UNIVERSITY OF CALIFORNIA SAN DIEGO

Leveraging human genetics and functional genomics to investigate insulin resistance related disorders

A Dissertation submitted in partial satisfaction of the requirements for the degree Doctor of Philosophy

in

Biomedical Sciences with a Specialization in Bioinformatics

by

Natalie Mae DeForest

Committee in charge:

Professor Amit Majithia, Chair
Professor Hannah Carter
Professor Melissa Gymrek
Professor Graham McVicker

2023

The Dissertation of Natalie Mae DeForest is approved, and it is acceptable in quality and form for publication on microfilm and electronically.

University of California San Diego

2023

DEDICATION

To family and friends, steadfast and true,
Whose love and support saw me through.

To Rosie, my sweet feline dear,
Whose comforting presence was always near.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGEMENTS

I would like to thank my Committee for the helpful discussions and critical feedback on the dissertation. I would also like to thank Bruce Hamilton for the scientific insights and lessons learned during my time in the Genetics Training Program, as well as funding from the Ruth L. Kirschstein Institutional National Research Service Award T32 GM008666 from the National Institute of General Medical Sciences.

Chapter 1, in full, is an edited reprint of the material as it appears in Current Diabetes Reports 2022. DeForest, Natalie; Majithia, Amit R., Springer Nature, 2022. The dissertation author was the primary investigator and author of this paper.

Chapter 2, in full, is an edited reprint of the material as it appears in Cell Genomics 2023. DeForest, Natalie; Kavitha, Babu; Hu, Siqi; Isaac, Roi; Krohn, Lynne; Wang, Minxian; Du, Xiaomi; Saldanha, Camila De Arruda; Gylys, Jenny; Merli, Edoardo; Abagyan, Ruben; Najmi, Laeya; Mohan, Viswanathan; Alnylam Human Genetics; AMP-T2D Consortium; Flannick, Jason; Peloso, Gina M.; Gordts, Philip L.S.M.; Heinz, Sven; Deaton, Aimee M.; Khera, Amit V.; Olefsky, Jerrold; Radha, Venkatesan; Majithia, Amit R., Cell Press, 2023. The dissertation author was the primary investigator and author of this paper.

Chapter 4, in full, is currently being prepared for submission for publication of the material. DeForest, Natalie; Zhu, Zhiyi; Dron, Jacqueline S.; Natarajan, Pradeep; Flannick, Jason; Amariuta, Tiffany; Peloso, Gina M.; Majithia, Amit R. The dissertation author was the primary investigator and author of this paper.

VITA

2018 Bachelor of Science in Pharmaceutical Chemistry with a Minor in Quantitative Biology and Bioinformatics, University of California Davis

2023 Doctor of Philosophy in Biomedical Sciences with a Specialization in Bioinformatics, University of California San Diego

PUBLICATIONS

DeForest N, Kavitha B, Hu S, Isaac R, Krohn L, Wang M, Du X, Saldanha CDA, Gylys J, Merli E, Abagyan R, Najimi L, Mohan V, Alnylam Human Genetics, AMP-T2D Consortium, Flannick J, Peloso GM, Heinz S, Gordts P, Khera AV, Olefsky J, Radha V, Majithia AR. Human gain-of-function variants in HNF1A confer protection from diabetes but independently increase hepatic secretion of atherogenic lipoproteins. Cell Genomics. May 30, 2023; 100339. doi: 10.1016/j.xgen.2023.100339.

Trieger GW, Pessentheiner AR, Purcell SC, Green CR, DeForest N, Willert K, Majithia AR, Metallo CM, Godula K, Gordts PLSM. Glycocalyx engineering with heparan sulfate mimetics attenuates Wnt activity during adipogenesis to promote glucose uptake and metabolism. Journal of Biological Chemistry. 2023 Mar 15;104611. doi: 10.1016/j.jbc.2023.104611. PMID: 36931394.

Isaac R, Vinik Y, Mikl M, Nadav-Eliyahu S, Shatz-Azoulay H, Yaakobi A, DeForest N, Majithia AR, Webster NJG, Shav-Tal Y, Elhanany E, Zick Y. A seven-transmembrane protein-TM7SF3, resides in nuclear speckles and regulates alternative splicing. iScience. 2022 Oct 4;25(11):105270. doi: 10.1016/j.isci.2022.105270. PMID: 36304109; PMCID: PMC9593240.

DeForest N, Majithia AR. Genetics of Type 2 Diabetes: Implications from Large-Scale Studies. Curr Diab Rep. 2022 May;22(5):227-235. doi: 10.1007/s11892-022-01462-3. Epub 2022 Mar 19. PMID: 35305202; PMCID: PMC9072491.

Du X, DeForest N, Majithia AR. Human Genetics to Identify Therapeutic Targets for NAFLD: Challenges and Opportunities. Front Endocrinol. 2021 Dec 7;12:777075. doi: 10.3389/fendo.2021.777075. PMID: 34950105; PMCID: PMC8688763.

Liang X, Park Y, DeForest N, Hao J, Zhao X, Niu C, Wang K, Smith B, Lai Y. In Vitro Hepatic Uptake in Human and Monkey Hepatocytes in the Presence and Absence of Serum Protein and Its In Vitro to In Vivo Extrapolation. Drug Metab Dispos. 2020 Dec;48(12):1283-1292. doi: 10.1124/dmd.120.000163. Epub 2020 Oct 9. PMID: 33037043.

ABSTRACT OF THE DISSERTATION

Leveraging human genetics and functional genomics to investigate insulin resistance related disorders

by

Natalie Mae DeForest

Doctor of Philosophy in Biomedical Sciences with a Specialization in Bioinformatics

University of California San Diego, 2023

Professor Amit Majithia, Chair

Insulin resistance plays a central role in the development of multiple epidemic metabolic disorders such as type 2 diabetes (T2D) and cardiovascular disease. In this dissertation, I leverage human genetics and functional genomics to unravel the molecular mechanisms underlying insulin resistance and its related disorders, as well as identify novel key effector genes for potential therapeutic applications. The work is divided into four chapters:

Chapter 1 provides an overview of the recent large-scale T2D genetic studies with particular focus on the effects of sample size and ancestral diversity on genetic discovery as well as discusses recent work on the use and limitations of genetic risk scores for T2D risk prediction.

Chapter 2 describes deep mutational scanning of 11,970 protein-coding HNF1A variants in human hepatocytes and clinical correlation with 553,246 exome-sequenced individuals which identifies gain-of-function mutations in HNF1A which exert pleiotropic effects, conferring protection against diabetes while increasing the hepatic secretion of atherogenic lipoproteins.

Chapter 3 utilizes genetic determinants of circulating protein levels to perform proteome-wide Mendelian Randomization which identifies RSPO3 as a critical mediator of non-alcoholic steatohepatitis (NASH) pathogenesis.

Chapter 4 presents a genome-wide association study for the serum triglyceride to HDL-cholesterol ratio and subsequent integrative genomics which systematically prioritizes novel insulin resistance effector genes including PLA2G12A, PLA2G6, VGLL3, and TNFAIP8 with insights into tissues of action and directional effects on metabolic disease.

Collectively, in this dissertation I demonstrate the power of integrating large-scale human genetics with functional genomics in elucidating the intricate mechanisms contributing to insulin resistance-related disorders. These findings offer novel insights into the pathogenesis of insulin resistance and its related disorders, and further prioritize potential candidate genes for further investigation and therapeutic targeting.

INTRODUCTION

Insulin resistance represents one of the largest modern threats to human health, affecting approximately one in three Americans(*1*) and causing multiple epidemic diseases including type 2 diabetes (T2D), cardiovascular disease (CVD), and non-alcoholic fatty liver disease (NAFLD)(*2*). Insulin resistance is a physiological condition in which the body's cells, particularly those in the metabolic liver, muscle, and adipose tissues, become less responsive to the effects of the hormone insulin(*3*). Insulin resistance disrupts the normal balance of glucose metabolism, resulting in persistently higher blood glucose levels and leading to various metabolic abnormalities. When an insulin-resistant individual is no longer able to secrete sufficient insulin to compensate for the elevated glucose levels, T2D develops(*4*).

Moreover, insulin resistance is closely linked to central obesity, hypertension, and dyslipidemia(*5*). These cardiometabolic risk factors cluster together to define the metabolic syndrome, and together contribute to a higher likelihood of atherosclerosis and related cardiovascular events(*5*). Additionally, insulin resistance and metabolic syndrome play a pivotal role in the pathogenesis of NAFLD, a condition marked by the accumulation of fat in the liver, which through fibrosis and inflammation can progress to more severe liver diseases such as non-alcoholic steatohepatitis (NASH) or cirrhosis(*6*). Insulin resistance develops gradually over time and is assessed through various tests including the gold standard hyperinsulinemic euglycemic clamp or fasting blood glucose and insulin levels, however it often remains undetected until related disease and subsequent complications arise. Metabolic disease outcomes collectively represent a significant public health burden as management strategies such as lifestyle modifications and

pharmacological or surgical interventions are available but few definitive cures exist. As the global rates of metabolic disease are rising,(*7*), there is an urgent need for continued research and therapeutic innovations to treat insulin resistance and prevent resulting disease progression.

Though lifestyle factors such as physical inactivity and poor diet contribute to metabolic disease, an individual's risk of developing insulin resistance and its subsequent disorders is also determined by genetic factors(*8*). Investigations into the genetic basis of insulin resistance related disorders through familial studies have identified single gene defects causing severe insulin resistance and related metabolic disorders, however such studies have provided limited insights as families with clinical presentations of insulin resistance are rare and difficult to identify(*9*–*11*). Genome-wide association studies (GWAS) have implicated genes with metabolic dysfunction(*12*–*15*), however insights into the molecular mechanisms underlying genetic associations (i.e. the directionality of identified associations, tissues of action) remain unknown without functional experimentation(*16*). Furthermore, the majority of GWAS signals involve common variants within noncoding regions which can lead to challenges identifying causal genes(*17*).

Human genetic investigations present a valuable tool to investigate disease mechanisms and potential therapeutic targets. With the widespread application of genome sequencing and more recently large-scale proteomics(*18*), large biobank-based genetic studies have the power to evaluate a large number of putative genes for causal effects on disease phenotypes and outcomes within a single study. These studies additionally provide findings with direct human relevance, thus circumventing the common

issue involving translatability of *in vitro* or animal model findings to human patients(*19*). Furthermore, it has been shown that drugs with direct human genetic evidence are approximately twice as likely to succeed through clinical trials and successfully approved for treatment(*20*). In this dissertation, I present the following four chapters in which I employ human genetic and integrative genomic approaches to prioritize novel insulin resistance effector genes and investigate the tissue-specific molecular mechanisms of insulin resistance and its related disorders.

In Chapter 1, I demonstrate how population genetics studies have played a crucial role in identifying genetic variants associated with T2D, an epidemic insulin resistance related disorder, via systematic review of recent large-scale GWAS for T2D. In this chapter I detail how GWAS and large-scale sequencing projects have revealed numerous T2D genetic risk loci which have provided valuable insights into the genetic architecture and pathways underlying T2D, and present how genetic understanding of T2D may hold limited promise for T2D risk prediction.

In Chapter 2, I implement a genotype-function-phenotype approach to investigate the clinical consequences of the full functional spectrum of HNF1A on insulin resistance outcomes. In this chapter I utilize a method of high-throughput functional characterization termed “deep mutational scanning” which assesses the full functional spectrum, including both loss and gain-of-function, of all possible amino acid changes at each position along a protein in a single assay(*21*). This method harnesses protein-coding variants which, unlike non-coding variants, can directly implicate genes and enables rare variants to be grouped by function and then associated with phenotype, i.e. genotype-function-phenotype. In this chapter, I describe a deep mutational scan of 11,970 protein-coding

HNF1A variants in human hepatocytes and the subsequent clinical correlations with 553,246 exome-sequenced individuals to characterize the molecular mechanisms of HNF1A on insulin resistance outcomes T2D and CVD.

In Chapter 3, I employ plasma proteome-wide Mendelian randomization to identify proteins likely causal for NASH using genetic determinants of circulating protein levels (protein quantitative trait loci, pQTLs). This Mendelian randomization framework is able to determine the causality and directionality of association of each protein with phenotype which genome-wide association studies are unable to assess. From this scan, I prioritize RSPO3 to be a top causal effector protein for NASH and further interrogate its role in NASH pathogenesis using complementary Mendelian randomization analyses.

In Chapter 4, I conduct a GWAS for the triglycerides to HDL-cholesterol ratio, a surrogate marker of insulin resistance, in the UK Biobank cohort and nominate likely causal insulin resistance genes through statistical fine-mapping and functional genomic annotations. To characterize the role of the novel prioritized insulin resistance genes, I integrate several orthogonal sources of gene-level evidence including genetic variation in protein coding sequences, common variation associated with changes in gene expression, and transcriptomic analysis of participant tissue samples from insulin sensitizing clinical trials. In summary, the work detailed in this chapter aims to identify new modulators of insulin resistance with potential to serve as biomarkers or pharmacological targets.

In summary, my dissertation delineates a genotype-function-phenotype framework which integrates high-throughput functional characterization and large-scale population cohorts in order to elucidate the molecular mechanisms of nominated insulin resistance

effector genes. Additionally, I conduct plasma proteome-wide Mendelian randomization and genome-wide discovery in combination with integrative genomic characterization to prioritize novel insulin resistance genes which can be investigated further using the established genotype-function-phenotype characterization approach. Taken together, this work integrates large-scale population cohorts, functional genomics, and statistical genetic techniques to establish reproducible genetic discovery workflows to prioritize effector genes which may be promising insulin resistance biomarkers or therapeutic targets.

References

1. A. S. Go, D. Mozaffarian, V. L. Roger, E. J. Benjamin, J. D. Berry, W. B. Borden, D. M. Bravata, S. Dai, E. S. Ford, C. S. Fox, S. Franco, H. J. Fullerton, C. Gillespie, S. M. Hailpern, J. A. Heit, V. J. Howard, M. D. Huffman, B. M. Kissela, S. J. Kittner, D. T. Lackland, J. H. Lichtman, L. D. Lisabeth, D. Magid, G. M. Marcus, A. Marelli, D. B. Matchar, D. K. McGuire, E. R. Mohler, C. S. Moy, M. E. Mussolino, G. Nichol, N. P. Paynter, P. J. Schreiner, P. D. Sorlie, J. Stein, T. N. Turan, S. S. Virani, N. D. Wong, D. Woo, M. B. Turner, American Heart Association Statistics Committee and Stroke Statistics Subcommittee, Executive summary: heart disease and stroke statistics--2013 update: a report from the American Heart Association. *Circulation*. **127**, 143–152 (2013).

2. G. Wilcox, Insulin and insulin resistance. *Clin. Biochem. Rev.* **26**, 19–39 (2005).

3. A. R. Saltiel, C. R. Kahn, Insulin signalling and the regulation of glucose and lipid metabolism. *Nature*. **414**, 799–806 (2001).

4. American Diabetes Association, 2. Classification and Diagnosis of Diabetes: Standards of Medical Care in Diabetes—2021. *Diabetes Care*. **44**, S15–S33 (2021).

5. C. K. Roberts, A. L. Hevener, R. J. Barnard, Metabolic syndrome and insulin resistance: underlying causes and modification by exercise training. *Compr. Physiol.* **3**, 1–58 (2013).

6. H. Kitade, G. Chen, Y. Ni, T. Ota, Nonalcoholic Fatty Liver Disease and Insulin Resistance: New Insights and Potential New Treatments. *Nutrients*. **9** (2017), doi:10.3390/nu9040387.

7. M. G. Saklayen, The Global Epidemic of the Metabolic Syndrome. *Curr. Hypertens. Rep.* **20**, 12 (2018).

8. A. E. Brown, M. Walker, Genetics of Insulin Resistance and the Metabolic Syndrome. *Curr. Cardiol. Rep.* **18**, 75 (2016).

9. R. M. Watanabe, T. Valle, E. R. Hauser, S. Ghosh, J. Eriksson, K. Kohtamäki, C. Ehnholm, J. Tuomilehto, F. S. Collins, R. N. Bergman, M. Boehnke, Familiality of quantitative metabolic traits in Finnish families with non-insulin-dependent diabetes mellitus. Finland-United States Investigation of NIDDM Genetics (FUSION) Study investigators. *Hum. Hered.* **49**, 159–168 (1999).

10. M. Lehtovirta, J. Kaprio, C. Forsblom, J. Eriksson, J. Tuomilehto, L. Groop, Insulin sensitivity and insulin secretion in monozygotic and dizygotic twins. *Diabetologia*. **43**, 285–293 (2000).

11. D. Carmelli, L. R. Cardon, R. Fabsitz, Clustering of hypertension, diabetes, and

obesity in adult male twins: same genes or same environments? *Am. J. Hum. Genet.* **55**, 566–573 (1994).

12. K. L. Mohlke, M. Boehnke, Recent advances in understanding the genetic architecture of type 2 diabetes. *Hum. Mol. Genet.* **24**, R85–92 (2015).

13. J. Dupuis, C. Langenberg, I. Prokopenko, R. Saxena, N. Soranzo, A. U. Jackson, E. Wheeler, N. L. Glazer, N. Bouatia-Naji, A. L. Gloyn, C. M. Lindgren, R. Mägi, A. P. Morris, J. Randall, T. Johnson, P. Elliott, D. Rybin, G. Thorleifsson, V. Steinthorsdottir, P. Henneman, H. Grallert, A. Dehghan, J. J. Hottenga, C. S. Franklin, P. Navarro, K. Song, A. Goel, J. R. B. Perry, J. M. Egan, T. Lajunen, N. Grarup, T. Sparsø, A. Doney, B. F. Voight, H. M. Stringham, M. Li, S. Kanoni, P. Shrader, C. Cavalcanti-Proença, M. Kumari, L. Qi, N. J. Timpson, C. Gieger, C. Zabena, G. Rocheleau, E. Ingelsson, P. An, J. O'Connell, J. 'an Luan, A. Elliott, S. A. McCarroll, F. Payne, R. M. Roccasecca, F. Pattou, P. Sethupathy, K. Ardlie, Y. Ariyurek, B. Balkau, P. Barter, J. P. Beilby, Y. Ben-Shlomo, R. Benediktsson, A. J. Bennett, S. Bergmann, M. Bochud, E. Boerwinkle, A. Bonnefond, L. L. Bonnycastle, K. Borch-Johnsen, Y. Böttcher, E. Brunner, S. J. Bumpstead, G. Charpentier, Y.-D. I. Chen, P. Chines, R. Clarke, L. J. M. Coin, M. N. Cooper, M. Cornelis, G. Crawford, L. Crisponi, I. N. M. Day, E. J. C. de Geus, J. Delplanque, C. Dina, M. R. Erdos, A. C. Fedson, A. Fischer-Rosinsky, N. G. Forouhi, C. S. Fox, R. Frants, M. G. Franzosi, P. Galan, M. O. Goodarzi, J. Graessler, C. J. Groves, S. Grundy, R. Gwilliam, U. Gyllensten, S. Hadjadj, G. Hallmans, N. Hammond, X. Han, A.-L. Hartikainen, N. Hassanali, C. Hayward, S. C. Heath, S. Hercberg, C. Herder, A. A. Hicks, D. R. Hillman, A. D. Hingorani, A. Hofman, J. Hui, J. Hung, B. Isomaa, P. R. V. Johnson, T. Jørgensen, A. Jula, M. Kaakinen, J. Kaprio, Y. A. Kesaniemi, M. Kivimaki, B. Knight, S. Koskinen, P. Kovacs, K. O. Kyvik, G. M. Lathrop, D. A. Lawlor, O. Le Bacquer, C. Lecoeur, Y. Li, V. Lyssenko, R. Mahley, M. Mangino, A. K. Manning, M. T. Martínez-Larrad, J. B. McAteer, L. J. McCulloch, R. McPherson, C. Meisinger, D. Melzer, D. Meyre, B. D. Mitchell, M. A. Morken, S. Mukherjee, S. Naitza, N. Narisu, M. J. Neville, B. A. Oostra, M. Orrù, R. Pakyz, C. N. A. Palmer, G. Paolisso, C. Pattaro, D. Pearson, J. F. Peden, N. L. Pedersen, M. Perola, A. F. H. Pfeiffer, I. Pichler, O. Polasek, D. Posthuma, S. C. Potter, A. Pouta, M. A. Province, B. M. Psaty, W. Rathmann, N. W. Rayner, K. Rice, S. Ripatti, F. Rivadeneira, M. Roden, O. Rolandsson, A. Sandbaek, M. Sandhu, S. Sanna, A. A. Sayer, P. Scheet, L. J. Scott, U. Seedorf, S. J. Sharp, B. Shields, G. Sigurethsson, E. J. G. Sijbrands, A. Silveira, L. Simpson, A. Singleton, N. L. Smith, U. Sovio, A. Swift, H. Syddall, A.-C. Syvänen, T. Tanaka, B. Thorand, J. Tichet, A. Tönjes, T. Tuomi, A. G. Uitterlinden, K. W. van Dijk, M. van Hoek, D. Varma, S. Visvikis-Siest, V. Vitart, N. Vogelzangs, G. Waeber, P. J. Wagner, A. Walley, G. B. Walters, K. L. Ward, H. Watkins, M. N. Weedon, S. H. Wild, G. Willemsen, J. C. M. Witteman, J. W. G. Yarnell, E. Zeggini, D. Zelenika, B. Zethelius, G. Zhai, J. H. Zhao, M. C. Zillikens, DIAGRAM Consortium, GIANT Consortium, Global BPgen Consortium, I. B. Borecki, R. J. F. Loos, P. Meneton, P. K. E. Magnusson, D. M. Nathan, G. H. Williams, A. T. Hattersley, K. Silander, V. Salomaa, G. D. Smith, S. R. Bornstein, P. Schwarz, J. Spranger, F. Karpe, A. R. Shuldiner, C. Cooper, G. V. Dedoussis, M. Serrano-Ríos, A. D. Morris, L. Lind, L. J. Palmer, F. B. Hu, P. W. Franks, S.

