Mutagenesis Kit (Stratagene). Excessively large plasmids such as PCW-*HNF1A*-puroR (**Figure 4.4B**) are difficult to mutagenise efficiently by PCR. Therefore, I introduced the allelic series variants into *HNF1A* cDNA that was cloned into a smaller vector (pUC57-kan 2579bp, GenScript) using QuikChange Site-Directed XL Mutagenesis Kit (Stratagene) and the mutagenic primers listed in **Table 4.1**. Mutagenesis was verified by Sanger sequencing (Eurofins). I then cloned successfully mutagenised sequences into PCW-puroR (**Figure 4.4B)** via BamHI/EcoRI (NEB) using Quick Ligation™ Kit (NEB cat no. M2200S).

4.2.3 Doxycycline dose induction

I induced *HNF1A* expression in the tet-ON construct (**subsection 4.2.2**; **Figure 4.4B)** by adding various doses of doxycycline to cells in culture. Doxycycline (dox) hyclate (Sigma, product no. D9891) was stored at 4°C and protected from light. For each independent FACS experiment, I dissolved 50 mgs of dox in sterile water in an amber microcentrifuge tube to create a fresh 50 mg/ml solution for use in culture. I exposed cells to a range of dox concentrations (0-5 µg/ml) and re-spiked with dox-supplemented growth media every 48 hours during an incubation period.

Table 4.1. Mutagenic primers used to synthesise *HNF1A* allelic series (recoded sequence) [4.3.2(v)].

Variant	cDNA position recoded	Forward 5'-3'	Reverse 5'-3'	Tm °C
p.T260M	779/780 a**CA** > a**TG**	ctgggctcaaatctggtcatggaggtgcgagtgtacaatt	aattgtacactcgcacctccatgaccagatttgagcccag	79.1
p.P112L	335 c**C**a > c**T**a	ctgcaggaggacctatggcgagtggct	agccactcgccataggtcctcctgcag	78.61
p.P447L	1340 c**C**t > c**T**t	caggcccagtccgtgcttgtgattaacagtatg-	catactgttaatcacaagcacggactgggcctg	79.14
p.R203H	607/608/609 **AGG** > **CAC**	aactaagaaaggaaggcgaaatcactttaaatggggaccagcctcac	gtgaggctggtccccatttaaagtgatttcgccttcctttcttagtt	78.2
p.E508K	1522/1524 **G**a**G** > **A**a**A**	ctgtacagccacaagcctaaagtggcacagtacactcat	atgagtgtactgtgccactttaggcttgtggctgtacag	79.04
p. S535R	1605 ag**T** > ag**A**	gcgccctggctagactgacaccaacca	tggttggtgtcagtctagccagggcgc	78.61
p.A161T	481 **G**cc > **A**cc	agactcagaaacgcgctaccctgtacacttgg	ccaagtgtacagggtagcgcgtttctgagtct	79.06
p.Q466X	1396 **C**ag > **T**ag	ccgtccagttctcctagccactgcatccc	gggatgcagtggctaggagaactggacgg	80.22

Primers were designed using QuikChange® Primer Design Program (Agilent Genomics) and obtained from Integrated DNA Technologies (IDT).
Tm = primer melting temperature.

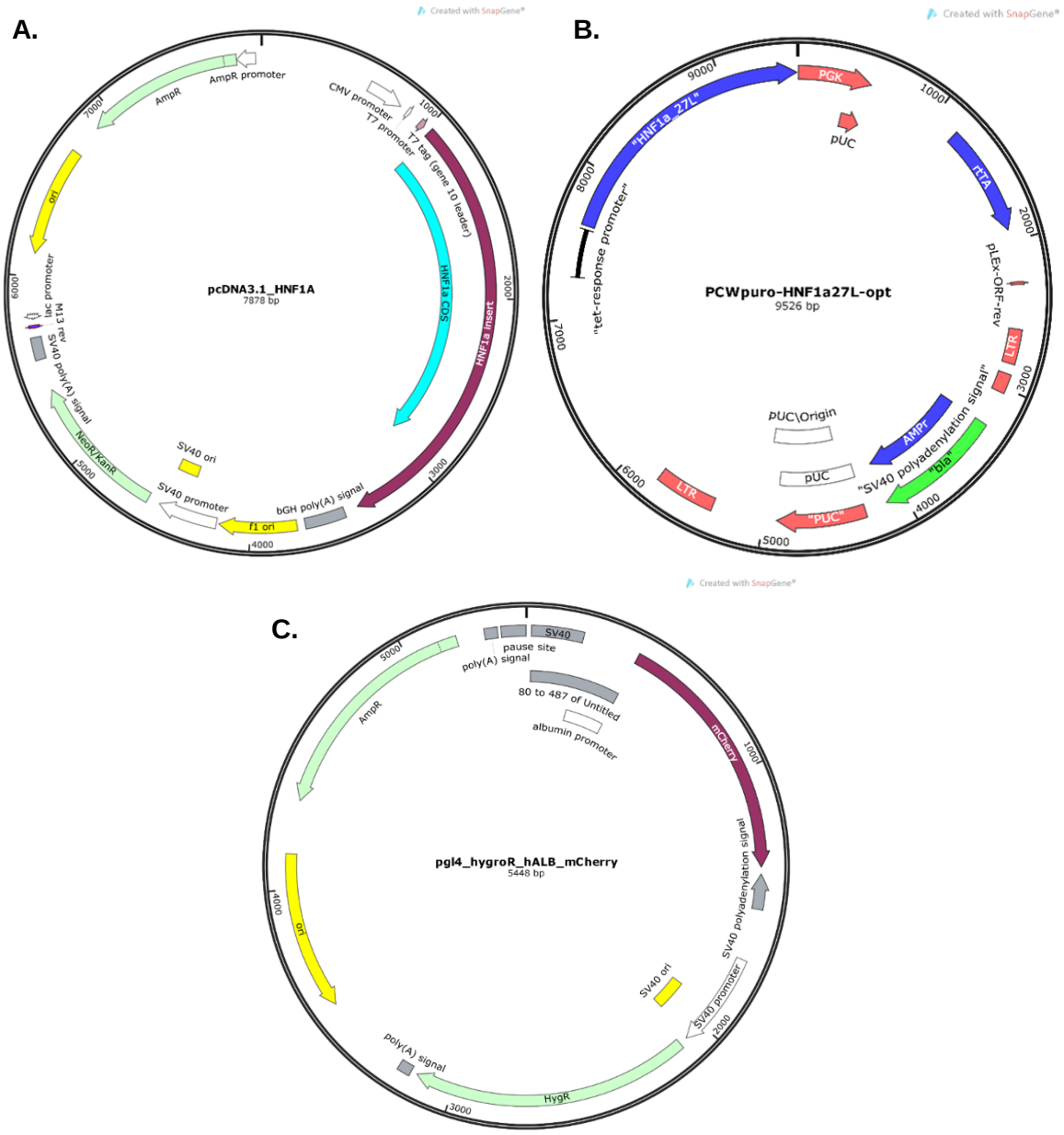


Figure 4.4. Maps of plasmid DNA (pDNA) used in MAVE development experiments. (A) For all transient transfections, I used *HNF1A* cDNA cloned into pcDNA3.1/HisC vector (supplied by the University of Bergen), in which expression is controlled by a constitutive CMV promoter. **(B)** For all lentiviral infections, I used a 2nd generation lentiviral transgene expression vector pCW57.1 Dest (pLXI_TRC401) (https://portals.broadinstitute.org/gpp/public/vector/details?vector=pLXI_TRC401) into which I cloned WT and recoded *HNF1A* cDNA via BamHI/EcoRI (NEB). **(C)** For artificial *HNF1A* reporter gene experiments, I used pgl4.26 [luc2/minP/Hygro] (Promega) into which I cloned human albumin promoter and mCherry (red fluorescent) reporter sequences via KpnI/BamHI (NEB). Maps generated with SnapGene Viewer v2.6.2**.**

4.2.4 RNA Sequencing

RNA sequencing was used to impartially survey HepG2 cells for differentially expressed HNF-1a target transcripts that could serve as endogenous MAVE-compatible molecular markers of *HNF1A* dysfunction **4.3.2(iii) (Figure 4.3)**.

I infected HepG2 cells with lentiviral particles harbouring either native or recoded *HNF1A* (27L) sequences. I then selected positive colonies (successfully transduced cells) with puromycin (2 µg/ml) and exposed them to a range of doxycycline concentrations (0, 0.1, 1, and 5 µg/ml), n=3 biological replicates per concentration for each sequence type. After 48 hours of dox stimulation, I isolated RNA using TRIzol® Reagent (Life Technologies 15596–018). RNA quality (RIN score) was assessed using the RNA 6000 Nano Kit (Agilent 5067–1511) and 2100 Bioanalyser. All samples passed quality control (RIN score =10.0).

Library preparation and sequencing was performed at the Oxford Genomics Centre (Wellcome Trust Centre for Human Genetics, University of Oxford). Polyadenylated transcripts were enriched using TruSeq Stranded mRNA HT Sample Prep Kit (Illumina, RS-122-2103) with custom 12bp indexes. Libraries were multiplexed and sequenced with the TruSeq PE Cluster Generation Kit v3 (Illumina, PE-401-3001) and TruSeq SBS Kit v3 (Illumina, FC401-3001) over 2 lanes of an Illumina HiSeq4000 as 75 nucleotide paired-end reads to a total depth of 20 million reads per sample.

Analysts at the Oxford Genomics Centre used TopHat2 (Kim et al, 2013) to align raw sequencing reads to the human genome reference GRCh37.p13 on default setting using GENCODE version 7 (Ensembl 62) as the transcriptome reference (https://www.broadinstitute.org/cancer/cga/sites/default/files/data/tools/rnaseqc/

gencode.v7.annotation_goodContig.gtf.gz). They also quantified gene level reads counts in a strand-specific manner with featureCounts (Liao et al, 2014) from the Subread package (http://subread.sourceforge.net/).

4.2.5 Differential expression analysis

I conducted differential expression analysis to statistically rank sequenced transcripts on the basis of *HNF1A* regulation to identify and prioritise MAVE-compatible targets **4.3.2(iii) (Figure 4.3)**. I performed the analysis in R version 3.3.1, using limma(voom) (Law et al, 2014) and edgeR (Robinson et al, 2010) on voom-normalised gene counts for all autosomal protein-coding genes that were expressed at > 1 count per million (CPM). I fitted a linear model with a treatment (dox) factor to the data, and determined significance by moderated empirical Bayes *F*-tests. Differentially regulated genes were defined by a false discovery rate (FDR)-adjusted *P* value of 0.01.

4.2.6 Bioinformatic analysis

To gain biological insights into the regulatory role of *HNF1A* in human liver from the endogenous reporter identification effort (**subsection 4.3.3**), I conducted several exploratory computational analyses:

(i) Motif analysis

For top differentially expressed transcripts with little-to-no *a priori* evidence of *HNF1A* regulation, I computationally scanned genomic sequences 5kb upstream their transcription start sites (TSS) for *HNF1A* DNA binding motifs.

I conducted motif analysis using R packages MotifDb, motifStack GenomicFeatures, GenomicRanges, TxDb.Hsapiens.UCSC.hg19.knownGene, and BSgenome.Hsapiens.UCSC.hg19.

I queried MotifDb for position frequency matrices (pfm) of *HNF1A* DNA binding motifs (from public sources accompanied by metadata). The query provided 3 *HNF1A* pfms, two identical ones from JASPAR_CORE and JASPAR_2014 (Vertebrata-JASPAR_CORE-HNF1A-MA0046.1 and Vertebrata-JASPAR_2014-HNF1A-MA0046.1) and one from jolma2013 (Hsapiens-jolma2013-HNF1A). These pfms were transformed to position count matrices (pcm), and used via the matchPWM function to scan promoter regions (extracted using getPromoterSeq function) upstream TSS annotations in TxDb.Hsapiens.UCSC.hg19.knownGene for sequences which matched *HNF1A* pfms with a score >0.80.

(ii) Pathway analysis

To gain mechanistic insight into the master regulatory role of *HNF1A* in the human hepatocyte, I analysed the differentially regulated gene set for enrichment of discrete cellular/biological pathways.

I conducted Gene Ontology (GO) and Kyoto Encyclopaedia of Genes and Genomes (KEGG) Pathway analysis using functions goana and kegga from the R package ‘limma’ (version 3.34.5). I tested the set of differentially expressed genes with FDR-adjusted p values ≤ 0.05 for over-representation of GO terms and KEGG pathways. I then extracted the most significantly enriched annotations using the topGP and topKEGG functions from ‘limma’. The R packages ‘AnnotationDbi’ and ‘KEGGREST’ were used to extract gene names from top KEGG pathways and GO terms for deeper exploration into specific pathway or

ontology annotations (eg MODY enrichment or HNF1A-regulated liver 'secrete-ome' via GO terms with extracellular/peripheral membrane annotations). I merged the extracted gene lists with the DE set by gene name to produce a final table of DE genes driving GO term or KEGG pathway enrichment.

4.2.7 Genome editing

To enhance signal-to-noise of potentially MAVE-compatible endogenous HNF-1a targets, I attempted to generate an *HNF1A*-free HepG2 system **4.3.2(i)**.

I infected HepG2 cells with CRISPR-Cas9 editing construct pHKO_hSpCas9_HygroR expressing human codon optimised Cas9 (hSpCas9), a small guide RNA sequence [agcagcacaacatcccacag] targeting a region in exon 2 of HNF1A, and a hygro resistance cassette (**Figure 4.5**). I selected cell colonies with stably integrated LTR flanked sequence by exposure to 1mg/ml Hygromycin B (ThermoFisher cat no. 10687010) for 7 days. To determine editing efficiency, I extracted genomic DNA from HepG2 cells at 3, 6, and 8 weeks post-selection using DNeasy Blood & Tissue Kit (Qiagen cat no. 69504) per manufacturer's instructions. I then amplified a 500-600 bp genomic region flanking the sgRNA-targeted sequence in *HNF1A* by PCR (**Table 4.2**) using Herculase II Fusion DNA polymerase (Agilent cat no. 600675), purified using QIAquick PCR purification Kit (Qiagen cat no.28104) and analysed on a 2% agarose gel (**Figure 4.6**). Purified amplicons from edited and WT *HNF1A* were Sanger sequenced (Eurofins). I evaluated genomic editing efficiency using the online tool **TIDE**: **T**racking of **I**ndels by **DE**composition (https://tide-calculator.nki.nl/) (**Figure 4.7**).

4.2.8 Site Saturation Variant Library (SSVL) synthesis, high-throughput cloning and verification

A saturation mutagenesis library of all possible missense permutations in *HNF1A* was designed by collaborators at the Broad Institute using recoded *HNF1A* cDNA as the template sequence **(subsections 4.2.1 and 4.4.2)**. The library was designed to include every possible single-amino acid change at each position in addition to a single synonymous substitution:

631 amino acids × 20 possible substitutions per position = 12,620 alleles theoretically (12,615 designed)

DNA oligonucleotides were synthesised, cloned and verified by the commercial biotechnology company Twist Bioscience (San Francisco, CA, USA). Library fabrication was performed using a site-directed multiplexed synthetic mutagenesis method similar to the Mutagenesis by Integrated TilEs (MITE) approach described by Melnikov et al (2014). Completeness and quality of the linear double-stranded DNA library was validated by NGS. It was delivered in 96 well plates with 2 site-saturated positions per well, 50 ng each.

To enable lentiviral delivery of the library, oligos were high-throughput cloned into the PCW-puroR lentiviral transgene vector **(Figure 4.4B)** via EcoRI/BamHI, based roughly on the pDNA library amplification protocol published by the GPP at the Broad Institute https://portals.broadinstitute.org/gpp/public/resources/protocols. To confirm representation and circularisation of all alleles post-cloning, NGS of *HNF1A* cDNA (Nextera, MiSeq) was undertaken after positive amp-resistant colonies were pooled, and DNA was isolated from them and linearised to size separate

HNF1A via EcoRI/BamHI restriction digest and gel size selection. Quality control summary reports were shared and included pass/fail scores for all amino acid substitutions and NGS read counts for each allele **4.4.2(i).**

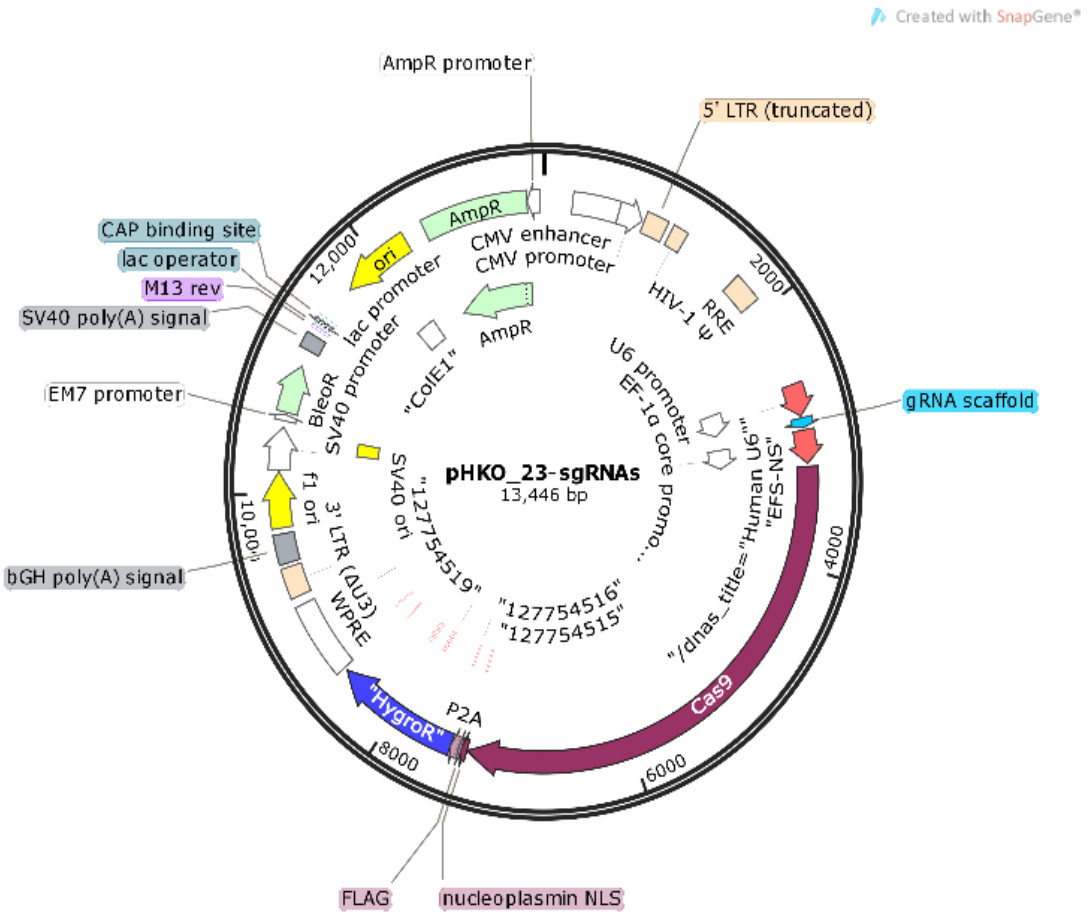


Figure 4.5. Map of pHKO_hSpCas9_HygroR construct expressing sgRNA targeting 20bp region of *HNF1A* exon 2. SnapGene Viewer v2.6.2.

Table 4.2. Primers for sequencing sgRNA targeted region of *HNF1A* exon 2.

	Forward 5' – 3'	**Reverse 5' – 3'**
HNF1A_CRISPR KO_A	ATGTGTCTGTGTCCCTGTGTC	AAGAAAGCAGGGACCCATTTCA
HNF1A_CRISPR KO_B	GACCATACCTCACCGTCCCT	GTTCAGGACCAAGTCTACTCCC

reagents	***ul***
DNA template	1 (200 ng)
5X Herculase buffer	10
Primer F (10uM)	1.25
Primer R (10uM)	1.25
dNTP mix (25 mM each dNTP)	0.5
DMSO	1
Herculase II fusion	0.5
water	34.5

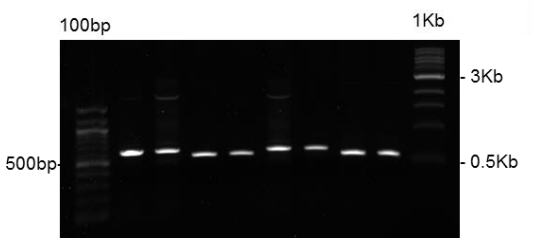


Figure 4.6. Amplification of genomic DNA (gDNA) region flanking sgRNA-targeted sequence in *HNF1A* exon 2. Left: Herculase II PCR amplification recipe; Right: 2% DNA agarose gel showing amplified *HNF1A* gDNA fragments (500-600bp) using primers in Table 2 (n=4 per primer) against 100bp and 1Kb DNA ladders.

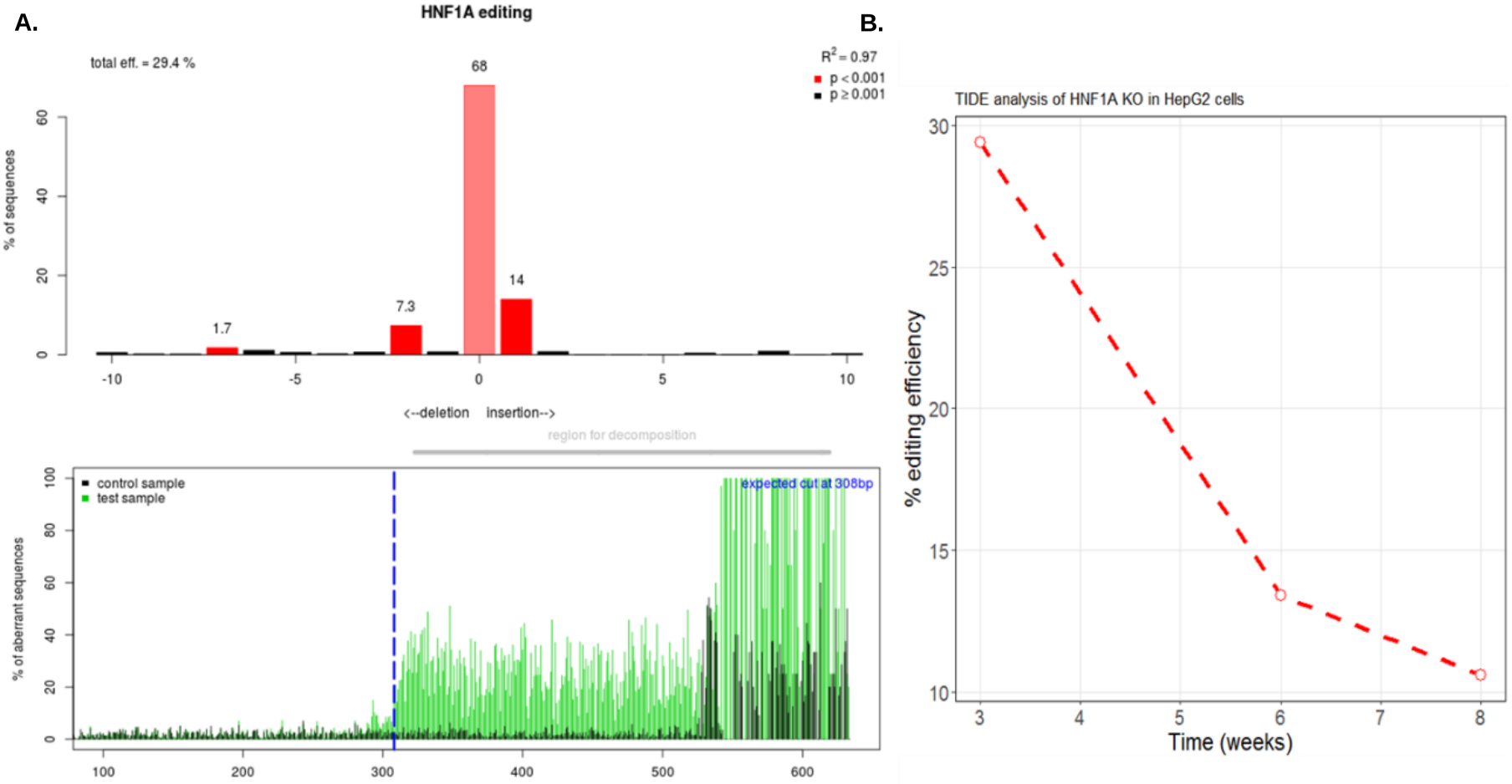


Figure 4.7. Analysis of Cas9-mediated *HNF1A* editing efficiency using the Tracking of Indels by Decomposition (TIDE) web tool. (A) Quantitative spectrum of introduced genomic sequence aberrations (indels) around Cas9 cut site (top) and alignment of Sanger sequence traces of wild-type versus edited *HNF1A* exon 2 amplicon (bottom) **(B)** Percent TIDE-analysed editing efficiency of HNF1A in HepG2 cells at 3, 6, and 8 weeks post-integration of sgRNA-Cas9 sequence.

4.3 Results

In this section, I describe my efforts towards building a transactivation-centric MAVE for *HNF1A*. The main hypothesis was that an *in vitro* transcriptional reporter gene system for *HNF1A*, either endogenous in a physiologically-relevant cell model or artificial, would demonstrate sensitive, allele-specific discrimination, in the form of fluorescence readouts, represented by laser-induced emission curves in fluorescence-activated cell sorting (FACS) acquisition data, upon induction by an *HNF1A* saturation mutagenesis library **(Figure 4.3).** The primary objective was to develop such a system. This required thorough evaluation of the sensitivity of each of the proposed strategies with a series of preliminary experiments and confirmatory analyses.

4.3.1 Artificial reporter

I attempted to build a MAVE for *HNF1A* which relies on an artificial fluorescent reporter (mCherry) downstream of a high binding consensus promoter sequence (human albumin) as a functional marker. I prioritised the development of this system because it mimicked, to a large degree, classic low-throughput transactivation assessment pipelines.

For an artificial reporter cell system, I selected HeLa cells by virtue of their robustness and lack of endogenous *HNF1A* to allow, in the first instance, an assessment of the experimental approach in an efficient and easy-to-handle *in vitro* model with minimal endogenous background noise. Additionally, there exists an enormous pool of data using the combination of albumin promoter in HeLa cells as a system for reporter-based molecular interrogation of *HNF1A* sequence variants (Galan et al, 2011; Najmi et al, 2017; Consortium STD, 2014;

Balamurugan et al, 2016; **results presented in Chapter 3**). The large-scale molecular phenotyping effort described in Chapter 3 is one prominent example of the use of an albumin promoter/HeLa cell combination for evaluating transcriptional activity of *HNF1A* sequence variants. The guiding rationale was that an identical artificial reporter assay would be more directly comparable and easier to calibrate against *HNF1A* luciferase data already generated in HeLa cells both in our lab and elsewhere.

(i) Phase I: transient transfections

Before embarking on experimental set-ups of higher complexity, which more closely resemble the high-throughput screen, I decided to test the various individual components of an artificial reporter system under the simplest conditions: artificially express *HNF1A* **(Figure 4.4A)** and the promoter-reporter gene sequence (hALB_mCherry) **(Figure 4.4C)** in HeLa cells, and measure mCherry induction levels by FACS.

a) Background (signal-to-noise)

To evaluate levels of mCherry reporter gene induction by *HNF1A*, I transfected HeLa cells with the pgl4.26_hALB_mCherry reporter construct and either a) empty pcDNA3.1 expression vector b) WT *HNF1A* pcDNA3.1 or c) a pcDNA3.1 expression vector with an *HNF1A* sequence harbouring the deleterious allele p.T260M **(Figure 4.8)**. An untransfected HeLa cell population was included as a negative, and also FACS-gating, control. Expectedly, mCherry reporter was induced in an *HNF1A* sequence-dependent manner, with highest degrees of detected fluorescence on the PE-Texas-Red-A emission spectra from cells transfected with WT *HNF1A*, followed by *HNF1A* p.T260M, which induced

mCherry expression in a manner almost indistinguishable from that of the empty vector control (i.e. basal levels of mCherry induced by other albumin promoter-binding proteins in the HeLa cell) **(Figure 4.8).** The small proportion of mCherry-positive HeLa populations out of total assayed HeLa cells, represented by 'mCherry(+)' in **Figure 4.8**, highlights both the limitations of lipid-based DNA delivery method (i.e. transfections versus lentiviral infection) and the high sensitivity of FACS-based detection despite suboptimal conditions **(Figure 4.8)**.

b) Dynamic range

In an effort to enhance HNF-1a-driven mCherry signal intensity and to keep the number of introduced variables per run at a minimum, I created a HeLa cell line with a stably integrated reporter using the hygromycin resistance cassette in the pgl4.26_hALB_mCherry vector to select positive cell clones. I expected an amplification in fluorescent reporter signal from polyclonal HeLa_pgl4.26_hALB_mCherry cells based on a guarantee of at least a single mCherry integration event per cell.

To evaluate the success of stable cell line generation and further examine the sensitivity of the FACS pipeline, I selected a range of technical and clinical control variants which span the spectrum of *HNF1A* functional effects **(Figure 4.9)**. Empirical data on the molecular behaviour of the set of variants I selected is available; in fact, they were selected as controls, in part, based on luciferase transactivation data, generated as functional follow-up of the exome identified *HNF1A* variants in the multi-ethnic T2D case-control cohort, using two promoters (rat albumin and *HNF4A*_P2) and two cell lines (HeLa and INS-1) **(Figure 4.9C) (see Chapter 3)**. These data, in addition to previously published sequence-based

characterization efforts and clinical information associated with carriers of some of these variants, show that this particular allelic series does indeed display a subtle range of HNF-1a functional effects and frequencies:

- benign WT-like (p.A161T)
- moderate impact on transactivation (p.N397S)
- common associated with T2D risk (p.A98V)
- low frequency associated with T2D risk (p.E508K) and
- MODY-causal loss-of-function (p.T260M)

Therefore, I transfected HeLa cells with stably integrated human albumin_mCherry reporter sequence with WT *HNF1A* and the set of variants, and analysed for mCherry induction by FACS. As an appropriate negative control, normal, untransfected HeLa cells were analysed alongside the experimental samples. There was an evident shift in the emission peak of the total viable HeLa_pgl4.26_hALB_mCherry cell populations from that of the standard HeLa cells with neither stably integrated nor transfected reporter (red curve **Figure 4.9A**), suggesting successful incorporation of hALB_mCherry into antibiotic selected cells **(Figure 4.9A)**. Differences in mutation-dependent induction of mCherry expression were evident though difficult to resolve without zooming in on the subset of gated mCherry-positive populations from each sample **(as shown in Figure 4.9B)**.

Upon viewing mCherry positive emission curves from each independent transfection, the variants exhibited functional trends nearly identical to those observed in luciferase assays **(Figure 4.9B).** Again, despite low transfection-based induction efficiency, FACS fluorescence signal detection proved extremely sensitive and able to discriminate subtle effects. **Figure 4.9D** shows median

fluorescence intensity values (MFI) of gated mCherry positive subpopulations from each sample. Generally, statistics generated on flow sort data are not considered robust measures due to the non-normal distribution of the cell populations and the relatively subjective nature with which functions such as gating are performed. Nonetheless, MFI values also appeared to demonstrate a functional trend consistent with previous findings and reports **(Figure 4.9C and 4.9D)**.

(ii) Phase II: lentiviral infection

Results presented in subsection **4.3.1(i)**, particularly in **Figure 4.9**, were promising, and encouraged further pursuit of the artificial transcriptional reporter system for higher level validation that would better replicate the conditions of the MAVE. High-throughput multiplexed functional assays require tight control of gene integration and expression which is achievable with a combination of tet-ON expression **(Figure 4.4B; subsection 4.2.2)** and lentiviral delivery (LTR-mediated genomic integration, and titres can guarantee 1 variant entry per cell; **Figure 4.1**). Conducting control experiments which more closely resemble the screen conditions required first cloning WT *HNF1A* sequence into a 2nd generation lentiviral transgene expression vector pCW57.1 Dest (PCW_puroR). Transcriptional activation of the transgene in this viral plasmid is controlled by doxycycline. The vector also contains a puromycin resistance marker for selection of successfully transduced cells. A puromycin kill curve was generated on HeLa_pgl4.26_hALB_mCherry cells to determine the dose required to kill all non-resistant cells within a 48 hour period.

To assess the impact of lentiviral delivery of *HNF1A* sequences on mCherry signal intensity, I infected populations of HeLa_pgl4.26_hALB_mCherry cells with lentivirus harbouring WT *HNF1A* **(Figure 4.10A)** *HNF1A*-p.T260M (LoF) **(Figure 4.10B)**. Transduced cells with successfully integrated sequence were selected with puromycin and exposed to a range of doxycycline concentrations to induce varied levels of *HNF1A* expression. Standard HeLa cells were also infected, selected and induced with 5 µg/ml doxycycline, as a negative control (lacking mCherry reporter). As evident from the FACS acquisition plots, there was a noticeable shift in the emission peak of the total viable populations from experimental versus control (red curve) samples **(Figure 4.10)**. Dox dose- and mutation- dependency were also observed across all samples; however, these trends were noticed by increases in the area under each curve rather than complete shifts needed to attain the discriminatory power required for the functional classification of thousands of genetic perturbations in parallel.

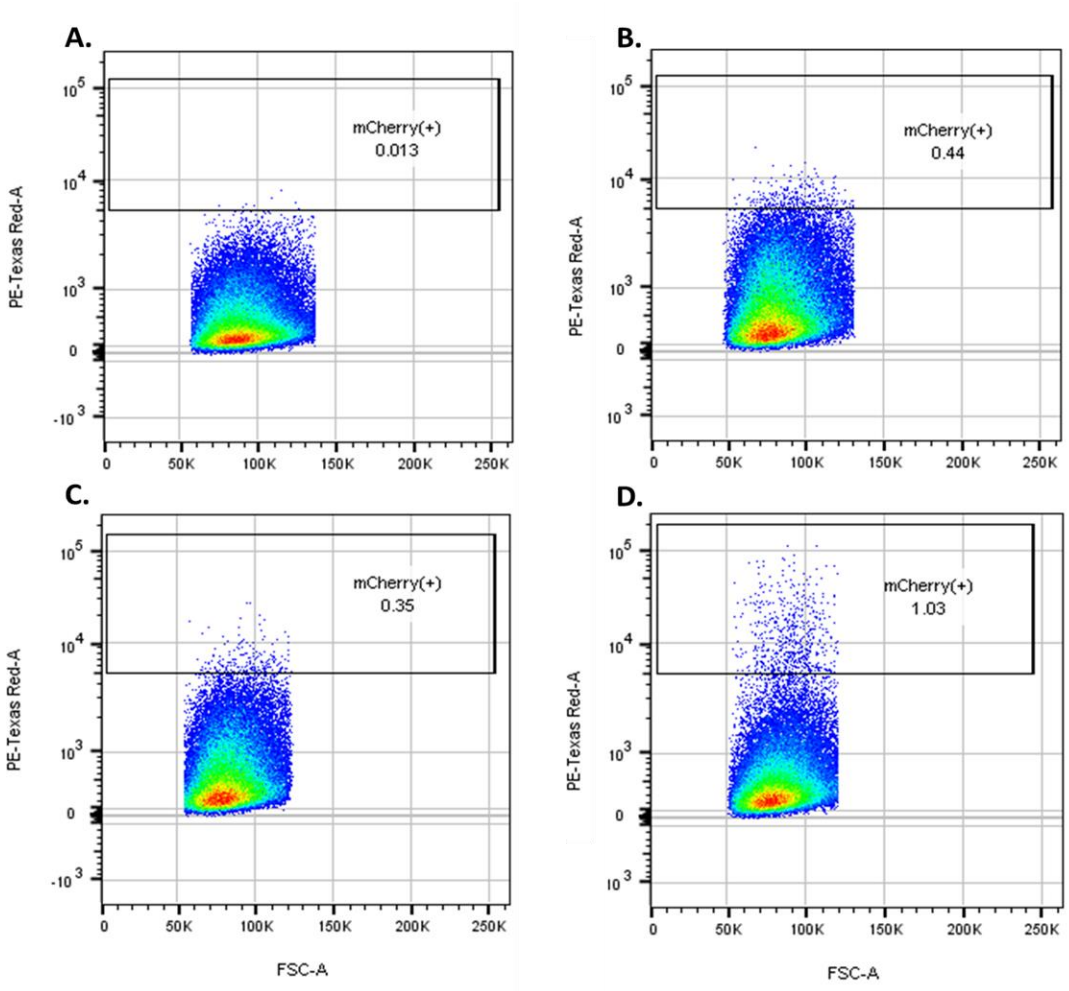


Figure 4.8. Flow cytometry acquisition plots showing varied degrees of mCherry reporter gene induction as a function of HNF-1A transcriptional activity in HeLa cell populations. 100,000 events were collected from each sample and analysed on a BD LSRFortessa™ cell analyser and FlowJo (version 10.0.8)**.** Populations were enriched for viable cells by excluding cell death and debris via forward/side scatter. **A.** Untransfected HeLa cells used to set the gate for mCherry-positive subpopulations**. B.** HeLa cells co-transfected with empty pcDNA3.1 vector and pgl4.26_hALB_mCherry reporter construct **C.** HeLa cells co-transfected with *HNF1A* p.T260M pcDNA3.1 expression plasmid and pgl4.26_hALB_mCherry reporter construct. **D.** HeLa cells co-transfected with *HNF1A* wild-type pcDNA3.1 expression plasmid and pgl4.26_hALB_mCherry reporter construct.

