

Role of rare *HNF1A* variants function in monogenic and type 2 diabetes

Laeya Najmi

Thesis for the Degree of Philosophiae Doctor (PhD)
University of Bergen, Norway
2018

UNIVERSITY OF BERGEN



Role of rare HNF1A variants function in monogenic and type 2 diabetes

Laeya Najmi



Thesis for the Degree of Philosophiae Doctor (PhD)
at the University of Bergen

2018

Date of defence: 26.10.2018

© Copyright Laeya Najmi

The material in this publication is covered by the provisions of the Copyright Act.

Year: 2018

Title: Role of rare HNF1A variants function in monogenic and type 2 diabetes

Name: Laeya Najmi

Print: Skipnes Kommunikasjon / University of Bergen

Scientific environment

This work was carried out at:

KG Jebsen Center for Diabetes Research
Department for Clinical Science
University of Bergen
Bergen, Norway

Department of Medical Genetics
Haukeland University Hospital
Bergen, Norway

Medical and Population Genetic Research
Broad Institute of MIT and Harvard
Cambridge, MA, USA

This work is dedicated to my family. They instilled in me everything it took to get here.

AKNOWLEDGMENT

It's been a long road, but here I am at the end, but there are so many wonderful people to whom thanks I extend!

First and foremost, I am profoundly indebted to Prof. Lise Bjørkhaug, my main supervisor, for her fundamental role in my doctoral work. There are no words commensurate with my gratitude for all her teaching, support and inspiration during the several years of my PhD program. Lise sustained me by not only providing consistently wise and creative academic mentorship over almost six years, but also with generosity and compassion during the rough road to finish this thesis. Her warmth and affability have always made me feel at ease and I will forever cherish what I learned from her about how to be a scientist and how to be a person. I am sincerely grateful to Prof. Pål R. Njølstad, my co-supervisor, for his continuous support, guidance, input, leadership and encouragement. He provided me the opportunity to enhance this research abroad, which was extremely powerful for the full realization of the work. I owe a great debt to Ingvild Aukrust, my co-supervisor, with her expansive knowledge, thoughtful guidance, generous availability and constructive suggestions were invaluable to the completion of the work presented in this thesis. I am grateful to Prof. Anders Molven, my co-supervisor, he provided great input and ideas throughout my course of study and was instrumental in helping me develop ways to effectively write about and communicate my research.

I am deeply in debt to Prof. Jose Florez for being an incredible leader. He is the kind of leader groups dream of, and his extreme competence and clarity of vision naturally inspire everyone around him. Thank you for welcoming me into your lab at the Broad Institute of Harvard and MIT. I am very grateful to Amit R. Majithia for his scientific input and guidance. His impact is woven into these pages and it would not have been the same without him.

I would like to acknowledge all collaborators and co-authors for their essential contributions, especially Prof. Stefan Johansson, Janne Molnes and Haydee Artaza. I am thankful to all the colleagues in the MODY group and at the Medical Genetics Center in Haukeland hospital. I appreciate all of your help and support. To my friends and

colleagues at the Broad Institute, I appreciate you all so much for your friendship and support, and for creating a comfortable working environment. Especially Li Wang, Xiaolan Zhang, and Robert Rice, it was so much fun to work with you and I'll always remember the times we had.

Special words of gratitude go to my friends who have always been there for me. Special one goes to Zahra Khorsand, I am so blessed to have you in my life.

I owe my deepest gratitude to my wonderful parents, brothers and sisters. Their eternal support and understanding for my goals and aspirations, their infallible love and endless encouragement has always been my strength. Their patience and sacrifice will remain my greatest inspiration throughout my life. I'm proud of having you and I would not be here without you.

I gratefully acknowledge the University of Bergen, the Norwegian Society of Endocrinology, the European Society for Pediatric Endocrinology, Det Alminnelig Medisinske Forskningsfond and Diabetesforbundet for providing me the financial support required to carry out my research.

Sincerely, Laeya Najmi
Bergen, 2018

ABSTRACT

Variants in the transcription factor gene *HNF1A* have been identified in subjects with maturity-onset diabetes of the young (MODY) type 3, type 2 diabetes, as well as in children with apparent type 1 diabetes. One of the challenges in a clinical setting is distinguishing MODY patients from those with type 2 diabetes, as there is considerable overlap in terms of clinical features. Mapping *HNF1A* variants to the correct clinical phenotype requires functional characterization of variants effects on hepatocyte nuclear factor – 1A (HNF-1A) function.

Large whole-exome sequencing of an Mexican and American Latino population, reported in Paper I, identified a low frequency rare variant p.(E508K) in *HNF1A* that confers increased risk for type 2 diabetes up to 5 fold (odds ratio (OR)=5.48; $P=4.4 \times 10^{-7}$). Functional investigation of this p.(E508K) HNF-1A protein variant demonstrated reduced transactivation activity <50%, low protein level expression and slightly impaired nuclear localization. These findings suggest that the p.(E508K) *HNF1A* variant mediates a mild-loss of function of HNF-1A and represents a risk variant for type 2 diabetes in the Mexican and American Latino population.

Exome sequencing of *HNF1A* in 4,115 well-phenotyped individuals from the general population (Framingham Heart Study cohort, the Jackson Heart Study cohort, and type 2 diabetes case and control patients from the extreme type 2 diabetes cohort) have previously shown that 1/50 individuals with diabetes harbors a missense variant in the *HNF1A* gene. 27 rare *HNF1A* missense variants were identified. In Paper II we show that after using bioinformatics prediction tools to determine predicted pathogenic effect, none of the 27 *HNF1A* variants that were classified as pathogenic were associated with risk for type 2 diabetes in the population cohorts (OR=2.02; 95% CI 0.73-5.60; $P=0.18$). We further evaluated the functional consequences of individual variants by their effect on HNF-1A

transcriptional activation, DNA binding, and subcellular (cytoplasmic/nuclear) localization. Furthermore, association between type 2 diabetes and different functional assay models was assessed. A transcriptional activity with a threshold of <60% compared to wild-type HNF-1A activity was able to best predict type 2 diabetes association with carrier type 2 diabetes phenotype (OR=5.04; 95% ; $P=0.0007$), and indicate that 0.44% of the population carry HNF1A variants that results in substantially increased risk for developing the disease.

To improve the diagnostic interpretation of the increasing number of *HNF1A* variants identified by next-generation sequencing, there is a future demand for robust and reliable high-throughput functional investigations of variant effects on normal HNF-1A protein function. In Paper III, we searched systematically for endogenous regulated HNF-1A transcripts as possible markers for investigating diabetes associated *HNF1A* variants effects. For this purpose we generated *HNF1A*-free liver specific cell lines (HuH7 and HepB3) by knocking out endogenous *HNF1A* using CRISPR/Cas9, prior to the controlled re-expression (doxycycline induced) of wild-type HNF-1A versus HNF-1A variants (MODY3, type 2 diabetes). The gene expression profile analyzed by RNA sequencing identified significantly differentially expressed genes upon overexpressing HNF-1A, of which the top 20 upregulated genes were further investigated and many found down-regulated by MODY3-causing *HNF1A* variants. Of these, *ABCC2*, *FABP1* and *HABP2* genes in HuH7, and *HKDC1*, *HRG* and *KL* genes in Hep3B cell lines, were considered as potential targets for future large-scale and high-throughput investigations of numerous *HNF1A* variants.

Table of Contents

ABSTRACT	7
LIST OF PUBLICATIONS	11
ABBREVIATIONS.....	13
1. INTRODUCTION	14
1.1 Glucose homeostasis	14
1.2 Function of the pancreas	15
1.3 Beta-cell and glucose-stimulated insulin secretion	16
1.4 Impaired glucose homeostasis and diabetes diagnosis.....	18
1.5 Classification of diabetes	20
1.5.1 Type 1 diabetes mellitus	20
1.5.2 T1D and risk genes	20
1.5.3 Type 2 diabetes mellitus	21
1.5.4 T2D and risk genes	22
1.5.5 Gestational diabetes mellitus	23
1.5.6 Monogenic forms of diabetes and maturity-onset diabetes of the young	24
1.6 MODY genes and subtypes.....	25
1.6.1 HNF4A-MODY (MODY1).....	26
1.6.2 GCK-MODY (MODY2).....	27
1.6.3 HNF1A-MODY (MODY3).....	27
1.7 HNF-1A protein and function	29
1.7.1 <i>HNF1A</i> transcripts and protein isoforms	29
1.7.2 Subcellular localization of HNF-1A	30
1.7.3 DNA-binding of HNF-1A.....	31
1.7.4 Transcriptional activation by HNF-1A.....	32
1.7.5 HNF-1A and target gene regulation.....	33
1.7.6 HNF-1A and gene network regulation.....	34
1.8 Deciphering the genetic architecture of T2D.....	35
1.8.1 The candidate gene and linkage analysis approach	36
1.8.2 Genome-wide association studies approach	36
1.8.3 Rare variants and missing heritability	37
2. AIM OF STUDY	41
3. SUMMARY OF RESULT	42
3.1 Paper I:	42
3.2 Paper II:.....	43
3.3 Paper III:.....	44
4. DISCUSSION	46
4.1 Prevalence and effect of common and rare <i>HNF1A</i> variants in T2D diabetes populations	46

4.2	Prevalence and relevance of rare <i>HNF1A</i> variants as risk factor for T2D diabetes in the general population	48
4.3	<i>HNF1A</i> variant effect on nuclear localization and disease risk prediction	50
4.4	<i>HNF1A</i> variant effect on transcriptional activity and disease risk prediction ...	51
4.5	Genotype-phenotype correlation by <i>HNF1A</i> and treatment options	53
4.6	Systematic search for HNF-1A regulated transcripts for developing <i>HNF1A</i> high throughput assay.....	54
5.	CONCLUDING REMARKS AND FUTURE PERSPECTIVE	58
6.	REFERENCES.....	60

LIST OF PUBLICATIONS

Paper I The SIGMA Type 2 Diabetes Consortium. Association of a low-frequency variant in *HNFA1A* with type 2 diabetes in a Latino population. *JAMA*. 2014; 311(22): 2305-2314. Author list by *JAMA* teams: Functional Studies: Laeya A. Najmi., Ingvild Aukrust, Lise Bjørkhaug, Suzanne B. R. Jacobs, Pål R. Njølstad.

Paper II Najmi, L.A., Aukrust, I., Flannick, J., Molnes, J., Burt, N., Molven, A., Groop, L., Altshuler, D., Johansson, S., Bjørkhaug, L., Njølstad, P.R. Functional investigations of *HNFA1A* identify rare variants as risk factors for type 2 diabetes in the general population. *Diabetes*. 2016; 66 (2): 335-346.

Paper III Najmi, L.A., Amit R. Majithia, Haydee Artaza, Xiaolan Zhang, Ingvild Aukrust, Anders Molven, Stefan Johansson, Bjørkhaug, L., Njølstad, P.R. **Liver cell models for systematic and effective search for HNF-1A-regulated transcripts.**
Manuscript in preparation.

Related articles not included in the thesis

Lek M, Karczewski KJ, Minikel EV, Samocha KE, Banks E, Fennell T, O'Donnell-Luria AH, Ware JS, Hill AJ, Cummings BB, Tukiainen T, Birnbaum DP, Kosmicki JA, Duncan LE, Estrada K, Zhao F, Zou J, Pierce-Hoffman E, Berghout J, Cooper DN, Deflaux N, DePristo M, Do R, Flannick J, Fromer M, Gauthier L, Goldstein J, Gupta N, Howrigan D, Kiezun A, Kurki MI, Moonshine AL, Natarajan P, Orozco L, Peloso GM, Poplin R, Rivas MA, Ruano-Rubio V, Rose SA, Ruderfer DM, Shakir K, Stenson PD, Stevens C, Thomas BP, Tiao G, Tusie-Luna MT, Weisburd B, Won HH, Yu D, Altshuler DM, Ardissino D, Boehnke M, Danesh J, Donnelly S, Elosua R, Florez JC, Gabriel SB, Getz G, Glatt SJ, Hultman CM, Kathiresan S, Laakso M, McCarroll S, McCarthy MI, McGovern D, McPherson R, Neale BM, Palotie A, Purcell SM, Saleheen D, Scharf JM, Sklar P, Sullivan PF, Tuomilehto J, Tsuang MT, Watkins HC, Wilson JG, Daly MJ, MacArthur DG; **Exome Aggregation Consortium (including Najmi L.A.)**. Analysis of protein-coding genetic variation in 60,706 humans. *Nature*. (2016) Aug 18;536(7616):285-91. doi: 10.1038/nature19057.

Rozenkova K, Malikova J, Nessa A, Dusatkova L, Bjørkhaug L, Obermannova B, Dusatkova P, Kytarova J, Aukrust I, **Najmi L.A**, Rypackova B, Sumnik Z, Lebl J, Njølstad PR, Hussain K, Pruhova S.(2015).High incidence of heterozygous ABCC8 and HNF1A mutations in Czech Patients with congenital hyperinsulinism. *J Clin Endocrinol Metab.* 100:E1540-9. doi: 10.1210/jc.2015-2763.

ABBREVIATIONS

ATP	Adenosine triphosphate
CRISPR/Cas9	Clustered Regularly Interspaced Short Palindromic Repeats/CRISPR associated protein 9
ExAC	Exome Aggregation Consortium
FC	Fold change
FDR	False discovery rate
G6P	Glucose-6-phosphate
GCK	Glucokinase
GDM	Gestational diabetes mellitus
GLUT2	Glucose transporter 2
GnomAD	Genome Aggregation Database
GSIS	Glucose-stimulated insulin secretion
GTP	Guanosine triphosphate
GWAS	Genome wide association study
HGMD	Human Gene Mutation Database
HLA	Human leukocyte antigen
HNF	Hepatocyte nuclear factor
INS	Insulin
KO	Knock-out
MAF	Minor Allele Frequency
MAVE	Multiplexed assays for variants effect
MODY3	Maturity-onset diabetes of the young type 3
mRNA	messenger RNA
NL	Nuclear localization
NLS	Nuclear localization signal
PCR	Polymerase-chain-reaction
Ran	Ras-related nuclear protein
T1D	Type 1 diabetes
T2D	Type 2 diabetes
TA	Transactivation activity
VUS	Variants of uncertain significance
WHO	World health organization
WT	Wild-type

1. INTRODUCTION

1.1 Glucose homeostasis

Glucose is the source of immediate energy for cells within the body. The brain and red blood cells are solely dependent upon glucose. The physiological process that maintains blood glucose levels in a steady-state is called glucose homeostasis. An important contribution to the maintenance of glucose homeostasis is achieved through hormone regulation of peripheral glucose uptake, which occurs through carbohydrate ingestion and breakdown of endogenous glycogen stores within the liver [1]. High levels of glucose trigger insulin secretion from the pancreatic beta-cells. The major function of insulin is to accelerate glucose transport and uptake in muscle and adipose tissues by reducing hepatic glucose output through decreased gluconeogenesis and glycogenesis, as well as by reducing glucagon secretion to inhibit hyperglycemia.

Between meals, or during sleep, blood glucose levels are normally low. This triggers secretion of glucagon from the alpha-cells in the pancreas. Glucagon is the main hormone that increases and maintains glucose levels during periods of high demand, by increasing glucose output from the liver. It therefore has an opposite function compared to insulin function (**Figure 1**). Thus, to avoid hyperglycemia or hypoglycemia, the body maintains blood glucose levels by secretion of the two primary hormones; insulin and glucagon, that work in balanced opposition of each other [1, 2]. In addition to these hormones, several other hormones are secreted from the gastrointestinal tract during digestion and absorption of food. These gastrointestinal hormones, which are called incretin hormones, are secreted from endocrine cells located within the stomach, small intestine, and large intestine. Ingestion of food results in higher levels of insulin secretion, as compared to glucose infused intravenously, due to the effect of these incretin hormones.

Therefore, incretin hormones also have a role in the regulation of glucose homeostasis [3].

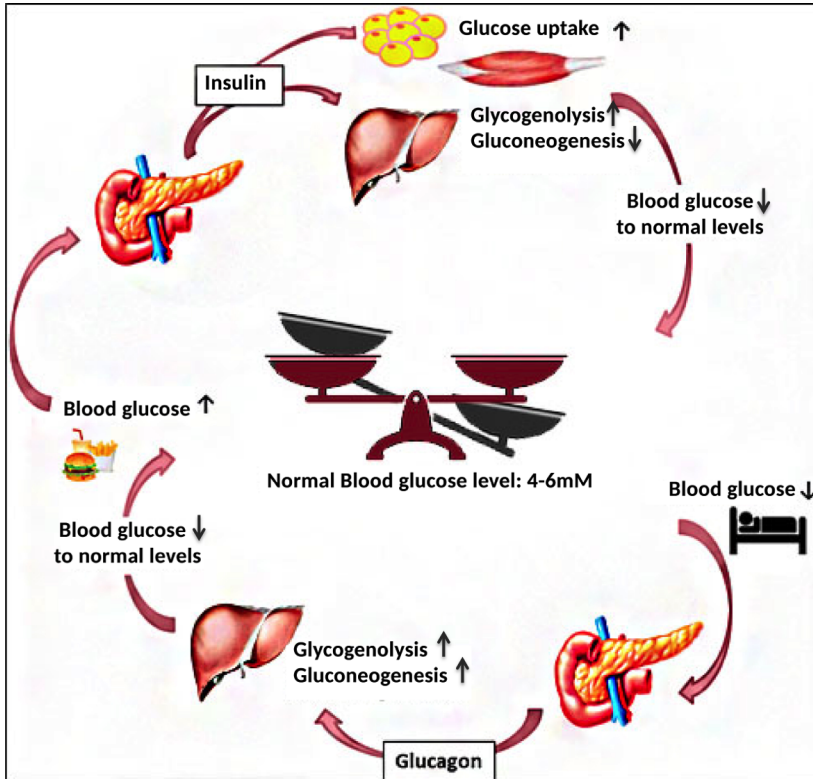


Figure 1. Regulation and maintenance of blood glucose levels by the hormones glucagon and insulin. After a meal, the pancreas secretes insulin as blood glucose levels tend to increase, and as a response, insulin increases glucose uptake in muscle and adipose tissues, accelerating glycogenesis. During the night, as blood levels tend to decrease, the pancreas secretes glucagon, which then increases blood glucose levels triggering glycogenolysis from the liver. Figure is adapted and modified from [3].

1.2 Function of the pancreas

The pancreas is part of the gastrointestinal system and plays a dual role in metabolic

homeostasis as well as in the digestion of macronutrients. The human pancreas is a hormone gland consisting of exocrine and endocrine tissue. The exocrine tissue accounts for >80% of the pancreatic volume with acinar cells and duct cells that secrete pancreatic juice, which contains digestive enzymes such as amylase, pancreatic lipase, and trypsinogen, into the pancreatic duct. The endocrine tissue, represented by less than <2% of total pancreatic volume, releases pancreatic hormones in an endocrine manner into the bloodstream. The endocrine cells are clustered together and form the Islets of Langerhans. Within these islets, there exist five different cell types producing different endocrine hormones: alpha cells [4] producing glucagon accounting for \approx 15-20% of the total islet cells, beta-cells [4] secreting insulin representing \approx 65-80% of the islet cells, gamma-cells [5] producing pancreatic polypeptide (PP) being \approx 3-5% of the islet cells, delta-cells [4] secreting somatostatin, \approx 3-10% of total islet cells, and finally the epsilon-cells [6] producing ghrelin and accounting for <1% of total islet cells. Each of the pancreatic hormones has distinct signaling functions and altogether regulates glucose homeostasis [4, 7]

1.3 Beta-cell and glucose-stimulated insulin secretion

Pancreatic endocrine cells secrete hormones in response to external signals such as nutrient intake, stress, neural, and/or hormonal release. These signals are transferred through intracellular molecular networks and result in the release of hormones, known as stimulus-secretion coupling. Glucose is the primary stimulus for insulin release from the beta-cells through glucose-stimulated insulin secretion (GSIS). Upon blood glucose levels increase, glucose is taken up by the beta-cells, mediated by the glucose transporter located on the beta cell plasma membrane, known as the GLUT2 (*SLC2A2*) transporter. Glucose entry is sensed by the metabolic enzyme glucokinase (GCK) and is converted to glucose-6-phosphate (G6P) by

phosphorylation. G6P then enters the glycolysis cycle, which results in adenosine triphosphate (ATP) production and causing an ATP/ADP ratio alteration in the beta-cells, and further blockage of the K^+ -ATP channel in the beta-cell membrane. Normally this channel is open in order to ensure maintenance of the membrane hyper-polarization by transporting of K^+ -ions. When the K^+ -ATP channel is closed, there is a subsequent decrease in K^+ permeability, which elicits the depolarization of the beta-cell membrane and mediate opening of the voltage-dependent- Ca^{2+} -channel. A subsequent increase in intracellular calcium levels eventually triggers exocytosis of insulin-containing beta-cell granules (**Figure 2**). The insulin secretion process is biphasic; starting with an initial rapid phase of insulin secretion, where the majority of total insulin is being released, and followed by a slower “second phase”, where a less intense and more sustained insulin secretion occurs [3, 8].

In addition to GSIS, glucose also increases metabolic-cAMP (cyclic adenosine monophosphate) levels via metabolic activation of protein kinase A (PKA) within the beta-cells, also stimulating insulin secretion. In general, with the exception of glucose, there are several other mediators for insulin secretion. Incretins, such as glucagon-like peptide 1 (GLP-1), gastric inhibitory polypeptide (GIP), and vasoactive intestinal peptide (VIP), have shown to play significant roles in the second phase of insulin secretion. The effect of non-nutrient modulators on insulin secretion is through neural stimuli, adrenergic pathways, peptide hormones and cationic acids [3, 9].

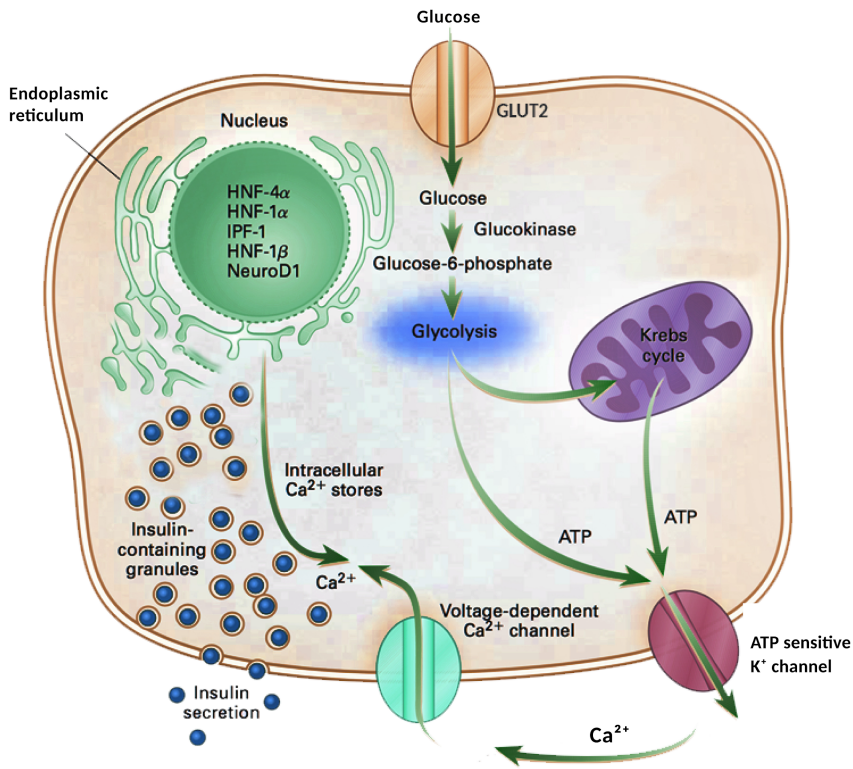


Figure 2. Glucose stimulated insulin secretion (GSIS) from the beta-cells. A rise in blood glucose level drives oxidative phosphorylation and the production of ATP, resulting in the closure of K⁺-ATP channels, plasma membrane depolarization, calcium influx and secretion of insulin vesicle by exocytosis. Figure is adapted and modified from [10].

1.4 Impaired glucose homeostasis and diabetes diagnosis

Hyperglycemia is the term for high blood glucose levels and can be the outcome of high glucose uptake from food, intravenous infusion, enhanced hepatic and renal glucose production, hepatic or peripheral insulin resistance, or due to reduced beta-cell insulin secretion. Chronic hyperglycemia is a hallmark of diabetes and can lead to severe symptoms including polyuria, polydipsia, weight loss, polyphagia and

blurred vision. Long-term complications of diabetes can also cause dysfunction of kidneys, heart, eyes, nerves, and blood vessels [11].

The world prevalence of diabetes was estimated to 6.4% among those >20 years of age (285 million people) in 2010 (Global estimates of the prevalence), and predicted to increase to 7.7% (592 million) by 2030 [12, 13]. Diagnosis of diabetes is based on plasma glucose criteria, the fasting plasma glucose (FPG) or 2-h plasma glucose level, or glycated hemoglobin (HbA1c). According to the world Health Organization (WHO) the cutoff for these levels is shown in Table 1 [14, 15].

Table 1. Criteria for diagnosis of diabetes

<p>1. HbA1C \geq6.5% (48 mmol/mol). The test should be performed in a laboratory using a method that is National Glycohemoglobin Standardization Program certified and standardized to the Diabetes Control and Complications Trial Assay*.</p> <p style="text-align: center;">OR</p> <p>2. FPG \geq126mg/dL (7.0 mmol/L). Fasting is defined as no caloric intake for at least 8h*.</p> <p style="text-align: center;">OR</p> <p>3. 2-h plasma glucose \geq200mg/dL (11.1 mmol/L) during an Oral Glucose Tolerance Test (OGTT). The test should be performed as described by the World Health Organization, using a glucose load containing the equivalent of 75g anhydrous glucose dissolved in water*.</p> <p style="text-align: center;">OR</p> <p>4. In a patient with classic symptoms of hyperglycemia or hyperglycemic crisis, a random plasma glucose \geq200mg/dL (11.1 mmol/L).</p>
--

* In the absence of unequivocal hyperglycemia, the result should be confirmed by repeat testing.

1.5 Classification of diabetes

1.5.1 Type 1 diabetes mellitus

Type 1 diabetes (T1D) accounts for 10% of all diabetes cases and was previously termed juvenile or insulin-dependent diabetes. T1D is most often diagnosed in children and young adults (<35 years old), and is caused by complete absence of insulin secretion secondary to autoimmune destruction of pancreatic beta-cells. An auto-reactivation of CD4⁺ and CD8⁺ T cells, and autoantibody-producing B-lymphocytes, activates the innate immune system and destroys the insulin producing beta-cells. Impaired insulin secretion in T1D patients have shown to be detected years before hyperglycemia, while patients with new onset T1D display 80-90% deficit in beta-cell mass. Long-standing T1D patients have lost close to 99% of their beta-cells [16, 17]. In addition to loss of beta-cells, beta-cell dysfunction also seems to play role in the pathogenesis of T1D since the degree of impaired beta-cells exceeds the amount of beta-cell lost.

Most T1D cases are positive for one or more autoimmune markers such as insulin autoantibody (IAAs), islet cell antibody (ICAs), glutamic acid decarboxylase (GAD), and the zinc transport (ZnT8). These markers are commonly present in 85-90% of T1D cases and classified as type 1A. The remaining 5% of suspected cases that lack these autoimmune markers are classified as type 1B or idiopathic [1]. Insulin is the required treatment for T1D, which has dramatically increased the quality of life of patients, and has been in clinical practice since the discovery of insulin in the early 1920s [18].

1.5.2 T1D and risk genes

T1D is a complex disease and assumed developed by the interaction of both genetic

and environmental triggers. The inheritance pattern of T1D is important and several lines of evidence indicate the presence of a strong genetic component that causes susceptibility to the disease. Although more than 85% of cases do not have a family history of the disease, there is a high familial clustering with a prevalence of 6% in siblings, compared to lifetime risk of 0.4% in the general population [19]. Furthermore, if the child's mother has diabetes the risk is 1-2%, compared to 3-7% if the child's father has diabetes. Moreover, risk in dizygotic twins is estimated to 6-10% while 30-50% in monozygotic twins [1].

Genetic studies have thus been able to explain 80% of the genetic architecture of T1D [20]. The predominant genetic susceptibility of T1D has been linked to the human leukocyte antigen (*HLA*) (encoding the major histocompatibility complex (MHC) proteins) on the short arm of chromosome 6, and accounts for up to 50% of genetic risk of T1D. The *HLA* locus is critical for distinguishing self and non-self peptides [1, 21, 22]. Numerous new susceptibility loci have been identified since, including the insulin (*INS*) gene, protein tyrosine phosphatase, cytotoxic T-lymphocyte-associated protein 4 (*CTLA4*), non-receptor type 22 (lymphoid) (*PTPN22*), the interleukin 2 receptor, as well as others that have recently been discovered from genome wide association studies (GWAS), but none of them are associated as strongly as the HLA region [23].

1.5.3 Type 2 diabetes mellitus

Type 2 diabetes (T2D), previously known as insulin-independent diabetes mellitus, is the most common form of diabetes and accounts for [90% of all cases of diabetes. T2D is a complex heterogeneous metabolic disorder, which is characterized by hyperglycemia, insulin deficiency and insulin resistance. The chronic metabolic imbalance and hyperglycemia effects vasculature and may lead to nephropathy causing renal failure, retinopathy causing loss of vision, and

complications such as stroke and atherosclerosis. Although T2D occurs mostly in adults and the incidences increase with age (>40 years), this trend is currently changing and the prevalence of T2D is now increasing in adolescents and even children. According to the International Diabetes Federation, around 1 in 11 adults aged 20-79 years, equivalent to 415 million, had T2D diabetes globally in 2015 [24].

T2D has a multifactorial trait where individual risk is defined by a complex interaction between genetic and environmental factors. Environmental factors such as physical inactivity, obesity and sedentary lifestyle, greatly influence disease development. Obesity measured by high body mass index (BMI), is the strongest risk factor for diabetes that results in insulin resistance. Incidence varies among different geographical regions. Asia shows a rapidly developing T2D epidemic, while China and India have the highest global epidemic for T2D. In these two countries diabetes is associated with lower BMI and younger age compared to Western populations. USA has the third highest diabetes population [24, 25].

1.5.4 T2D and risk genes

T2D is a complex and heritable disease. A genetic component is an important contributing factor and estimated to explain 30-70% of T2D risk. While lifetime risk in the general population is ~7%, risk in offspring with one diabetic parent is ~35%, and with two diabetic parents as high as 70%. Among monozygotic twins risk for T2D is close to 100%, and thus illustrate the strong inheritable pattern of the disease [11, 26]. T2D is a polygenic and heterogeneous disease, where multiple genes and different combination of genes play a role in the disease development in different individuals. Predicting the exact cause of T2D is challenging due to the complexity of genetic architecture of disease, varies in age of onset, low penetrance,

and heterogeneity of the disease [27]. During the last decades enormous efforts have been made to identify the risk genes through hypothesis-free Genome-wide association studies (GWAS) and candidate genes approach. In total, GWAS have identified 153 variants mapped to >120 loci [28]. Most of these genes play a role in beta-cell functions that highlight the theory that genetic factors primarily cause beta-cell dysfunction, while insulin resistance mainly result from environmental factors [29]. In other words, T2D develops when environmental factors trigger insulin resistance in the context of a genetically impaired beta-cell background.

1.5.5 Gestational diabetes mellitus

Gestational diabetes mellitus (GDM) is known as glucose intolerance with onset or first occurrence during pregnancy. GDM is a major health problem because of its prevalence, complications during pregnancy, and its risk association with T2D later in life. It is one of the most common complications of pregnancy and prevalence varies in the range of 1-14% depending on ethnicity as well as diagnostic criteria [30]. This type of diabetes includes pregnancies where diabetes might exist prior to pregnancy, but not recognized, and insulin therapy is needed. Insulin resistance is the major physiological alteration during pregnancy, which if accompanied with beta-cell dysfunction may lead to GDM. According to the WHO, diagnostic criteria includes FPG >126 mg/dL and 75 g-glucose tolerance test >140 mg/dL [30]. The pathophysiology of GDM is, however, controversial. Some believe that predisposition to T2D is underline of GDM occurring during pregnancy. Others suggest that GDM is an outcome of the extreme manifestation of metabolic changes during pregnancy.