Ebrahim, M. Marmot, W. H. L. Kao, J. S. Pankow, M. J. Sampson, J. Kuusisto, M. Laakso, T. Hansen, O. Pedersen, P. P. Pramstaller, H. E. Wichmann, T. Illig, I. Rudan, A. F. Wright, M. Stumvoll, H. Campbell, J. F. Wilson, Anders Hamsten on behalf of Procardis Consortium, MAGIC investigators, R. N. Bergman, T. A. Buchanan, F. S. Collins, K. L. Mohlke, J. Tuomilehto, T. T. Valle, D. Altshuler, J. I. Rotter, D. S. Siscovick, B. W. J. H. Penninx, D. I. Boomsma, P. Deloukas, T. D. Spector, T. M. Frayling, L. Ferrucci, A. Kong, U. Thorsteinsdottir, K. Stefansson, C. M. van Duijn, Y. S. Aulchenko, A. Cao, A. Scuteri, D. Schlessinger, M. Uda, A. Ruokonen, M.-R. Jarvelin, D. M. Waterworth, P. Vollenweider, L. Peltonen, V. Mooser, G. R. Abecasis, N. J. Wareham, R. Sladek, P. Froguel, R. M. Watanabe, J. B. Meigs, L. Groop, M. Boehnke, M. I. McCarthy, J. C. Florez, I. Barroso, New genetic loci implicated in fasting glucose homeostasis and their impact on type 2 diabetes risk. *Nat. Genet.* **42**, 105–116 (2010).

14. A. K. Manning, M.-F. Hivert, R. A. Scott, J. L. Grimsby, N. Bouatia-Naji, H. Chen, D. Rybin, C.-T. Liu, L. F. Bielak, I. Prokopenko, N. Amin, D. Barnes, G. Cadby, J.-J. Hottenga, E. Ingelsson, A. U. Jackson, T. Johnson, S. Kanoni, C. Ladenvall, V. Lagou, J. Lahti, C. Lecoeur, Y. Liu, M. T. Martinez-Larrad, M. E. Montasser, P. Navarro, J. R. B. Perry, L. J. Rasmussen-Torvik, P. Salo, N. Sattar, D. Shungin, R. J. Strawbridge, T. Tanaka, C. M. van Duijn, P. An, M. de Andrade, J. S. Andrews, T. Aspelund, M. Atalay, Y. Aulchenko, B. Balkau, S. Bandinelli, J. S. Beckmann, J. P. Beilby, C. Bellis, R. N. Bergman, J. Blangero, M. Boban, M. Boehnke, E. Boerwinkle, L. L. Bonnycastle, D. I. Boomsma, I. B. Borecki, Y. Böttcher, C. Bouchard, E. Brunner, D. Budimir, H. Campbell, O. Carlson, P. S. Chines, R. Clarke, F. S. Collins, A. Corbatón-Anchuelo, D. Couper, U. de Faire, G. V. Dedoussis, P. Deloukas, M. Dimitriou, J. M. Egan, G. Eiriksdottir, M. R. Erdos, J. G. Eriksson, E. Eury, L. Ferrucci, I. Ford, N. G. Forouhi, C. S. Fox, M. G. Franzosi, P. W. Franks, T. M. Frayling, P. Froguel, P. Galan, E. de Geus, B. Gigante, N. L. Glazer, A. Goel, L. Groop, V. Gudnason, G. Hallmans, A. Hamsten, O. Hansson, T. B. Harris, C. Hayward, S. Heath, S. Hercberg, A. A. Hicks, A. Hingorani, A. Hofman, J. Hui, J. Hung, M.-R. Jarvelin, M. A. Jhun, P. C. D. Johnson, J. W. Jukema, A. Jula, W. H. Kao, J. Kaprio, S. L. R. Kardia, S. Keinanen-Kiukaanniemi, M. Kivimaki, I. Kolcic, P. Kovacs, M. Kumari, J. Kuusisto, K. O. Kyvik, M. Laakso, T. Lakka, L. Lannfelt, G. M. Lathrop, L. J. Launer, K. Leander, G. Li, L. Lind, J. Lindstrom, S. Lobbens, R. J. F. Loos, J. 'an Luan, V. Lyssenko, R. Mägi, P. K. E. Magnusson, M. Marmot, P. Meneton, K. L. Mohlke, V. Mooser, M. A. Morken, I. Miljkovic, N. Narisu, J. O'Connell, K. K. Ong, B. A. Oostra, L. J. Palmer, A. Palotie, J. S. Pankow, J. F. Peden, N. L. Pedersen, M. Pehlic, L. Peltonen, B. Penninx, M. Pericic, M. Perola, L. Perusse, P. A. Peyser, O. Polasek, P. P. Pramstaller, M. A. Province, K. Räikkönen, R. Rauramaa, E. Rehnberg, K. Rice, J. I. Rotter, I. Rudan, A. Ruokonen, T. Saaristo, M. Sabater-Lleal, V. Salomaa, D. B. Savage, R. Saxena, P. Schwarz, U. Seedorf, B. Sennblad, M. Serrano-Rios, A. R. Shuldiner, E. J. G. Sijbrands, D. S. Siscovick, J. H. Smit, K. S. Small, N. L. Smith, A. V. Smith, A. Stančáková, K. Stirrups, M. Stumvoll, Y. V. Sun, A. J. Swift, A. Tönjes, J. Tuomilehto, S. Trompet, A. G. Uitterlinden, M. Uusitupa, M. Vikström, V. Vitart, M.-C. Vohl, B. F. Voight, P. Vollenweider, G. Waeber, D. M. Waterworth, H. Watkins, E. Wheeler, E. Widen, S. H. Wild, S. M. Willems, G. Willemsen, J. F. Wilson, J. C.

M. Witteman, A. F. Wright, H. Yaghootkar, D. Zelenika, T. Zemunik, L. Zgaga, DIAbetes Genetics Replication And Meta-analysis (DIAGRAM) Consortium, Multiple Tissue Human Expression Resource (MUTHER) Consortium, N. J. Wareham, M. I. McCarthy, I. Barroso, R. M. Watanabe, J. C. Florez, J. Dupuis, J. B. Meigs, C. Langenberg, A genome-wide approach accounting for body mass index identifies genetic variants influencing fasting glycemic traits and insulin resistance. *Nat. Genet.* **44**, 659–669 (2012).

15. R. A. Scott, V. Lagou, R. P. Welch, E. Wheeler, M. E. Montasser, J. 'an Luan, R. Mägi, R. J. Strawbridge, E. Rehnberg, S. Gustafsson, S. Kanoni, L. J. Rasmussen-Torvik, L. Yengo, C. Lecoeur, D. Shungin, S. Sanna, C. Sidore, P. C. D. Johnson, J. W. Jukema, T. Johnson, A. Mahajan, N. Verweij, G. Thorleifsson, J.-J. Hottenga, S. Shah, A. V. Smith, B. Sennblad, C. Gieger, P. Salo, M. Perola, N. J. Timpson, D. M. Evans, B. S. Pourcain, Y. Wu, J. S. Andrews, J. Hui, L. F. Bielak, W. Zhao, M. Horikoshi, P. Navarro, A. Isaacs, J. R. O'Connell, K. Stirrups, V. Vitart, C. Hayward, T. Esko, E. Mihailov, R. M. Fraser, T. Fall, B. F. Voight, S. Raychaudhuri, H. Chen, C. M. Lindgren, A. P. Morris, N. W. Rayner, N. Robertson, D. Rybin, C.-T. Liu, J. S. Beckmann, S. M. Willems, P. S. Chines, A. U. Jackson, H. M. Kang, H. M. Stringham, K. Song, T. Tanaka, J. F. Peden, A. Goel, A. A. Hicks, P. An, M. Müller-Nurasyid, A. Franco-Cereceda, L. Folkersen, L. Marullo, H. Jansen, A. J. Oldehinkel, M. Bruinenberg, J. S. Pankow, K. E. North, N. G. Forouhi, R. J. F. Loos, S. Edkins, T. V. Varga, G. Hallmans, H. Oksa, M. Antonella, R. Nagaraja, S. Trompet, I. Ford, S. J. L. Bakker, A. Kong, M. Kumari, B. Gigante, C. Herder, P. B. Munroe, M. Caulfield, J. Antti, M. Mangino, K. Small, I. Miljkovic, Y. Liu, M. Atalay, W. Kiess, A. L. James, F. Rivadeneira, A. G. Uitterlinden, C. N. A. Palmer, A. S. F. Doney, G. Willemsen, J. H. Smit, S. Campbell, O. Polasek, L. L. Bonnycastle, S. Hercberg, M. Dimitriou, J. L. Bolton, G. R. Fowkes, P. Kovacs, J. Lindström, T. Zemunik, S. Bandinelli, S. H. Wild, H. V. Basart, W. Rathmann, H. Grallert, DIAbetes Genetics Replication and Meta-analysis (DIAGRAM) Consortium, W. Maerz, M. E. Kleber, B. O. Boehm, A. Peters, P. P. Pramstaller, M. A. Province, I. B. Borecki, N. D. Hastie, I. Rudan, H. Campbell, H. Watkins, M. Farrall, M. Stumvoll, L. Ferrucci, D. M. Waterworth, R. N. Bergman, F. S. Collins, J. Tuomilehto, R. M. Watanabe, E. J. C. de Geus, B. W. Penninx, A. Hofman, B. A. Oostra, B. M. Psaty, P. Vollenweider, J. F. Wilson, A. F. Wright, G. K. Hovingh, A. Metspalu, M. Uusitupa, P. K. E. Magnusson, K. O. Kyvik, J. Kaprio, J. F. Price, G. V. Dedoussis, P. Deloukas, P. Meneton, L. Lind, M. Boehnke, A. R. Shuldiner, C. M. van Duijn, A. D. Morris, A. Toenjes, P. A. Peyser, J. P. Beilby, A. Körner, J. Kuusisto, M. Laakso, S. R. Bornstein, P. E. H. Schwarz, T. A. Lakka, R. Rauramaa, L. S. Adair, G. D. Smith, T. D. Spector, T. Illig, U. de Faire, A. Hamsten, V. Gudnason, M. Kivimaki, A. Hingorani, S. M. Keinanen-Kiukaanniemi, T. E. Saaristo, D. I. Boomsma, K. Stefansson, P. van der Harst, J. Dupuis, N. L. Pedersen, N. Sattar, T. B. Harris, F. Cucca, S. Ripatti, V. Salomaa, K. L. Mohlke, B. Balkau, P. Froguel, A. Pouta, M.-R. Jarvelin, N. J. Wareham, N. Bouatia-Naji, M. I. McCarthy, P. W. Franks, J. B. Meigs, T. M. Teslovich, J. C. Florez, C. Langenberg, E. Ingelsson, I. Prokopenko, I. Barroso, Large-scale association analyses identify new loci influencing glycemic traits and provide insight into the underlying biological pathways. *Nat. Genet.* **44**, 991–1005 (2012).

16. L. M. Starita, N. Ahituv, M. J. Dunham, J. O. Kitzman, F. P. Roth, G. Seelig, J. Shendure, D. M. Fowler, Variant Interpretation: Functional Assays to the Rescue. *The American Journal of Human Genetics*. **101** (2017), pp. 315–325.

17. F. Zhang, J. R. Lupski, Non-coding genetic variants in human disease. *Hum. Mol. Genet.* **24**, R102–R110 (2015).

18. K. Suhre, M. I. McCarthy, J. M. Schwenk, Genetics meets proteomics: perspectives for large population-based studies. *Nat. Rev. Genet.* **22**, 19–37 (2021).

19. I. W. Mak, N. Evaniew, M. Ghert, Lost in translation: animal models and clinical trials in cancer treatment. *Am. J. Transl. Res.* **6**, 114–118 (2014).

20. M. R. Nelson, H. Tipney, J. L. Painter, J. Shen, P. Nicoletti, Y. Shen, A. Floratos, P. C. Sham, M. J. Li, J. Wang, L. R. Cardon, J. C. Whittaker, P. Sanseau, The support of human genetic evidence for approved drug indications. *Nat. Genet.* **47**, 856–860 (2015).

21. D. M. Fowler, S. Fields, Deep mutational scanning: a new style of protein science. *Nat. Methods*. **11**, 801–807 (2014).

CHAPTER 1

Genetics of Type 2 Diabetes: Implications from Large-Scale Studies

Introduction

Type 2 Diabetes (T2D) is a multifactorial, heritable syndrome characterized by dysregulated glucose homeostasis that results in impaired insulin secretion and insulin resistance. Recent large-scale, multi-ancestry genetic studies of T2D have identified over 500 novel risk loci. In this chapter, I provide an overview of the recent T2D genetic studies from the past three years with particular focus on the effects of sample size and ancestral diversity on genetic discovery as well as discuss recent work on the use and limitations of genetic risk scores (GRS) for T2D risk prediction.

Type 2 Diabetes is a heterogeneous but heritable syndrome

Type 2 Diabetes (T2D) is characterized by impaired glucose metabolism arising from defects in insulin resistance and secretion(*1*). In clinical practice, T2D is diagnosed by elevated blood glucose levels most commonly assessed via point measurements in the fasting state or averaged over months via glycated hemoglobin (HbA1c) tests. Clinical presentation and disease progression may vary considerably among individuals, and the prevalence of T2D varies between different ethnic groups; for example, Hispanic and Black populations have higher age-adjusted T2D prevalence compared to White and Asian groups(*2*, *3*). Clinical complications of T2D include microvascular complications such as retinopathy, neuropathy, and nephropathy as well as macrovascular complications such as myocardial infarction and stroke(*4*). Cardiovascular disease (CVD)

is the leading cause of death in people with T2D who have up to a three-fold increase in CVD risk as compared to people without T2D(*5*).

The pathogenesis of T2D involves both environmental and genetic causes. Environmental factors including obesity, stress, and lifestyle choices such as an unhealthy, energy-dense diet and a sedentary lifestyle have been closely associated with the development of T2D(*6*). The heritability of T2D ranges from 30-70%(*7*) and family history of T2D is a significant risk factor, with an approximate two-fold relative risk for siblings(*8*) and a three-fold increased risk for first-degree relatives of a T2D individual(*9*). A handful of robust disease genes were identified by early small-scale genetic association studies for T2D(*10*, *11*) and related Mendelian diabetes syndromes(*12*). With the advent of genotyping arrays and the systematic cataloging of common genetic variation by the International HapMap project, population-scale genome-wide association studies (GWAS) became feasible, leading to the identification of hundreds of T2D associated loci(*13*). This review focuses on the T2D genetic association studies conducted over the past 3 years.

T2D risk loci marked by common genetic variants are mostly shared across ancestries

In the early 2000s, collaborative efforts spanning multiple institutions across the globe coalesced into several international consortia focused on genetic mapping of T2D along with its related traits and even complications (summarized in Table 1.1). Initially, consortia such as DIAGRAM(*14*) and MAGIC(*15*) aggregated participants of a single ancestry (mostly northern European), but more recently they have included participants

across a variety of ancestries(*16*, *17*). As of 2018, the list of T2D associations included over 200 independent loci(*18*) (Table 1.2). Subsequent studies over the past three years have built upon earlier work by meta-analyzing previously collected samples with samples obtained across multiple ancestries to identify an additional 500 T2D risk loci, defined as those >500 kb and r^2 linkage disequilibrium (LD) < 0.05 from previously reported loci (Table 1.2).