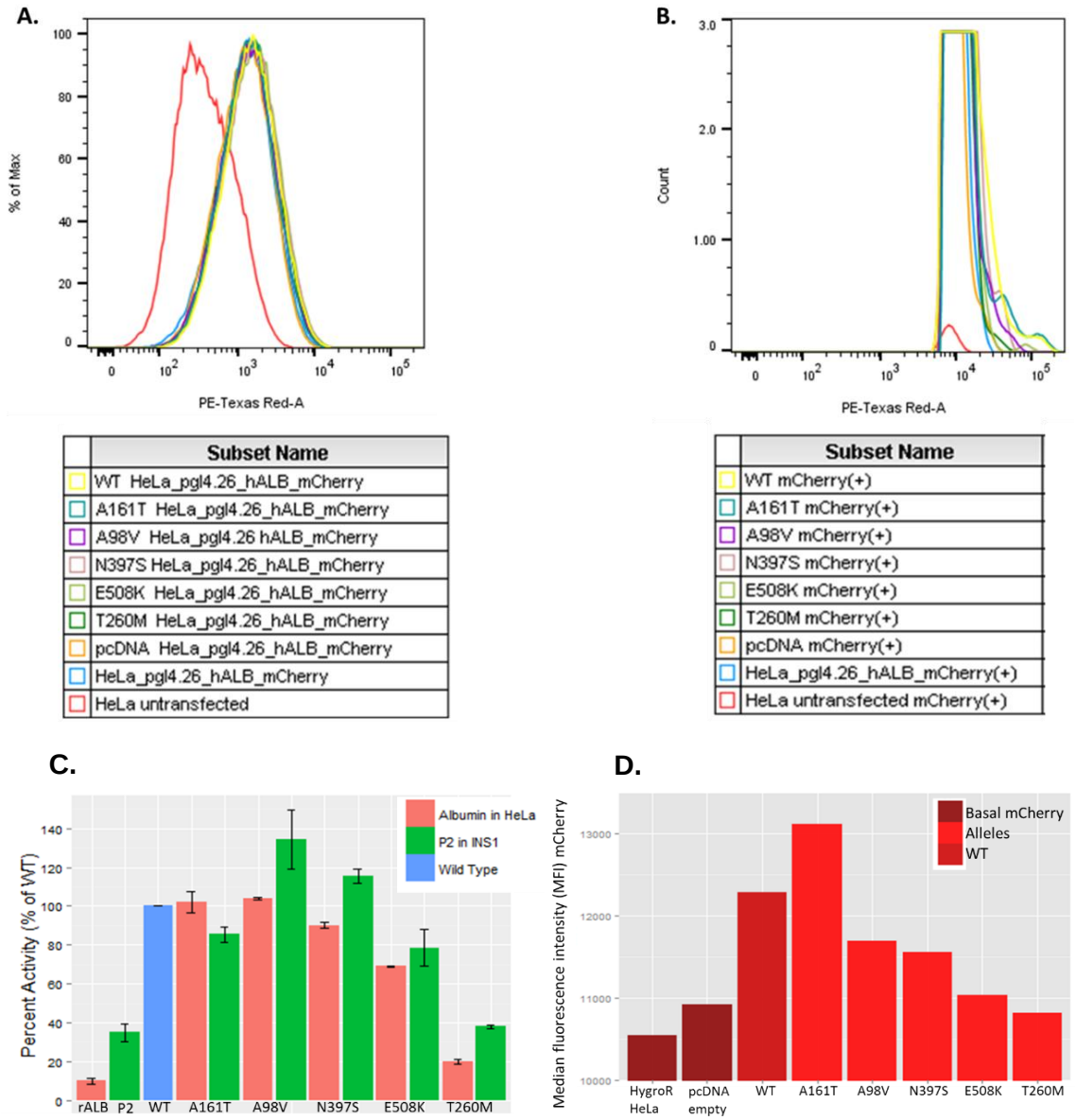


Figure 4.9. FACS analysis of hygromycin resistant (HygroR) HeLa cells with stably integrated mCherry reporter (HeLa_pgl4.26_hALB_mCherry) transfected with wild-type *HNF1A* cDNA and an HNF1A allelic series. 100,000 events were collected from each sample and analysed on a BD LSRFortessa™ cell analyser and FlowJo (version 10.0.8). Populations were enriched for viable cells by excluding cell death and debris via forward/side scatter. **A.** Histogram overlay of total viable HygroR HeLa cell populations transfected with wild-type and mutant *HNF1A*, pcDNA3.1 empty vector, or standard untransfected HeLa cells (red). **B.** Histogram overlay of mCherry-positive HygroR HeLa cell populations transfected with wild-type and mutant *HNF1A*, pcDNA3.1 empty vector control transfections and standard untransfected HeLa cells (red). **C.** Luciferase transactivation data (generated by Dr. Jana Rundle and Dr. Amanda Bennett) from WT HNF-1A and the allelic series. The variants, WT, and empty pcDNA3.1 vectors were co-transfected with luciferase reporter gene constructs downstream either HNF4A_P2 promoter in INS1 cells or Rat Albumin (rALB) in HeLa cells. Measurements are shown as percentage activity relative to wild-type ± SEM (n=3 biological replicates). **D.** Median fluorescence intensity (MFI) values of gated mCherry positive HygroR HeLa cell populations shown in **B,** where basal levels of mCherry induction independent of HNF-1A transcriptional activity are represented by untransfected HygroR HeLa cells (HeLa_pgl4.26_hALB_mCherry) and pcDNA3.1 empty vector transfections**.**

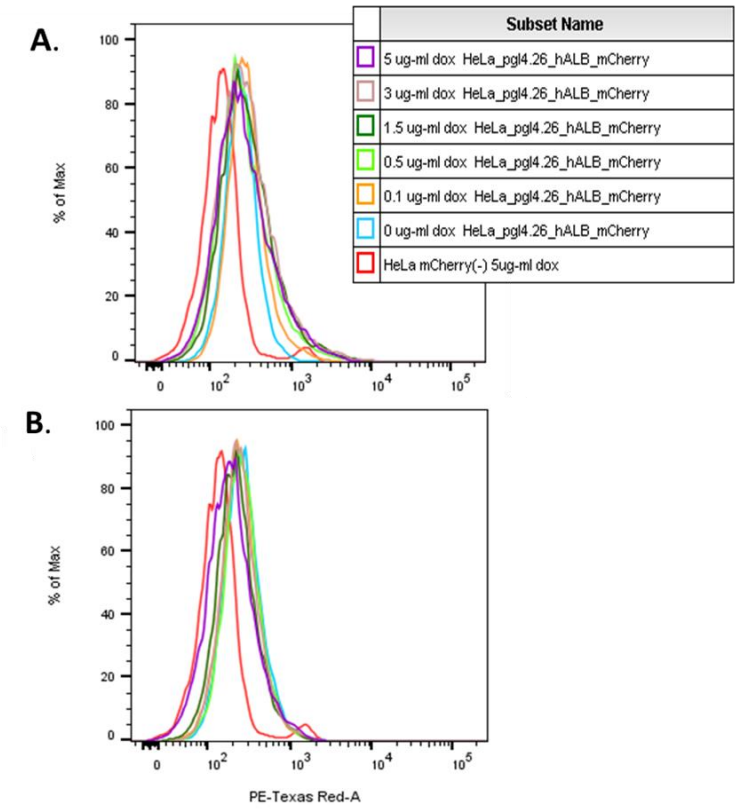


Figure 4.10. FACS analysis of hygromycin resistant HeLa cell populations (HygroR) HeLa_pgl4.26_hALB_mCherry transduced with A. Recoded WT *HNF1A* **B.** Loss-of-function *HNF1A* mutant p.T260M. Transduced puromycin-selected (1 µg/ml) HeLa_pgl4.26_hALB_mCherry were induced to express HNF-1A with a range of doxycycline concentrations (0-5 µg/ml) over a 48 hours. Approximately 10,000 events were collected from each sample and analysed on a BD LSRFortessa™ cell analyser and FlowJo (version 10.0.8)**.**

4.3.2 Endogenous reporter

The challenges associated with the artificial reporter gene strategy, particularly as they pertain to achieving high-throughput-compatible signal-to-noise levels, prompted a shift towards exploring an endogenous response gene approach.

The key aim in this section was to identify and validate a set of robust endogenous cell surface-expressed HNF-1a target transcripts in a physiologically relevant cell model. Cell-surface expression is an important criterion for an endogenous target as fluorescent staining of cell surface markers does not require permeabilisation of the cells, which severely comprises genomic DNA integrity. Next-generation sequencing of gDNA, which demands high quality input material, is an essential component of the MAVE workflow, as it is on the basis of read counts per FAC-sorted function group that each allele is assigned a function score **(Figure 4.1)**.

Below, I describe my efforts to systematically identify high resolution cell-surface markers of HNF-1a dysfunction and evaluate their suitability for massively parallel functional screening as per the methodological pipeline depicted in **Figure 4.3**.

(i) Cellular model selection and characterisation

I selected the human hepatoma cell line HepG2 as an *in vitro* model for an *HNF1A* MAVE for several technical and biological reasons which include: a) its physiological relevance for screening *HNF1A* perturbation as the hepatocyte is among the main tissue types in which *HNF1A* is predominantly expressed, b) ease of maintenance and manipulation compared to immortalised human pancreatic beta cell lines and iPSC-derived beta cells, which would otherwise provide the most informative systems for studying MODY/T2D-relevant genes, and c) its human origin which facilitates interpretability and translatability of

findings compared to data derived from, for example, the pancreatic beta cell lines INS-1 and Min6, which are of rat and mouse origin, respectively.

I characterised HepG2 cells for cytogenetic, transcript- and protein- level expression of *HNF1A* **(Figure 4.11A, B and C)**. The hepatoma line endogenously expresses *HNF1A* transcript and protein, as expected, and is tetraploid for *HNF1A* (chr12:121,416,371-121,440,314) based on FISH characterisation. The next step was to transduce the cells with *HNF1A*-harbouring lentivirus and stimulate expression of the gene with dox. This could be done in two different HepG2 systems: one of overexpression, which would need to be sufficiently high to render background noise from endogenous *HNF1A* function negligible, or one in which *HNF1A* would be stimulated in an *HNF1A*-depleted system to eliminate background expression (noise) of targets driven by endogenous *HNF1A*.

In case of problematic background noise from endogenous HNF1A background expression, I attempted to generate an *HNF1A*-depleted HepG2 system using CRISPR-Cas9 gene editing technology. I used a human codon optimised Cas9 (hSpCas9) lentiviral expression construct (2^{nd} generation) pHKO_hSpCas9_HygroR (methods) into which sgRNA targeting HNF1A exon 2 was cloned. The sgRNA was designed and tested by collaborator Laeya Najmi (University of Bergen) who has achieved complete knockout of HNF1A in Hep3b and HUH-7 human liver lines using this guide sequence (unpublished). Determining the hygromycin selection concentration for HepG2 cells was necessary to use this editing construct as successfully transduced cells are selected via integrated hygro resistance sequence. HepG2 cells exhibited a naturally high resistance to hygromycin as I was only able to observe near-complete cell death at a concentration of 1 mg/ml after 7 days of exposure to the

antibiotic in media (standard hygro selection cell culture concentration range is 200-500 ug/ml over a period of 5 days).

I transduced HepG2 cells with the CRISPR-Cas9 construct, which I packaged into lentiviral particles using PAX and VSV-G, and selected positive clones using 1 mg/ml of hygromycin over 7 days. Hygro resistance of the surviving cells suggests integration of the LTR flanked sgRNA-hSpCas9-hygroR sequence into the HepG2 genome. In these cells, *HNF1A* editing should occur perpetually with potentially increasing efficiency over time depending on sgRNA performance. To assess CRISPR induced HNF1A KO, I extracted genomic DNA (gDNA) from HepG2 cells 3, 6 and 8 weeks post-selection; I then amplified the 500-600bp region encompassing the targeted sequence in exon 2 by PCR, which was Sanger sequenced; finally, I analysed the DNA sequence by TIDE for cutting efficiency. TIDE analysis showed the efficiency with which indels were being introduced dropped from roughly 30% at 3 weeks after a hygro-resistant population was selected to 10.8% after 8 weeks of culture. These data were unsurprising considering the appreciable amount of cell death observed with each passage.

Altogether, decreasing editing efficiency and cell survival over time suggested *HNF1A* essentiality in HepG2 cells. In light of RNAseq data presented later in this section which support an essential master regulatory role of *HNF1A* in HepG2 cells, these results were not unexpected. Based on these results, I decided to abandon the depletion approach and pursue overexpression In HepG2 cells.

(ii) Expression HNF1A cDNA in a dox dose-dependent manner

To identify high resolution MAVE-compatible endogenous markers of HNF-1a transactivation, HepG2 cells were infected with *HNF1A* cDNA under tight regulation of doxycycline (dox) and exposed to a range of dox doses (0, 0.1, 1, 5 µg/ml) over a 48-hour period. Dox dose-dependent induction of *HNF1A* sequences was confirmed by quantitative PCR (qPCR) [fold change in ΔCt = 112.7 ± 32.04 and 3.6 ± 1.08 for recoded and native *HNF1A* sequences, respectively, from 0 to 5 µg/ml dox] and immunoblotting **(Figure 4.11B and C)**.

(iii) Target identification in HNF1A-infected cells

Transcriptomic profiling of samples that showed dox dose-dependent induction of exogenous *HNF1A* was undertaken by 75bp paired-end RNA-sequencing of multiplexed polyA cDNA libraries on a single lane of a HiSeq4000 (Nextera, Illumina). Differential expression (DE) analysis revealed 729 up- or down-regulated transcripts (FDR-adjusted P ≤ 0.01) in response to *HNF1A* dose induction (P=2.57e-09) **(Figure 4.11D)**. The expression trends of the top 30 up- and down-regulated genes across the tested dox dose range are displayed as heat maps in **Figure 4.11D**. Beyond 1 µg/ml dox, the linear dose response relationship demonstrated by the vast majority of differentially expressed genes was no longer maintained, which suggests extreme synthetic perturbation at 5 µg/ml dox (despite linear dose response of exogenous *HNF1A*). This induction trend was observed with both the exogenous native and recoded *HNF1A* cDNA sequences, which exhibited highest correlation in $\log_2$FC gene expression at 1 µg/ml dox (Spearman's rho = 0.56; P < 2.2e-16) **(Figure 4.12)**.

(iv) Target prioritisation and validation

I used several criteria to prioritise HNF-1a response genes for MAVE compatibility from the DE list:

- significant differential expression (FDR 1%, log2FC < or > 0.5)
- linear dose response to dox-stimulated *HNF1A* across tested doses
- cell-surface expression for easy access
- availability of well-validated antibodies

Under these criteria, I prioritised *CUBN* ($P=1.48x10^{-7}$), *CLDN2* ($P=2.60x10^{-6}$), *TMEM27* ($P=3.07x10^{-3}$), and *TM4SF4* ($P=2.59x10^{-3}$) for target validation. I used qPCR to validate RNA-seq expression trends of the prioritised gene set, all of which exhibited high fold changes in transcript abundance in response to dox induced *HNF1A* overexpression **(Figure 4.13A)**. To control for potentially confounding effects on expression by dox, I performed the same experiment without infecting the cells with *HNF1A* and evaluated expression of the shortlisted gene set by qPCR – that is, HepG2 cells were exposed to the same range of dox doses over 48 hours and the expression of *CUBN*, *CLDN2*, *TMEM27* and *TM4SF4* was assessed as a function of exposure to dox. This experiment eliminated dox as a confounding factor due to unchanging expression trends across doses without *HNF1A* stimulation and overexpression (undetectable *CUBN* transcript levels without *HNF1A* stimulation) **(Figure 4.13C)**.

Dose-dependent increases in expression were observed at the protein level for both CLDN2 and TM4SF4 after a 48-hour dox stimulation, whereas CUBN and TMEM27 were undetectable on a Western blot (not unexpected considering barely detectable transcript levels for these genes at 0 µg/ml dox) **(Figure 4.13B)**.

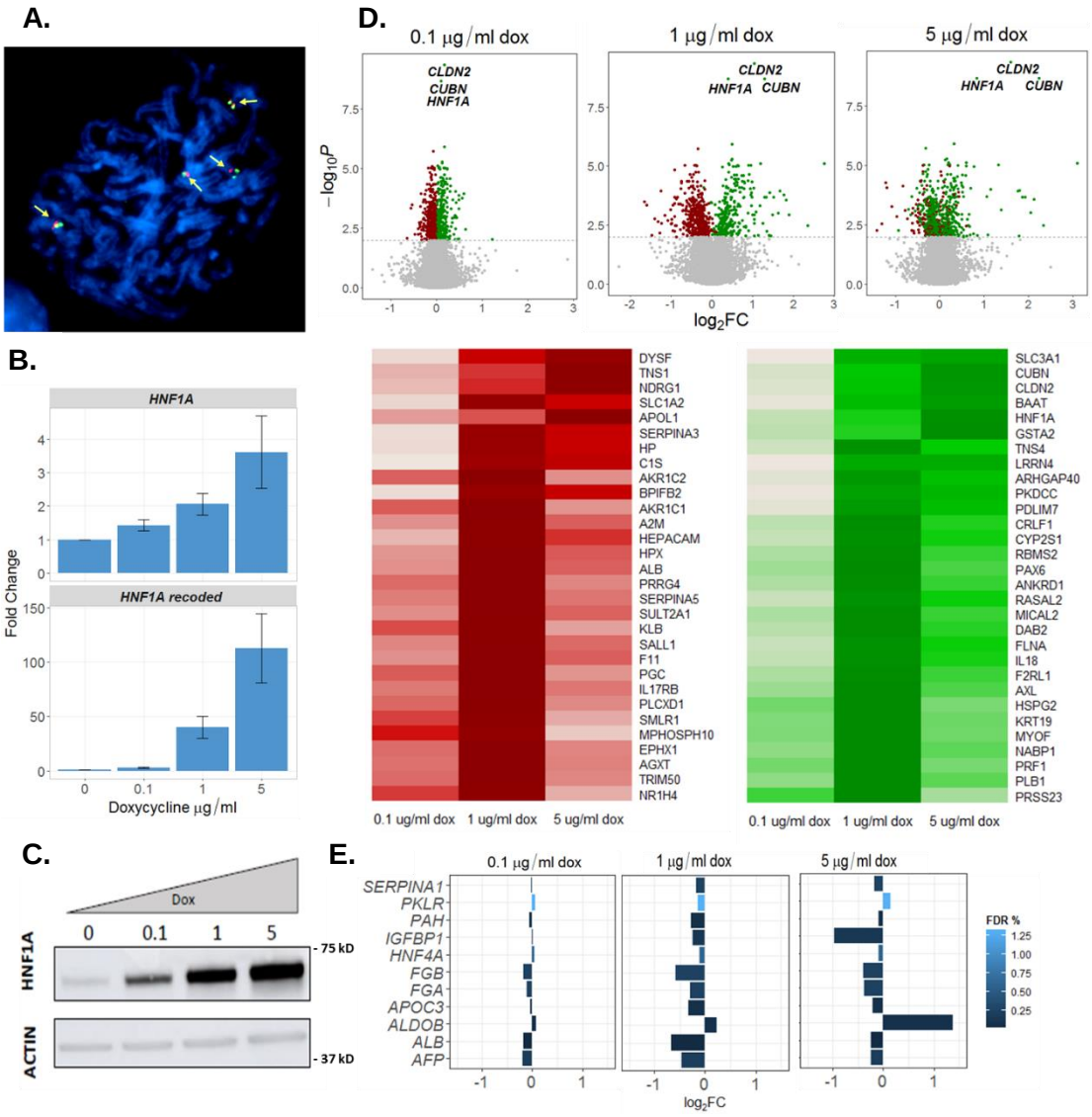


Figure 4.11. Transcriptomic profiling of HepG2 cells transduced with doxycycline-regulated *HNF1A* sequences. (A) FISH on HepG2 metaphase spreads, with BAC CTD-3028F22 (red signal) covering *HNF1A*, RP11-148L11 (green signal) on chromosome 12 telomere. Yellow arrows point to probe hybridization sites for *HNF1A*. **(B)** Quantitative PCR (qPCR) analysis of mRNA expression of wild-type and recoded *HNF1A* sequences in response to dox dose induction over 48 hours. *PPIA* was used as the internal control. PPIA was used as the internal control. Results are the average fold change in relative expression ± SEM of biological triplicates. **(C)** Immunoblot analysis of HNF-1a expression induced by a range of dox doses in HepG2 cells over 48 hours. **(D)** Volcano plots of transcript profiles from HepG2 cells transduced with wild-type *HNF1A* and induced with a range of dox doses. Red and green dots represent down- and up-regulated differentially expressed transcripts across tested doses (FDR 1%). Heat maps show expression (log2FC) of top 10% differentially regulated transcripts; green and red bars represent up- and down-regulation, respectively (FDR 1%). **(E)** Differential expression trends of canonical *HNF1A* target transcripts across tested doses.

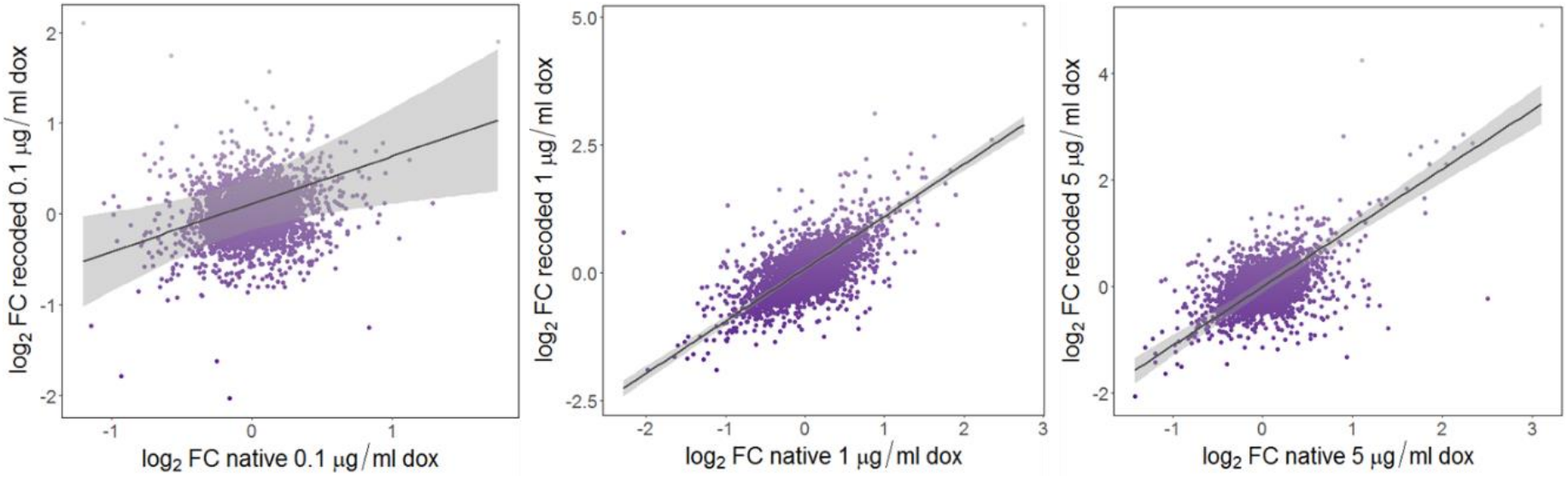


Figure 4.12. Correlation between expression profiles of samples transduced with wild type or recoded *HNF1A* sequences. Linear regression was performed on the subset of transcripts with a log2FC ≥ 0.5 compared to non-induced samples at 0.1 µg/ml dox (Spearman's rho correlation coefficient = 0.10; $P < 2.2x10^{-16}$), 1 µg/ml dox (Spearman's rho correlation coefficient = 0.56; $P < 2.2x10^{-16}$), and 5 µg/ml dox (Spearman's rho correlation coefficient = 0.39; $P < 2.2x10^{-16}$).

Dose response of CLDN2 appeared more promising on a Western blot compared to seemingly unchanged expression of TM4SF4 across tested doses; however, in flow cytometric assays, the monoclonal fluo-conjugated antibody targeted against human TM4SF4 outperformed double-staining efforts of CLDN2 as evident by appreciable shifts in fluorescence emission peaks accompanied by a log-scale trace after 48 hours of *HNF1A* dox stimulation **(Figure 4.13B)**. A crucially determining factor of MAVE suitability is for the selected target to exhibit log-scale shifts of fluorescence emission peaks that would accommodate library variability between positive and negative control staining conditions. As such, TM4SF4 was selected for extensive follow-up and optimisation for *HNF1A* saturation mutagenesis screening.

The discriminatory capacity of TM4SF4 as an *HNF1A* MAVE marker was interrogated on two levels: **1)** signal-to-noise by comparing TM4SF4+ HepG2 populations as a function of WT *HNF1A* or the induction of *HNF1A* harbouring a severe loss-of-function missense allele (p.T260M), and **2)** dynamic range of TM4SF4 fluorescent signal, evaluated by an *HNF1A* allelic series that spans the functional and phenotypic spectra of the gene. I observed a roughly 27-fold increase in TM4SF4+ HepG2 cell populations under dox stimulation at 1 and 5 µg/ml as a response to WT *HNF1A* cDNA compared to the LoF variant **(Figure 4.13D)**. Though impressive, the fluorescent signal from TM4SF4 after 48 hours of induction (roughly 20% TM4SF4+ HepG2 populations) is considered submaximal for the purposes of high-throughput screening. To achieve highest signal-to-noise, I evaluated TM4SF4 dox dose response in flow cytometric assays over a 7-day time course **(Figure 4.14)**. Extending dox stimulation to one

week resulted in a log shift in WT *HNF1A*-driven TM4SF4-PE emission compared to unstimulated controls, and an increase to ~70% TM4SF4+ HepG2 populations.

(v) Evaluation of dynamic range with an HNF1A allelic series

To evaluate the dynamic range of an anti-TM4SF4 HepG2 MAVE for *HNF1A*, I interrogated TM4SF4 response to a well characterised *HNF1A* allelic series with robust genetic, clinical and functional evidence. The functional and phenotypic consequences of the chosen alleles span the continuum of effects associated with *HNF1A* **(Figure 4.15)**. This series differs slightly from the one I tested in subsection **4.3.1(i)(b) (Figure 4.9)**; while equally wide-ranging, it is more comprehensive and captures a more subtle gradient of *in vitro* function effects along the spectrum.

The spectrum is defined by benign WT and alleles with varied degrees of WT-like *in vitro* behaviour (p.A161T & p.S535R) on one end, loss-of-function MODY-causal alleles on the other (p.T260M, p.P447L, p.P112L, p.R203H), and an intermediate functional T2D risk modifier (p.E508K) **(Figure 4.15)**. Luciferase transactivation data was generated using these alleles in two different cell systems using two different promoters: rat albumin upstream luciferase in HeLa cells and *HNF4A*-P2 promoter upstream luciferase in INS-1 cells **(Figure 4.15)**. Functional data from the latter are influenced by endogenous levels of HNF-1a, which are absent from HeLa cells.

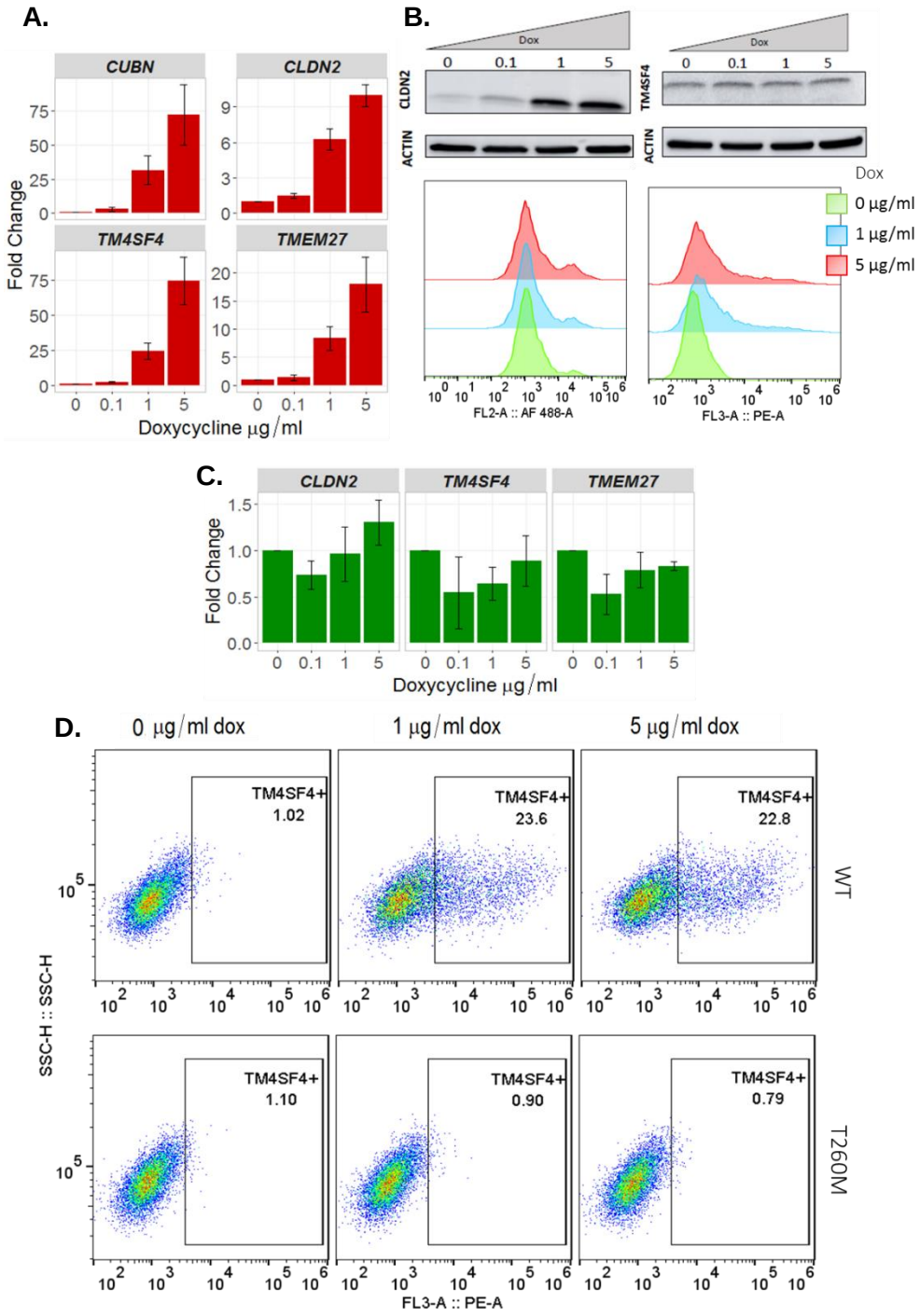


Figure 4.13. HNF-1a transcriptional target validation. Quantitative PCR (qPCR) analysis of mRNA expression of *CUBN*, *CLDN2*, *TM4SF4*, and *TMEM27* in HepG2 cells **(A)** transduced with dox-inducible *HNF1A* sequence and **(C)** non-transduced, exposed to a range of dox concentrations over a 48-hr period. *PPIA* was used as the internal control. Results are the average fold change in expression ± SEM of biological triplicates. **(B)** Proteomic validation of CLDN2 and TM4SF4 expression in HepG2 cells transduced with dox-inducible *HNF1A* and exposed to a range of dox concentrations over 48 hours via immunoblot analysis and fluorescent staining. **(D)** Flow cytometry acquisition plots showing TM4SF4 expression in response to either dox-inducible wild-type *HNF1A* (top) or dox-inducible *HNF1A* harbouring a deleterious MODY-causal variant p.T260M (bottom) in HepG2 populations following 48 hrs of dox induction.

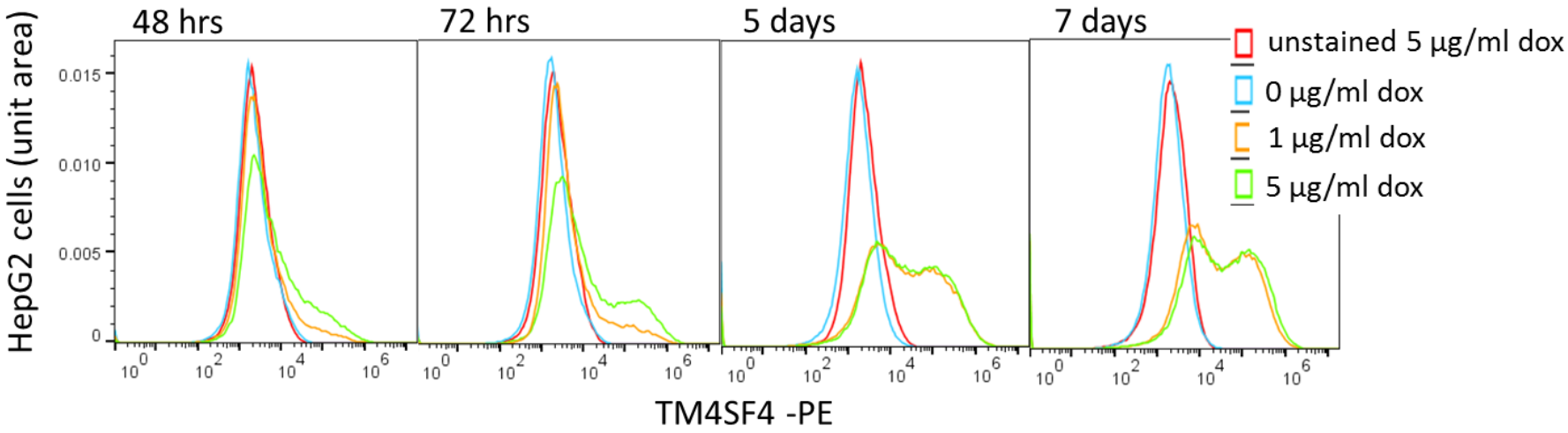


Figure 4.14. Seven-day time course of HNF-1a--mediated TM4SF4 induction in HepG2 cells. *HNF1A*-infected HepG2 cells were stained for TM4SF4 cell surface expression using monoclonal human anti-TM4SF4 PE-conjugated antibody. After 2, 3, 5, and 7 days of *HNF1A* dox-stimulation, the cells were analysed for TM4SF4-PE signal intensity. To understand the impact of TM4SF4 background levels of expression at 0 µg/ml dox (blue curve), HepG2 cells unstained for TM4SF4 were analysed with each experiment as a control (red curve). Approximately 10,000 events were collected from each sample and analysed on a BD Accuri C6™ cytometer at the Broad Institute of MIT and FlowJo (version 10.0.8).

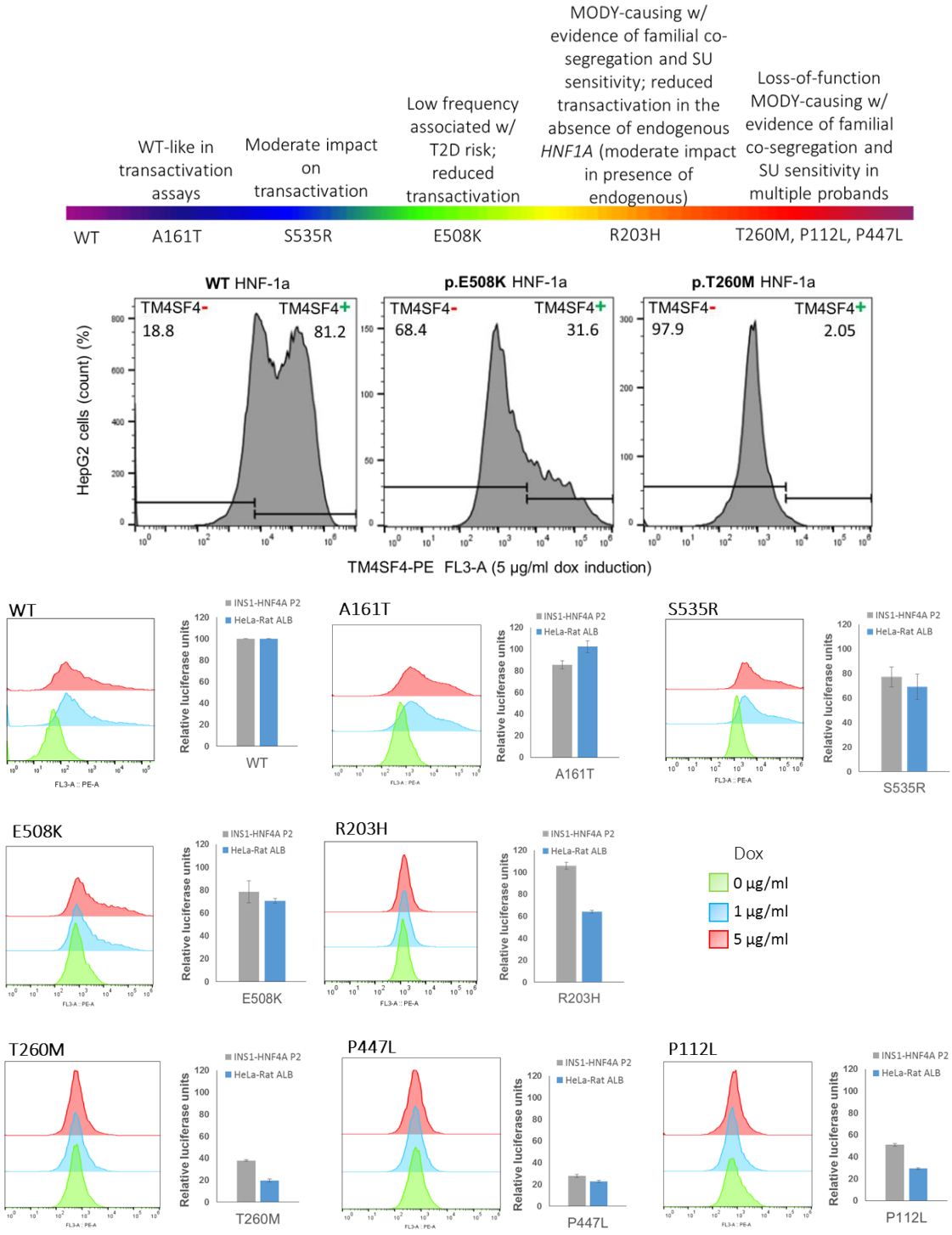


Figure 4.15. FACS analysis of HepG2 cells stained with anti-humanTM4SF4-PE following 7-day dox-dose induction of *HNF1A* allelic series. Each *HNF1A* allele was used to infect HepG2 cells with an MOI of 0.8-1.4 (based on titration experiments) to mimic library infection conditions. Cells were exposed to 0, 1, or 5 µg/ml dox over the course of 7 days and stained with monoclonal human anti-TM4SF4 PE-conjugated antibody. The stained cell populations were analysed for HNF-1a-mediated TM4SF4 induction as reflected by PE emission spectra (x axis on FACS histograms). Roughly 100,000 events were collected from each sample and analysed on a SH800S Sony Cell Sorter at the Wellcome Centre for Human Genetics and FlowJo v10.0.8. Each FACS plot is juxtaposed with the performance of the allele in luciferase reporter gene assays (from effort described in Chapter 3) in two different cells lines using two different promoters (INS1 & HNF-4a_P2 promoter, HeLa and rat albumin promoter).