GDM is a heterogeneous disorder where age, obesity, and genetic background play a role in the development of the disease [31]. It has further been estimated that 17-63% of GDM women have increased risk for developing T2D within 5-16 years

after pregnancy [32]. Furthermore, women with GDM can have low insulin sensitivity, dyslipidemia, insulin resistance, and high serum level of triacylglycerol. In GDM condition, the fetus produces high levels of insulin in exposition to the mother's hyperglycemia. Consequently, the fetus has high risk of developing macrosomia, neonatal hypoglycemia, hyperbilirubinemia, and has an increased risk of developing diabetes and obesity later in childhood and adulthood [1].

GDM and T2D have similar pathogenesis and as expected they share some common genetic risk factors. Several GWAS and meta analysis studies have confirmed that multiple genes related to beta-cell function are conserved between GDM and T2D, including the *TCF7L2*, *CDKAL1*, *KCNJ11*, *GCK*, *IGF2BP2*, *KCNQ1* [33], *CDKN2A/2B*, *FTO*, *HHEX*, *SLC30A8*, *KCNJ11*, *PPARG* [34] and *MTNR1B* [33, 35] genes.

1.5.6 Monogenic forms of diabetes and maturity-onset diabetes of the young

Monogenic diabetes includes all subtypes of diabetes that are caused by a defect in single genes and include mitochondrial diabetes, neonatal diabetes and maturity-onset diabetes. To date, about 30 genes have been identified related to monogenic diabetes [36]. The genetic modifications causing monogenic diabetes ultimately result in beta-cell dysfunction and diabetes.

The most common form of monogenic diabetes is Maturity-Onset Diabetes of the Young (MODY). MODY is not single entity; it is a heterogeneous group of monogenic diabetes caused by mutations in different MODY genes, which play roles in normal beta-cell development and function [37, 38]. MODY is characterized by an autosomal dominant pattern of inheritance, a family history of diabetes, early onset usually before age 25, absence of beta-cell autoimmunity, measurable C-peptide, and absence of insulin resistance [39]. MODY accounts for

□1-2% of all diabetes cases occurring in children and young adults. Accurate prevalence estimation of MODY is difficult due to some overlapping features of MODY with T1D or T2D [40]. This can result in misdiagnosis of MODY patients and inappropriate treatment. Similarities and differences in subtypes of MODY, T1D and T2D are given in Table 2.

Table 2: Clinical and biochemical characteristic features associated with T1D, T2D and common subtypes of MODY.

Features	T1D	T2D	MODY1/MODY3	MODY2
Age of onset	10-40	>35	< 25	< 25
Diabetic ketoacidosis	Common	Rare	Rare	Rare
Insulin dependent	Yes	No	No	No
Family history	<15%	>50% in young onset	Usually minimum three generations affected	Usually minimum three generations affected
Obesity	Uncommon	Common	Uncommon	Uncommon
Insulin resistance	Uncommon	Common	Uncommon	Uncommon
Pancreatic auto-antibody	Positive	Negative	Rare	Rare
C-peptide level	Low	Normal/high	Normal	Normal
First line treatment	Insulin	Metformin	Sulfonylurea	None

1.6 MODY genes and subtypes

To date, Online Mendelian Inheritance in Man (OMIM) lists mutations in 13 different genes as cause of MODY (MODY1-MODY13). These genes include the hepatocyte nuclear factor 4- α (*HNF4A*, MODY1) [41], glucokinase (*GCK*, MODY2) [42], hepatocyte nuclear factor 1- α (*HNF1A*, MODY3) [43], pancreatic and duodenal homeobox 1 (*PDX1*, MODY4) [44], hepatocyte nuclear factor 1- β (*HNF1 β* , MODY5) [45], neurogenic differentiation 1 (*NEUROD1*, MODY6)[46], kruppel-like factor 11 (*KLF11*, MODY7) [47], bile salt dependent lipase (BSDL),

also known as carboxyl ester lipase (*CEL*, MODY8) [48], paired box gene 4 (*PAX4*, MODY9) [49], insulin (*INS*, MODY10) [50], B lymphocyte kinase (*BLK*, MODY11) [51], ATP-binding cassette, subfamily C (*ABCC8*, MODY12) [52] and the potassium channel inwardly rectifying subfamily J (*KCNJ11*, MODY13)[53]. Unknown MODY loci (MODY-X) account for 20-50% of the cases and remain to be discovered [54]. Some of these genes are however debated whether they indeed are MODY genes, mainly due to being extremely rare and found in few families. Mutations in the *HNF1A*, *HNF4A* and *GCK* genes are the most common causes of MODY, whereby *HNF1A* and *GCK* mutations account for ~70% of all MODY causes [39].

1.6.1 HNF4A-MODY (MODY1)

Heterozygous mutations in *HNF4A* causing MODY1, is relatively rare and account for only 3-5% of MODY causes [55]. *HNF4A* is located in chromosome 20 and encodes for HNF-4A, which is a transcription factor expressed in the liver, pancreas, kidney and small intestine. It binds to the DNA of target genes as a homodimer and activates transcription of its target genes. HNF-4A is member of nuclear receptor superfamily and all members contain a ligand-binding site. HNF-4A is a member of the orphan family and few specific ligands have been described, with the exception of long chain fatty acids, which are known as modulator of HNF-4A transcription activity binding to ligand domain of HNF-4A [54, 56] The HNF-4A protein consists of several functional domains; an N-terminal transactivation domain (AF-1), a DNA binding domain, and C-terminal complex that form ligand binding domain, dimerization domain, and transactivation domain. The *HNF4A* gene has two distinct promoters (P1 and P2), The P1 promoter is utilized in hepatocytes and P2 in pancreatic beta-cells. Expression through these two promoters result in nine isoforms of *HNF4A* where some isoforms are present either in the liver or in the pancreas [57, 58]. The clinical phenotype is in principle

indistinguishable from HNF1A-MODY (MODY3).

1.6.2 GCK-MODY (MODY2)

The second most common type of MODY is MODY2, which is caused by mutations in the *GCK* gene. *GCK* is located on chromosome 7 and encodes for the metabolic enzyme glucokinase (GCK), and is expressed mainly in the pancreas, liver and brain. Glucokinase converts glucose to G6P; a key and rate-controlling step in glucose metabolism. Therefore, GCK acts as a glucose sensor in pancreatic beta-cells and controls insulin secretion [10]. In the liver it plays an important role in the ability of the organ to store glucose as glycogen. Heterozygous loss-of-function mutations in *GCK* result in partial deficiency of the enzyme and are associated with MODY2 [42]. Homozygous or compound heterozygous mutations of *GCK* result in complete deficiency of the enzyme and cause severe permanent neonatal diabetes [59, 60]. MODY2 is generally asymptomatic with mild hyperglycemia, which commonly is diagnosed during routine screening or during pregnancy, and typically does not lead to long-term diabetes complications. MODY2 cases have the same risk as the general population for developing polygenic T2D [39]. Distinguishing MODY2 from other causes of GDM is important because it will save patients from unnecessary treatment [61]. In the case of MODY2 associated GDM, a mother does not have an increased risk for T2D and will not need treatment after pregnancy, since she is not at risk of developing later onset diabetes complications [39].

1.6.3 HNF1A-MODY (MODY3)

The hepatocyte nuclear factor 1- α (HNF-1A), known also as HNF1 or TCF1, is a transcription factor encoded by the *HNF1A* gene that is expressed in the liver, pancreas, stomach, small intestine and kidney [62-64]. In the pancreas, both

endocrine and exocrine cells express *HNF1A* during development [65]. The *HNF1A* gene is located on the long arm of chromosome 12, and consists of 10 exons, encoding the 631 amino acid long HNF-1A protein.

Heterozygous mutations in *HNF1A* result in the most common cause of MODY (MODY3) in most populations [66]. The obvious clinical feature of MODY3 is impaired insulin secretion, which result in rapid deterioration of glycemia level by age, and requires treatment. According to the Human Gene Mutation Database (HGMD), around 500 different diabetes associated mutations in *HNF1A* have been reported [67]. *HNF1A* mutations present with high penetrance, in which 63% of *HNF1A* mutation carriers develop diabetes before age of 25, 79% before age 35, and as many as 96% before age 55 years [39]. Although mutations have been identified in all exons of the *HNF1A* gene, mutations in exons 1 to 6 have been reported to give earlier onset compared to mutations in exon 8 to 10 [68].

HNF1A mutations can cause diabetes through haplo-insufficiency or dominant negative effect. The level of HNF-1A activity is important for beta-cell function and defines the severity of the disease. MODY3 patients have shown to be sensitive to the oral hypoglycemic medication, sulfonylurea [55], where some patients have demonstrated good glycemic control upon sulfonylurea treatment up to 49 years after diagnosis [69]. Therefore, confirming a genetic diagnosis of MODY is important, which provides the knowledge to classify the subtype, predict the prognosis of the patient, select the precise treatment, and estimate the risk in patient relatives [66].

Not surprisingly, the clinical implications of MODY3 and MODY1 are similar. They both demonstrate mild increase in fasting blood glucose levels in childhood, but develop progressive hyperglycemia and finally diabetes in adulthood. Patients may also develop diabetes complications such as microangiopathic and microvascular complications. Despite the similarities, it is possible to distinguish

these two types of MODY by the ability of glucose to stimulate insulin secretion through GSIS, since this ability is retained in MODY3 cases while lost in MODY1 [10]. Also MODY3 patients have higher levels of high-density lipoprotein (HDL) concentrations while MODY1 patients present low levels of HDL and triglyceride, but high levels of low-density lipoprotein (LDL) [39].

1.7 HNF-1A protein and function

1.7.1 *HNF1A* transcripts and protein isoforms

Alternative splicing of the *HNF1A* gene generates three transcripts encoding three different protein isoforms (A, B and C) [70]. These isoforms are identical in their N-terminal end while differs in their C-terminal. Here, isoform B is truncated at exon 6 and isoform C terminates within exon 7. The pattern of isoform expression differs during pancreas development. Isoform A is the dominant form in fetal pancreas while isoform B is the main isoform in the adult pancreas (55% total expression) [70]. In liver tissue, isoform A is the predominant form, comprising around 54% of total HNF-1A expression in adult. Transactivation studies indicate that all three isoforms of *HNF1A* are active, however isoform B and C being 5-fold more active than isoform A [66, 70].

The HNF-1A protein further consists of an N-terminal dimerization domain (amino acid 1-32), a DNA-binding domain consisting of a POU domain (100-199 amino acid) and homeodomain (199-286 amino acid), and a C-terminal transactivation domain (287-631 amino acid) [70-72]. An overview of the HNF-1A protein functional domains is shown in **Figure 3**.

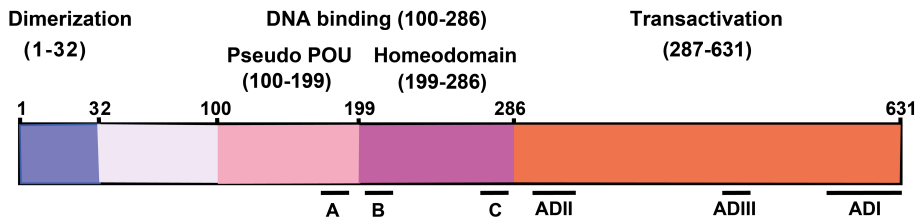


Figure 3. HNF-1A protein functional domains. Illustration of localization of the dimerization, DNA binding, and transactivation domains in HNF-1A. A, B and C; representing the putative nuclear localization signals A, B and C identified in HNF-1A. ADI, ADII and ADIII; representing the reported most important transactivation regions I, II and III within the transactivation domain.

1.7.2 Subcellular localization of HNF-1A

The nucleus is a distinct compartment of the eukaryotic cell, facilitating diverse cellular processes like gene expression, signaling pathways, and cell cycle regulation. It separates the DNA content and transcriptional machinery from protein synthesis and metabolic pathways within the cell cytoplasm. The nuclear membrane contains large structures known as Nuclear Pore Complexes (NPCs), which allow transport of small molecules, <60 kDa in size, while macromolecules are transported through an active transport mechanism regulated by the GTPase protein (Guanosine triphosphatase), Ran (Ras-related nuclear protein), which switches between GDP- and GTP-bound states through its regulation by regulatory proteins Ran guanine nucleotide exchange factor (Ran-GEF) and Ran GTPase activation protein (Ran-GAP). Ran-GEF is localized in the nucleus and Ran-GAP is localized in the cytoplasm; an arrangement which plays a role in determining the direction of nuclear transport by creation of a Ran-GDP/ Ran-GTP gradient in cytoplasm/nucleus [73].

The main element regulating the nuclear localization of a protein is the Nuclear Localization Signal (NLS). The NLS sequence within a target protein designates selective transport and accumulation of the protein in the nucleus, through its

recognition by an adaptor protein, importin- α , which has a binding motif for the importin- β protein. The import mechanism occurs by the target protein (cargo) binding an importin- α - β heterodimer complex. The affinity of binding determines the efficiency of transport. Generally, the importin- α - β heterodimer complex binds to its cargo (target protein) in the cytoplasm, and releases it in the nucleus, through a Ran-GTP-induced dissociation [73, 74]. Since HNF-1A is a large protein (\approx 75kDa), it is transported through an active transport- and NLS-mediated mechanism. Studies have shown that there exist three regions in the HNF-1A protein structure that are similar to NLSs; region A (158-171 amino acid), region B (197-205 amino acid) and region C (271-282 amino acid) [75, 76]. Regions B and C are mainly involved in nuclear translocation of HNF-1A; in which region B is the main essential element whereas region C has significant contribution in the process [77].

1.7.3 DNA-binding of HNF-1A

Transcription factors have specific DNA-binding domains, which has high affinity for binding to a specific sequence within target gene promoters. Crystal structures of the DNA binding domain of HNF-1A has been elucidated and showed that HNF1-A binds to a palindromic consensus sequence, GTTAATNATTAAC, on the target promoter either as a homodimer (with itself) or a heterodimer with HNF-1B [78-80]. This domain (residues 100-286) consists of two parts; a POU like (pseudo POU; (residues 100-199) and homeodomain (residues 199-286) motifs. HNF-1A binds to DNA through an atypical homeodomain that spans exon 3 and exon 4, and has a unique structure. It contains a loop of an extra 21-amino acid insertion between helices 2 and 3, not common in other homeodomains. The POU domain further consists of two sub-domains: POU_S (specific POU) and POU_H (homeodomain). The POU_S is an integral element of HNF-1A that plays a role in the stability of the protein, while POU_H initiates protein-DNA interaction. Structural

studies indicate that the POU_H domain interacts with the 21-amino acid loop of POU_H and form a boundary between these DNA binding domains [81]. The DNA binding domain of HNF-1A has the highest mutation rate (0.15 per nucleotide) compared with transactivation domain (0.03 per nucleotide)[81]. According to the Human Gene Mutation Database (HGMD), 39 diabetes-associated missense and five nonsense mutations have been localized to the POU_H domain of HNF-1A [82], and functional studies have shown that those causing MODY3 severely impair the DNA binding and HNF-1A regulation of target genes.

1.7.4 Transcriptional activation by HNF-1A

The C-terminal transactivation domain of HNF-1A, located to residues 287-631, contains three specific regions: ADI (residue 546-628), ADII (residues 281-318), and ADIII (residues 440-506) (**Figure 3**). Studies have reported that *in vivo* transactivation is mainly achieved by the action of the two regions ADI and ADII [83], while *in vitro* transactivation activity required the combination of ADI and ADIII [84]. Furthermore, the interaction of HNF-1A with cofactors for transcriptional activation has been established. Various domains within HNF-1A have been found capable of interacting with certain coactivators including the dimerization cofactor of HNF-1A (DCoH), also known as protein-4-alpha-carbinolamine dehydratase 1(PCBD1), which, stabilize HNF-1A homodimers, leading to an increased transcriptional activity of HNF-1A. DCoH is also known as a phenylalanine hydroxylation enzyme, however it is not clear whether dehydratase activity of DCoH is essential for HNF-1A activity or not [85]. The High Mobility Group Protein-B1 (HMGB1) has also been identified as an HNF-1A cofactor. HMGB1 is a non-histone chromosomal protein that stabilizes nucleosomes and facilitates DNA binding for transcription. These two proteins interact through HMG box domains of HMGB1 and the homeodomain of HNF-1A, and HMGB1 promotes the DNA-binding ability of HNF-1A and increases its transcriptional activity [86].

Additional, studies have confirmed the binding of HNF-1A to several other coactivators such as the Histone Acetyl Transferases (HATs), CREB-Binding Protein (CBP), p300/CBP-associated factor (P/CAF), SRC-1, and RAC3 [87]. Of these, the co-activator p300/CBP interacts with the DNA-binding and transactivation domain of HNF-1A, while P/CAF, SRC-1 and RAC3 have been shown to interact only with the HNF-1A transactivation domain [87].

1.7.5 HNF-1A and target gene regulation

HNF-1A has been associated with the regulated expression of more than 222 target genes in liver hepatocytes central for hepatic functions including carbohydrate metabolism and storage, lipid metabolism (cholesterol synthesis and apolipoproteins), detoxification (synthesis of cytochrome P450) and synthesis of serum proteins (albumin, complements and coagulation) [88]. Some of these genes include *SLC2A2* (*GLUT2*), *ALB* (albumin), *FGB* (β -fibrinogen), *AAT* (α -trypsin), *AFP* (α -fetoprotein), *FGA* (α -fibrinogen), *A1AT* (SERPINA1), *TTR* (transthyretin) and *ALDOB* (aldolase B) [70, 86].

In pancreatic beta-cells, HNF-1A has been associated with regulating the expression of more than 100 genes, of which 30% of these genes are identical to those found in hepatocytes [85]. For instance, HNF-1A regulates the expression of the *GLUT2* gene (*SLC2A2*) that encodes the solute carrier family 2 member 2 that ensures glucose transport in beta-cells, the HNF-4A gene (*HNF4A*) through P2 promoter, the L-type pyruvate kinase gene (*PKL*), which encodes a rate limiting kinase of glycolysis, and the insulin gene (*INS*) [65, 89]. In addition, HNF-1A has been shown to regulate the expressions of several other genes associated with cell proliferation and apoptosis such as the *IGF-1*, *cyclin E* and *BCL-XL* genes. Reduced beta-cell proliferation rate and decreased beta-cell number has been reported in

HNF1A knockout mice and transgenic mice carrying dominant negative mutation in *HNF1A* [65].

1.7.6 HNF-1A and gene network regulation

There is strong evidence that the five MODY genes and transcription factors (HNF-1A, HNF-4A, HNF-1B, PDX-1 and NEUROD1) act synergistically to support normal pancreatic development and play an important role in normal function of beta-cells and insulin secretion [10]. In the pancreas, HNF-1A is a member of a transcription factor network that regulates the expression of various genes in glucose transport, metabolism and insulin secretion. For instance, it has been suggested that HNF-3B in the pancreas regulates the expression of *HNF1A*, *HNF4A* and *IPF*, whereas HNF-6 regulates expression of *HNF3B* (**Figure. 4**) [65, 90]. In the liver, *HNF1A* expression is regulated by transcription factors including HNF-3A, HNF-3B and HNF-4A. While, HNF-1A is dependent on HNF-4A for normal transcription in the liver, HNF-1A controls *HNF4A* gene expression in differentiated pancreatic cells [54].

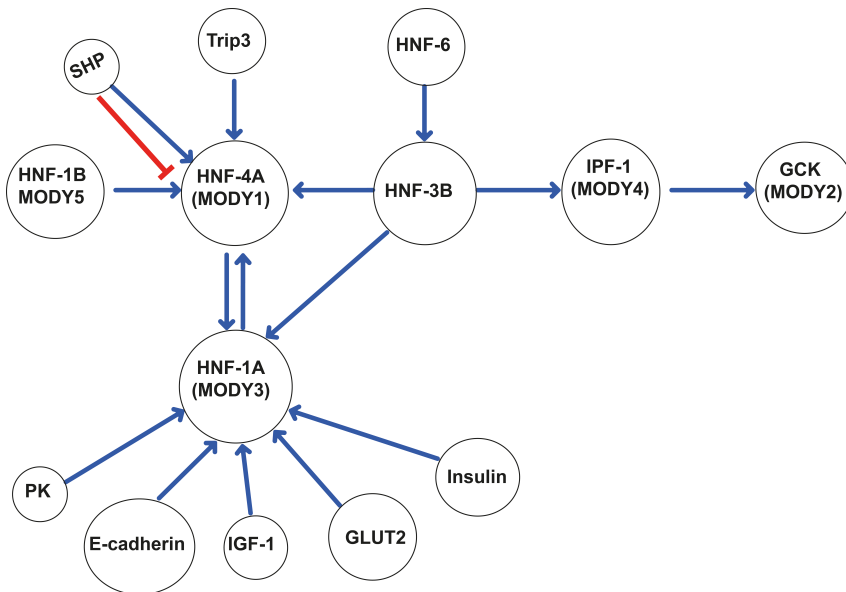


Figure 4. HNF transcription network in the pancreatic beta-cell. HNF-1A regulates expression of *HNF4A*. HNF-1B acts as a homodimer or heterodimer with HNF-1A. Trip enhances HNF-4A function while SHP suppresses HNF-4A function. This figure is adapted and modified from [65].

1.8 Deciphering the genetic architecture of T2D

Many common diseases are known as heterogeneous diseases, which cluster with families and are influenced by both genetic and environmental factors. Understanding the genetic architecture of complex diseases is important because the individual differences in susceptibility to disease are believed to have genetic reasons. Identifying these genetic factors could aid correct diagnosis, precise treatment and better risk estimation and prevention. Several approaches have been applied to understand the genetic basis of T2D.

1.8.1 The candidate gene and linkage analysis approach

Early efforts for the genetic mapping of T2D began with the candidate gene and linkage studies. Candidate gene approach searches for the association between T2D and the variants in or nearby target genes, and comparing their frequencies of identified variants in cases and controls or parents-offspring [27]. In linkage study the segments of DNA are traced in families to identify a locus, which co-segregates with T2D. Although the candidate gene- and linkage study approaches were able to identify several genes such as *PPARG* [91], *KCJN11*[92] and *TCF7L2* [93], it has failed to explain more than 95% of the genetic basis of T2D, suggesting that the majority of susceptibility of T2D might arise from multiple loci with small effect size [94].

1.8.2 Genome-wide association studies approach

The next attempts for discovering the genetic basis of T2D was motivated by the “common disease, common variants” hypothesis indicating that common disease arise from common variants. GWAS as approach for genetic mapping compare the frequency of variants in large case-control cohorts [94]. GWAS has been a powerful tool and have identified around 150 novel loci in the predisposition of T2D and including three genes previously identified by a candidate genes approach (*PPARG*, *KCJN11* and *TCF7L2*) [95]. Recent studies have also discovered loci pointing to monogenic diabetes-associated genes; *HNFI1A*, *HNFI1B*, *GCK*, *PDX1*, *GLIS3*, *WFS1*, *PAX4* and *LMNA* [96]. Despite GWAS success, the identified association loci can explain less than 10% of the genetic background of T2D [97].

The existing catalogue of human variants allows GWAS to detect only common variants (Minor Allele Frequency (MAF) >5%). Apparently, common variants with

small effect do not contribute to the majority of the heritability of T2D [94]. Furthermore, limitations of GWAS are that they only discover the association loci and do not have sufficient resolution for distinguishing the causal variants. Therefore in-depth sequencing may be more ideal to identify causal variants [95].

1.8.3 Rare variants and missing heritability

The majority of heritability of complex disease is remaining unexplained and hence the term “missing heritability”. It has been hypothesized that rare variants with large effect may exist and could explain the missing heritability. The frequencies of such variants are low; therefore they could not be captured by current GWAS arrays. GWAS could capture common variants, which usually caused by common allele ($MAF \geq 5\%$) with small effect size. Due to small effect size of common variants they could not be detected in families through linkage analysis studies. In contrast, rare Mendelian diseases are usually caused by rare variants with larger effect size and can be detected by linkage analysis **Figure 5**. When the $MAF < 0.5\%$, the chance of association detection is unlikely unless there is large effect size like monogenic situation [98].

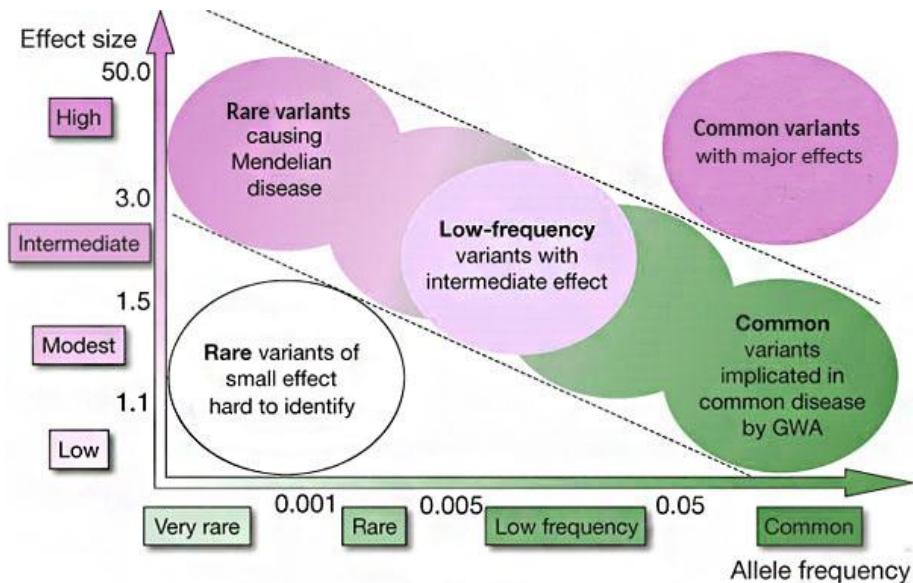


Figure 5. Feasibility of identifying genetic variants by risk allele frequency and strength of genetic effect. Common disease caused by common alleles with minor allele frequency greater than 0.05 and small effects sizes, shown in bottom right of graph. Rare Mendelian disease causes by rare variants with large effects sizes, shown in upper left of graph. Low frequency variants with modest effect could contribute to missing heritability of common disease, shown in center of graph. This graph is adapted and modified from [98].

The next-generation sequencing (NGS) technique provides the possibility of investigating low frequency and rare variants related to complex disease. Studies of rare variants are more recent and the first exome sequencing was conducted in 2012 that identified five rare variants; two reported before *SGSM2* and *MADD* and three novel variants *TBC1D30*, *KANK1* and *PAM* in association with glycaemic trait and T2D [99]. Several other sequencing studies have identified low frequency variants in *PAM* and *PDX1* associated with T2D or related traits [100]. Although these reports demonstrate the ability of NGS in identifying low frequency variants associated with T2D, the major challenge in rare variants studies is however to find

the association of variants to the disease. There should be enough copies of each allele to be able to authorize statistic analysis, while the majority of variants only presents in a single or few individuals and makes it therefore difficult to conclude any association. The required sample size is increasing linearly with the ratio of $1/MAF$. In addition, the sample size for association studies increases as odds ratio (is the unit for defining the effect size, the probability of individuals having risk allele versus not having the risk allele) drop. Therefore, low frequency and rare variants should have higher odds ratio to be detected [98].

Altogether, results suggest that the low frequency variants alone may play a limited role in the genetic basis of T2D [97]. Consequently, it seems that GWAS will probably remain as an efficient approach in identifying additional loci, since rare variants association studies require much larger comprehensive sample size. To overcome limitation of rare variants studies, several factors have been suggested to consider in the study design like extreme phenotype sampling, studying isolated populations, and proper choice of statistical methods to test rare variants association [101, 102].

Genome/exome sequencing increasingly generates the number of DNA sequence variants and extensive efforts are needed to interpret their true consequence. Variant classification is however challenging and care must be taken in applying criteria for interpretation of variants. In 2015, the American College of Medical Genetics and Genomics (ACMG) published important guidelines for variant interpretation of monogenic diseases (not suitable for complex diseases) [103].

The ACMG guidelines give a range of criteria in order to help classify a variant into five different groups i.e. benign (class 1), likely benign (class 2), variant of unknown significance (class 3), likely pathogenic (class 4) and pathogenic (class 5). The different criteria are divided into four main groups for classification of either pathogenic or benign variants. Those variants that could not meet the criteria for

pathogenic and benign are classified as variant of unknown significance (VUS). For instance if the variant is a null variant (e.g. nonsense, frameshift) this is a very strong evidence for pathogenicity, for genes where loss-of-function is already a known mechanism for the disease. Some examples of other criteria for evidence of pathogenicity are if functional studies supports a deleterious effect of the gene (strong evidence), the variant involves the same amino acid change that have been known as pathogenic previously (moderate evidence), if the variant is absent from control populations (e.g. 1000 Genomes Project, Exome Aggregation Consortium (ExAC), Genome Aggregation Database (GnomAD)) (moderate evidence) and if the variant cosegregates in several affected family members (supporting evidence) [103, 104].

Several computational (*in silico*) tools, that are publicly or commercially available, have throughout the years been developed in order to support variant interpretation. However, the accuracy of *in silico* tools for prediction of missense variants in Mendelian disease is 65-80% and is not recommend to use as sole evidence to make a clinical decision [105].

Despite the guidelines described above, variant interpretation is still not perfect, and has been shown to be inconsistent between different laboratories (and even within the same laboratory) [106]. Robust functional assessment is a strong evidence for pathogenicity for variant interpretation. In order to efficiently map variants effect as cause of a clinical phenotype, functional assays are needed at the same scale and speed as variants are being identified. Therefore, there is an essential need to develop high-throughput functional assays that can efficiently cope with numerous variants simultaneously, rather than the “one-by-one”-strategy that is the most commonly used approach today.

2. AIM OF STUDY

The overall aim of this thesis was the functional interpretation of numerous rare missense *HNF1A* variants to find causative variants or risk factors for MODY3 or T2D.

Specific aims:

1. To characterize the functional consequences of rare missense *HNF1A* variant associated with risk of T2D in a Mexican and Latino population (Paper I)
2. To understand the functional consequences of rare protein coding *HNF1A* variants identified in the general population as potential risk factors for T2D (Paper II)
3. To identify novel HNF-1A regulated transcripts as potential read-out for future functional interpretation of numerous *HNF1A* variant alleles (Paper III)

3. SUMMARY OF RESULT

3.1 Paper I:

Association of a low-frequency variant in *HNF1A* with T2D in a Latino population. The Mexican and Latino population have one of the highest prevalence of T2D worldwide. To investigate the genetic basis of T2D in this population, whole-exome sequencing was performed in 3756 Mexican and American Latino individuals (1794 with type 2 diabetes and 1962 without diabetes). The c.1522G>A (p.E508K) variant in the *HNF1A* gene was observed in 0.36 % of control subjects and in 2.1 % individuals with T2D. The data were replicated in a multiethnic data set (T2D-Genes and Go-T2D; ~15 000 individuals, T2D/controls) confirming the presence of *HNF1A* p.(E508K) variant only in Latino patients. The pathogenic effect of this rare *HNF1A* variant on normal HNF-1A function was examined by assessing transactivation activity, DNA binding, subcellular localization, and cellular protein level. The p.(E508K) variant demonstrated reduced transactivation activity (TA) on a target rat *albumin* promoter (TA<50%), however its activity was not as low as severe MODY3 variants (p.(P447L), p.(P379sdelCT), p.(R229Q)) included as controls. Reduced TA of the p.(E508K) was also confirmed by other target promoters (*GLUT2* and *HNF4AP2*) in a beta-cell like mouse model (MIN6) (TA<60%). Furthermore, the p.(E508K) variant showed slightly impaired nuclear localization, but not as significant as MODY3 control variants. Although the variant did not affect HNF-1A DNA binding the protein expression level of p.(E508K) was significantly lower compared to wild-type HNF-1A, indicating that this missense variant might cause an unstable p.(E508K) protein.

3.2 Paper II:

Functional investigations of *HNF1A* identify rare variants as risk factors for T2D in the general population. Exome sequencing of *HNF1A* in population-sized cohorts have shown that 1/50 individuals with diabetes harbors a missense variant in the *HNF1A* gene. As many as 27 rare *HNF1A* missense variants (MAF < 1%) were identified in well-phenotyped population (n=4,115). We used five different bioinformatics prediction tools including SIFT, PolyPhen-2, Combined Annotation Dependent Depletion, MutationTaster and Align Grantham Variation and Grantham Deviation for variants classification. None of the variants classified as pathogenic were associated with risk for T2D (OR=2.02; 95% CI 0.73-5.60; $P=0.18$). To better understand a pathogenic relevance and variant effect on normal HNF-1A function, we evaluated the functional consequences of individual variants experimentally. Effect on transcriptional activity was assessed using a rat *albumin* promoter driven of luciferase reporter assay in transfected HeLa cells and 11 of 27 variants were identified with reduced TA activity (<60%). The effects of the variants on the subcellular localization of HNF-1A demonstrated that most protein variants had normal nuclear versus cytoplasmic localization compared to wild-type HNF-1A, while three variants (p.(R131Q), p.(E508K), p.(H514R)) had slightly reduced nuclear localization. When assessing which functional assay could best predict T2D association based on carrier phenotypes, best association was shown for transactivation impaired variants (<60%) and T2D (OR=5.04; 95% CI 1.99-12.80; $P=0.0007$), while, variants with transactivation activity >60% conferred no additional risk for T2D (OR=0.77; 95% CI 0.35-1.72; $P=0.53$). In aggregate, these impaired *HNF1A* missense variants conferred about a 6-fold increased risk of T2D in carriers. This study concluded that functional assays could identify variants better with a true pathogenic effect, and potential risk factor for the development of T2D, compared to bioinformatics prediction tools.