Table 1.1: Overview of T2D-specific and disease agnostic large-scale consortia

Abbreviation	Name	Phenotype(s)	Sample Size	Ancestry
DIAGRAM(*14*)	DIAbetes Genetics Replication And Meta-analysis	T2D	26,676 T2D cases and 132,532 controls	European
DIAMANTE(*19*)	DIAbetes Meta-ANalysis of Trans-Ethnic association studies	T2D	European: 74,124 T2D cases and 824,006 controls	African–American, East Asian, European, Hispanic, and South Asian
MAGIC(*17*)	Meta-Analyses of Glucose and Insulin-related traits Consortium	Glycemic traits (fasting glucose, fasting insulin, 2hour glucose, HbA1c)	281, 416 non-diabetic individuals	East Asian, Hispanic, African American, South Asian, and sub-Saharan African
SIGMA(*20*)	Slim Initiative in Genomic Medicine for the Americas	T2D, cancer, kidney disease	8,227 T2D cases and 12,966 controls	Hispanic
AGEN(*21*)	Asian Genetic Epidemiology Network	T2D, cardiovascular disease	77,418 T2D cases and 356,122 controls	East Asian
T2D-GENES(*22*)	Type 2 Diabetes Genetic Exploration by Next-generation sequencing in multi-Ethnic Samples	T2D	20,791 T2D cases and 24,440 controls	Hispanic/Latino, European, African-American, East-Asian, and South-Asian
MEDIA(*22*)	African Americans from the MEta-analysis of type 2 DIabetes in African Americans	T2D	8,284 T2D cases and 15,543 controls	African American
MVP(*23*)	Million Veteran Program	Disease agnostic	825,000	Predominantly European, also African American, Hispanic, and Asian
UKB(*24*)	UK Biobank	Disease agnostic	Genotype: 500,000; Whole exome: 450,000; Whole genome: 150,000	Predominantly European, also African American, Hispanic, Asian, African, and Carribbean
BBJ(*25*)	BioBank Japan	Disease agnostic	Genotype: 200,000; Whole genome: 2,000	Japanese
PAGE(*26*)	Population Architecture Genomics and Epidemiology	Disease agnostic	50,000	European, African American, Asian, Native Hawaiian, and Hispanic

Table 1.2: Summary of recent large-scale T2D genetic association studies

BBJ: Biobank Japan; ToMMo: Tohoku Medical Megabank Organization ; IMM: Iwate Tohoku Medical Megabank Organization; JPHC: Japan Public Health Center–based Prospective; J-MICC: Japan Multi-Institutional Collaborative Cohort; AGEN: Asian Genetic Epidemiology Network; DIAMANTE: DIAbetes Meta-ANalysis of Trans-Ethnic association studies; CKB: China Kadoorie Biobank; KBA: Korea Biobank Array; SIGMA: Slim Initiative in Genomic Medicine for the Americas; MEDIA: African Americans from the MEta-analysis of type 2 DIabetes in African Americans

Citation	Cohort(s)	Meta-analysis sample size	Ancestry	Phenotype(s)	Total risk loci	Novel risk loci
Mahajan et al. *Nat. Gen* 2018	DIAGRAM	74,124 T2D cases and 824,006 controls	European	T2D	243	135
Suzuki et al. *Nat. Gen* 2019	BBJ, Osaka-Midousuji Rotary Club, Pharma SNP, ToMMo, IIMM, JPHC, and J-MICC	36,614 T2D cases and 155,150 controls	Japanese	T2D	88	28
Spracklen et al. *Nature* 2020	AGEN, DIAMANTE, CKB, KBA, and BBJ	77,418 T2D cases and 356,122 controls	East Asian	T2D	183	61
Vujkovic et al. *Nat. Gen* 2020	MVP, DIAMANTE, Penn Medicine Biobank, Pakistan Genomic Resource, BBJ, Malmö Diet and Cancer Study, Medstar, and PennCath	228,499 T2D cases and 1,178,783 controls	European, African American, Hispanic, South Asian, and East Asian	T2D	568	318
Polfus et al. *HGG Advances* 2021	PAGE, DIAGRAM; Replication: 23andMe, DIAMANTE, SIGMA, AGEN, MEDIA	53,102 T2D cases and 193,679 controls	European, African American, Hispanic, Asian, and Native Hawaiian	T2D	39	4; 2 of which replicated
Chen et al. *Nat. Gen* 2021	MAGIC	281,416 non-T2D individuals	European, African American, sub-Saharan African, Hispanic, South Asian, and East Asian	Fasting glucose, fasting insulin, 2-hour glucose, HbA1c	242	72 (99 at time of writing)

The largest T2D genetic association study to date meta-analyzed GWAS from eight cohorts including population-based biobanks such as the Million Veteran Program (MVP) and Biobank Japan as well as dedicated T2D case-control cohorts such as DIAMANTE(*27*). These cohorts contained individuals from five different ancestral groups (European, African American, Hispanic, South Asian, and East Asian) for a total of 228,499 T2D cases and 1,178,783 controls. A total of 568 T2D risk loci were identified at genome-wide significance, 293 of which were novel in this study(*27*). These newly identified loci had smaller effect sizes (average beta regression coefficient of 0.032 ± 0.012 per allele) than previously discovered T2D risk loci (average beta of 0.054 ± 0.045 per allele), demonstrating that increased sample size enhanced statistical power to detect association signals with smaller biological effects. Additionally, within the MVP cohort, Vujkovic et al. performed ancestry-specific GWAS which identified an additional 21 loci in Europeans and 4 loci in African Americans not initially identified in the original meta-analysis. A few loci demonstrated higher effect sizes for T2D in African Americans compared with Europeans, but the majority of loci (92.1%) showed no significant heterogeneity in effect estimates between Europeans and African Americans.

The most recently published multi-ancestry T2D case-control genetic study illustrates the dominant effect of sample size in driving locus discovery(*28*). Polfus et al. conducted a GWAS meta-analysis of 53,102 T2D cases and 193,679 controls from the multi-ethnic Population Architecture Genomics and Epidemiology (PAGE) consortium along with the DIAGRAM consortium, and replicated their findings in independent ancestry-specific samples from multiple T2D consortia including DIAMANTE, Asian Genetic Epidemiology Network (AGEN), Slim Initiative in Genomic Medicine for the

Americas (SIGMA), and African Americans from the MEta-analysis of type 2 DIabetes in African Americans (MEDIA)(*28*). They identified four novel loci from the discovery PAGE + DIAGRAM GWAS, two of which replicated in single ancestry replication GWAS: 1) rs11466334 near the transforming growth factor beta-1 (*TGFB1*) gene and 2) rs13052926 near beta-secretase 2 (*BACE2*). Only the *TGFB1* locus (rs11466334) was an ancestry-specific variant occurring more commonly in African (Minor allele frequency (MAF)=6.8%) and Hispanic populations (MAF=1.3%) as compared with other ancestries (MAF<1%). The single nucleotide polymorphism (SNP) was also predicted to be functionally consequential via disrupting a CCCTC-binding factor (CTCF) binding motif potentially leading to altered enhancer - promoter interactions. Although this study identified four novel loci, it did not re-identify over 90% of the genome-wide significant loci identified in previous studies(*18*, *27*) (Table 1.2). The critical distinguishing factor was the sample size, highlighting this as the major determinant of genetic discovery in common variant association studies for T2D.

To examine the effect of ancestry on loci associated with glycemic traits (fasting glucose, fasting insulin, 2-hour glucose, and HbA1c) in non-diabetic individuals, Chen et al. and the MAGIC investigators first conducted meta-analyses of GWAS within a single-ancestry population: European, African American, Hispanic, East Asian, or South Asian populations. They then meta-analyzed these "single-ancestry GWAS" in a "trans-ancestry" GWAS consisting of a total of 281,416 non-diabetic individuals(*17*). From the trans-ancestry GWAS, they identified 235 loci associated with at least one glycemic trait, and 7 additional loci from the single-ancestry GWAS that did not rise to genome-wide significance in the trans-ancestry analysis. Interestingly, the single-ancestry loci had

similar allele frequencies across the sampled ancestries, potentially suggesting epistatic effects with other ancestry-specific variants or that they rose to significance in a particular single-ancestry analysis simply by chance.

Of the 235 trans-ancestry glycemic trait-associated loci, 93 were novel at the time of publication and Chen et. al. performed an instructive simulation to quantify the benefit of including multiple ancestries as opposed to simply increasing sample size to enhance novel locus discovery. By re-scaling the standard errors of the European single-ancestry GWAS to simulate the trans-ancestry sample size, Chen et al. found that that 21 out of the 93 (22.6%) newly discovered trans-ancestry loci would not have been identified in a GWAS restricted to European ancestry. This suggests that while the majority of novel loci were identified due to increase in sample size, a modest benefit was obtained by including non-European samples.

Furthermore, this study examined the effect of single- versus trans-ancestry analyses on the resolution of genetic fine-mapping to identify causal variants. To do this, the authors identified 98 locus-trait associations that had a single causal variant from both single- and trans-ancestry fine-mapping and found that 72 (73%) locus-trait associations showed improvements in the resolution of fine-mapping, as quantified by a decreased number of variants in the 99% credible sets. Of these 72 locus-trait associations, 53% were improved due to larger sample size in the trans-ancestry analysis and 47% were improved due to the inclusion of diverse ancestries as demonstrated by a decrease in the median number of variants in the 99% credible sets from 24 to 15 variants (37.5% median reduction). Thus, for about half the loci identified, inclusion of diverse ancestries enabled a reduction of about 10 variants from the final 99% credible sets for the causal variant.

In addition to the above described multi-ancestry studies, recent large-scale T2D genetic studies have also been performed in East Asian populations which have been previously under-represented in GWAS. In a T2D case-control GWAS meta-analysis including Biobank Japan participants, Suzuki et al. examined 36,614 T2D cases and 155,150 controls of Japanese ancestry and identified 88 T2D risk loci, 28 of which were novel(*29*). The majority (77%) of the identified lead variants are common (MAF>0.05) in both Japanese and European populations, and Suzuki et al. demonstrated that effect sizes are strongly correlated (Pearson's $r = 0.83$, $P = 8.7e\text{-}51$) and directly consistent (94%) between the Japanese GWAS and a comparable T2D European GWAS, indicating the majority of genetic susceptibility between Japanese and European ancestry is shared. In addition to this study in Japanese individuals, the largest meta-analysis of T2D GWAS in individuals of East Asian ancestry to date examined 77,418 T2D cases and 356,122 controls across 23 studies including AGEN and Biobank Japan to identify 183 loci, of which 61 were novel(*21*). Upon comparison with a previously published T2D GWAS in European individuals of similar sample size, Spracklen et al. demonstrated that effect sizes of variants significantly associated with T2D in both East Asian and European ancestry were strongly correlated (r=0.87). Furthermore, the authors find that only 8.4% of variants showed significant heterogeneity in effect size between the East Asian and European GWAS results, and the variants which have the greatest differences in effect sizes between the two populations are those that are common or low-frequency in East Asians but rare in Europeans (MAF<0.1%). Overall, these recent T2D genetic association studies in East Asian ancestry cohorts underscore the finding that genetic susceptibility to T2D captured by common genetic variation is mostly shared across ancestries.

Genetic risk scores for T2D do not substantially enhance risk prediction over traditional clinical risk factors

While over 700 loci identified by common variant association studies (i.e. GWAS) combine to explain almost 20% of T2D heritability(*27*), each individual common variant (i.e. SNP) has a small to modest effect (10-30%) on disease risk as compared to simply knowing family history of T2D, which if present in a parent confers a large increase in risk (~two-three fold)(*30*). Combining multiple variants genotyped in a single person into a genetic risk score (GRS, also commonly referred to as polygenic risk score) is a logical strategy to enhance the clinical utility of genetic information from common variants to identify individuals at high risk(*31*). GRS combining multiple loci were initially tested in the early 2000s with the first T2D GWAS studies. One of the first studies calculated a T2D GRS from a combination of 18 loci finding that genetic information minimally enhanced risk prediction when combined with traditional clinical risk factors such as age, sex, or family history of diabetes(*32*). In the past few years there has been a resurgence of interest in GRS leveraging many more loci identified from large-scale, multi-ancestry cohorts.

In the largest T2D GWAS to date (Table 1.2), Vujkovic et al. used results from a previous European GWAS(*18*) to calculate GRS for participants in the MVP and demonstrated that individuals with the highest T2D GRS (90–100% GRS percentile) presented the highest risk for T2D (OR = 5.21, 95% CI=4.94–5.49) compared to those with the lowest T2D GRS (0-10% GRS percentile)(*27*). Using the GWAS effect estimates from the T2D GWAS conducted by Vujkovic et al., Polfus et al. computed a GRS for T2D

in a multi-ethnic cohort(*28*). From this they found that GRS constructed from multi-ethnic computed weights demonstrated nominal increases in predictive power compared to single-ancestry computed weights, and observed strongly significant heterogeneity across ancestries for accuracy of T2D risk prediction. For instance, the multi-ethnic GRS without adjustment for clinical risk factors performed best in European and East Asian populations (AUC=0.66 and 0.63 respectively) and most poorly in African Americans (AUC=0.57).

These recent studies which have generated GRS for T2D and its related phenotypes have demonstrated that GRS has the highest discriminative ability when applied to European populations and that performance is subsequently improved in non-European ancestries when GRS is computed using multi-ancestry weights(*28*, *33*). However, even after a decade of methodological refinement as well as an increase in the number of loci to calculate GRS, the predictive power of GRS for T2D is comparable to discrimination by clinical risk factors alone (Figure 1.1). However, GRS may have a role in detecting individuals at high risk before clinical risk factors become apparent. Whether information from GRS can motivate preventative therapy to meaningfully reduce rates of future incident T2D remains to be studied. GRS have also been applied widely beyond T2D to other heritable diseases such as heart disease and cancer(*34*, *35*) and even been offered as a tool for embryo screening during in vitro fertilization(*36*–*38*). But the current consensus among geneticists, ethicists, and clinicians is that the scientific and technical uncertainty in GRS and their limited predictive power should limit their use in genetic screening(*39*).

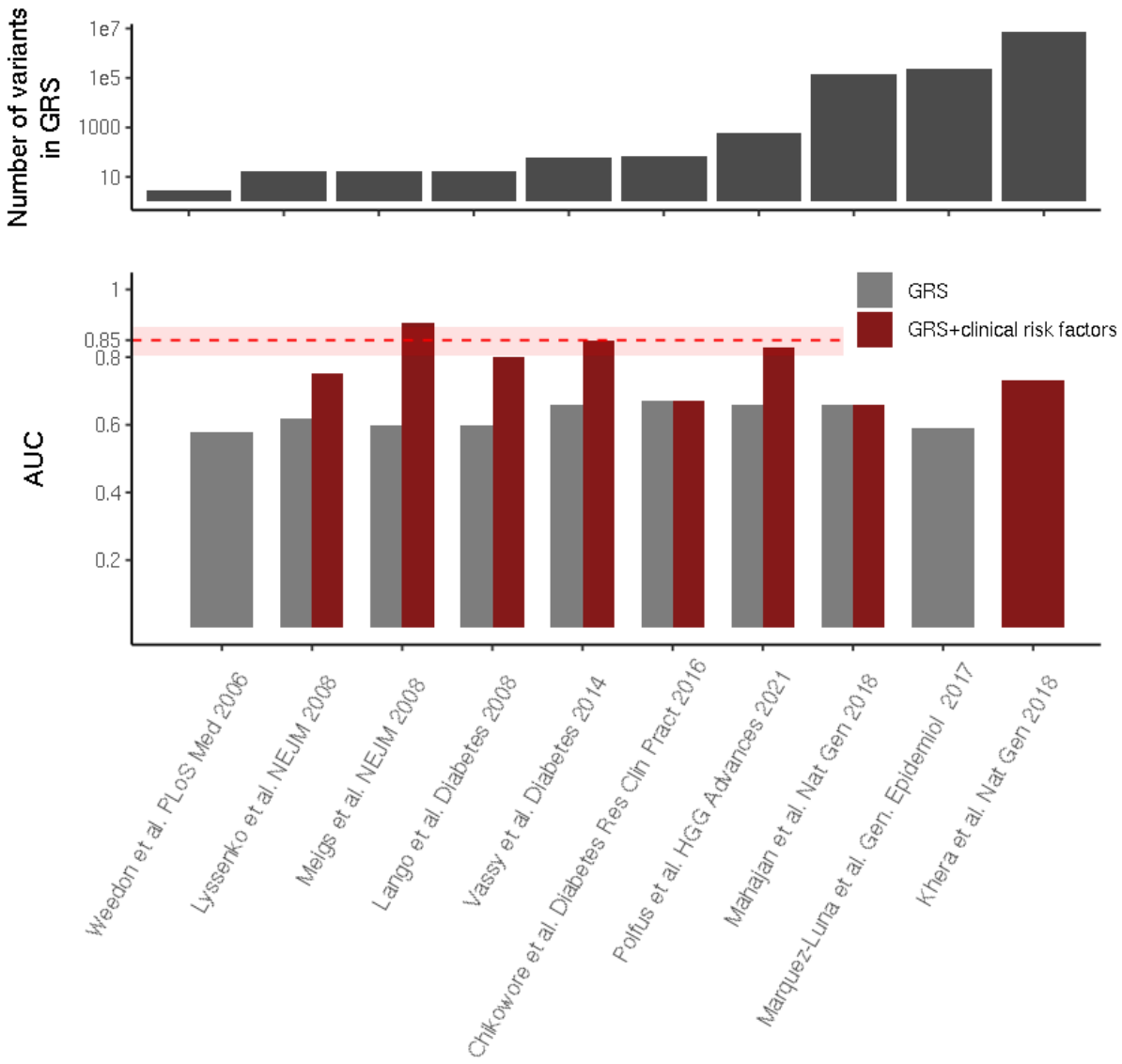


Figure 1.1. Genetic risk scores for T2D do not substantially enhance risk prediction over traditional clinical risk factors.

Outcomes from independent, large-scale studies which have constructed genetic risk scores (GRS) for T2D(*18*, *28*, *32–34*, *40–44*). Studies are shown along the x-axis, ordered by the number of variants used to construct the genetic risk scores (top panel). (bottom panel) The accuracy of the GRS alone and the GRS with clinical T2D risk factors to predict T2D from each study, quantified by the area under the receiver operating characteristic curve (AUC). The dashed red line and shaded red box represent the current predictive power and 95% confidence interval respectively of T2D clinical risk factors (age, sex, parental T2D, BMI, systolic blood pressure, fasting glucose, HDL cholesterol, and triglycerides) to predict T2D(*43*).

Perspective on future genetic mapping studies in T2D

With over 700 T2D risk loci identified by common variant genetic association studies (i.e. GWAS), decades of follow-up biological studies in cellular and organismal model systems will be required to fully understand the causal genes and molecular mechanisms of disease pathogenesis. Thus, it is unlikely that simply aggregating larger T2D case:control cohorts for association analysis will provide scientific and clinical insight into T2D. Here, we expect that an enhanced focus on T2D complications, which are the leading cause of death in T2D(*5*) and are independently heritable of diabetes(*45*), using common variant association methodology will advance understanding and treatment as has been ongoing for T1D(*46*).

In contrast to common variants which were generated millions of years ago in an genetically equilibrated ancestral human population, rare genetic variants (MAF<<0.01) which arose during the “out of Africa” human population expansion(*47*) potentially offer different mechanisms of disease causation. As exome and whole genome sequencing are becoming more commonplace, investigators have begun to examine rare variant associations with T2D(*48*). The challenge with rare variant association studies is that the sample size requirement vastly increases due to the low allele frequency and increase in multiple hypothesis testing burden from the large number of rare variants(*49*). Using a combination of methodological enhancements such as “burden tests” which aggregate rare variants across a gene to reduce the multiple hypothesis testing burden and population-scale biobanks like the UK Biobank to increase sample size, investigators have identified novel T2D loci such as GIGYF1(*50*) and FAM234A(*51*) which were not marked by common variant signals.

In summary, we expect that large scale exome and whole-genome sequencing of population scale biobanks will facilitate rare-variant association studies of T2D to identify novel loci beyond what has been identified by common variant association studies thus far. Additionally, focusing genetic mapping efforts on micro and macrovascular diabetes complications is likely to maximize the value of novel locus discovery to further understand and treat T2D.

Funding

This work was supported by grants from the National Institute of Diabetes and Digestive and Kidney Diseases (1R01DK123422-01 to A.R.M), a UCSD/UCLA Pilot and Feasibility grant (P30 DK063491 to A.R.M), and a Ruth L. Kirschstein Institutional National Research Service Award T32 GM008666 from the National Institute of General Medical Sciences (to N.D.).

Author Contributions

N.D. and A.R.M. conceived the manuscript and figures. N.D. drafted the manuscript. All authors were involved in critical manuscript revision.

Chapter 1, in full, is an edited reprint of the material as it appears in Current Diabetes Reports 2022. DeForest, Natalie; Majithia, Amit R., Springer Nature, 2022. The dissertation author was the primary investigator and author of this paper.

References

1. R. A. DeFronzo, E. Ferrannini, L. Groop, R. R. Henry, W. H. Herman, J. J. Holst, F. B. Hu, C. R. Kahn, I. Raz, G. I. Shulman, D. C. Simonson, M. A. Testa, R. Weiss, Type 2 diabetes mellitus. *Nat Rev Dis Primers*. **1**, 15019 (2015).