To simulate MAVE conditions, the alleles were synthesised in recoded *HNF1A* cDNA sequence identical to that of the saturation library and used to infect HepG2 cells at a limiting dilution (MOI 0.8-1.4) to ensure a maximum of a single integrated sequence per cell. Each allele was used in an individual transduction experiment (i.e. not pooled) to facilitate immediate resolution of allele-specific TM4SF4 induction by flow cytometric cell analysis. HepG2 cells with successfully integrated *HNF1A* sequences (survival after 48 hours of selection with 3 µg/ml puromycin) were exposed to either 0, 1, or 5 µg/ml for 7 days. The abundance of TM4SF4-postive populations matched expected molecular behaviour based on relative allele position along the functional spectrum of *HNF1A*: 2-8% from pathogenic damaging alleles, 30% from the low frequency allele associated with T2D risk, and >60% from WT-like and WT *HNF1A*, at 1 and 5 µg/ml dox compared to unstimulated control samples **(Figure 4.15)**. The log-scale activity-specific shifts in TM4SF4-PE fluorescence emission after 7 days of *HNF1A* dox induction suggest that this assay system is highly compatible for massively parallel functional scanning.

4.3.3 Biological insights from transcriptomic profiling of *HNF1A* MAVE model

(i) Known and novel HNF-1a transcriptional targets

The biological relevance of the endogenous reporter MAVE is supported by significant differential expression patterns of canonical HNF-1a target transcripts (FDR-adjusted P ≤ 0.01) across the tested dox doses, including *HNF4A*, *ALB*, *ALDOB*, *AFP*, and *IGFBP1* **(Figure 4.11E)**. The top two differentially regulated transcripts, *CUBN* (P= 2.57e-09) and *CLDN2* (P= 4.90e-10), have not been

previously reported as HNF-1a targets in human liver **(Figure 4.11D)**. I computationally scanned genomic sequence 5kb upstream the TSS of *CUBN* and *CLDN2* for an 80% match with HNF-1a DNA binding motifs as annotated in the public motif databases JASPAR and jolma et al 2013 (MotifDb) **(Figure 4.16)**. A total of 5 and 3 DNA-binding motif sequences were identified upstream TSS of *CUBN* and *CLDN2*, respectively **(Table 4.3)**. The matching motif sequence identified upstream the TSS of the longest CLDN2 transcript variant overlaps an ENCODE-annotated DNase I hypersensitivity peak in HepG2 cells.

CLDN2 **(Figure 4.17)** (MIM #300520) or Claudin-2 encodes an integral membrane protein that localises in and regulates tight junctions of kidney, liver and intestine (Sakaguchi et al, 2002). Sakaguchi and colleagues have shown that the 5'- flanking region of *CLDN2* contains HNF-1a as well as homeodomain protein (CDX-1 and CDX-2) binding sites that are conserved in mouse and human. Using *HNF1A*-deficient mice, they showed that *HNF1A* is required for the expression of *CLDN2* in the crypt-villus axis of the ileum and liver. Expression data using Affymetrix Mouse Genome 430 2.0 arrays from hnf1a -/- by Servitja et al also supported hnf1a-dependant *CLDN2* expression (downregulated 14-fold) (Servitja et al, 2009).

CUBN **(Figure 4.17)** (MIM #602997) or cubilin (Intrinsic Factor Cobalamin-Receptor) encodes a large (460 kDa) peripheral membrane protein that functions as the physiological receptor for intrinsic factor-vitamin B12 complexes (Seetharam et al, 1997). It has been shown to bind to apolipoprotein A-I (APOA1) and high density lipoprotein (HDL) with high affinity, thereby facilitating HDL uptake (Kozyraki et al, 1999). *CUBN* is predominately expressed in the intestine,

renal proximal tubule where it is crucial for normal reabsorption, and embryonic yolk sac (Seetharam et al, 1997; Kozyraki et al, 1999).

(ii) Pathway analysis shows enrichment for MODY and other physiologically relevant metabolic processes

A differentially regulated gene set defined by a more relaxed significance threshold (FDR-adjusted P ≤ 0.05) was used to search for enriched biological processes and molecular pathways in response to *HNF1A* overexpression in HepG2 cells. Gene ontology (GO) and KEGG pathway analyses on this subset revealed discrete cellular and metabolic processes associated with HNF1A dose induction and consistent with existing knowledge of *HNF1A* regulatory biology; these included Maturity Onset Diabetes of the Young (MODY), bile secretion, PPAR signalling, insulin resistance, and gluconeogenesis **(Figures 4.18A and 4.19A)**.

I extracted the names of differentially expressed genes driving MODY enrichment ($P= 3.80x10^{-3}$) in KEGG pathway analysis. Unsurprisingly, *HNF4A* (P=0.006), which is responsible for HNF4A-MODY (MODY1; MIM #125850) and is both a regulatory target and regulator of *HNF1A* was among the differentially expressed genes underlying the MODY annotation signal (Boj et al, 2001; Ferrer et al, 2002) **(Figure 4.18B)**. In addition to *HNF1A* and *HNF4A*, 3 of the 8 genes influencing the enrichment of MODY belong to different hepatocyte nuclear factor (HNF) gene families and encode transcription factors which play a key role in pancreatic beta-cell development and maturation: *HNF6A* or *ONECUT1* (P=0.008) and *HNF4G* (P=0.009), *FOXA3* or *HNF3G* (P=0.048) **(Figure 4.18B)**.

There is a substantial body of *in vitro* and *in vivo* evidence which suggests that *PKLR* (P=0.001), a gene which encodes pyruvate kinase (liver and red blood cell isozymes), is a direct transcriptional target of HNF-1a in liver (Párrizas et al, 2001; Boj et al, 2001; Servitja et al, 2009). Also, increased expression of *PKLR* by nearly 3-fold has been observed in the liver of diabetic mice compared to non-diabetic controls (Zhang et al, 2013).

The transcription factor-encoding genes *HHEX* (P=0.009) and *PAX6* (P= 4.48×10^{-5}) are both development regulators. *PAX6* controls the specification and differentiation of endocrine hormone producing cells and helps maintain the functional identity (insulin biosynthesis, sensitivity and glucose-stimulated secretion) of pancreatic beta-cells by repressing the expression of other islet cell genes (Swisa et al, 2017; Gosmain et al, 2012; Hart et al, 2013). *HHEX* encodes hematopoietically expressed homeobox protein, which is expressed in anterior visceral endoderm cells and controls their positioning in early development; after organogenesis, it is expressed predominantly in the liver, thyroid and ventral pancreas buds (Bort et al, 2004). Several common alleles near and within the *HHEX* locus have been associated with T2D susceptibility and beta-cell dysfunction in European populations (Zeggini et al, 2007; van Vliet-Ostaptchouk et al, 2008; Staiger et al, 2008).

To develop insight into the *HNF1A*-regulated hepatocyte secretome, I extracted the subset of up- or down-regulated genes driving a significant enrichment of GO terms with extracellular and/or peripheral membrane localisation annotations (i.e. GO:0043230, GO:1903561, GO:0070062, GO:0065010, GO:0044421, GO:0005576, GO:0005615) **(Figure 4.19A)**. Approximately 40% of total differentially regulated genes (FDR-adjusted $P \leq 0.05$) across tested doses

encode proteins which are secreted, harbour extracellular domains, and interact or associate with plasma membrane, and are thus likely involved in signalling/secretion pathways broadly regulated by *HNF1A*. Separate KEGG pathway analysis on the secretome gene-set showed enrichment for biological processes ($P < 0.01$) which include vitamin digestion and absorption, lipid metabolism (fat digestion and absorption, PPAR signalling), galactose metabolism, and glycolysis/gluconeogenesis **(Figure 4.19B)**.

Table 4.3. *in silico* survey of *HNF1A* DNA binding motifs upstream *CUBN* and *CLDN2*

Gene	TxDb.Hsapiens.UCSC.hg19 chromosomal location	JASPAR (5kb upstream TSS)	Jolma et al, 2013 (5kb upstream TSS)
CUBN	[1] chr10 [16865965, 17171816]	[1] [AGTTAATTTTTGCA] 4561-4574	[1] [AATTAATGATATACA] 3323-3337 [AGATAATGATTAAGT] 4799-4813
	[2] chr10 [16893234, 16949747]	-	[2/3] [ATATAATGATTATCA] 1050-1064 [CCTTAATTATCAACT] 1198-1212
	[3] chr10 [16911635, 16949747]	-	
CLDN2	[1] chrX [106143394, 106174091]	[1] [GGTTAAATTTTGAC]* 3555-3568 *overlaps Dnase I hypersensitivity peak from ENCODE analysis in HepG2 cells.	-
	[2] chrX [106161590, 106174091]	[2] [GGTTTATAATTTTT] 1321-1334	[2/3] [TTTAAATAATTTACA] 2201-2215/185-199
	[3] chrX [106163606, 106174091]	-	

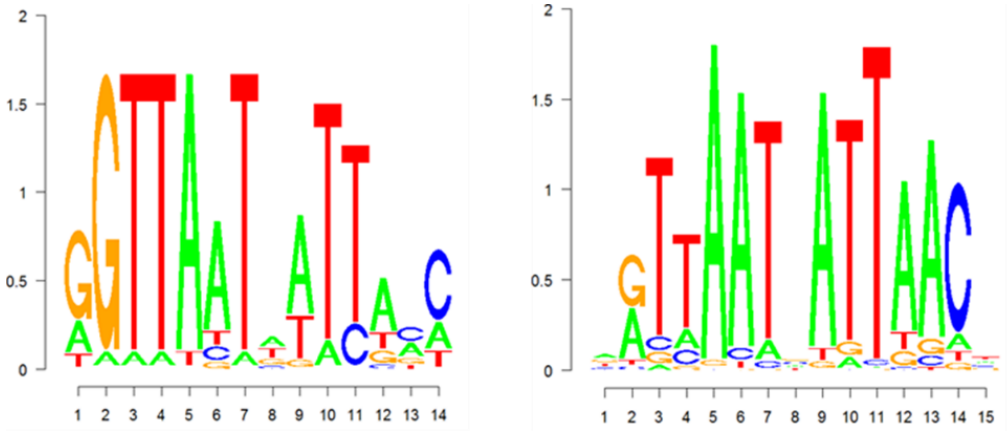


Figure 4.16. Position frequency matrices of *HNF1A* DNA binding motif sequences per the motif databases Jaspar (left) and Jolma et al 2013 (right) (MotifDB, R).

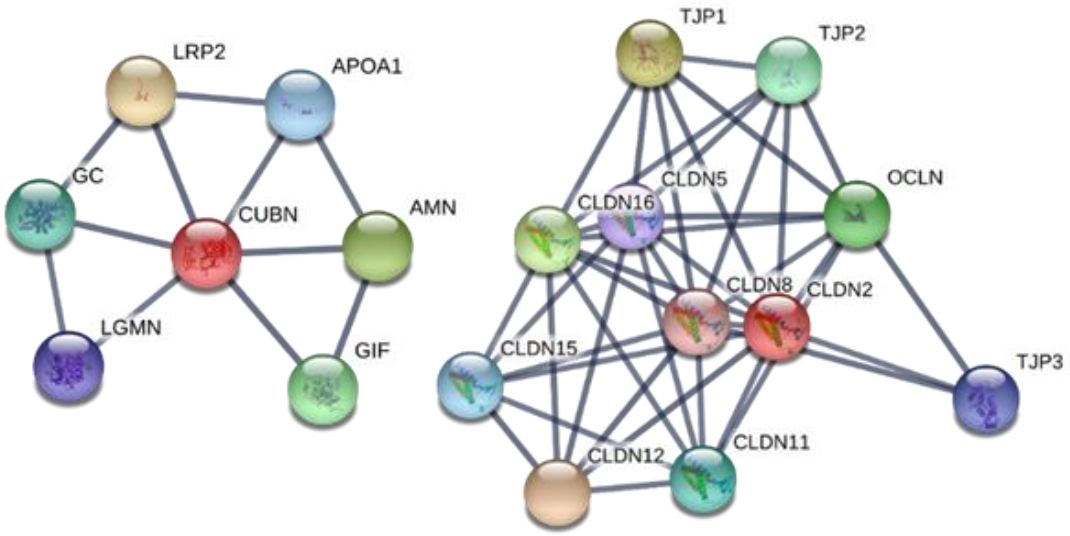


Figure 4.17. Protein-protein interaction networks of CUBN (left; PPI enrichment p-value 0.0439) and CLDN2 (right; PPI enrichment p-value <1.0e-16) from STRING (Search Tool for the Retrieval of Interacting Genes/Proteins) database v10.5). Network edges defined by highest confidence interaction score (0.900) based on 7 evidence channels or interaction sources (textmining, experiments, databases, co-expression, neighbourhood, gene fusion, co-occurrence).

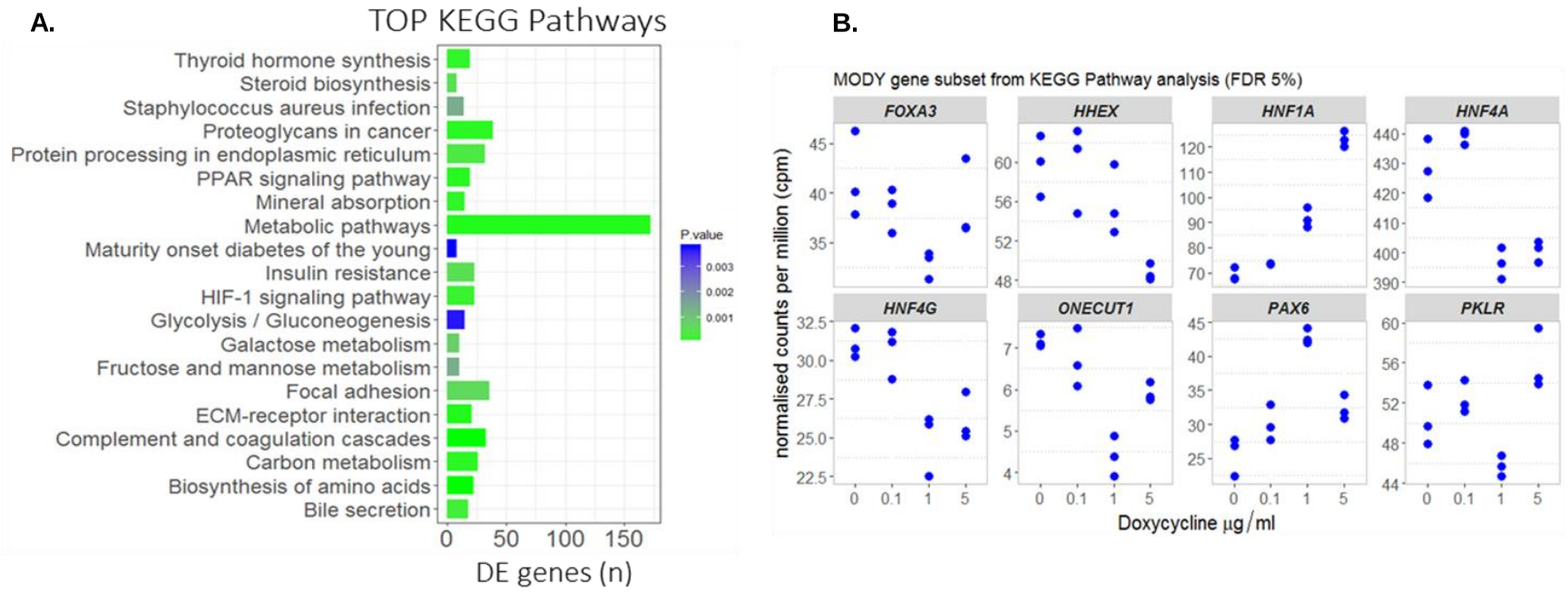


Figure 4.18. Pathway enrichment analysis (A) KEGG Pathway analysis using a subset of differentially expressed transcripts genes (FDR 5%) across all tested dox doses, illustrated as a function of the number of differentially expressed genes contributing to each annotated pathway and coloured with a gradient defined by the statistical significance of pathway enrichment. **(B)** Differentially expressed genes extracted from the KEGG Pathway annotation MODY which is enriched in Top KEGG analysis

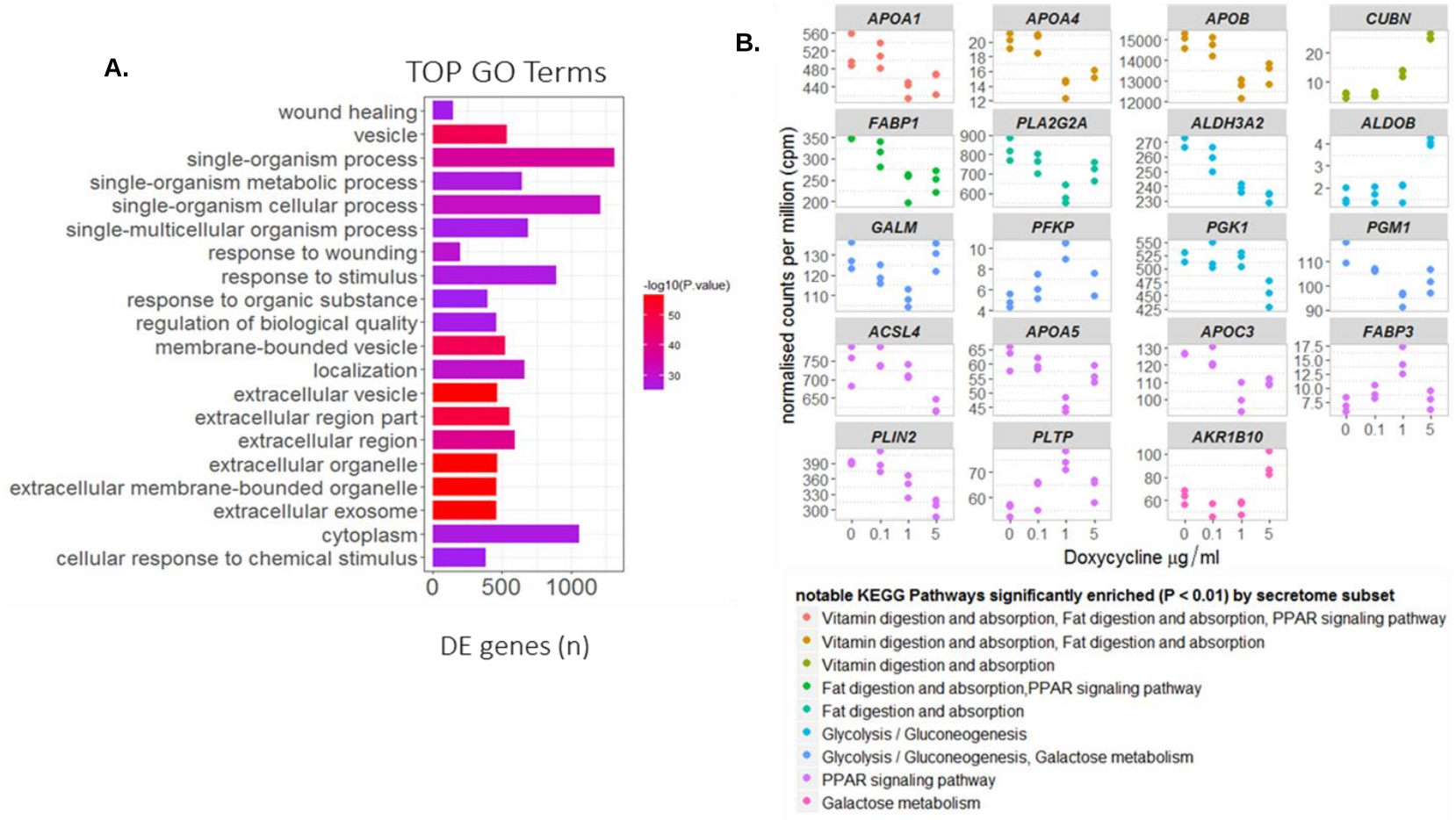


Figure 4.19. ***HNF1A*-regulated hepatocyte secretome (A)** Go Term analysis using a subset of differentially expressed transcripts genes (FDR 5%) across all tested dox doses, illustrated as a function of the number of differentially expressed genes contributing to each annotated pathway and coloured with a gradient defined by the statistical significance of pathway enrichment. **(B)** Differentially expressed transcripts (FDR 5%) involved in major absorption, secretion and signalling pathways extracted from Top Go terms with extracellular/membrane-bound annotations

4.4 Discussion

4.4.1 Summary and conclusions

The aim of this project was to develop a MAVE for *HNF1A* that would enable massively parallel molecular annotation of >12K alleles on the basis of transcriptional activity **(Figure 4.1)**. To this end, I explored two different assay development strategies which varied on the basis of transactivation readout **(Figure 4.3)**:

Strategy A

An **artificial (fluorescent) reporter gene** downstream a high binding-affinity promoter sequence in an easy-to-manipulate cell system.

With this strategy, I was unable to achieve the level of signal intensity required to segregate thousands of allele at once. I used HeLa cells as an *in vitro* model as the majority of *HNF1A* functional characterization data have been generated using luciferase reporter assays in these cells. One obvious limitation here is that HeLa cells do not endogenously express *HNF1A*, and thus likely lack the regulatory machinery (e.g. co-factors) required to enhance *HNF1A* transcriptional activity.

My preliminary experiments which involved lipid-based delivery **4.3.1(i)** of expression (*HNF1A*) and promoter-reporter gene (hALB_mCherry) sequences demonstrated enough promise to proceed with lentiviral-based delivery methods, which more closely mimic MAVE conditions and meet high-throughput expression control requirements **4.3.1(ii)**. The lentiviral delivery experiment using recoded *HNF1A* sequences showed discrimination that was both allele- and doxycycline dose- dependent **(Figure 4.10; subsection 4.3.1(ii))**, though not

with the degree of sensitivity required to render the mCherry reporter system translatable to the high-throughput context. Altogether, the experiments provided solid validation of the foundational components of the screen, yet suggested that an alternative approach (i.e. strategy B) was needed to attain higher signal-to-noise and maximal induction by HNF-1a to accommodate a broad and subtle range of functional effects.

Strategy B

Fluorescent staining of an **endogenous cell surface-expressed reporter gene** which exhibits high fold-change inducibility by *HNF1A* in a physiologically relevant cell system.

The limitations associated with strategy A prompted a shift towards strategy B (endogenous reporter gene readout). I developed a systematic and generalizable pipeline for building a transactivation-centric MAVE for *HNF1A* using endogenous reporter genes as discriminators of molecular dysfunction; my pipeline is as follows **(Figure 4.3)**:

- *in vitro* model selection and characterisation
- dox dose-dependent stimulation of the transcription factor-encoding gene
- identification of differentially regulated transcripts by RNA-seq
- prioritisation of a subset of cell-surface expressed endogenous targets and validation for MAVE-compatibility:
    - signal-to-noise (null mutant vs WT)
    - dynamic range in response to an allelic series

I chose the human hepatoma line HepG2 as a model for this strategy as it fulfils the criteria of biological context and experimental feasibility. I showed that *HNF1A* is robustly expressed on the cytogenomic, transcript, and protein level in HepG2

cells **(Figure 4.11A, B & C; subsection 4.3.2(i))**. The biological relevance of human hepatoma HepG2 cells for an *HNF1A* MAVE is supported by multiple lines of evidence which included significant differential expression of canonical HNF-1a target transcripts **(Figure 4.11E)** and the enrichment of physiologically and clinically relevant molecular processes and pathways **(Figure 4.18).** A HepG2 MAVE model facilitated exploratory analysis of the HNF1A-regulated liver secretome: a list of differentially regulated secretary/peripheral membrane-associated protein (FDR 5%), with potential to serve as clinical biomarkers of *HNF1A* dysfunction **(Figure 4.19)**. In addition to revealing known targets and disease-relevant transcriptional networks, the *HNF1A* MAVE development pipeline uncovered two novel target transcripts which had not previously been described as *HNF1A* transcriptional targets in human liver: *CUBN* and *CLDN2* **(Figure 4.11D).**

To identify endogenous cell surface-expressed *HNF1A*-regulated transcripts, I induced *HNF1A* expression in a dox dose-dependent manner in HepG2 cells **4.3.2(ii) (Figure 4.11B)**. RNA from samples that showed dose-dependent induction of *HNF1A* was sequenced (Oxford Genomics Centre, WHG) to identify differentially regulated targets. A total of 729 transcripts were differentially expressed (FDR 1%) across dox doses (0-5 µg/ml) **4.3.2(iii) (Figure 4.11D)**. I validated HNF-1a regulation of cell surface-expressed targets, *CUBN*, *CLDN2*, *TMEM27*, and *TM4SF4* by qPCR and western blot analysis **4.3.2(iv) (Figure 4.13)** and evaluated their ability to sensitively discriminate a series of HNF1A alleles (with strong genetic, clinical, and functional evidence) on the basis of molecular function **4.3.2(v).** I achieved robust allelic segregation using TM4SF4 as a readout in flow cytometric analyses **(Figure 4.15)**. Currently, TM4SF4, is

being assessed for its ability to discriminate >12K alleles in the HepG2 MAVE **4.4.2(ii)**.

The endogenous reporter MAVE-building pipeline I have developed demonstrates that transcriptomic profiling in high-throughput-compatible systems yields biologic insights (expression of cell-surface transmembrane glycoprotein TM4SF4 regulated by HNF-1a) that can be immediately translated towards rapid and scalable functional applications (TM4SF4 robust marker for massively parallel functional scanning of *HNF1A* alleles).

The pipeline can be used to develop endogenous reporter-based MAVEs for other actionable, disease-relevant transcription factor-encoding genes, such as *HNF4A*. Alleles in *HNF4A* are rarer (~5% compared to 20-50% in *HNF1A*) though, much like *HNF1A* variant carriers, individuals with diabetes and a deleterious allele in *HNF4A* can also be spared from insulin injections and transfer successfully to treatment with sulphonylurea (Pearson et al 2005; Shepherd et al, 2003; Ellard et al, 2008). An accurate molecular genetic diagnosis also informs pregnancy management as damaging *HNF4A* alleles are associated with fetal macrosomia (increase in birth weight by ~800g) and transient neonatal hypoglycaemia (diazoxide-responsive) (Pearson et al, 2007). *HNF4A* and *HNF1A* are members of a complex cross-regulatory transcriptional circuit (Boj et al, 2001; Ferrer et al, 2002). Based on an *HNF1A* and *HNF4A* mutation update published in 2013, of the total 517 variants identified in the 2 genes, *in vitro* or *in vivo* functional characterisation is available for only about 60 (Colclough et al 2013). A comprehensive catalogue of MAVE-based functional data would be useful for carriers of alleles in *HNF4A*, where high fidelity sequence-function relationships can impact treatment and management.

By virtue of their novelty and technical complexity, MAVEs (particularly those which interrogate all possible natural genetic perturbations) remain largely limited to a single functional readout (Starita et al, 2017; Majithia et al, 2016). Transcriptional activity represents one of multiple biochemical functions (protein abundance, protein stability, DNA binding, nuclear localisation) through which a disease-causing variant within a transcription factor might act. MAVEs testing multiple disease-related functional elements and molecular phenotypes, including those driven by splice and regulatory alleles, would certainly yield more robust estimates of pathogenicity (Starita et al, 2017). Until this is technologically, analytically and financially more tractable in the high-throughput context, rigorous prioritisation of functional elements (*HNF1A* cDNA) and signatures (transactivation) is required.

4.4.2 Ongoing efforts: application of an anti-TM4SF4 HepG2 MAVE for *HNF1A*

(i) Collaboration status

Collaborators on the *HNF1A* MAVE-building project have not progressed as far along the experimental pipeline **(Figure 4.3)**. During her time at the Broad Institute, Laeya Najmi focused her efforts on achieving complete Cas9-mediated *HNF1A* knockout in two other human hepatoma lines (Hep3b and Huh-7). This work was prioritised over *HNF1A* target identification in an overexpression context. Upon achieving successful generation of *HNF1A* knockout liver lines, Najmi induced *HNF1A* expression in a dox dose-dependent manner and conducted RNA-seq experiments as I have done with HepG2 cells. Majithia has assisted in differential expression analysis of her dataset, from which she has yet

to prioritise and validate targets for *in situ* RNA-FISH. Although, prior to RNA-seq, Najmi was able to demonstrate via qPCR that the targets I which I identified and prioritised in HepG2 cells **4.3.2(iv) (Figure 4.13A)**, demonstrated dose responsive expression to *HNF1A* in the cell lines she selected.

(ii) HNF1A site saturation variant library (n=12,615 permutations)

The work described in this chapter represents the most arduous phase of the MAVE workflow. I am leading the next phase of implementation which involves preparing the allelic library and using it in carefully designed TM4SF4-induction experiments in HepG2 cells for massively parallel functional stratification.

As described in subsection **4.2.8**, a US-based oligo synthesis company, Twist Bioscience, was responsible for generating a high quality *HNF1A* site saturation variant library based on codon sequence design files supplied by Dr. Majithia. Twist had synthesised and delivered an *HNF1A* library in Spring of 2017. Although, upon analysis of the NGS data that was generated to verify cloning of each allele into the lentiviral vector **(Figure 4.4B)**, I discovered major discrepancies between the original design reference files and the sequence data **(Figure 4.20).** Not only were there 19,549 non-designed codon alterations (which occurred at higher frequency than designed ones) **(Figure 4.20)**, but also, 30% of the designed events were completely missing from the data. After further investigation it became apparent that Twist had used the wrong design files and as a consequence mistakenly synthesised the wrong product, due to internal miscommunication.

Failure on behalf of Twist to generate a usable product on the first round of production meant that I would not be able to interrogate the library in the MAVE

system I developed before writing my thesis. Twist delivered a new library in November of 2017 which they had cloned into the viral vector *en masse*, and supplied the associated NGS read count data. This time, Twist used the correct design files for sequence alignment. I have graphically depicted library completeness and distribution based on read counts and pass/fail quality scores **(Figure 4.21).** Over 95% of the designed alleles are represented in this library with an average of 30 reads per allele after pooled high-throughput cloning **(Figure 4.21).**

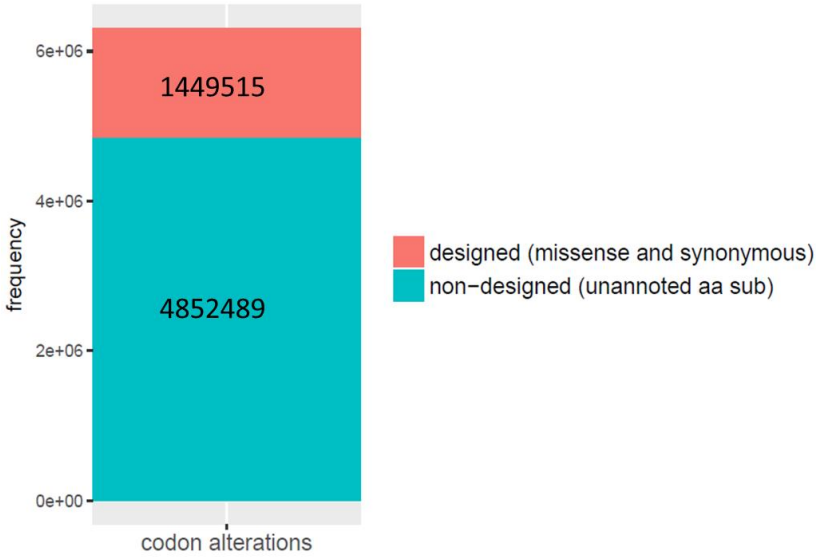


Figure 4.20. Frequency of designed versus non-designed codon alterations from the first HNF1A SSVL synthesised by Twist Bioscience (Spring 2017).

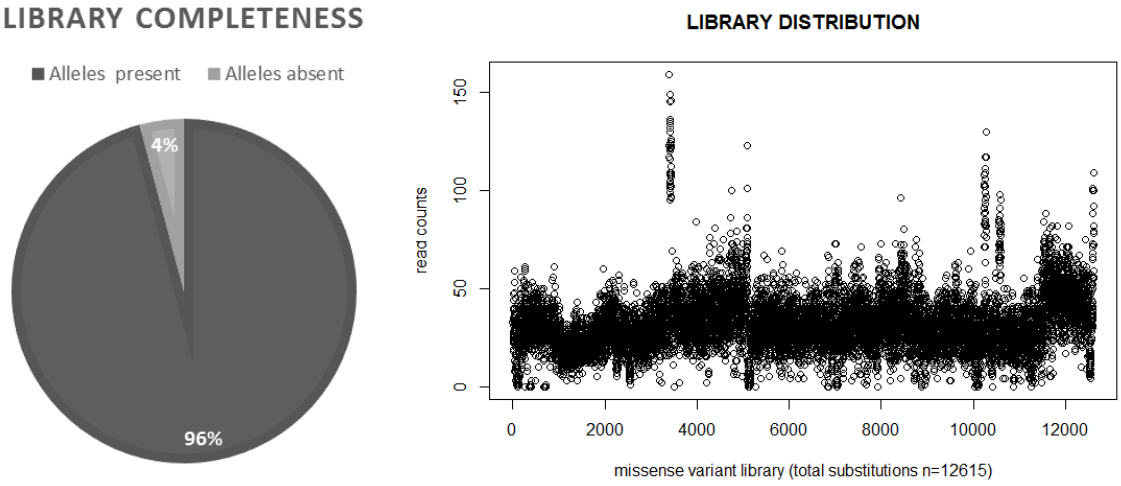


Figure 4.21. Verification of library completeness (designed alleles present v designed alleles undetectable; left) and library distribution (abundance of each designed substitution at each position based on read counts after high-throughput cloning; right); Second HNF1A SSVL synthesised by Twist Bioscience (Autumn 2017).

(iii) High-throughput-compatibility of anti-TM4SF4 HepG2 MAVE for HNF1A

Judging by the various technical challenges I have faced throughout the course of this project, it is indeed difficult to predict how well the library from Twist will perform and also how effective an anti-TM4SF4 HepG2 MAVE will be at segregating >12K alleles.

However, based on the distributions of *HNF1A*-induced TM4SF4 fluorescent signals in FACS experiments **(Figures 4.14 and 4.15)**, I have modelled some of my own predictions **(Figure 4.22)**. I have assumed that the *HNF1A* library TM4SF4 induction trend will span the full range of the *HNF1A* allelic series I tested **4.3.2(v)**. If this proves to be the case upon screening the library, then, at the highest dose of *HNF1A* induction (5 µg/ml dox), a three-way cell sort based on three activity bins, with cut-offs roughly corresponding null (p.T260M/unstimulated WT), intermediate/moderately active (p.E508K) and fully active (WT) TM4SF4 induction trends, would be possible **(Figure 4.22)**. The appropriate control samples would need to be run in parallel to the pooled library transduction to facilitate this FACS-gating (i.e. independent inductions of WT *HNF1A* for WT-like gating, HNF1A-p.E508K for intermediate gating, and WT *HNF1A* at zero dox or loss-of-function allele *HNF1A* for background gating).

If the library application is successful and the system performs as predicted, raw function score will be assigned based on frequency of each allele per activity bin using eBayes-moderated log-likelihood functions. The scores from different induction conditions (dox doses, replicates) will be combined and used to build and train a high performance *HNF1A* variant classifier by supervised machine learning.

We have a plethora of genomic, functional, and clinical *HNF1A* datasets at our disposal in Oxford to both test and enhance the performance of a variant classifier built using high-throughput experimental data. The pathogenicity odds derived from *HNF1A* MAVE data will be calibrated against function annotations from gold standard comprehensive low-throughput assays such as those described in Chapter 3. Well-calibrated MAVE-data derived variant scores will enable a re-evaluation of every single documented clinical diagnosis made based on an *HNF1A* missense allele.