3.3 Paper III:

Generation of liver cell models for improved identification of novel HNF-1A target genes using CRISPR/Cas9 technology. To improve the diagnostic interpretation of numerous *HNF1A* variants and their pathogenic effect, a robust and reliable high-throughput functional assay is needed. For this purpose, we searched for strongly regulated HNF-1A transcripts in liver cell models for the purpose of massively parallel allelic screening of *HNF1A* variant dysfunction. To evaluate variants function in clean genetic background, we also generated *HNF1A*- free cell lines by knocking out endogenous *HNF1A* in the HuH7 and Hep3B liver cell lines using CRISPR/Cas9 technology. To identify HNF-1A endogenous targets, the *HNF1A*- free cell lines were transduced with doxycycline-inducible recoded WT *HNF1A*, followed by RNA sequencing and transcript analysis. A total of 62 and 367 genes in HuH7 and Hep3B cell lines, respectively, were significantly upregulated (FDR (false discovery rate) < 0.01 and log₂FC (fold change) ≥ 2) by WT HNF-1A induction at 5 µg/ml doxycycline.

In order to select relevant targets that could discriminate *HNF1A* variants, expression of series of *HNF1A* alleles (with known clinical and functional evidence) was equally performed in HuH7 and Hep3B cell line, followed by RNA sequencing. The expression levels of top 20 upregulated genes in WT HNF-1A condition are nominated in allelic series analysis. We found the majority of the top 20 upregulated genes (induced by WT HNF-1A) were significantly down-regulated in severe MODY3 variants (p.(P112L) and p.(P447L)), while they upregulated or didn't show significant differential expression in mild-loss of function variant p.(E508K) condition. Among top 20 hits, by considering their ability in discrimination of *HNF1A* variants and the level of expression of genes, *ABCC2*, *FABP1* and *HABP2* genes in HuH7 and *HKDC1*, *HRG* and *KL* in Hep3B selected as potential targets of HNF1A.

This study has thus identified several potential novel HNF-1A regulated transcripts in HuH7 and Hep3B cell lines, which can be used as potential markers for future high-throughput functional characterization of *HNF1A* allelic variants.

4. DISCUSSION

4.1 Prevalence and effect of common and rare *HNFI1A* variants in T2D diabetes populations

Identified genes responsible for monogenic forms of diabetes along with discoveries of T2D-loci by GWAS provide strong evidence that monogenic diabetes genes may be involved in the susceptibility of the polygenic forms of diabetes as well [96]. These findings are a strong indication of the hypothesis that different variants in the same gene may cause various diabetes phenotypes, from monogenic forms to common forms of diabetes disease. Genome sequencing of target genes can thus identify individuals in the general population who carry coding variants in genes for Mendelian disorders and who may consequently have increased disease risk.

Common variants in the *HNFI1A* gene have been reported as risk factor for T2D. Studies have shown that the common SNPs p.(I27L), p.(A98V) and p.(S487N), alter HNF-1A transcriptional activity *in vitro* and insulin secretion *in vivo*. Furthermore, p.(I27L) has been associated with increased risk of diabetes (OR=1.5, p=0.002) [89, 107, 108]. Others, however, have not been able to confirm these associations [109]. This may be explained by such variants only presenting a modest effect, or a role of other unknown genetic or environmental factors in other and different ethnic populations.

T2D is anticipated to be a heterogeneous disease. The heterogeneity of a complex disease within a population raise the question how genetic risk variants can be translated between populations, and is a fundamental aspect to consider in common disease risk prediction [110]. To illustrate this, screening for rare coding variants associated with T2D in a Mexican and Latin American population (4000 individuals) identified the low frequency rare p.(E508K) variant in *HNFI1A* (OR=5.48; 95% CI 2.83-10.61; P=4.4x10⁻⁷) (**Paper I**). This variant is private to the

Mexican and Latin American population and has so far not been found in other diabetes populations. Functional investigation of this variant demonstrated a mild reduction in HNF-1A transactivation activity <50%, compared to MODY3 variants (p.(P447L), p.(P379sdelCT), p.(R229Q)), that has severely reduced transactivation activity <20% (**Paper I**). Confirmation of reduced transcriptional activity of the strongly T2D associated variant p.(E508K) in transfected HeLa cells was also shown in a mouse insulinoma (MIN6) cell line, and by testing other HNF-1A target promoters (*GLUT2* and *HNF4A P2*) (**Paper I**). The p.(E508K) variant also showed slightly impaired nuclear localization compared to a MODY3 variant (p.(Q466X)) that highly accumulates in the cytoplasm compared to nucleus. In addition, the p.(E508K) variant demonstrated significantly lower protein levels (≈50%) compared to wild-type HNF-1A (100%). Since the DNA binding of p.(E508K) was normal, the low level of transactivation activity might be explained by conformation changes that could cause an unstable protein and reduced transactivation activity due to this variant's position in the transactivation domain of the protein sequence. Therefore, the rare protein coding p.(E508K) variant, which demonstrates mild-loss of function, is considered a risk variant for T2D in the Mexican and Latin American population. Furthermore, the p.(E508K) variant being not as functionally severe as classic MODY3 variant, is thus thought to influence diabetes onset later in life.

Another private missense variant in *HNF1A*, p.(G319S), has been reported in the Canadian Oji-Cree T2D Indian population (117 individuals). This variant was found in as many as 40% of patients with diabetes, and was found to increase the onset of T2D by seven years [111]. Similarly to p.(E508K), p.(G319S) is located in the transactivation domain of HNF-1A and reduces transactivation by nearly 50% [112]. Both variants have shown incomplete penetrance, suggesting that they are moderate diabetes alleles and different from the true MODY3 variants that demonstrate high disease penetrance[113]. These variants, however, demonstrate mild loss-of-function effect by functional assessments, which might contribute to a

polygenic background and predispose some individuals from specific ethnic populations to T2D. To investigate this further, the prevalence and effect of *HNF1A* gene variants from large multi-ethnic based T2D cohorts (around 13 000 individuals) are currently being investigated by collaborators and us. So far, around 80 different *HNF1A* variants have been identified in cohort individuals (case/controls). Assessing their gradient of pathogenic effect, based on *in vitro* functional assays, will be important both in terms of determining the presence of ethnicity-specific *HNF1A* variants, and for the identification of variant-specific phenotypic markers.

4.2 Prevalence and relevance of rare *HNF1A* variants as risk factor for T2D diabetes in the general population

Genome sequencing of individuals of extreme phenotypes may display ascertainment bias in identifying causal rare variants associated with the disease. Such studies, however, may report an upward bias in the estimated effect sizes in healthy individuals carrying such risk variants. Therefore, care must be taken in analysis of these variants by using proper statistical tests to avoid falsely predicting individuals as being at risk for Mendelian disease. Genome sequencing of the most common MODY genes in a well-phenotyped population (4 003 individuals) was recently performed by Flannick and colleagues [114], and was the basis for our work in **Paper II**. In this study, the spectrum of rare variants in the seven most common MODY genes was investigated in three population cohorts (the Framingham Heart Study, The Jackson Heart Study, and extremes of type 2 diabetes) [114]. They reported that around 5% of individuals in this general population carry a low frequency non-synonymous variant in one of these seven MODY genes. Among these, 1.5% (in the Framingham Heart study) and 0.5% (in the Jackson Heart Study) carry variants previously reported as MODY, or classified as rare, conserved, and protein damaging by bioinformatics predictions tools.

Among these, 27 rare non-synonymous variants were identified in *HNF1A*; the most common MODY gene.

In **Paper II**, we adopted to assess the pathogenic relevance of each and every one of these 27 *HNF1A* variants using two different approaches: 1) bioinformatics analysis tools commonly used in medical genetics laboratories, and 2) by functional protein analyses. Bioinformatics evaluation of the 27 rare *HNF1A* variants classified 11 variants as likely pathogenic, and these variants did not show significant association with T2D (OR=2.02; 95% CI 0.73-5.60; P=0.18). Using our second approach, we did however find that the functionally evaluated impaired variants (activity < 60%) showed significant association with T2D (OR=5.04; 95% CI 1.99-12.80; P=0.0007). Based on this, we concluded that bioinformatics tools can aid in variants interpretation regarding predicted pathogenic effect, however they are often inconclusive (resulting in a large number of variants with unknown significant; VUSs) and prone to erroneous results. The general accuracy of bioinformatics prediction tools has been estimated to 75-80% by others [115]. Such tools are mostly capable of predicting the two-end side of the spectrum; benign and pathogenic variants, but are not accurate enough to classify mild loss-of-function variants.

Based on numerous reports on MODY3 associated *HNF1A* variants [77, 116-118] and our own experience from our studies presented in **Paper I** and **Paper II**, we believe that functional characterization of variants effect is the ideal method to truly understand the phenotypic consequences of genome sequence variants on HNF-1A protein function (genotype-function-phenotype). Different variants in *HNF1A* present different phenotypic profiles. This, to some extent, depends on the type of amino acid substitution and location in the protein sequence, affecting different properties and function of the HNF-1A protein. The experimental methods for evaluation of variants effect on HNF-1A function generally include assessing the

effect on HNF-1A expression/stability, subcellular localization, transcriptional activity, and DNA binding ability, through a “one-by-one” analysis of individual variants (not high-throughput) (**Paper I** and **Paper II**).

4.3 *HNF1A* variant effect on nuclear localization and disease risk prediction

Previous studies have identified variants that result in impaired (reduced) nuclear localization of HNF-1A [77, 116, 118]. Such loss in proper nuclear translocation may result in reduced activation of HNF-1A target genes in an *in vivo* situation. Most of the 27 rare *HNF1A* variants identified in the general population [114] demonstrated nuclear localization levels similar to wild-type HNF-1A. Only three variants (p.(R131Q), p.(E508K), p.(H514R)) showed slightly impaired nuclear localization, however not to the same extent as the previously described MODY3 variant p.(Q466X) [118](**Paper I** and **II**). Regarding the p.(R131Q) variant, the amino acid arginine in position 131 is highly conserved and exposed, and this residue interacts with DNA according to a previous report [82]. This study further demonstrated that the p.(R131Q) substitution within HNF-1A has a destabilizing effect on the protein that may explain the low protein expression, low DNA binding, and low transcriptional activity of p.(R131Q) variant observed by us (**Paper II**). Of the four variants studied (p.(R131Q), p.(E508K), p.(H514R) and p.(Q466X), none are localized in the previously determined NLS in HNF-1A [77], indicating that prediction of variant effect on nuclear translocation cannot only be determined based on localization of the variant within the HNF-1A protein sequence. Other variant mediating factors like altering HNF-1A protein structure and possibly masking an NLS, or causing aggregation and cytosolic HNF-1A accumulation, are hypotheses for reduced nuclear localization that may be at play for such variants.

When we used the functional approach for studying the effect of 27 *HNF1A* variants on nuclear localization we assessed whether this assay (nuclear localization) could provide as model for predicting impaired variants as risk factors for T2D (**Paper II**). Variants were subdivided based on different thresholds of nuclear localization level (<80%, <70%, <60%) and associations of carriers with T2D were calculated. A model of subcellular localization <70% compare to WT HNF-1A showed significant association with diabetes ($P=0.0007$), however, with this threshold we could only detect two variants as impaired variants. Although, this assay and model could help to understand the mechanism behind dysfunctionality caused by *HNF1A* variants, our study concluded that this assay is not recommended to use alone as a functional predictor for the evaluation of *HNF1A* variants as T2D risk variants (for improved model, see end of section below).

4.4 *HNF1A* variant effect on transcriptional activity and disease risk prediction

In order to evaluate transcriptional activity of *HNF1A* variants alone without interfering with any contribution from endogenous HNF-1A protein, the HeLa cell line is commonly used as model in an HNF-1A linked luciferase reporter assay. Furthermore, as HNF-1A target and gene promoter the rat *albumin* promoter is also commonly used, as it shows strong regulation by HNF-1A and is frequently used for read-out of *HNF1A* variant effect, in combination with analyses of other relevant HNF-1A gene promoters (*INS*, *GLUT2* and *HNF4AP2*). In the functional evaluation of the 27 rare *HNF1A* variants in **Paper II**, the variants were subdivided based on their effect on transcriptional activity, assessing regulation of the rat *albumin* promoter in HeLa cells, by different thresholds of their activity level (various levels ranging from <80% to <40%). Level of transcriptional activity <60% compared to wild-type HNF-1A (100%) showed the strongest association with T2D in the cohorts representing the general population ($P=0.0007$) [114]. With

this model, a total of 11 *HNF1A* variants were marked as impaired (<60% activity). Variants affecting HNF-1A transactivation to a lesser extent than this (<60%) were not equally significant (association with T2D), and most likely due to the few number of carriers with these variants (**Paper II**). Nine of the 11 functionally impaired variants (<60% activity) are located in transactivation domain and may thus explain their reducing effect on HNF-1A activity. Furthermore, four variants among the 11 transactivation impaired variants showed significantly lower protein level (p.(Y322C), p.(E508K), p.(H514R) and p.(T515K)), indicating that loss in HNF-1A protein may also explain loss in total cellular measured activity. Further analysis of three variants located in the DNA binding domain confirmed that one variant p.(E275del) significantly reduced the DNA binding of HNF-1A, while two variants p.(R131Q) and p.(V103M) reduced DNA binding, but not to a significant level, which may explain their loss in transactivation potential (**Paper II**). These DNA binding defective variants may mediate their effect on HNF-1A directly (loss of DNA binding residue), or indirectly through conformational changes within the protein, which leads to some degree of loss of ability to bind to target DNA [78]. The variant positively driving most of the association in **Paper II** is p.(E508K). This was shown in the replication analysis in the Type2 Diabetes Knowledge Portal (type2diabetesgenetics.org) publicly available dataset, where the T2D associated *P*-value increased to non-significant after excluding this Mexican risk variant in the association test.

Worth noting regarding our outlined T2D disease prediction models (**Paper II**) is that by adding the best (most T2D significant) model of nuclear localization (threshold of NL<60%) to the most significant model by transactivation (threshold TA<60%), the combined model (NL<60% + TA<60%) did not increase association of impaired variants with T2D. Furthermore, this combined model did not change the list of impaired variants when comparing to the impaired variants extracted from the transactivation model alone. Our end conclusion was therefore that the

transactivation assay itself could discriminate between functionally impaired variants associated with T2D risk versus benign variants, and providing a robust single assay for a functional assessment of *HNFI1A* variants.

4.5 Genotype-phenotype correlation by *HNFI1A* and treatment options

With increasingly effective sequencing methods and the growth in the number of identified rare *HNFI1A* variants with unknown significance, functional assays have shown to be essential to best determine the true pathogenicity of such variants. The ability to instantaneously interpret the function of any *HNFI1A* missense variant would enable the use of genetic information and assess diabetes risk, predict disease biomarkers, and perhaps offer more precise treatment for risk/disease causing variant carriers. Certain diabetes patients (MODY3) carrying severe pathogenic variants in *HNFI1A* have shown hypersensitivity to sulfonylureas treatment [119, 120]. This response is also greater compared with T2D patients. In a study, MODY3 patients showed 5.2-fold greater response to sulfonylurea than to metformin, as well as a 3.9-fold greater response to sulfonylurea compared to T2D patients [113]. Hence, classification of the genetic cause of a patient's hyperglycemia provides a significant implication for precise treatment. However, pharmacodynamic efficacy to sulfonylurea varies in T2D patients, presumably due to other genetic and environmental factors [121, 122].

Since 2% of the Mexican and Latin American T2D population carries the *HNFI1A* p.(E508K) variant (**Paper I**), it is hypothesized that these individuals could benefit from treatment modification by sulfonylurea. To test this hypothesis a study has been conducted involving researchers at the Broad Institute of Harvard and MIT in Cambridge, USA. The result of this study, which may have been underpowered, was that it was negative for the hypothesized increased sensitivity to sulfonylureas in p.(E508K) carriers in the Mexican population (personal communication,

Martagón et al., unpublished data). A study with a more sizable number of this, and other T2D *HNF1A* risk variants carriers, is most likely needed to provide a higher resolution for determining carrier sensitivity to sulfonylureas.

4.6 Systematic search for HNF-1A regulated transcripts for developing *HNF1A* high throughput assay

Variants analysis is currently a big challenge and many interpreted variants do not have strong evidence to be classified as benign or pathogenic, and are therefore categorized as VUSs. These variants are not actionable and cannot be used in guiding diagnosis or disease management. Of the 4.6 million missense variants identified in ~140,000 exomes and genomes in the GnomAD dataset (<http://gnomad.broadinstitute.org>), 99% of these variants are rare and with a MAF < 0.5% (**Figure 6**). Although the majority of these variants occur in genes previously related to different diseases, only 2% have a clinical interpretation in the ClinVar database (<https://www.ncbi.nlm.nih.gov/clinvar/>).

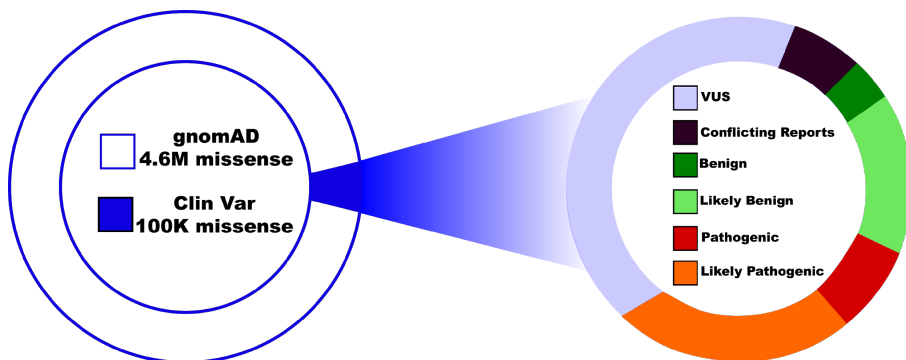


Figure 6. Missense variants discovery predicted clinical effect. The Genome Aggregation Database (gnomAD) has reported 4.6 million missense variants (left figure), among them only 100K have clinical interpretation in ClinVar (right figure). Data used for

this plot was taken from gnomED (February 28,2017) and ClinVar (April 5, 2017) This figure is adapted and modified from [123].

To illustrate the recent burst in variant numbers identified in *HNFI1A*: As reported in one of our studies (**Paper II**), around 0.4% of the general population carry impaired variants in *HNFI1A* that could increase the risk of T2D up to 6-fold. A more recent study and unpublished data based on sequencing multi-ethnic cohorts of 13,000 T2D cases/controls (T2D knowledge portal T2D-GENES and Go T2D consortia) have additionally identified around 80 missense variants in *HNFI1A* [124]. Furthermore, the genome aggregation database (gnomAD) has reported 472 missense variants in *HNFI1A* from sequencing of 123,136 exomes and 15,496 genomes of unrelated individuals from various disease populations [125]. This has thus in the last few years resulted in a large list of rare and novel *HNFI1A* variants with unknown functional and clinical consequence.

An ideal strategy in variant evaluation is to improve the efficiency of the functional assessment in well-validated assays. However, current functional assays are low-throughput, laborious and too slow to fulfill the demand (“one-by-one” approach). This challenge emphasizes the essential need for developing high-throughput functional assays in a comprehensive and systematic fashion. Multiplexed assays for investigating variants effect (MAVEs) measure the function of a comprehensive number of variants (ideally all possible variants in a gene) in a single experiment. MAVEs directly link the genotype of each variant to its effect in well-validated assays. Common framework in MAVEs is to synthesize variants, introduce them into an assay system, and select for a phenotype of interest [123]. MAVEs can then be used to prospectively produce lookup tables of variants functional effects for accurate pathogenicity prediction in disease related genes. The MAVEs strategy was recently used to comprehensively measure the effects of all possible missense variants of protein coding in *PPARG* [126].

Based on the relevance of *HNF1A* gene variants in HNF-1A-MODY as well as T2D (**Paper I and II**), and on the increasing number of HNF1A variants identified recently by NGS strategies, this gene and protein is an ideal candidate for MAVE analysis. Prior to this, a screening of robust and reliable targets for HNF-1A regulation is needed in order to distinguish levels of pathogenic effect induced by different *HNF1A* variants. This strategy encouraged us to pursue the search for other HNF-1A regulated genes and presented in **paper III** (manuscript). Here, we systematically searched for novel and strongly regulated HNF-1A transcripts that could represent more ideal and precise read-outs of HNF-1A dysfunction, and furthermore, to be compatible for a pooled *HNF1A* allelic screening in a high throughput (MAVE) assay, in the future.

In this study, we assessed the full complexity of the HNF-1A transcriptional response, in liver cell models, to find multiple endogenous target genes with altered expression. Ideally, such investigations should have been performed in a beta-cell model, but due to the lack of a robust human beta-cell line, the physiologically relevant human liver cell lines; HuH7 and Hep3B cell lines were selected. In order to be able to study *HNF1A* variants effects in these cell lines, without the interference of endogenous HNF-1A, we firstly generated *HNF1A*- free cell lines by knocking out endogenous HNF-1A using CRISPR/Cas9 (**Paper III**). To boost the HNF-1A regulated transcripts in the *HNF1A*- free cell lines, recoded *HNF1A* (resistance to CRISPR/Cas9) was re-introduced into the knock out cell lines at 5 µg/ml doxycycline doses (induce HNF-1A expression), prior to deciphering the landscape of differentially expressed transcripts by RNA sequencing.

A total of 62 genes in HuH7 and 367 genes in Hep3B cell lines were significantly upregulated (FDR < 0.01 and log₂ FC ≥ 2). In order to select suitable transcript as

markers for HNF-1A dysfunction, we further analyzed the expression of the top 20 significantly upregulated genes also in HuH7 and HepB3 cell lines transfected with carefully selected *HNF1A* variants known to cause HNF-1A MODY or be strongly associated with T2D (p.(P112L),p.(P447L) and p.(E508K), respectively). The expression of the majority of these 20 genes were significantly down-regulated in MODY3 variants conditions, whereas most were upregulated or did not show significant differential expression in mild-loss of function variant p.(E508K) condition. In order to further select relevant candidate targets by considering their ability in their discrimination of different diabetes-associated *HNF1A* variants, as well as sufficient level of expression, the genes *ABCC2*, *FABP1* and *HABP2* genes in HuH7, and *HKDC1*, *HRG* and *KL* in Hep3B cell lines, were suggested in **Paper III** as potential targets for readout of *HNF1A* variant effect in a future MAVE.

5. CONCLUDING REMARKS AND FUTURE PERSPECTIVE

Variants in *HNFI A* are associated with different diabetes phenotypes and can be identified in subjects with MODY3, T2D, and T1D. Phenotypic features range from severe loss-of-function in MODY3, to mild loss of function in late-onset diabetes. By this, there is no direct link between a specific sequence variant in *HNFI A* and causing phenotype, as variants exist in both Mendelian and complex forms of diabetes. Historically, T2D affects individuals in middle age or later, but this trend is changing as prevalence of T2D has increased in children and adolescents over the years. Moreover, the diversity of the clinical presentation of MODY3, and overlapping clinical features of T2D, makes the distinguishing of these two types complicated.

The overall goal of our studies (**Papers I, II and III**) was to investigate whether novel, rare variants in *HNFI A* could present as risk factors for T2D, and search for novel cell-based markers that can represent as read-out for massively parallel allelic screening of *HNFI A* variants in the future high throughput assay (MAVE). Variants investigated in Paper **I** and **II**, based on a “one-at-a-time” analysis approach, clearly distinguished the severe MODY3 variants from the (impaired), mild loss-of-function T2D variants, and the latter from the more benign variants, based on an activity model (threshold <60%) and loss of transcriptional activity induced by *HNFI A* variants.

Generally, the functional significance of rare variants identified in large-scale population sequencing lack sufficient statistical power because they are mostly observed in one or few individuals, unless variant prevalence is as high (2%) as the rare variant p.(E508K) identified in the Mexican and Latin American population, with strong odd ratios of 5.48 (95% CI 2.83-10.61) with T2D, and with a functionally confirmed mild loss-of-function effect. For more rare and low

frequency variants the phenotypic effect cannot be achieved by association tests. Thus, the ability to discriminate between milder causal versus benign sequence variants, to increase the power of sequencing studies, is more difficult for assessing contribution in complex disease.

Most of the speculation of a missing heritability of complex disease has focused on the contribution of rare and low frequency variants in complex disease. Next-generation sequencing is a commonly used sequencing approach now and has revolutionized the amount of genetic information available and increased our knowledge of human genome variation. However, the vast majority of coding variants are rare, novel, and with unknown functional consequence. The key to interpreting these variants for clinical use depends on their functional evaluation and preferably by a MAVEs approach that is less time consuming than a “one-by-one” variant analysis approach. This need motivated us for establishing new and improved cell models for a future high throughput MAVE assay for functional characterization of every possible *HNF1A* variant. Based on the candidate target(s) identified in Paper III, an ideal functional assay system should be chosen. For this, factors including the cellular localization of selected target(s) and target protein expression detection by flow cytometry (antibody staining) or FISH-Flow (mRNA staining) must be taken into careful consideration. Our study (**Paper III**), however, represent a good basis for the subsequent development of such a high throughput disease-relevant gene function assays. The outcome of such an assay, if successful, is a lookup table of variant effect, which can ultimately guide clinicians in diabetes diagnosis of current and future patients.

6. References

1. Szablewski, L., *Glucose Homeostasis – Mechanism and Defects*, in *Diabetes – Damages and Treatments*, E.C. Rigobelo, Editor. 2011.
2. Gerich, M.Z.S.E., *Normal Glucose Homeostasis*, in *Principles of Diabetes Mellitus*. 2010.
3. Roder, P.V., et al., *Pancreatic regulation of glucose homeostasis*. *Exp Mol Med*, 2016. **48**: p. e219.
4. Brissova, M., et al., *Assessment of human pancreatic islet architecture and composition by laser scanning confocal microscopy*. *J Histochem Cytochem*, 2005. **53**(9): p. 1087-97.
5. Katsuura, G., A. Asakawa, and A. Inui, *Roles of pancreatic polypeptide in regulation of food intake*. *Peptides*, 2002. **23**(2): p. 323-9.
6. Wierup, N., et al., *The ghrelin cell: a novel developmentally regulated islet cell in the human pancreas*. *Regul Pept*, 2002. **107**(1-3): p. 63-9.
7. Longnecker, D.S., *Anatomy and Histology of the Pancreas*, in *Pancreapedia: Exocrine Pancreas Knowledge Base*. 2014, Pancreapedia.
8. Meda, P. and F. Schuit, *Glucose-stimulated insulin secretion: the hierarchy of its multiple cellular and subcellular mechanisms*. *Diabetologia*, 2013. **56**(12): p. 2552-5.
9. Wilcox, G., *Insulin and insulin resistance*. *Clin Biochem Rev*, 2005. **26**(2): p. 19-39.
10. Fajans, S.S., G.I. Bell, and K.S. Polonsky, *Molecular mechanisms and clinical pathophysiology of maturity-onset diabetes of the young*. *N Engl J Med*, 2001. **345**(13): p. 971-80.
11. Florez, G.A.W.J.C., *Type 2 Diabetes and Genetics, 2010: Translating Knowledge into Understanding*. *Curr Cardio Risk Rep*, 2010. **4**: p. 437-445.
12. Shaw, J.E., R.A. Sicree, and P.Z. Zimmet, *Global estimates of the prevalence of diabetes for 2010 and 2030*. *Diabetes Res Clin Pract*, 2010. **87**(1): p. 4-14.
13. Ogurtsova, K., et al., *IDF Diabetes Atlas: Global estimates for the prevalence of diabetes for 2015 and 2040*. *Diabetes Res Clin Pract*, 2017. **128**: p. 40-50.
14. American Diabetes, A., *2. Classification and Diagnosis of Diabetes*. *Diabetes Care*, 2017. **40**(Suppl 1): p. S11-S24.
15. Organization, W.H., *Definition and diagnosis of diabetes mellitus and intermediate hyperglycaemia*, in *Definition and diagnosis of diabetes mellitus and intermediate hyperglycaemia*. 2006, WHO: WHO.
16. Sreenan, S., et al., *Increased beta-cell proliferation and reduced mass before diabetes onset in the nonobese diabetic mouse*. *Diabetes*, 1999. **48**(5): p. 989-96.
17. Meier, J.J., *Beta cell mass in diabetes: a realistic therapeutic target?* *Diabetologia*, 2008. **51**(5): p. 703-13.

18. Karamitsos, D.T., *The story of insulin discovery*. Diabetes Res Clin Pract, 2011. **93 Suppl 1**: p. S2-8.
19. Haller, M.J., M.A. Atkinson, and D. Schatz, *Type 1 diabetes mellitus: etiology, presentation, and management*. Pediatr Clin North Am, 2005. **52**(6): p. 1553-78.
20. Stankov, K., D. Benc, and D. Draskovic, *Genetic and epigenetic factors in etiology of diabetes mellitus type 1*. Pediatrics, 2013. **132**(6): p. 1112-22.
21. Mehers, K.L. and K.M. Gillespie, *The genetic basis for type 1 diabetes*. Br Med Bull, 2008. **88**(1): p. 115-29.
22. Pociot, F. and M.F. McDermott, *Genetics of type 1 diabetes mellitus*. Genes Immun, 2002. **3**(5): p. 235-49.
23. Steck, A.K. and M.J. Rewers, *Genetics of type 1 diabetes*. Clin Chem, 2011. **57**(2): p. 176-85.
24. Zheng, Y., S.H. Ley, and F.B. Hu, *Global aetiology and epidemiology of type 2 diabetes mellitus and its complications*. Nat Rev Endocrinol, 2018. **14**(2): p. 88-98.
25. Zimmet, P., K.G. Alberti, and J. Shaw, *Global and societal implications of the diabetes epidemic*. Nature, 2001. **414**(6865): p. 782-7.
26. Rother, K.I., *Diabetes treatment--bridging the divide*. N Engl J Med, 2007. **356**(15): p. 1499-501.
27. Das, S.K. and S.C. Elbein, *The Genetic Basis of Type 2 Diabetes*. Cellscience, 2006. **2**(4): p. 100-131.
28. Prasad, R.B. and L. Groop, *Genetics of type 2 diabetes-pitfalls and possibilities*. Genes (Basel), 2015. **6**(1): p. 87-123.
29. Florez, J.C., *Newly identified loci highlight beta cell dysfunction as a key cause of type 2 diabetes: where are the insulin resistance genes?* Diabetologia, 2008. **51**(7): p. 1100-10.
30. Mpondo, B.C., A. Ernest, and H.E. Dee, *Gestational diabetes mellitus: challenges in diagnosis and management*. J Diabetes Metab Disord, 2015. **14**: p. 42.
31. Cheung, N.W. and D. Helmink, *Gestational diabetes: the significance of persistent fasting hyperglycemia for the subsequent development of diabetes mellitus*. J Diabetes Complications, 2006. **20**(1): p. 21-5.
32. Robitaille, J. and A.M. Grant, *The genetics of gestational diabetes mellitus: evidence for relationship with type 2 diabetes mellitus*. Genet Med, 2008. **10**(4): p. 240-50.
33. Kleinberger, J.W., K.A. Maloney, and T.I. Pollin, *The Genetic Architecture of Diabetes in Pregnancy: Implications for Clinical Practice*. Am J Perinatol, 2016. **33**(13): p. 1319-1326.
34. Cho, Y.M., et al., *Type 2 diabetes-associated genetic variants discovered in the recent genome-wide association studies are related to gestational diabetes mellitus in the Korean population*. Diabetologia, 2009. **52**(2): p. 253-61.