2. American Diabetes Association, 2. Classification and Diagnosis of Diabetes: Standards of Medical Care in Diabetes—2021. *Diabetes Care*. **44**, S15–S33 (2021).

3. E. K. Spanakis, S. H. Golden, Race/ethnic difference in diabetes and diabetic complications. *Curr. Diab. Rep.* **13**, 814–823 (2013).

4. N. J. Morrish, S. L. Wang, L. K. Stevens, J. H. Fuller, H. Keen, Mortality and causes of death in the WHO Multinational Study of Vascular Disease in Diabetes. *Diabetologia*. **44 Suppl 2**, S14–21 (2001).

5. C. S. Fox, S. Coady, P. D. Sorlie, R. B. D'Agostino Sr, M. J. Pencina, R. S. Vasan, J. B. Meigs, D. Levy, P. J. Savage, Increasing cardiovascular disease burden due to diabetes mellitus: the Framingham Heart Study. *Circulation*. **115**, 1544–1550 (2007).

6. S. Chatterjee, K. Khunti, M. J. Davies, Type 2 diabetes. *Lancet*. **389**, 2239–2251 (2017).

7. P. Almgren, M. Lehtovirta, B. Isomaa, L. Sarelin, M. R. Taskinen, V. Lyssenko, T. Tuomi, L. Groop, Botnia Study Group, Heritability and familiality of type 2 diabetes and related quantitative traits in the Botnia Study. *Diabetologia*. **54**, 2811–2819 (2011).

8. K. Hemminki, X. Li, K. Sundquist, J. Sundquist, Familial risks for type 2 diabetes in Sweden. *Diabetes Care*. **33**, 293–297 (2010).

9. V. Lyssenko, P. Almgren, D. Anevski, R. Perfekt, K. Lahti, M. Nissén, B. Isomaa, B. Forsen, N. Homström, C. Saloranta, M.-R. Taskinen, L. Groop, T. Tuomi, Botnia study group, Predictors of and longitudinal changes in insulin sensitivity and secretion preceding onset of type 2 diabetes. *Diabetes*. **54**, 166–174 (2005).

10. D. Altshuler, J. N. Hirschhorn, M. Klannemark, C. M. Lindgren, M.-C. Vohl, J. Nemesh, C. R. Lane, S. F. Schaffner, S. Bolk, C. Brewer, T. Tuomi, D. Gaudet, T. J. Hudson, M. Daly, L. Groop, E. S. Lander, The common PPARγ Pro12Ala polymorphism is associated with decreased risk of type 2 diabetes. *Nature Genetics*. **26** (2000), pp. 76–80.

11. S. F. A. Grant, G. Thorleifsson, I. Reynisdottir, R. Benediktsson, A. Manolescu, J. Sainz, A. Helgason, H. Stefansson, V. Emilsson, A. Helgadottir, U. Styrkarsdottir, K. P. Magnusson, G. B. Walters, E. Palsdottir, T. Jonsdottir, T. Gudmundsdottir, A.

Gylfason, J. Saemundsdottir, R. L. Wilensky, M. P. Reilly, D. J. Rader, Y. Bagger, C. Christiansen, V. Gudnason, G. Sigurdsson, U. Thorsteinsdottir, J. R. Gulcher, A. Kong, K. Stefansson, Variant of transcription factor 7-like 2 (TCF7L2) gene confers risk of type 2 diabetes. *Nat. Genet.* **38**, 320–323 (2006).

12. A. Bonnefond, P. Froguel, Rare and common genetic events in type 2 diabetes: what should biologists know? *Cell Metab.* **21**, 357–368 (2015).

13. J. Flannick, J. C. Florez, Type 2 diabetes: genetic data sharing to advance complex disease research. *Nat. Rev. Genet.* **17**, 535–549 (2016).

14. R. A. Scott, L. J. Scott, R. Mägi, L. Marullo, K. J. Gaulton, M. Kaakinen, N. Pervjakova, T. H. Pers, A. D. Johnson, J. D. Eicher, A. U. Jackson, T. Ferreira, Y. Lee, C. Ma, V. Steinthorsdottir, G. Thorleifsson, L. Qi, N. R. Van Zuydam, A. Mahajan, H. Chen, P. Almgren, B. F. Voight, H. Grallert, M. Müller-Nurasyid, J. S. Ried, N. W. Rayner, N. Robertson, L. C. Karssen, E. M. van Leeuwen, S. M. Willems, C. Fuchsberger, P. Kwan, T. M. Teslovich, P. Chanda, M. Li, Y. Lu, C. Dina, D. Thuillier, L. Yengo, L. Jiang, T. Sparso, H. A. Kestler, H. Chheda, L. Eisele, S. Gustafsson, M. Frånberg, R. J. Strawbridge, R. Benediktsson, A. B. Hreidarsson, A. Kong, G. Sigurðsson, N. D. Kerrison, J. 'an Luan, L. Liang, T. Meitinger, M. Roden, B. Thorand, T. Esko, E. Mihailov, C. Fox, C.-T. Liu, D. Rybin, B. Isomaa, V. Lyssenko, T. Tuomi, D. J. Couper, J. S. Pankow, N. Grarup, C. T. Have, M. E. Jørgensen, T. Jørgensen, A. Linneberg, M. C. Cornelis, R. M. van Dam, D. J. Hunter, P. Kraft, Q. Sun, S. Edkins, K. R. Owen, J. R. B. Perry, A. R. Wood, E. Zeggini, J. Tajes-Fernandes, G. R. Abecasis, L. L. Bonnycastle, P. S. Chines, H. M. Stringham, H. A. Koistinen, L. Kinnunen, B. Sennblad, T. W. Mühleisen, M. M. Nöthen, S. Pechlivanis, D. Baldassarre, K. Gertow, S. E. Humphries, E. Tremoli, N. Klopp, J. Meyer, G. Steinbach, R. Wennauer, J. G. Eriksson, S. Männistö, L. Peltonen, E. Tikkanen, G. Charpentier, E. Eury, S. Lobbens, B. Gigante, K. Leander, O. McLeod, E. P. Bottinger, O. Gottesman, D. Ruderfer, M. Blüher, P. Kovacs, A. Tonjes, N. M. Maruthur, C. Scapoli, R. Erbel, K.-H. Jöckel, S. Moebus, U. de Faire, A. Hamsten, M. Stumvoll, P. Deloukas, P. J. Donnelly, T. M. Frayling, A. T. Hattersley, S. Ripatti, V. Salomaa, N. L. Pedersen, B. O. Boehm, R. N. Bergman, F. S. Collins, K. L. Mohlke, J. Tuomilehto, T. Hansen, O. Pedersen, I. Barroso, L. Lannfelt, E. Ingelsson, L. Lind, C. M. Lindgren, S. Cauchi, P. Froguel, R. J. F. Loos, B. Balkau, H. Boeing, P. W. Franks, A. Barricarte Gurrea, D. Palli, Y. T. van der Schouw, D. Altshuler, L. C. Groop, C. Langenberg, N. J. Wareham, E. Sijbrands, C. M. van Duijn, J. C. Florez, J. B. Meigs, E. Boerwinkle, C. Gieger, K. Strauch, A. Metspalu, A. D. Morris, C. N. A. Palmer, F. B. Hu, U. Thorsteinsdottir, K. Stefansson, J. Dupuis, A. P. Morris, M. Boehnke, M. I. McCarthy, I. Prokopenko, DIAbetes Genetics Replication And Meta-analysis (DIAGRAM) Consortium, An Expanded Genome-Wide Association Study of Type 2 Diabetes in Europeans. *Diabetes*. **66**, 2888–2902 (2017).

15. J. Dupuis, C. Langenberg, I. Prokopenko, R. Saxena, N. Soranzo, A. U. Jackson, E. Wheeler, N. L. Glazer, N. Bouatia-Naji, A. L. Gloyn, C. M. Lindgren, R. Mägi, A. P. Morris, J. Randall, T. Johnson, P. Elliott, D. Rybin, G. Thorleifsson, V.

Steinthorsdottir, P. Henneman, H. Grallert, A. Dehghan, J. J. Hottenga, C. S. Franklin, P. Navarro, K. Song, A. Goel, J. R. B. Perry, J. M. Egan, T. Lajunen, N. Grarup, T. Sparsø, A. Doney, B. F. Voight, H. M. Stringham, M. Li, S. Kanoni, P. Shrader, C. Cavalcanti-Proença, M. Kumari, L. Qi, N. J. Timpson, C. Gieger, C. Zabena, G. Rocheleau, E. Ingelsson, P. An, J. O'Connell, J. 'an Luan, A. Elliott, S. A. McCarroll, F. Payne, R. M. Roccasecca, F. Pattou, P. Sethupathy, K. Ardlie, Y. Ariyurek, B. Balkau, P. Barter, J. P. Beilby, Y. Ben-Shlomo, R. Benediktsson, A. J. Bennett, S. Bergmann, M. Bochud, E. Boerwinkle, A. Bonnefond, L. L. Bonnycastle, K. Borch-Johnsen, Y. Böttcher, E. Brunner, S. J. Bumpstead, G. Charpentier, Y.-D. I. Chen, P. Chines, R. Clarke, L. J. M. Coin, M. N. Cooper, M. Cornelis, G. Crawford, L. Crisponi, I. N. M. Day, E. J. C. de Geus, J. Delplanque, C. Dina, M. R. Erdos, A. C. Fedson, A. Fischer-Rosinsky, N. G. Forouhi, C. S. Fox, R. Frants, M. G. Franzosi, P. Galan, M. O. Goodarzi, J. Graessler, C. J. Groves, S. Grundy, R. Gwilliam, U. Gyllensten, S. Hadjadj, G. Hallmans, N. Hammond, X. Han, A.-L. Hartikainen, N. Hassanali, C. Hayward, S. C. Heath, S. Hercberg, C. Herder, A. A. Hicks, D. R. Hillman, A. D. Hingorani, A. Hofman, J. Hui, J. Hung, B. Isomaa, P. R. V. Johnson, T. Jørgensen, A. Jula, M. Kaakinen, J. Kaprio, Y. A. Kesaniemi, M. Kivimaki, B. Knight, S. Koskinen, P. Kovacs, K. O. Kyvik, G. M. Lathrop, D. A. Lawlor, O. Le Bacquer, C. Lecoeur, Y. Li, V. Lyssenko, R. Mahley, M. Mangino, A. K. Manning, M. T. Martínez-Larrad, J. B. McAteer, L. J. McCulloch, R. McPherson, C. Meisinger, D. Melzer, D. Meyre, B. D. Mitchell, M. A. Morken, S. Mukherjee, S. Naitza, N. Narisu, M. J. Neville, B. A. Oostra, M. Orrù, R. Pakyz, C. N. A. Palmer, G. Paolisso, C. Pattaro, D. Pearson, J. F. Peden, N. L. Pedersen, M. Perola, A. F. H. Pfeiffer, I. Pichler, O. Polasek, D. Posthuma, S. C. Potter, A. Pouta, M. A. Province, B. M. Psaty, W. Rathmann, N. W. Rayner, K. Rice, S. Ripatti, F. Rivadeneira, M. Roden, O. Rolandsson, A. Sandbaek, M. Sandhu, S. Sanna, A. A. Sayer, P. Scheet, L. J. Scott, U. Seedorf, S. J. Sharp, B. Shields, G. Sigurethsson, E. J. G. Sijbrands, A. Silveira, L. Simpson, A. Singleton, N. L. Smith, U. Sovio, A. Swift, H. Syddall, A.-C. Syvänen, T. Tanaka, B. Thorand, J. Tichet, A. Tönjes, T. Tuomi, A. G. Uitterlinden, K. W. van Dijk, M. van Hoek, D. Varma, S. Visvikis-Siest, V. Vitart, N. Vogelzangs, G. Waeber, P. J. Wagner, A. Walley, G. B. Walters, K. L. Ward, H. Watkins, M. N. Weedon, S. H. Wild, G. Willemsen, J. C. M. Witteman, J. W. G. Yarnell, E. Zeggini, D. Zelenika, B. Zethelius, G. Zhai, J. H. Zhao, M. C. Zillikens, DIAGRAM Consortium, GIANT Consortium, Global BPgen Consortium, I. B. Borecki, R. J. F. Loos, P. Meneton, P. K. E. Magnusson, D. M. Nathan, G. H. Williams, A. T. Hattersley, K. Silander, V. Salomaa, G. D. Smith, S. R. Bornstein, P. Schwarz, J. Spranger, F. Karpe, A. R. Shuldiner, C. Cooper, G. V. Dedoussis, M. Serrano-Ríos, A. D. Morris, L. Lind, L. J. Palmer, F. B. Hu, P. W. Franks, S. Ebrahim, M. Marmot, W. H. L. Kao, J. S. Pankow, M. J. Sampson, J. Kuusisto, M. Laakso, T. Hansen, O. Pedersen, P. P. Pramstaller, H. E. Wichmann, T. Illig, I. Rudan, A. F. Wright, M. Stumvoll, H. Campbell, J. F. Wilson, Anders Hamsten on behalf of Procardis Consortium, MAGIC investigators, R. N. Bergman, T. A. Buchanan, F. S. Collins, K. L. Mohlke, J. Tuomilehto, T. T. Valle, D. Altshuler, J. I. Rotter, D. S. Siscovick, B. W. J. H. Penninx, D. I. Boomsma, P. Deloukas, T. D. Spector, T. M. Frayling, L. Ferrucci, A. Kong, U. Thorsteinsdottir, K. Stefansson, C. M. van Duijn, Y. S. Aulchenko, A. Cao, A. Scuteri, D. Schlessinger, M. Uda, A.

Ruokonen, M.-R. Jarvelin, D. M. Waterworth, P. Vollenweider, L. Peltonen, V. Mooser, G. R. Abecasis, N. J. Wareham, R. Sladek, P. Froguel, R. M. Watanabe, J. B. Meigs, L. Groop, M. Boehnke, M. I. McCarthy, J. C. Florez, I. Barroso, New genetic loci implicated in fasting glucose homeostasis and their impact on type 2 diabetes risk. *Nat. Genet.* **42**, 105–116 (2010).

16. A. Mahajan, C. N. Spracklen, W. Zhang, M. C. Y. Ng, L. E. Petty, H. Kitajima, G. Z. Yu, S. Rüeger, L. Speidel, Y. J. Kim, M. Horikoshi, J. M. Mercader, D. Taliun, S. Moon, S.-H. Kwak, N. R. Robertson, N. W. Rayner, M. Loh, B.-J. Kim, J. Chiou, I. Miguel-Escalada, P. D. Briotta Parolo, K. Lin, F. Bragg, M. H. Preuss, F. Takeuchi, J. Nano, X. Guo, A. Lamri, M. Nakatochi, R. A. Scott, J.-J. Lee, A. Huerta-Chagoya, M. Graff, J.-F. Chai, E. J. Parra, J. Yao, L. F. Bielak, Y. Tabara, Y. Hai, V. Steinthorsdottir, J. P. Cook, M. Kals, N. Grarup, E. M. Schmidt, I. Pan, T. Sofer, M. Wuttke, C. Sarnowski, C. Gieger, D. Nousome, S. Trompet, J. Long, M. Sun, L. Tong, W.-M. Chen, M. Ahmad, R. Noordam, V. J. Y. Lim, C. H. T. Tam, Y. Y. Joo, C.-H. Chen, L. M. Raffield, C. Lecoeur, N. M. Maruthur, B. P. Prins, A. Nicolas, L. R. Yanek, G. Chen, R. A. Jensen, S. Tajuddin, E. Kabagambe, P. An, A. H. Xiang, H. S. Choi, B. E. Cade, J. Tan, F. Abaitua, L. S. Adair, A. Adeyemo, C. A. Aguilar-Salinas, M. Akiyama, S. S. Anand, A. Bertoni, Z. Bian, J. Bork-Jensen, I. Brandslund, J. A. Brody, C. M. Brummett, T. A. Buchanan, M. Canouil, J. C. N. Chan, L.-C. Chang, M.-L. Chee, J. Chen, S.-H. Chen, Y.-T. Chen, Z. Chen, L.-M. Chuang, M. Cushman, S. K. Das, H. J. de Silva, G. Dedoussis, L. Dimitrov, A. P. Doumatey, S. Du, Q. Duan, K.-U. Eckardt, L. S. Emery, D. S. Evans, M. K. Evans, K. Fischer, J. S. Floyd, I. Ford, M. Fornage, O. H. Franco, T. M. Frayling, B. I. Freedman, C. Fuchsberger, P. Genter, H. C. Gerstein, V. Giedraitis, C. González-Villalpando, M. E. González-Villalpando, M. O. Goodarzi, P. Gordon-Larsen, D. Gorkin, M. Gross, Y. Guo, S. Hackinger, S. Han, A. T. Hattersley, C. Herder, A.-G. Howard, W. Hsueh, M. Huang, W. Huang, Y.-J. Hung, M. Y. Hwang, C.-M. Hwu, S. Ichihara, M. A. Ikram, M. Ingelsson, M. T. Islam, M. Isono, H.-M. Jang, F. Jasmine, G. Jiang, J. B. Jonas, M. E. Jørgensen, T. Jørgensen, Y. Kamatani, F. R. Kandeel, A. Kasturiratne, T. Katsuya, V. Kaur, T. Kawaguchi, J. M. Keaton, A. N. Kho, C.-C. Khor, M. G. Kibriya, D.-H. Kim, K. Kohara, J. Kriebel, F. Kronenberg, J. Kuusisto, K. Läll, L. A. Lange, M.-S. Lee, N. R. Lee, A. Leong, L. Li, Y. Li, R. Li-Gao, S. Ligthart, C. M. Lindgren, A. Linneberg, C.-T. Liu, J. Liu, A. E. Locke, T. Louie, J. Luan, A. O. Luk, X. Luo, J. Lv, V. Lyssenko, V. Mamakou, K. R. Mani, T. Meitinger, A. Metspalu, A. D. Morris, G. N. Nadkarni, J. L. Nadler, M. A. Nalls, U. Nayak, I. Ntalla, Y. Okada, L. Orozco, S. R. Patel, M. A. Pereira, A. Peters, F. J. Pirie, B. Porneala, G. Prasad, S. Preissl, L. J. Rasmussen-Torvik, A. P. Reiner, M. Roden, R. Rohde, K. Roll, C. Sabanayagam, M. Sander, K. Sandow, N. Sattar, S. Schönherr, C. Schurmann, M. Shahriar, J. Shi, D. M. Shin, D. Shriner, J. A. Smith, W. Y. So, A. Stančáková, A. M. Stilp, K. Strauch, K. Suzuki, A. Takahashi, K. D. Taylor, B. Thorand, G. Thorleifsson, U. Thorsteinsdottir, B. Tomlinson, J. M. Torres, F.-J. Tsai, J. Tuomilehto, T. Tusie-Luna, M. S. Udler, A. Valladares-Salgado, R. M. van Dam, J. B. van Klinken, R. Varma, M. Vujkovic, N. Wacher-Rodarte, E. Wheeler, E. A. Whitsel, A. R. Wickremasinghe, K. W. van Dijk, D. R. Witte, C. S. Yajnik, K. Yamamoto, T. Yamauchi, L. Yengo, K. Yoon, C. Yu, J.-M. Yuan, S. Yusuf, L. Zhang, W. Zheng, F. Gen, L. J. Raffel, M. Igase, E. Ipp, S. Redline, Y. S.