The ultimate aim is to build a multivariate probabilistic model based on a combination of a) MAVE-derived high-throughput functional data (ideally multiple MAVEs for a single gene in multiple cell systems and genetic backgrounds), b) smaller scale comprehensive molecular assessment efforts (**Chapter 3**), c) *in silico* prediction, d) allele frequency, e) and clinical datasets with information on BMI, age-of-diagnosis, family history, biomarker profiles, SU-sensitivity of HNF1A variant carriers. Likelihood estimates from such a model would be made accessible, publically available, user-friendly, and interactive to help researchers and clinicians, variant carriers and their families, understand and make informed judgements of ascertained variants in *HNF1A*.

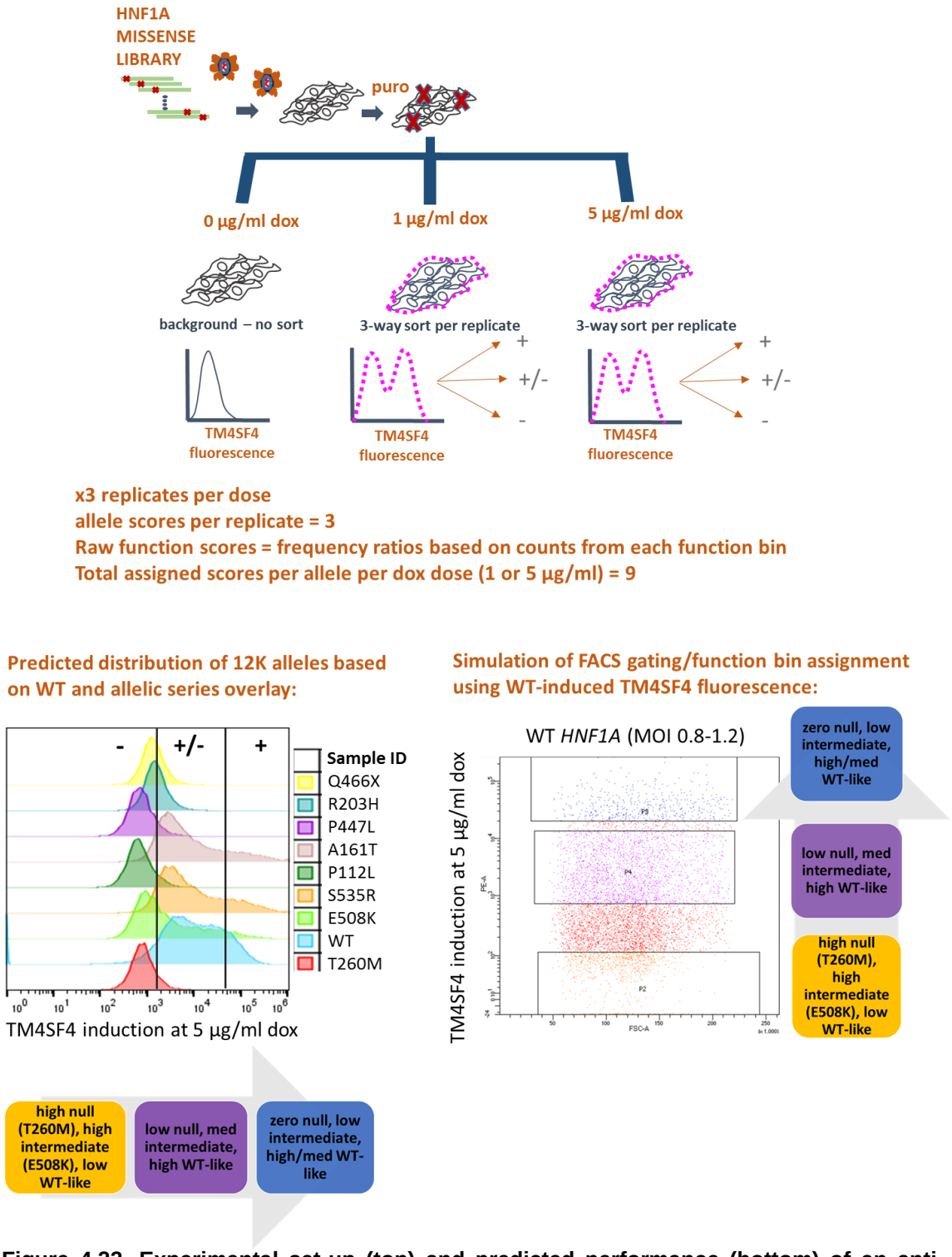


Figure 4.22. Experimental set-up (top) and predicted performance (bottom) of an anti-TM4SF4 MAVE for *HNF1A* in HepG2 cells. Transduction of HepG2 cells with *HNF1A* saturation mutagenesis library using a limiting dilution (MOI 0.8-1.2) and at a scale which ensures high independent representation of each allele (x100 events per allele or more). Dox doses serve as technical replicates since rates of TM4SF4 induction are relatively equal at these concentrations. This is followed by a three-way sort of *HNF1A* library-infected HepG2 cells based on three activity bins with cut-offs roughly corresponding null (p.T260M/unstimulated WT), intermediate/moderately active (p.E508K) and fully active (WT) TM4SF4 induction.

5

Comprehensive functional assessment of *GCK* missense alleles identified in clinical genetic testing

5.1 Background

The glucokinase (*GCK*) gene (MIM #138079) encodes a glycolytic enzyme that is a structurally and functionally distinct member of the hexokinase family (hexokinase I-IV) (Magnuson 1990; Iynedjian 1993). It operates as a glucose sensor in pancreatic islet beta-cells and tightly controls the rate of glucose metabolism in the hepatocyte (glycogen and glucose production) (Magnuson 1990; Iynedjian 1993; Matschinsky et al, 2002). Glucokinase initiates the first step in glycolysis by phosphorylating glucose – its primary substrate – on the sixth carbon with Mg-ATP as a co-substrate, to generate glucose-6-phosphate (G6P) (Magnuson 1990; Iynedjian 1993; Matschinsky et al, 2002).

5.1.1 Glucokinase mutational spectrum

Rare genetic variation in *GCK* gives rise to a range of pathologies. A correct molecular diagnosis helps a) distinguish glucokinase diabetes from other diabetes subtypes (T1D, T2D, other forms of monogenic diabetes) and b) decide treatment and management strategies as they vary with each disorder.

(i) GCK-MODY

Individuals with GCK-MODY carry an autosomal dominantly inherited heterozygous inactivating variant in *GCK* (Hattersley et al, 1992; Froguel et al, 1993). They have stable, moderate asymptomatic fasting hyperglycaemia that is present from birth (5.5–8.3 mmol/L) and an HbA1c 40–60 mmol/mol (~5.8–7.6%) (Steele et al, 2013). It is also characterised by a small OGTT increment [(2hr glucose) – (fasting glucose)] < 3 mmol/l (Ellard et al, 2008; Stride et al, 2002). There is only slight deterioration of glycaemic status with age that is consistent with what is observed in the general population, and no microvascular

complications (hyperglycaemia is mild, and lower than threshold above which the risk for diabetes-related complications increases) (Velho et al, 1997; Chakera et al, 2015). GCK-MODY does not require pharmacological intervention for the vast majority of cases (Stride et al, 2014). In fact, individuals with GCK-MODY do not respond to insulin or oral glucose lowering agents due to tight homeostatic regulation of glycaemia at an elevated glucose 'set-point' (Steele et al, 2013; Chakera et al, 2015).

A GCK-MODY diagnosis also influences clinical management during pregnancy. In GCK-MODY pregnancy, foetal genotype determines response to maternal hyperglycaemia with foetal insulin; and since insulin acts as growth factor in utero, this translates to increased risk of macrosomia (Hattersley et al, 1998; Group HSCR et al, 2008; Chakera et al, 2015). If the foetus inherits the variant (50% chance), it will grow normally (Hattersley et al, 1998). On the other hand, if the foetus does not inherit the GCK-MODY allele, it will produce excess insulin to respond to maternal hyperglycaemia, resulting in excess growth (Hattersley et al, 1998). In this case, insulin treatment is recommended for the mother (Spyer et al, 2001). Pregnant women with GCK-MODY are advised to monitor foetal growth and evaluate risk of macrosomia via fortnightly scans from 26 weeks' gestation in which foetal abdominal circumference is measured (Spyer et al, 2009; Chakera et al, 2015).

(ii) Permanent Neonatal Diabetes Mellitus (PNDM)

This glucokinase diabetes subtype is caused by homozygous or compound heterozygous *GCK* biallelic inactivation (Njolstad et al, 2001). It is present from birth and diagnosed typically within the first 6 months of life, and is a rare severe

form of diabetes which requires lifelong insulin therapy (Njolstad et al, 2001; Bennett et al, 2011; Durmaz et al, 2012; Rubio-Cabezas et al, 2008). A total of 38 homozygous *GCK* cases have been reported (Raimondo et al, 2014). Homozygous inactivation of *GCK* is associated with very low birth weight (Gloyn 2006).

(iii) Hyperinsulinemic Hypoglycemia (HH)

Autosomal dominant inheritance of heterozygous activating (gain-of-function, GoF) mutations in *GCK* gives rise to Hyperinsulinemic Hypoglycemia, also known as Persistent Hyperinsulinemic Hypoglycemia of infancy (PHHI) (Glaser et al, 1998). They increase the enzyme's affinity for glucose and drastically reduce the pancreatic beta-cell glucose threshold for glucose-stimulated insulin secretion (Matschinsky et al, 2002; Gloyn 2005). Most individuals with GCK-HH respond well to treatment with diazoxide (K_{ATP} channel activator) (Gloyn 2003).

5.1.2 Glucokinase and type 2 diabetes (T2D)

Prior to the advent of GWAS, the only ascertained genomic association with fasting glucose (FG) was at the *GCK* locus (Weedon et al, 2005; Weedon et al, 2006). Large-scale T2D association studies and meta-analyses of glycaemic trait genetics, in European ancestry populations, have revealed multiple genome-wide significant associations between common single nucleotide polymorphisms (SNPs) at the *GCK* locus and T2D as well as various metabolic and anthropometric traits (FG, FI [fasting insulin], HbA1c, BMI, 2hr glucose) (lead T2D association SNP at *GCK*: rs10278336, MAF 50%, OR 1.09; glycaemic trait intronic *GCK* SNP: rs730497, MAF 16.2%) (Gaulton et al, 2015; Morris et al, 2012, Mahajan et al, 2017; Dupuis et al, 2010; Manning et al, 2012).

The critical role *GCK* plays in glucose homeostasis has rendered it an attractive anti-diabetic drug target, particularly in the context of T2D where glucose sensing (glucose-stimulated insulin secretion) in the pancreatic beta cells and hepatic glucose uptake and storage are typically impaired (Grimsby et al, 2003; Matschinsky et al, 2009). Synthetic small molecule glucokinase activators (GKAs), which bind the allosteric activator site of the enzyme and increase glucose affinity, have demonstrated great promise in animals and early human trials in T2D patients and controls (Matschinsky, 2009). Results from Phase II clinical trials (Piragliatin, Roche-RO4389620) suggest that GKA alternatives need to be explored considering their shortcomings with respect to long term efficacy, and adverse effects including hypoglycaemia and increase in serum triglyceride levels (Bonadonna et al, 2010; Nakamura and Terauchi, 2015).

A novel compound targeting GCK has emerged recently as an attractive pharmacotherapy after demonstrating an ability to restore mutant GCK kinetic activity to wild type levels without, unlike GKAs, altering substrate affinity [34, 45]. Phosphorylated alpha helices of the BH3 domain of proapoptotic human BCL-2 family protein BAD bind non-allosterically to GCK, restore activity, enhance glucose metabolism and insulin secretion, and increase beta cell mass and survival *in vitro* and *in vivo* (Ljubicic et al, 2015; Szlyk et al, 2014).

5.1.3 Chapter preview

A correct assessment of the impact of *GCK* alleles on glycaemic status requires understanding the impact of the identified allele on protein activity. An accurate molecular genetic diagnosis does not only optimise prognostic evaluation and treatment strategy but also allows for appropriate genetic counselling and

screening. In this chapter, I describe my efforts to assist in the interpretation of two novel *GCK* alleles (p.F133L/N and p.E157K/p.E157K) - identified in two distinct, phenotypically complex pedigrees – by providing comprehensive functional follow-up.

To understand the molecular impact of p.F133L and p.E157K on GCK activity, I adopted a rigorous multi-tiered functional characterisation pipeline, which interrogates catalytic activity and protein stability reliably *in vitro* **(Figure 5.1)**. The workflow is outlined in **Figure 5.1**: first, variants are introduced to human GCK cDNA sequence in an amplification reaction using mutagenic primers **(Table 5.1)**; then, wild-type variant-harbouring *GCK* cDNA **(subsection 5.2.1)** is used to synthesise recombinant GCK protein *in vitro*; finally, kinetic activity (affinity for glucose and ATP) is measured using glucose 6-phosphate dehydrogenase-coupled assays, and over a temperature titration in thermostability assessments.

An additional molecular mechanism was explored in the case of the homozygous missense allele p.E157K: activation by pharmacological GCK activator (Roche RO0281675) (Grimsby et al, 2003; Matschinsky, 2009). The GKA used in this study was the first published and most thoroughly researched: it has not reached clinical trials though has been shown to a) increase K_{cat}, ATP_{km}, lower glucose $S_{0.5}$ and Hill coefficient *in vitro* at 5mM basal glucose, b) improve glucose stimulated insulin secretion (GSIS) in isolated rodent islets, c) reduce blood glucose to hypo- and normo-glycaemic ranges levels in normal and diabetic animal models, respectively, via enhanced insulin secretion and reduced hepatic gluconeogeneisis (Matschinsky, 2009; Grimsby et al, 2003).

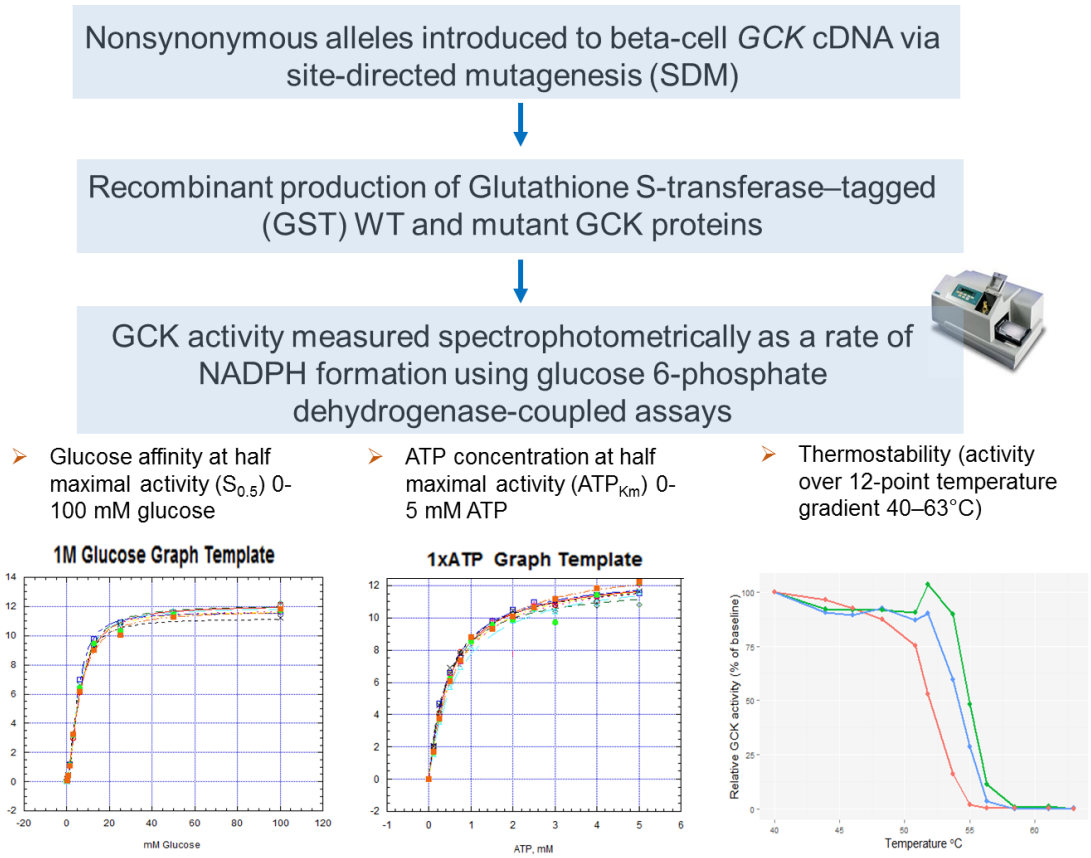


Figure 5.1. Experimental pipeline for *in vitro* functional characterisation of *GCK* variants. The pipeline begins with incorporating the allele(s) of interest into human *GCK* cDNA (NM_000162.3; beta cell isoform) via site-directed mutagenesis using relevant mutagenic primers. This is followed by *in vitro* recombinant production of allele-harbouring glucokinase protein GCK. The GST tag is used to separate and purify the protein. GCK activity is then evaluated based on affinity for glucose (0–100 mmol/l) and ATP (0–5 mmol/l) in glucose-6-phosphate dehydrogenase-linked assays. GCK stability is also evaluated over a 12-point temperature gradient spanning 40–63°C at a final glucose concentration 8 mmol/l using the same spectrophotometric assay. Standard assay ATP and glucose concentrations may vary depending on the molecular severity of the variant(s) under interrogation.

5.1.3 Case reports

(i) GCK p.F133L (c.397T>C)

Referral from Professor Louis Philipson, UChicago Medicine (2016); the proband and his father were studied at Seattle Children's Hospital (2012). Clinical features and family history limited to what was available and provided upon request to assist in interpretation via functional testing.

The proband (III.1) was diagnosed at 6 years of age with mild fasting hyperglycaemia **(Figure 5.2)**. At the year of referral, he was 17 years old with a BMI of 33.3 kg/m^2 and mild fasting hyperglycaemia (~7-8 mmol/L). He tested negative for pancreatic autoantibodies [GAD 0, ICA-512 0; ZnT8 0.004 (ref <0.03), insulin 0.012 (ref <0.013)]. HbA1c levels were marginally elevated 45-51 mmol/mol (6.3-6.8%). From 2012, and after years of insulin therapy (Lantus and Novolog), treatment was changed to Metformin (500 mg, once daily).

The proband's paternal grandfather (I.2) was clinically diagnosed with T1D, although he had not been tested for presence of islet autoantibodies **(Figure 5.2)**. The proband's father (II.2), 55yo at year of referral, was diagnosed with diabetes at age 15y and has since been on insulin pump treatment (Humalog, fast-acting mealtime insulin). Like his son, he tested negative for islet autoantibodies. Clinical examination results from 2015 show he had a BMI of 30.5 kg/m^2, HbA1c levels over 64 mmol mol^{-1}, blood glucose of ~14 mmol/L, and low-level C-peptide at 99 pmol/L. Blood glucose monitoring revealed average blood sugar level of 10 mmol/L, ranging from ~7 mmol/L between 9pm and midnight and an average of ~11.5 mmol/L between 3-6pm. He is also on lipid-lowering treatment (statins) and treatment for hypertension.

GCK sequencing identified a heterozygous nonsynonymous missense variant in the proband and his father (c.397T>C, p.F133L/N). The proband was also tested for presence of *HNF1A* and/or *HNF4A* alleles; no variants were reported and it is unclear whether the father had undergone similar testing. After discovering the *GCK* variant, there were attempts to enhance endogenous insulin secretion of the father - thought to be downregulated by high doses of insulin - by reducing the pump (Total Daily Dose = 0.73 units/kg/day). These were moderately successful as pump was reduced by 50% though he remains on basal-bolus insulin regimen in addition to treatment with oral agents: sulphonylurea (glimepiride twice-daily, 4mg) and Metformin (500mg, 2 tablets twice-daily).

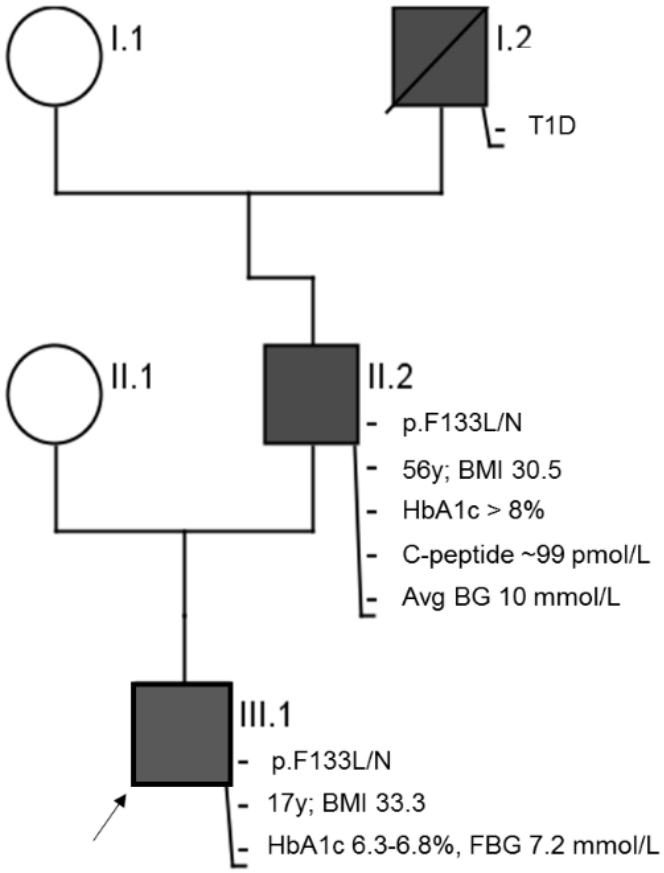


Figure 5.2. Pedigree of the proband with heterozygous *GCK*-p.F133L variant. Circles and squares represent females and males, respectively. Filled shapes represent hyperglycaemic individuals. An arrow indicates the proband. At the year of referral, the proband was 17 years of age and exhibited mild fasting hyperglycaemia (~7-8 mmol/L). Reported clinical information on each of the family members is provided underneath each symbol, including body mass index (BMI), FBG (fasting blood glucose), HbA1c (glycated haemoglobin), C-peptide, type of diabetes with which the individual was diagnosed (T1D = type 1 diabetes), as well as age at diagnosis.

(ii) GCK p.E157K c.469G>A

Referral from Professor Sian Ellard and Mr. Kevin Colclough, Molecular Genetics Laboratory, Royal Devon and Exeter NHS Foundation Trust (2016); the proband was studied at QEQM Hospital, Margate (2015). Clinical features and family history limited to what was available and provided upon request to assist in interpretation via functional testing.

The proband (IV.5) was diagnosed with T1D at 5 years of age – she was 18 at the year of referral. She had a reported history of polyuria and polydipsia. At first presentation, she was not acidotic. Subcutaneous insulin (Mixtard twice-daily regimen) was prescribed as treatment. At 12y, this was replaced with basal-bolus regimen. Blood tests were undertaken as part of a clinic visit at age 15 and revealed the following: blood glucose 13 mmol/l, blood ketones 0.3 mmol/l, blood C-peptide levels 497 pmol/L, negative for GAD and islet antigen 2 autoantibodies. Retinopathy screen, also conducted at 15y, showed mild-moderate non-proliferative diabetic retinopathy. At the year she was examined (2015), her BMI was 22.2 kg/m^2 (slim). It was also reported that she has maintained Hb1Ac levels of 70-85 $mmol\ mol^{-1}$ over the years and a glucose profile which seemed to vary between 9-18 mmol/L **(Figure 5.3)**.

At age 16, genetic testing was undertaken for the proband. *GCK* sequencing identified a homozygous nonsynonymous missense variant in the proband (c.469G>A, p.E157K/p.E157K). There is no information on whether other monogenic diabetes genes have been tested in the proband. At the time of referral, none of the proband's family members had been genetically tested.

The proband was presumed to have monogenic diabetes after discovery of the *GCK* variant. She was started on oral hypoglycaemic agent (OHA) sulphonylurea Gliclazide, with which she did not seem compliant. Treatment was then changed to Levemir (insulin detemir; long-acting insulin) to achieve better control, after which the proband was transitioned to adult diabetic services for further care. There is a positive family history of hyperglycaemia – including cases of type 2 diabetes (T2D) and gestational diabetes (GD) in her maternal lineage **(Figure 5.3).** The proband's maternal grandparents (II.I and II.2) are first cousins whose children have all been diagnosed with some form of diabetes, and she has cousins reported to be on insulin therapy **(Figure 5.3)**.

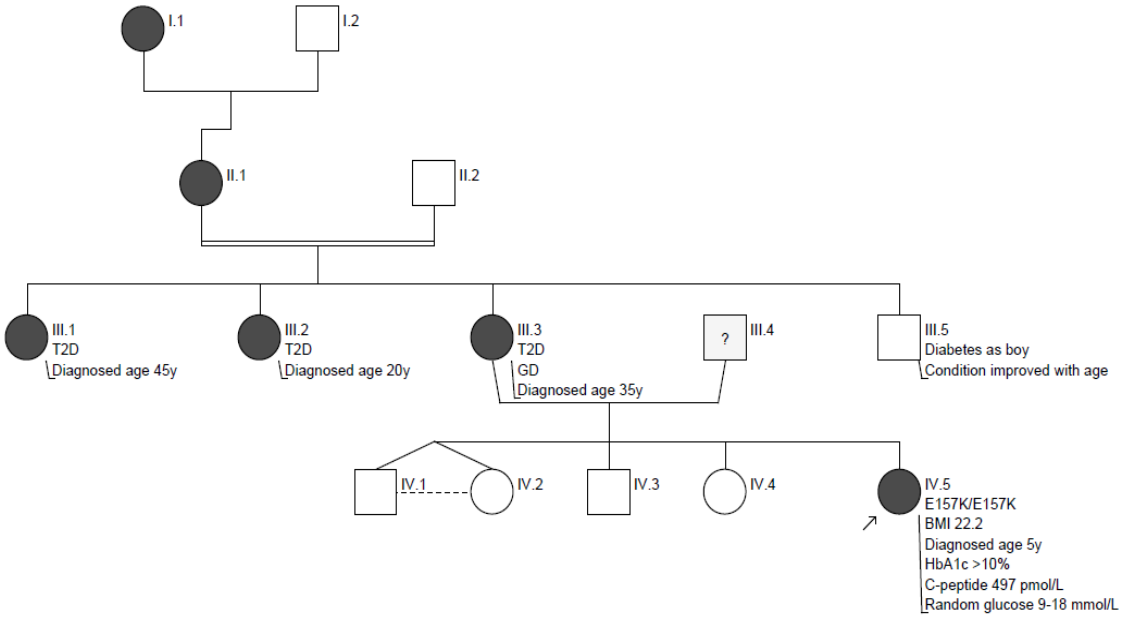


Figure 5.3. Pedigree of the proband with homozygous *GCK*-p.E175K variant. Circles and squares represent females and males, respectively. Filled shapes represent hyperglycaemic individuals. Shapes which enclose a question mark (?) denote individuals with an unknown phenotype. An arrow indicates the proband. At the year of referral, the proband was 18 years of age and exhibited poor glycaemic control (HbA1c of 70-85 mmol mol^{-1} and a glucose profile which varied between 9-18 mmol/L). Reported clinical information on each of the family members is provided underneath each symbol, including body mass index (BMI), HbA1c (glycated haemoglobin), C-peptide, randomly monitored glucose, type of diabetes with which the individual was diagnosed (type 2 diabetes T2D; gestational diabetes GD), age at diagnosis. Double horizontal line indicates consanguineous union.

5.2 Methods

5.2.1 Glucokinase mutagenesis and recombinant protein production

To investigate the function of the *GCK* alleles identified in clinical genetic testing (**Figures 5.2 and 5.3**) *in vitro*, the nonsynonymous missense variants were introduced to pancreatic beta-cell isoform of *GCK* (accession number NM_000162.3) in pGEX-3X-GST-hBGCK (**Figure 5.4**) using the QuikChange Site-Directed XL Mutagenesis Kit (Stratagene). Incorporation of desired nucleotide substitutions was confirmed by Sanger sequencing. Glutathione S-transferase–tagged human wild-type and mutant GCK were produced and purified as previously described (Liang et al, 1995; Beer et al, 2009) (**see Chapter 2 for details**)**.**

5.2.2 Functional assays

The adopted functional assessment pipeline is depicted in **Figure 5.1**. Glucose affinity was measured using a concentration range of 0-100 mM. This concentration was increased to 0-200 mmol/L for p.E157K. Affinity to adenosine triphosphate (ATP) was also measured using a concentration range of 0-5 mM. This concentration was increased to 0-20 mmol/L for p.F133L. Activation of GCK via allosteric binding of a synthetic glucokinase activators (GKA) (Roche R00281675, Patent number WO 0058293 [2000]) (Grimsby et al, 2003) was assessed in the presence of 0, 0.3, 1, 3, 9, and 27 µM GKA. Thermostability assessments of WT and mutant *GCK* were performed as previously described (Davis et al, 1999; Kesavan et al, 1997). The protein stability assay scheme is similar to that of glucose-dependent activity measurements (**Chapter 2**), except that the enzymes are analysed using a stable glucose concentration over a 12-

point temperature gradient spanning 40–63°C (in a tetrad thermal cycler; DNA Engine, MJ Research) (differences in stability between WT and mutant proteins are captured *in vitro* between 50-60°C) (Kesavan et al, 1997). In p.F133L stability assessments, the final glucose concentration was 8 mmol/L; in p.E157K tests the final glucose concentration was 22 mmol/L.

Table 5.1. Mutagenic primers used to synthesise *GCK* alleles

Mutagenesis primers	Forward 5'-3'	Reverse 5'-3'
p.E157K c.469Gaa>Aaa	cctttcctgtgaggcacaaagacatcgataagggc	gcccttatcgatgtctttgtgcctcacaggaaagg
p.F133L c.397Ttc > Ctc	gagtgcatctccgacctcctggacaagcatc	gatgcttgtccaggaggtcggagatgcactc

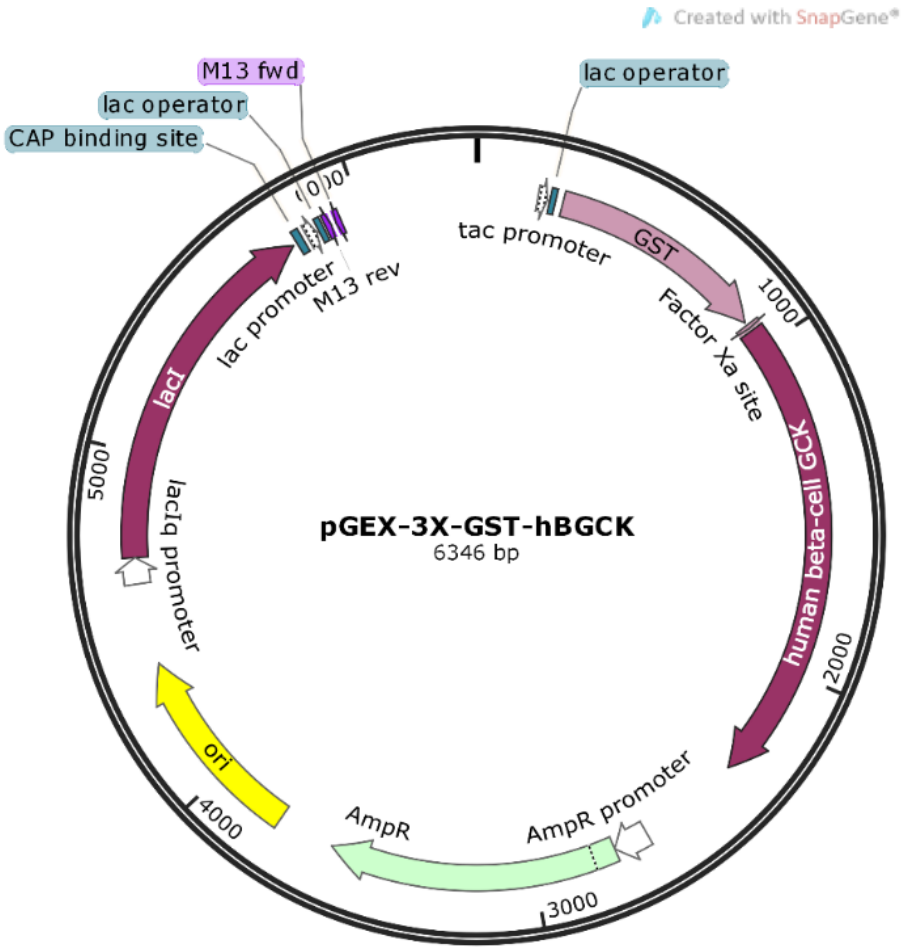


Figure 5.4. Map of plasmid DNA construct harbouring glutathione S-transferase–tagged human beta-cell *GCK* cDNA (hBGCK). Generated using SnapGene Viewer v2.6.2.

5.2.3 Bioinformatic analysis

Alamut Visual v.2.10.0 (Interactive Biosoftware) was used for integrative analysis of missense variant pathogenicity. Alamut® Visual produces a comprehensive digital report on pathogenicity clues for queried variants (HGVS Nomenclature [v15.11]) which includes:

- Functional protein domain
- Nucleotide conservation (phyloP and phastCons scores)
- Orthologues assignment
- Physiochemical difference by Grantham distance, which is a scoring system that reflects the evolutionary distance between amino acids based on physiochemical (dis)similarity. Scores range from 5 (tolerated/similar) to 215 (deleterious/dissimilar) (Grantham, 1974).
- Scores from missense prediction tools SIFT (predictions based on protein sequence conservation), PolyPhen-2 (impact of amino acid substitution on protein stability/structure and function), MutationTaster (integrative tool which predicts on basis of sequence conservation and data from ClinVar, ENCODE, 1000 Genomes Project), AlignGVGD (predictions based on integrated Grantham Variation, Grantham Deviation scores and multiple sequence alignments; GV score reflects observed range of evolutionary biochemical variation at each alignment position and GD score reflects biochemical difference between those properties and the interrogated substitution) (Eilbeck et al, 2017; Hicks et al, 2011).

- Annotation/presence in public databases dbSNP (Single Nucleotide Polymorphism database https://www.ncbi.nlm.nih.gov/projects/SNP/), ClinVar (https://www.ncbi.nlm.nih.gov/clinvar/), gnomAD (genome Aggregation Database http://gnomad.broadinstitute.org/), COSMIC (Catalogue of Somatic Mutations in Cancer http://cancer.sanger.ac.uk/cosmic), ESP (NHLBI GO Exome Sequencing Project http://evs.gs.washington.edu/EVS/)

The software complies with the ACMG/AMP Variant Interpretation Standards and Guidelines (http://www.interactive-biosoftware.com/alamut-visual/).

5.2.4 Graphical and statistical analysis

Glucose affinity ($S_{0.5}$), Hill number (n_H) and turnover number (K_{cat}) values were calculated using the Hill equation. ATP affinity (ATP_{Km}) was calculated using the Michaelis–Menten equation. All data fits were performed using Kaleidagraph v3.52 (Synergy Software). Relative activity indices (RAI) were calculated as previously described (Christesen et al, 2002). Relative stability indices (RSI) were derived from (TA50(mutant) -TA50(min))/(TA50(WT) - TA50(min)), where TA50 refers to the temperature point at which each protein displayed 50% of its activity at 40°C in a thermostability assessment. A five-parameter logistic regression model was used to fit thermostability data and calculate TA50 values for each variant in GraphPad Prism 6.0a (http://www.graphpad.com/).

Pedigrees were generated using CeGaT Pedigree Chart Designer (http://pedigree.cegat.de/). Hierarchical clustering was performed using Ward's minimum variance method (ward.d2) on Euclidean distances with the hclust() function (R statistical package http://www.r-project.org).

5.2.5 Structural modelling

Alleles were mapped onto the crystal structure of human glucokinase bound to glucose (closed form, Protein DataBank entry 1V4S, Kamata et al, 2004) using PyMOL v.0.99.

5.3 Results

5.3.1 *GCK* p.F133L (c.397T>C)

(i) Bioinformatic analysis and structural modelling

This p.F133L alteration in *GCK* is novel; it is not reported in the literature or general databases nor is it listed as a polymorphism (https://www.ncbi.nlm.nih.gov/projects/SNP/) **(Table 5.2)**. Therefore, the clinical significance of the variant was uncertain. It is predicted to be damaging/pathogenic by various *in silico* tools **(Table 5.2)**. I mapped the variant onto the crystal structure of beta-cell GCK and found that the residue lies near to the adenosine triphosphate (ATP) binding site **(Figure 5.5)**.