35. Kwak, S.H., H.C. Jang, and K.S. Park, *Finding genetic risk factors of gestational diabetes*. Genomics Inform, 2012. **10**(4): p. 239-43.
36. Molven, A. and P.R. Njolstad, *Role of molecular genetics in transforming diagnosis of diabetes mellitus*. Expert Rev Mol Diagn, 2011. **11**(3): p. 313-20.
37. Kim, S.H., *Maturity-Onset Diabetes of the Young: What Do Clinicians Need to Know?* Diabetes Metab J, 2015. **39**(6): p. 468-77.
38. Bonnefond, A., et al., *Whole-exome sequencing and high throughput genotyping identified KCNJ11 as the thirteenth MODY gene*. PLoS One, 2012. **7**(6): p. e37423.
39. Amed, S. and R. Oram, *Maturity-Onset Diabetes of the Young (MODY): Making the Right Diagnosis to Optimize Treatment*. Can J Diabetes, 2016. **40**(5): p. 449-454.
40. Thanabalasingham, G. and K.R. Owen, *Diagnosis and management of maturity onset diabetes of the young (MODY)*. BMJ, 2011. **343**: p. d6044.
41. Yamagata, K., et al., *Mutations in the hepatocyte nuclear factor-4alpha gene in maturity-onset diabetes of the young (MODY1)*. Nature, 1996. **384**(6608): p. 458-60.
42. Froguel, P., et al., *Familial hyperglycemia due to mutations in glucokinase. Definition of a subtype of diabetes mellitus*. N Engl J Med, 1993. **328**(10): p. 697-702.
43. Vaxillaire, M., et al., *A gene for maturity onset diabetes of the young (MODY) maps to chromosome 12q*. Nat Genet, 1995. **9**(4): p. 418-23.
44. Stoffers, D.A., et al., *Early-onset type-II diabetes mellitus (MODY4) linked to IPFI*. Nat Genet, 1997. **17**(2): p. 138-9.
45. Horikawa, Y., et al., *Mutation in hepatocyte nuclear factor-1 beta gene (TCF2) associated with MODY*. Nat Genet, 1997. **17**(4): p. 384-5.
46. Malecki, M.T., et al., *Mutations in NEUROD1 are associated with the development of type 2 diabetes mellitus*. Nat Genet, 1999. **23**(3): p. 323-8.
47. Neve, B., et al., *Role of transcription factor KLF11 and its diabetes-associated gene variants in pancreatic beta cell function*. Proc Natl Acad Sci U S A, 2005. **102**(13): p. 4807-12.
48. Raeder, H., et al., *Mutations in the CEL VNTR cause a syndrome of diabetes and pancreatic exocrine dysfunction*. Nat Genet, 2006. **38**(1): p. 54-62.
49. Plengvidhya, N., et al., *PAX4 mutations in Thais with maturity onset diabetes of the young*. J Clin Endocrinol Metab, 2007. **92**(7): p. 2821-6.
50. Edghill, E.L., et al., *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, 2008. **57**(4): p. 1034-42.
51. Kim, S.H., et al., *Identification of a locus for maturity-onset diabetes of the young on chromosome 8p23*. Diabetes, 2004. **53**(5): p. 1375-84.

52. Patch, A.M., et al., *Mutations in the ABCC8 gene encoding the SUR1 subunit of the KATP channel cause transient neonatal diabetes, permanent neonatal diabetes or permanent diabetes diagnosed outside the neonatal period.* Diabetes Obes Metab, 2007. **9 Suppl 2**: p. 28-39.
53. Flanagan, S.E., et al., *Mutations in ATP-sensitive K⁺ channel genes cause transient neonatal diabetes and permanent diabetes in childhood or adulthood.* Diabetes, 2007. **56**(7): p. 1930-7.
54. Vaxillaire, M. and P. Froguel, *Monogenic diabetes in the young, pharmacogenetics and relevance to multifactorial forms of type 2 diabetes.* Endocr Rev, 2008. **29**(3): p. 254-64.
55. McDonald, T.J. and S. Ellard, *Maturity onset diabetes of the young: identification and diagnosis.* Ann Clin Biochem, 2013. **50**(Pt 5): p. 403-15.
56. Wisely, G.B., et al., *Hepatocyte nuclear factor 4 is a transcription factor that constitutively binds fatty acids.* Structure, 2002. **10**(9): p. 1225-34.
57. Bogan, A.A., et al., *Analysis of protein dimerization and ligand binding of orphan receptor HNF4alpha.* J Mol Biol, 2000. **302**(4): p. 831-51.
58. Huang, J., L.L. Levitsky, and D.B. Rhoads, *Novel P2 promoter-derived HNF4alpha isoforms with different N-terminus generated by alternate exon insertion.* Exp Cell Res, 2009. **315**(7): p. 1200-11.
59. Gloyn, A.L., *Glucokinase (GCK) mutations in hyper- and hypoglycemia: maturity-onset diabetes of the young, permanent neonatal diabetes, and hyperinsulinemia of infancy.* Hum Mutat, 2003. **22**(5): p. 353-62.
60. Njolstad, P.R., et al., *Permanent neonatal diabetes caused by glucokinase deficiency: inborn error of the glucose-insulin signaling pathway.* Diabetes, 2003. **52**(11): p. 2854-60.
61. Sagen, J.V., et al., *Diagnostic screening of MODY2/GCK mutations in the Norwegian MODY Registry.* Pediatr Diabetes, 2008. **9**(5): p. 442-9.
62. Frain, M., et al., *The liver-specific transcription factor LF-B1 contains a highly diverged homeobox DNA binding domain.* Cell, 1989. **59**(1): p. 145-57.
63. Miquerol, L., et al., *Expression of the L-type pyruvate kinase gene and the hepatocyte nuclear factor 4 transcription factor in exocrine and endocrine pancreas.* J Biol Chem, 1994. **269**(12): p. 8944-51.
64. Baumhueter, S., et al., *HNF-1 shares three sequence motifs with the POU domain proteins and is identical to LF-B1 and APF.* Genes Dev, 1990. **4**(3): p. 372-9.
65. Yamagata, K., *Regulation of pancreatic beta-cell function by the HNF transcription network: lessons from maturity-onset diabetes of the young (MODY).* Endocr J, 2003. **50**(5): p. 491-9.
66. Colclough, K., et al., *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, 2013. **34**(5): p. 669-85.

67. Stenson, P.D., et al., *Human Gene Mutation Database (HGMD): 2003 update*. Hum Mutat, 2003. **21**(6): p. 577-81.
68. Frayling, T.M., et al., *beta-cell genes and diabetes: molecular and clinical characterization of mutations in transcription factors*. Diabetes, 2001. **50 Suppl 1**: p. S94-100.
69. Bellanne-Chantelot, C., et al., *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, 2011. **96**(8): p. E1346-51.
70. Harries, L.W., et al., *Isomers of the TCF1 gene encoding hepatocyte nuclear factor-1 alpha show differential expression in the pancreas and define the relationship between mutation position and clinical phenotype in monogenic diabetes*. Hum Mol Genet, 2006. **15**(14): p. 2216-24.
71. Bach, I., M. Pontoglio, and M. Yaniv, *Structure of the gene encoding hepatocyte nuclear factor 1 (HNF1)*. Nucleic Acids Res, 1992. **20**(16): p. 4199-204.
72. Mendel, D.B. and G.R. Crabtree, *HNF-1, a member of a novel class of dimerizing homeodomain proteins*. J Biol Chem, 1991. **266**(2): p. 677-80.
73. Marfori, M., et al., *Molecular basis for specificity of nuclear import and prediction of nuclear localization*. Biochim Biophys Acta, 2011. **1813**(9): p. 1562-77.
74. Nachury, M.V. and K. Weis, *The direction of transport through the nuclear pore can be inverted*. Proc Natl Acad Sci U S A, 1999. **96**(17): p. 9622-7.
75. Hessabi, B., et al., *The nuclear localization signal (NLS) of PDX-1 is part of the homeodomain and represents a novel type of NLS*. Eur J Biochem, 1999. **263**(1): p. 170-7.
76. Sock, E., et al., *Identification of the nuclear localization signal of the POU domain protein Tst-1/Oct6*. J Biol Chem, 1996. **271**(29): p. 17512-8.
77. Bjorkhaug, L., et al., *Functional dissection of the HNF-1alpha transcription factor: a study on nuclear localization and transcriptional activation*. DNA Cell Biol, 2005. **24**(11): p. 661-9.
78. Rose, R.B., et al., *Structural basis of dimerization, coactivator recognition and MODY3 mutations in HNF-1alpha*. Nat Struct Biol, 2000. **7**(9): p. 744-8.
79. Narayana, N., Q. Hua, and M.A. Weiss, *The dimerization domain of HNF-1alpha: structure and plasticity of an intertwined four-helix bundle with application to diabetes mellitus*. J Mol Biol, 2001. **310**(3): p. 635-58.
80. Cereghini, S., *Liver-enriched transcription factors and hepatocyte differentiation*. FASEB J, 1996. **10**(2): p. 267-82.
81. Chi, Y.I., et al., *Diabetes mutations delineate an atypical POU domain in HNF-1alpha*. Mol Cell, 2002. **10**(5): p. 1129-37.
82. Sneha P., et al., *Determining the role of missense mutations in the POU domain of HNF1A that reduce the DNA-binding affinity: A computational approach*. PLoS One, 2017. **12**(4): p. e0174953.

83. Nicosia, A., et al., *A myosin-like dimerization helix and an extra-large homeodomain are essential elements of the tripartite DNA binding structure of LFB1*. Cell, 1990. **61**(7): p. 1225-36.
84. Toniatti, C., et al., *A bipartite activation domain is responsible for the activity of transcription factor HNF1/LFB1 in cells of hepatic and nonhepatic origin*. DNA Cell Biol, 1993. **12**(3): p. 199-208.
85. Calkhoven, C.F. and G. Ab, *Multiple steps in the regulation of transcription-factor level and activity*. Biochem J, 1996. **317** (Pt 2): p. 329-42.
86. Yu, M., et al., *Proteomic screen defines the hepatocyte nuclear factor 1alpha-binding partners and identifies HMGB1 as a new cofactor of HNF1alpha*. Nucleic Acids Res, 2008. **36**(4): p. 1209-19.
87. Soutoglou, E., et al., *Transcriptional activation by hepatocyte nuclear factor-1 requires synergism between multiple coactivator proteins*. J Biol Chem, 2000. **275**(17): p. 12515-20.
88. Odom, D.T., et al., *Control of pancreas and liver gene expression by HNF transcription factors*. Science, 2004. **303**(5662): p. 1378-81.
89. Mitchell, S.M. and T.M. Frayling, *The role of transcription factors in maturity-onset diabetes of the young*. Mol Genet Metab, 2002. **77**(1-2): p. 35-43.
90. Cerf, M.E., *Transcription factors regulating beta-cell function*. Eur J Endocrinol, 2006. **155**(5): p. 671-9.
91. Altshuler, D., et al., *The common PPARgamma Pro12Ala polymorphism is associated with decreased risk of type 2 diabetes*. Nat Genet, 2000. **26**(1): p. 76-80.
92. Gloyn, A.L., et al., *Large-scale association studies of variants in genes encoding the pancreatic beta-cell KATP channel subunits Kir6.2 (KCNJ11) and SUR1 (ABCC8) confirm that the KCNJ11 E23K variant is associated with type 2 diabetes*. Diabetes, 2003. **52**(2): p. 568-72.
93. Florez, J.C., *The new type 2 diabetes gene TCF7L2*. Curr Opin Clin Nutr Metab Care, 2007. **10**(4): p. 391-6.
94. Majithia, A.R. and J.C. Florez, *Clinical translation of genetic predictors for type 2 diabetes*. Curr Opin Endocrinol Diabetes Obes, 2009. **16**(2): p. 100-6.
95. Prokopenko, I., M.I. McCarthy, and C.M. Lindgren, *Type 2 diabetes: new genes, new understanding*. Trends Genet, 2008. **24**(12): p. 613-21.
96. Tallapragada, D.S., S. Bhaskar, and G.R. Chandak, *New insights from monogenic diabetes for "common" type 2 diabetes*. Front Genet, 2015. **6**: p. 251.
97. Flannick, J. and J.C. Florez, *Type 2 diabetes: genetic data sharing to advance complex disease research*. Nat Rev Genet, 2016. **17**(9): p. 535-49.
98. Manolio, T.A., et al., *Finding the missing heritability of complex diseases*. Nature, 2009. **461**(7265): p. 747-53.


99. Huyghe, J.R., et al., *Exome array analysis identifies new loci and low-frequency variants influencing insulin processing and secretion*. Nat Genet, 2013. **45**(2): p. 197-201.
100. Steinthorsdottir, V., et al., *Identification of low-frequency and rare sequence variants associated with elevated or reduced risk of type 2 diabetes*. Nat Genet, 2014. **46**(3): p. 294-8.
101. Bomba, L., K. Walter, and N. Soranzo, *The impact of rare and low-frequency genetic variants in common disease*. Genome Biol, 2017. **18**(1): p. 77.
102. Auer, P.L. and G. Lettre, *Rare variant association studies: considerations, challenges and opportunities*. Genome Med, 2015. **7**(1): p. 16.
103. Richards, S., et al., *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, 2015. **17**(5): p. 405-24.
104. Calabrese, R., et al., *Functional annotations improve the predictive score of human disease-related mutations in proteins*. Hum Mutat, 2009. **30**(8): p. 1237-44.
105. Thusberg, J., A. Olatubosun, and M. Vihinen, *Performance of mutation pathogenicity prediction methods on missense variants*. Hum Mutat, 2011. **32**(4): p. 358-68.
106. Amendola, L.M., et al., *Actionable exomic incidental findings in 6503 participants: challenges of variant classification*. Genome Res, 2015. **25**(3): p. 305-15.
107. Holmkvist, J., et al., *Common variants in HNF-1 alpha and risk of type 2 diabetes*. Diabetologia, 2006. **49**(12): p. 2882-91.
108. Tao Chen a, X.C.b., Yang Long c, Xiangxun Zhang c, Honglin Yu c, Jin Xu d, Ting Yu d, Haoming Tian a, *I27L Polymorphism in hepatocyte nuclear factor-1a gene and type 2 diabetes mellitus: A meta-analysis of studies about orient population (Chinese and Japanese)*. International Journal of Diabetes, 2010. **2**: p. 28-31.
109. Winckler, W., et al., *Association of common variation in the HNF1alpha gene region with risk of type 2 diabetes*. Diabetes, 2005. **54**(8): p. 2336-42.
110. Fu, J., E.A. Festen, and C. Wijmenga, *Multi-ethnic studies in complex traits*. Hum Mol Genet, 2011. **20**(R2): p. R206-13.
111. Hegele, R.A., et al., *The hepatic nuclear factor-1alpha G319S variant is associated with early-onset type 2 diabetes in Canadian Oji-Cree*. J Clin Endocrinol Metab, 1999. **84**(3): p. 1077-82.
112. Triggs-Raine, B.L., et al., *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, 2002. **99**(7): p. 4614-9.
113. Pearson, E.R., et al., *Genetic cause of hyperglycaemia and response to treatment in diabetes*. Lancet, 2003. **362**(9392): p. 1275-81.

114. Flannick, J., et al., *Assessing the phenotypic effects in the general population of rare variants in genes for a dominant Mendelian form of diabetes*. *Nat Genet*, 2013. **45**(11): p. 1380-5.
115. Sunyaev, S.R., *Inferring causality and functional significance of human coding DNA variants*. *Hum Mol Genet*, 2012. **21**(R1): p. R10-7.
116. Bjorkhaug, L., et al., *Hepatocyte nuclear factor-1 alpha gene mutations and diabetes in Norway*. *J Clin Endocrinol Metab*, 2003. **88**(2): p. 920-31.
117. Vaxillaire, M., et al., *Anatomy of a homeoprotein revealed by the analysis of human MODY3 mutations*. *J Biol Chem*, 1999. **274**(50): p. 35639-46.
118. Bjorkhaug, L., et al., *MODY associated with two novel hepatocyte nuclear factor-1alpha loss-of-function mutations (P112L and Q466X)*. *Biochem Biophys Res Commun*, 2000. **279**(3): p. 792-8.
119. Sovik, O., et al., *Hyperexcitability to sulphonylurea in MODY3*. *Diabetologia*, 1998. **41**(5): p. 607-8.
120. Sagen, J.V., et al., *Preserved insulin response to tolbutamide in hepatocyte nuclear factor-1alpha mutation carriers*. *Diabet Med*, 2005. **22**(4): p. 406-9.
121. Fukui, M., et al., *Antibodies to glutamic acid decarboxylase in Japanese diabetic patients with secondary failure of oral hypoglycaemic therapy*. *Diabet Med*, 1997. **14**(2): p. 148-52.
122. Levy, J., et al., *Beta-cell deterioration determines the onset and rate of progression of secondary dietary failure in type 2 diabetes mellitus: the 10-year follow-up of the Belfast Diet Study*. *Diabet Med*, 1998. **15**(4): p. 290-6.
123. Starita, L.M., et al., *Variant Interpretation: Functional Assays to the Rescue*. *Am J Hum Genet*, 2017. **101**(3): p. 315-325.
124. Portal, T.D.K., *Type 2 Diabetes Knowledge Portal*. 2017: Broad Institute.
125. Lek, M., et al., *Analysis of protein-coding genetic variation in 60,706 humans*. *Nature*, 2016. **536**(7616): p. 285-91.
126. Majithia, A.R., et al., *Prospective functional classification of all possible missense variants in PPARG*. *Nat Genet*, 2016. **48**(12): p. 1570-1575.

Original Investigation

Association of a Low-Frequency Variant in *HNF1A* With Type 2 Diabetes in a Latino Population

The SIGMA Type 2 Diabetes Consortium

 Supplemental content at jama.com

IMPORTANCE Latino populations have one of the highest prevalences of type 2 diabetes worldwide.

OBJECTIVES To investigate the association between rare protein-coding genetic variants and prevalence of type 2 diabetes in a large Latino population and to explore potential molecular and physiological mechanisms for the observed relationships.

DESIGN, SETTING, AND PARTICIPANTS Whole-exome sequencing was performed on DNA samples from 3756 Mexican and US Latino individuals (1794 with type 2 diabetes and 1962 without diabetes) recruited from 1993 to 2013. One variant was further tested for allele frequency and association with type 2 diabetes in large multiethnic data sets of 14 276 participants and characterized in experimental assays.

MAIN OUTCOME AND MEASURES Prevalence of type 2 diabetes. Secondary outcomes included age of onset, body mass index, and effect on protein function.

RESULTS A single rare missense variant (c.1522G>A [p.E508K]) was associated with type 2 diabetes prevalence (odds ratio [OR], 5.48; 95% CI, 2.83-10.61; $P = 4.4 \times 10^{-7}$) in hepatocyte nuclear factor 1- α (*HNF1A*), the gene responsible for maturity onset diabetes of the young type 3 (MODY3). This variant was observed in 0.36% of participants without type 2 diabetes and 2.1% of participants with it. In multiethnic replication data sets, the p.E508K variant was seen only in Latino patients ($n = 1443$ with type 2 diabetes and 1673 without it) and was associated with type 2 diabetes (OR, 4.16; 95% CI, 1.75-9.92; $P = .0013$). In experimental assays, HNF-1A protein encoding the p.E508K mutant demonstrated reduced transactivation activity of its target promoter compared with a wild-type protein. In our data, carriers and noncarriers of the p.E508K mutation with type 2 diabetes had no significant differences in compared clinical characteristics, including age at onset. The mean (SD) age for carriers was 45.3 years (11.2) vs 47.5 years (11.5) for noncarriers ($P = .49$) and the mean (SD) BMI for carriers was 28.2 (5.5) vs 29.3 (5.3) for noncarriers ($P = .19$).

CONCLUSIONS AND RELEVANCE Using whole-exome sequencing, we identified a single low-frequency variant in the MODY3-causing gene *HNF1A* that is associated with type 2 diabetes in Latino populations and may affect protein function. This finding may have implications for screening and therapeutic modification in this population, but additional studies are required.

JAMA. 2014;311(22):2305-2314. doi:10.1001/jama.2014.6511

The Authors and other collaborators of the SIGMA Type 2 Diabetes Consortium are listed at the end of this article.

Corresponding Author: Jose C. Florez, MD, PhD, Center for Human Genetic Research, Diabetes Unit, Department of Medicine, Massachusetts General Hospital, Boston, MA 02114 (jcflorez@partners.org).

The estimated prevalence of type 2 diabetes in Mexican adults was 14.4% in 2006,¹ making it one of the leading causes of death in Mexico.² Based on statistics from 1999-2002, the standardized prevalence of diagnosed diabetes was 10% in Mexican Americans and 5.2% in whites.³ Although environmental factors such as lifestyle and diet likely explain the majority of this health disparity, it was recently found that genetic variants in the gene *SLC16A11* (NCBI NC_000017.11) were associated with higher rates of type 2 diabetes in Latinos.⁴ Latinos, defined as persons who trace their origin to Central and South America, and other Spanish cultures, fall on a continuum of Native American and European genetic ancestry.⁴ Identifying genetic factors associated with type 2 diabetes in Latino populations could increase understanding of its pathophysiology, improve risk prediction, and focus treatment choice based on knowledge of the underlying biology of the disease.

Type 2 diabetes is typically diagnosed after age 40 years, is caused by the combined action of genetic susceptibility and environmental factors, is associated with obesity, and is polygenic. Genome-wide association studies for typical type 2 diabetes forms have identified more than 70 distinct genetic loci carrying common variants that are associated with modest differences in prevalence of the disease.⁵⁻⁷ Because these common variants explain a small fraction of the estimated heritability, it is hypothesized that low-frequency or rare variants of strong effects, not captured by genome-wide association studies but amenable to sequencing approaches, contribute in a meaningful proportion to the genetic architecture of the dis-

ease. To date, low-frequency variants with near-complete penetrance have not been found in whole-exome sequencing studies of type 2 diabetes,^{8,9} although a recent whole-genome sequencing study found rare variants associated with type 2 diabetes prevalence in an Icelandic population.¹⁰

To explore the association of rare protein-coding genetic variants with type 2 diabetes in the Latino population, we performed whole-exome sequencing (which captures both common and rare genetic variants in the protein-coding regions of genes) on case-control studies composed of individuals of Mexican or another Latino ancestry, with replication in a separate multiethnic data set.

Methods

Study Design and Patients

This study was performed as part of the Slim Initiative in Genomic Medicine for the Americas (SIGMA) Type 2 Diabetes Consortium, whose goal is to characterize the genetic basis of type 2 diabetes in Mexican and Latin American populations drawn from 4 studies^{4,11-13} (Table 1, details of these studies are provided in the Supplement). All participants had either Mexican or other Latino ancestry based on self-report and verification using principal component analysis of genotype data. Replication studies included individuals from a multiethnic study (Type 2 Diabetes Genetic Exploration by Next-Generation Sequencing in Multi-Ethnic Samples [T2D-GENES] and Genetics of T2D [GoT2D]) and an ongoing col-

Table 1. Characteristics of Cohorts Comprising the SIGMA Type 2 Diabetes Whole-Exome Sequence Project

Source	Sample Location	Study Design	No. of Participants	No (%) of Men	Mean (SD)			Proportion With Native American Ancestry		
					Age, y	Age of Onset, y	BMI			
UNAM/INCMNSZ Diabetes Study, ⁴ 2014	Mexico City, Mexico	Prospective cohort	Controls	539	206 (38.2)	55.0 (9.4)		28.4 (3.8)	86.4 (7.2)	0.75 (0.10)
			Type 2 diabetes	533	216 (40.5)	55.3 (12.5)	43.8 (11.2)	28.5 (4.4)		0.78 (0.11)
Diabetes in Mexico Study, ⁴ 2014	Mexico City, Mexico	Prospective cohort	Controls	459	119 (25.9)	52.4 (7.7)		28.0 (4.6)	90.1 (7.2)	0.67 (0.18)
			Type 2 diabetes	509	168 (33.0)	55.5 (11.1)	47.2 (10.6)	29.0 (5.4)		0.79 (0.12)
Mexico City Diabetes Study, ^{11,12} 2005 and 2011	Mexico City, Mexico	Prospective cohort	Controls	526	204 (38.8)	62.3 (7.5)		29.4 (4.8)	90.1 (9.0)	0.69 (0.14)
			Type 2 diabetes	270	110 (40.7)	64.0 (7.5)	55.0 (9.7)	29.9 (5.5)		0.67 (0.15)
Multiethnic Cohort, ¹ 2000	Los Angeles, California	Prospective cohort	Controls	438	212 (48.5)	59.3 (7.2)		26.9 (4.3)		0.53 (0.09)
			Type 2 diabetes	482	227 (47.0)	58.7 (7.2)	NA	29.8 (5.7)	NA	0.58 (0.08)
Overall SIGMA										
			Controls	1962	742 (37.8)	57.3 (8.9)		28.3 (4.5)	88.2 (9.0)	0.67 (0.15)
			Type 2 diabetes	1794	719 (40.1)	57.6 (10.6)	47.5 (11.5)	29.1 (5.2)		0.71 (0.15)

Abbreviations: BMI, body mass index, calculated as weight in kilograms divided by height in meters squared; NA, not available; UNAM/INCMNSZ, Instituto Nacional de Ciencias Médicas y Nutrición Salvador Zubirán, Universidad Nacional Autónoma de México.

SI conversion factor: To convert fasting glucose from mg/dL to mmol/L, multiply by 0.0555.

lection of Mexican participants from 18 indigenous groups for genetic studies (Diabetes in Mexico Study 2[DMS2]) (eTable 1, details of these studies are provided in Supplement). Diagnosis of type 2 diabetes followed the American Diabetes Association criteria. Each participant provided written informed consent for genetic investigation. All contributing studies were approved by their respective local ethics committees.

Genetic Studies

Sample Selection and Whole-Exome Sequencing

In total, 3862 samples were selected for whole-exome sequencing from a larger data set of 8214 samples previously genotyped with the OMNI 2.5 array (Illumina).⁴ To increase representation of genetic variation not queried in studies of European populations, selection criteria for whole-exome sequencing was based on the proportion of Native American ancestry estimated from principal component analysis of genotype data (eMethods section and eFigures 1 and 2 in the Supplement). Whole-exome sequencing was performed on blood DNA from these samples using Sure-Select Human All Exon v2.0 (Illumina), 44-Mb-baited target. Raw reads were mapped with the Burrows-Wheeler Aligner, reprocessed with Picard to recalibrate base quality scores and perform local realignment around known indels. Genetic variants were called with the Genome Analysis Toolkit Unified Genotyper module¹⁴ and were filtered to remove likely artifacts using several quality-control metrics such as mean coverage, concordance of nonreference genotypes with array data, and missing rate as specified in the eMethods section in the Supplement. Independent replication was sought in whole-exome sequence data from the T2D-GENES and GoT2D projects, which together sequenced 13 098 individuals from 5 ethnic groups (Europeans, East Asians, African Americans, South Asians, and Latinos).

Statistical Analyses

We used the liability threshold model, which models participants as having an unobserved continuous phenotype called liability.¹⁵ We computed the residual value of the liability after accounting for the part that can be predicted by each participant's age and body mass index (BMI) using LTSOFT software (<http://www.hsph.harvard.edu/alkes-price/software>).¹⁶ Significance was evaluated with the residual liabilities as outcome using an expedited mixed linear model,¹⁷ which adjusts for sex, ancestry (eFigure 3 in the Supplement), and relatedness via a variance-component matrix with 2-sided tests. Odds ratios (ORs) were estimated using logistic regression models on type 2 diabetes status adjusting for age, BMI, and ancestry as specified in the eMethods section in the Supplement. The experiment-wide statistical significance threshold was set to $P < 5 \times 10^{-8}$ to adjust for the number of variants evaluated. In addition to single-variant testing, the sequence kernel association test¹⁸ and collapsing tests¹⁹ were used to test the possibility of genes and groups of genes associated to disease susceptibility via aggregation of rare variants.

Results of all functional experiments are expressed as means (SDs), and experiments were performed on at least 3 independent occasions unless otherwise specified. Statistical analyses were performed using the 2-tailed *t* test, and $P < .05$ was considered significant for these functional studies.

Functional Studies

Plasmids, Cell Culture, and Transfections

Details of functional studies are specified in the eMethods section in the Supplement. The human liver hepatocyte nuclear factor 1a (*HNF1A*) complementary DNA in expression vector pcDNA3.1/HisC (NCBI Entrez Gene BC104910.1) was used for all cell studies.²⁰ Firefly luciferase reporter vectors (pGL3) included promoter sequences for the rat albumin (pGL3-RA), human *HNF4A* (NCBI Entrez Gene 3172) P2 (pGL3-HNF4AP2), and mouse *Glut2* (pGL3-GLUT2) genes. Renilla luciferase reporter construct pRL-SV40 (GenBank AF025845.2) was used as an internal control. The HNF-1A mutants were made using the QuikChange Site-Directed XL Mutagenesis Kit (Stratagene). HeLa cells and MIN6 β -cells were grown as previously described,^{20,21} and transfected according to manufacturers' recommendations using the Metafectene Pro (Biontex-USA) or Lipofectamine 2000 (Life Technologies), respectively.

Transactivation and Protein Expression Analyses

Transcriptional activity was measured 24 hours after transfection using the Dual-Luciferase Reporter Assay System (Promega Biotech) on a Chameleon luminometer (Hidex). To measure HNF-1A protein levels, transfected HeLa cells were lysed in passive lysis buffer (Promega Biotech) and proteins were analyzed (from 2.5 μ g of total protein) by SDS-PAGE and immunoblotting using an HNF-1A-tag (anti-Xpress antibody, Life Technologies).

DNA Binding Studies

The HNF-1A protein was produced in a coupled in vitro transcription/translation System (TnT-T7, Promega Biotech). The level of binding of HNF-1A proteins to radiolabeled rat albumin oligonucleotide was investigated by electrophoretic mobility shift assays as previously described.²²

Immunofluorescence

Analysis of nuclear vs cytosol localization of HNF-1A proteins was performed in 500 cells using an HNF-1A-tag (anti-Xpress antibody) and Alexa Fluor 488 (Life Technologies) essentially as reported previously.²⁰

Results

Study Participants

Demographic and clinical characteristics of the 3756 participants in the discovery cohort are shown in Table 1. Only 2% of type 2 diabetes cases had onset before 25 years, and 81% of them were overweight or obese (BMI >25, calculated as weight in kilograms divided by height in meters squared).

Genetic Studies

Exome-wide Search for Low-Frequency Variants Associated With Type 2 Diabetes

Our hybrid selection libraries covered 76% of sequenced targets at 20x depth of coverage with a mean of 67.17x. The concordance of nonreference genotypes between the sequence data and the array data was 0.995. After quality control of se-

quence data, 1 190 196 variants were observed in the whole-exome sequencing data of 3756 samples (1794 type 2 diabetes cases and 1962 controls; eTable 2 in the Supplement). Of these, 264 995 variants were observed in at least 2 of our samples but absent in the 1000 Genomes Project²³ and the Exome Sequencing Project²⁴ (eTable 3 in the Supplement).

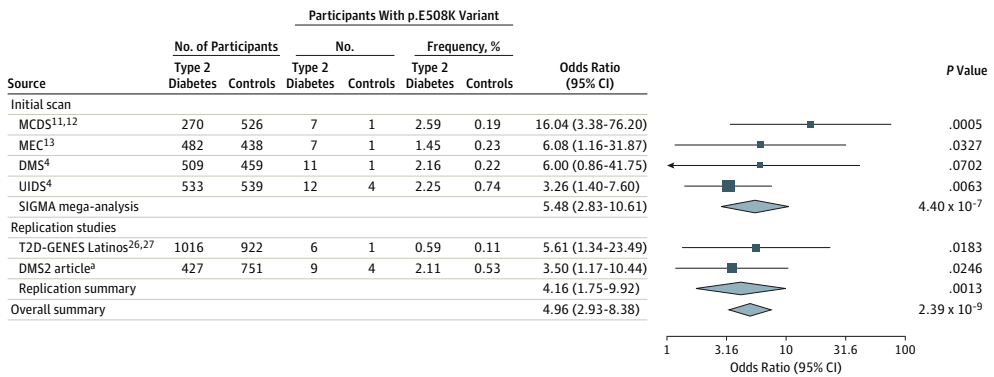
In our single-variant association analyses, a cluster of linked common missense variants in *SLC16A11* were consistently associated with type 2 diabetes prevalence ($P = 2.08 \times 10^{-10}$) as had been previously reported in genome-wide association studies by the SIGMA T2D Consortium and others (eFigure 4A and eTable 4 in the Supplement).^{4,25}

Among variants with minor allele frequency of less than 5%, a single missense variant departed from the null distribution (eFigure 4B in the Supplement). This variant encoded an

NCBI NP_000536.5:p.E508K (p.E508K) substitution (NCBI NC_000012.12:c.1522G>A; chr12:121437091_G>A) in exon 8 of *HNF1A*, the gene responsible for the maturity onset diabetes of the young type 3 (MODY3) subtype of MODY3 (Mendelian Inheritance in Man No. 142410). The p.E508K variant was observed in 37 type 2 diabetes cases (1 in homozygous form) and in 7 participants without diabetes (OR, 5.48; 95% CI, 2.83-10.61; $P = 4.4 \times 10^{-7}$; Figure 1 and Figure 2 and eFigure 5 in the Supplement).