Cho, L. Lind, M. A. Province, C. L. Hanis, P. A. Peyser, E. Ingelsson, A. B. Zonderman, B. M. Psaty, Y.-X. Wang, C. N. Rotimi, D. M. Becker, F. Matsuda, Y. Liu, E. Zeggini, M. Yokota, S. S. Rich, C. Kooperberg, J. S. Pankow, J. C. Engert, Y.-D. I. Chen, P. Froguel, J. G. Wilson, W. H. H. Sheu, S. L. R. Kardia, J.-Y. Wu, M. G. Hayes, R. C. W. Ma, T.-Y. Wong, L. Groop, D. O. Mook-Kanamori, G. R. Chandak, F. S. Collins, D. Bharadwaj, G. Paré, M. M. Sale, H. Ahsan, A. A. Motala, X.-O. Shu, K.-S. Park, J. W. Jukema, M. Cruz, R. McKean-Cowdin, H. Grallert, C.-Y. Cheng, E. P. Bottinger, A. Dehghan, E.-S. Tai, J. Dupuis, N. Kato, M. Laakso, A. Köttgen, W.-P. Koh, C. N. A. Palmer, S. Liu, G. Abecasis, J. S. Kooner, R. J. F. Loos, K. E. North, C. A. Haiman, J. C. Florez, D. Saleheen, T. Hansen, O. Pedersen, R. Mägi, C. Langenberg, N. J. Wareham, S. Maeda, T. Kadowaki, J. Lee, I. Y. Millwood, R. G. Walters, K. Stefansson, S. R. Myers, J. Ferrer, K. J. Gaulton, J. B. Meigs, K. L. Mohlke, A. L. Gloyn, D. W. Bowden, J. E. Below, J. C. Chambers, X. Sim, M. Boehnke, J. I. Rotter, M. I. McCarthy, A. P. Morris, Trans-ancestry genetic study of type 2 diabetes highlights the power of diverse populations for discovery and translation. *bioRxiv* (2020), , doi:10.1101/2020.09.22.20198937.

17. J. Chen, C. N. Spracklen, G. Marenne, A. Varshney, L. J. Corbin, J. ’an Luan, S. M. Willems, Y. Wu, X. Zhang, M. Horikoshi, T. S. Boutin, R. Mägi, J. Waage, R. Li-Gao, K. H. K. Chan, J. Yao, M. D. Anasanti, A. Y. Chu, A. Claringbould, J. Heikkinen, J. Hong, J.-J. Hottenga, S. Huo, M. A. Kaakinen, T. Louie, W. März, H. Moreno-Macias, A. Ndungu, S. C. Nelson, I. M. Nolte, K. E. North, C. K. Raulerson, D. Ray, R. Rohde, D. Rybin, C. Schurmann, X. Sim, L. Southam, I. D. Stewart, C. A. Wang, Y. Wang, P. Wu, W. Zhang, T. S. Ahluwalia, E. V. R. Appel, L. F. Bielak, J. A. Brody, N. P. Burtt, C. P. Cabrera, B. E. Cade, J. F. Chai, X. Chai, L.-C. Chang, C.-H. Chen, B. H. Chen, K. N. Chitrala, Y.-F. Chiu, H. G. de Haan, G. E. Delgado, A. Demirkan, Q. Duan, J. Engmann, S. A. Fatumo, J. Gayán, F. Giulianini, J. H. Gong, S. Gustafsson, Y. Hai, F. P. Hartwig, J. He, Y. Heianza, T. Huang, A. Huerta-Chagoya, M. Y. Hwang, R. A. Jensen, T. Kawaguchi, K. A. Kentistou, Y. J. Kim, M. E. Kleber, I. K. Kooner, S. Lai, L. A. Lange, C. D. Langefeld, M. Lauzon, M. Li, S. Ligthart, J. Liu, M. Loh, J. Long, V. Lyssenko, M. Mangino, C. Marzi, M. E. Montasser, A. Nag, M. Nakatochi, D. Noce, R. Noordam, G. Pistis, M. Preuss, L. Raffield, L. J. Rasmussen-Torvik, S. S. Rich, N. R. Robertson, R. Rueedi, K. Ryan, S. Sanna, R. Saxena, K. E. Schraut, B. Sennblad, K. Setoh, A. V. Smith, T. Sparsø, R. J. Strawbridge, F. Takeuchi, J. Tan, S. Trompet, E. van den Akker, P. J. van der Most, N. Verweij, M. Vogel, H. Wang, C. Wang, N. Wang, H. R. Warren, W. Wen, T. Wilsgaard, A. Wong, A. R. Wood, T. Xie, M. H. Zafarmand, J.-H. Zhao, W. Zhao, N. Amin, Z. Arzumanyan, A. Astrup, S. J. L. Bakker, D. Baldassarre, M. Beekman, R. N. Bergman, A. Bertoni, M. Blüher, L. L. Bonnycastle, S. R. Bornstein, D. W. Bowden, Q. Cai, A. Campbell, H. Campbell, Y. C. Chang, E. J. C. de Geus, A. Dehghan, S. Du, G. Eiriksdottir, A. E. Farmaki, M. Frånberg, C. Fuchsberger, Y. Gao, A. P. Gjesing, A. Goel, S. Han, C. A. Hartman, C. Herder, A. A. Hicks, C.-H. Hsieh, W. A. Hsueh, S. Ichihara, M. Igase, M. A. Ikram, W. C. Johnson, M. E. Jørgensen, P. K. Joshi, R. R. Kalyani, F. R. Kandeel, T. Katsuya, C. C. Khor, W. Kiess, I. Kolcic, T. Kuulasmaa, J. Kuusisto, K. Läll, K. Lam, D. A. Lawlor, N. R. Lee, R. N. Lemaitre, H. Li, Lifelines Cohort Study,

S.-Y. Lin, J. Lindström, A. Linneberg, J. Liu, C. Lorenzo, T. Matsubara, F. Matsuda, G. Mingrone, S. Mooijaart, S. Moon, T. Nabika, G. N. Nadkarni, J. L. Nadler, M. Nelis, M. J. Neville, J. M. Norris, Y. Ohyagi, A. Peters, P. A. Peyser, O. Polasek, Q. Qi, D. Raven, D. F. Reilly, A. Reiner, F. Rivideneira, K. Roll, I. Rudan, C. Sabanayagam, K. Sandow, N. Sattar, A. Schürmann, J. Shi, H. M. Stringham, K. D. Taylor, T. M. Teslovich, B. Thuesen, P. R. H. J. Timmers, E. Tremoli, M. Y. Tsai, A. Uitterlinden, R. M. van Dam, D. van Heemst, A. van Hylckama Vlieg, J. V. van Vliet-Ostaptchouk, J. Vangipurapu, H. Vestergaard, T. Wang, K. Willems van Dijk, T. Zemunik, G. R. Abecasis, L. S. Adair, C. A. Aguilar-Salinas, M. E. Alarcón-Riquelme, P. An, L. Aviles-Santa, D. M. Becker, L. J. Beilin, S. Bergmann, H. Bisgaard, C. Black, M. Boehnke, E. Boerwinkle, B. O. Böhm, K. Bønnelykke, D. I. Boomsma, E. P. Bottinger, T. A. Buchanan, M. Canouil, M. J. Caulfield, J. C. Chambers, D. I. Chasman, Y.-D. I. Chen, C.-Y. Cheng, F. S. Collins, A. Correa, F. Cucca, H. J. de Silva, G. Dedoussis, S. Elmståhl, M. K. Evans, E. Ferrannini, L. Ferrucci, J. C. Florez, P. W. Franks, T. M. Frayling, P. Froguel, B. Gigante, M. O. Goodarzi, P. Gordon-Larsen, H. Grallert, N. Grarup, S. Grimsgaard, L. Groop, V. Gudnason, X. Guo, A. Hamsten, T. Hansen, C. Hayward, S. R. Heckbert, B. L. Horta, W. Huang, E. Ingelsson, P. S. James, M.-R. Jarvelin, J. B. Jonas, J. W. Jukema, P. Kaleebu, R. Kaplan, S. L. R. Kardia, N. Kato, S. M. Keinanen-Kiukaanniemi, B.-J. Kim, M. Kivimaki, H. A. Koistinen, J. S. Kooner, A. Körner, P. Kovacs, D. Kuh, M. Kumari, Z. Kutalik, M. Laakso, T. A. Lakka, L. J. Launer, K. Leander, H. Li, X. Lin, L. Lind, C. Lindgren, S. Liu, R. J. F. Loos, P. K. E. Magnusson, A. Mahajan, A. Metspalu, D. O. Mook-Kanamori, T. A. Mori, P. B. Munroe, I. Njølstad, J. R. O'Connell, A. J. Oldehinkel, K. K. Ong, S. Padmanabhan, C. N. A. Palmer, N. D. Palmer, O. Pedersen, C. E. Pennell, D. J. Porteous, P. P. Pramstaller, M. A. Province, B. M. Psaty, L. Qi, L. J. Raffel, R. Rauramaa, S. Redline, P. M. Ridker, F. R. Rosendaal, T. E. Saaristo, M. Sandhu, J. Saramies, N. Schneiderman, P. Schwarz, L. J. Scott, E. Selvin, P. Sever, X.-O. Shu, P. E. Slagboom, K. S. Small, B. H. Smith, H. Snieder, T. Sofer, T. I. A. Sørensen, T. D. Spector, A. Stanton, C. J. Steves, M. Stumvoll, L. Sun, Y. Tabara, E. S. Tai, N. J. Timpson, A. Tönjes, J. Tuomilehto, T. Tusie, M. Uusitupa, P. van der Harst, C. van Duijn, V. Vitart, P. Vollenweider, T. G. M. Vrijkotte, L. E. Wagenknecht, M. Walker, Y. X. Wang, N. J. Wareham, R. M. Watanabe, H. Watkins, W. B. Wei, A. R. Wickremasinghe, G. Willemsen, J. F. Wilson, T.-Y. Wong, J.-Y. Wu, A. H. Xiang, L. R. Yanek, L. Yengo, M. Yokota, E. Zeggini, W. Zheng, A. B. Zonderman, J. I. Rotter, A. L. Gloyn, M. I. McCarthy, J. Dupuis, J. B. Meigs, R. A. Scott, I. Prokopenko, A. Leong, C.-T. Liu, S. C. J. Parker, K. L. Mohlke, C. Langenberg, E. Wheeler, A. P. Morris, I. Barroso, Meta-Analysis of Glucose and Insulin-related Traits Consortium (MAGIC), The trans-ancestral genomic architecture of glycemic traits. *Nat. Genet.* **53**, 840–860 (2021).

18. A. Mahajan, D. Taliun, M. Thurner, N. R. Robertson, J. M. Torres, N. W. Rayner, A. J. Payne, V. Steinthorsdottir, R. A. Scott, N. Grarup, J. P. Cook, E. M. Schmidt, M. Wuttke, C. Sarnowski, R. Mägi, J. Nano, C. Gieger, S. Trompet, C. Lecoeur, M. H. Preuss, B. P. Prins, X. Guo, L. F. Bielak, J. E. Below, D. W. Bowden, J. C. Chambers, Y. J. Kim, M. C. Y. Ng, L. E. Petty, X. Sim, W. Zhang, A. J. Bennett, J. Bork-Jensen, C. M. Brummett, M. Canouil, E. K. Ku, K. Fischer, S. L. R. Kardia, F.

Kronenberg, K. Läll, C. T. Liu, A. E. Locke, J. Luan, I. Ntalla, V. Nylander, S. Schönherr, C. Schurmann, L. Yengo, E. P. Bottinger, I. Brandslund, C. Christensen, G. Dedoussis, J. C. Florez, I. Ford, O. H. Franco, T. M. Frayling, V. Giedraitis, S. Hackinger, A. T. Hattersley, C. Herder, M. A. Ikram, M. Ingelsson, M. E. Jørgensen, T. Jørgensen, J. Kriebel, J. Kuusisto, S. Ligthart, C. M. Lindgren, A. Linneberg, V. Lyssenko, V. Mamakou, T. Meitinger, K. L. Mohlke, A. D. Morris, G. Nadkarni, J. S. Pankow, A. Peters, N. Sattar, A. Stančáková, K. Strauch, K. D. Taylor, B. Thorand, G. Thorleifsson, U. Thorsteinsdottir, J. Tuomilehto, D. R. Witte, J. Dupuis, P. A. Peyser, E. Zeggini, R. J. F. Loos, P. Froguel, E. Ingelsson, L. Lind, L. Groop, M. Laakso, F. S. Collins, J. W. Jukema, C. N. A. Palmer, H. Grallert, A. Metspalu, A. Dehghan, A. Köttgen, G. R. Abecasis, J. B. Meigs, J. I. Rotter, J. Marchini, O. Pedersen, T. Hansen, C. Langenberg, N. J. Wareham, K. Stefansson, A. L. Gloyn, A. P. Morris, M. Boehnke, M. I. McCarthy, Fine-mapping type 2 diabetes loci to single-variant resolution using high-density imputation and islet-specific epigenome maps. *Nat. Genet.* **50** (2018), doi:10.1038/s41588-018-0241-6.

19. A. Mahajan, J. Wessel, S. M. Willems, W. Zhao, N. R. Robertson, A. Y. Chu, W. Gan, H. Kitajima, D. Taliun, N. W. Rayner, X. Guo, Y. Lu, M. Li, R. A. Jensen, Y. Hu, S. Huo, K. K. Lohman, W. Zhang, J. P. Cook, B. P. Prins, J. Flannick, N. Grarup, V. V. Trubetskoy, J. Kravic, Y. J. Kim, D. V. Rybin, H. Yaghootkar, M. Müller-Nurasyid, K. Meidtner, R. Li-Gao, T. V. Varga, J. Marten, J. Li, A. V. Smith, P. An, S. Ligthart, S. Gustafsson, G. Malerba, A. Demirkan, J. F. Tajes, V. Steinthorsdottir, M. Wuttke, C. Lecoeur, M. Preuss, L. F. Bielak, M. Graff, H. M. Highland, A. E. Justice, D. J. Liu, E. Marouli, G. M. Peloso, H. R. Warren, ExomeBP Consortium, MAGIC Consortium, GIANT Consortium, S. Afaq, S. Afzal, E. Ahlqvist, P. Almgren, N. Amin, L. B. Bang, A. G. Bertoni, C. Bombieri, J. Bork-Jensen, I. Brandslund, J. A. Brody, N. P. Burtt, M. Canouil, Y.-D. I. Chen, Y. S. Cho, C. Christensen, S. V. Eastwood, K.-U. Eckardt, K. Fischer, G. Gambaro, V. Giedraitis, M. L. Grove, H. G. de Haan, S. Hackinger, Y. Hai, S. Han, A. Tybjærg-Hansen, M.-F. Hivert, B. Isomaa, S. Jäger, M. E. Jørgensen, T. Jørgensen, A. Käräjämäki, B.-J. Kim, S. S. Kim, H. A. Koistinen, P. Kovacs, J. Kriebel, F. Kronenberg, K. Läll, L. A. Lange, J.-J. Lee, B. Lehne, H. Li, K.-H. Lin, A. Linneberg, C.-T. Liu, J. Liu, M. Loh, R. Mägi, V. Mamakou, R. McKean-Cowdin, G. Nadkarni, M. Neville, S. F. Nielsen, I. Ntalla, P. A. Peyser, W. Rathmann, K. Rice, S. S. Rich, L. Rode, O. Rolandsson, S. Schönherr, E. Selvin, K. S. Small, A. Stančáková, P. Surendran, K. D. Taylor, T. M. Teslovich, B. Thorand, G. Thorleifsson, A. Tin, A. Tönjes, A. Varbo, D. R. Witte, A. R. Wood, P. Yajnik, J. Yao, L. Yengo, R. Young, P. Amouyel, H. Boeing, E. Boerwinkle, E. P. Bottinger, R. Chowdhury, F. S. Collins, G. Dedoussis, A. Dehghan, P. Deloukas, M. M. Ferrario, J. Ferrières, J. C. Florez, P. Frossard, V. Gudnason, T. B. Harris, S. R. Heckbert, J. M. M. Howson, M. Ingelsson, S. Kathiresan, F. Kee, J. Kuusisto, C. Langenberg, L. J. Launer, C. M. Lindgren, S. Männistö, T. Meitinger, O. Melander, K. L. Mohlke, M. Moitry, A. D. Morris, A. D. Murray, R. de Mutsert, M. Orho-Melander, K. R. Owen, M. Perola, A. Peters, M. A. Province, A. Rasheed, P. M. Ridker, F. Rivadineira, F. R. Rosendaal, A. H. Rosengren, V. Salomaa, W. H.-H. Sheu, R. Sladek, B. H. Smith, K. Strauch, A. G. Uitterlinden, R. Varma, C. J. Willer, M. Blüher, A. S. Butterworth, J. C. Chambers, D. I. Chasman, J. Danesh, C. van Duijn, J. Dupuis, O. H. Franco, P. W.

Franks, P. Froguel, H. Grallert, L. Groop, B.-G. Han, T. Hansen, A. T. Hattersley, C. Hayward, E. Ingelsson, S. L. R. Kardia, F. Karpe, J. S. Kooner, A. Köttgen, K. Kuulasmaa, M. Laakso, X. Lin, L. Lind, Y. Liu, R. J. F. Loos, J. Marchini, A. Metspalu, D. Mook-Kanamori, B. G. Nordestgaard, C. N. A. Palmer, J. S. Pankow, O. Pedersen, B. M. Psaty, R. Rauramaa, N. Sattar, M. B. Schulze, N. Soranzo, T. D. Spector, K. Stefansson, M. Stumvoll, U. Thorsteinsdottir, T. Tuomi, J. Tuomilehto, N. J. Wareham, J. G. Wilson, E. Zeggini, R. A. Scott, I. Barroso, T. M. Frayling, M. O. Goodarzi, J. B. Meigs, M. Boehnke, D. Saleheen, A. P. Morris, J. I. Rotter, M. I. McCarthy, Refining the accuracy of validated target identification through coding variant fine-mapping in type 2 diabetes. *Nat. Genet.* **50**, 559–571 (2018).

20. SIGMA Type 2 Diabetes Consortium, A. L. Williams, S. B. R. Jacobs, H. Moreno-Macías, A. Huerta-Chagoya, C. Churchhouse, C. Márquez-Luna, H. García-Ortíz, M. J. Gómez-Vázquez, N. P. Burtt, C. A. Aguilar-Salinas, C. González-Villalpando, J. C. Florez, L. Orozco, C. A. Haiman, T. Tusié-Luna, D. Altshuler, Sequence variants in SLC16A11 are a common risk factor for type 2 diabetes in Mexico. *Nature*. **506**, 97–101 (2014).