Table 5.2. GCK p.F133L bioinformatic predictions of pathogenicity (Alamut Visual v.2.10.0)

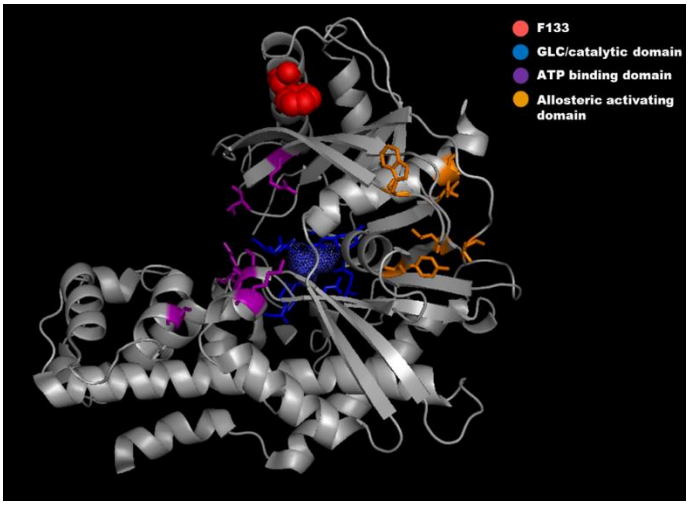
Variant	cDNA Level: **NM_000162.3:c.397T>C** gDNA Level: **Chr7(GRCh37):g.44190641A>G** Protein Level: **p.(Phe133Leu)**
Pathogenicity clues	**Nucleotide conservation:** highly conserved (phyloP: 4.81 [-14.1;6.4]) **Amino acid conservation:** highly conserved, up to Zebrafish (considering 29 species) **Physiochemical difference between Phe and Leu:** Grantham distance of 22 **Protein domain:** Hexokinase, N-terminal **Align GVGD (v2007):** C0 (GV 21.6; GD 21.82) **SIFT (v6.2.0):** Deleterious (score: 0.01, median 4.32) **PolyPhen-2:** possibly damaging (0.675) **MutationTaster (v2013):** disease causing (p-value 1) Not present in ClinVar, dbSNP, ESP, gnomAD

(ii) Functional characterisation

Recombinant GST-tagged GCK protein harbouring the p.F133L alteration showed both markedly decreased protein stability and affinity for ATP **(Table 5.3)**. Affinity for ATP was evaluated under a 4-fold increase in ATP substrate concentration (0-20 mmol/L compared to standard 0-5 mmol/L range) due to failure to saturate at normal conditions (~0.4 mmol/L). Considering the proximity of the residue to the ATP binding site, impaired ATP turnover (K_{cat} b 22.09 ± 0.20) is not unexpected **(Figure 5.5; Table 5.3)**. The enzyme also exhibits cooperativity with glucose and a decrease in the maximal specific activity (K_{cat} a) determined from glucose $S_{0.5}$ measurements. Despite generally compromised activity, the relative activity index (RAI) of GCK-p.F133L is almost 3-fold greater than that of WT. This discrepancy is largely driven by a slight decrease in half-saturation glucose concentration (i.e. greater affinity for glucose) for GCK-p.F133L compared to WT GCK **(Table 5.3).** Indeed, RAI values are biased towards glucose affinity and glucose-dependant activity and are less influenced by impact on ATP turnover (K_{cat} b not included in model).

To determine the stability profile of the mutant protein, GCK-p.F133L activity was assessed along with WT and a neutral variant p.T342P (Steele et al 2011; Raimondo et al, 2014) over a temperature gradient 40–63°C. Of note, GCK-p.T342P was previously reported as a pathogenic LoF variant, a classification that was disproved by Steele and colleagues, who showed that the variant demonstrated normal catalytic activity and did not co-segregate with a GCK-MODY phenotype in the two families in which it was identified, more in line with the benign *in silico* predictions for p.T342P. Activity data generated by Raimondo and colleagues (2014) on a highly unstable GCK-PNDM variant p.H50D [RSI =

0.00; TA50 = 49.30°C] were overlaid as a negative control reference. Across the tested temperature gradient, GCK-p.F133L exhibited markedly decreased protein stability (RSI of 0.53) compared to both WT and p.T324P **(Figure 5.6)**.



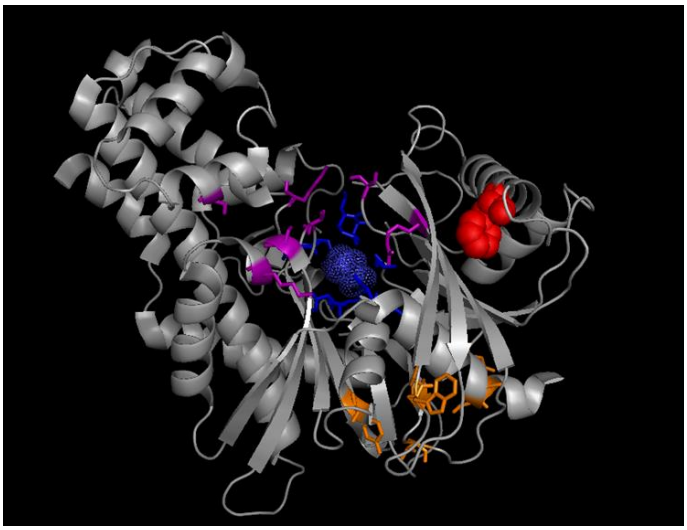


Figure 5.5. Ribbon model of human GCK in the active closed confirmation (glucose-bound) illustrating the position of lysine residue 133 (F133) (red spheres). Bound glucose (GLC) is shown in dot form in the centre of the active site which is highlighted as blue sticks. The ATP-binding and GKA-binding allosteric active site domains are indicated in purple and orange sticks, respectively. Protein DataBank entry 1V4S (Kamata et al, 2004). Generated with PyMOL v.0.99.

Table 5.3. Kinetic characterization of wild type (WT) and p.F133L GST-GCK protein

Protein	Yield (mg/ml)	$S_{0.5}$ (mmol/L)	Hill (nH)	ATP_{Km} (mmol/L)	K_{cat} a (s-1)	K_{cat} b (s-1)	RAI	RSI
WT	1.29	7.74 ± 0.11	1.56 ± 0.02	0.74 ± 0.07	55.87 ± 0.36	46.26 ± 0.32	1.00	1.00
F133L	1.63	5.33 ± 0.08	1.22 ± 0.01	1.96 ± 0.07	45.09 ± 0.34	22.09 ± 0.30	2.63	0.53

Glucose affinity (S0.5), Hill number (nH), and turnover number (Kcat) values were determined in the presence of 0-100mmol/L glucose. Affinity for ATP (ATPKm) was determined in the presence of 0-20mmol/L ATP. Glucose $S_{0.5}$ and ATP_{km} assays were used to determine K_{cat} a and k_{cat} b values, respectively. Relative activity indices (RAI) were calculated as previously described (Christesen et al, 2002). Relative stability indices (RSI) were derived from (TA50(mutant) - TA50(min))/(TA50(WT) 2 TA50(min)), where TA50 refers to the temperature point at which each protein displayed 50% of its activity at 40°C in a thermostability assessment. The TA50(min) value, which refers to the minimum observed TA50, that was used to calculate WT and p.F133L RSI was that of an unstable clinical GCK-PNDM variant, H50D (49.18°C). Data shown are mean ±SEM (Standard error of the mean of n=2 biological replicates, each comprised of n=3 technical replicates per glucose or ATP concentration).

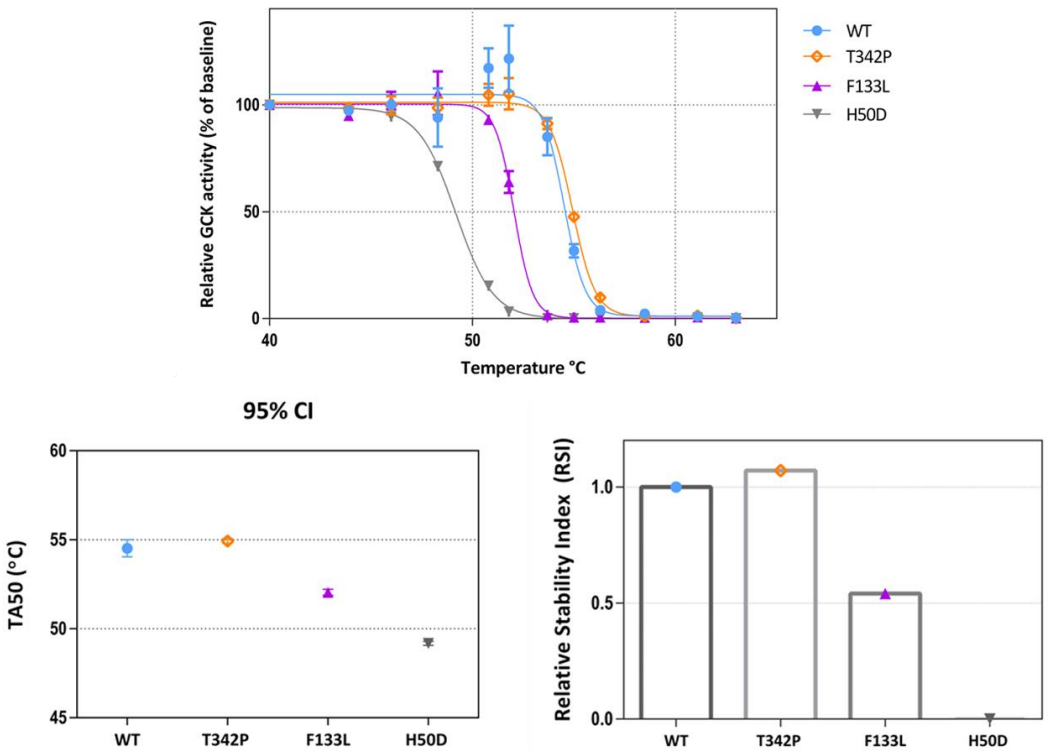


Figure 5.6. Assessment of thermostability for WT and mutant p.F133L human GST-GCK proteins. Logistic regression modelling was used to fit an activity curve to each of the assessed proteins (top). Data were normalized to the baseline level of activity for each protein at 40°C. GCK allele p.T342P was used as a WT-like stable control variant; GCK protein p.H50D was used as a clinical GCK-PNDM unstable control variant. The logistic regression model was used to calculate the temperature point at which each protein displayed 50% of its activity at 40°C (TA50) (bottom left). TA50 values were used to calculate stability indices relative to WT (RSI) using the equation previously described (Raimondo et al., 2014) (bottom right). Data shown are mean ±SEM (Standard error of the mean of n=2 biological replicates, each comprised of n=4 technical replicates per temperature point) for WT, T342P and F133L.

5.3.2 *GCK* p.E157K (c.469G>A)

(i) Bioinformatic analysis and structural modelling

This alteration is not present in general databases nor is it listed as a polymorphism (https://www.ncbi.nlm.nih.gov/projects/SNP/) **(Table 5.4)**. Mapping of the variant onto the crystal structure of beta-cell GCK shows that the residue lies within extremely close proximity of the allosteric activating site **(Figure 5.7)**. It was identified once in a cohort of Italian children with clinical characteristics of MODY (Massa et al, 2001). It is predicted to be damaging/pathogenic by multiple *in silico* prediction algorithms **(Table 5.4)**. Clinical significance of the mutation was uncertain though it was presumed to be hypomorphic due to severe diabetes phenotype diagnosed outside of the neonatal period.

Table 5.4. GCK p.E157K bioinformatic predictions of pathogenicity (Alamut Visual v.2.10.0)

Variant	cDNA Level: **NM_000162.3:c.469G>A** gDNA Level: **Chr7(GRCh37):g.44190569C>T** Protein Level: **p.(Glu157Lys)**
Pathogenicity clues	**Nucleotide conservation:** moderately conserved (phyloP: 4.24 [-14.1;6.4]) **Amino acid conservation:** highly conserved, up to Zebrafish (considering 29 species) **Physiochemical difference between Glu and Lys:** Grantham distance of 56 **Protein domain:** Hexokinase, N-terminal **Align GVGD (v2007):** C55 (GV: 0.00 - GD: 56.87) **SIFT (v6.2.0):** Deleterious (score: 0, median: 4.32) **PolyPhen-2:** probably damaging (1.000) **MutationTaster (v2013):** disease causing (p-value: 1) Not present in ClinVar, dbSNP, ESP, gnomAD

(ii) Functional characterisation

GCK-p.E157K exhibited severe catalytic loss-of-function as reflected by the low relative activity index of 0.15 compared to WT, and by reduced substrate cooperativity, turnover and affinity for glucose (normally assessed at 0–100 mmo/L, but in this case increased to 0-200 mmol/L due to poor affinity) **(Table 5.5).** Since p.E157K resides near the GCK allosteric activating domain, I tested its response to a range of glucokinase activator (GKA) doses (0, 0.3, 1, 3, 9, 27 µM) **(Table 5.6 and Figure 5.8)**. As illustrated by the fold changes in glucose affinity as a function of GKA concentrations, the variant harbouring protein responded well to treatment. The average half-saturating glucose concentration for p.E157K and WT, between 1 µM and 27 µM GKA, was 0.98 mmol/L and 1.12 mmol/L, respectively **(Table 5.6 and Figure 5.8).**

To evaluate the impact of the variant on protein stability, catalytic activity of GCK-p.E157K was assessed along with WT and the WT-like stable control, p.T342P **(Figure 5.9).** Activity data from the unstable variant, p.H50D, evaluated as part of the GCK-PNDM mutation catalogue, was plotted as a negative control reference (Raimondo et al, 2014). Compared to WT and the (in)stability control variants, GCK-p.E157K showed only mildly reduced protein stability across the tested temperature gradient (RSI of 0.84). At the lower end of the temperature gradient, GCK-p.E157K demonstrated greater stability than WT, particularly at 48.3°C and 50.8°C **(Figure 5.9).** The stability profile of p.E157K closely resembles that of previously assessed homozygous GCK alleles identified also in cases diagnosed with diabetes outside of infancy (RSI of 0.98 and 0.81 for p.D160N and p.V226M, respectively; activity curves generated by Dr. Raimondo and overlaid in Figure 5.9) (Raimondo et al, 2014) **(Figure 5.9)**.

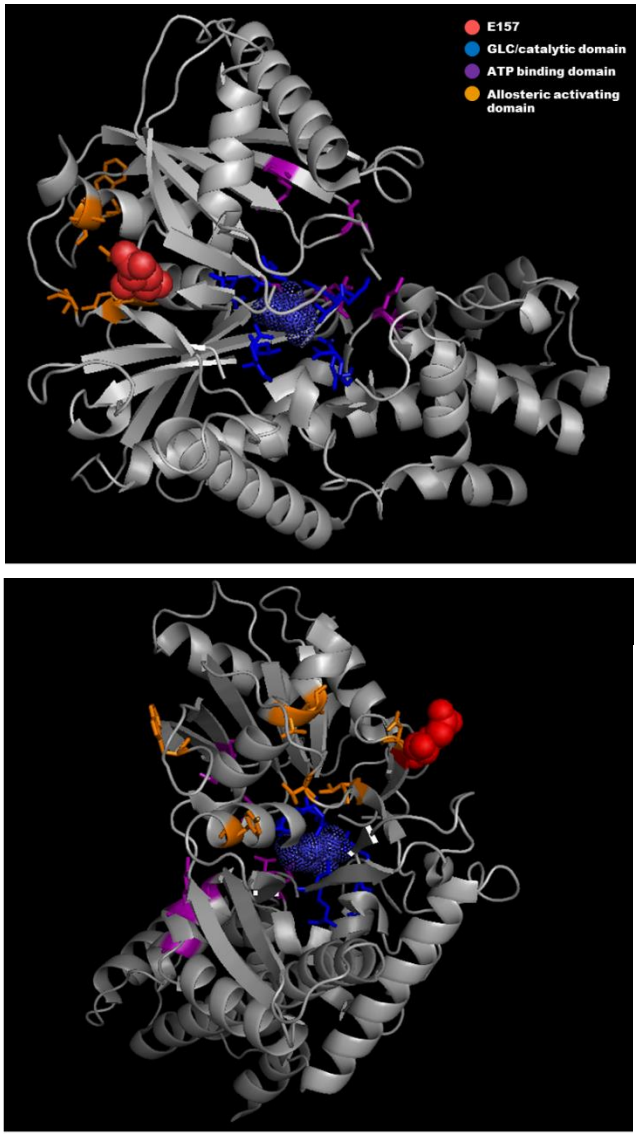


Figure 5.7. Ribbon model of human GCK in the active closed confirmation (glucose-bound) illustrating the position of lysine residue 157 (E157) (red spheres). Bound glucose (GLC) is shown in dot form in the centre of the active site which is highlighted as blue sticks. The ATP-binding and GKA-binding allosteric active site domains are indicated in purple and orange sticks, respectively. Protein DataBank entry 1V4S (Kamata et al, 2004). Generated with PyMOL v.0.99.

Table 5.5. Kinetic characterization of wild type (WT) and p.E157K GST-GCK proteins

Protein	Yield (mg/ml)	$S_{0.5}$ (mmol/L)	Hill (nH)	ATP_{Km} (mmol/L)	K_{cat} a (s-1)	K_{cat} b (s-1)	RAI	RSI
WT	1.29	8.57 ± 0.21	1.69 ± 0.06	0.45 ± 0.02	53.69 ± 0.72	47.90 ± 0.95	1.00	1.00
E157K	0.05	20.19 ± 0.36	1.42 ± 0.27	0.34 ± 0.02	34.37 ± 0.54	23.62 ± 0.60	0.15	0.84

Glucose affinity (S0.5), Hill number (nH), and turnover number (Kcat) values were determined in the presence of 0-200mmol/L glucose. Affinity for ATP (ATPKm) was determined in the presence of 0-5mmol/L ATP. Glucose $S_{0.5}$ and ATP_{km} assays were used to determine K_{cat} a and k_{cat} b values, respectively. Relative activity indices (RAI) were calculated as previously described (Christesen et al, 2002). Relative stability indices (RSI) were derived from (TA50(mutant) - TA50(min))/(TA50(WT) - TA50(min)), where TA50 refers to the temperature point at which each protein displayed 50% of its activity at 40°C in a thermostability assessment. The TA50(min) value, which refers to the minimum observed TA50, that was used to calculate WT and p.E157K RSI was that of an unstable clinical GCK-PNDM variant, H50D (49.18°C). Data shown are mean ±SEM (Standard error of the mean of n=3 biological replicates, each comprised of n=3 technical replicates per glucose or ATP concentration).

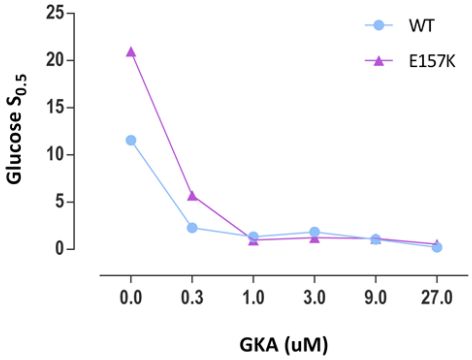


Figure 5.8. Wild type (WT) and p.E157K GST-GCK affinity to glucose in response to a glucokinase activator (GKA) R00281675 (Roche) concentration gradient. Glucose affinity ($S_{0.5}$) values were determined in the presence of 0-200mmol/L glucose. Enzymes at 0 uM GKA conditions were treated with DMSO (vehicle).

Table 5.6. Effect of glucokinase activator (GKA) R00281675 on wild type (WT) and p.E157K GST-GCK

Protein	Glucose $S_{0.5}$ (mmol/L)	$S_{0.5}$ vehicle/$S_{0.5}$ GKA-treated
WT + vehicle	11.56	1.00
WT + GKA 0.3 uM	2.29	5.05
WT + GKA 1 uM	1.32	8.76
WT + GKA 3 uM	1.85	6.25
WT + GKA 9 uM	1.06	10.91
WT + GKA 27 uM	0.23	50.26
E157K + vehicle	20.99	1.00
E157K + GKA 0.3 uM	5.74	3.69
E157K + GKA 1 uM	0.98	21.44
E157K + GKA 3 uM	1.24	16.94
E157K + GKA 9 uM	1.15	18.18
E157K + GKA 27 uM	0.55	38.05

Glucose affinity ($S_{0.5}$) values were determined in the presence of 0-200mmol/L glucose. $S_{0.5}$ vehicle/$S_{0.5}$ GKA-treated values represent fold change in enzymatic affinity to glucose.

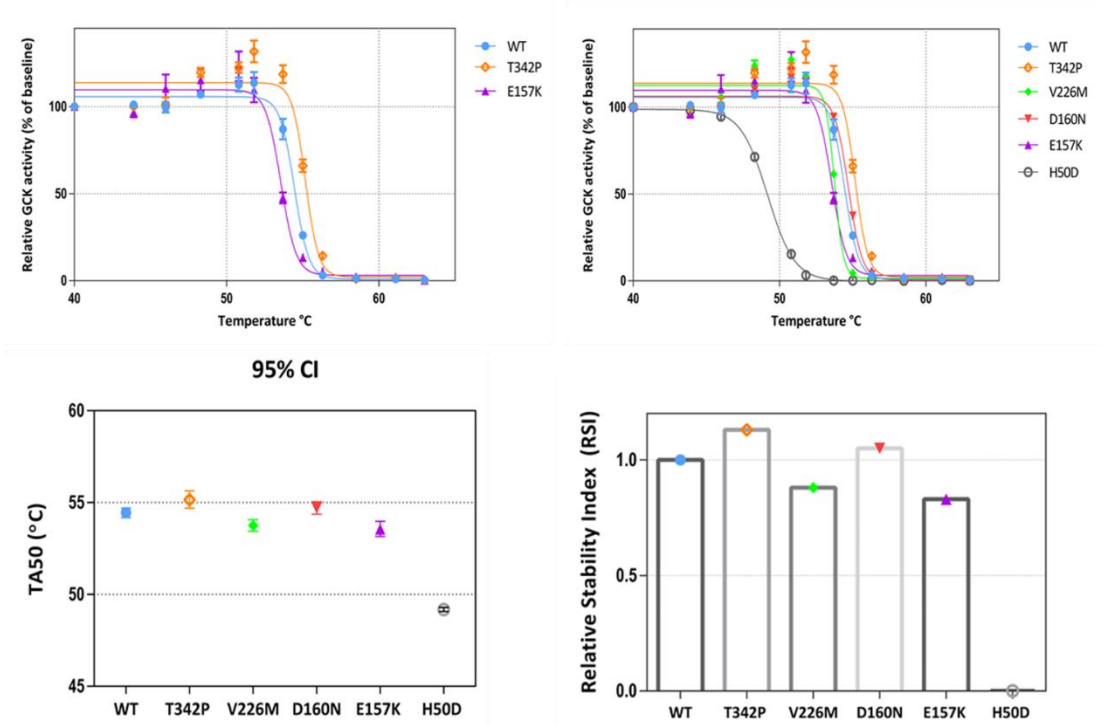


Figure 5.9. Thermostability assessment of wild type (WT) and mutant p.E157K human GST-GCK proteins. Logistic regression modelling was used to fit an activity curve to each of the assessed proteins (top). Data were normalized to the baseline level of activity for each protein at 40°C. GCK -p.T342P was used as a WT-like stable control variant; GCK- p.H50D was used as a clinical GCK-PNDM unstable control. The logistic regression model was used to calculate the temperature point at which each protein displayed 50% of its activity at 40°C (TA50) (bottom left). TA50 values were used to calculate stability indices relative to WT (RSI) using the equation previously described (Raimondo et al., 2014) (bottom right). Data shown are mean ±SEM (Standard error of the mean of n=3 biological replicates, each comprised of n=4 technical replicates per temperature point).

5.4 Discussion

The functional and clinical impact of the *GCK* variants, p.F133L and p.E157K, was difficult to interpret based only on family history, clinical phenotype, and computational variant-effect prediction. Per the American College of Medical Genetics (ACMG) standards and guidelines, p.F133L and p.E157K would be classified as variants of uncertain significance (**Class 3 – VUS**) (Richards et al, 2015).

While the variants do not meet any criteria for **supporting**, **strong**, or **stand-alone** evidence of benign function, they meet a few for evidence of pathogenicity though these are insufficient for a confident pathogenic/likely pathogenic classification based on ACMG guidelines. For instance, both alleles are absent in population databases, a criterion that provides **moderate** evidence of pathogenicity (rule PM2 under moderate evidence for pathogenicity) based on ACMG guidelines. Also, for both alleles, multiple lines of *in silico* evidence support loss-of-function (rule PP3 under supporting evidence for pathogenicity) (**Table 5.2 and Table 5.3**) and the patient's phenotype or family history is specific for single genetic aetiology (rule PP4, under supporting evidence for pathogenicity) (**Figure 5.2 and 5.3**). These two lines of evidence are considered **supporting** for a pathogenic classification.

According to ACMG guidelines, robust *in vitro* or *in vivo* functional characterisation studies which support a damaging effect qualify as a **strong** line of evidence for pathogenicity (rule PS3) (Richards et al, 2015). The weight of this criterion combined with 1-2 **moderate**, or ≥ 2 **supporting** lines of evidence, would be sufficient to classify the GCK alleles as likely pathogenic. In this case,

functional studies can potentially provide molecular diagnostic clarity for allele carriers and their families. Although, it is worth mentioning that, even with functional data supportive of deleterious impact, without meeting any criteria under the **very strong** evidence category, more observations weighted as strong, moderate and/or supporting are required for the variants to be classified as pathogenic.

5.4.1 GCK-p.F133L Interpretation and reporting

Based on multiple arms of functional testing, I have shown that GCK-p.F133L is associated with decreased K_{cat} values, poor affinity for ATP, and a stability index half that of WT-GCK **(Table 5.3; Figure 5.6).** These functional characteristics are highly consistent with reported biochemical properties of GCK-MODY causing variants and suggest that the mutant protein is damaging (Osbak et al, 2009).

A couple of novel variants near the 133 residue in GCK (p.L146P, p.S151T) were detected in a GCK-PNDM cohort, and were functionally interrogated using the same experimental pipeline as the one described in **Figure 5.1** (Raimondo et al, 2014). Interestingly, data on the kinetic and stability effects of these variants were unobtainable due to extremely poor affinity for glucose and/or ATP in functional assays. Another *GCK* allele closer to the variant, p.D132N, was identified in a T2D case-control exome sequencing study (T2D Genes/GoT2D) (Fuchsberger et al, 2016). Functional evaluation p.D132N showed no discernible impact on catalytic activity or stability: the variant appeared to behave like WT GCK (unpublished data from Gloyn group generated by Dr. Anne Raimondo and Mrs. Amy Barrett).

The phenotype of the proband (**Figure 5.2** III.1) also fits with the clinical picture of GCK-MODY: mild fasting hyperglycaemia (detected since aged 6 and likely present from birth) and marginal elevation of HbA1c. Besides high BMI, the proband does not appear to be suffering from confounding metabolic complications that would point towards T2D, unlike his father. *HNF1A*- and *HNF4A*- MODY can also be eliminated as a) the clinical features of the proband do not match those of transcription factor MODY and b) these genes were sequenced in the proband, and no variants were reported. Given also the lack of indication of T1D (normal C-peptide and absence of GAD and pancreatic autoantibodies), the proband most likely has GCK-MODY. As such, he could potentially go on without pharmacological intervention like the majority of GCK-MODY cases.

Taken together: (i) consensus across multiple *in silico* prediction tools on deleteriousness, (ii) clinical phenotype consistent with a single gene aetiology, (iii) absence of the variant from population databases, (iv) damaging effect on protein function in *in vitro* functional assays, the variant p.F133L in GCK is **likely pathogenic** (Class 4) based on ACMG classification guidelines (Richards et al, 2015).

Unfortunately, functional annotation does not provide the same level of diagnostic clarity for the proband's father (**Figure 5.2** II.1). He was diagnosed with diabetes at age 15, is overweight and has had challenges controlling blood sugar levels. After years of insulin pump therapy, he has HbA1c levels well over the range associated with GCK-MODY and lower than normal C-peptide levels. The phenotype is too severe to be attributed solely to the GCK-p.F133L variant.

Testing negative for autoantibodies does not immediately eliminate a combination of T1D (potentially inherited from his father (I.I)) and GCK-MODY. In a cohort of individuals with a clinical diagnosis of T1D, Hameed and colleagues found that 10% presented as autoantibody-negative at diagnosis, and 5.2% remained persistently autoantibody-negative (PAN) on repeat testing (Hameed et al, 2011). In terms of insulin production and secretion, multiple studies have reported persistent low-level insulin secretion (C-peptide < 30 pmol/L) in a majority of individuals with long-duration T1D, which is enhanced after meal stimulus (Oram et al 2014; Wang et al, 2012). Oram and colleagues have dubbed these individuals insulin microsecretors. While C-peptide levels in the proband's father are much higher than the T1D microsecretor range, they certainly fall below the >200 pmol/L cut-off used to define significant endogenous insulin secretion (Oram et al, 2014; DCCTRG 1998).

Although, considering lack of autoimmunity and both elevated BMI and levels of endogenous insulin secretion compared to T1D microsecretors, it is more likely that the father has a combination of early-onset familial T2D and GCK-MODY (Doria et al, 2008; Raghavan et al, 2015). The proband's grandfather, who was diagnosed with T1D without autoantibody testing, could have suffered from the same combination of diseases. It is unclear whether the proband's father was tested for variation in other MODY genes, which may be influencing his diabetic phenotype in addition to the damaging GCK-p.F133L allele.

Determining the father's T1D genetic risk score would perhaps assist in making a definitive diagnosis (Oram et al, 2015; Patel et al, 2016). Oram and colleagues have developed a T1D genetic risk score (GRS) based on 30 common risk variants robustly associated with T1D, from HLA and non-HLA loci, identified in

genome-wide association studies (Oram et al, 2015). Scores are calculated on an individual basis by genotyping the variants and combining the risk of each inherited allele (Oram et al, 2015; Patel et al, 2016). T1D-GRS were generated for n=1,963 T1D subjects from the Wellcome Trust Case Control Consortium (WTCCC) and n=805 MODY patients of white European ancestry (415 patients with HNF1A-MODY, 346 with GCK-MODY, 42 with HNF4A-MODY, and 2 with HNF1B MODY), and were shown to be highly discriminatory (ROC area under the curve 0.87, 95% CI 0.86–0.89) (Patel et al, 2016). A T1D-GRS >0.280 (>50th T1D centile) indicates T1D with 94% specificity and 50% sensitivity (Oram et al, 2015; Patel et al, 2016).

The initial study which reported development of the risk scores showed that T1D-GRS can also highly discriminate clinically defined T1D and T2D from WTCCC (AUC 0.88 [95% CI 0.87–0.89]; P < 0.0001). A stable DNA-based test such as the T1D-GRS can powerfully complement existing diagnostic criteria such as clinical features and autoimmunity markers, and is especially useful considering ascending obesity rates among young adults.

5.4.2 GCK-p.E157K Interpretation and reporting

My biochemical assessments of GCK-p.E157K showed that is a kinetically defective protein associated with decreased K_{cat} values, severally poor affinity for glucose, and a mildly affected stability index (**Table 5.5 and Figure 5.9**). These functional characteristics are highly consistent with other hypomorphic homozygous *GCK* mutations that cause severe diabetes diagnosed in childhood rather than infancy (Raimondo et al, 2014).

Upon evaluating the impact of 19 unique homozygous *GCK* alleles (causing neonatal or childhood-onset diabetes) on catalytic function and protein stability, Raimondo and colleagues found that protein stability is far better correlated with clinical severity compared to kinetic loss-of-function (Raimondo et al, 2014). This is further supported by the fact that the two variants in the study identified in individuals with childhood-onset diabetes (p.D160N, c.478G>A and p.V226M, c.676G>A) were among the most stable of the studied suite of variants despite residual kinetic activity. That is, high/WT-like relative stability might explain later manifestation of diabetes despite homozygosity typically associated with a neonatal diagnosis (GCK-PNDM).

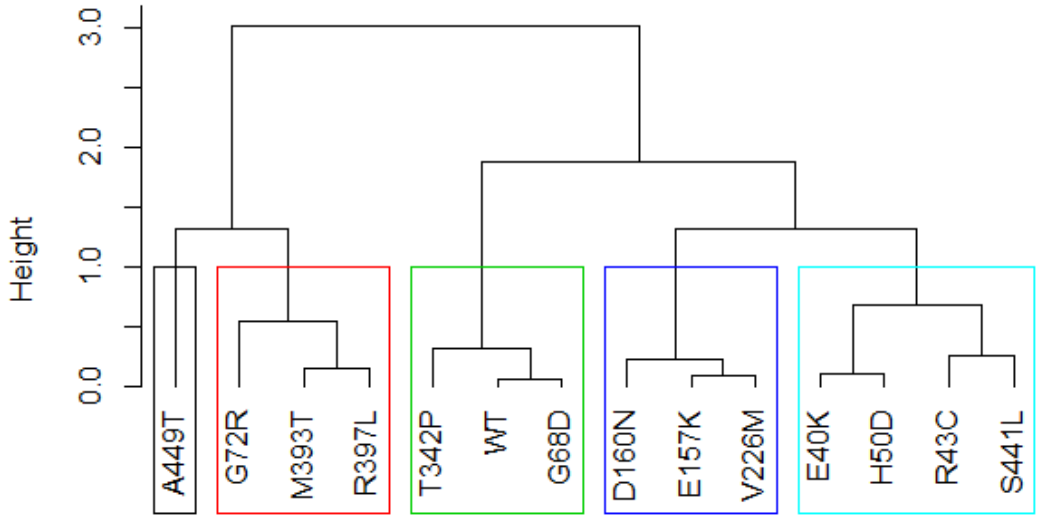


Figure 5.10. Hierarchical clustering of functionally interrogated homozygous missense *GCK* alleles and WT GCK based on RAI and RSI values. Subgroups defined by height cut-off value of 1.0.

Both variants were detected in white Canadians, one of whom was diagnosed at age 9yo (p.D160N/p.D160N) and the other at age 15yo (p.V226M/p.V226M). Assessments of p.V226M activity were conducted under exceptionally elevated glucose and ATP concentrations: 0-500 mmol/L in glucose $S_{0.5}$ assays, 0-25 mmol/L ATP in ATP_{km} assays, and a final glucose concentration of 45 mmol/L in thermostability assays (Raimondo et al, 2014). Indeed, not only do the childhood-onset alleles resemble p.E157K stability profile, but their activity trends are also closely matched. In fact, hierarchical clustering of p.E157K with 12 of childhood/neonatal *GCK* alleles based on RAI and RSI (for 7/19 variants excluded from the analysis, either RAI or RSI unbootable) shows that p.D160N, p.V226M, and p.E157K form an independent subgroup **(Figure 5.10)**. Within the cluster, p.E157K and p.V226M are more closely linked **(Figure 5.10).**

The phenotype of the proband (**Figure 5.4** III.1) also fits with the clinical picture of severe diabetes diagnosed outside neonatal period, associated with hypomorphic homozygous *GCK* alleles. She was not responsive to treatment with oral hypoglycaemic agents (OHAs). To date, enhanced insulin production with sulphonylurea in GCK-PNDM has only been reported once (Turkkahraman et al, 2008). Individuals with biallelic loss-of-function in *GCK* require lifelong insulin therapy. Despite history of presumed T2D in her family, the proband is slim (BMI 22) and was diagnosed with diabetes at the age of 5y, so it is unlikely that her altered glycaemia is caused by T2D. She has tested negative anti-islet autoantibodies and has presented with elevated levels of C-peptide (up to ~500 pmol/L) which is sufficient to eliminate a diagnosis of T1D.

Altogether: (i) consensus across multiple *in silico* prediction tools on deleteriousness of the allele, (ii) clinical phenotype consistent with a single gene

aetiology, (iii) absence of the variant from population databases, (iv) damaging effect on protein function in *in vitro* molecular assays, the variant p.E157K in *GCK* is **likely pathogenic** (Class 4) based on ACMG classification guidelines (Richards et al, 2015). In its heterozygous form, it also likely to cause GCK-MODY.

5.4.3 Summary and conclusions

The main objective behind the efforts described in this chapter was to assist in the medical diagnostic interpretation of two novel missense variants in *GCK*, identified in clinical genetic testing, by characterising them functionally.

Multiple bioinformatic tools which predict variant pathogenicity on the basis of protein sequence conservation, structure, function, amino acid biophysical properties, and annotations in public databases, agreed that the substitutions in question (p.F133L, p.E157K) are likely to be damaging and disease-causing **(Tables 5.2 and 5.3).** Modelling of the variants onto GCK structure (Protein DataBank entry 1V4S [Kamata et al, 2004]) helped further predict the likely impact of the variants on protein activity. F133 sits in close proximity of the enzyme's ATP binding domain and was thus expected to impair ATP-associated kinetic metrics **(Figure 5.5)**. E157 resides near the enzyme's allosteric activating site to which pharmacological GKAs bind (as well endogenous activators) and where HH GoF mutations cluster **(Figure 5.7)**. It was thus predicted that GCK-p.E157K would respond well to GKA treatment. The clinical phenotypes of the probands in which *GCK* genetic testing was performed appeared consistent with their *GCK* mutation status:

- p.F133L/N appeared to be responsible for mild fasting hyperglycaemia in the proband (i.e. GCK-MODY) (see **subsection 5.1.3(i)**)
- p.E157/p.E157K appeared to be consistent with severe childhood-onset diabetes due to a hypomorphic mutation (see **subsection 5.1.3(ii)**)

All of this information on each of the identified *GCK* alleles was still insufficient to make a confident assignment of pathogenicity according to ACMG guidelines. I thus conducted functional assays interrogating multiple mutational mechanisms to determine variant pathogenicity. I evaluated the impact of each allele on:

- affinity, cooperativity and turnover of *GCK* substrates glucose and ATP
- protein stability in response to a temperature titration at basal glucose concentrations

For p.E157K, I tested response to GKA treatment given its structural proximity to the allosteric activator site.

The heterozygous allele p.F133L, which seemed to be causing GCK-MODY in the proband, was impaired in kinetic assays as reflected by an expected poor affinity for ATP and reduced substrate turnover measurements, and also demonstrated reduced protein stability compared to WT and a stable control allele (**Table 5.3 and Figure 5.6)**. The homozygous allele p.E157K, which seemed to be behind severe insulin-dependent diabetes diagnosed outside the neonatal period in the proband, was a) responsive to GKA treatment as demonstrated by enhanced glucose affinity (**Table 5.6 and Figure 5.8),** b) catalytically dysfunctional with extremely poor affinity for glucose and an activity index of 0.15 (**Table 5.5**). Although, despite residual kinetic activity, the GCK-p.E157K demonstrated a relatively WT-like stability profile (**Figure 5.9**), a

functional trend which also mimicked that of other ‘super-stable’ homozygous childhood-onset *GCK* alleles (Raimondo et al, 2014).