In our replication effort, the p.E508K variant was found in the T2D-GENES Latino group^{26,27} but entirely absent in all other populations, showing a nominally significant association with increased prevalence for type 2 diabetes (7 affected carriers and 1 nonaffected carrier; OR, 5.61; 95% CI, 1.34-23.49; $P = .0013$). After de novo genotyping 1178 additional

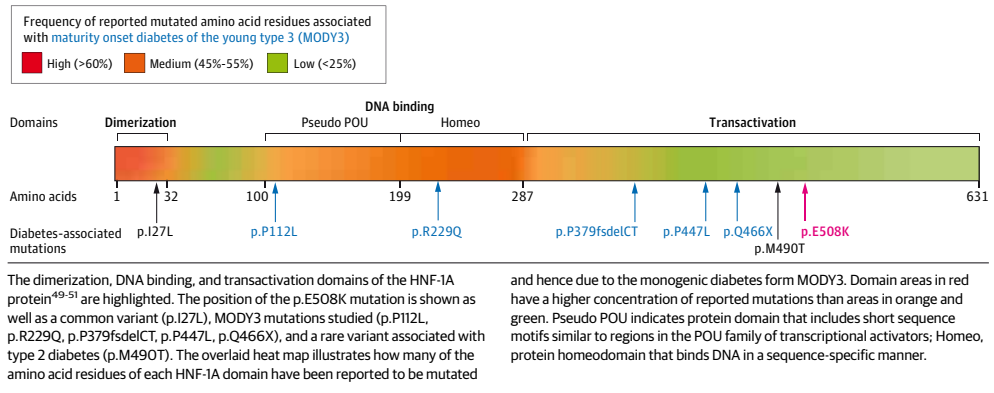
Figure 1. Discovery and Replication of the *HNF1A* p.E508K Variant



Forest plot showing odds ratio estimates and 95% confidence intervals at p.E508K (squared boxes) from the 4 SIGMA studies, the SIGMA pooled mega-analysis, the replication studies, and the overall meta-analysis. Odds ratios for the meta-analyses are represented with a diamond. SIGMA mega-analysis represents the combined results from the 4 SIGMA studies. DMS indicates Diabetes in Mexico Study; MCDS, Mexico City Diabetes Study; MEC,

Multiethnic Cohort; UIDS, Universidad Nacional Autónoma de México/Instituto Nacional de Ciencias Médicas y Nutrición Salvador Zubirán Diabetes Study; T2D-GENES, Type 2 Diabetes Genetic Exploration by Next-Generation Sequencing in Multi-Ethnic Samples. ^a Represents data from the current article.

Figure 2. The *HNF1A* Protein With a Heat Map of Diabetes-Associated Mutations



Mexican self-identified indigenous individuals (DMS2, further details are provided in the Supplement), we observed 9 affected carriers and 4 nonaffected carriers (OR, 3.50; 95% CI, 1.17-10.44; $P = .0183$). Combined, the 2 replication studies identified 15 affected carriers and 5 nonaffected controls (OR, 4.16; 95% CI, 1.75-9.92; $P = .0013$). Combining all available data yielded 52 affected carriers and 12 nonaffected controls (OR, 4.96; 95% CI, 2.93-8.38; an experiment-wide $P = 2.39 \times 10^{-9}$; Figure 1).

We found no evidence for p.E508K in the 1092 samples of the 1000 Genomes Project,²³ the 6503 samples in the Exome Sequencing Project²⁴ or in 11 160 non-Latino samples in the T2D-GENES and GoT2D data sets. Analysis of local ancestry in our data indicates that all p.E508K carriers in our studies carry at least 1 segment of inferred Native American ancestry (eTable 5 in the Supplement).

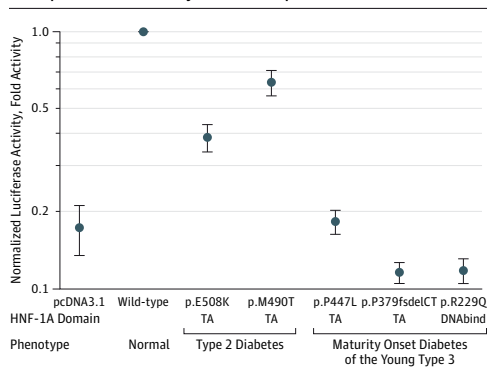
In group tests that included combinations of rare (MAF <1%) nonsynonymous, loss-of-function variants, or both in up to 15 469 genes (eTables 6 and 7 in the Supplement), we found no significant associations after removing the effect of the *HNFI1A* p.E508K variant. The aggregated effect of these potentially functional variants in 2 gene-sets of 13 *MODY* genes and 70 previously implicated type 2 diabetes genes were similarly negative after removing the effect of the *HNFI1A* p.E508K variant (eTables 8 and 9 in the Supplement).

Functional Studies

Mutations in *HNFI1A* that cause *MODY* diabetes alter protein function through reduced transactivation, decreased binding to DNA, or disrupted nuclear localization.²⁰ Because p.E508K is located in the HNF-1A transactivation domain, we investigated its effect on transactivation using a reporter construct assay in HeLa cells. Protein carrying p.E508K was compared with a wild-type HNF-1A variant as well as 4 other HNF-1A variants in the DNA-binding or transactivation domains: p.M490T, which has been observed in 1 patient with type 2 diabetes,²⁸ and 3 mutations (p.P447L, p.P379fsdelCT, and p.R229Q) previously identified in patients with *MODY3*.²⁹ The p.E508K mutant demonstrated lower transcriptional activity on the HNF-1A-responsive rat albumin promoter than wild-type HNF-1A ($P < .0001$) or p.M490T. However, the 3 *MODY3* mutants showed greater reductions in transactivation (Figure 3). Similar reductions in p.E508K transcriptional activation were found in MIN6 cells (eFigure 6A in the Supplement), and using 2 different reporter constructs (*GLUT2* and *HNF4A* promoters; eFigure 6B in the Supplement). The p.E508K mutant protein bound to an HNF-1A binding site-containing oligonucleotide with equal affinity to the wild-type protein (Figure 4 and eFigure 6C in the Supplement), whereas 2 *MODY3*-associated mutants with mutations in the DNA-binding domain, p.P112L and p.R229Q, demonstrated impaired DNA binding (Figure 4).²⁰

Compared with wild-type HNF-1A, the p.E508K mutant demonstrated slightly impaired nuclear targeting, with an increased proportion of cells displaying both cytosolic and nuclear staining. The shift in nuclear localization was less than that observed using the cytosol-retained HNF-1A mutant p.Q466X (Figure 5 and eFigure 6D in the Supplement). Express-

Figure 3. Transcriptional Activation of HNF-1A p.E508K as Measured by the Expression of the Firefly Luciferase Reporter Gene



HeLa cells were transiently transfected with nonmutant or mutant *HNFI1A* plasmids and reporter plasmids pGL3-RA and pRL-SV40. Measurements are given in fold activity relative to wild-type. Each point represents the mean (error bars indicate 95% CIs) of 9 readings. TA indicates variants that affect the transactivation domain; DNABind, the DNA binding domain; and pcDNA3.1, the empty pcDNA3.1 vector. All values were $P < .05$ compared with wild-type activity.

sion of the p.E508K protein was 47.5% lower than that of wild-type HNF-1A ($P = 1.03 \times 10^{-5}$; eFigure 6E in the Supplement).

Clinical Characteristics of p.E508K Carriers

When comparing p.E508K carriers with noncarriers among the 3756 participants in our study, we did not observe statistically significant differences in the mean (SD) age of diabetes onset: 45.3 (11.2) years vs 47.5 (11.5) years, $P = .49$; BMI, 28.2; (5.5) vs 29.3 (5.3), $P = .19$; waist circumference in men, 92.9 (7.0) cm vs 99.3 (11.0) cm, $P = .14$ or women, 98.0 (13.9) cm vs 99.7 (13.9) cm, $P = .64$; or in fasting glucose levels, 176.5 (84.6) mg/dL vs 165.7 (75.6) mg/dL, $P = .43$ (To convert fasting glucose from mg/dL to mmol/L, multiply by 0.0555; Table 2 and Figure 6).

Discussion

We performed whole-exome sequencing in 3756 individuals of Mexican and Mexican American ancestry and performed an exome-wide search for low-frequency and rare variants associated with type 2 diabetes. The only rare variant with a significant association with type 2 diabetes prevalence was the p.E508K variant in *HNFI1A*, the gene responsible for *MODY3*. The effect size of the variant (OR, 4.96; 95% CI, 2.93-8.38) was the largest observed to date for any diabetes variant with a frequency more than 1 in 1000. This association was replicated in 2 independent cohorts of Latinos and Mexicans with an OR of similar magnitude. We also demonstrated, using transiently transfected cell models, reduced levels of transactivation activity for p.E508K compared with wild-type HNF-1A. As shown in binding assays, this reduction in activity was not

driven by differences in DNA-binding affinity but may be attributable to reduced protein expression and altered nuclear localization of the mutant protein.

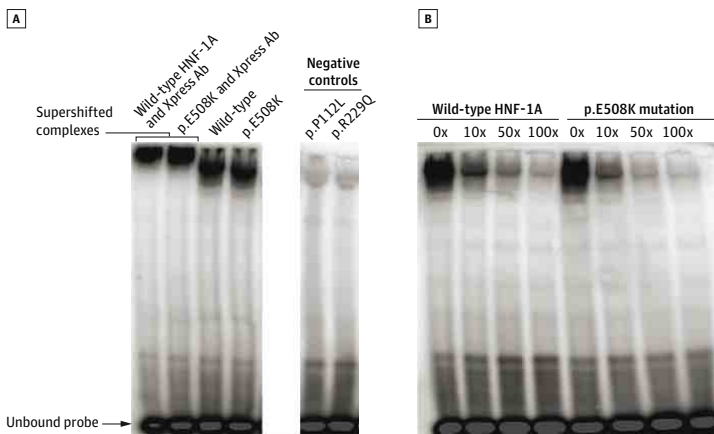
MODY is a monogenic cause of diabetes, which usually manifests at earlier ages (<25 years) and presents in nonobese patients.³⁰ Each MODY family carries a rare coding mutation in 1 of 13 genes that has an autosomal dominant pattern of transmission.³⁰ Mutations in the known MODY genes are thought to explain between 0.18% and 1.8% of all type 2 diabetes cases.³¹⁻³⁴

The p.E508K variant has been reported in 2 published articles,^{35,36} both reporting on individuals with MODY. In 1 case, a family member had early onset diabetes (age 17 years), and carried both *HNF1A* p.E508K and a mutation in *HNF4A*, p.R80Q. The father from whom p.E508K was inherited was diagnosed with type 2 diabetes at age 57 years.^{35,36} The finding

of these variants in patients with MODY suggested that they might be high-penetrance alleles. Our study in large populations without ascertainment bias for early-onset showed that p.E508K was associated with a 5-fold increase in prevalence, but incomplete penetrance. Moreover, in our study, carriers of p.E508K did not show early-onset of type 2 diabetes, were indistinguishable from the wider type 2 diabetes population in adiposity or glycemia, and thus did not fulfill classical MODY3 diagnostic criteria (Table 2, Figure 6). These data are consistent with the possibility that p.E508K is a weaker allele than some other MODY3 mutations and that ascertainment bias may have led to overestimation of the effects of this and other MODY mutations, as suggested previously.²⁸

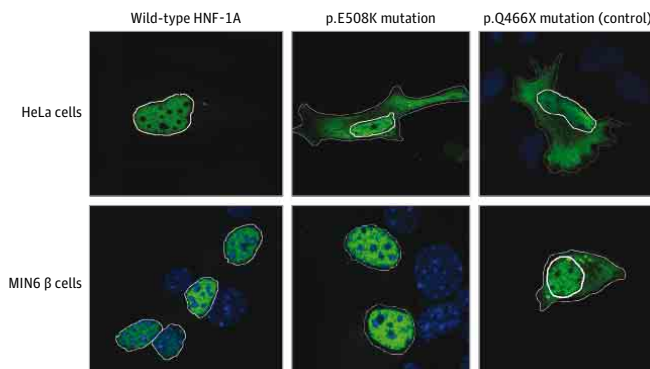
A private mutation (G319S) in *HNF1A* has been found in Oji-Cree populations associated with early-onset type 2 diabetes.³⁷ Also, a very rare frameshift deletion in *HNF1A*,

Figure 4. DNA Binding of HNF-1A p.E508K to the Rat Albumin Promoter as Studied by Electrophoretic Mobility Shift Assay



Xpress-epitope-tagged wild-type and p.E508K mutant proteins incubated with a radiolabeled DNA fragment containing the HNF-1A-binding site in the rat albumin promoter. A, Two HNF-1A mutants (p.P112L and p.R229Q) with impaired DNA-binding were included as negative controls. Addition of the anti-Xpress antibody induced a supershift (a reduction in mobility of protein-DNA complex alone) for the DNA-protein complexes, confirming the identity of HNF-1A within the complexes. B, A competition assay was performed by adding increasing amounts (0x, 10x, 50x, or 100x) of radiolabeled DNA fragment, confirming the identity of the radiolabeled probe.

Figure 5. Intracellular Localization of HNF-1A p.E508K in Transiently Transfected HeLa cells and MIN6 β cells



Cells were transfected for 48 hours and Xpress-epitope-tagged HNF-1A proteins detected with anti-Xpress antibody and Alexa488 (green). DNA staining (DAPI) is shown in blue. A previously reported HNF-1A mutant, p.Q466X, with impaired nuclear localization was included as a control. For the purpose of clarity, the nuclei have been marked with a solid white line. To illustrate cytosolic accumulation, the cell membrane has been marked with a dotted white line for mutants p.E508K and p.Q466X.

290fsdelC, was recently associated with MODY and type 2 diabetes in the Icelandic population.^{10,38}

Our study surveyed variants across the majority of protein-coding exons in a sizable population, providing the highest-resolution scan to date of the contribution of protein-coding genetic variation to type 2 diabetes. Our study had 80% power to detect variants with the OR and carrier frequency of p.E508K (5-fold and 1% in the population). For variants of higher frequency, our power was sufficient to detect a smaller effect (80% power for variants with frequency >2% and OR>3.3). We performed both single-variant analysis and burden tests that combined rare variants in each gene. Only 1 rare coding variant and

1 gene showed significant association with type 2 diabetes prevalence. These data suggest that low-frequency variants in coding regions explain only a small fraction of the heritability of type 2 diabetes.

Our study has limitations. Current exome-capture methods are imperfect. Additional low-frequency variants associated with type 2 diabetes might have been missed due to incomplete coverage of all human exons, and, by design, this technology does not detect variants in the noncoding majority of the genome. Although a 2% frequency of p.E508K among type 2 diabetes cases could translate into more than 100 000 carriers in Mexico alone, this number is still far from explain-

Table 2. Phenotypic Characteristics of 3756 Participants From the SIGMA Studies According to Type 2 Diabetes Status and p.E508K Carrier Status

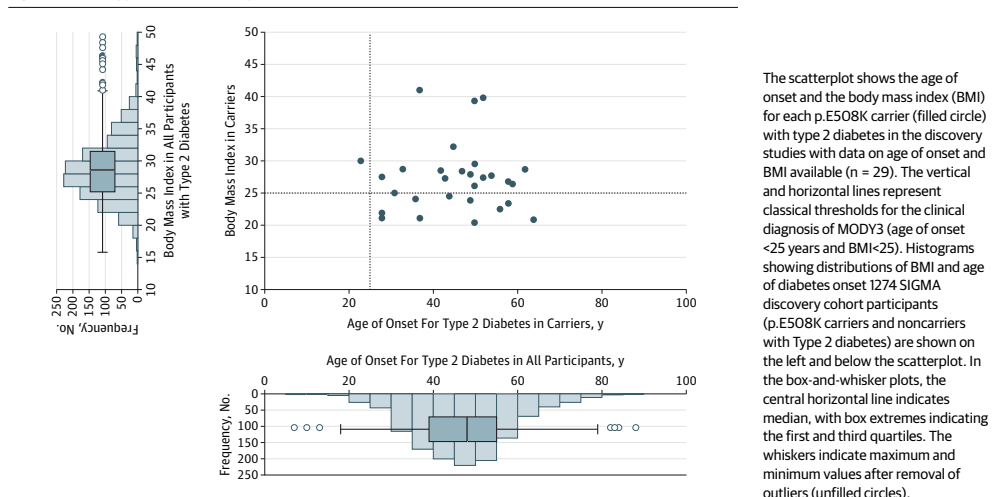
	Mean (SD)				P Value	
	Type 2 Diabetes		Controls		Carriers vs Noncarriers	
	p.E508K (n = 37)	p.E508 (n = 1757)	p.E508K (n = 7)	p.E508 (n = 1955)	Type 2 Diabetes	Controls
Age, y	55.9 (9.6)	57.6 (10.7)	54.3 (9.2)	57.3 (8.9)	.34	.34
Age at onset, y	45.3 (11.2)	47.5 (11.5)			.49	
Men	11	707	3	739		
Women	26	1050	4	1216		
Fasting glucose, mg/dL	176.5 (84.6)	165.7 (75.6)	86.4 (9.0)	88.2 (9.0)	.43	.37
BMI	28.2 (5.5)	29.3 (5.3)	27.1 (3.5)	28.3 (4.5)	.19	.55
Waist, cm						
Men	92.9 (7.0)	99.3 (11.1)	90.5 (19.8)	97.6 (9.7)	.14	.64
Women	98.0 (13.9)	99.7 (13.9)	95.5 (7.8)	94.9 (13.3)	.64	.88
Waist to hip ratio, cm						
Men	0.96 (0.05)	0.97 (0.07)	0.96 (NA) ^a	0.97 (0.10)	.54	.88
Women	0.93 (0.07)	0.92 (0.08)	0.91 (0.05)	0.90 (0.09)	.90	.85

Abbreviations: BMI, body mass index, calculated as weight in kilograms divided by height in meters squared; NA, not applicable.

^a Only 1 participant with this measurement.

SI conversion factor: To convert fasting glucose from mg/dL to mmol/L, multiply by 0.0555.

Figure 6. Phenotypic Distribution of p.E508K Carriers



The scatterplot shows the age of onset and the body mass index (BMI) for each p.E508K carrier (filled circle) with type 2 diabetes in the discovery studies with data on age of onset and BMI available (n = 29). The vertical and horizontal lines represent classical thresholds for the clinical diagnosis of MODY3 (age of onset <25 years and BMI <25). Histograms showing distributions of BMI and age of diabetes onset 1274 SIGMA discovery cohort participants (p.E508K carriers and noncarriers with Type 2 diabetes) are shown on the left and below the scatterplot. In the box-and-whisker plots, the central horizontal line indicates median, with box extremes indicating the first and third quartiles. The whiskers indicate maximum and minimum values after removal of outliers (unfilled circles).

ing the expected overall genetic contribution to type 2 diabetes. Although our study represents the largest published exome-based survey of type 2 diabetes to date, larger sample sizes will be needed to perform an adequately powered survey of variants at frequencies lower than 1%.^{39,40}

The current study and a recent publication reporting an association of common variants in *SLC16A11* with type 2 diabetes in Latinos⁴ demonstrate the value of studying diverse populations. The *HNF1A* p.E508K variant has not been reported in other whole-exome sequencing or candidate gene association studies for type 2 diabetes of European^{9,10,41} and Asian⁴²⁻⁴⁵ ancestry. We surveyed a total of 25 663 exomes in this study, both from our own study and collaborating consortia. The p.E508K variant was identified only in individuals from Mexico or in Latinos from the southern United States, indicating that this variant is only found at appreciable frequency in a tightly restricted subset of human populations. Further studies will be required to characterize the fine-scale geographic distribution of p.E508K and its association with type 2 diabetes prevalence in other Latino populations. Our results emphasize that systematic discovery of the genetic determinants of complex disease, especially for rare variants, will require surveys across a wide range of human populations.

The association of the p.E508K variant with type 2 diabetes prevalence in the Latino population has potential clinical implications. Approximately 4 in a thousand people in Latino populations carry p.E508K, and these individuals have a 5-fold increase in prevalence for type 2 diabetes (2.1% in cases, 0.35% in controls). Second, it is known that patients with MODY3 are sensitive to sulfonylureas,⁴⁶ experiencing improved metabolic con-

trol on sulfonylurea therapy compared with insulin,⁴⁷ in addition to improved quality of life due to reduced injections and capillary glucose measurements. Also, these patients have a 5-fold higher response to the sulfonylurea gliclazide than to metformin, which is the first-line drug of choice for the treatment of type 2 diabetes.⁴⁸ If this was shown to be the case for carriers of p.E508K, it could motivate choice of sulfonylurea therapy for the estimated 2% of all Latino patients with type 2 diabetes who carry this variant. Because this response may be dependent on additional genetic or environmental factors, further studies are needed to determine whether metformin or a sulfonylurea should be the first line of treatment in these patients.

Conclusions

Using whole-exome sequencing, we identified a single low-frequency missense variant (p.E508K) in *HNF1A*, the gene responsible for a monogenic, early-onset form of diabetes (MODY3), that was associated with type 2 diabetes prevalence in general populations of Latinos. This rare variant was associated with a 5-fold increase in the prevalence of type 2 diabetes, but it was not associated with an early-onset form of diabetes, and, in our data, affected carriers were clinically indistinguishable from the wider type 2 diabetes population. In vitro, p.E508K negatively affected transcriptional activation, protein expression, and nuclear localization. Further research is warranted to evaluate the clinical relevance of these findings, including the benefits of selective population screening and the choice of genotype-guided therapeutic regimens.

ARTICLE INFORMATION

Authors: The following investigators of the SIGMA Type 2 Diabetes Consortium take authorship responsibility for the study results: Karol Estrada, PhD; Ingvið Aukrust, PhD; Lise Bjørkhaug, PhD; Noël P. Burt, PhD; Josep M. Mercader, PhD; Humberto García-Ortiz, PhD; Alicia Huerta-Chagoya, MSc; Hortensia Moreno-Macias, PhD; Geoffrey Walford, MD; Jason Flannick, PhD; Amy L. Williams, PhD; María J. Gómez-Vázquez, BSc; Juan C. Fernandez-Lopez, MSc; Angélica Martínez-Hernández, PhD; Silvia Jiménez-Morales, PhD; Federico Centeno-Cruz, PhD; Elvia Mendoza-Caamal, MD; Cristina Revilla-Monsalve, PhD; Sergio Islas-Andrade, MD, PhD; Emilio J. Córdova, PhD; Xavier Soberón, PhD; María E. González-Villalpando, MD; E. Henderson, MD; Lynne R. Wilkens, DrPH; Loïc Le Marchand, MD, PhD; Olimpia Arellano-Campos, MD, PhD; María L. Ordóñez-Sánchez, BSc; Maribel Rodríguez-Torres, BSc; Rosario Rodríguez-Guillén, MSc; Laura Riba, MSc; Laeya A. Najmi, MSc; Suzanne B.R. Jacobs, PhD; Timothy Fennell, BSc; Stacey Gabriel, PhD; Pierre Fontanillas, PhD; Craig L. Hanis, PhD; Donna M. Lehman, PhD; Christopher P. Jenkinson, PhD; Hanna E. Abboud, MD; Graeme I. Bell, PhD; María L. Cortes, PhD; Michael Boehnke, PhD; Clicerio González-Villalpando, MD; Lorena Orozco, MD, PhD; Christopher A. Haiman, ScD; Teresa Tusié-Luna, MD, PhD; Carlos A. Aguilar-Salinas, MD, PhD; David Altshuler, MD, PhD; Pål R. Njølstad, MD, PhD; Jose C. Florez, MD, PhD; Daniel G. MacArthur, PhD.

Affiliations of Authors: Program in Medical and Population Genetics, Broad Institute of Harvard and MIT, Cambridge, Massachusetts (Estrada, Burt, Mercader, Flannick, Williams, Jacobs, Fontanillas, Altshuler, Florez, MacArthur); Analytic and Translational Genetics Unit, Massachusetts General Hospital, Boston (Estrada); Department of Medicine, Harvard Medical School, Boston, Massachusetts (Estrada, Walford, Altshuler, Florez, MacArthur); KG Jebsen Center for Diabetes Research, Department of Clinical Science, University of Bergen, Bergen, Norway (Aukrust, Bjørkhaug, Najmi, Njølstad); Department of Pediatrics, Haukeland University Hospital, Bergen, Norway (Bjørkhaug, Njølstad); Department of Biomedicine, University of Bergen, Bergen, Norway (Aukrust); Center for Human Genetic Research and Diabetes Research Center (Diabetes Unit), Massachusetts General Hospital, Boston (Mercader, Walford, Altshuler, Florez); Joint BSC-CRG-IRB Research Program in Computational Biology, Barcelona Supercomputing Center, Barcelona, Spain (Mercader); Instituto Nacional de Medicina Genómica, Tlalpan, Mexico City, Mexico (García-Ortiz, Fernandez-Lopez, Martínez-Hernández, Jiménez-Morales, Centeno-Cruz, Mendoza-Caamal, Córdova, Soberón, Orozco); Instituto de Investigaciones Biomédicas, UNAM Unidad de Biología Molecular y Medicina Genómica, UNAM/INCNSZ, Coyoacán, Mexico City, Mexico (Huerta-Chagoya, Riba, Tusié-Luna); Universidad Autónoma Metropolitana, Tlalpan, Mexico City, Mexico (Moreno-Macias);

Centro de Estudios en Diabetes, Unidad de Investigación en Diabetes y Riesgo Cardiovascular, Centro de Investigación en Salud Poblacional, Instituto Nacional de Salud Pública, Mexico City, Mexico (M. E. González-Villalpando, C. González-Villalpando); Department of Molecular Biology, Harvard Medical School, Boston, Massachusetts (Flannick, Altshuler); Department of Biological Sciences, Columbia University, New York, New York (Williams); Department of Biostatistics, Center for Statistical Genetics, University of Michigan, Ann Arbor (Boehnke); Department of Preventive Medicine, Keck School of Medicine, University of Southern California, Los Angeles (Henderson, Haiman); Instituto Nacional de Ciencias Médicas y Nutrición Salvador Zubirán, Sección XVI, Tlalpan, Mexico City, Mexico (Gómez-Vázquez, Arellano-Campos, Ordóñez-Sánchez, Rodríguez-Torres, Rodríguez-Guillén, Tusié-Luna, Aguilar-Salinas); Department of Genetics, Harvard Medical School, Boston, Massachusetts (Altshuler); Center for Human Genetic Research, Massachusetts General Hospital, Boston (Altshuler); Department of Biology, Massachusetts Institute of Technology, Cambridge (Altshuler); Unidad de Investigación Médica en Enfermedades Metabólicas, CMN SXXI, Instituto Mexicano del Seguro Social, Mexico City (Revilla-Monsalve, Islas-Andrade); Epidemiology Program, University of Hawaii Cancer Center, Honolulu (Wilkens, Le Marchand); Center for Medical Genetics and Molecular Medicine,

Haukeland University Hospital, Bergen, Norway (Najmi); The Genomics Platform, The Broad Institute of Harvard and MIT, Cambridge, Massachusetts (Fennell, Gabriel); Human Genetics Center, University of Texas Health Science Center at Houston (Hanis); Department of Medicine, University of Texas Health Science Center at San Antonio (Lehman, Jenkinson, Abboud); Department of Human Genetics, University of Chicago, Chicago, Illinois (Bell); Department of Medicine, University of Chicago, Chicago, Illinois (Bell); Broad Institute of Harvard and MIT, Cambridge, Massachusetts (Cortes).

Author Contributions: Dr Estrada had full access to all of the data in the study and takes responsibility for the integrity of the data and the accuracy of the data analysis.

Study concept and design: Estrada, Aukrust, Bjørkhaug, Burt, Orozco, Haiman, Tusié-Luna, Altschuler, Njølstad, MacArthur, Williams, Islas-Andrade, M. González-Villalpando, Hanis, Florez, Boehnke.

Acquisition, analysis, or interpretation of data: Estrada, Aukrust, Bjørkhaug, Burt, Mercader, García-Ortiz, Huerta-Chagoya, Moreno-Macías, C. González-Villalpando, Orozco, Salinas, Altschuler, Njølstad, MacArthur, Flannick, Cortes, Williams, Gómez-Vázquez, Fernandez-Lopez, Martínez-Hernández, Centeno-Cruz, Mendoza-Caamal, Revilla-Monsalve, Córdova, Soberón, Henderson, Wilkens, Marchand, Arellano-Campos, Ordóñez-Sánchez, Torres, Rodríguez-Guillén, Riba, Walford, Najmi, Jacobs, Fennell, Gabriel, Fontanillas, Jiménez-Morales, Hanis, Florez, Lehman, Jenkinson, Abboud, Bell, Boehnke.

Drafting of the manuscript: Estrada, Mercader, García-Ortiz, Huerta-Chagoya, Moreno-Macías, Orozco, Altschuler, Njølstad, MacArthur, Cortes, Martínez-Hernández, Centeno-Cruz, Islas-Andrade, Córdova, Henderson, Arellano-Campos, Najmi, Gabriel, Jiménez-Morales.

Critical revision of the manuscript for important intellectual content: Estrada, Aukrust, Bjørkhaug, Burt, Mercader, C. González-Villalpando, Orozco, Haiman, Tusié-Luna, Salinas, Altschuler, Njølstad, MacArthur, Flannick, Williams, Gómez-Vázquez, Fernandez-Lopez, Mendoza-Caamal, Revilla-Monsalve, Soberón, M. González-Villalpando, Wilkens, Marchand, Torres, Rodríguez-Guillén, Riba, Walford, Jacobs, Fennell, Gabriel, Fontanillas, Hanis, Florez, Lehman, Jenkinson, Abboud, Bell, Boehnke.

Statistical analysis: Estrada, Mercader, García-Ortiz, Huerta-Chagoya, Moreno-Macías, Orozco, Haiman, Altschuler, MacArthur, Flannick, Williams, Gómez-Vázquez, Fernandez-Lopez, Walford, Najmi, Fennell, Fontanillas, Boehnke.

Obtained funding: Orozco, Tusié-Luna, Altschuler, Njølstad, Cortes, Soberón, Wilkens, Hanis, Florez, Lehman, Boehnke.

Administrative, technical, or material support: Aukrust, Bjørkhaug, Burt, Orozco, Tusié-Luna, Salinas, Altschuler, Njølstad, MacArthur, Flannick, Cortes, Fernandez-Lopez, Martínez-Hernández, Centeno-Cruz, Mendoza-Caamal, Revilla-Monsalve, Islas-Andrade, Córdova, Ordóñez-Sánchez, Torres, Rodríguez-Guillén, Riba, Jiménez-Morales, Florez, Lehman, Jenkinson, Abboud, Bell.

Study supervision: Aukrust, Bjørkhaug, Burt, C. González-Villalpando, Orozco, Tusié-Luna, Altschuler, Njølstad, MacArthur, M. González-Villalpando, Riba, Gabriel, Florez.

Conflict of Interest Disclosures: All authors have completed and submitted the ICMJE Form for Disclosure of Potential Conflicts of Interest and none were reported.

Funding/Support: The work was conducted as part of the Slim Initiative for Genomic Medicine, a project funded by the Carlos Slim Health Institute in Mexico. The UNAM/INCMSZ Diabetes Study was supported by Consejo Nacional de Ciencia y Tecnología grants 138826, 128877, CONACT-SALUD 2009-01-115250, and a grant from Dirección General de Asuntos del Personal Académico, UNAM, IT 214711. The Diabetes in Mexico Study was supported by Consejo Nacional de Ciencia y Tecnología grant 86867 and by Instituto Carlos Slim de la Salud, A.C. The Mexico City Diabetes Study was supported by National Institutes of Health (NIH) grant RO1HL24799 and by the Consejo Nacional de Ciencia y Tecnología grants 2092, M9303, F677-M9407, 251M, and 2005-C01-14502, SALUD 2010-2-151165. The Multiethnic Cohort was supported by NIH grants CA164973, CA054281, and CA063464. The Singapore Chinese Health Study was funded by the National Medical Research Council of Singapore under its individual research grant scheme and by NIH grants R01 CA55069, R35 CA53890, R01 CA80205, and R01 CA144034. The Type 2 Diabetes Genetic Exploration by Next-generation sequencing in multi-Ethnic Samples (T2D-GENES) project was supported by NIH grants U01DK085526 and U01DK085501. The San Antonio Mexican American Family Studies (SAMAFFS) were supported by R01 DK042273, R01 DK047482, R01DK053889, R01 DK057295, P01 HL045522, and a Veterans Administration Epidemiologic grant (R.A.D.). The University of Bergen, Research Council of Norway, KG Jebsen Foundation, Helse Vest, and European Research Council funded the Norwegian team. Dr Mercader was supported by Sara Borrell Fellowship from the Instituto Carlos III, Spain. Dr Estrada was supported by The Netherlands Organization for Scientific Research under the Rubicon fellowship 825.12.023.