21. C. N. Spracklen, M. Horikoshi, Y. J. Kim, K. Lin, F. Bragg, S. Moon, K. Suzuki, C. H. T. Tam, Y. Tabara, S.-H. Kwak, F. Takeuchi, J. Long, V. J. Y. Lim, J.-F. Chai, C.-H. Chen, M. Nakatochi, J. Yao, H. S. Choi, A. K. Iyengar, H. J. Perrin, S. M. Brotman, M. van de Bunt, A. L. Gloyn, J. E. Below, M. Boehnke, D. W. Bowden, J. C. Chambers, A. Mahajan, M. I. McCarthy, M. C. Y. Ng, L. E. Petty, W. Zhang, A. P. Morris, L. S. Adair, M. Akiyama, Z. Bian, J. C. N. Chan, L.-C. Chang, M.-L. Chee, Y.-D. I. Chen, Y.-T. Chen, Z. Chen, L.-M. Chuang, S. Du, P. Gordon-Larsen, M. Gross, X. Guo, Y. Guo, S. Han, A.-G. Howard, W. Huang, Y.-J. Hung, M. Y. Hwang, C.-M. Hwu, S. Ichihara, M. Isono, H.-M. Jang, G. Jiang, J. B. Jonas, Y. Kamatani, T. Katsuya, T. Kawaguchi, C.-C. Khor, K. Kohara, M.-S. Lee, N. R. Lee, L. Li, J. Liu, A. O. Luk, J. Lv, Y. Okada, M. A. Pereira, C. Sabanayagam, J. Shi, D. M. Shin, W. Y. So, A. Takahashi, B. Tomlinson, F.-J. Tsai, R. M. van Dam, Y.-B. Xiang, K. Yamamoto, T. Yamauchi, K. Yoon, C. Yu, J.-M. Yuan, L. Zhang, W. Zheng, M. Igase, Y. S. Cho, J. I. Rotter, Y.-X. Wang, W. H. H. Sheu, M. Yokota, J.-Y. Wu, C.-Y. Cheng, T.-Y. Wong, X.-O. Shu, N. Kato, K.-S. Park, E.-S. Tai, F. Matsuda, W.-P. Koh, R. C. W. Ma, S. Maeda, I. Y. Millwood, J. Lee, T. Kadowaki, R. G. Walters, B.-J. Kim, K. L. Mohlke, X. Sim, Identification of type 2 diabetes loci in 433,540 East Asian individuals. *Nature*. **582**, 240–245 (2020).

22. J. Flannick, J. M. Mercader, C. Fuchsberger, M. S. Udler, A. Mahajan, J. Wessel, T. M. Teslovich, L. Caulkins, R. Koesterer, F. Barajas-Olmos, T. W. Blackwell, E. Boerwinkle, J. A. Brody, F. Centeno-Cruz, L. Chen, S. Chen, C. Contreras-Cubas, E. Córdova, A. Correa, M. Cortes, R. A. DeFronzo, L. Dolan, K. L. Drews, A. Elliott, J. S. Floyd, S. Gabriel, M. E. Garay-Sevilla, H. García-Ortiz, M. Gross, S. Han, N. L. Heard-Costa, A. U. Jackson, M. E. Jørgensen, H. M. Kang, M. Kelsey, B.-J. Kim, H. A. Koistinen, J. Kuusisto, J. B. Leader, A. Linneberg, C.-T. Liu, J. Liu, V. Lyssenko, A. K. Manning, A. Marcketta, J. M. Malacara-Hernandez, A. Martínez-Hernández, K. Matsuo, E. Mayer-Davis, E. Mendoza-Caamal, K. L. Mohlke, A. C.

Morrison, A. Ndungu, M. C. Y. Ng, C. O'Dushlaine, A. J. Payne, C. Pihoker, Broad Genomics Platform, W. S. Post, M. Preuss, B. M. Psaty, R. S. Vasan, N. W. Rayner, A. P. Reiner, C. Revilla-Monsalve, N. R. Robertson, N. Santoro, C. Schurmann, W. Y. So, X. Soberón, H. M. Stringham, T. M. Strom, C. H. T. Tam, F. Thameem, B. Tomlinson, J. M. Torres, R. P. Tracy, R. M. van Dam, M. Vujkovic, S. Wang, R. P. Welch, D. R. Witte, T.-Y. Wong, G. Atzmon, N. Barzilai, J. Blangero, L. L. Bonnycastle, D. W. Bowden, J. C. Chambers, E. Chan, C.-Y. Cheng, Y. S. Cho, F. S. Collins, P. S. de Vries, R. Duggirala, B. Glaser, C. Gonzalez, M. E. Gonzalez, L. Groop, J. S. Kooner, S. H. Kwak, M. Laakso, D. M. Lehman, P. Nilsson, T. D. Spector, E. S. Tai, T. Tuomi, J. Tuomilehto, J. G. Wilson, C. A. Aguilar-Salinas, E. Bottinger, B. Burke, D. J. Carey, J. C. N. Chan, J. Dupuis, P. Frossard, S. R. Heckbert, M. Y. Hwang, Y. J. Kim, H. L. Kirchner, J.-Y. Lee, J. Lee, R. J. F. Loos, R. C. W. Ma, A. D. Morris, C. J. O'Donnell, C. N. A. Palmer, J. Pankow, K. S. Park, A. Rasheed, D. Saleheen, X. Sim, K. S. Small, Y. Y. Teo, C. Haiman, C. L. Hanis, B. E. Henderson, L. Orozco, T. Tusié-Luna, F. E. Dewey, A. Baras, C. Gieger, T. Meitinger, K. Strauch, L. Lange, N. Grarup, T. Hansen, O. Pedersen, P. Zeitler, D. Dabelea, G. Abecasis, G. I. Bell, N. J. Cox, M. Seielstad, R. Sladek, J. B. Meigs, S. S. Rich, J. I. Rotter, DiscovEHR Collaboration, CHARGE, LuCamp, ProDiGY, GoT2D, ESP, SIGMA-T2D, T2D-GENES, AMP-T2D-GENES, D. Altshuler, N. P. Burtt, L. J. Scott, A. P. Morris, J. C. Florez, M. I. McCarthy, M. Boehnke, Exome sequencing of 20,791 cases of type 2 diabetes and 24,440 controls. *Nature*. **570**, 71–76 (2019).

23. J. M. Gaziano, J. Concato, M. Brophy, L. Fiore, S. Pyarajan, J. Breeling, S. Whitbourne, J. Deen, C. Shannon, D. Humphries, P. Guarino, M. Aslan, D. Anderson, R. LaFleur, T. Hammond, K. Schaa, J. Moser, G. Huang, S. Muralidhar, R. Przygodzki, T. J. O'Leary, Million Veteran Program: A mega-biobank to study genetic influences on health and disease. *J. Clin. Epidemiol.* **70**, 214–223 (2016).

24. C. Bycroft, C. Freeman, D. Petkova, G. Band, L. T. Elliott, K. Sharp, A. Motyer, D. Vukcevic, O. Delaneau, J. O'Connell, A. Cortes, S. Welsh, A. Young, M. Effingham, G. McVean, S. Leslie, N. Allen, P. Donnelly, J. Marchini, The UK Biobank resource with deep phenotyping and genomic data. *Nature*. **562**, 203–209 (2018).

25. A. Nagai, M. Hirata, Y. Kamatani, K. Muto, K. Matsuda, Y. Kiyohara, T. Ninomiya, A. Tamakoshi, Z. Yamagata, T. Mushiroda, Y. Murakami, K. Yuji, Y. Furukawa, H. Zembutsu, T. Tanaka, Y. Ohnishi, Y. Nakamura, BioBank Japan Cooperative Hospital Group, M. Kubo, Overview of the BioBank Japan Project: Study design and profile. *J. Epidemiol.* **27**, S2–S8 (2017).

26. G. L. Wojcik, M. Graff, K. K. Nishimura, R. Tao, J. Haessler, C. R. Gignoux, H. M. Highland, Y. M. Patel, E. P. Sorokin, C. L. Avery, G. M. Belbin, S. A. Bien, I. Cheng, S. Cullina, C. J. Hodonsky, Y. Hu, L. M. Huckins, J. Jeff, A. E. Justice, J. M. Kocarnik, U. Lim, B. M. Lin, Y. Lu, S. C. Nelson, S.-S. L. Park, H. Poisner, M. H. Preuss, M. A. Richard, C. Schurmann, V. W. Setiawan, A. Sockell, K. Vahi, M. Verbanck, A. Vishnu, R. W. Walker, K. L. Young, N. Zubair, V. Acuña-Alonso, J. L. Ambite, K. C. Barnes, E. Boerwinkle, E. P. Bottinger, C. D. Bustamante, C.

Caberto, S. Canizales-Quinteros, M. P. Conomos, E. Deelman, R. Do, K. Doheny, L. Fernández-Rhodes, M. Fornage, B. Hailu, G. Heiss, B. M. Henn, L. A. Hindorff, R. D. Jackson, C. A. Laurie, C. C. Laurie, Y. Li, D.-Y. Lin, A. Moreno-Estrada, G. Nadkarni, P. J. Norman, L. C. Pooler, A. P. Reiner, J. Romm, C. Sabatti, K. Sandoval, X. Sheng, E. A. Stahl, D. O. Stram, T. A. Thornton, C. L. Wassel, L. R. Wilkens, C. A. Winkler, S. Yoneyama, S. Buyske, C. A. Haiman, C. Kooperberg, L. Le Marchand, R. J. F. Loos, T. C. Matise, K. E. North, U. Peters, E. E. Kenny, C. S. Carlson, Genetic analyses of diverse populations improves discovery for complex traits. *Nature*. **570**, 514–518 (2019).

27. M. Vujkovic, J. M. Keaton, J. A. Lynch, D. R. Miller, J. Zhou, C. Tcheandjieu, J. E. Huffman, T. L. Assimes, K. Lorenz, X. Zhu, A. T. Hilliard, R. L. Judy, J. Huang, K. M. Lee, D. Klarin, S. Pyarajan, J. Danesh, O. Melander, A. Rasheed, N. H. Mallick, S. Hameed, I. H. Qureshi, M. N. Afzal, U. Malik, A. Jalal, S. Abbas, X. Sheng, L. Gao, K. H. Kaestner, K. Susztak, Y. V. Sun, S. L. DuVall, K. Cho, J. S. Lee, J. M. Gaziano, L. S. Phillips, J. B. Meigs, P. D. Reaven, P. W. Wilson, T. L. Edwards, D. J. Rader, S. M. Damrauer, C. J. O'Donnell, P. S. Tsao, K.-M. Chang, B. F. Voight, D. Saleheen, Discovery of 318 new risk loci for type 2 diabetes and related vascular outcomes among 1.4 million participants in a multi-ancestry meta-analysis. *Nat. Genet.* **52**, 680–691 (2020).

28. L. M. Polfus, B. F. Darst, H. Highland, X. Sheng, M. C. Y. Ng, J. E. Below, L. Petty, S. Bien, X. Sim, W. Wang, P. Fontanillas, Y. Patel, 23andMe Research Team, DIAMANTE Hispanic/Latino Consortium, MEta-analysis of type 2 DIabetes in African Americans Consortium, M. Preuss, C. Schurmann, Z. Du, Y. Lu, S. K. Rhie, J. M. Mercader, T. Tusie-Luna, C. González-Villalpando, L. Orozco, C. N. Spracklen, B. E. Cade, R. A. Jensen, M. Sun, Y. Y. Joo, P. An, L. R. Yanek, L. F. Bielak, S. Tajuddin, A. Nicolas, G. Chen, L. Raffield, X. Guo, W.-M. Chen, G. N. Nadkarni, M. Graff, R. Tao, J. S. Pankow, M. Daviglus, Q. Qi, E. A. Boerwinkle, S. Liu, L. S. Phillips, U. Peters, C. Carlson, L. R. Wikens, L. L. Marchand, K. E. North, S. Buyske, C. Kooperberg, R. J. F. Loos, D. O. Stram, C. A. Haiman, Genetic discovery and risk characterization in type 2 diabetes across diverse populations. *HGG Adv*. **2** (2021), doi:10.1016/j.xhgg.2021.100029.

29. K. Suzuki, M. Akiyama, K. Ishigaki, M. Kanai, J. Hosoe, N. Shojima, A. Hozawa, A. Kadota, K. Kuriki, M. Naito, K. Tanno, Y. Ishigaki, M. Hirata, K. Matsuda, N. Iwata, M. Ikeda, N. Sawada, T. Yamaji, M. Iwasaki, S. Ikegawa, S. Maeda, Y. Murakami, K. Wakai, S. Tsugane, M. Sasaki, M. Yamamoto, Y. Okada, M. Kubo, Y. Kamatani, M. Horikoshi, T. Yamauchi, T. Kadowaki, Identification of 28 new susceptibility loci for type 2 diabetes in the Japanese population. *Nat. Genet.* **51**, 379–386 (2019).

30. J. B. Meigs, L. A. Cupples, P. W. Wilson, Parental transmission of type 2 diabetes: the Framingham Offspring Study. *Diabetes*. **49**, 2201–2207 (2000).

31. N. R. Wray, M. E. Goddard, P. M. Visscher, Prediction of individual genetic risk to disease from genome-wide association studies. *Genome Res.* **17**, 1520–1528 (2007).

32. J. B. Meigs, P. Shrader, L. M. Sullivan, J. B. McAteer, C. S. Fox, J. Dupuis, A. K. Manning, J. C. Florez, P. W. F. Wilson, R. B. D'Agostino Sr, L. A. Cupples, Genotype score in addition to common risk factors for prediction of type 2 diabetes. *N. Engl. J. Med.* **359**, 2208–2219 (2008).

33. C. Márquez-Luna, P.-R. Loh, South Asian Type 2 Diabetes (SAT2D) Consortium, SIGMA Type 2 Diabetes Consortium, A. L. Price, Multiethnic polygenic risk scores improve risk prediction in diverse populations. *Genet. Epidemiol.* **41**, 811–823 (2017).

34. A. V. Khera, M. Chaffin, K. G. Aragam, M. E. Haas, C. Roselli, S. H. Choi, P. Natarajan, E. S. Lander, S. A. Lubitz, P. T. Ellinor, S. Kathiresan, Genome-wide polygenic scores for common diseases identify individuals with risk equivalent to monogenic mutations. *Nat. Genet.* **50**, 1219–1224 (2018).

35. Y. D. Zhang, A. N. Hurson, H. Zhang, P. P. Choudhury, D. F. Easton, R. L. Milne, J. Simard, P. Hall, K. Michailidou, J. Dennis, M. K. Schmidt, J. Chang-Claude, P. Gharahkhani, D. Whiteman, P. T. Campbell, M. Hoffmeister, M. Jenkins, U. Peters, L. Hsu, S. B. Gruber, G. Casey, S. L. Schmit, T. A. O'Mara, A. B. Spurdle, D. J. Thompson, I. Tomlinson, I. De Vivo, M. T. Landi, M. H. Law, M. M. Iles, F. Demenais, R. Kumar, S. MacGregor, D. T. Bishop, S. V. Ward, M. L. Bondy, R. Houlston, J. K. Wiencke, B. Melin, J. Barnholtz-Sloan, B. Kinnersley, M. R. Wrensch, C. I. Amos, R. J. Hung, P. Brennan, J. McKay, N. E. Caporaso, S. I. Berndt, B. M. Birmann, N. J. Camp, P. Kraft, N. Rothman, S. L. Slager, A. Berchuck, P. D. P. Pharoah, T. A. Sellers, S. A. Gayther, C. L. Pearce, E. L. Goode, J. M. Schildkraut, K. B. Moysich, L. T. Amundadottir, E. J. Jacobs, A. P. Klein, G. M. Petersen, H. A. Risch, R. Z. Stolzenberg-Solomon, B. M. Wolpin, D. Li, R. A. Eeles, C. A. Haiman, Z. Kote-Jarai, F. R. Schumacher, A. A. Al Olama, M. P. Purdue, G. Scelo, M. D. Dalgaard, M. H. Greene, T. Grotmol, P. A. Kanetsky, K. A. McGlynn, K. L. Nathanson, C. Turnbull, F. Wiklund, Breast Cancer Association Consortium (BCAC), Barrett's and Esophageal Adenocarcinoma Consortium (BEACON), Colon Cancer Family Registry (CCFR), Transdisciplinary Studies of Genetic Variation in Colorectal Cancer (CORECT), Endometrial Cancer Association Consortium (ECAC), Genetics and Epidemiology of Colorectal Cancer Consortium (GECCO), Melanoma Genetics Consortium (GenoMEL), Glioma International Case-Control Study (GICC), International Lung Cancer Consortium (ILCCO), Integrative Analysis of Lung Cancer Etiology and Risk (INTEGRAL) Consortium, International Consortium of Investigators Working on Non-Hodgkin's Lymphoma Epidemiologic Studies (InterLymph), Ovarian Cancer Association Consortium (OCAC), Oral Cancer GWAS, Pancreatic Cancer Case-Control Consortium (PanC4), Pancreatic Cancer Cohort Consortium (PanScan), Prostate Cancer Association Group to Investigate Cancer Associated Alterations in the Genome (PRACTICAL), Renal Cancer GWAS, Testicular Cancer Consortium (TECAC), S. J. Chanock, N. Chatterjee, M. Garcia-Closas, Assessment of polygenic architecture and risk prediction based on common variants across fourteen cancers. *Nat. Commun.* **11**, 3353 (2020).

36. LifeView: Successful pregnancy, Healthy baby, (available at https://genomicprediction.com/).

37. Orchid: Have healthy babies, (available at https://www.orchidhealth.com/).

38. MyOme – Harnessing the True Power of Genetics, (available at https://myome.com/).

39. P. Turley, M. N. Meyer, N. Wang, D. Cesarini, E. Hammonds, A. R. Martin, B. M. Neale, H. L. Rehm, L. Wilkins-Haug, D. J. Benjamin, S. Hyman, D. Laibson, P. M. Visscher, Problems with Using Polygenic Scores to Select Embryos. *N. Engl. J. Med.* **385**, 78–86 (2021).

40. M. N. Weedon, M. I. McCarthy, G. Hitman, M. Walker, C. J. Groves, E. Zeggini, N. W. Rayner, B. Shields, K. R. Owen, A. T. Hattersley, T. M. Frayling, Combining information from common type 2 diabetes risk polymorphisms improves disease prediction. *PLoS Med.* **3**, e374 (2006).

41. V. Lyssenko, A. Jonsson, P. Almgren, N. Pulizzi, B. Isomaa, T. Tuomi, G. Berglund, D. Altshuler, P. Nilsson, L. Groop, Clinical risk factors, DNA variants, and the development of type 2 diabetes. *N. Engl. J. Med.* **359**, 2220–2232 (2008).

42. H. Lango, UK Type 2 Diabetes Genetics Consortium, C. N. A. Palmer, A. D. Morris, E. Zeggini, A. T. Hattersley, M. I. McCarthy, T. M. Frayling, M. N. Weedon, Assessing the combined impact of 18 common genetic variants of modest effect sizes on type 2 diabetes risk. *Diabetes*. **57**, 3129–3135 (2008).

43. J. L. Vassy, M.-F. Hivert, B. Porneala, M. Dauriz, J. C. Florez, J. Dupuis, D. S. Siscovick, M. Fornage, L. J. Rasmussen-Torvik, C. Bouchard, J. B. Meigs, Polygenic type 2 diabetes prediction at the limit of common variant detection. *Diabetes*. **63**, 2172–2182 (2014).

44. T. Chikowore, T. van Zyl, E. J. M. Feskens, K. R. Conradie, Predictive utility of a genetic risk score of common variants associated with type 2 diabetes in a black South African population. *Diabetes Res. Clin. Pract.* **122**, 1–8 (2016).

45. E. R. Seaquist, F. C. Goetz, S. Rich, J. Barbosa, Familial Clustering of Diabetic Kidney Disease. *N. Engl. J. Med.* **320**, 1161–1165 (1989).

46. J. B. Cole, J. C. Florez, Genetics of diabetes mellitus and diabetes complications. *Nature Reviews Nephrology*. **16** (2020), pp. 377–390.

47. A. R. Majithia, D. Altshuler, J. N. Hirschhorn, Genetics of Endocrinology. *Williams Textbook of Endocrinology* (2016), pp. 49–68.