In conclusion, functional characterisation data of GCK-p.F133L and GCK-p.E157K, generated using the experimental pipeline depicted in **Figure 5.1,** substantiate the computationally-predicted deleterious effects of these variants and suggest a **likely pathogenic** classification based on ACMG guidelines.

6

General Discussion

The rapid reduction in the cost of high-throughput genomic technologies has been associated with increased application and data generation, in both the research and clinical arenas, uncovering a host of new analytical and ethical questions. Already, genetic sequencing data have refuted the longstanding notion that Mendelian disease variants pose less of an interpretive challenge than variants in polygenic diseases.

The genetics of MODY perfectly exemplify the extent to which genotype-phenotype mapping in the monogenic disease context is not as straightforward as previously assumed. Variant discovery efforts have shifted from being phenotype-driven to being led by large-scale genomic sequencing in multiethnic populations and case-control cohorts. In the context of diabetes, this has exposed a) overlap between T2D-associated genomic loci and monogenic diabetes genes/alleles, b) alleles of higher penetrance for T2D and lower penetrance for MODY in monogenic diabetes genes and c) rare alleles previously classified as disease-causing in euglycemic individuals (Flannick et al, 2013; Flannick et al, 2016).

Refined estimates of variant pathogenicity demand an integrative analysis of sequence data by accounting for demographic variables, molecular function in relevant experimental systems, evolutionary conservation, clinical and biochemical phenotypes, familial segregation, and allelic penetrance. Developing such a contextual understanding of variant behavior, rapidly and at scale, will better position scientists and clinicians to evaluate the implication of novel and incidentally detected variants, especially in known disease-associated genes.

About 70% of all MODY cases are attributable to variation in *HNF1A* and *GCK* (Ellard et al, 2008). Missense mutations comprise 65% of genetic alterations in GCK and about 55% in *HNF1A* (Chakera et al 2015; Colclough et al, 2013). Variants of uncertain clinical significance in *GCK* and *HNF1A* are often detected and the majority of identified sequence variants in these genes are very rare and novel (Colclough et al, 2014). Only ~12% of all reported *GCK*, *HNF1A* and *HNF4A* variants have been functionally annotated (Osbak et al, 2009; Colclough et al, 2013). This means that for the majority of known sequence variants, impact on molecular function is unknown, and they are thus medically uninformative.

For variants in *GCK*, protein stability correlates well with clinical severity, and *in vitro* catalytic activity trends which deviate substantially from wild-type *GCK* function have been reported for inactivating heterozygous and homozygous mutations (Raimondo et al, 2014; Osbak et al, 2009). For variants in *HNF1A*, understanding the extent of molecular dysfunction means understanding where a variant sits along the pathophysiological continuum defined by familial early-onset hyperglycemia (SU-responsive) driven by large effects of causal variants on the one end and T2D predisposition on the other (lifestyle interventions) driven by modest effects of risk-modifying alleles.

6.1 Summary of findings

For the clinically actionable MODY genes *HNF1A* and *GCK*, I have developed and employed a range of functional genomic and computational approaches to resolve the complexity of genotype-phenotype relationships and assist in medical diagnostic interpretation.

A large collaborative multi-dimensional functional dataset had been generated by research groups at the Universities of Oxford (Prof Anna Gloyn and Prof Mark McCarthy) and Bergen (Prof Pal Njølstad) as follow-up for ~80 *HNF1A* missense alleles identified in the exomes of ~13K T2D cases and controls. My aim was to develop an analytical pipeline that would leverage all of the tested functional parameters to stratify alleles along the *HNF1A* phenotypic continuum without imposing arbitrary molecular cut-offs or prescribed labels onto the dataset.

To this end, I employed a combination of unsupervised statistical models (principal component analysis and hierarchical clustering) to obtain high resolution clusters using the dimensions of functional data (transactivation, protein abundance, nuclear translocation) that were available for all alleles. To refine the spatial organisation of clusters in multivariate space, I assigned weights for spatial variation of tissue-specific isoform expression (liver vs pancreatic beta cell isoform) and functional domain (dimerization, DNA binding, transactivation and undefined). The structure of dimension reduced multi-parametric functional data was primarily shaped by transcriptional activity, which suggests that it might be the most informative mechanism for assessing *HNF1A* molecular dysfunction. Results from this analysis facilitated a clinically impactful variant-phenotype mapping effort which led to the reclassification of n=7 *HNF1A* missense alleles (p.D526N, p.A174V, p.G606S, p.H469Y, p.P291S, p.N62S, p.A161T) from pathogenic/likely pathogenic to VUS/likely benign in the UK MODY diagnostic registry **(Chapter 3; Figure 6.1)**.

To enable comprehensive annotation of all possible missense alleles in *HNF1A*, I developed a systematic and generalizable pipeline for an *HNF1A* Multiplexed

(functional) Assay of Variant Effects (MAVE) which relies on transactivation as a molecular signature.

My pipeline begins with *in vitro* model selection and characterisation, followed by dose-dependent stimulation of the transcription factor-encoding gene, generation of a catalogue of dose-responsive differentially expressed transcripts by RNA-seq, and finally prioritising a subset of cell surface expressed endogenous targets for validation of MAVE-compatibility by signal-to-noise and dynamic range flow cytometric assessments. Since the assay development pipeline is based upon identification of endogenous target transcripts to serve as high resolution molecular discriminators, it not only provides an array of technical markers but also mechanistic insights into the regulatory biology of the disease-relevant transcription factor under interrogation.

I used the human hepatoma line HepG2 as an *in vitro* model. Through whole transcriptome profiling of HepG2 samples transduced with dox-inducible *HNF1A*, I was able to identify 729 up- or down-regulated transcripts (FDR 1%) across tested dox doses after 48 hours of induction. Following multiple layers of validation including qPCR, western blotting, and flow cytometric analysis, the cell surface expressed gene *TM4SF4* was selected as an HNF-1a activity-specific MAVE-compatible marker from the differential expression gene list. With TM4SF4 as a readout in flow cytometric analyses, I achieved robust segregation of an *HNF1A* allelic series: 2-8% from MODY-causal loss-of-function alleles, 30% from a low frequency allele associated with T2D risk, >60-70% from WT and neutral alleles.

I have verified the quality and completeness of an *HNF1A* saturation mutagenesis library (n=12,615 substitutions) synthesised by Twist Bioscience (San Francisco, CA, USA). Based on the allelic series induced TM4SF4 fluorescent distribution trends, I have modelled library distribution, and predict it would be possible to generate 3 different functional bins into which allele-harbouring HepG2 cells can be sorted by FACS (WT+, intermediate p.E508K +/-, and null -) for function score assignment **(Chapter 4; Figure 6.2)**.

With respect to *GCK*, my objective was to assist in the clinical diagnostic interpretation of two missense alleles identified through genetic testing in two complex pedigrees. To enable this pursuit, I subjected the alleles to a rigorous and comprehensive *in vitro* functional characterization pipeline. One of the variants was detected in the homozygous state (p.E157K/p.E157K) in a proband who was diagnosed with type 1 diabetes at age 6 (later tested negative for negative for GAD and islet autoantibodies), is on insulin therapy, has an extended family history of diabetes, and Hb1Ac and glucose levels (8.6-9.9% and 9-18 mmol/l, respectively) suggesting a severe diabetes phenotype. The other variant was detected in the heterozygous state (p.F133L/N) in a proband and his father, both of whom were diagnosed with diabetes. Though the proband's phenotype appeared to be consistent with GCK-MODY present from birth (HbA1c 6.3-6.8% and FBG 7.2 mmol/l), the phenotype of the father is much more complex. The alleles were absent from public databases (gnomAD, COSMIC, dbSNP, 1000 genomes) and multiple *in silico* tools (SIFT, PolyPhen, Align GVGD, MutationTaster) predicted the alleles to be deleterious/disease-causing.

Through a series of assays which interrogated catalytic function (glucose and ATP affinity and turnover) and protein stability, I was able to provide a strong line

of evidence (i.e. functional impact) to support an assessment of variant pathogenicity (Richards et al, 2015). Based on family history, clinical features, bioinformatic prediction, and functional impact *in vitro*, I classified both alleles as **likely pathogenic**. In the heterozygous state, p.F133L and p.E157K are likely to cause GCK-MODY as both alleles render the protein kinetically defective. In the homozygous state, p.E157K appears to be behind the severe diabetes phenotype of the proband diagnosed outside of the neonatal period. In stability assessments, p.F133L was half as stable as WT and WT-like control alleles (RSI 0.5); whereas p.E157K, despite residual kinetic activity, only mildly impacted stability (RSI 0.89), which was consistent with the functional trends of other homozygous *GCK* alleles associated with childhood-onset insulin-dependent diabetes (Raimondo et al, 2014) **(Chapter 5; Figure 6.3)**.

6.2 Context and limitations

Revisiting clinical variant interpretations using functionally annotated *HNF1A* missense variant clusters led to the reclassification of ~5% of the all *HNF1A* missense alleles in the clinical database at Exeter (and ~23% of alleles which overlapped between the database and the functional characterisation effort). Misinterpretation or conflicting pathogenicity assessments across clinical laboratories is a problem that is being addressed by efforts such as unification of guidelines and standards for interpreting genetic variants by the American College of Medical Genetics and Genomics (ACMG) (Richards et al, 205; Rehm et al, 2015). An estimated 17% of variant submissions to the ClinVar database have conflicting and inconsistent interpretations by submitters. Further, ClinVar has identified 220 alleles in known disease-causing genes, presumed to be

pathogenic and annotated as such in the OMIM database, that are currently being reclassified and reinterpreted as benign/likely benign or VUS (Rehm et al, 2015).

Rapid, high-throughput, prospective variant annotation technologies (MAVEs), will help mitigate the variant interpretation crisis. Successful and productive implementation of these technologies requires overcoming the technical and analytical complexities associated with scaling up, which represent the most significant barrier in the face of closing the chasm between variant resolution and variant interpretation.

MAVEs are technically complex experimental systems that require diligent design and multiple layers of testing and validation to achieve maximal discrimination of a multiplexed oligonucleotide library on the basis of molecular dysfunction. In their current implementations, MAVEs are noisy, complicated (both application and interpretation of data), and limited in terms of a) biochemical phenotypes or molecular signatures per assay, b) context fidelity (of *in vitro* applications most amenable to multiplexing), and c) generalisability and reproducibility (Starita et al, 2017).

To date, MAVEs have only been applied in *in vitro* contexts (cell lines, yeast two-hybrid, phage display) and will likely be limited to such experimentally feasible set-ups until new technologies emerge to enable multicellular *in vivo* systems-wide interrogations (mice, zebrafish, drosophila). Advances in genome-editing tools and single-cell genomic assays can help design and implement generalisable MAVEs which capture endogenous genomic contexts. Although, until scale and efficiency of such systems becomes compatible with massively parallel applications, physiologically-relevant and easy-to-manipulate cell

systems represent the best alternative. A panel of different cell lines and assay architectures can be implemented to make up for compromised context, as is being done currently with the *HNF1A* MAVE collaboration. Our aim is to ultimately test two libraries synthesised on two different genetic backgrounds (I27 and L27 haplotypes) using a panel of 3 human liver cell lines (HepG2, Huh-7, Hep3B).

Several MAVEs interrogating protein-coding sequence libraries have been implemented at various scales. Majithia and colleagues developed a *PPARG*-specific MAVE using anti-CD36 THP1 macrophage system to discriminate Mendelian lipodystrophy alleles (FPLD3 [MIM #604367]) from neutral variants (Majithia et al, 2016). *PPARG* expression was enhanced for maximal discrimination through stimulating exogenous *PPARG* expression with agonists (rosiglitazone and PGJ2) and doxycycline, and depleting endogenic *PPARG* by CRISPR-Cas9 editing (Majithia et al, 2016). They were able to achieve maximal discrimination between CD36+ (induced by WT PPARG) and CD36- (deleterious PPARG), and a bimodal library distribution with peaks at both extremes (Majithia et al, 2016).

In their *BRCA1* MAVE, Starita and colleagues limited their interrogation to missense alleles in the RING domain of the protein where the majority (~60%) of pathogenic alleles occur (Starita et al, 2015). They compensated for the limitation in scale (non-saturation) by interrogating two disease-relevant biochemical phenotypes in yeast and phage display assays (Starita et al, 2015). Most recently, a MAVE to screen a *PTEN* saturation mutagenesis library (n=7,891 missense, nonsense, and single residue deletion) was implemented, and the data are currently undergoing analysis (Mighell et al, 2017). The assay is based on a humanized yeast model wherein yeast growth is contingent upon the catalytic

competence of *PTEN*. Accurate interpretation of *PTEN* alleles is important as somatic variants drive tumorigenesis and germline variants are associated with Cowden syndrome and increased risk of autism (Liaw et al, 1997; Butler et al, 2005; Hollander et al, 2011).

One limitation of protein-coding cDNA saturation libraries is that they fail to, for example, capture pathogenicity via splicing effects. A number of MAVEs have been designed and implemented to interrogate splice variants as well the effects of non-coding sequence on mRNA stability and protein expression (Rosenberg et al, 2015; Findlay et al, 2014; Kosuri et al, 2013; Shalem et al, 2015). Alleles in candidate regulatory elements have also been annotated in parallel through Massively Parallel Reporter Gene Assays (MPRAs) (Inoue and Ahituv, 2015; Tewhey et al, 2016; Ulirsch et al, 2016; Zhao et al, 2014). These assays can be used to guide multiplexed interrogations of non-coding alleles in *HNF1A* (intronic splicing and promoter and 5'UTR variants [~9% and ~2%, respectively]), many of which have been identified in single families and have been implicated in diabetes pathogenesis (Godart et al, 2000; Colclough et al, 2013; Frayling et al, 2001; Bellanne-Chantelot et al, 2008).

Variation in the distal regulatory elements of monogenic disease genes has been detected (Zhang et al, 2015; Weedon et al, 2014; Sakabe et al, 2012; VanderMeer et al, 2011). A prime example is recessively inherited distal cis-regulatory variation in the developmental enhancer of *PFT1A* (encodes pancreas-specific transcription factor 1a) identified through linkage analysis and whole genome sequencing in individuals with isolated pancreatic agenesis (Weedon et al, 2014). The ~400bp sequence located 25 Kb downstream *PFT1A* in which the alleles (n=6) were identified was characterised as a developmental

enhancer of the gene *in vitro* using a combination of luciferase assays and chromatin conformation capture (3C) in human embryonic stem cell (hESC)-derived pancreatic progenitor cells (Weedon et al, 2014). Data from such studies and efforts like the ENCODE (Encyclopedia of DNA Elements) project, which has been instrumental in the identification of tissue-specific functional elements throughout the non-coding genome, can help prioritise regulatory segments for interrogation by MPRAs (Consortium EP 2012; Starita et al, 2017).

6.3 Looking forward

In their early and initial phases of implementation, MAVEs will facilitate the creation of an atlas of functional consequences for disease-relevant loci. Eventually, with increased application and technical and financial tractability, an atlas of MAVE-derived effects for the entire human genome will become a reality, if not directly through experimental interrogations then via comprehensive high performance predictive models built using MAVE data. Progress in this arena requires public availability and accessibility of large MAVE datasets along with dissemination of well annotated step-by-step protocols to share expertise and enhance reproducibility.

While important, informative and powerful first lines of evidence, MAVE-derived function scores should not be treated as superior or stand-alone determinants of variant pathogenicity, nor should the relationship between functional impact and phenotypic consequence be mistaken for a linear one. The reality of phenotypic variant manifestation is often complex, context-dependant, non-linear and spectrum-based. Indeed, coding alleles exhibit variable penetrance depending on genomic (regulatory alleles in *cis* or *trans*, or haplotype epistasis) and

environmental (epigenomic) contexts both of which are difficult to capture in functional assessments (Lappalainen et al, 2011; Castel et al, 2017; Manning and Cooper, 2017; Couch et al, 2013).

The p.D526N allele in *HNF1A* is a perfect example which illustrates the need for nuanced evaluations of variant effects despite the availability of multiple layers of functional annotation from various cell systems. The allele was classified as pathogenic/likely pathogenic clinically (UK MODY diagnostic registry) and exhibited impaired *in vitro* functional activity (shared a cluster with other known MODY-causal alleles). Given its impact on molecular function, it would probably yield biomarker profiles (hsCRP and glycans) suggestive of HNF1A-MODY. Yet, re-evaluation of the clinical features of allele carriers, and the fact that it is present only in the longest *HNF1A* transcript isoform expressed predominantly in liver, suggest that it is more likely to be a contributing factor to common multifactorial diabetes rather than a primary driver of early-onset SU-responsive familial hyperglycaemia.

Developing a contextual and thorough understanding of variant behaviour from diverse functional, clinical, biochemical and demographic datasets is necessary to facilitate highly accurate interpretation. Large-scale reference datasets of human genomic sequence variation (e.g. gnomAD), which will continue to grow in size and in ethnic and demographic complexity, provide valuable insights into variant effects (Lek et al, 2016). Alongside improved coding annotations and isoform specification, ultimately, these datasets will enable individual characterisation of nucleotides by providing clues into selective constraint from, for example, more accurate allele frequency estimates (Lek et al, 2016; Samocha et al, 2014; Petrovski et al, 2013; Samocha et al, 2017).

Herein also lies the value of gold standard multi-tiered low-throughput functional assessments. They have enabled robust multi-level annotation of ~80 exome-detected missense alleles in *HNF1A* which included bioinformatic prediction, multi-dimensional *in vitro* function, allele frequency, clinical features, and ethnicity. The performance and predictive utility of any model built using *HNF1A* MAVE-derived function scores would be enhanced immeasurably upon calibration against these multi-layered data.

Well-calibrated MAVE-derived estimates of pathogenicity can be incorporated into existing multi-factorial probability models, and in the case of medically actionable disease genes such as *HNF1A*, they can be used to develop models that predict drug response. For example, an *HNF1A* MAVE-data derived lookup table can be used to create a multivariate clinical prediction tool for sulphonylurea response in individuals with *HNF1A* missense variants. The algorithm can be designed to generate a probability score for SU response based on input clinical features (e.g. age, duration of diabetes since SU-responsive allele carriers eventually progress to requiring insulin, hsCRP or glycan profile) and calibrated MAVE-based function score from an *HNF1A* lookup table.

Additionally, MAVE-derived variant classifications can be used to improve the performance of *in silico* prediction tools based on multi-data integration such as CADD (Combined Annotation-Dependent Depletion), MutationTaster, FitCons (fitness consequence) and VAAST (Variant Annotation, Analysis and Search Tool) (Eilbeck et al, 2017). They can also be incorporated into rigorous and collaborative multi-level annotation efforts led by the Clinical Genome Resource (ClinGen) program and evidence-based disease-specific variant classification databases such as InSiGHT (variant database for the International Society for

Gastrointestinal Hereditary Tumours) and CFTR2 (Clinical and Functional Translation of CFTR) (Rehm et al, 2015; Sosnay et al, 2013; Thompson et al, 2014).

6.4 Conclusion

The ever-growing amount of data from resequencing efforts of large case-control cohorts unselected for phenotypic extremes, multi-ethnic populations, isolated founder populations, and populations enriched for autozygosity by virtue of consanguinity continues to shift the paradigm and challenge discrete pre-genomic divides between healthy, common complex and Mendelian disease states. Such endeavours have not only identified presumed monogenic disease-causing alleles in healthy individuals, but have also led to the discovery of biallelic loss-of-function (LoF) events in the genome that confer disease-protective phenotypes (Flannick et al, 2013; Alkuraya 2015).

The challenge is no longer in accumulating sequence data, but in annotating and interpreting the effects of known and novel alleles, particularly those which are therapeutically and diagnostically consequential. Rapidly advancing functional genomic technologies (both experimental and analytic) are making it possible to map sequence-function relationships reliably, comprehensively, and at scale (Starita et al, 2017). These will assist in developing an atlas of variant effects for the human genome, which can be consulted by clinicians and researchers to make well-informed initial judgements of allele pathogenicity in the context of individual phenotypic presentations.

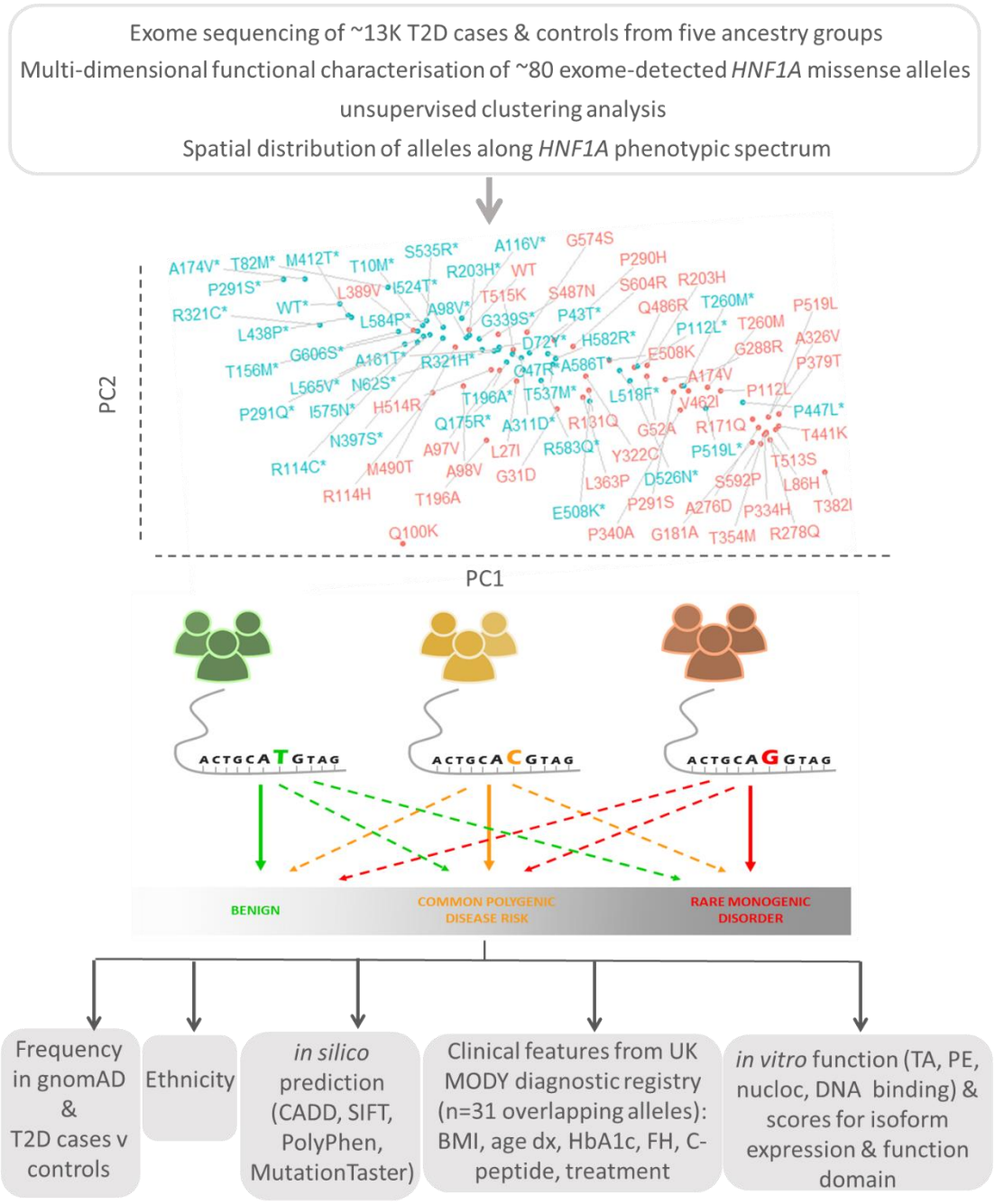


Figure 6.1. Summary of Chapter 3: Unsupervised clustering of exome-detected *HNF1A* alleles using multi-dimensional functional data.

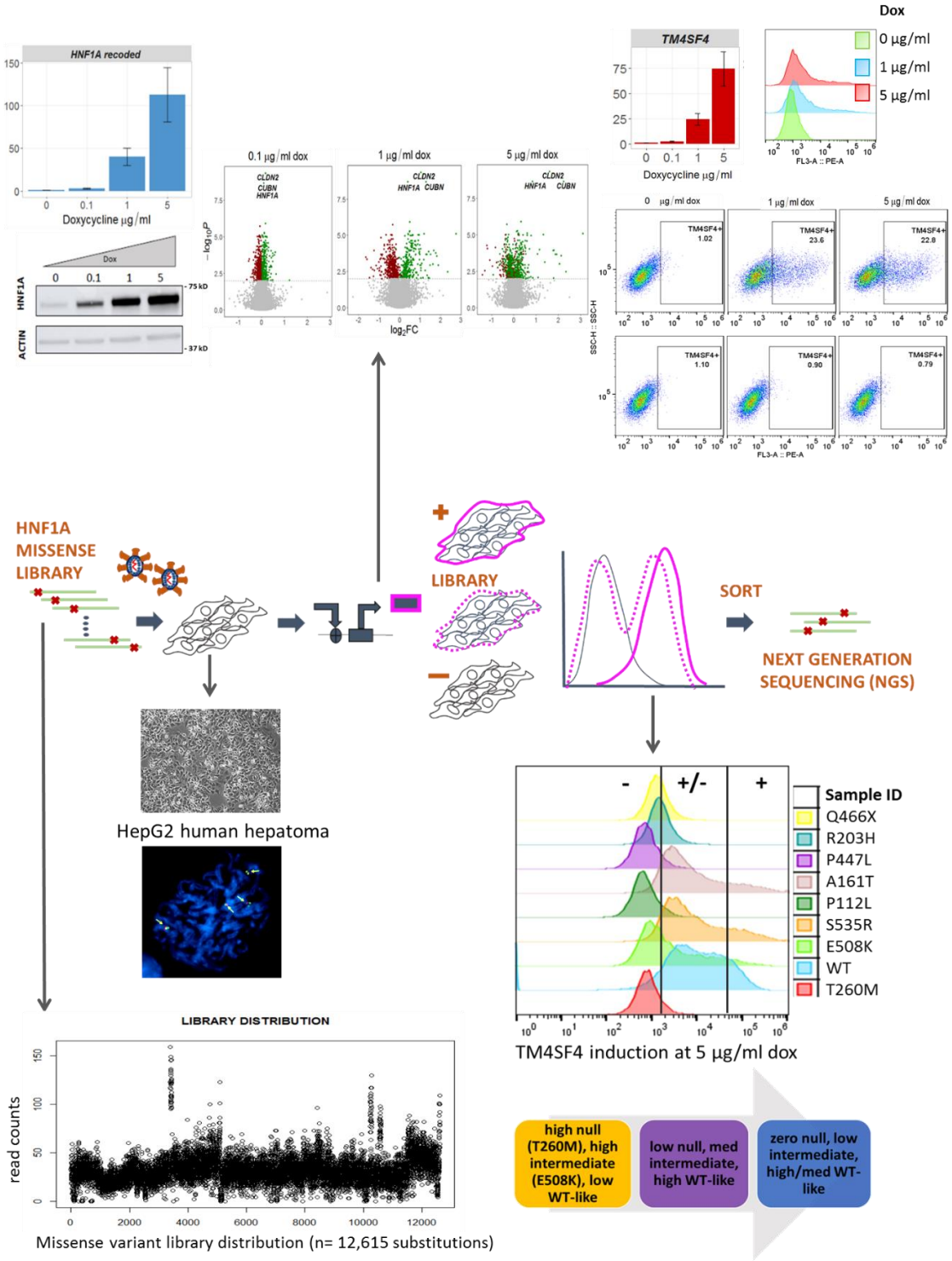


Figure 6.2. Summary of Chapter 4: Building a Multiplexed Assay for Variant Effects (MAVE) to screen all possible *HNF1A* missense alleles**.**

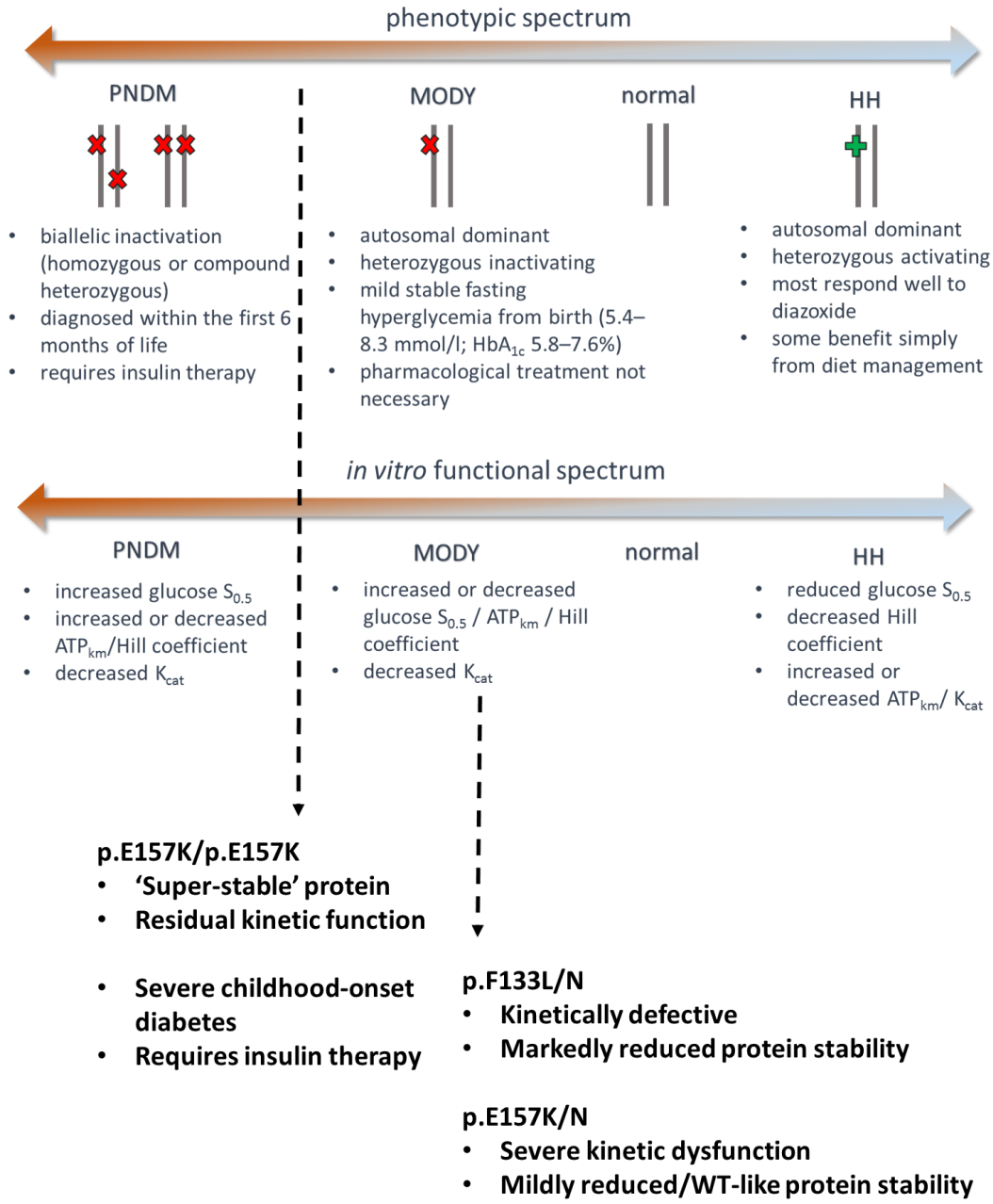


Figure 6.3. Summary of Chapter 5: Comprehensive functional assessment of *GCK* missense alleles identified in clinical genetic testing**.**

Appendix

References

Adzhubei I, Jordan DM, Sunyaev SR. 2013. Predicting functional effect of human missense mutations using PolyPhen-2. *Curr Protoc Hum Genet* **Chapter 7**: Unit7 20.

Aguilar-Bryan L, Nichols CG, Wechsler SW, Clement JPt, Boyd AE, 3rd, Gonzalez G, Herrera-Sosa H, Nguy K, Bryan J, Nelson DA. 1995. Cloning of the beta cell high-affinity sulfonylurea receptor: a regulator of insulin secretion. *Science* **268**: 423-426.

Alkuraya FS. 2015. Human knockout research: new horizons and opportunities. *Trends Genet* **31**: 108-115.

Angov E. 2011. Codon usage: nature's roadmap to expression and folding of proteins. *Biotechnol J* **6**: 650-659.

Arden C, Trainer A, de la Iglesia N, Scougall KT, Gloyn AL, Lange AJ, Shaw JA, Matschinsky FM, Agius L. 2007. Cell biology assessment of glucokinase mutations V62M and G72R in pancreatic beta-cells: evidence for cellular instability of catalytic activity. *Diabetes* **56**: 1773-1782.

Awata T, Kurihara S, Inoue K, Inoue I, Takei S, Ishii C, Negishi K, Namai K, Kanazawa Y, Kuzuya T et al. 1998. A novel missense mutation in the homeodomain of the hepatocyte nuclear factor-1alpha/maturity-onset diabetes of the young 3 in a Japanese early-onset type 2 diabetic patient and time-course of glucose-stimulated insulin secretion. *Diabetes Care* **21**: 1569-1571.

Babenko AP, Polak M, Cave H, Busiah K, Czernichow P, Scharfmann R, Bryan J, Aguilar-Bryan L, Vaxillaire M, Froguel P. 2006. Activating mutations in the ABCC8 gene in neonatal diabetes mellitus. *N Engl J Med* **355**: 456-466.

Bach I, 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**: 4229-4242.

Balamurugan K, Bjorkhaug L, Mahajan S, Kanthimathi S, Njolstad PR, Srinivasan N, Mohan V, Radha V. 2016. Structure-function studies of HNF1A (MODY3) gene mutations in South Indian patients with monogenic diabetes. *Clin Genet* **90**: 486-495.

Bamshad MJ, Ng SB, Bigham AW, Tabor HK, Emond MJ, Nickerson DA, Shendure J. 2011. Exome sequencing as a tool for Mendelian disease gene discovery. *Nat Rev Genet* **12**: 745-755.

Barrett TG, Bundey SE, Macleod AF. 1995. Neurodegeneration and diabetes: UK nationwide study of Wolfram (DIDMOAD) syndrome. *Lancet* **346**: 1458-1463.

Bazalova Z, Rypackova B, Broz J, Brunerova L, Polak J, Rusavy Z, Treslova L, Andel M. 2010. Three novel mutations in MODY and its phenotype in three different Czech families. *Diabetes Res Clin Pract* **88**: 132-138.

Becker RA, Chambers, JM, Wilks, AR. 1988. The New S Language. Wadsworth & Brooks/Cole.

Beer NL, Osbak KK, van de Bunt M, Tribble ND, Steele AM, Wensley KJ, Edghill EL, Colcough K, Barrett A, Valentinova L et al. 2012. Insights into the pathogenicity of rare missense GCK variants from the identification and functional characterization of compound heterozygous and double mutations inherited in cis. *Diabetes Care* **35**: 1482-1484.

Beer NL, Tribble ND, McCulloch LJ, Roos C, Johnson PRV, Orho-Melander M, Gloyn AL. 2009. The P446L variant in GCKR associated with fasting plasma glucose and triglyceride levels exerts its effect through increased glucokinase activity in liver. *Human Molecular Genetics* **18**: 4081-4088.

Begg CB. 2002. On the use of familial aggregation in population-based case probands for calculating penetrance. *J Natl Cancer Inst* **94**: 1221-1226.

Bell GI, Xiang KS, Newman MV, Wu SH, Wright LG, Fajans SS, Spielman RS, Cox NJ. 1991. Gene for non-insulin-dependent diabetes mellitus (maturity-onset diabetes of the young subtype) is linked to DNA polymorphism on human chromosome 20q. *Proc Natl Acad Sci U S A* **88**: 1484-1488.

Bellanne-Chantelot C, Carette C, Riveline JP, Valero R, Gautier JF, Larger E, Reznik Y, Ducluzeau PH, Sola A, Hartemann-Heurtier A 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**: 503-508.

Bellanne-Chantelot C, Levy DJ, Carette C, Saint-Martin C, Riveline JP, Larger E, Valero R, Gautier JF, Reznik Y, Sola A et al. 2011. Clinical characteristics and diagnostic criteria of maturity-onset diabetes of the young (MODY) due to molecular anomalies of the HNF1A gene. *J Clin Endocrinol Metab* **96**: E1346-1351.

Bennett K, James C, Mutair A, Al-Shaikh H, Sinani A, Hussain K. 2011. Four novel cases of permanent neonatal diabetes mellitus caused by homozygous mutations in the glucokinase gene. *Pediatr Diabetes* **12**: 192-196.

Bingham C, Hattersley AT. 2004. Renal cysts and diabetes syndrome resulting from mutations in hepatocyte nuclear factor-1beta. *Nephrol Dial Transplant* **19**: 2703-2708.

Bjorkhaug L, Sagen JV, Thorsby P, Sovik O, Molven A, Njolstad PR. 2003. Hepatocyte nuclear factor-1 alpha gene mutations and diabetes in Norway. *J Clin Endocrinol Metab* **88**: 920-931.

Bjorkhaug L, Ye H, Horikawa Y, Sovik O, Molven A, Njolstad PR. 2000. MODY associated with two novel hepatocyte nuclear factor-1alpha loss-of-function mutations (P112L and Q466X). *Biochem Biophys Res Commun* **279**: 792-798.