Role of the Sponsors: The funding sources had no role in the design and conduct of the study; collection, management, analysis, and interpretation of the data; and preparation, review, or approval of the manuscript; and decision to submit the manuscript for publication.

The SIGMA Type 2 Diabetes Consortium: Writing Team: Karol Estrada, PhD, Ingvild Aukrust, PhD, Lise Bjørkhaug, PhD, Noël P. Burt, PhD, Josep M. Mercader, PhD, Humberto García-Ortiz, PhD, Alicia Huerta-Chagoya, MSc, Hortensia Moreno-Macías, PhD, Geoffrey Walford, MD, Jason Flannick, PhD, Amy L. Williams, PhD, Michael Boehnke, PhD, Clicerio González-Villalpando, MD, Lorena Orozco, MD, PhD, Christopher A. Haiman, ScD, Teresa Tusié-Luna, MD, PhD, Carlos A. Aguilar-Salinas, MD, PhD, David Altschuler, MD, PhD, Pål R. Njølstad, MD, PhD, Jose C. Florez, MD, PhD, Daniel G. MacArthur, PhD.

Analysis Team: Karol Estrada, PhD, Alicia Huerta-Chagoya, MSc, Humberto García-Ortiz, PhD, Hortensia Moreno-Macías, PhD, Josep M. Mercader, PhD, Jason Flannick, PhD, Amy L. Williams, PhD, María J. Gómez-Vázquez, BSc, Juan C. Fernandez-Lopez, MSc, Noël P. Burt, PhD, Carlos A. Aguilar-Salinas, MD, PhD, Lorena Orozco, MD, PhD, Teresa Tusié-Luna, MD, PhD, David Altschuler, MD, PhD, Jose C. Florez, MD, PhD, Daniel G. MacArthur, PhD; Whole-Exome Sequenced cohorts: *Diabetes in*

Mexico Study: Humberto García-Ortiz, PhD, Angélica Martínez-Hernández, PhD, Federico Centeno-Cruz, PhD, Elvia Mendoza-Caamal, MD, Cristina Revilla-Monsalve, PhD, Sergio Islas-Andrade, MD, PhD, Emilio J. Córdova, PhD, Xavier Soberón, PhD, Lorena Orozco, MD, PhD. *Mexico City diabetes study:* Clicerio González-Villalpando, MD, María E. González-Villalpando, MD. *Multiethnic cohort study:* Christopher A. Haiman, ScD, Brian E. Henderson, MD, Lynne R. Wilkens, DrPH, Loïc Le Marchand, MD, PhD. *UNAM/INCMSZ diabetes study:* Olimpia Arellano-Campos, MD, PhD, Alicia Huerta-Chagoya, MSc, María L. Ordóñez-Sánchez, BSc, Maribel Rodríguez-Torres, BSc, Rosario Rodríguez-Guillén, MSc, Laura Riba, MSc, Teresa Tusié-Luna, MD, PhD, Carlos A. Aguilar-Salinas, MD, PhD.

Functional Studies: Laeya A. Najmi, MSc, Ingvild Aukrust, PhD, Lise Bjørkhaug, PhD, Suzanne B. R. Jacobs, PhD, Pål R. Njølstad, MD, PhD.

Whole-Exome Sequencing: Noël P. Burt, PhD, Timothy Fennell, BSc, Broad Genomics Platform, Stacey Gabriel, PhD.

Replication Studies: T2D-GENES Consortium: Jason Flannick, PhD, Pierre Fontanillas, PhD, Craig L. Hanis, PhD, Donna M. Lehman, PhD, Christopher P. Jenkinson, PhD, Hanna E. Abboud, MD, Graeme I. Bell, PhD, Jose C. Florez, MD, PhD, David Altschuler, MD, PhD, Michael Boehnke, PhD. *Diabetes in Mexico study 2:* Humberto García-Ortiz, PhD, Angélica Martínez-Hernández, PhD, Emilio J. Córdova, PhD, Silvia Jiménez-Morales, PhD, Federico Centeno-Cruz, PhD, Elvia Mendoza-Caamal, MD, Cristina Revilla-Monsalve, PhD, Sergio Islas-Andrade, MD, PhD, Xavier Soberón, PhD, Lorena Orozco, MD, PhD.

Scientific and Project Management: Noël P. Burt, PhD, María L. Cortes, PhD.

Steering Committee: David Altschuler, MD, PhD, Jose C. Florez, MD, PhD, Christopher A. Haiman, ScD, Carlos A. Aguilar-Salinas, MD, PhD, Clicerio González-Villalpando, MD, Lorena Orozco, MD, PhD, Teresa Tusié-Luna, MD, PhD.

Additional Information: The members of the SIGMA Type 2 Diabetes Consortium mourn the sudden passing of coauthor Laura Riba, a good friend, respected colleague and lab manager with outstanding contributions to the research of type 2 diabetes in Mexico. We dedicate this article to her memory.

Additional Contribution: Researchers of the DMS2 study thank Olaf Iván Corro Labra and José Luis de Jesus García Ruiz from the "Comisión Nacional para el Desarrollo de los Pueblos Indígenas" for their support in sample collection, for which they were not compensated.

Correction: The authors added a tribute on August 20, 2014 to a colleague who had died unexpectedly and added the name of an author who was not included in the byline.

REFERENCES

- Villalpando S, de la Cruz V, Rojas R, et al. Prevalence and distribution of type 2 diabetes mellitus in Mexican adult population. *Salud Publica Mex.* 2010;52(suppl 1):S19-S26.
- Barquera S, Tovar-Guzmán V, Campos-Nonato I, González-Villalpando C, Rivera-Dommarco J. Geography of diabetes mellitus mortality in Mexico. *Arch Med Res.* 2003;34(5):407-414.

3. Cowie CC, Rust KF, Byrd-Holt DD, et al. Prevalence of diabetes and impaired fasting glucose in adults in the U.S. population: National Health and Nutrition Examination Survey 1999-2002. *Diabetes Care*. 2006;29(6):1263-1268.
4. Williams AL, Jacobs SB, Moreno-Macias H, et al. Sequence variants in SLC16A11 are a common risk factor for type 2 diabetes in Mexico. *Nature*. 2014;506(7486):97-101.
5. Morris AP, Voight BF, Teslovich TM, et al; Large-scale association analysis provides insights into the genetic architecture and pathophysiology of type 2 diabetes. *Nat Genet*. 2012;44(9):981-990.
6. Voight BF, Scott LJ, Steinthorsdottir V, et al. Twelve type 2 diabetes susceptibility loci identified through large-scale association analysis. *Nat Genet*. 2010;42(7):579-589.
7. Diabetes Genetics Replication and Meta-analysis (DIAGRAM) Consortium. Genome-wide trans-ancestry meta-analysis provides insight into the genetic architecture of type 2 diabetes susceptibility. *Nat Genet*. 2014;46(3):234-244.
8. Albrechtsen A, Grarup N, Li Y, et al. Exome sequencing-driven discovery of coding polymorphisms associated with common metabolic phenotypes. *Diabetologia*. 2013;56(2):298-310.
9. Lohmueller KE, Sparso T, Li Q, et al. Whole-exome sequencing of 2000 Danish individuals and the role of rare coding variants in type 2 diabetes. *Am J Hum Genet*. 2013;93(6):1072-1086.
10. Steinthorsdottir V, Thorleifsson G, Sulem P, et al. Identification of low-frequency and rare sequence variants associated with elevated or reduced risk of type 2 diabetes. *Nat Genet*. 2014;46(3):294-298.
11. Lorenzo C, Williams K, Gonzalez-Villalpando C, Haffner SM. The prevalence of the metabolic syndrome did not increase in Mexico City between 1990-1992 and 1997-1999 despite more central obesity. *Diabetes Care*. 2005;28(10):2480-2485.
12. Hunt KJ, Gonzalez ME, Lopez R, Haffner SM, Stern MP, Gonzalez-Villalpando C. Diabetes is more lethal in Mexicans and Mexican-Americans compared to Non-Hispanic whites. *Ann Epidemiol*. 2011;21(12):899-906.
13. Kolonel LN, Henderson BE, Hankin JH, et al. A multiethnic cohort in Hawaii and Los Angeles. *Am J Epidemiol*. 2000;151(4):346-357.
14. DePristo MA, Banks E, Poplin R, et al. A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nat Genet*. 2011;43(5):491-498.
15. Falconer DS. The inheritance of liability to diseases with variable age of onset, with particular reference to diabetes mellitus. *Ann Hum Genet*. 1967;31(1):1-20.
16. Zaitlen N, Pasaniuc B, Patterson N, et al. Analysis of case-control association studies with known risk variants. *Bioinformatics*. 2012;28(13):1729-1737.
17. Kang HM, Sul JH, Service SK, et al. Variance component model to account for sample structure in genome-wide association studies. *Nat Genet*. 2010;42(4):348-354.
18. Lee S, Wu MC, Lin X. Optimal tests for rare variant effects in sequencing association studies. *Biostatistics*. 2012;13(4):762-775.
19. Li B, Leal SM. Methods for detecting associations with rare variants for common diseases. *Am J Hum Genet*. 2008;83(3):311-321.
20. Björkhaug L, Sagen JV, Thorsby P, Søvik O, Molven A, Njølstad PR. Hepatocyte nuclear factor-1 alpha gene mutations and diabetes in Norway. *J Clin Endocrinol Metab*. 2003;88(2):920-931.
21. Aukrust I, Björkhaug L, Negahdar M, et al. SUMOylation of pancreatic glucokinase regulates its cellular stability and activity. *J Biol Chem*. 2013;288(8):5951-5962.
22. Björkhaug L, Ye H, Horikawa Y, Søvik O, Molven A, Njølstad PR. MODY associated with two novel hepatocyte nuclear factor-1alpha loss-of-function mutations (P112L and Q466X). *Biochem Biophys Res Commun*. 2000;279(3):792-798.
23. Abecasis GR, Auton A, Brooks LD, et al; 1000 Genomes Project Consortium. An integrated map of genetic variation from 1,092 human genomes. *Nature*. 2012;491(7422):56-65.
24. Tennessen JA, Bigham AW, O'Connor TD, et al. Evolution and functional impact of rare coding variation from deep sequencing of human exomes. *Science*. 2012;337(6090):64-69.
25. Hara K, Fujita H, Johnson TA, et al; DIAGRAM consortium. Genome-wide association study identifies three novel loci for type 2 diabetes. *Hum Mol Genet*. 2014;23(1):239-246.
26. Mitchell BD, Kammerer CM, Blangero J, et al. Genetic and environmental contributions to cardiovascular risk factors in Mexican Americans. *Circulation*. 1996;94(9):2159-2170.
27. Hanis CL, Ferrell RE, Barton SA, et al. Diabetes among Mexican Americans in Starr County, Texas. *Am J Epidemiol*. 1983;118(5):659-672.
28. Flannick J, Beer NL, Bick AG, et al. Assessing the phenotypic effects in the general population of rare variants in genes for a dominant Mendelian form of diabetes. *Nat Genet*. 2013;45(11):1380-1385.
29. Ellard S. Hepatocyte nuclear factor 1 alpha (HNF1 alpha) mutations in maturity-onset diabetes of the young. *Hum Mutat*. 2000;16(5):377-385.
30. Molven A, Njølstad PR. Role of molecular genetics in transforming diagnosis of diabetes mellitus. *Expert Rev Mol Diagn*. 2011;11(3):313-320.
31. Eide SA, Raeder H, Johansson S, et al. Prevalence of HNF1A (MODY3) mutations in a Norwegian population. *Diabetic Med*. 2008;25(7):775-781.
32. Kropff J, Selwood MP, McCarthy MI, Farmer AJ, Owen KR. Prevalence of monogenic diabetes in young adults. *Diabetologia*. 2011;54(5):1261-1263.
33. Ledermann HM. Maturity-onset diabetes of the young (MODY) at least ten times more common in Europe than previously assumed? *Diabetologia*. 1995;38(12):1482.
34. Shields BM, Hicks S, Shepherd MH, Colclough K, Hattersley AT, Ellard S. Maturity-onset diabetes of the young (MODY)? *Diabetologia*. 2010;53(12):2504-2508.
35. Bellanné-Chantelot C, Carette C, Riveline JP, et al. The type and the position of HNF1A mutation modulate age at diagnosis of diabetes in patients with maturity-onset diabetes of the young (MODY)-3. *Diabetes*. 2008;57(2):503-508.
36. Forlani G, Zucchini S, Di Rocco A, et al. Double heterozygous mutations involving both HNF1A/MODY3 and HNF4A/MODY1 genes. *Diabetes Care*. 2010;33(11):2336-2338.
37. Hegele RA, Cao H, Harris SB, Hanley AJ, Zinman B. The hepatic nuclear factor-1alpha G319S variant is associated with early-onset type 2 diabetes in Canadian Oji-Cree. *J Clin Endocrinol Metab*. 1999;84(3):1077-1082.
38. Kristinsson SY, Thorolfsson ET, Talseth B, et al. Maturity-onset diabetes in Iceland is associated with mutations in HNF-1alpha and a novel mutation in NeuroD1. *Diabetologia*. 2001;44(11):2098-2103.
39. Zulk O, Schaffner SF, Samocha K, et al. Searching for missing heritability. *Proc Natl Acad Sci U S A*. 2014;111(4):455-464.
40. Altshuler D, Daly MJ, Lander ES. Genetic mapping in human disease. *Science*. 2008;322(5903):881-888.
41. Winkler W, Weedon MN, Graham RR, et al. Evaluation of common variants in the six known maturity-onset diabetes of the young (MODY) genes for association with type 2 diabetes. *Diabetes*. 2007;56(3):685-693.
42. Fang QC, Zhang R, Wang CR, Lin X, Xiang KS. Scanning HNF-1 alpha gene mutation in Chinese early-onset and/or multiplex diabetes pedigrees [in Chinese]. *Zhonghua Yi Xue Yi Chuan Xue Za Zhi*. 2004;21(4):329-334.
43. Nishigori H, Yamada S, Kohama T, et al. Mutations in the hepatocyte nuclear factor-1 alpha gene (MODY3) are not a major cause of early-onset non-insulin-dependent (type 2) diabetes mellitus in Japanese. *J Hum Genet*. 1998;43(2):107-110.
44. Tonooka N, Tomura H, Takahashi Y, et al. 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*. 2002;45(12):1709-1712.
45. Yorifuji T, Fujimaru R, Hosokawa Y, et al. Comprehensive molecular analysis of Japanese patients with pediatric-onset MODY-type diabetes mellitus. *Pediatr Diabet*. 2012;13(1):26-32.
46. Søvik O, Njølstad P, Følling I, Sagen J, Cockburn BN, Bell GI. Hyperexcitability to sulphonylurea in MODY3. *Diabetologia*. 1998;41(5):607-608.
47. Shepherd M, Shields B, Ellard S, Rubio-Cabezas O, Hattersley AT. A genetic diagnosis of HNF1A diabetes alters treatment and improves glycaemic control in the majority of insulin-treated patients. *Diabet Med*. 2009;26(4):437-441.
48. Pearson ER, Starkey BJ, Powell RJ, Gribble FM, Clark PM, Hattersley AT. Genetic cause of hyperglycaemia and response to treatment in diabetes. *Lancet*. 2003;362(9392):1275-1281.
49. Baumhueter S, Mendel DB, Conley PB, et al. HNF-1 shares three sequence motifs with the POU domain proteins and is identical to LF-B1 and APF. *Genes Dev*. 1990;4(3):372-379.
50. Choudat T, Blumenfeld M, Bach I, Vandekerckhove J, Cereghini S, Yaniv M. A distal dimerization domain is essential for DNA-binding by the atypical HNF1 homeodomain. *Nucleic Acids Res*. 1990;18(19):5853-5863.
51. Tronche F, Yaniv M. HNF1, a homeoprotein member of the hepatic transcription regulatory network. *BioEssays*. 1992;14(9):579-587.

Supplementary Online Content

The SIGMA Type 2 Diabetes Consortium. Association of a Low-Frequency Variant in *HNF1A* With Diabetes Prevalence in a Latino population. *JAMA*. doi:10.1001/jama.2014.6511

SUPPLEMENTARY METHODS	2
Study participants of the discovery stage	2
Genetic studies	4
Selection of samples for exome-sequencing	4
Whole-exome sequencing of SIGMA samples	4
Variant calling of exome-sequence data.....	4
Sequencing or genotyping of individuals used in the replication stage.....	5
Statistical analyzes	6
Functional characterization of <i>HNF1A</i> p.E508K variant	7
SUPPLEMENTARY FIGURES	9
eFigure 1. Principal component analysis including parental samples.	10
eFigure 2. Proportions of Native American ancestry.	11
eFigure 3. Principal component analysis of exome-sequenced samples in SIGMA.	12
eFigure 4. Quantile-Quantile plot of observed vs expected test statistics	13
eFigure 5. Regional association plot of the <i>HNF1A</i> locus.	14
eFigure 6. Transactivation and subcellular localization experiments	16
SUPPLEMENTARY TABLES	17
eTable 1. Study descriptives of replication studies	17
eTable 2. Functional annotation of variants identified in the discovery cohort	18
eTable 3. Intersection of known and novel variants ascertained by the SIGMA T2D exome project	18
eTable 4. Discovery stage results for all significant markers in the exome	19
eTable 5. Local ancestry results near the <i>HNF1A</i> p.E508K variant	20
eTable 6. Top burden association tests results for non-synonymous variants with MAF < 1%	21
eTable 7. Top burden tests for loss of function variants with MAF<1%	21
eTable 8. Gene-set association test results for non-synonymous variants with MAF < 1%	22
eTable 9. Gene-set association test results for non-synonymous variants with MAF < 1%	22
SUBCONSORTIA INVESTIGATORS	23
REFERENCES	28

This supplementary material has been provided by the authors to give readers additional information about their work.

© 2013 American Medical Association. All rights reserved.

SUPPLEMENTARY METHODS

Study participants of the discovery stage

Diabetes in Mexico Study (DMS):

Participants were recruited between 2010 and 2011 from two tertiary level institutions (IMSS and ISSSTE) located in Mexico City.¹ Phenotyping was done centrally and type 2 diabetes (T2D) was diagnosed based on American Diabetes Association (ADA) criteria. 811 unrelated healthy subjects older than 45 years and with fasting glucose levels below 100 mg/dL were classified as controls. 569 unrelated individuals, older than 18 years, with either previous T2D diagnosis or fasting glucose levels above 125 mg/dL were included as T2D cases. Individuals with fasting glucose levels between 100-125 mg/dL were excluded. Informed consent was obtained from all participants. The study was conducted with the approval of the Ethics and Research Committees of all institutions involved. Genomic DNA was purified from whole blood samples using a modified salting-out precipitation method (Genra Puregene, Qiagen Systems, Inc., Valencia, CA, USA).

Mexico City Diabetes Study (MCDS):

The Mexico City Diabetes Study is a population based prospective investigation.^{2,3} All 35-64 years of age men and non-pregnant women residing in the study site (low income neighborhoods equivalent to 6 census tracts with a total population of 15,000 inhabitants) were interviewed and invited to participate in the study. We had a response rate of 67% for the initial exam. Participant follow-up information included in this study was collected in 2008. Diagnostic criteria for T2D were as recommended by the ADA: a fasting glucose of 126 mg/dL or greater, or a 2 hr post 75 gr of glucose load of 200 mg/dl or greater. If a participant was diagnosed with diabetes by a physician and was under pharmacologic therapy for diabetes s/he was considered as having diabetes regardless the blood glucose levels. The study was conducted with the approval of the Ethics and Research Committees of all institutions. Informed consent was obtained from all participants. Genomic DNA was extracted from whole blood using the QIAmp 96 DNA Blood Ki5 (12) (Qiagen, Cat. No. 51162).

Multiethnic Cohort (MEC):

The MEC consists of 215,251 men and women in Hawaii and Los Angeles, and is comprised of mainly five self-reported racial/ethnic populations: African Americans, Japanese Americans, Latinos, Native Hawaiians and European Americans⁴. Between 1993 and 1996, adults between 45 and 75 years old were enrolled by completing a 26-page, self-administered questionnaire asking detailed information about dietary habits, demographic factors, level of education, personal behaviors, and history of prior medical conditions (e.g., diabetes). Potential cohort members were identified through Department of Motor Vehicles drivers' license files, voter registration files and Health Care Financing Administration data files. In 2001, a short follow-up questionnaire was sent to update information on dietary habits, as well as to obtain information about new diagnoses of medical conditions since recruitment. Between 2003 and 2007, we re-administered a modified version of the baseline questionnaire. All questionnaires inquired about history of diabetes, without specification as to type (1 vs. 2). Between 1995 and 2004, blood specimens were collected from ~67,000 MEC participants at which time a short questionnaire was administered to update certain exposures, and collect current information about medication use.

Cohort members in California are linked each year to the California Office of Statewide Health Planning and Development (OSHPD) hospitalization discharge database which consists of mandatory records of all in-patient hospitalizations at most acute-care facilities in California. Records include information on the principal diagnosis plus up to 24 other diagnoses (coded according to ICD-9), including type 1 diabetes (T1D) and T2D. In Hawaii cohort members have been linked with the diabetes care registries for subjects with Hawaii Medical Service Association (HMSA) and Kaiser Permanente Hawaii (KPH) health plans (~90% of the Hawaii population has one of these two plans). Information from these additional databases has been utilized to assess the percentage of T2D controls (as defined below) with undiagnosed T2D, as well as the percentage of identified diabetes cases with T1D rather than T2D. Based on the OSHPD database <3% of T2D cases had a previous diagnosis of T1D. We did not use these sources

to identify T2D cases because they did not include information on diabetes medications, one of our inclusion criteria for cases (see below).

In the MEC, diabetic cases were defined using the following criteria: (a) a self-report of diabetes on the baseline questionnaire, 2nd questionnaire or 3rd questionnaire; and (b) self-report of taking medication for T2D at the time of blood draw; and (c) no diagnosis of T1D in the absence of a T2D diagnosis from the OSHPD (California Residents). Controls were defined as: (a) no self-report of diabetes on any of the questionnaires while having completed a minimum of 2 of the 3 (~80% of controls returned all 3 questionnaires); and (b) no use of medications for T2D at the time of blood draw; and (c) no diabetes diagnosis (type 1 or 2) from the OSHPD, HMSA or KPH registries. To preserve DNA for genetic studies of cancer in the MEC, subjects with an incident cancer diagnosis at time of selection for this study were excluded. Controls were frequency matched to cases on sex, ethnicity and age at entry into the cohort (5-year age groups) and for Latinos, place of birth (U.S. vs. Mexico, South or Central America), oversampling African American, Native Hawaiian and European American controls to increase statistical power. Many of the T2D variants have also been evaluated in studies of cancer in the MEC which allowed for inclusion of additional controls who met the criteria above. Informed consent was obtained from all participants. The study was conducted with the approval of the Ethics and Research Committees of all institutions. Genomic DNA extraction was done using Qiagen from buffy coat.

UNAM/INCMNSZ Diabetes Study (UIDS):

Cases were recruited between 2011 and 2013 at the outpatient diabetes clinic of the Department of Endocrinology and Metabolism of the Instituto Nacional de Ciencias Médicas y Nutrición Salvador Zubirán (INCMNSZ)¹. All Mexican-mestizo individuals were invited to participate in the study. Diagnosis of T2D was done centrally following the ADA criteria, i.e., fasting plasma glucose values ≥ 126 mg/dL, current treatment with a hypoglycemic agent, or casual glucose values ≥ 200 mg/dL. Control subjects were recruited from 2007 to 2013 from a cohort of adults aged 45 years or older among government employees, blue collar workers and subjects seeking for attention in medical units for any condition besides those considered as exclusion criteria (see below). Normoglycemic status was defined as having a fasting plasma glucose concentration < 100 mg/dl and no previous history of hyperglycemia, gestational diabetes or use of metformin.

Patients were interviewed following a standardized questionnaire; it included the medical history, a previously validated, three days food record and a physical activity registry. In addition a blood sample (after 9-12 hours of fasting) was obtained. The questionnaire included demographic, socio-economic and medical history of the patients and their family. Blood pressure, height, waist circumference and weight must be measured in the same visit. For taking blood pressure, systolic and diastolic pressure were recorded using a mercury sphygmomanometer; subjects remained seated and at rest for five minutes before measuring.

Inclusion criteria: Men or women aged 25 years or older, with BMI greater than 20 but lower than 40 kg/m².

Exclusion criteria: Diabetes, coronary heart disease, stroke, transient ischemic attack, lower limb amputations, alcoholism (more than 10 servings of alcohol per week) or any disease that in opinion of the researcher may limit life expectancy to less than 2 years. Subjects that planned to move out of town permanently during the next three years were also excluded. Pregnant women, individuals with drug addictions, the use of systemic corticosteroids in pharmacologic doses (intravenous, oral or injectable, including injections in the joints) were exclusion criteria also. Replacement dosage of systemic corticosteroids (up 7.5 mg/day of prednisone or 30 mg/day of hydrocortisone or its equivalent; as well as inhaled or topical corticosteroids) was allowed into the study. Other exclusion criteria were: active liver disease (defined as AST (SGOT) or ALT (SGPT) > 2.0 x upper limit of the normal range, alkaline phosphatase (ALK-P) > 1.5 x upper limit of the normal range or total bilirubin > 1.5 x upper limit of the normal range), significant renal dysfunction (defined as serum creatinine > 1.7 upper limit of the normal

range or nephrotic syndrome), any history of malignancy (except for basal cell skin carcinoma) and uncontrolled depression or psychosis.

Informed consent was obtained from all participants. The study was conducted with the approval of the Ethics and Research Committees of all institutions. Genomic DNA was extracted from whole blood using the QIAmp 96 DNA Blood Ki5 (12) (Qiagen, Cat. No. 51162).

We used a Mann-Whitney U test, also called Wilcoxon test, to compare the clinical characteristics between the p.E508K carriers and non-carriers, as well as to compare T2D cases vs controls.

Genetic studies

Selection of samples for exome-sequencing

All sequenced individuals are part of a larger Mexican and Latinos cohort, consisting of 8,214 subjects. Genome-wide association study (GWAS) single nucleotide polymorphism (SNP) data have been collected on all samples, using the Illumina Omni 2.5 platform. Samples with gender mismatches based on X chromosome, duplicate samples, samples with high African or East Asian ancestry, or samples with more than 10% of relatedness were previously removed.¹ From this cohort, we selected 3,862 individuals for exome sequencing. We selected samples from each cohort to maximize the percentage of Native American ancestry while keeping ~ 1,000 samples in each of the four studies included in the discovery stage. To estimate the percentage of Native American ancestry we first calculated principal components projecting SIGMA samples onto HGDP Yoruba, French, Karitiana and Han (Chinese) populations to obtain global ancestry estimates. Then we used the extreme values to define the range of possible ancestry proportions (0 to 1).

Whole-exome sequencing of SIGMA samples

Exome sequencing was performed at the Broad Institute's Genomics Platform. For input DNA we used >250 ng of DNA, at >2ng/μl. Our exome-sequencing pipeline included sample plating, library preparation (2-plexing of samples per hybridization), hybrid capture, sequencing (76bp paired reads), sample identification QC check, and data storage. The exome sequencing data was de-multiplexed and each sample's sequence data were aggregated into a single Picard BAM file. Reads were mapped to the human genome hg19 with the BWA algorithm⁵ and processed with the Genome Analysis Toolkit (GATK)⁶ to recalibrate base quality-scores and perform local realignment around known indels.

Variant calling of exome-sequence data

Target coverage for each sample was computed with the GATK. Single nucleotide variants (SNVs) and small insertions and deletions (indels) were called with the Unified Genotyper module of the GATK and filtered to remove SNVs with annotations indicative of technical artifacts (such as strand-bias, low variant call quality, or homopolymer runs). We applied default filters to SNP and indel calls using the GATK's Variant Quality Score Recalibration (VQSR) approach. Samples with fewer than 76% of targeted bases covered to 20x, with an abnormally high number of non-reference alleles or heterozygosity, or with an abnormally low concordance with prior SNP array genotypes (based on the distribution across all samples) were excluded from analysis. Any sample genotype at a site with fewer than 10x coverage in the sample was ignored (e.g. set as missing). Variants were annotated with the Variant Effect Predictor.⁷ SNPs with differential call rates between cases and controls ($p < 5 \times 10^{-6}$), less than 50% call rate, or with a Hardy-

Weinberg equilibrium p-value $< 5 \times 10^{-8}$ in controls were excluded from association analysis.

Sequencing or genotyping of individuals used in the replication stage

T2D-GENES Study

The exons of *HNFI1A* were sequenced in 13,098 additional individuals as part of the whole-exome sequencing studies performed through the Genetics of Type 2 Diabetes (GoT2D) and Type 2 Diabetes Genetic Exploration by Next-generation sequencing in multi-Ethnic Samples (T2D-GENES) consortia. Individuals were selected spanning 5 ethnicities: European (the FUSION study⁸ [FUSION], the METSIM study⁹ [METSIM], KORA-gen¹⁰ [KORA], the WTCCC/UKT2D consortium^{11,12} and the UK Adult Twin Registry¹³ [UKT2D]), as well as Ashkenazi individuals recruited from the metropolitan New York region¹⁴ [Ashkenazim] and small number of individuals from the Finnish [Botnia] and Swedish [Malmö] prospective cohorts used for the initial sequencing experiment¹⁵⁻²³), African-American (the Jackson Heart Study (JHS) cohort [JHS] as well as additional individuals recruited from North Carolina, South Carolina, Georgia, Tennessee, or Virginia²³ [WFS]), South Asian (the London Life Sciences Prospective Population Study (LOLIPOP)^{24,25} [LOLIPOP] and Singapore Indian Eye Study (SINDI)²⁶ [Singapore Indians]), East Asian (the Korean Association Resource (KARE)²⁷ [KARE] as well as the Singapore Diabetes Cohort Study (SDCS) and Singapore Prospective Study Program²⁸⁻³⁰ [Singapore Chinese]), and Latinos/Hispanic (the San Antonio Family Heart Study (FHS)³¹, the San Antonio Family Diabetes/Gallbladder Study (SAFDGS)³², the Veterans Administration Genetic Epidemiology Study (VAGES)³³, the Family Investigation of Nephropathy and Diabetes (FIND)³⁴, San Antonio component [San Antonio], and individuals from Starr County, TX³⁵ [Starr County]). Data generation and processing was performed in an identical fashion as for the initial sequencing experiment, although target capture was performed with the Agilent SureSelect Human All Exon platform rather than a custom hybrid capture array.

Diabetes in Mexico Study 2 (DMS2):

Participants were recruited between 2011 and 2012 from 18 different ethnic groups along Mexico (Tarahumara, Yaqui, Mayo, Mixteco, Nahuatl, Otomi, Chinanteco, Mixe, Zapoteco, Mazateco, Totonaco, Huasteco, Maya, Kanjobal, Mame, Poptijaltec, Kaqchikel, Tojolabal). Inclusion criteria were that they identified themselves as indigenous, both parents or their four grandparents speak the same native language and were born in the same community. Phenotyping was done centrally and diagnosis of T2D was made based on ADA criteria. A total of 751 unrelated healthy subjects older than 45 years and with fasting glucose levels below 100 mg/dL were classified as controls. We also included 427 unrelated individuals, older than 18 years with either previous T2D diagnosis or fasting glucose levels above 125 mg/dL, as T2D cases. Individuals with fasting glycemia between 100-125 mg/dL were excluded. Informed consent was obtained from all participants. The study was conducted with the approval of the Ethics and Research Committees of all institutions involved. Genomic DNA was purified from whole blood samples using a modified salting-out precipitation method (Gentra Puregene, Qiagen Systems, Inc., Valencia, CA, USA). Genotyping of the *HNFI1A* p.E508K variant was performed using a custom TaqMan SNP Genotyping Assay (Applied Biosystems, Foster City, CA, USA) and genotype of each sample was assigned automatically by SDS 2.3 software (Applied Biosystems, Foster City, CA, USA). For the genotyping quality control, 5% of samples were randomly selected and measured in duplicates. TaqMan probes: Allele 'G' - **TT**CACAAGCCCGAGGTG. Allele 'A' FAM- CACAAGCCCAAGGTG. Five positive controls were added to all plates and verified that their genotype matched the expected.