48. J. A. Tennessen, A. W. Bigham, T. D. O'Connor, W. Fu, E. E. Kenny, S. Gravel, S. McGee, R. Do, X. Liu, G. Jun, H. M. Kang, D. Jordan, S. M. Leal, S. Gabriel, M. J. Rieder, G. Abecasis, D. Altshuler, D. A. Nickerson, E. Boerwinkle, S. Sunyaev, C.

D. Bustamante, M. J. Bamshad, J. M. Akey, G. O. Broad, G. O. Seattle, N. E. S. Project, Evolution and functional impact of rare coding variation from deep sequencing of human exomes. *Science*. **337**, 64–69 (2012).

49. O. Zuk, S. F. Schaffner, K. Samocha, R. Do, E. Hechter, S. Kathiresan, M. J. Daly, B. M. Neale, S. R. Sunyaev, E. S. Lander, Searching for missing heritability: designing rare variant association studies. *Proc. Natl. Acad. Sci. U. S. A.* **111**, E455–64 (2014).

50. A. M. Deaton, M. M. Parker, L. D. Ward, A. O. Flynn-Carroll, L. BonDurant, G. Hinkle, P. Akbari, L. A. Lotta, A. Baras, P. Nioi, Gene-level analysis of rare variants in 379,066 whole exome sequences identifies an association of GIGYF1 loss of function with type 2 diabetes. *Sci. Rep.* **11**, 1–16 (2021).

51. J. D. Backman, A. H. Li, A. Marcketta, D. Sun, J. Mbatchou, M. D. Kessler, C. Benner, D. Liu, A. E. Locke, S. Balasubramanian, A. Yadav, N. Banerjee, C. E. Gillies, A. Damask, S. Liu, X. Bai, A. Hawes, E. Maxwell, L. Gurski, K. Watanabe, J. A. Kosmicki, V. Rajagopal, J. Mighty, M. Jones, L. Mitnaul, E. Stahl, G. Coppola, E. Jorgenson, L. Habegger, W. J. Salerno, A. R. Shuldiner, L. A. Lotta, J. D. Overton, M. N. Cantor, J. G. Reid, G. Yancopoulos, H. M. Kang, J. Marchini, A. Baras, G. R. Abecasis, M. A. R. Ferreira, Exome sequencing and analysis of 454,787 UK Biobank participants. *Nature*. **599**, 628–634 (2021).

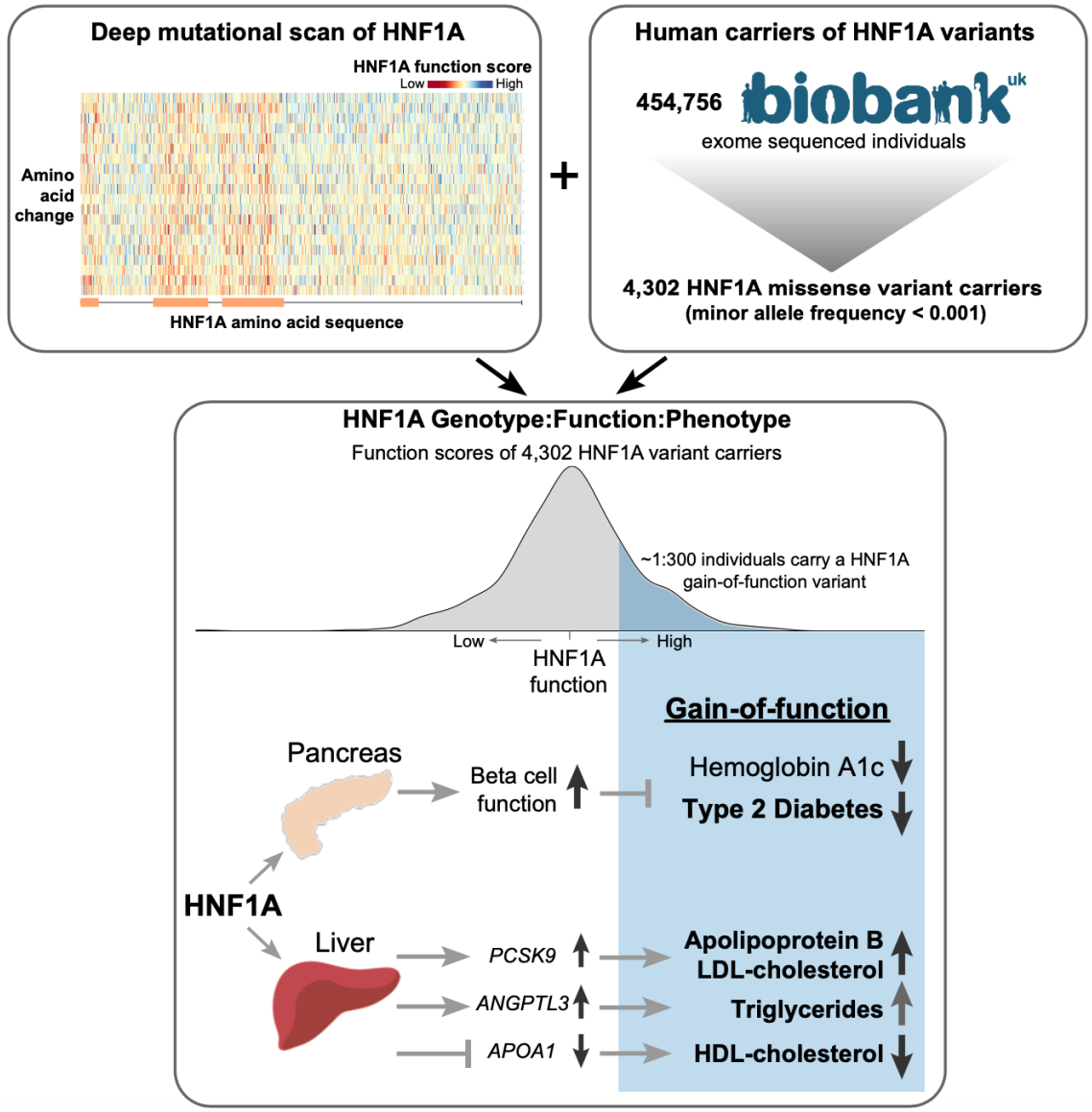
CHAPTER 2

Human gain-of-function variants in HNF1A confer protection from diabetes but independently increase hepatic secretion of atherogenic lipoproteins

Abstract

Loss-of-function mutations in Hepatocyte Nuclear Factor 1A (HNF1A) are known to cause rare forms of diabetes and alter hepatic physiology through unclear mechanisms. In the general population, 1:100 individuals carry a rare, protein-coding *HNF1A* variant, most of unknown functional consequence. To characterize the full allelic series, we performed deep mutational scanning of 11,970 protein-coding *HNF1A* variants in human hepatocytes and clinical correlation with 553,246 exome-sequenced individuals. Surprisingly, we found ~1:5 rare protein-coding *HNF1A* variants in the general population cause molecular gain-of-function (GOF), increasing the transcriptional activity of HNF1A by up to 50% and conferring protection from type 2 diabetes (OR=0.77, p=0.007). Increased hepatic expression of *HNF1A* promoted a pro-atherogenic serum profile mediated in part by enhanced transcription of risk genes including *ANGPTL3* and *PCSK9*. In summary, ~1:300 individuals carry a GOF variant in *HNF1A* that protects carriers from diabetes but enhances hepatic secretion of atherogenic lipoproteins.

Graphical Abstract



Highlights

- Deep mutational scan of HNF1A characterized the function of 11,970 HNF1A variants
- Gain-of-function HNF1A variants identified that increase transcriptional activity
- 1 in 300 rare HNF1A missense variants in humans confer gain of function
- Human carriers of HNF1A gain-of-function variants are protected from T2D

Introduction

Hepatocyte Nuclear Factor 1 alpha (*HNF1A*) is a lineage-determining transcription factor expressed in several tissues including the liver and pancreas(*1*). Rare, complete loss-of-function mutations (LOF) in *HNF1A* have been shown to cause an autosomal dominant monogenic form of diabetes (MODY3)(*2*, *3*) through deficiencies in pancreatic insulin secretion via a haploinsufficient genetic mechanism(*4*). MODY3 patients have been observed to carry altered levels of serum lipoproteins(*5*) and lower levels of the inflammatory marker high-sensitivity C-reactive protein (hsCRP)(*6*) than matched patients with type 2 diabetes (T2D), suggesting a potential liver-mediated functional role for HNF1A in metabolic disease pathogenesis. However, diabetes itself through underlying insulin resistance or elevated blood glucose can dysregulate serum lipids(*7*) and systemic inflammation(*8*), and thus these observations in diabetic individuals with MODY3 cannot evidentiate HNF1A as a direct hepatic regulator of serum lipids or inflammation. Murine LOF studies have implicated HNF1A in bile acid secretion and cholesterol metabolism but are also confounded by a diabetic phenotype(*9*). Population-based genetic association studies have corroborated associations between *HNF1A* and T2D(*10*), serum lipoproteins(*11*), and hsCRP(*12*), but do not provide insight into mechanism as most of the associated variants (SNPs) are either non-coding or have minimal experimentally observable consequences on protein function(*13*). Finally, common HNF1A SNPs have been associated with coronary artery disease (CAD) risk(*14*, *15*) but these SNPs are also associated with T2D and serum lipids, both of which are known, independent CAD mediators(*16*, *17*). Thus, the relationship between HNF1A

function, these risk factors, and the specific mechanisms of disease pathogenesis remains unclear.

Characterization of an allelic series of functional variants for the hepatic functions of *HNF1A* with clinical correlation would enable a mechanistic understanding of the complex role for *HNF1A* in driving metabolic disease in multiple tissues including the pancreas (via diabetes) and liver (serum lipids) with their respective directions-of-effect. In the past few years, dozens of novel protein-coding alleles in *HNF1A* have been identified from population-based exome sequencing(*18*). These variants are individually rare (minor allele frequency (MAF) < 0.001) and mostly of unknown functional consequence, but in aggregate are carried by over one percent of the population. The much greater allelic diversity of rare variants presents an opportunity to dissect the role of HNF1A in metabolic disease, if these variants could be functionally annotated at scale.

To address this gap in knowledge, we examined the full allelic series of protein-coding variants in *HNF1A* using deep mutational scanning, a high-throughput approach that has been successfully utilized to characterize protein-coding variants in clinically important genes(*19–21*). All possible single amino acid substitutions in HNF1A were tested for transcriptional activity in human hepatocytes. These comprehensive data were intersected with carriers of rare protein-coding variants (MAF<0.001) identified from among 553,246 sequenced individuals in order to relate variant to function to phenotype for metabolic disease associated factors and disease outcomes including T2D and CAD. We discover gain-of-function (GOF) variants which while conferring protection from T2D promote pro-inflammatory/thrombogenic gene expression, and an atherogenic lipid profile through liver specific enhancement of HNF1A target genes.

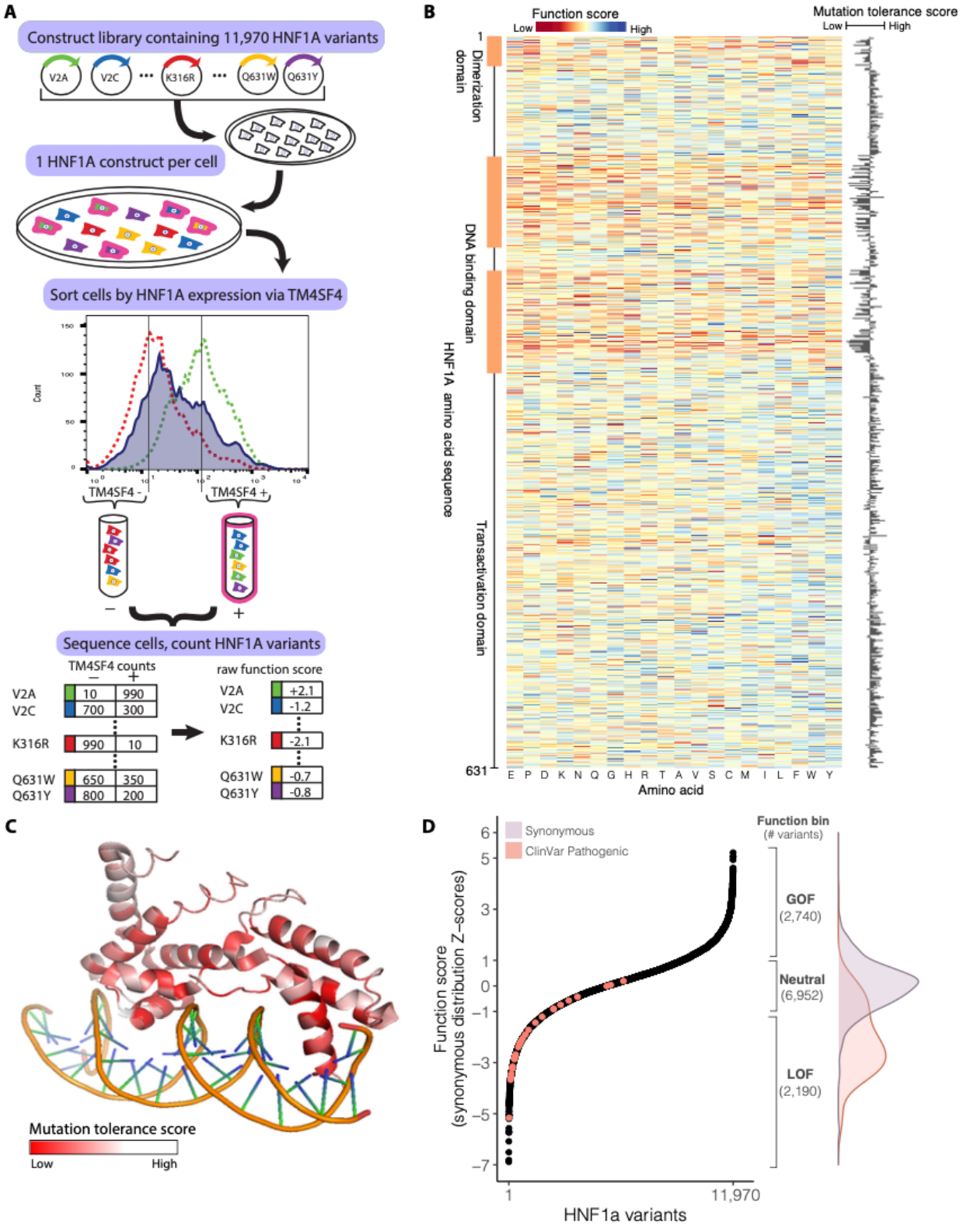
Results

Identification of gain-of-function substitutions from comprehensive functional testing of 11,970 HNF1A amino acid variants in human hepatocytes

To comprehensively assess the hepatic function of protein-coding variants in *HNF1A*, we synthesized a cDNA library comprised of all possible single amino acid (aa) permutations of HNF1A (630 aa protein x 19 aa changes = 11,970) such that each transgene encoded a single amino acid substitution. The library was introduced into a human hepatocyte cell line previously engineered to lack endogenous *HNF1A* and at a dilution maximizing the number of cells receiving only a single copy of *HNF1A* transgene. The *HNF1A* transgenes were induced for six days via a doxycycline-inducible promoter and cells were sorted via fluorescence-activated cell sorting (FACS) based on their protein expression level of TM4SF4, a cell surface protein shown to be a direct transcriptional target of HNF1A in pancreas and liver(*22*, *23*) (Figure 2.1A). The population of cells bearing *HNF1A* transgenes was partitioned into two bins separated by a ten-fold expression intensity difference in TM4SF4: $TM4SF4^{low}$ and $TM4SF4^{high}$ (Figure 2.1A). To recover and quantify the HNF1A variants in each TM4SF4 expression bin, *HNF1A* transgenes from the $TM4SF4^{low}$ and $TM4SF4^{high}$ cells were sequenced by targeted next-generation sequencing. Raw function scores were generated for each amino acid substitution at each site in HNF1A by determining the frequency of appearance of each variant in the $TM4SF4^{low}$ and $TM4SF4^{high}$ bins as we and others have previously done(*20*, *24*). Over 99% of the originally designed variants were recovered in sufficient quantity from the pooled screen to be assigned a function score.

Figure 2.1. Identification of gain-of-function variants from comprehensive functional testing of 11,970 HNF1A variants in human hepatocytes.

A) A library of 11,970 HNF1A constructs was synthesized, with each construct encoding a single amino acid substitution. The construct library was introduced into HUH7 hepatocytes (deleted for endogenous HNF1A) at a dilution of one construct per cell. The resulting polyclonal population of HUH7 hepatocytes was separated via FACS according to the expression of the known HNF1A transcriptional target TM4SF4 and sorted into low (-) and high (+) bins of HNF1A activity. Activity cutoffs were established through flow cytometry experiments of HNF1A KO cells (dashed red line) and WT cells (dashed green line). Each bin of cells was sequenced at the transgenic HNF1A locus to identify and tabulate the introduced variants. Each HNF1A variant was assigned a function score based on its abundance in the low and high TM4SF4 expression bins. B) Heatmap of 11,970 HNF1A variant function scores, arranged according to the primary amino acid sequence (rows). Function scores lower than WT are shaded red and function scores greater than WT are shaded blue. Function scores averaged (mean) at each amino acid position are plotted to the right, showing the level of tolerance for any amino acid substitution away from WT at each position. C) Mutation tolerance scores as described in B) overlaid on the crystalized protein structure of HNF1A DNA-binding domain (PF04814). Positions intolerant of amino acid changes (i.e. lower function scores) are shaded red. Helices that make direct contacts with the DNA are the most intolerant of mutations. D) (left panel) HNF1A function scores ranked for all 11,970 amino acid variants tested, and ClinVar annotated pathogenic variants (n=29) are highlighted in red. (right panel) Function bins correspond to variants with function scores above (GOF), within (neutral), or below (LOF) +/- 1 Z-score of the synonymous distribution, shown with the total number of variants per bin. Overlaid are the function score distributions of the 613 synonymous HNF1A variants tested (purple) and the 29 ClinVar pathogenic variants (red).


A
Construct library containing 11,970 HNF1A variants
V2A
V2C
K316R
Q631W
Q631Y
1 HNF1A construct per cell
Sort cells by HNF1A expression via TM4SF4
Count
TM4SF4 -
TM4SF4 +
Sequence cells, count HNF1A variants
TM4SF4 counts
raw function score
V2A 10 990 +2.1
V2C 700 300 -1.2
K316R 990 10 -2.1
Q631W 650 350 -0.7
Q631Y 800 200 -0.8
B
Function score
Low
High
Mutation tolerance score
Dimerization domain
DNA binding domain
Transactivation domain
HNF1A amino acid sequence
1
631
E P D K N Q G H R T A V S C M I L F W Y
Amino acid
C
Mutation tolerance score
Low
High
D
Synonymous
ClinVar Pathogenic
Function bin
(# variants)
GOF
(2,740)
Neutral
(6,952)
LOF
(2,190)
Function score
(synonymous distribution Z-scores)
6
5
3
1
0
−1
−3
−5
−7
1
11,970
HNF1a variants

The raw function scores for all protein-coding *HNF1A* variants were compared with the known sequence:function relationships of HNF1A. First, the function scores were averaged (mean) at each of the 631 amino acid positions along HNF1A to obtain a "mutation tolerance" score representing the effect on function of substituting any non-WT amino acid at that position; i.e. low "mutation tolerance" represents a functionally important residue (Figure 2.1B). Several clusters of amino acids along the primary sequence exhibiting low mutation tolerance scores were observed and colocalized with known domains critical for HNF1A molecular function including the dimerization domain and the DNA binding domain(*25*, *26*). The mutation tolerance scores were also overlaid onto the HNF1A DNA binding domain crystal structure to evaluate the relationship between function and higher-order structure (Figure 2.1C). Within the extent of the available crystal structure, the amino acid residues comprising the alpha helices in the DNA binding domain that directly contact the DNA double helix exhibit the lowest mutation tolerance scores, consistent with a high degree of functional conservation (Figure 2.1C).