Boj SF, Parrizas M, Maestro MA, Ferrer J. 2001. A transcription factor regulatory circuit in differentiated pancreatic cells. *Proc Natl Acad Sci U S A* **98**: 14481-14486.

Boj SF, van Es JH, Huch M, Li VS, Jose A, Hatzis P, Mokry M, Haegebarth A, van den Born M, Chambon P et al. 2012. Diabetes risk gene and Wnt effector Tcf7l2/TCF4 controls hepatic response to perinatal and adult metabolic demand. *Cell* **151**: 1595-1607.

Bonnefond A, Philippe J, Durand E, Dechaume A, Huyvaert M, Montagne L, Marre M, Balkau B, Fajardy I, Vambergue A et al. 2012. Whole-exome sequencing and high throughput genotyping identified KCNJ11 as the thirteenth MODY gene. *PLoS One* **7**: e37423.

Bort R, Martinez-Barbera JP, Beddington RS, Zaret KS. 2004. Hex homeobox gene-dependent tissue positioning is required for organogenesis of the ventral pancreas. *Development* **131**: 797-806.

Bowman P, Flanagan SE, Edghill EL, Damhuis A, Shepherd MH, Paisey R, Hattersley AT, Ellard S. 2012. Heterozygous ABCC8 mutations are a cause of MODY. *Diabetologia* **55**: 123-127.

Burke CV, Buettger CW, Davis EA, McClane SJ, Matschinsky FM, Raper SE. 1999. Cell-biological assessment of human glucokinase mutants causing maturity-onset diabetes of the young type 2 (MODY-2) or glucokinase-linked hyperinsulinaemia (GK-HI). *Biochem J* **342 (Pt 2)**: 345-352.

Butler MG, Dasouki MJ, Zhou XP, Talebizadeh Z, Brown M, Takahashi TN, Miles JH, Wang CH, Stratton R, Pilarski R et al. 2005. Subset of individuals with autism spectrum disorders and extreme macrocephaly associated with germline PTEN tumour suppressor gene mutations. *J Med Genet* **42**: 318-321.

Byrne MM, Sturis J, Clement K, Vionnet N, Pueyo ME, Stoffel M, Takeda J, Passa P, Cohen D, Bell GI et al. 1994. Insulin secretory abnormalities in subjects with hyperglycemia due to glucokinase mutations. *J Clin Invest* **93**: 1120-1130.

Castel SE, Cervera A, Mohammadi P, Aguet F, Reverter F, Wolman A, Guigo R, Vasileva A, Lappalainen T. 2017. Modified penetrance of coding variants by cis-regulatory variation shapes human traits. *bioRxiv* doi:10.1101/190397.

Chakera AJ, Spyer G, Vincent N, Ellard S, Hattersley AT, Dunne FP. 2014. The 0.1% of the population with glucokinase monogenic diabetes can be recognized by clinical characteristics in pregnancy: the Atlantic Diabetes in Pregnancy cohort. *Diabetes Care* **37**: 1230-1236.

Chakera AJ, Steele AM, Gloyn AL, Shepherd MH, Shields B, Ellard S, Hattersley AT. 2015. Recognition and Management of Individuals With Hyperglycemia Because of a Heterozygous Glucokinase Mutation. *Diabetes Care* **38**: 1383-1392.

Chandran S, Peng FY, Rajadurai VS, Lu YT, Chang KT, Flanagan SE, Ellard S, Hussain K. 2013. Paternally inherited ABCC8 mutation causing diffuse congenital hyperinsulinism. *Endocrinol Diabetes Metab Case Rep* **2013**: 130041.

Charrad M, Ghazzali N, Boiteau V, Niknafs A. 2014. NbClust: An R Package for Determining the Relevant Number of Clusters in a Data Set. Journal of Statistical Software, **61**(6), 1-36. URL http://www.jstatsoft.org/v61/i06/

Chen R, Shi L, Hakenberg J, Naughton B, Sklar P, Zhang J, Zhou H, Tian L, Prakash O, Lemire M et al. 2016. Analysis of 589,306 genomes identifies individuals resilient to severe Mendelian childhood diseases. *Nat Biotechnol* **34**: 531-538.

Christesen HB, Jacobsen BB, Odili S, Buettger C, Cuesta-Munoz A, Hansen T, Brusgaard K, Massa O, Magnuson MA, Shiota C et al. 2002. The second activating glucokinase mutation (A456V): implications for glucose homeostasis and diabetes therapy. *Diabetes* **51**: 1240-1246.

Christesen HB, Tribble ND, Molven A, Siddiqui J, Sandal T, Brusgaard K, Ellard S, Njolstad PR, Alm J, Brock Jacobsen B et al. 2008. Activating glucokinase (GCK) mutations as a cause of medically responsive congenital hyperinsulinism: prevalence in children and characterisation of a novel GCK mutation. *Eur J Endocrinol* **159**: 27-34.

Colclough K, Bellanne-Chantelot C, Saint-Martin C, Flanagan SE, Ellard S. 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**: 669-685.

Colclough K, Saint-Martin C, Timsit J, Ellard S, Bellanne-Chantelot C. 2014. Clinical utility gene card for: Maturity-onset diabetes of the young. *Eur J Hum Genet* **22**.

Colombo C, Porzio O, Liu M, Massa O, Vasta M, Salardi S, Beccaria L, Monciotti C, Toni S, Pedersen O et al. 2008. Seven mutations in the human insulin gene linked to permanent neonatal/infancy-onset diabetes mellitus. *J Clin Invest* **118**: 2148-2156.

Consortium EP. 2012. An integrated encyclopedia of DNA elements in the human genome. *Nature* **489**: 57-74.

Consortium STD, Estrada K, Aukrust I, Bjorkhaug L, Burtt NP, Mercader JM, Garcia-Ortiz H, Huerta-Chagoya A, Moreno-Macias H, Walford G et al. 2014. Association of a low-frequency variant in HNF1A with type 2 diabetes in a Latino population. *JAMA* **311**: 2305-2314.

Cooper GM. 2015. Parlez-vous VUS? *Genome Res* **25**: 1423-1426.

Costa A, Bescos M, Velho G, Chevre J, Vidal J, Sesmilo G, Bellanne-Chantelot C, Froguel P, Casamitjana R, Rivera-Fillat F et al. 2000. Genetic and clinical characterisation of maturity-onset diabetes of the young in Spanish families. *Eur J Endocrinol* **142**: 380-386.

Couch FJ Wang X McGuffog L Lee A Olswold C Kuchenbaecker KB Soucy P Fredericksen Z Barrowdale D Dennis J et al. 2013. Genome-wide association study in BRCA1 mutation carriers identifies novel loci associated with breast and ovarian cancer risk. *PLoS Genet* **9**: e1003212.

Cutting GR. 2014. Annotating DNA variants is the next major goal for human genetics. *Am J Hum Genet* **94**: 5-10.

Daber R, Stayrook S, Rosenberg A, Lewis M. 2007. Structural analysis of lac repressor bound to allosteric effectors. *J Mol Biol* **370**: 609-619.

Davis EA, Cuesta-Munoz A, Raoul M, Buettger C, Sweet I, Moates M, Magnuson MA, Matschinsky FM. 1999. Mutants of glucokinase cause hypoglycaemia- and hyperglycaemia syndromes and their analysis illuminates fundamental quantitative concepts of glucose homeostasis. *Diabetologia* **42**: 1175-1186.

Davis TME, Kauhanen J, Davis WA. 2017. Pneumococcal vaccination and incident hospitalisation for pneumonia in type 2 diabetes: the Fremantle Diabetes Study Phase II. *Intern Med J* **47**: 1206-1210.

DCCTRG. 1998. Effect of intensive therapy on residual beta-cell function in patients with type 1 diabetes in the diabetes control and complications trial. A randomized, controlled trial. The Diabetes Control and Complications Trial Research Group. *Ann Intern Med* **128**: 517-523.

De Franco E, Shaw-Smith C, Flanagan SE, Edghill EL, Wolf J, Otte V, Ebinger F, Varthakavi P, Vasanthi T, Edvardsson S et al. 2013. Biallelic PDX1 (insulin promoter factor 1) mutations causing neonatal diabetes without exocrine pancreatic insufficiency. *Diabet Med* **30**: e197-200.

Doria A, Patti ME, Kahn CR. 2008. The emerging genetic architecture of type 2 diabetes. *Cell Metab* **8**: 186-200.

Dunne MJ, Cosgrove KE, Shepherd RM, Aynsley-Green A, Lindley KJ. 2004. Hyperinsulinism in infancy: from basic science to clinical disease. *Physiol Rev* **84**: 239-275.

Dupuis J Langenberg C Prokopenko I Saxena R Soranzo N Jackson AU Wheeler E Glazer NL Bouatia-Naji N Gloyn AL et al. 2010. New genetic loci implicated in fasting glucose homeostasis and their impact on type 2 diabetes risk. *Nat Genet* **42**: 105-116.

Durmaz E, Flanagan S, Berdeli A, Semiz S, Akcurin S, Ellard S, Bircan I. 2012. Variability in the age at diagnosis of diabetes in two unrelated patients with a homozygous glucokinase gene mutation. *J Pediatr Endocrinol Metab* **25**: 805-808.

Edghill EL, Bingham C, Ellard S, Hattersley AT. 2006. Mutations in hepatocyte nuclear factor-1beta and their related phenotypes. *J Med Genet* **43**: 84-90.

Edghill EL, Flanagan SE, Patch AM, Boustred C, Parrish A, Shields B, Shepherd MH, Hussain K, Kapoor RR, Malecki M 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**: 1034-1042.

Eide SA, Raeder H, Johansson S, Midthjell K, Sovik O, Njolstad PR, Molven A. 2008. Prevalence of HNF1A (MODY3) mutations in a Norwegian population (the HUNT2 Study). *Diabet Med* **25**: 775-781.

Eilbeck K, Quinlan A, Yandell M. 2017. Settling the score: variant prioritization and Mendelian disease. *Nat Rev Genet* **18**: 599-612.

Ellard S, Beards F, Allen LI, Shepherd M, Ballantyne E, Harvey R, Hattersley AT. 2000. A high prevalence of glucokinase mutations in gestational diabetic subjects selected by clinical criteria. *Diabetologia* **43**: 250-253.

Ellard S, Bellanne-Chantelot C, Hattersley AT, European Molecular Genetics Quality Network Mg. 2008. Best practice guidelines for the molecular genetic diagnosis of maturity-onset diabetes of the young. *Diabetologia* **51**: 546-553.

Ellard S, Lango Allen H, De Franco E, Flanagan SE, Hysenaj G, Colclough K, Houghton JA, Shepherd M, Hattersley AT, Weedon MN et al. 2013. Improved genetic testing for monogenic diabetes using targeted next-generation sequencing. *Diabetologia* **56**: 1958-1963.

Estalella I, Rica I, Perez de Nanclares G, Bilbao JR, Vazquez JA, San Pedro JI, Busturia MA, Castano L, Spanish MG. 2007. Mutations in GCK and HNF-1alpha explain the majority of cases with clinical diagnosis of MODY in Spain. *Clin Endocrinol (Oxf)* **67**: 538-546.

Fajans SS, Bell GI, Bowden DW, Halter JB, Polonsky KS. 1994. Maturity-onset diabetes of the young. *Life Sci* **55**: 413-422.

Fajans SS, Bell GI, Polonsky KS. 2001. Molecular mechanisms and clinical pathophysiology of maturity-onset diabetes of the young. *N Engl J Med* **345**: 971-980.

Fendler W, Borowiec M, Baranowska-Jazwiecka A, Szadkowska A, Skala-Zamorowska E, Deja G, Jarosz-Chobot P, Techmanska I, Bautembach-Minkowska J, Mysliwiec M et al. 2012. Prevalence of monogenic diabetes amongst Polish children after a nationwide genetic screening campaign. *Diabetologia* **55**: 2631-2635.

Ferrer J. 2002. A genetic switch in pancreatic beta-cells: implications for differentiation and haploinsufficiency. *Diabetes* **51**: 2355-2362.

Findlay GM, Boyle EA, Hause RJ, Klein JC, Shendure J. 2014. Saturation editing of genomic regions by multiplex homology-directed repair. *Nature* **513**: 120-123.

Flanagan SE, Edghill EL, Gloyn AL, Ellard S, Hattersley AT. 2006. Mutations in KCNJ11, which encodes Kir6.2, are a common cause of diabetes diagnosed in the first 6 months of life, with the phenotype determined by genotype. *Diabetologia* **49**: 1190-1197.

Flanagan SE, Kapoor RR, Mali G, Cody D, Murphy N, Schwahn B, Siahanidou T, Banerjee I, Akcay T, Rubio-Cabezas O et al. 2010. Diazoxide-responsive hyperinsulinemic hypoglycemia caused by HNF4A gene mutations. *Eur J Endocrinol* **162**: 987-992.

Flannick J, Beer NL, Bick AG, Agarwala V, Molnes J, Gupta N, Burtt NP, Florez JC, Meigs JB, Taylor H et al. 2013. Assessing the phenotypic effects in the general population of rare variants in genes for a dominant Mendelian form of diabetes. *Nat Genet* **45**: 1380-1385.

Flannick J, Johansson S, Njolstad PR. 2016. Common and rare forms of diabetes mellitus: towards a continuum of diabetes subtypes. *Nat Rev Endocrinol* **12**: 394-406.

Frayling TM, Bulman MP, Appleton M, Hattersley AT, Ellard S. 1997. A rapid screening method for hepatocyte nuclear factor 1 alpha frameshift mutations; prevalence in maturity-onset diabetes of the young and late-onset non-insulin dependent diabetes. *Hum Genet* **101**: 351-354.

Frayling TM, Evans JC, Bulman MP, Pearson E, Allen L, Owen K, Bingham C, Hannemann M, Shepherd M, Ellard S et al. 2001. beta-cell genes and diabetes: molecular and clinical characterization of mutations in transcription factors. *Diabetes* **50 Suppl 1**: S94-100.

Froguel P, Vaxillaire M, Sun F, Velho G, Zouali H, Butel MO, Lesage S, Vionnet N, Clement K, Fougerousse F et al. 1992. Close linkage of glucokinase locus on chromosome 7p to early-onset non-insulin-dependent diabetes mellitus. *Nature* **356**: 162-164.

Fuchsberger C Flannick J Teslovich TM Mahajan A Agarwala V Gaulton KJ Ma C Fontanillas P Moutsianas L McCarthy DJ et al. 2016. The genetic architecture of type 2 diabetes. *Nature* **536**: 41-47.

Galan M, Garcia-Herrero CM, Azriel S, Gargallo M, Duran M, Gorgojo JJ, Andia VM, Navas MA. 2011. Differential effects of HNF-1alpha mutations associated with familial young-onset diabetes on target gene regulation. *Mol Med* **17**: 256-265.

Garin I, Edghill EL, Akerman I, Rubio-Cabezas O, Rica I, Locke JM, Maestro MA, Alshaikh A, Bundak R, del Castillo G et al. 2010. Recessive mutations in the INS gene result in neonatal diabetes through reduced insulin biosynthesis. *Proc Natl Acad Sci U S A* **107**: 3105-3110.

Gasperini M, Starita L, Shendure J. 2016. The power of multiplexed functional analysis of genetic variants. *Nat Protoc* **11**: 1782-1787.

Gaulton KJ Ferreira T Lee Y Raimondo A Magi R Reschen ME Mahajan A Locke A William Rayner N Robertson N et al. 2015. Genetic fine mapping and genomic

annotation defines causal mechanisms at type 2 diabetes susceptibility loci. *Nat Genet* doi:10.1038/ng.3437.

Glaser B, Kesavan P, Heyman M, Davis E, Cuesta A, Buchs A, Stanley CA, Thornton PS, Permutt MA, Matschinsky FM et al. 1998. Familial hyperinsulinism caused by an activating glucokinase mutation. *N Engl J Med* **338**: 226-230.

Glaser B, Thornton P, Otonkoski T, Junien C. 2000. Genetics of neonatal hyperinsulinism. *Arch Dis Child Fetal Neonatal Ed* **82**: F79-86.

Gloyn AL. 2003. Glucokinase (GCK) mutations in hyper- and hypoglycemia: maturity-onset diabetes of the young, permanent neonatal diabetes, and hyperinsulinemia of infancy. *Hum Mutat* **22**: 353-362.

Gloyn AL, Odili S, Zelent D, Buettger C, Castleden HA, Steele AM, Stride A, Shiota C, Magnuson MA, Lorini R et al. 2005. Insights into the structure and regulation of glucokinase from a novel mutation (V62M), which causes maturity-onset diabetes of the young. *J Biol Chem* **280**: 14105-14113.

Gloyn AL, Pearson ER, Antcliff JF, Proks P, Bruining GJ, Slingerland AS, Howard N, Srinivasan S, Silva JM, Molnes J et al. 2004. Activating mutations in the gene encoding the ATP-sensitive potassium-channel subunit Kir6.2 and permanent neonatal diabetes. *N Engl J Med* **350**: 1838-1849.

Gloyn AL, Siddiqui J, Ellard S. 2006. Mutations in the genes encoding the pancreatic beta-cell KATP channel subunits Kir6.2 (KCNJ11) and SUR1 (ABCC8) in diabetes mellitus and hyperinsulinism. *Hum Mutat* **27**: 220-231.

Glucksmann MA, Lehto M, Tayber O, Scotti S, Berkemeier L, Pulido JC, Wu Y, Nir WJ, Fang L, Markel P et al. 1997. Novel mutations and a mutational hotspot in the MODY3 gene. *Diabetes* **46**: 1081-1086.

Godart F, Bellanne-Chantelot C, Clauin S, Gragnoli C, Abderrahmani A, Blanche H, Boutin P, Chevre JC, Froguel P, Bailleul B. 2000. Identification of seven novel nucleotide variants in the hepatocyte nuclear factor-1alpha (TCF1) promoter region in MODY patients. *Hum Mutat* **15**: 173-180.

Goring HH, Terwilliger JD, Blangero J. 2001. Large upward bias in estimation of locus-specific effects from genomewide scans. *Am J Hum Genet* **69**: 1357-1369.

Gosmain Y, Katz LS, Masson MH, Cheyssac C, Poisson C, Philippe J. 2012. Pax6 is crucial for beta-cell function, insulin biosynthesis, and glucose-induced insulin secretion. *Mol Endocrinol* **26**: 696-709.

Gossen M, Freundlieb S, Bender G, Muller G, Hillen W, Bujard H. 1995. Transcriptional activation by tetracyclines in mammalian cells. *Science* **268**: 1766-1769.

Grantham R. 1974. Amino acid difference formula to help explain protein evolution. *Science* **185**: 862-864.

Grimm DG, Azencott CA, Aicheler F, Gieraths U, MacArthur DG, Samocha KE, Cooper DN, Stenson PD, Daly MJ, Smoller JW et al. 2015. The evaluation of tools used to predict the impact of missense variants is hindered by two types of circularity. *Hum Mutat* **36**: 513-523.

Grimsby J, Sarabu R, Corbett WL, Haynes NE, Bizzarro FT, Coffey JW, Guertin KR, Hilliard DW, Kester RF, Mahaney PE et al. 2003. Allosteric activators of glucokinase: potential role in diabetes therapy. *Science* **301**: 370-373.

Group HSCR, Metzger BE, Lowe LP, Dyer AR, Trimble ER, Chaovarindr U, Coustan DR, Hadden DR, McCance DR, Hod M et al. 2008. Hyperglycemia and adverse pregnancy outcomes. *N Engl J Med* **358**: 1991-2002.

Grupe A, Hultgren B, Ryan A, Ma YH, Bauer M, Stewart TA. 1995. Transgenic knockouts reveal a critical requirement for pancreatic beta cell glucokinase in maintaining glucose homeostasis. *Cell* **83**: 69-78.

Guenat E, Seematter G, Philippe J, Temler E, Jequier E, Tappy L. 2000. Counterregulatory responses to hypoglycemia in patients with glucokinase gene mutations. *Diabetes Metab* **26**: 377-384.

Hameed S, Ellard S, Woodhead HJ, Neville KA, Walker JL, Craig ME, Armstrong T, Yu L, Eisenbarth GS, Hattersley AT et al. 2011. Persistently autoantibody negative (PAN) type 1 diabetes mellitus in children. *Pediatr Diabetes* **12**: 142-149.

Hansen T, Eiberg H, Rouard M, Vaxillaire M, Moller AM, Rasmussen SK, Fridberg M, Urhammer SA, Holst JJ, Almind K et al. 1997. Novel MODY3 mutations in the hepatocyte nuclear factor-1alpha gene: evidence for a hyperexcitability of pancreatic beta-cells to intravenous secretagogues in a glucose-tolerant carrier of a P447L mutation. *Diabetes* **46**: 726-730.

Harries LW, Ellard S, Stride A, Morgan NG, Hattersley AT. 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**: 2216-2224.

Harries LW, Sloman MJ, Sellers EA, Hattersley AT, Ellard S. 2008. Diabetes susceptibility in the Canadian Oji-Cree population is moderated by abnormal mRNA processing of HNF1A G319S transcripts. *Diabetes* **57**: 1978-1982.

Hart AW, Mella S, Mendrychowski J, van Heyningen V, Kleinjan DA. 2013. The developmental regulator Pax6 is essential for maintenance of islet cell function in the adult mouse pancreas. *PLoS One* **8**: e54173.

Hattersley AT. 1998. Maturity-onset diabetes of the young: clinical heterogeneity explained by genetic heterogeneity. *Diabet Med* **15**: 15-24.

Hattersley AT, Beards F, Ballantyne E, Appleton M, Harvey R, Ellard S. 1998. Mutations in the glucokinase gene of the fetus result in reduced birth weight. *Nat Genet* **19**: 268-270.

Hattersley AT, Patel KA. 2017. Precision diabetes: learning from monogenic diabetes. *Diabetologia* **60**: 769-777.

Hattersley AT, Turner RC, Permutt MA, Patel P, Tanizawa Y, Chiu KC, O'Rahilly S, Watkins PJ, Wainscoat JS. 1992. Linkage of type 2 diabetes to the glucokinase gene. *Lancet* **339**: 1307-1310.

Hegele RA, Cao H, Harris SB, Hanley AJ, Zinman B. 1999. The hepatic nuclear factor-1alpha G319S variant is associated with early-onset type 2 diabetes in Canadian Oji-Cree. *J Clin Endocrinol Metab* **84**: 1077-1082.

Hicks S, Wheeler DA, Plon SE, Kimmel M. 2011. Prediction of missense mutation functionality depends on both the algorithm and sequence alignment employed. *Hum Mutat* **32**: 661-668.

Hill AV. 1910. The possible effects of the aggregation of the molecules of haemoglobin on its dissociation curves. *Journal of Physiology* **40**: Proceedings iv–vii.

Hollander MC, Blumenthal GM, Dennis PA. 2011. PTEN loss in the continuum of common cancers, rare syndromes and mouse models. *Nat Rev Cancer* **11**: 289-301.

Holmkvist J, Cervin C, Lyssenko V, Winckler W, Anevski D, Cilio C, Almgren P, Berglund G, Nilsson P, Tuomi T et al. 2006. Common variants in HNF-1 alpha and risk of type 2 diabetes. *Diabetologia* **49**: 2882-2891.

Horikawa Y, Iwasaki N, Hara M, Furuta H, Hinokio Y, Cockburn BN, Lindner T, Yamagata K, Ogata M, Tomonaga O et al. 1997. Mutation in hepatocyte nuclear factor-1 beta gene (TCF2) associated with MODY. *Nat Genet* **17**: 384-385.

Hsu PD, Lander ES, Zhang F. 2014. Development and applications of CRISPR-Cas9 for genome engineering. *Cell* **157**: 1262-1278.

Hua H, Shang L, Martinez H, Freeby M, Gallagher MP, Ludwig T, Deng L, Greenberg E, Leduc C, Chung WK et al. 2013. iPSC-derived beta cells model diabetes due to glucokinase deficiency. *J Clin Invest* **123**: 3146-3153.

Hwang JS, Shin CH, Yang SW, Jung SY, Huh N. 2006. Genetic and clinical characteristics of Korean maturity-onset diabetes of the young (MODY) patients. *Diabetes Res Clin Pract* **74**: 75-81.

Ikemura T. 1982. Correlation between the abundance of yeast transfer RNAs and the occurrence of the respective codons in protein genes. Differences in synonymous

codon choice patterns of yeast and Escherichia coli with reference to the abundance of isoaccepting transfer RNAs. *J Mol Biol* **158**: 573-597.
Inagaki N, Gonoi T, Clement JPt, Namba N, Inazawa J, Gonzalez G, Aguilar-Bryan L, Seino S, Bryan J. 1995. Reconstitution of IKATP: an inward rectifier subunit plus the sulfonylurea receptor. *Science* **270**: 1166-1170.
Ingolia NT, Ghaemmaghami S, Newman JR, Weissman JS. 2009. Genome-wide analysis in vivo of translation with nucleotide resolution using ribosome profiling. *Science* **324**: 218-223.
Inoue F, Ahituv N. 2015. Decoding enhancers using massively parallel reporter assays. *Genomics* **106**: 159-164.
Irgens HU, Molnes J, Johansson BB, Ringdal M, Skrivarhaug T, Undlien DE, Sovik O, Joner G, Molven A, Njolstad PR. 2013. Prevalence of monogenic diabetes in the population-based Norwegian Childhood Diabetes Registry. *Diabetologia* **56**: 1512-1519.
Isomaa B, Henricsson M, Lehto M, Forsblom C, Karanko S, Sarelin L, Haggblom M, Groop L. 1998. Chronic diabetic complications in patients with MODY3 diabetes. *Diabetologia* **41**: 467-473.
Iynedjian PB. 1993. Mammalian glucokinase and its gene. *Biochem J* **293 (Pt 1)**: 1-13.
Jain AK, Dubes RC. 1988. Algorithms for clustering data. Prentice Hall, New Jersey.
Jain AK, Murty MN, Flynn PJ. 1999. Data clustering: A review. *ACM Computing Surveys* **31**(3):264-323
Johansen A, Ek J, Mortensen HB, Pedersen O, Hansen T. 2005. Half of clinically defined maturity-onset diabetes of the young patients in Denmark do not have mutations in HNF4A, GCK, and TCF1. *J Clin Endocrinol Metab* **90**: 4607-4614.
Jolma A, Yan J, Whitington T, Toivonen J, Nitta KR, Rastas P, Morgunova E, Enge M, Taipale M, Wei G et al. 2013. DNA-binding specificities of human transcription factors. *Cell* **152**: 327-339.
Kamata K, Mitsuya M, Nishimura T, Eiki J, Nagata Y. 2004. Structural basis for allosteric regulation of the monomeric allosteric enzyme human glucokinase. *Structure* **12**: 429-438.
Kaufman L, Rousseeuw PJ. 1990. *Finding groups in data : an introduction to cluster analysis*. Wiley, New York ; Chichester.
Kesavan P, Wang L, Davis E, Cuesta A, Sweet I, Niswender K, Magnuson MA, Matschinsky FM. 1997. Structural instability of mutant beta-cell glucokinase: implications for the molecular pathogenesis of maturity-onset diabetes of the young (type-2). *Biochem J* **322 (Pt 1)**: 57-63.
Kircher M, Witten DM, Jain P, O'Roak BJ, Cooper GM, Shendure J. 2014. A general framework for estimating the relative pathogenicity of human genetic variants. *Nat Genet* **46**: 310-315.
Klupa T, Warram JH, Antonellis A, Pezzolesi M, Nam M, Malecki MT, Doria A, Rich SS, Krolewski AS. 2002. Determinants of the development of diabetes (maturity-onset diabetes of the young-3) in carriers of HNF-1alpha mutations: evidence for parent-of-origin effect. *Diabetes Care* **25**: 2292-2301.
Kosuri S, Goodman DB, Cambray G, Mutalik VK, Gao Y, Arkin AP, Endy D, Church GM. 2013. Composability of regulatory sequences controlling transcription and translation in Escherichia coli. *Proc Natl Acad Sci U S A* **110**: 14024-14029.
Kousta E, Ellard S, Allen LI, Saker PJ, Huxtable SJ, Hattersley AT, McCarthy MI. 2001. Glucokinase mutations in a phenotypically selected multiethnic group of women with a history of gestational diabetes. *Diabet Med* **18**: 683-684.
Kozyraki R, Fyfe J, Kristiansen M, Gerdes C, Jacobsen C, Cui S, Christensen EI, Aminoff M, de la Chapelle A, Krahe R et al. 1999. The intrinsic factor-vitamin B12 receptor, cubilin, is a high-affinity apolipoprotein A-I receptor facilitating endocytosis of high-density lipoprotein. *Nat Med* **5**: 656-661.
Kristinsson SY, Thorolfsdottir ET, Talseth B, Steingrimsson E, Thorsson AV, Helgason T, Hreidarsson AB, Arngrimsson R. 2001. MODY in Iceland is associated with

mutations in HNF-1alpha and a novel mutation in NeuroD1. *Diabetologia* **44**: 2098-2103.

Kropff J, Selwood MP, McCarthy MI, Farmer AJ, Owen KR. 2011. Prevalence of monogenic diabetes in young adults: a community-based, cross-sectional study in Oxfordshire, UK. *Diabetologia* **54**: 1261-1263.

Kryukov GV, Pennacchio LA, Sunyaev SR. 2007. Most rare missense alleles are deleterious in humans: implications for complex disease and association studies. *Am J Hum Genet* **80**: 727-739.

Kumar P, Henikoff S, Ng PC. 2009. Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm. *Nat Protoc* **4**: 1073-1081.

Lappalainen T, Montgomery SB, Nica AC, Dermitzakis ET. 2011. Epistatic selection between coding and regulatory variation in human evolution and disease. *Am J Hum Genet* **89**: 459-463.

Lauc G, Essafi A, Huffman JE, Hayward C, Knezevic A, Kattla JJ, Polasek O, Gornik O, Vitart V, Abrahams JL 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**: e1001256.

Law CW, Chen Y, Shi W, Smyth GK. 2014. voom: Precision weights unlock linear model analysis tools for RNA-seq read counts. *Genome Biol* **15**: R29.

Ledermann HM. 1995. Is maturity onset diabetes at young age (MODY) more common in Europe than previously assumed? *Lancet* **345**: 648.

Legendre P, Legendre L, Legendre L. 1998. *Numerical ecology*. Elsevier, Amsterdam ; Oxford.

Lehto M, Wipemo C, Ivarsson SA, Lindgren C, Lipsanen-Nyman M, Weng J, Wibell L, Widen E, Tuomi T, Groop L. 1999. High frequency of mutations in MODY and mitochondrial genes in Scandinavian patients with familial early-onset diabetes. *Diabetologia* **42**: 1131-1137.

Lek M, Karczewski KJ, Minikel EV, Samocha KE, Banks E, Fennell T, O'Donnell-Luria AH, Ware JS, Hill AJ, Cummings BB et al. 2016. Analysis of protein-coding genetic variation in 60,706 humans. *Nature* **536**: 285-291.

Liang Y, Kesavan P, Wang LQ, Niswender K, Tanizawa Y, Permutt MA, Magnuson MA, Matschinsky FM. 1995. Variable Effects of Maturity-Onset-Diabetes-of-Youth (Mody)-Associated Glucokinase Mutations on Substrate Interactions and Stability of the Enzyme. *Biochemical Journal* **309**: 167-173.

Liang Y, Najafi H, Smith RM, Zimmerman EC, Magnuson MA, Tal M, Matschinsky FM. 1992. Concordant glucose induction of glucokinase, glucose usage, and glucose-stimulated insulin release in pancreatic islets maintained in organ culture. *Diabetes* **41**: 792-806.

Liao Y, Smyth GK, Shi W. 2014. featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. *Bioinformatics* **30**: 923-930.

Liaw D, Marsh DJ, Li J, Dahia PL, Wang SI, Zheng Z, Bose S, Call KM, Tsou HC, Peacocke M et al. 1997. Germline mutations of the PTEN gene in Cowden disease, an inherited breast and thyroid cancer syndrome. *Nat Genet* **16**: 64-67.

Ljubicic S, Polak K, Fu A, Wiwczar J, Szlyk B, Chang Y, Alvarez-Perez JC, Bird GH, Walensky LD, Garcia-Ocana A et al. 2015. Phospho-BAD BH3 mimicry protects beta cells and restores functional beta cell mass in diabetes. *Cell Rep* **10**: 497-504.

Lorini R, Klersy C, d'Annunzio G, Massa O, Minuto N, Iafusco D, Bellanne-Chantelot C, Frongia AP, Toni S, Meschi F et al. 2009. Maturity-onset diabetes of the young in children with incidental hyperglycemia: a multicenter Italian study of 172 families. *Diabetes Care* **32**: 1864-1866.

MacArthur DG, Balasubramanian S, Frankish A, Huang N, Morris J, Walter K, Jostins L, Habegger L, Pickrell JK, Montgomery SB et al. 2012. A systematic survey of loss-of-function variants in human protein-coding genes. *Science* **335**: 823-828.

MacArthur DG, Manolio TA, Dimmock DP, Rehm HL, Shendure J, Abecasis GR, Adams DR, Altman RB, Antonarakis SE, Ashley EA et al. 2014. Guidelines for investigating causality of sequence variants in human disease. *Nature* **508**: 469-476.

Magnuson MA. 1990. Glucokinase gene structure. Functional implications of molecular genetic studies. *Diabetes* **39**: 523-527.

Mahajan A Taliun D Thurner M Robertson NR Torres JM Rayner NW Steinthorsdottir V Scott RA Grarup N Cook JP et al. 2018. Fine-mapping of an expanded set of type 2 diabetes loci to single-variant resolution using high-density imputation and islet-specific epigenome maps. *bioRxiv* doi:10.1101/245506.

Mahajan A, Wessel J, Willems S, Zhao W, Robertson NR, Chu AY, Gan W, Kitajima H, Scott RA, Barroso I et al. 2017. Refining The Accuracy Of Validated Target Identification Through Coding Variant Fine-Mapping In Type 2 Diabetes. *bioRxiv* doi:10.1101/144410.

Majithia AR, Flannick J, Shahinian P, Guo M, Bray MA, Fontanillas P, Gabriel SB, Go TDC, Project NJFAS, Consortium STD 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**: 13127-13132.

Majithia AR, Tsuda B, Agostini M, Gnanapradeepan K, Rice R, Peloso G, Patel KA, Zhang X, Broekema MF, Patterson N et al. 2016. Prospective functional classification of all possible missense variants in PPARG. *Nat Genet* **48**: 1570-1575.

Mali P, Yang L, Esvelt KM, Aach J, Guell M, DiCarlo JE, Norville JE, Church GM. 2013. RNA-guided human genome engineering via Cas9. *Science* **339**: 823-826.

Manning AK Hivert MF Scott RA Grimsby JL Bouatia-Naji N Chen H Rybin D Liu CT Bielak LF Prokopenko I et al. 2012. A genome-wide approach accounting for body mass index identifies genetic variants influencing fasting glycemic traits and insulin resistance. *Nat Genet* **44**: 659-669.

Manning KS, Cooper TA. 2017. The roles of RNA processing in translating genotype to phenotype. *Nat Rev Mol Cell Biol* **18**: 102-114.

Manrai AK, Funke BH, Rehm HL, Olesen MS, Maron BA, Szolovits P, Margulies DM, Loscalzo J, Kohane IS. 2016. Genetic Misdiagnoses and the Potential for Health Disparities. *N Engl J Med* **375**: 655-665.

Marshall BA, Permutt MA, Paciorkowski AR, Hoekel J, Karzon R, Wasson J, Viehover A, White NH, Shimony JS, Manwaring L et al. 2013. Phenotypic characteristics of early Wolfram syndrome. *Orphanet J Rare Dis* **8**: 64.

Mardia KV, Kent JT, Bibby JM. 1979. Multivariate Analysis, London: Academic Press.

Massa O, Meschi F, Cuesta-Munoz A, Caumo A, Cerutti F, Toni S, Cherubini V, Guazzarotti L, Sulli N, Matschinsky FM et al. 2001. High prevalence of glucokinase mutations in Italian children with MODY. Influence on glucose tolerance, first-phase insulin response, insulin sensitivity and BMI. *Diabetologia* **44**: 898-905.

Matschinsky FM. 2002. Regulation of pancreatic beta-cell glucokinase: from basics to therapeutics. *Diabetes* **51 Suppl 3**: S394-404.

Matschinsky FM. 2009. Assessing the potential of glucokinase activators in diabetes therapy. *Nat Rev Drug Discov* **8**: 399-416.

Matschinsky FM, Zelent B, Doliba N, Li C, Vanderkooi JM, Naji A, Sarabu R, Grimsby J. 2011. Glucokinase activators for diabetes therapy: May 2010 status report. *Diabetes Care* **34 Suppl 2**: S236-243.

McCarthy MI. 2015. Genomic medicine at the heart of diabetes management. *Diabetologia* **58**: 1725-1729.

McDonald TJ, Ellard S. 2013. Maturity onset diabetes of the young: identification and diagnosis. *Ann Clin Biochem* **50**: 403-415.

McDonald TJ, McEneny J, Pearson ER, Thanabalasingham G, Szopa M, Shields BM, Ellard S, Owen KR, Malecki MT, Hattersley AT et al. 2012. Lipoprotein

composition in HNF1A-MODY: differentiating between HNF1A-MODY and type 2 diabetes. *Clin Chim Acta* **413**: 927-932.

McDonald TJ, Shields BM, Lawry J, Owen KR, Gloyn AL, Ellard S, Hattersley AT. 2011. High-sensitivity CRP discriminates HNF1A-MODY from other subtypes of diabetes. *Diabetes Care* **34**: 1860-1862.