Statistical analyzes

Single variant analysis

Association analysis of each single variant was computed first in all SIGMA samples as well as replication analysis of p.E508K in the additional sequenced individuals (from the T2D-Genes consortium) were done using a mixed linear model implemented in EMMAX on the LTsoft transformed phenotype.³⁶ EMMAX implements a mixed linear model that accounts for different layers of sample structure, including population stratification and sample relatedness. A kinship matrix was first computed using independent SNPs (MAF >1%) using EMMAX; association p-values were then computed for p.E508K. As EMMAX and LTsoft transformed phenotypes do not produce effect size estimates for dichotomous traits, point estimates for odds ratios were computed via a standard logistic regression as implemented in PLINK³⁷ with age, gender, BMI and 10 principal components as covariates (computed via the EIGENSTRAT³⁸ software package from the same SNPs as for the kinship matrix). These were then transformed into 95% Wald confidence intervals using standard error estimates back-calculated from the p-values produced by the linear mixed model. In the DMS2 dataset a linear regression implemented on PLINK was used on the LTsoft transformed phenotype to estimate significance. The 95% OR confidence intervals were estimated in the same way described above.

The resulting association statistics from the discovery and replication studies were combined via an inverse variance based fixed-effects meta-analysis (as implemented in the METAL software package³⁹) to obtain an estimated odds ratio and *P*-value for association of p.E508K with type 2 diabetes.

Local ancestry analysis

Local ancestry estimation was performed using LAMP-LD⁴⁰ version 1.0 with previously generated array data on the same SIGMA samples¹. Panels for inference included a diverse collection of 227 Native American samples from Central America and Mexico^{41,42}, 72 Southern Europeans from HGDP and 12 Spanish individuals⁴³, and 109 Yoruba Africans (YRI) from HapMap⁴⁴. Prior to performing local ancestry inference, the panels were merged, yielding an intersected SNP set of 252,402 markers. The panels were then jointly phased using SHAPEIT⁴⁵ version 1.532. Next the SNPs from the panels and the SIGMA data were intersected, yielding 235,660 SNPs, and LAMP-LD was run to infer local ancestry.

Conditional analyzes

Common variants (rs1169288, rs7957197) in *HNFI1A* have previously been reported to be associated with T2D.^{11,46} We examined both variants in our cohort and found neither to be significantly associated with T2D risk (rs1169288 *P*=0.07, rs7957197 *P*=0.19).

We conducted statistical conditional analyzes to evaluate the influence of known T2D variants (rs7957197, rs1169288) in the *HNFI1A* gene to the novel variant. Genotypes from the same subjects within GWAS dataset previously described¹ were used for rs7957197. Sixty one samples did not have genotype information for this variant and therefore the conditional analysis was based on 3730 samples. The conditional analyzes was carried out with a logistic regression using age, sex, BMI, 10 PCs and the variant to condition on implemented in PLINK. The effect of *HNFI1A* p.E508K was unchanged after adjustment for each of these variants and therefore represents an independent signal (data not shown).

Burden and Gene-set tests

In addition to single variant testing, Sequence Kernel Association Test (SKAT)⁴⁷ and collapsing tests⁴⁸ were used to test the possibility of genes and groups of genes contributing to disease susceptibility via combinations of rare variants (burden tests). Burden tests included rare (MAF <1%) non-synonymous and/or loss-of-function variants in up to 15,469 genes.

We also tested the cumulative effect of these potentially functional variants in two gene-sets: 13 MODY genes: *HNFI1A*, *HNFI1B*, *KCNJ11*, *ABCC8*, *BLK*, *INS*, *NEUROD1*, *PDX1*, *INS-IGF2*, *HNFI4A*, *GCK*, *KLF11*, *CEL*; and a second gene-set of genes in 70 previously implicated T2D loci: *ADAMTS9*, *ADCY5*, *ANK1*, *ANKRD55*, *AP3S2*, *ARAPI1*, *BCAR1*, *BCL11A*, *BCL2*, *C2CD4A*, *CAMK1D*, *CCND2*, *CDC123*, *CDKAL1*, *CDKN2A*, *CDKN2B*, *CILP2*, *DGKB*, *DUSP8*, *DUSP9*, *FITM2*, *FTO*, *GCC1*, *GCK*, *GCKR*, *GIPR*, *GLIS3*, *GRB14*, *GRK5*, *HHEX*, *HMG20A*, *HMGA2*, *HNFI1A*, *HNFI1B*, *HNFI4A*, *IDE*, *IGF2BP2*, *IRSI*, *JAZF1*, *KCNJ11*, *KCNK16*, *KCNQ1*, *KLF14*, *KLHDC5*, *LAMA1*, *LGR5*, *MACF1*, *MAEA*, *MC4R*, *MTNR1B*, *NOTCH2*, *PEPD*, *PPARG*, *PRC1*, *PROX1*, *PSMD6*, *PTPRD*, *R3HDML*, *RASGRP1*, *RBMS1*, *RND3*, *SGCG*, *SLC16A11*, *SLC16A13*, *SLC30A8*, *SPRY2*, *SRR*, *ST6GAL1*, *TCF7L2*, *THADA*, *TLE1*, *TLE4*, *TMEM163*, *TP53INP1*, *TSPAN8*, *UBE2E2*, *VPS26A*, *WFS1*, *ZBED3*, *ZFAND3*, *ZFAND6*, *ZMIZ1*.⁴⁹ Non-synonymous and loss-of-function variants with a minor allele frequency <1% were extracted from these genes and were used as input for association with T2D.

Functional characterization of *HNFI1A* p.E508K variant

Plasmids, Cell Culture and Transfections

We used the human liver *HNFI1A* cDNA in the expression vector pcDNA3.1/HisC for all cell studies⁵⁰. *HNFI1A* mutants were made using the QuikChange Site-Directed XL Mutagenesis Kit (Stratagene). All sequences were verified by Sanger DNA sequencing. The firefly luciferase reporter gene construct pGL3-RA, containing the promoter of the rat albumin gene, and the pRL-SV40 reporter vector encoding the renilla luciferase gene, was kindly provided by Professor Graeme I. Bell, University of Chicago, Chicago, IL. The reporter gene constructs pGL3-HNF4AP2 and pGL3-GLUT2 (containing human *HNFI4A* P2 promoter and mouse *Glut2* promoter, respectively) were kindly provided by Dr. Maria-Angeles Navas, Complutense University of Madrid, Spain.⁵¹ HeLa cells and MIN6 β -cells were grown as previously described^{50,52} and transfected using the Metafectene Pro (Biontexas-USA, Dan Diego, CA) or Lipofectamine 2000 (Life Technologies, Carlsbad, CA), respectively.

Transactivation and Protein Expression Analyses

HeLa or MIN6 cells were transiently transfected with nonmutant and/or mutant *HNFI1A* plasmids together with pRL-SV40 and the reporter plasmids pGL3-RA, pGL3-HNF4AP2 or pGL3-GLUT2. Transcriptional activity was measured 24 h after transfection using the Dual-Luciferase Reporter Assay System (Promega Biotech, Madison, WI) on a Chameleon luminometer (Hidex, Turku, Finland). To measure expression levels of nonmutant and p.E508K HNF-1A proteins, we analyzed cell lysates (2.5 μ g total protein) from transfected HeLa cells lysed in passive lysis buffer (Promega Biotech) by SDS-PAGE and immunoblotting using an anti-Xpress antibody (Life Technologies). HNF-1A protein expression levels were quantitated by densitometric analyses and were normalized to actin (Santa Cruz Biotechnology, Dallas, TX).

DNA Binding Studies

HNF-1A proteins were expressed in an *in vitro* coupled transcription/translation system (TnT-T7 transcription/translation system, Promega Biotech). The level of HNF-1A binding to a radiolabeled rat albumin oligonucleotide was investigated by electrophoretic mobility shift assays (EMSA) as previously described.⁵³ Protein-DNA samples were analyzed by 6% non-denaturing polyacrylamide gel electrophoresis and subsequent autoradiography. We quantified the level of binding by measuring the intensity of the HNF-1A-oligonucleotide complexes. Competition assays were performed by adding increasing amounts of non-labeled oligonucleotides and supershift analyses by the addition of HNF-1A-tag specific antibody (anti-Xpress antibody) to the binding reaction.

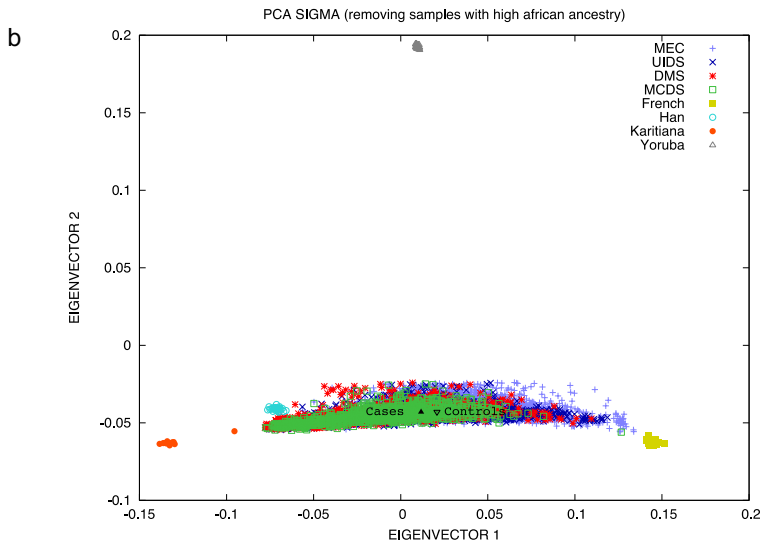
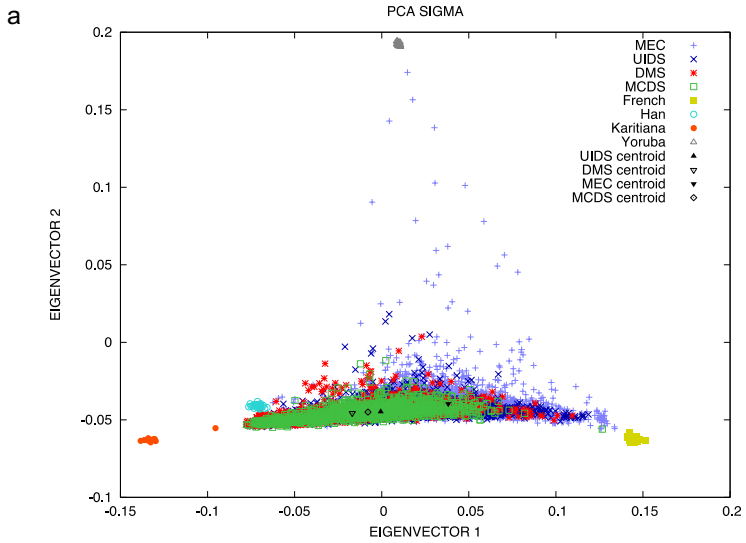
Immuofluorescence

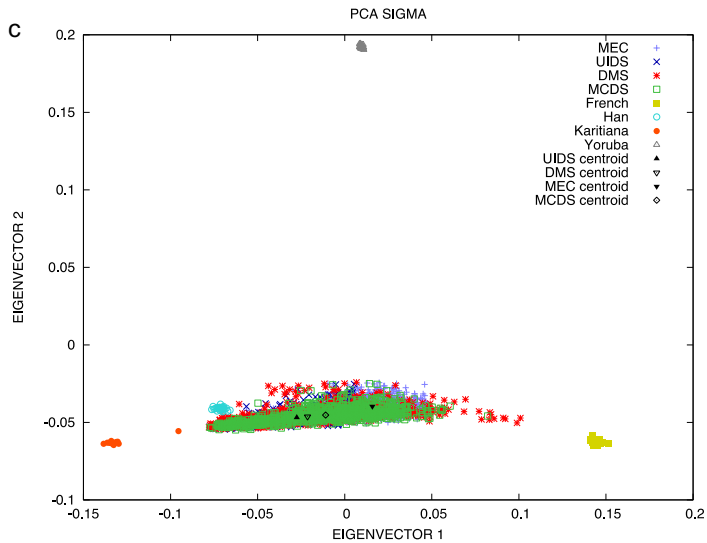
We transiently transfected HeLa and MIN6 cells with nonmutant or mutant *HNF1A*. The p.Q466X mutant was included as a positive control for impaired nuclear localization in HeLa cells, as previously reported.⁵³ The nuclear/cytosol distribution of HNF-1A proteins was detected by the HNF-1A-tag (anti-Xpress) antibody and Alexa Fluor 488 (green) essentially as reported before.⁵⁰

Statistical Analysis

All data are expressed as means \pm SD, and experiments were performed at least on three independent occasions unless otherwise specified. Statistical analyses were performed using 2-tailed Student's *t* test, and a P value <0.05 was considered significant.

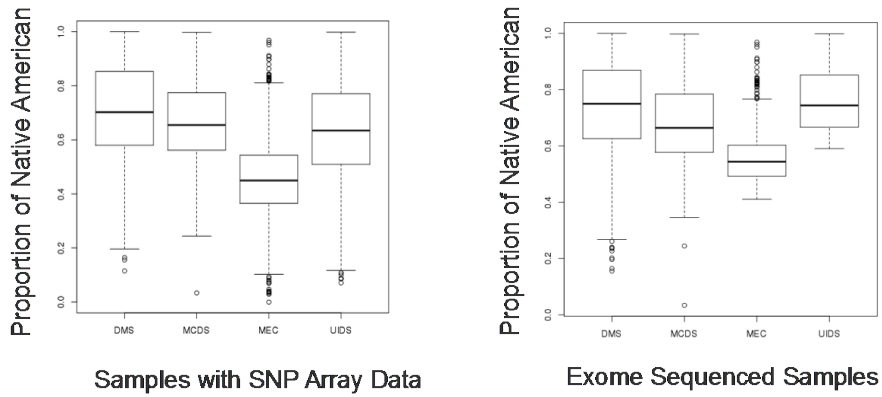
Supplementary Figures





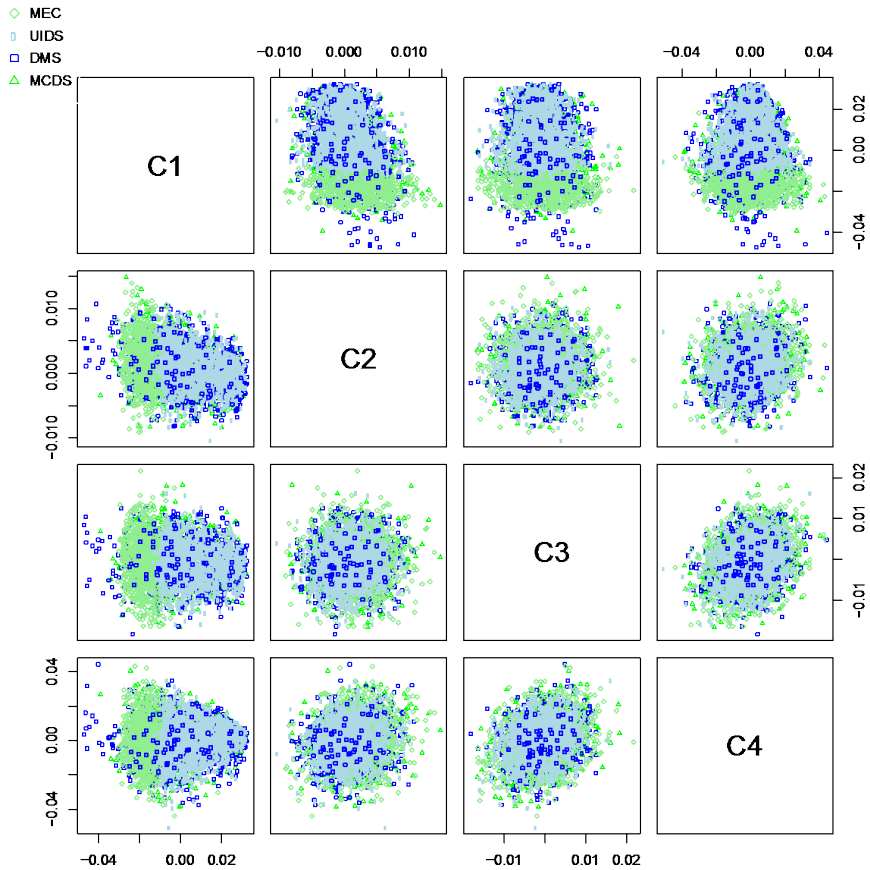
eFigure 1. Principal component analysis including parental samples.

Principal components calculated using data from samples collected by the Human Genome Diversity Project (HGDP) and 1000 Genomes Project. PCA projection of SIGMA onto HGDP Yoruba, French, Karitiana and Han (Chinese) populations before ancestry quality control filters were applied (a), with cohort centroids as indicated, and after all quality control filters were applied (b), and after selection of individuals with highest Native American ancestry (c), this dataset was used for whole-exome sequencing (n=3,756)..



eFigure 2. Proportions of Native American ancestry.

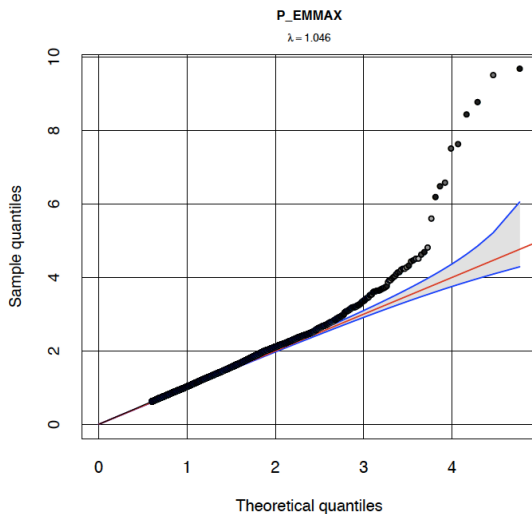
Boxplots representing proportions of Native American ancestry for each SIGMA study included in the discovery stage. A. SIGMA samples before selection of high Native American ancestry. B. SIGMA samples after selection of high Native American ancestry (n=3,756). Convention for box-and-whisker plots: the central horizontal lines indicate the median; the extremes of the boxes indicate the 1st and 3rd quartile; the top whisker indicate maximum value after removing outliers; bottom whisker indicate minimum value after removing outliers; outliers are represented as unfilled circles.



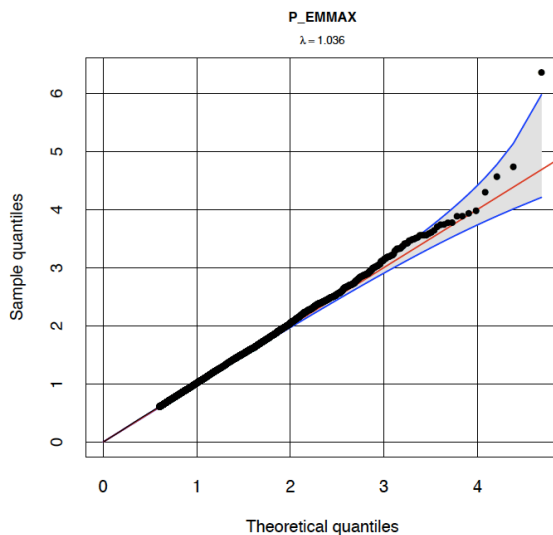
eFigure 3. Principal component analysis of exome-sequenced samples in SIGMA.

Scatterplots for different combinations of principal components (i.e., C1 vs C2; C1 vs C3; C2 vs C3; C3 vs C4). Each dot represents a participant of each one of the four SIGMA studies included in the discovery stage (n=3,756). Figures in the same columns and rows share the same axes and are therefore not shown.

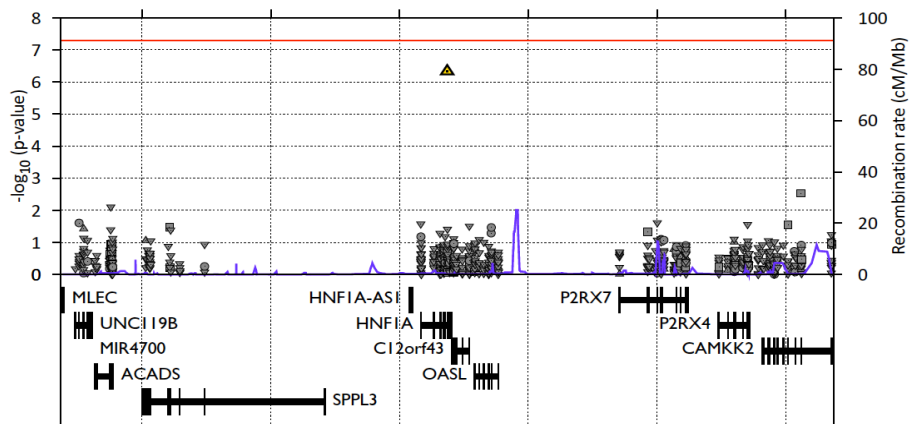
a All variant associations for T2D (LT-SOFT): MAF > .05



b All variant associations for T2D (LT-SOFT): MAC > 15, MAF < .05



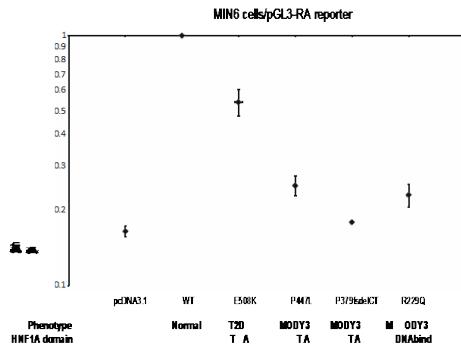
eFigure 4. Quantile-Quantile plot of observed vs expected test statistics
Quantile-Quantile (QQ) plot showing observed (y-axis) vs expected (x-axis) of the test statistics of the SIGMA T2D discovery study (n=3,756). A) Common variants (MAF > 0.05), B) Low frequent and rare variants (MAC > 15 and MAF < 0.05). Abbreviations: T2D, Type 2 Diabetes; LT-SOFT, liability threshold transformed phenotype; MAC, Minor allele count; MAF, Minor allele frequency; P_EMMAX, P value estimated with the software EMMAX.



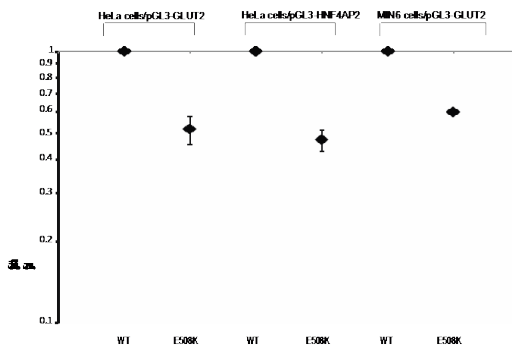
eFigure 5. Regional association plot of the *HNF1A* locus.

The x-axis represents genomic position in chromosome 12 and the y-axis represents the $-\log_{10} P$ of association with type 2 diabetes in the SIGMA discovery cohort ($n=3,756$). Black boxes represent exons interconnected by introns (lines). In purple are local recombination rates. Red horizontal line indicates $P < 5 \times 10^{-8}$.

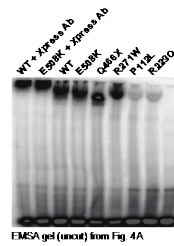
A



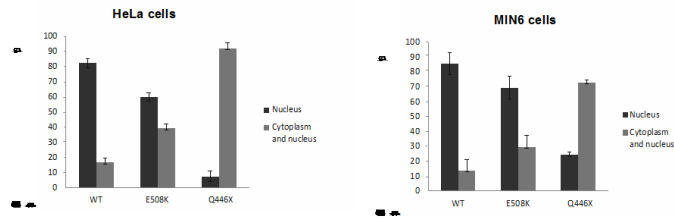
B



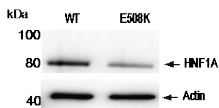
C



D



E



eFigure 6. Transactivation and subcellular localization experiments

A-B. Transcriptional activation of WT and p.E508K HNF-1A as measured by the expression of the firefly luciferase reporter gene in HeLa or MIN6 cells. Cells were transiently transfected with the indicated plasmids together with reporter plasmids pGL3-RA and pRL-SV40. Measurements are given in fold activity relative to wild-type activity. Each point represents the mean (error bars indicate 95% confidence intervals) of nine readings ($n = 9$), for the wild-type and each mutant. All values were statistically significant ($P < .05$) compared to wild-type activity. TA denotes transactivation domain, DNABind the DNA binding domain, WT meaning wild-type and pcDNA3.1 meaning empty pcDNA3.1 vector. B). Transactivation detected with reporter constructs pGL3-GLUT2 or pGL3-HNF4AP2 in HeLa and MIN6 cells. C) Uncut EMSA gel from figure 4A. D) Quantification of the subcellular localization of wild-type and HNF-1A p.E508K in HeLa and MIN6 cells. Error bars indicate the variance in percent (%) number of cells demonstrating cytosolic versus nuclear HNF1A accumulation, obtained from 3 individual experiments. E) Cell lysate from transfected HeLa cells used in transactivation assay in Figure 3 were here analyzed by SDS-PAGE and immunoblotting (anti-Xpress antibody). The level of HNF-1A protein expression was normalized to actin.

SUPPLEMENTARY TABLES

eTable 1. Study descriptives of replication studies

Study / References	Ancestry	Type	Number of samples	Per cent male	Age (years)	Age-of-onset (years)	BMI (kg/m ²)	Fasting Plasma glucose (mmol/l)	
T2D-GENES / Ref 6-33	African-American	controls	1063	40.4	53.5 (11.7)	NA	30.9 (6.8)	4.8 (0.3)	
		cases	1037	36.9	61 (10.2)	47.9 (10.4)	31.3 (6.8)	7.5 (3.2)	
	East-Asian	controls	1159	40.2	60.7 (6.2)	NA	23.3 (3.3)	4.6 (0.5)	
		cases	1018	51.3	55.8 (8.7)	44.9 (9.2)	25.7 (3.5)	7.1 (2.6)	
	European	controls	2247	57.1	64.6 (11.6)	NA	28.7 (5.2)	5.2 (0.6)	
		cases	2415	62.2	58.4 (9.9)	51.4 (7.7)	28.1 (4.7)	8.1 (2.3)	
	Hispanic/Latinos	controls	922	31	42.3 (12.6)	NA	30.2 (6.4)	4.8 (0.5)	
		cases	1016	40.9	56.7 (12)	46.2 (12.1)	31.9 (6.6)	9.8 (3.9)	
	South-Asian	controls	1126	66.8	59.6 (10.3)	NA	26.2 (4.3)	5.2 (0.4)	
		cases	1095	70.2	56.9 (8.9)	50.7 (10.6)	26.8 (4.2)	9.1 (3.2)	
	DMS2 / This paper	Mexican	controls	751	26.4	59.57 (11.42)	NA	27.21 (4.99)	4.77 (0.54)
			cases	427	38.5	57.29 (12.62)	49.15 (11.87)	28.00 (4.71)	9.59 (4.53)

eTable2. Functional annotation of variants identified in the discovery cohort

Variant Category	Before QC	After QC	MAF>=0.01	MAF<0.01	Private (MAC>=2)*
loss_of_function	28 986	18 813	267	18 546	4 945
stop	10 639	9 442	151	9 291	1 681
splice	5 963	4 742	72	4 670	1 047
frameshift	12 384	4 629	44	4 585	2 217
missense	483 018	431 777	20 113	411 664	80 259
probably_Damaging	122 497	111 624	2 252	109 372	21 225
possibly_Damaging	83 696	75 398	2 301	73 097	14 388
benign	254 928	228 636	14 690	213 946	41 213
unknown	22 431	16 785	900	15 885	3 544
coding_syn	300 629	269 071	23 246	245 825	44 493
UTR3prime	32 143	26 157	2 187	23 970	5 666
UTR5prime	26 040	18 167	1 534	16 633	4 667
intronic	511 773	415 133	36 919	378 214	120 549
intergenic	10 782	7 448	672	6 776	2 234
inframe_indels	7 773	2 803	60	2 743	2 021
init_codon	971	827	31	796	161
Total	1 402 115	1 190 196	85 028	1 105 167	264 995

Variants in the SIGMA T2D discovery study (n=3,756). Abbreviations: QC, Quality Control; MAF, Minor Allele Frequency; MAC, Minor allele Count.

*Variants observed at least two times in the SIGMA project (MAC>=2), and not present in 1000 Genomes or exome sequencing project (ESP).

eTable 3. Intersection of known and novel variants ascertained by the SIGMA T2D exome project

Sequencing Project	Number of variants
SIGMA T2D specific	741145
SIGMA T2D specific with a MAC>=2	264,995
1KG specific	36,564,476
ESP specific	806,726
SIGMA T2D and ESP	133,733
SIGMA T2D and 1KG	157,959
1KG and ESP	96,314
SIGMA T2D and 1KG and ESP	180,355

Variants in the SIGMA T2D discovery study (n=3,756) compared to their presence in other public datasets. Abbreviations: SIGMA T2D, SIGMA T2D Exome project; ESP, Exome Sequencing Project, 1KG, 1000 Genomes Project; MAC, Minor allele count

Table 4. Discovery stage results for all significant markers in the exome

ID	GENE	CHROM	POS	Consequence	AA Change	REF	ALT	MAF A (%)	MAF U (%)	OR	95% CI	P
rs13342692	SLC16A11	17	6946287	missense	p.D127G	T	C	42.93	34.18	1.31	1.2 - 1.42	2.08E-10
rs2292351	SLC16A11	17	6946921	5-prime-UTR	NA	C	G	41.33	32.34	1.33	1.22 - 1.45	3.10E-10
rs76070643	SLC16A13	17	6941625	synonymous	p.Y166Y	C	T	41.44	32.76	1.31	1.2 - 1.43	1.68E-09
rs75493593	SLC16A11	17	6945087	missense	p.P443T	G	T	41.31	32.78	1.31	1.2 - 1.43	3.66E-09
rs4796576	SLC16A13	17	6943266	synonymous	p.G422G	G	A	40.37	47.57	0.78	0.72 - 0.85	2.34E-08
rs75418188	SLC16A11	17	6945483	missense	p.G340S	C	T	41.83	33.57	1.32	1.19 - 1.45	3.08E-08
rs117767867	SLC16A11	17	6946330	missense	p.V113I	C	T	40.69	33.01	1.32	1.19 - 1.47	2.64E-07
var_12_121437091	HNF1A	12	121437091	missense	p.E508K	G	A	1.05	0.18	5.48	2.83 - 10.61	4.40E-07

Abbreviations: CHROM, Chromosome; POS, Position in genome build 37; AA, Amino Acid change; REF, Reference allele; ALT, Alternative allele;

MAFA, Minor allele frequency in affecteds; MAFU, Minor allele frequency in unaffected; OR, odds ratio from logistic regression and its 95% confidence interval; P, P-value estimates.

Results from the SIGMA T2D discovery study (n=3,756). The odds ratios are per allele copy for T2D difference in prevalence among cases as compared with controls

eTable 5. Local ancestry results near the HNF1A p.E508K variant

HFN1A E508K	Cases			Controls		
	N=0	N=1	N=2	N=0	N=1	N=2
A=0	207 100.0%	729 98.5%	823 96.8%	282 100.0%	876 99.9%	797 99.3%
A>=1	0 0.00%	11 1.49%	26 3.06%	0 0.00%	1 0.11%	6 0.75%

Native American local ancestry counts (N) stratified by A allele counts (A) at the p.E508K variant from the SIGMA T2D discovery study (n=3,756).