Melnikov A, Murugan A, Zhang X, Tesileanu T, Wang L, Rogov P, Feizi S, Gnirke A, Callan CG, Jr., Kinney JB et al. 2012. Systematic dissection and optimization of inducible enhancers in human cells using a massively parallel reporter assay. *Nat Biotechnol* **30**: 271-277.

Melnikov A, Zhang X, Rogov P, Wang L, Mikkelsen TS. 2014. Massively parallel reporter assays in cultured mammalian cells. *J Vis Exp* doi:10.3791/51719.

Michaelis L, Menten MM. 2013. The kinetics of invertin action. 1913. *FEBS Lett* **587**: 2712-2720.

Mighell TL, Evans S, O'Roak BJ. High-throughput functional analysis of PTEN variants reveals genotype-phenotype relationships; (PgmNr 154), Presented at the 67th Annual Meeting of The American Society of Human Genetics, October 20, 2017, Orlando, FL, USA.

Miosge LA, Field MA, Sontani Y, Cho V, Johnson S, Palkova A, Balakishnan B, Liang R, Zhang Y, Lyon S et al. 2015. Comparison of predicted and actual consequences of missense mutations. *Proc Natl Acad Sci U S A* **112**: E5189-5198.

Molven A, Njolstad PR. 2011. Role of molecular genetics in transforming diagnosis of diabetes mellitus. *Expert Rev Mol Diagn* **11**: 313-320.

Molven A, Ringdal M, Nordbo AM, Raeder H, Stoy J, Lipkind GM, Steiner DF, Philipson LH, Bergmann I, Aarskog D et al. 2008. Mutations in the insulin gene can cause MODY and autoantibody-negative type 1 diabetes. *Diabetes* **57**: 1131-1135.

Morris AP Voight BF Teslovich TM Ferreira T Segre AV Steinthorsdottir V Strawbridge RJ Khan H Grallert H Mahajan A et al. 2012. Large-scale association analysis provides insights into the genetic architecture and pathophysiology of type 2 diabetes. *Nat Genet* **44**: 981-990.

Murphy R, Ellard S, Hattersley AT. 2008. Clinical implications of a molecular genetic classification of monogenic beta-cell diabetes. *Nat Clin Pract Endocrinol Metab* **4**: 200-213.

Murtagh F, Legendre P. 2014. Ward's Hierarchical Agglomerative Clustering Method: Which Algorithms Implement Ward's Criterion? Journal of Classification, **31**, 274-295

Najmi LA, Aukrust I, Flannick J, Molnes J, Burtt N, Molven A, Groop L, Altshuler D, Johansson S, Bjorkhaug L et al. 2017. Functional Investigations of HNF1A Identify Rare Variants as Risk Factors for Type 2 Diabetes in the General Population. *Diabetes* **66**: 335-346.

Nakamura A, Terauchi Y. 2015. Present status of clinical deployment of glucokinase activators. *J Diabetes Investig* **6**: 124-132.

Neu A, Feldhahn L, Ehehalt S, Hub R, Ranke MB, Baden-Wurttemberg Dg. 2009. Type 2 diabetes mellitus in children and adolescents is still a rare disease in Germany: a population-based assessment of the prevalence of type 2 diabetes and MODY in patients aged 0-20 years. *Pediatr Diabetes* **10**: 468-473.

Ng MC, Cockburn BN, Lindner TH, Yeung VT, Chow CC, So WY, Li JK, Lo YM, Lee ZS, Cockram CS et al. 1999. Molecular genetics of diabetes mellitus in Chinese subjects: identification of mutations in glucokinase and hepatocyte nuclear factor-1alpha genes in patients with early-onset type 2 diabetes mellitus/MODY. *Diabet Med* **16**: 956-963.

Njolstad PR, Sovik O, Cuesta-Munoz A, Bjorkhaug L, Massa O, Barbetti F, Undlien DE, Shiota C, Magnuson MA, Molven A et al. 2001. Neonatal diabetes mellitus due to complete glucokinase deficiency. *N Engl J Med* **344**: 1588-1592.

O'Sullivan EP, Avalos G, O'Reilly M, Dennedy MC, Gaffney G, Dunne F, Atlantic DIPc. 2011. Atlantic Diabetes in Pregnancy (DIP): the prevalence and outcomes of

gestational diabetes mellitus using new diagnostic criteria. *Diabetologia* **54**: 1670-1675.

Oram RA, Jones AG, Besser RE, Knight BA, Shields BM, Brown RJ, Hattersley AT, McDonald TJ. 2014. The majority of patients with long-duration type 1 diabetes are insulin microsecretors and have functioning beta cells. *Diabetologia* **57**: 187-191.

Oram RA, McDonald TJ, Shields BM, Hudson MM, Shepherd MH, Hammersley S, Pearson ER, Hattersley AT, Team U. 2015. Most people with long-duration type 1 diabetes in a large population-based study are insulin microsecretors. *Diabetes Care* **38**: 323-328.

Oram RA, Patel K, Hill A, Shields B, McDonald TJ, Jones A, Hattersley AT, Weedon MN. 2016. A Type 1 Diabetes Genetic Risk Score Can Aid Discrimination Between Type 1 and Type 2 Diabetes in Young Adults. *Diabetes Care* **39**: 337-344.

Osbak KK, Colclough K, Saint-Martin C, Beer NL, Bellanne-Chantelot C, Ellard S, Gloyn AL. 2009. Update on mutations in glucokinase (GCK), which cause maturity-onset diabetes of the young, permanent neonatal diabetes, and hyperinsulinemic hypoglycemia. *Hum Mutat* **30**: 1512-1526.

Owen KR, Thanabalasingham G, James TJ, Karpe F, Farmer AJ, McCarthy MI, Gloyn AL. 2010. Assessment of high-sensitivity C-reactive protein levels as diagnostic discriminator of maturity-onset diabetes of the young due to HNF1A mutations. *Diabetes Care* **33**: 1919-1924.

Pagliuca FW, Millman JR, Gurtler M, Segel M, Van Dervort A, Ryu JH, Peterson QP, Greiner D, Melton DA. 2014. Generation of functional human pancreatic beta cells in vitro. *Cell* **159**: 428-439.

Pal A, Farmer AJ, Dudley C, Selwood MP, Barrow BA, Klyne R, Grew JP, McCarthy MI, Gloyn AL, Owen KR. 2010. Evaluation of serum 1,5 anhydroglucitol levels as a clinical test to differentiate subtypes of diabetes. *Diabetes Care* **33**: 252-257.

Parrizas M, Maestro MA, Boj SF, Paniagua A, Casamitjana R, Gomis R, Rivera F, Ferrer J. 2001. Hepatic nuclear factor 1-alpha directs nucleosomal hyperacetylation to its tissue-specific transcriptional targets. *Mol Cell Biol* **21**: 3234-3243.

Patel KA, Kettunen J, Laakso M, Stancakova A, Laver TW, Colclough K, Johnson MB, Abramowicz M, Groop L, Miettinen PJ et al. 2017. Heterozygous RFX6 protein truncating variants are associated with MODY with reduced penetrance. *Nat Commun* **8**: 888.

Patel KA, Oram RA, Flanagan SE, De Franco E, Colclough K, Shepherd M, Ellard S, Weedon MN, Hattersley AT. 2016. Type 1 Diabetes Genetic Risk Score: A Novel Tool to Discriminate Monogenic and Type 1 Diabetes. *Diabetes* **65**: 2094-2099.

Pearson ER, Boj SF, Steele AM, Barrett T, Stals K, Shield JP, Ellard S, Ferrer J, Hattersley AT. 2007. Macrosomia and hyperinsulinaemic hypoglycaemia in patients with heterozygous mutations in the HNF4A gene. *PLoS Med* **4**: e118.

Pearson ER, Liddell WG, Shepherd M, Corrall RJ, Hattersley AT. 2000. Sensitivity to sulphonylureas in patients with hepatocyte nuclear factor-1alpha gene mutations: evidence for pharmacogenetics in diabetes. *Diabet Med* **17**: 543-545.

Pearson ER, Pruhova S, Tack CJ, Johansen A, Castleden HA, Lumb PJ, Wierzbicki AS, Clark PM, Lebl J, Pedersen O et al. 2005. Molecular genetics and phenotypic characteristics of MODY caused by hepatocyte nuclear factor 4alpha mutations in a large European collection. *Diabetologia* **48**: 878-885.

Pearson ER, Starkey BJ, Powell RJ, Gribble FM, Clark PM, Hattersley AT. 2003. Genetic cause of hyperglycaemia and response to treatment in diabetes. *Lancet* **362**: 1275-1281.

Pearson ER, Velho G, Clark P, Stride A, Shepherd M, Frayling TM, Bulman MP, Ellard S, Froguel P, Hattersley AT. 2001. beta-cell genes and diabetes: quantitative and qualitative differences in the pathophysiology of hepatic nuclear factor-1alpha and glucokinase mutations. *Diabetes* **50 Suppl 1**: S101-107.

Petrovski S, Wang Q, Heinzen EL, Allen AS, Goldstein DB. 2013. Genic intolerance to functional variation and the interpretation of personal genomes. *PLoS Genet* **9**: e1003709.

Pihoker C, Gilliam LK, Ellard S, Dabelea D, Davis C, Dolan LM, Greenbaum CJ, Imperatore G, Lawrence JM, Marcovina SM et al. 2013. Prevalence, characteristics and clinical diagnosis of maturity onset diabetes of the young due to mutations in HNF1A, HNF4A, and glucokinase: results from the SEARCH for Diabetes in Youth. *J Clin Endocrinol Metab* **98**: 4055-4062.

Plenge RM, Scolnick EM, Altshuler D. 2013. Validating therapeutic targets through human genetics. *Nat Rev Drug Discov* **12**: 581-594.

Polak M, Dechaume A, Cave H, Nimri R, Crosnier H, Sulmont V, de Kerdanet M, Scharfmann R, Lebenthal Y, Froguel P et al. 2008. Heterozygous missense mutations in the insulin gene are linked to permanent diabetes appearing in the neonatal period or in early infancy: a report from the French ND (Neonatal Diabetes) Study Group. *Diabetes* **57**: 1115-1119.

Ponting CP, Hardison RC. 2011. What fraction of the human genome is functional? *Genome Res* **21**: 1769-1776.

Pontoglio M, Sreenan S, Roe M, Pugh W, Ostrega D, Doyen A, Pick AJ, Baldwin A, Velho G, Froguel P et al. 1998. Defective insulin secretion in hepatocyte nuclear factor 1alpha-deficient mice. *J Clin Invest* **101**: 2215-2222.

Postic C, Shiota M, Niswender KD, Jetton TL, Chen Y, Moates JM, Shelton KD, Lindner J, Cherrington AD, Magnuson MA. 1999. Dual roles for glucokinase in glucose homeostasis as determined by liver and pancreatic beta cell-specific gene knock-outs using Cre recombinase. *J Biol Chem* **274**: 305-315.

Pouwer F, Hermanns N. 2009. Insulin therapy and quality of life. A review. *Diabetes Metab Res Rev* **25 Suppl 1**: S4-S10.

Prisco F, Iafusco D, Franzese A, Sulli N, Barbetti F. 2000. MODY 2 presenting as neonatal hyperglycaemia: a need to reshape the definition of "neonatal diabetes"? *Diabetologia* **43**: 1331-1332.

Pruhova S, Ek J, Lebl J, Sumnik Z, Saudek F, Andel M, Pedersen O, Hansen T. 2003. Genetic epidemiology of MODY in the Czech republic: new mutations in the MODY genes HNF-4alpha, GCK and HNF-1alpha. *Diabetologia* **46**: 291-295.

Raeder H, Johansson S, Holm PI, Haldorsen IS, Mas E, Sbarra V, Nermoen I, Eide SA, Grevle L, Bjorkhaug L et al. 2006. Mutations in the CEL VNTR cause a syndrome of diabetes and pancreatic exocrine dysfunction. *Nat Genet* **38**: 54-62.

Raeder H, McAllister FE, Tjora E, Bhatt S, Haldorsen I, Hu J, Willems SM, Vesterhus M, El Ouaamari A, Liu M et al. 2014. Carboxyl-ester lipase maturity-onset diabetes of the young is associated with development of pancreatic cysts and upregulated MAPK signaling in secretin-stimulated duodenal fluid. *Diabetes* **63**: 259-269.

Raghavan S, Porneala B, McKeown N, Fox CS, Dupuis J, Meigs JB. 2015. Metabolic factors and genetic risk mediate familial type 2 diabetes risk in the Framingham Heart Study. *Diabetologia* **58**: 988-996.

Raimondo A, Chakera AJ, Thomsen SK, Colclough K, Barrett A, De Franco E, Chatelas A, Demirbilek H, Akcay T, Alawneh H et al. 2014. Phenotypic severity of homozygous GCK mutations causing neonatal or childhood-onset diabetes is primarily mediated through effects on protein stability. *Hum Mol Genet* **23**: 6432-6440.

Rands CM, Meader S, Ponting CP, Lunter G. 2014. 8.2% of the Human genome is constrained: variation in rates of turnover across functional element classes in the human lineage. *PLoS Genet* **10**: e1004525.

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

Rees MG, Raimondo A, Wang J, Ban MR, Davis MI, Barrett A, Ranft J, Jagdhuhn D, Waterstradt R, Baltrusch S et al. 2014. Inheritance of rare functional GCKR

variants and their contribution to triglyceride levels in families. *Hum Mol Genet* **23**: 5570-5578.

Rehm HL, Berg JS, Brooks LD, Bustamante CD, Evans JP, Landrum MJ, Ledbetter DH, Maglott DR, Martin CL, Nussbaum RL et al. 2015. ClinGen--the Clinical Genome Resource. *N Engl J Med* **372**: 2235-2242.

Rezania A, Bruin JE, Arora P, Rubin A, Batushansky I, Asadi A, O'Dwyer S, Quiskamp N, Mojibian M, Albrecht T et al. 2014. Reversal of diabetes with insulin-producing cells derived in vitro from human pluripotent stem cells. *Nat Biotechnol* **32**: 1121-1133.

Richards S, Aziz N, Bale S, Bick D, Das S, Gastier-Foster J, Grody WW, Hegde M, Lyon E, Spector E et al. 2015. Standards and guidelines for the interpretation of sequence variants: a joint consensus recommendation of the American College of Medical Genetics and Genomics and the Association for Molecular Pathology. *Genet Med* **17**: 405-424.

Robinson MD, McCarthy DJ, Smyth GK. 2010. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* **26**: 139-140.

Rosenberg AB, Patwardhan RP, Shendure J, Seelig G. 2015. Learning the sequence determinants of alternative splicing from millions of random sequences. *Cell* **163**: 698-711.

Rubio-Cabezas O, Diaz Gonzalez F, Aragones A, Argente J, Campos-Barros A. 2008. Permanent neonatal diabetes caused by a homozygous nonsense mutation in the glucokinase gene. *Pediatr Diabetes* **9**: 245-249.

Rubio-Cabezas O, Minton JA, Kantor I, Williams D, Ellard S, Hattersley AT. 2010. Homozygous mutations in NEUROD1 are responsible for a novel syndrome of permanent neonatal diabetes and neurological abnormalities. *Diabetes* **59**: 2326-2331.

Sagen JV, Bjorkhaug L, Molnes J, Raeder H, Grevle L, Sovik O, Molven A, Njolstad PR. 2008. Diagnostic screening of MODY2/GCK mutations in the Norwegian MODY Registry. *Pediatr Diabetes* **9**: 442-449.

Sagen JV, Raeder H, Hathout E, Shehadeh N, Gudmundsson K, Baevre H, Abuelo D, Phornphutkul C, Molnes J, Bell GI et al. 2004. Permanent neonatal diabetes due to mutations in KCNJ11 encoding Kir6.2: patient characteristics and initial response to sulfonylurea therapy. *Diabetes* **53**: 2713-2718.

Sakabe NJ, Savic D, Nobrega MA. 2012. Transcriptional enhancers in development and disease. *Genome Biol* **13**: 238.

Sakaguchi T, Gu X, Golden HM, Suh E, Rhoads DB, Reinecker HC. 2002. Cloning of the human claudin-2 5'-flanking region revealed a TATA-less promoter with conserved binding sites in mouse and human for caudal-related homeodomain proteins and hepatocyte nuclear factor-1alpha. *J Biol Chem* **277**: 21361-21370.

Samocha KE, Kosmicki JA, Karczewski KJ, O'Donnell-Luria AH, Pierce-Hoffman E, MacArthur DG, Neale BM, Daly MJ. 2017. Regional missense constraint improves variant deleteriousness prediction. *bioRxiv* doi:10.1101/148353.

Samocha KE, Robinson EB, Sanders SJ, Stevens C, Sabo A, McGrath LM, Kosmicki JA, Rehnstrom K, Mallick S, Kirby A et al. 2014. A framework for the interpretation of de novo mutation in human disease. *Nat Genet* **46**: 944-950.

Scharfmann R, Pechberty S, Hazhouz Y, von Bulow M, Bricout-Neveu E, Grenier-Godard M, Guez F, Rachdi L, Lohmann M, Czernichow P et al. 2014. Development of a conditionally immortalized human pancreatic beta cell line. *J Clin Invest* **124**: 2087-2098.

Schober E, Rami B, Grabert M, Thon A, Kapellen T, Reinehr T, Holl RW, and DP-WIotGWGfPD. 2009. Phenotypical aspects of maturity-onset diabetes of the young (MODY diabetes) in comparison with Type 2 diabetes mellitus (T2DM) in children and adolescents: experience from a large multicentre database. *Diabet Med* **26**: 466-473.

Schwarz JM, Cooper DN, Schuelke M, Seelow D. 2014. MutationTaster2: mutation prediction for the deep-sequencing age. *Nat Methods* **11**: 361-362.

Seetharam B, Christensen EI, Moestrup SK, Hammond TG, Verroust PJ. 1997. Identification of rat yolk sac target protein of teratogenic antibodies, gp280, as intrinsic factor-cobalamin receptor. *J Clin Invest* **99**: 2317-2322.

Servitja JM, Pignatelli M, Maestro MA, Cardalda C, Boj SF, Lozano J, Blanco E, Lafuente A, McCarthy MI, Sumoy L 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**: 2945-2959.

Shalem O, Sanjana NE, Zhang F. 2015a. High-throughput functional genomics using CRISPR-Cas9. *Nat Rev Genet* **16**: 299-311.

Shalem O, Sharon E, Lubliner S, Regev I, Lotan-Pompan M, Yakhini Z, Segal E. 2015b. Systematic dissection of the sequence determinants of gene 3' end mediated expression control. *PLoS Genet* **11**: e1005147.

Shang L, Hua H, Foo K, Martinez H, Watanabe K, Zimmer M, Kahler DJ, Freeby M, Chung W, LeDuc C et al. 2014. beta-cell dysfunction due to increased ER stress in a stem cell model of Wolfram syndrome. *Diabetes* **63**: 923-933.

Shemer R, Avnon Ziv C, Laiba E, Zhou Q, Gay J, Tunovsky-Babaey S, Shyng SL, Glaser B, Zangen DH. 2012. Relative expression of a dominant mutated ABCC8 allele determines the clinical manifestation of congenital hyperinsulinism. *Diabetes* **61**: 258-263.

Shepherd M, Ellis I, Ahmad AM, Todd PJ, Bowen-Jones D, Mannion G, Ellard S, Sparkes AC, Hattersley AT. 2001. Predictive genetic testing in maturity-onset diabetes of the young (MODY). *Diabet Med* **18**: 417-421.

Shepherd M, Shields B, Ellard S, Rubio-Cabezas O, Hattersley AT. 2009. A genetic diagnosis of HNF1A diabetes alters treatment and improves glycaemic control in the majority of insulin-treated patients. *Diabet Med* **26**: 437-441.

Shields B, Colclough K. 2017. Towards a systematic nationwide screening strategy for MODY. *Diabetologia* **60**: 609-612.

Shields BM, Hicks S, Shepherd MH, Colclough K, Hattersley AT, Ellard S. 2010. Maturity-onset diabetes of the young (MODY): how many cases are we missing? *Diabetologia* **53**: 2504-2508.

Shields BM, Shepherd M, Hudson M, McDonald TJ, Colclough K, Peters J, Knight B, Hyde C, Ellard S, Pearson ER et al. 2017. Population-Based Assessment of a Biomarker-Based Screening Pathway to Aid Diagnosis of Monogenic Diabetes in Young-Onset Patients. *Diabetes Care* **40**: 1017-1025.

Singh W. 2008. Robustness of three hierarchical agglomerative clustering techniques for ecological data [M.S. thesis] Reykjavik, Iceland University of Iceland.

Sosnay PR, Siklosi KR, Van Goor F, Kaniecki K, Yu H, Sharma N, Ramalho AS, Amaral MD, Dorfman R, Zielenski J et al. 2013. Defining the disease liability of variants in the cystic fibrosis transmembrane conductance regulator gene. *Nat Genet* **45**: 1160-1167.

Spyer G, Macleod KM, Shepherd M, Ellard S, Hattersley AT. 2009. Pregnancy outcome in patients with raised blood glucose due to a heterozygous glucokinase gene mutation. *Diabet Med* **26**: 14-18.

Staiger H, Stancakova A, Zilinskaite J, Vanttinen M, Hansen T, Marini MA, Hammarstedt A, Jansson PA, Sesti G, Smith U et al. 2008. A candidate type 2 diabetes polymorphism near the HHEX locus affects acute glucose-stimulated insulin release in European populations: results from the EUGENE2 study. *Diabetes* **57**: 514-517.

Stanik J, Dusatkova P, Cinek O, Valentinova L, Huckova M, Skopkova M, Dusatkova L, Stanikova D, Pura M, Klimes I et al. 2014. De novo mutations of GCK, HNF1A and HNF4A may be more frequent in MODY than previously assumed. *Diabetologia* **57**: 480-484.

Starita LM, Ahituv N, Dunham MJ, Kitzman JO, Roth FP, Seelig G, Shendure J, Fowler DM. 2017. Variant Interpretation: Functional Assays to the Rescue. *Am J Hum Genet* **101**: 315-325.

Starita LM, Young DL, Islam M, Kitzman JO, Gullingsrud J, Hause RJ, Fowler DM, Parvin JD, Shendure J, Fields S. 2015. Massively Parallel Functional Analysis of BRCA1 RING Domain Variants. *Genetics* **200**: 413-422.

Steele AM, Shields BM, Wensley KJ, Colclough K, Ellard S, Hattersley AT. 2014. Prevalence of vascular complications among patients with glucokinase mutations and prolonged, mild hyperglycemia. *JAMA* **311**: 279-286.

Steele AM, Tribble ND, Caswell R, Wensley KJ, Hattersley AT, Gloyn AL, Ellard S. 2011. The previously reported T342P GCK missense variant is not a pathogenic mutation causing MODY. *Diabetologia* **54**: 2202-2205.

Steele AM, Wensley KJ, Ellard S, Murphy R, Shepherd M, Colclough K, Hattersley AT, Shields BM. 2013. Use of HbA1c in the identification of patients with hyperglycaemia caused by a glucokinase mutation: observational case control studies. *PLoS One* **8**: e65326.

Stenson PD, Ball EV, Howells K, Phillips AD, Mort M, Cooper DN. 2009. The Human Gene Mutation Database: providing a comprehensive central mutation database for molecular diagnostics and personalized genomics. *Hum Genomics* **4**: 69-72.

Stoffers DA, Ferrer J, Clarke WL, Habener JF. 1997. Early-onset type-II diabetes mellitus (MODY4) linked to IPF1. *Nat Genet* **17**: 138-139.

Stride A, Hattersley AT. 2002. Different genes, different diabetes: lessons from maturity-onset diabetes of the young. *Ann Med* **34**: 207-216.

Stride A, Shields B, Gill-Carey O, Chakera AJ, Colclough K, Ellard S, Hattersley AT. 2014. Cross-sectional and longitudinal studies suggest pharmacological treatment used in patients with glucokinase mutations does not alter glycaemia. *Diabetologia* **57**: 54-56.

Swisa A, Avrahami D, Eden N, Zhang J, Feleke E, Dahan T, Cohen-Tayar Y, Stolovich-Rain M, Kaestner KH, Glaser B et al. 2017. PAX6 maintains beta cell identity by repressing genes of alternative islet cell types. *J Clin Invest* **127**: 230-243.

Szklarczyk D, Morris JH, Cook H, Kuhn M, Wyder S, Simonovic M, Santos A, Doncheva NT, Roth A, Bork P et al. 2017. The STRING database in 2017: quality-controlled protein-protein association networks, made broadly accessible. *Nucleic Acids Res* **45**: D362-D368.

Szlyk B, Braun CR, Ljubicic S, Patton E, Bird GH, Osundiji MA, Matschinsky FM, Walensky LD, Danial NN. 2014. A phospho-BAD BH3 helix activates glucokinase by a mechanism distinct from that of allosteric activators. *Nat Struct Mol Biol* **21**: 36-42.

Tattersall RB. 1974. Mild familial diabetes with dominant inheritance. *Q J Med* **43**: 339-357.

Tattersall RB, Fajans SS. 1975. A difference between the inheritance of classical juvenile-onset and maturity-onset type diabetes of young people. *Diabetes* **24**: 44-53.

Teo AK, Windmueller R, Johansson BB, Dirice E, Njolstad PR, Tjora E, Raeder H, Kulkarni RN. 2013. Derivation of human induced pluripotent stem cells from patients with maturity onset diabetes of the young. *J Biol Chem* **288**: 5353-5356.

Terauchi Y, Sakura H, Yasuda K, Iwamoto K, Takahashi N, Ito K, Kasai H, Suzuki H, Ueda O, Kamada N et al. 1995. Pancreatic beta-cell-specific targeted disruption of glucokinase gene. Diabetes mellitus due to defective insulin secretion to glucose. *J Biol Chem* **270**: 30253-30256.

Tewhey R, Kotliar D, Park DS, Liu B, Winnicki S, Reilly SK, Andersen KG, Mikkelsen TS, Lander ES, Schaffner SF et al. 2016. Direct Identification of Hundreds of Expression-Modulating Variants using a Multiplexed Reporter Assay. *Cell* **165**: 1519-1529.

Thanabalasingham G, Huffman JE, Kattla JJ, Novokmet M, Rudan I, Gloyn AL, Hayward C, Adamczyk B, Reynolds RM, Muzinic A et al. 2013. Mutations in HNF1A result in marked alterations of plasma glycan profile. *Diabetes* **62**: 1329-1337.

Thanabalasingham G, Pal A, Selwood MP, Dudley C, Fisher K, Bingley PJ, Ellard S, Farmer AJ, McCarthy MI, Owen KR. 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**: 1206-1212.

Thanabalasingham G, Shah N, Vaxillaire M, Hansen T, Tuomi T, Gasperikova D, Szopa M, Tjora E, James TJ, Kokko P 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**: 2801-2810.

Thomas P, Ye Y, Lightner E. 1996. Mutation of the pancreatic islet inward rectifier Kir6.2 also leads to familial persistent hyperinsulinemic hypoglycemia of infancy. *Hum Mol Genet* **5**: 1809-1812.

Thompson BA, Spurdle AB, Plazzer JP, Greenblatt MS, Akagi K, Al-Mulla F, Bapat B, Bernstein I, Capella G, den Dunnen JT et al. 2014. Application of a 5-tiered scheme for standardized classification of 2,360 unique mismatch repair gene variants in the InSiGHT locus-specific database. *Nat Genet* **46**: 107-115.

Thormaehlen AS, Schuberth C, Won HH, Blattmann P, Joggerst-Thomalla B, Theiss S, Asselta R, Duga S, Merlini PA, Ardissino D 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**: e1004855.

Tibshirani, Walther, and Hastie. 2002. Estimating the number of clusters in a data set via the gap statistic. *Journal of the Royal Statistical Society* B **63**, 411-423.

Tonooka N, Tomura H, Takahashi Y, Onigata K, Kikuchi N, Horikawa Y, Mori M, Takeda J. 2002. High frequency of mutations in the HNF-1alpha gene in non-obese patients with diabetes of youth in Japanese and identification of a case of digenic inheritance. *Diabetologia* **45**: 1709-1712.

Torsvik J, Johansson S, Johansen A, Ek J, Minton J, Raeder H, Ellard S, Hattersley A, Pedersen O, Hansen T et al. 2010. Mutations in the VNTR of the carboxyl-ester lipase gene (CEL) are a rare cause of monogenic diabetes. *Hum Genet* **127**: 55-64.

Triggs-Raine BL, Kirkpatrick RD, Kelly SL, Norquay LD, Cattini PA, Yamagata K, Hanley AJ, Zinman B, Harris SB, Barrett PH 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**: 4614-4619.

Turkkahraman D, Bircan I, Tribble ND, Akcurin S, Ellard S, Gloyn AL. 2008. Permanent neonatal diabetes mellitus caused by a novel homozygous (T168A) glucokinase (GCK) mutation: initial response to oral sulphonylurea therapy. *J Pediatr* **153**: 122-126.

Ulirsch JC, Nandakumar SK, Wang L, Giani FC, Zhang X, Rogov P, Melnikov A, McDonel P, Do R, Mikkelsen TS et al. 2016. Systematic Functional Dissection of Common Genetic Variation Affecting Red Blood Cell Traits. *Cell* **165**: 1530-1545.

Urano F. 2014. Wolfram syndrome iPS cells: the first human cell model of endoplasmic reticulum disease. *Diabetes* **63**: 844-846.

van Vliet-Ostaptchouk JV, Onland-Moret NC, van Haeften TW, Franke L, Elbers CC, Shiri-Sverdlov R, van der Schouw YT, Hofker MH, Wijmenga C. 2008. HHEX gene polymorphisms are associated with type 2 diabetes in the Dutch Breda cohort. *Eur J Hum Genet* **16**: 652-656.

VanderMeer JE, Ahituv N. 2011. cis-regulatory mutations are a genetic cause of human limb malformations. *Dev Dyn* **240**: 920-930.

Vaxillaire M, Abderrahmani A, Boutin P, Bailleul B, Froguel P, Yaniv M, Pontoglio M. 1999. Anatomy of a homeoprotein revealed by the analysis of human MODY3 mutations. *J Biol Chem* **274**: 35639-35646.

Vaxillaire M, Boccio V, Philippi A, Vigouroux C, Terwilliger J, Passa P, Beckmann JS, Velho G, Lathrop GM, Froguel P. 1995. A gene for maturity onset diabetes of the young (MODY) maps to chromosome 12q. *Nat Genet* **9**: 418-423.

Vaxillaire M, Froguel P. 2008. Monogenic diabetes in the young, pharmacogenetics and relevance to multifactorial forms of type 2 diabetes. *Endocr Rev* **29**: 254-264.

Vaxillaire M, Populaire C, Busiah K, Cave H, Gloyn AL, Hattersley AT, Czernichow P, Froguel P, Polak M. 2004. Kir6.2 mutations are a common cause of permanent neonatal diabetes in a large cohort of French patients. *Diabetes* **53**: 2719-2722.

Velho G, Blanche H, Vaxillaire M, Bellanne-Chantelot C, Pardini VC, Timsit J, Passa P, Deschamps I, Robert JJ, Weber IT et al. 1997. Identification of 14 new glucokinase mutations and description of the clinical profile of 42 MODY-2 families. *Diabetologia* **40**: 217-224.

Velho G, Froguel P, Clement K, Pueyo ME, Rakotoambinina B, Zouali H, Passa P, Cohen D, Robert JJ. 1992. Primary pancreatic beta-cell secretory defect caused by mutations in glucokinase gene in kindreds of maturity onset diabetes of the young. *Lancet* **340**: 444-448.

Venables WN, Ripley BD. 2002. Modern Applied Statistics with S, Springer-Verlag.

Vesterhus M, Haldorsen IS, Raeder H, Molven A, Njolstad PR. 2008. Reduced pancreatic volume in hepatocyte nuclear factor 1A-maturity-onset diabetes of the young. *J Clin Endocrinol Metab* **93**: 3505-3509.

Voight BF Scott LJ Steinthorsdottir V Morris AP Dina C Welch RP Zeggini E Huth C Aulchenko YS Thorleifsson G et al. 2010. Twelve type 2 diabetes susceptibility loci identified through large-scale association analysis. *Nat Genet* **42**: 579-589.

Wang L, Lovejoy NF, Faustman DL. 2012. Persistence of prolonged C-peptide production in type 1 diabetes as measured with an ultrasensitive C-peptide assay. *Diabetes Care* **35**: 465-470.

Ward JH. 1963. Hierarchical groupings to optimize an objective function. Journal of the American Statistical Association, **58**, 234–24.

Weedon MN, Cebola I, Patch AM, Flanagan SE, De Franco E, Caswell R, Rodriguez-Segui SA, Shaw-Smith C, Cho CH, Allen HL et al. 2014. Recessive mutations in a distal PTF1A enhancer cause isolated pancreatic agenesis. *Nat Genet* **46**: 61-64.

Weedon MN, Clark VJ, Qian Y, Ben-Shlomo Y, Timpson N, Ebrahim S, Lawlor DA, Pembrey ME, Ring S, Wilkin TJ et al. 2006. A common haplotype of the glucokinase gene alters fasting glucose and birth weight: association in six studies and population-genetics analyses. *Am J Hum Genet* **79**: 991-1001.

Weedon MN, Owen KR, Shields B, Hitman G, Walker M, McCarthy MI, Hattersley AT, Frayling TM. 2005. A large-scale association analysis of common variation of the HNF1alpha gene with type 2 diabetes in the U.K. Caucasian population. *Diabetes* **54**: 2487-2491.

Weng J, Ekelund M, Lehto M, Li H, Ekberg G, Frid A, Aberg A, Groop LC, Berntorp K. 2002. Screening for MODY mutations, GAD antibodies, and type 1 diabetes--associated HLA genotypes in women with gestational diabetes mellitus. *Diabetes Care* **25**: 68-71.

Winckler W, Weedon MN, Graham RR, McCarroll SA, Purcell S, Almgren P, Tuomi T, Gaudet D, Bostrom KB, Walker M 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**: 685-693.

Xia W, Bringmann P, McClary J, Jones PP, Manzana W, Zhu Y, Wang S, Liu Y, Harvey S, Madlansacay MR et al. 2006. High levels of protein expression using different mammalian CMV promoters in several cell lines. *Protein Expr Purif* **45**: 115-124.

Xu JY, Chan V, Zhang WY, Wat NM, Lam KS. 2002. Mutations in the hepatocyte nuclear factor-1alpha gene in Chinese MODY families: prevalence and functional analysis. *Diabetologia* **45**: 744-746.

Xu JY, Dan QH, Chan V, Wat NM, Tam S, Tiu SC, Lee KF, Siu SC, Tsang MW, Fung LM et al. 2005. Genetic and clinical characteristics of maturity-onset diabetes of the young in Chinese patients. *Eur J Hum Genet* **13**: 422-427.

Xue Y, Chen Y, Ayub Q, Huang N, Ball EV, Mort M, Phillips AD, Shaw K, Stenson PD, Cooper DN et al. 2012. Deleterious- and disease-allele prevalence in healthy individuals: insights from current predictions, mutation databases, and population-scale resequencing. *Am J Hum Genet* **91**: 1022-1032.

Yamagata K, Furuta H, Oda N, Kaisaki PJ, Menzel S, Cox NJ, Fajans SS, Signorini S, Stoffel M, Bell GI. 1996a. Mutations in the hepatocyte nuclear factor-4alpha gene in maturity-onset diabetes of the young (MODY1). *Nature* **384**: 458-460.

Yamagata K, Oda N, Kaisaki PJ, Menzel S, Furuta H, Vaxillaire M, Southam L, Cox RD, Lathrop GM, Boriraj VV et al. 1996b. Mutations in the hepatocyte nuclear factor-1alpha gene in maturity-onset diabetes of the young (MODY3). *Nature* **384**: 455-458.

Yorifuji T, Fujimaru R, Hosokawa Y, Tamagawa N, Shiozaki M, Aizu K, Jinno K, Maruo Y, Nagasaka H, Tajima T et al. 2012. Comprehensive molecular analysis of Japanese patients with pediatric-onset MODY-type diabetes mellitus. *Pediatr Diabetes* **13**: 26-32.

Zeggini E, Weedon MN, Lindgren CM, Frayling TM, Elliott KS, Lango H, Timpson NJ, Perry JR, Rayner NW, Freathy RM et al. 2007. Replication of genome-wide association signals in UK samples reveals risk loci for type 2 diabetes. *Science* **316**: 1336-1341.

Zhang F, Lupski JR. 2015. Non-coding genetic variants in human disease. *Hum Mol Genet* **24**: R102-110.

Zhang F, Xu X, Zhang Y, Zhou B, He Z, Zhai Q. 2013. Gene expression profile analysis of type 2 diabetic mouse liver. *PLoS One* **8**: e57766.

Zhao W, Pollack JL, Blagev DP, Zaitlen N, McManus MT, Erle DJ. 2014. Massively parallel functional annotation of 3' untranslated regions. *Nat Biotechnol* **32**: 387-391.

Zucman-Rossi J, Jeannot E, Nhieu JT, Scoazec JY, Guettier C, Rebouissou S, Bacq Y, Leteurtre E, Paradis V, Michalak S et al. 2006. Genotype-phenotype correlation in hepatocellular adenoma: new classification and relationship with HCC. *Hepatology* **43**: 515-524.

Zuk O, Schaffner SF, Samocha K, Do R, Hechter E, Kathiresan S, Daly MJ, Neale BM, Sunyaev SR, Lander ES. 2014. Searching for missing heritability: designing rare variant association studies. *Proc Natl Acad Sci U S A* **111**: E455-464.