Also shown are percentages of samples within a given genotype class that have the a given Native American local ancestry count.

eTable 6. Top burden association tests results for non-synonymous variants with MAF < 1%

GENE	TRANSCRIPT	COLLAPSE	SKAT	MINA	MINU
BDKRB1	ENST00000216629	0.439	4.31E-05	81	93
HNF1A	ENST00000257555	0.016	5.03E-05	106	86
PRDM9	ENST00000296682	0.179	0.000257	60	51
PTAR1	ENST00000340434	0.0099	0.000356	30	14
SLCO5A1	ENST00000524945	0.697	0.000479	56	70
SLCO5A1	ENST00000260126	0.638	0.000555	64	79
USP29	ENST00000269834	0.00952	0.000767	95	76
ZIM3	NA	0.00952	0.000767	95	76
DEFB1	ENST00000297439	0.00332	0.000804	31	16
C5orf49	ENST00000399810	0.57	0.000824	63	73
CDKL1	ENST00000395834	2.75E-05	0.000963	31	100
SFRP5	ENST00000266066	0.000901	0.00208	63	34
TMEM120B	ENST00000449592	0.000666	0.00268	11	36
HABP2	ENST00000351270	1.99E-05	0.00371	125	78
HABP2	ENST00000542051	2.89E-05	0.00372	124	77
FLVCR2	ENST00000238667	0.000728	0.00415	50	26
SEP2	ENST00000360051	0.000318	0.00437	14	1
RGS9	ENST00000262406	1.12E-05	0.0247	48	111
GIPC2	ENST00000370759	0.00079	0.0265	16	50
NCAPG	ENST00000251496	0.000674	0.0738	55	28
AMBP	ENST00000265132	0.000424	0.0772	52	24

Results from the SIGMA T2D discovery study (n=3,756). Abbreviations: COLLAPSE: Collapsing test; SKAT: Sequence Kernel Association Test; ns, Non-synonymous variants, MINA, sum of total minor allele counts of all variants in the gene in Affecteds; MINU, sum of total minor allele counts of all variants in the gene in Unaffecteds

eTable 7. Top burden tests for loss of function variants with MAF<1%

GENE	TRANSCRIPT	COLLAPSE	SKAT	MINA	MINU
KIF9	ENST00000265529	0.0644	0.00055	6	16
KIF9	ENST00000444589	0.0159	0.000555	4	15
LNX1	ENST00000263925	0.00314	0.00223	44	14
FAM166B	NA	0.00191	0.00242	26	4
RUSC2	ENST00000399742	0.00191	0.00242	26	4
ZNF772	ENST00000343280	0.00289	0.00296	9	1
SMC5	ENST00000361138	0.00338	0.00332	4	26
C3orf26	ENST00000489081	0.00932	0.00627	38	13
FILIP1L	ENST00000421999	0.00932	0.00627	38	13
SIGLEC1	ENST00000344754	0.00552	0.00664	56	32
CWF19L2	ENST00000282251	0.003	0.00734	0	9
WDR17	ENST00000280190	0.00314	0.00795	1	17

BCS1L	NA	0.00539	0.00869	0	7
RNF25	ENST00000359273	0.00539	0.00869	0	7
DPYD	ENST00000370192	0.0032	0.0123	35	11
PZP	ENST00000261336	0.00325	0.0141	24	53
KIAA1586	ENST00000370733	0.00198	0.0147	11	33
OR5M3	ENST00000312240	0.00653	0.0202	1	10
ZNF425	ENST00000378061	0.00321	0.0238	13	2
CARD9	NA	0.00727	0.0445	9	1
SNAPC4	ENST00000371732	0.00727	0.0445	9	1
CCDC129	ENST00000407970	0.0075	0.0546	2	8
THUMPD3	ENST00000345094	0.0071	0.0667	12	2
LY9	ENST00000368041	0.00409	0.097	0	10
SOAT2	ENST00000301466	0.00616	0.122	2	14

Results from the SIGMA T2D discovery study (n=3,756). Abbreviations: COLLAPSE, Collapsing test; SKAT, Sequence Kernel Association Test; ns, Non-synonymous variants, MINA, sum of total minor allele counts of all variants in the gene in Affecteds; MINU, sum of total minor allele counts of all variants in the gene in Unaffecteds

eTable 8. Gene-set association test results for non-synonymous variants with MAF < 1%

Gene set	NS	FRAC_WITH_RARE	NUM_PASS_VARS	NUM_SING_VARS	P	P (removing p.E508K)
GWAS	3792	0.81487	2169	1341	0.007	0.16
MODY	3792	0.21915	280	165	0.001	0.26

Results from the SIGMA T2D discovery study (n=3,756). Abbreviations: NS, Number of SIGMA Samples on which this analysis was done; FRAC_WITH_RARE, Fraction of individual carrying rare variants below 1%; NUM_PASS_VARS, Number of variants passing filters; NUM_SING_VARS : Number of singletons among variants in NUM_PASS_VARS; P, P-value in Gene set

eTable 9. Gene-set association test results for non-synonymous variants with MAF < 1%

Gene set	NS	FRAC_WITH_RARE	NUM_PASS_VARS	NUM_SING_VARS	PVALUE
GWAS	3792	0.040612	62	43	0.15
MODY	3792	0.011076	11	7	0.86

Results from the SIGMA T2D discovery study (n=3,756). Abbreviations: NS, Number of SIGMA Samples on which this analysis was done; FRAC_WITH_RARE, Fraction of individual carrying rare variants below 1%; NUM_PASS_VARS, Number of variants passing filters; NUM_SING_VARS : Number of singletons among variants in NUM_PASS_VARS; P, P-value in Gene set

Subconsortia Investigators

Broad Genomics Platform: Adal Abebe¹, Justin Abreu¹, Kristin Anderka¹, Scott Anderson¹, Sarah Babchuck¹, Maria Baco¹, Samira Bahl¹, Danielle Bain¹, Kylee Bergin¹, Amy Biasella¹, Bill Biggs¹, Brendan Blumenstiel¹, Harry Bochner¹, Claude Bonnet¹, Wendy Brodeur¹, Joseph BuAbbud¹, Emily C. Davis¹, Jody Camarata¹, Jason Carey¹, Mauricio Carneiro¹, Brynne Cassidy¹, Clinton Chalk¹, Sheridan Channer¹, Andrew Cheney¹, Michelle Cipicchio¹, Kristen Connolly¹, Matthew Coole¹, Maura Costello¹, Miguel Covarrubias¹, Cassandra Crawford¹, Lindsay Croschier¹, Michael Dasilva¹, Matthew Defelice¹, Tim Desmet¹, Alexandra Dimitriou¹, Katerina Dimitriou¹, Michael Dinsmore¹, Danielle Dionne¹, Sheli Dookran¹, Teni Dowdell¹, Phil Dunlea¹, Cassandra Elie¹, M. Erik Husby¹, Emelia Failing¹, Yossi Farjoun¹, Timothy Fennell¹, Damien Fenske-Corbriere¹, Steven Ferreira¹, Sheila Fisher¹, Jennifer Franklin¹, Paul Frere¹, Shemifhar Freytes¹, Dennis Friedrich¹, Stacey Gabriel¹, Diane Gage¹, Christina Gearin¹, Jeff Gentry¹, Lizz Gottardi¹, Alexander Graff¹, George Grant¹, Lisa Green¹, Jonna Grimsby¹, Namrata Gupta¹, Kunsang Gyaltzen¹, Bertrand Haas¹, Susanna Hamilton¹, Maegan Harden¹, Ryan Hegarty¹, Desiree Hernandez¹, Andrew Hollinger¹, Laurie Holmes¹, Tracey Honan¹, Tom Howd¹, Maria Jenkins¹, Ryan Johnson¹, Andrew Johnson¹, Kevin Joseph¹, Fontina Kelley¹, Edward Kelliher¹, Cristyn Kells¹, Amanda Kennedy¹, Sharon Kim¹, Kevinson Kim¹, Samuel Kim¹, Catherine King¹, Charles Kivolowitz¹, Jessica Klopp¹, Anna Koutoulas¹, Massami Laird¹, Katie Larkin¹, Katie Larsson¹, Zach Leber¹, Matthew Lee¹, James Lee¹, Niall Lennon¹, Frances Letendre¹, Tsamla Lhanyitsang¹, Shuqiang Li¹, Kenneth Livak¹, Hayley Lyon¹, Alyssa Macbeth¹, Vasilina Magnisalis¹, Tsheko Makuwa¹, Lauren Margolin¹, Tamara Mason¹, Scott Matthews¹, Michael McCowan¹, Susan McDonough¹, Kaitlyn McGrath¹, James Meldrim¹, Atanas Mihalev¹, Mariela Mihaleva¹, Tyler Miselis¹, Ruchi Munshi¹, Gregory Nakashian¹, Jillian Nolan¹, Nyima Norbu¹, Deborah Norman Farlow¹, Sam Novod¹, Robert Onofrio¹, Veronika Oshero¹, Melissa Parkin¹, Danielle Perrin¹, Caroline Petersen¹, Prapti Pokharel¹, Eliot Polk¹, Samuel Pollock¹, Shannon Power¹, Katelin Pratt¹, Mark Puppo¹, Anthony Rachupka¹, Howard Rafal¹, Ashley Ray¹, Brian Reilly¹, Scott Rich¹, Dana Robbins¹, Joseph Rose¹, Carsten Russ¹, Dennis Ryan¹, Surayya Sana¹, Ahmed Sandakli¹, Michael Saylor¹, Sampath Settupalli¹, Philip Shapiro¹, Kara Slowik¹, Cherylyn Smith¹, Brian Sogoloff¹, Carrie Sougnez¹, Sharon Stavropoulos¹, Gregory Stoneham¹, Jordan Sullivan¹, Katherine Sullivan¹, Danielle Sutherby¹, Frederick Ta¹, Alvin Tam¹, Bradley Taylor¹, Jon Thompson¹, Kathleen Tibbetts¹, Charlotte Tolonen¹, Kristina Tracy¹, Austin Tzou¹, Gina Vicente¹, Fernando Vilorio¹, Andy Vo¹, Louisa Walker¹, John Walsh¹, Cole Walsh¹, Kendra West¹, Emily Wheeler¹, Jane Wilkinson¹, Michael Wilson¹, Ellen Winchester¹, Jennifer Wineski¹, Betty Woolf¹, Chin-Lee Wu¹, Alec Wysoker¹, Qing Yu¹, David Zdeb¹, Andrew Zimmer¹

The T2D-GENES Consortium: Gonçalo Abecasis², Marcio Almeida³, David Altshuler^{4,5,6,7,8,9,10}, Jennifer L. Asimit¹¹, Gil Atzmon¹², Mathew Barber¹³, Nicola L. Beer¹⁴, Graeme I. Bell^{13,15}, Jennifer Below¹⁶, Tom

Blackwell², John Blangero³, Michael Boehnke², Donald W. Bowden^{17,18,19,20}, Noël Burt⁴, John Chambers^{21,22,23}, Han Chen²⁴, Peng Chen²⁵, Peter S.Chines²⁶, Sungkyoung Choi²⁷, Claire Churchhouse⁴, Pablo Cingolani²⁸, Belinda K. Cornes²⁹, Nancy Cox^{13,15}, Aaron G. Day-Williams¹¹, Ravindranath Duggirala³, Josée Dupuis²⁴, Thomas Dyer³, Shuang Feng², Juan Fernandez-Tajes³⁰, Teresa Ferreira³⁰, Tasha E. Fingerlin³¹, Jason Flannick^{4,6}, Jose Florez^{4,6,7}, Pierre Fontanillas⁴, Timothy M. Frayling³², Christian Fuchsberger², Eric R. Gamazon¹⁵, Kyle Gaulton³⁰, Saurabh Ghosh, Anna Gloyn¹⁴, Robert L. Grossman^{15,33}, Jason Grundstad³³, Craig Hanis¹⁶, Allison Heath³³, Heather Highland¹⁶, Momoko Hirokoshi³⁰, Ik-Soo Huh²⁷, Jeroen R. Huyghe², Kamran Ikram^{34,29,35,36}, Kathleen A. Jablonski³⁷, Young Jin Kim³⁸, Goo Jun², Norihiro Kato³⁹, Jayoun Kim²⁷, C. Ryan King⁴⁰, Jaspal Kooner^{22,23,41}, Min-Seok Kwon²⁷, Hae Kyung Im⁴⁰, Markku Laakso⁴², Kevin Koi-Yau Lam²⁵, Jaehoon Lee²⁷, Selyeong Lee²⁷, Sungyoung Lee²⁷, Donna M. Lehman⁴³, Heng Li⁴, Cecilia M. Lindgren³⁰, Xuanyao Liu^{25,44}, Oren E. Livne¹³, Adam E. Locke², Anubha Mahajan³⁰, Julian B. Maller^{30,45}, Alisa K. Manning⁴, Taylor J. Maxwell¹⁶, Alexander Mazouze⁴⁶, Mark I. McCarthy^{30,14,47}, James B. Meigs^{7,48}, Byungju Min²⁷, Karen L. Mohlke⁴⁹, Andrew Morris⁵⁰, Solomon Musani⁵¹, Yoshihiko Nagai⁴⁶, Maggie C.Y. Ng^{17,18}, Dan Nicolae^{13,15,52}, Sohee Oh²⁷, Nicholette Palmer^{17,18,19}, Taesung Park²⁷, Toni I. Pollin⁵³, Inga Prokopenko^{30,54}, David Reich^{4,5}, Manuel A. Rivas³⁰, Laura J. Scott², Mark Seielstad⁵⁵, Yoon Shin Cho⁵⁶, E-Shyong Tai^{34,25,57}, Xueling Sim², Robert Sladek^{46,58}, Philip Smith⁵⁹, Ioanna Tachmazidou¹¹, Tanya M. Teslovich², Jason Torres^{13,15}, Vasily Trubetsky^{13,15}, Sara M. Willems⁶⁰, Amy L. Williams^{4,5}, James G. Wilson⁶¹, Steven Wiltshire⁶², Sungho Won⁶³, Andrew R. Wood³², Wang Xu⁵⁷, Yik Ying Teo^{64,65,66,67,68}, Joon Yoon²⁷, Jong-Young Lee⁶⁹, Matthew Zawistowski², Eleftheria Zeggini¹¹, Weihua Zhang²¹, Sebastian Zöllner^{2,70}

¹The Genomics Platform, The Broad Institute of Harvard and MIT, Cambridge, Massachusetts 02142, USA.

²Department of Biostatistics, Center for Statistical Genetics, University of Michigan, Ann Arbor, Michigan 48109, USA.

³Department of Genetics, Texas Biomedical Research Institute, San Antonio, Texas 78227, USA.

⁴Program in Medical and Population Genetics, Broad Institute of Harvard and MIT, Cambridge, Massachusetts 02142, USA.

⁵Department of Genetics, Harvard Medical School, Boston, Massachusetts 02115, USA.

⁶Center for Human Genetic Research and Diabetes Research Center (Diabetes Unit), Massachusetts General Hospital, Boston 02114, Massachusetts, USA.

⁷Department of Medicine, Harvard Medical School, Boston, Massachusetts 02115, USA.

⁸Center for Human Genetic Research, Massachusetts General Hospital, Boston, Massachusetts 02114, USA.

⁹Department of Molecular Biology, Harvard Medical School, Boston, Massachusetts 02114, USA.

- ¹⁰Department of Biology, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139, USA.
- ¹¹Wellcome Trust Sanger Institute, Hinxton, Cambridge, CB10 1HH, UK.
- ¹²Department of Medicine, Department of Genetics, Albert Einstein College of Medicine, Bronx, New York 10461, USA.
- ¹³Department of Human Genetics, University of Chicago, Chicago, Illinois 60637, USA.
- ¹⁴Oxford Centre for Diabetes, Endocrinology and Metabolism, University of Oxford, Oxford, OX3 7LJ, UK.
- ¹⁵Department of Medicine, University of Chicago, Chicago, Illinois 60637, USA.
- ¹⁶Human Genetics Center, University of Texas Health Science Center at Houston, Houston, Texas 77030, USA.
- ¹⁷Center for Genomics and Personalized Medicine Research, Wake Forest School of Medicine, Winston-Salem, North Carolina 27157, USA.
- ¹⁸Center for Diabetes Research, Wake Forest School of Medicine, Winston-Salem, North Carolina 27157, USA.
- ¹⁹Department of Biochemistry, Wake Forest School of Medicine, Winston-Salem, North Carolina 27157, USA.
- ²⁰Internal Medicine-Endocrinology, Wake Forest School of Medicine, Winston-Salem, North Carolina 27157, USA.
- ²¹Department of Epidemiology and Biostatistics, Imperial College London, London SW7 2AZ, UK.
- ²²Imperial College Healthcare NHS Trust, London W2 1NY, UK.
- ²³Ealing Hospital National Health Service (NHS) Trust, Middlesex UB1 3HW, UK.
- ²⁴Department of Biostatistics, Boston University School of Public Health, Boston, Massachusetts 02115, USA.
- ²⁵Saw Swee Hock School of Public Health, National University of Singapore, Singapore 117597, Singapore.
- ²⁶National Human Genome Research Institute, National Institutes of Health, Bethesda, MD 20892, USA.
- ²⁷Seoul National University, Seoul 110-799, South Korea.
- ²⁸McGill Centre for Bioinformatics, McGill University, Montréal, Quebec, H3G 0B1, Canada.
- ²⁹Singapore Eye Research Institute, Singapore National Eye Centre, Singapore 168751, Singapore.
- ³⁰Wellcome Trust Centre for Human Genetics, University of Oxford, Oxford OX3 7BN, UK.

- ³¹Department of Epidemiology, Colorado School of Public Health, Aurora, Colorado 80045, USA.
- ³²Genetics of Complex Traits, University of Exeter Medical School, Exeter, EX4 4SB UK.
- ³³Institute for Genomics and Systems Biology, University of Chicago, Chicago, Illinois 60637, USA.
- ³⁴Duke National University of Singapore Graduate Medical School, Singapore 169857, Singapore.
- ³⁵Department of Ophthalmology, National University of Singapore and National University Health System, Singapore 119228, Singapore.
- ³⁶Department of Ophthalmology, Erasmus Medical Center, Rotterdam 3000 CA, the Netherlands.
- ³⁷The Biostatistics Center, George Washington University, Rockville, Maryland 20852, USA.
- ³⁸Department of Neurology, Konkuk University School of Medicine, Seoul 143-701, South Korea.
- ³⁹Department of Gene Diagnostics and Therapeutics, Research Institute, National Center for Global Health and Medicine, Tokyo 162-8655, Japan.
- ⁴⁰Department of Health Studies, University of Chicago, Chicago, Illinois 60637, USA.
- ⁴¹National Heart and Lung Institute (NHLI), Imperial College London, Hammersmith Hospital, London W12 0HS, UK.
- ⁴²Department of Medicine, University of Eastern Finland, Kuopio Campus and Kuopio University Hospital, FI-70211 Kuopio, Finland.
- ⁴³Division of Clinical Epidemiology, Department of Medicine, University of Texas Health Science Center at San Antonio, San Antonio, Texas 78229, USA.
- ⁴⁴Graduate School for Integrative Science and Engineering, National University of Singapore, Singapore 117456, Singapore.
- ⁴⁵Department of Statistics, University of Oxford, Oxford, OX1 3TG UK.
- ⁴⁶McGill University, Montréal, Québec H3A 0G4, Canada.
- ⁴⁷Oxford NIHR Biomedical Research Centre, Churchill Hospital, Headington OX3 7LE, UK.
- ⁴⁸General Medicine Division, Massachusetts General Hospital, Boston, Massachusetts 02114, USA.
- ⁴⁹Department of Genetics, University of North Carolina-Chapel Hill, Chapel Hill, North Carolina 27599, USA.
- ⁵⁰Department of Genetic Medicine, Manchester Academic Health Sciences Centre, Manchester M13 9NT, UK.
- ⁵¹Department of Medicine, University of Mississippi Medical Center, Jackson, Mississippi 39126, USA.
- ⁵²Department of Statistics, University of Chicago, Chicago, Illinois 60637, USA.

- ⁵³Department of Medicine, Program for Personalized and Genomic Medicine, University of Maryland School of Medicine, Baltimore, Maryland 21201, USA.
- ⁵⁴Department of Medical Sciences, Molecular Epidemiology and Science for Life Laboratory, Uppsala University, 751 05 Uppsala, Sweden.
- ⁵⁵University of California San Francisco, San Francisco, California 94143, USA.
- ⁵⁶Department of Biomedical Science, Hallym University, Chuncheon, Gangwon-do, 200-702 South Korea.
- ⁵⁷Department of Medicine, Yong Loo Lin School of Medicine, National University of Singapore, Singapore 117597, Singapore.
- ⁵⁸Department of Medicine, Royal Victoria Hospital, Montréal, Québec H3A 1A1, Canada.
- ⁵⁹National Institute of Diabetes and Digestive and Kidney Disease, National Institutes of Health, Bethesda, MD 20817, USA.
- ⁶⁰Department of Genetic Epidemiology, Erasmus Medical Center, Rotterdam 3000 CA, the Netherlands.
- ⁶¹Department of Physiology and Biophysics, University of Mississippi Medical Center, Jackson, Mississippi 39216, USA.
- ⁶²Centre for Medical Research, Western Australian Institute for Medical Research, The University of Western Australia, Nedlands WA 6008, Australia.
- ⁶³Chung-Ang University, Seoul 156-756, South Korea.
- ⁶⁴Department of Epidemiology and Public Health, National University of Singapore, Singapore 117597, Singapore.
- ⁶⁵Centre for Molecular Epidemiology, National University of Singapore, Singapore 117456, Singapore.
- ⁶⁶Genome Institute of Singapore, Agency for Science, Technology and Research, Singapore 138672, Singapore.
- ⁶⁷Graduate School for Integrative Science and Engineering, National University of Singapore, Singapore 117456, Singapore.
- ⁶⁸Department of Statistics and Applied Probability, National University of Singapore, Singapore 117546, Singapore.
- ⁶⁹Center for Genome Science, Korea National Institute of Health, Osong Health Technology Administration Complex, Chungcheongbuk-do, 363-951, South Korea.
- ⁷⁰Department of Psychiatry, University of Michigan, Ann Arbor, Michigan 48109, USA.

REFERENCES

1. Consortium STD, Williams AL, Jacobs SB, et al. Sequence variants in SLC16A11 are a common risk factor for type 2 diabetes in Mexico. *Nature*. Feb 6 2014;506(7486):97-101.
2. Lorenzo C, Williams K, Gonzalez-Villalpando C, Haffner SM. The prevalence of the metabolic syndrome did not increase in Mexico City between 1990-1992 and 1997-1999 despite more central obesity. *Diabetes care*. Oct 2005;28(10):2480-2485.
3. Hunt KJ, Gonzalez ME, Lopez R, Haffner SM, Stern MP, Gonzalez-Villalpando C. Diabetes is more lethal in Mexicans and Mexican-Americans compared to Non-Hispanic whites. *Annals of epidemiology*. Dec 2011;21(12):899-906.
4. Kolonel LN, Henderson BE, Hankin JH, et al. A Multiethnic Cohort in Hawaii and Los Angeles: Baseline Characteristics. *American Journal of Epidemiology*. February 15, 2000 2000;151(4):346-357.
5. Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics*. Jul 15 2009;25(14):1754-1760.
6. DePristo MA, Banks E, Poplin R, et al. A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nature genetics*. May 2011;43(5):491-498.
7. McLaren W, Pritchard B, Rios D, Chen Y, Flicek P, Cunningham F. Deriving the consequences of genomic variants with the Ensembl API and SNP Effect Predictor. *Bioinformatics*. Aug 15 2010;26(16):2069-2070.
8. Scott LJ, Mohlke KL, Bonnycastle LL, et al. A genome-wide association study of type 2 diabetes in Finns detects multiple susceptibility variants. *Science*. Jun 1 2007;316(5829):1341-1345.
9. Stancakova A, Javorsky M, Kuulasmaa T, Haffner SM, Kuusisto J, Laakso M. Changes in insulin sensitivity and insulin release in relation to glycemia and glucose tolerance in 6,414 Finnish men. *Diabetes*. May 2009;58(5):1212-1221.
10. Wichmann HE, Gieger C, Illig T, Group MKS. KORA-gen--resource for population genetics, controls and a broad spectrum of disease phenotypes. *Gesundheitswesen*. Aug 2005;67 Suppl 1:S26-30.
11. Voight BF, Scott LJ, Steinthorsdottir V, et al. Twelve type 2 diabetes susceptibility loci identified through large-scale association analysis. *Nature genetics*. Jul 2010;42(7):579-589.
12. Wellcome Trust Case Control C. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature*. Jun 7 2007;447(7145):661-678.
13. Spector TD, Williams FM. The UK Adult Twin Registry (TwinsUK). *Twin research and human genetics : the official journal of the International Society for Twin Studies*. Dec 2006;9(6):899-906.

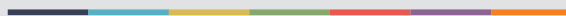
14. Atzmon G, Hao L, Pe'er I, et al. Abraham's children in the genome era: major Jewish diaspora populations comprise distinct genetic clusters with shared Middle Eastern Ancestry. *American journal of human genetics*. Jun 11 2010;86(6):850-859.
15. Berglund G, Elmstahl S, Janzon L, Larsson SA. The Malmo Diet and Cancer Study. Design and feasibility. *Journal of internal medicine*. Jan 1993;233(1):45-51.
16. Berglund G, Nilsson P, Eriksson KF, et al. Long-term outcome of the Malmo preventive project: mortality and cardiovascular morbidity. *Journal of internal medicine*. Jan 2000;247(1):19-29.
17. Bog-Hansen E, Lindblad U, Bengtsson K, Ranstam J, Melander A, Rastam L. Risk factor clustering in patients with hypertension and non-insulin-dependent diabetes mellitus. The Skaraborg Hypertension Project. *Journal of internal medicine*. Mar 1998;243(3):223-232.
18. Groop L, Forsblom C, Lehtovirta M, et al. Metabolic consequences of a family history of NIDDM (the Botnia study): evidence for sex-specific parental effects. *Diabetes*. Nov 1996;45(11):1585-1593.
19. Isomaa B, Forsen B, Lahti K, et al. A family history of diabetes is associated with reduced physical fitness in the Prevalence, Prediction and Prevention of Diabetes (PPP)-Botnia study. *Diabetologia*. Aug 2010;53(8):1709-1713.
20. Lindholm E, Agardh E, Tuomi T, Groop L, Agardh CD. Classifying diabetes according to the new WHO clinical stages. *European journal of epidemiology*. 2001;17(11):983-989.
21. Lyssenko V, Jonsson A, Almgren P, et al. Clinical risk factors, DNA variants, and the development of type 2 diabetes. *The New England journal of medicine*. Nov 20 2008;359(21):2220-2232.
22. Parker A, Meyer J, Lewitzky S, et al. A gene conferring susceptibility to type 2 diabetes in conjunction with obesity is located on chromosome 18p11. *Diabetes*. Mar 2001;50(3):675-680.
23. Yu H, Bowden DW, Spray BJ, Rich SS, Freedman BI. Linkage analysis between loci in the renin-angiotensin axis and end-stage renal disease in African Americans. *Journal of the American Society of Nephrology : JASN*. Dec 1996;7(12):2559-2564.
24. Chahal NS, Lim TK, Jain P, Chambers JC, Kooner JS, Senior R. Does subclinical atherosclerosis burden identify the increased risk of cardiovascular disease mortality among United Kingdom Indian Asians? A population study. *American heart journal*. Sep 2011;162(3):460-466.
25. Chahal NS, Lim TK, Jain P, Chambers JC, Kooner JS, Senior R. Ethnicity-related differences in left ventricular function, structure and geometry: a population study of UK Indian Asian and European white subjects. *Heart*. Mar 2010;96(6):466-471.
26. Lavanya R, Jeganathan VS, Zheng Y, et al. Methodology of the Singapore Indian Chinese Cohort (SICC) eye study: quantifying ethnic variations in the epidemiology of eye diseases in Asians. *Ophthalmic epidemiology*. Nov-Dec 2009;16(6):325-336.

27. Cho YS, Go MJ, Kim YJ, et al. A large-scale genome-wide association study of Asian populations uncovers genetic factors influencing eight quantitative traits. *Nature genetics*. May 2009;41(5):527-534.
28. Tan CE, Emmanuel SC, Tan BY, Jacob E. Prevalence of diabetes and ethnic differences in cardiovascular risk factors. The 1992 Singapore National Health Survey. *Diabetes care*. Feb 1999;22(2):241-247.
29. Hughes K, Yeo PP, Lun KC, et al. Cardiovascular diseases in Chinese, Malays, and Indians in Singapore. II. Differences in risk factor levels. *Journal of epidemiology and community health*. Mar 1990;44(1):29-35.
30. Hughes K, Aw TC, Kuperan P, Choo M. Central obesity, insulin resistance, syndrome X, lipoprotein(a), and cardiovascular risk in Indians, Malays, and Chinese in Singapore. *Journal of epidemiology and community health*. Aug 1997;51(4):394-399.
31. Mitchell BD, Kammerer CM, Blangero J, et al. Genetic and environmental contributions to cardiovascular risk factors in Mexican Americans. The San Antonio Family Heart Study. *Circulation*. Nov 1 1996;94(9):2159-2170.
32. Hunt KJ, Lehman DM, Arya R, et al. Genome-wide linkage analyses of type 2 diabetes in Mexican Americans: the San Antonio Family Diabetes/Gallbladder Study. *Diabetes*. Sep 2005;54(9):2655-2662.
33. Coletta DK, Schneider J, Hu SL, et al. Genome-wide linkage scan for genes influencing plasma triglyceride levels in the Veterans Administration Genetic Epidemiology Study. *Diabetes*. Jan 2009;58(1):279-284.
34. Knowler WC, Coresh J, Elston RC, et al. The Family Investigation of Nephropathy and Diabetes (FIND): design and methods. *Journal of diabetes and its complications*. Jan-Feb 2005;19(1):1-9.
35. Hanis CL, Ferrell RE, Barton SA, et al. Diabetes among Mexican Americans in Starr County, Texas. *American journal of epidemiology*. Nov 1983;118(5):659-672.
36. Kang HM, Sul JH, Service SK, et al. Variance component model to account for sample structure in genome-wide association studies. *Nature genetics*. Apr 2010;42(4):348-354.
37. Purcell S, Neale B, Todd-Brown K, et al. PLINK: a tool set for whole-genome association and population-based linkage analyses. *American journal of human genetics*. Sep 2007;81(3):559-575.
38. Price AL, Patterson NJ, Plenge RM, Weinblatt ME, Shadick NA, Reich D. Principal components analysis corrects for stratification in genome-wide association studies. *Nature genetics*. Aug 2006;38(8):904-909.
39. Willer CJ, Li Y, Abecasis GR. METAL: fast and efficient meta-analysis of genomewide association scans. *Bioinformatics*. Sep 1 2010;26(17):2190-2191.
40. Baran Y, Pasaniuc B, Sankararaman S, et al. Fast and accurate inference of local ancestry in Latino populations. *Bioinformatics*. May 15 2012;28(10):1359-1367.
41. Li JZ, Absher DM, Tang H, et al. Worldwide human relationships inferred from genome-wide patterns of variation. *Science*. Feb 22 2008;319(5866):1100-1104.

42. Reich D, Patterson N, Campbell D, et al. Reconstructing Native American population history. *Nature*. Aug 16 2012;488(7411):370-374.
43. Behar DM, Yunusbayev B, Metspalu M, et al. The genome-wide structure of the Jewish people. *Nature*. Jul 8 2010;466(7303):238-242.
44. International HapMap C, Altshuler DM, Gibbs RA, et al. Integrating common and rare genetic variation in diverse human populations. *Nature*. Sep 2 2010;467(7311):52-58.
45. Delaneau O, Marchini J, Zagury JF. A linear complexity phasing method for thousands of genomes. *Nature methods*. Feb 2012;9(2):179-181.
46. Holmkvist J, Almgren P, Lyssenko V, et al. Common variants in maturity-onset diabetes of the young genes and future risk of type 2 diabetes. *Diabetes*. Jun 2008;57(6):1738-1744.
47. Lee S, Wu MC, Lin X. Optimal tests for rare variant effects in sequencing association studies. *Biostatistics*. Sep 2012;13(4):762-775.
48. Li B, Leal SM. Methods for detecting associations with rare variants for common diseases: application to analysis of sequence data. *American journal of human genetics*. Sep 2008;83(3):311-321.
49. Morris AP, Voight BF, Teslovich TM, et al. Large-scale association analysis provides insights into the genetic architecture and pathophysiology of type 2 diabetes. *Nature genetics*. Sep 2012;44(9):981-990.
50. Bjorkhaug L, Sagen JV, Thorsby P, Sovik O, Molven A, Njolstad PR. Hepatocyte nuclear factor-1 alpha gene mutations and diabetes in Norway. *The Journal of clinical endocrinology and metabolism*. Feb 2003;88(2):920-931.
51. Galan M, Garcia-Herrero CM, Azriel S, et al. Differential effects of HNF-1alpha mutations associated with familial young-onset diabetes on target gene regulation. *Mol Med*. Mar-Apr 2011;17(3-4):256-265.
52. Aukrust I, Bjorkhaug L, Negahdar M, et al. SUMOylation of pancreatic glucokinase regulates its cellular stability and activity. *J Biol Chem*. Feb 22 2013;288(8):5951-5962.
53. Bjorkhaug L, Ye H, Horikawa Y, Sovik O, Molven A, Njolstad PR. MODY associated with two novel hepatocyte nuclear factor-1alpha loss-of-function mutations (P112L and Q466X). *Biochem Biophys Res Commun*. Dec 29 2000;279(3):792-798.



Graphic design: Communication Division, UIB / Print: Skjipes Kommunikasjon AS



uib.no

ISBN: 978-82-308-3654-5