Taking a conservative approach to minimize the effect of extreme values, the raw individual variant scores were added to the mutation tolerance score at that amino acid position (Figure 2.1B) and rescaled to the mean and standard deviation (SD) of the distribution of synonymous variants to provide an intuitive interpretation to the function scores (Figure 2.1D). Of the 11,970 variants tested, 2,190 fell into the putative LOF category (score < -1SD), 6,952 fell into the neutral category (-1SD <= score <= 1SD) and surprisingly 2,740 fell into a putative gain-of-function GOF (score > 1SD) category indicating that those amino acid substitutions increased the transcriptional activity of HNF1A over WT in our assay, a finding not previously reported in over two decades of

HNF1A functional variant characterization(*2*). As expected, variants categorized as putative LOF by our experiments occurred at codons that were more evolutionarily conserved in mammalian genomic sequences(*27*) than those classified as neutral ($p < 10^{-6}$ Wilcoxon rank-sum test). Variants categorized as putative GOF, however, could not be distinguished from neutral variants on the basis of evolutionary conservation suggesting that they would be poorly annotated by computational prediction algorithms leveraging conservation metrics(*28*, *29*). To further benchmark our experiments, we extracted the function scores of pathogenic human HNF1A variants from ClinVar(*30*), a curated database of clinically significant human genetic variants, (Figure 2.1D) finding that the majority, but not all (n=21/29) would be classified as LOF by our data. Thus, to explicitly quantify on a per variant basis the confidence of LOF/GOF, a posterior error probability(*31*) (PEP) was computed for every variant based on where its function score fell in the distribution of function scores for synonymous versus nonsynonymous variants (Figure 2.1D).

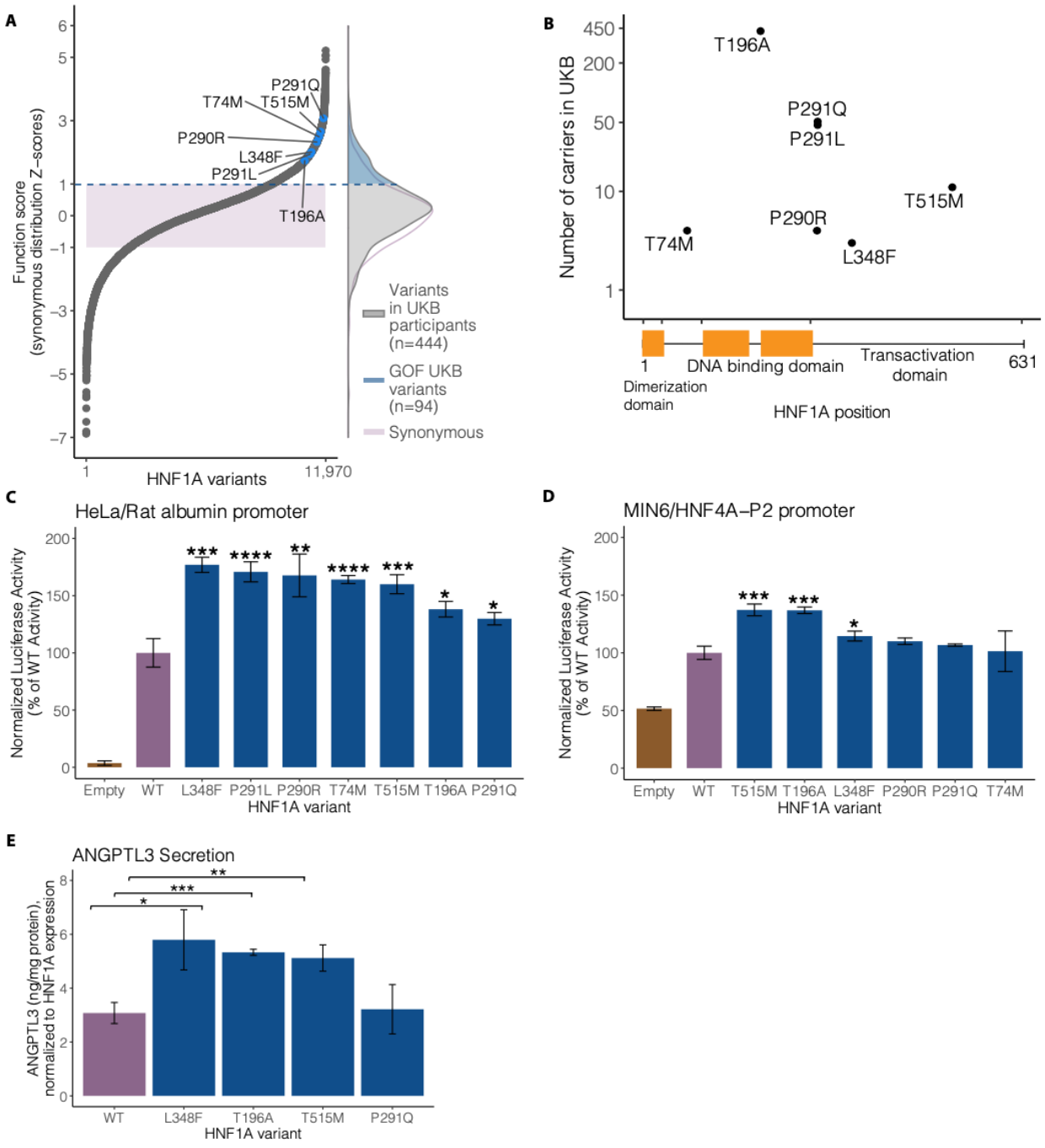
Gain-of-function *HNF1A* variants are carried in the general population and increase transcriptional activity in multiple cellular contexts.

To evaluate the human relevance of our deep mutational scan and assess if these putative GOF variants are present in people, we intersected our comprehensive dataset of HNF1A function scores with *HNF1A* protein-coding variants in the UK Biobank, a population-based cohort containing 454,756 exome sequenced individuals(*32*, *33*). From the exome sequences we extracted all *HNF1A* variants and identified 4,302 individuals who carried a protein-coding variant with minor allele frequency less than 0.001 (~1:100

individuals). Of the 444 unique protein-coding variants identified, 335 (75%) were absent from ClinVar(*30*). The distribution of HNF1A function scores in the UK Biobank peaked at the same level as the distribution of synonymous variants (Figure 2.2A), consistent with prior observations that most protein-coding variants are not functional(*34*). The distribution of function scores in the UK Biobank also notably lacked the lowest scoring pathogenic/MODY3 variants as would be expected for a generally healthy population-based cohort(*32*) and with the estimated UK prevalence of MODY3 being ~5 per 100,000(*35*). We did observe a 22% prevalence of diagnosed diabetes among carriers of HNF1A frameshifting mutations (n = 62 carriers) which exceeded the 8% prevalence of diabetes in the general UK Biobank cohort. The presence of LOF *HNF1A* variants in population-based cohorts has been described previously(*36*) but remarkably carriers of gain-of-function (GOF) variants have not previously been described in the many studies functionally characterizing human genetic variants in *HNF1A*.

Figure 2.2. Gain-of-function HNF1A variants are carried in the general population and increase transcriptional activity in multiple cellular contexts.

A) (left panel) Selected human HNF1A variants with function scores greater than WT highlighted in blue among the full distribution of all 11,970 amino acid variants tested. The shaded purple box represents +/- 1 Z-score of the distribution of 613 synonymous HNF1A variants tested. (right panel) The distributions of function scores of rare protein coding HNF1A variants (n=444) identified from 454,756 sequenced individuals in the UK Biobank (UKB) and of the 613 synonymous variants. An upper tail of function increasing variants in UK Biobank individuals is highlighted in blue (n=94 unique variants in 1,469 individuals). B) Location of top scoring variants selected for validation along the HNF1A protein primary amino acid sequence with number of human variant carriers identified in 454,756 exome sequences from the UK Biobank. C) Putative GOF variants were individually recreated and tested for transcriptional activation using luciferase reporters in HeLa cells (which lack endogenous HNF1A activity) on the rat albumin promoter. Activity measurements are shown as percentages of WT HNF1A activity +/- s.e.m; n=3. Basal promoter activity in cells is measured by transfection of vectors lacking the HNF1A transgene (empty). (* $p<0.05$, ** $p<0.01$, *** $p<0.001$, **** $p<0.0001$, linear mixed model) D) The same variants described in B) were tested for transcriptional activity in mouse insulinoma MIN6 cells on the HNF4A-P2 promoter. E) GOF HNF1A variant constructs were transfected into HNF1A deleted HUH7 hepatocytes, and ANGPTL3 levels were measured by ELISA 48 hours post confluency (n=3 per variant). Transfection with HNF1A p.L348F, p.T196A, and p.T515M increases ANGPTL3 secretion as compared to WT (*** $p<0.0005$,** $p<0.005$, * $p<0.05$, linear mixed model).


A
Function score (synonymous distribution Z−scores)
HNF1A variants
1
11,970
P291Q
T74M
T515M
P290R
L348F
P291L
T196A
Variants in UKB participants (n=444)
GOF UKB variants (n=94)
Synonymous
B
Number of carriers in UKB
T196A
P291Q
P291L
T515M
P290R
T74M
L348F
1
DNA binding domain
Transactivation domain
631
Dimerization domain
HNF1A position
C
HeLa/Rat albumin promoter
Normalized Luciferase Activity (% of WT Activity)
Empty
WT
L348F
P291L
P290R
T74M
T515M
T196A
P291Q
HNF1A variant
D
MIN6/HNF4A–P2 promoter
Normalized Luciferase Activity (% of WT Activity)
Empty
WT
T515M
T196A
L348F
P290R
P291Q
T74M
HNF1A variant
E
ANGPTL3 Secretion
ANGPTL3 (ng/mg protein), normalized to HNF1A expression
WT
L348F
T196A
T515M
P291Q
HNF1A variant

To confirm and further characterize these putative GOF *HNF1A* variants carried by humans in the general population, we selected a series of high-scoring GOF variants that spanned across the protein domains of HNF1A (Figure 2.2A). These seven variants ranged in frequency from 3 carriers (p.L348F) to 421 carriers (p.T196A) in the 454,756 sampled exomes (Figure 2.2B). Each variant was recreated, transfected, and assessed for transcriptional activity using luciferase reporter assays as we(*37*) and others(*38*) have previously performed in two separate experimental contexts: 1) HeLa cells using the rat albumin promoter (Figure 2.2C) and 2) MIN6 insulinoma cells using the HNF4A-P2 promoter (Figure 2.2D). In HeLa cells, all selected variants showed significant increases in transcriptional activity over WT (ranging from 30-70%) whereas in MIN6 cells the measured increase in transactivation was attenuated with only three of the tested variants exceeding the nominal threshold of significance. Notably, the background activity of HNF1A in Hela cells was nil, whereas in MIN6 cells it approached fifty percent of WT. Given that GOF was observed in two independent HNF1A-free cellular contexts (i.e. HeLa and HUH7 *HNF1A*-null cells) the attenuation in signal in the MIN6 cell system likely reflected decreased sensitivity in measuring increased HNF1A transcription in the setting of pre-existing high levels of endogenous HNF1A activity.

To assess the functional consequence of GOF *HNF1A* in hepatocytes we developed a secretion assay for ANGPTL3, an emerging cardiovascular disease risk factor(*39*) that encodes a hepatically secreted regulator of serum lipids(*40*) whose gene is bound by HNF1A(*41*). After confirming that HNF1A functionally regulates production of ANGPTL3 from cultured hepatocytes, we evaluated ANGPTL3 secretion in *HNF1A* null hepatocytes and performed reconstitution experiments with WT and the identified GOF

HNF1A variants. The p.L348F, p.T196A, and p.T515M variants but not the p.P291Q increased ANGPTL3 secretion when transfected into *HNF1A*-null HUH7 hepatocytes compared to WT (Figure 2.2E). Taken together, these cell culture experiments demonstrate that GOF HNF1A variants increase transcriptional activity in both pancreatic and hepatic cellular contexts.

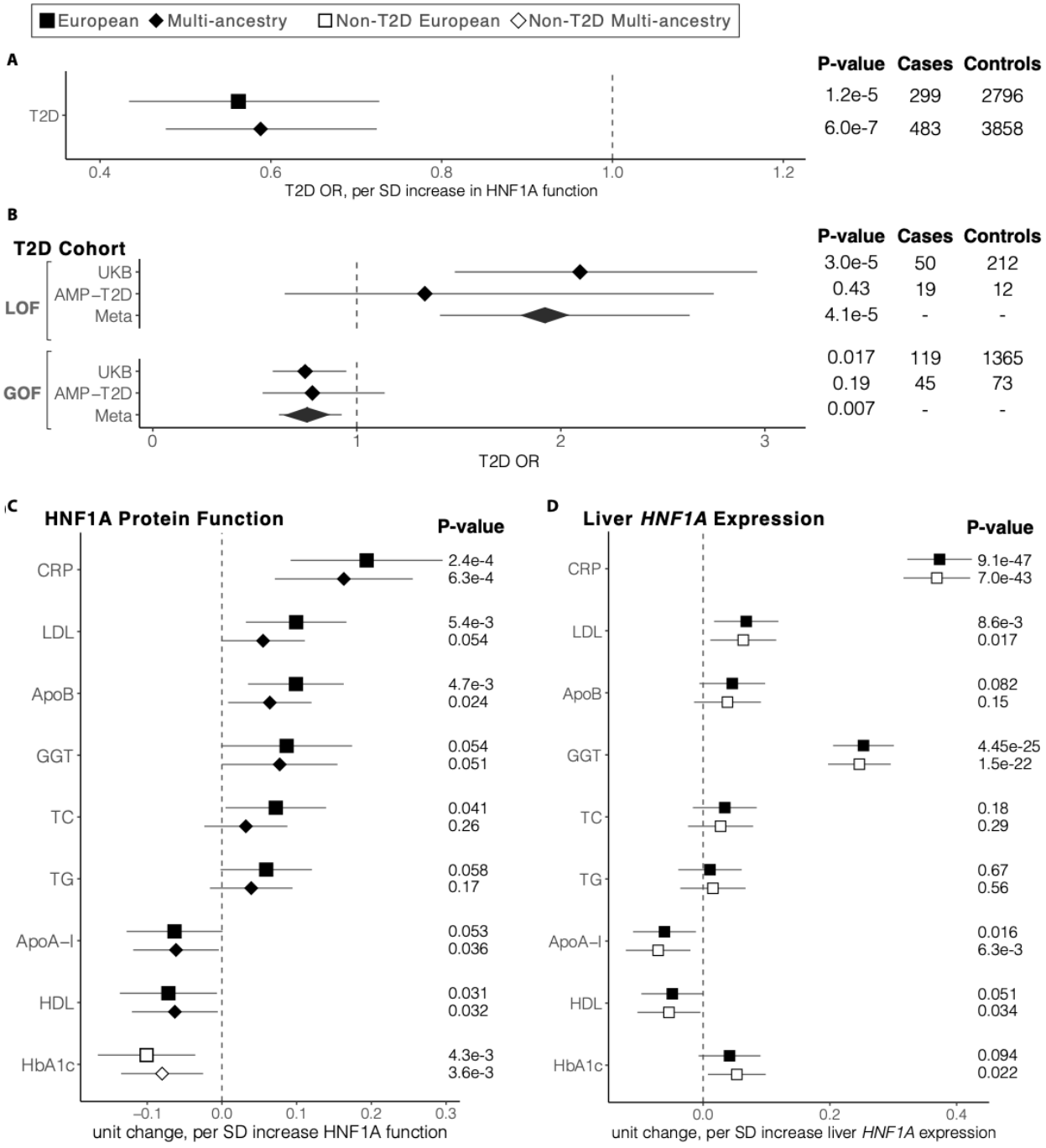
HNF1A GOF variant carriers in the population are protected from T2D but show elevated levels of serum atherogenic risk factors

We examined the relationship between HNF1A function and T2D in carriers of rare protein-coding *HNF1A* variants ascertained from the UK Biobank and found an inverse association between HNF1A function and T2D risk in the full multi-ancestry cohort (OR=0.59 per SD increase in HNF1A function, 95% CI=[0.48-0.72], $p=6.0 \times 10^{-7}$, n=4,341, logistic mixed effects regression) (Figure 2.3A). A similar association was obtained when restricting to European ancestry *HNF1A* rare variant carriers (OR=0.56 per SD increase in HNF1A function, 95% CI=[0.43-0.73], $p=1.2 \times 10^{-5}$, n=3,095, logistic mixed effects regression). These analyses included the 421 carriers of the p.T196A variant (Figure 2.2B). Given the previously known associations between LOF variants in *HNF1A* and diabetes(*2*, *36*, *42*), we performed a categorical analysis binning the *HNF1A* variants by function (GOF, neutral, LOF) and separately examined their association with T2D risk in the UK Biobank and in an independent, multi-ancestry T2D case:control cohort (AMP-T2D(*43*); 20,791 cases, 24,440 controls) to disentangle the effect of GOF variants from the known relationship of LOF variants and increased risk of T2D (Figure 2.3B). In the meta-analysis, LOF variant carriers compared to neutral variant carriers had an increased

risk of T2D as expected from prior studies (OR=1.9, 95% CI=[1.41-2.63], p=4.1 x 10^{-5}, logistic regression). Remarkably, GOF variant carriers were protected from T2D (UK Biobank GOF carriers: n=1,469, OR=0.75, 95% CI=[0.59-0.95], p=0.017, logistic regression), a signal that was strengthened when analyzed in combination with the AMP-T2D cohort (OR=0.76, 95% CI=[0.62-0.93], p=0.007, logistic regression; Figure 2.3B) evidentiating that genetically determined increases in HNF1A function have clinical consequences in humans. The association signal for GOF variants conferring protection from T2D was consistent over progressively increasing function score thresholds although at the most stringent thresholds we observed a loss of statistical power due to steep decrease in number of GOF variant carriers. Furthermore, correction for independent common protein-coding HNF1A variants I27L (rs1169288) and A98V (rs1800574) which have been associated with increased risk of T2D(*44*) had minimal impact on these associations.

Figure 2.3. HNF1A GOF variant carriers in the population are protected from T2D but show elevated levels of serum atherogenic risk factors.

Shape and fill indicate the ancestry and T2D status respectively of individuals included in each analysis. All phenotypes are adjusted for age, age^2, sex, and the first 10 principal components of ancestry. T2D, type 2 diabetes; SD, standard deviation; LOF, loss-of-function; GOF, gain-of-function; CRP, C-reactive protein; LDL, low-density lipoprotein cholesterol; ApoB, apolipoprotein B; GGT, gamma-glutamyl transferase; TC, total cholesterol; TG, triglycerides; ApoA-I, apolipoprotein A-I; HDL, high-density lipoprotein cholesterol; HbA1c, glycated hemoglobin. A) Association of HNF1A function score and T2D risk using logistic regression in the UK Biobank HNF1A rare protein-coding variant carriers identified among the 454,756 exome sequenced individuals. Odds ratios and the 95% confidence interval are shown. B) Association of LOF and GOF HNF1A variants with T2D in the UK Biobank and AMP-T2D (n=20,791 cases, 24,440 controls) using logistic regression. Odds ratios and the 95% confidence interval are shown, and fixed effects inverse-variance meta-analysis was used to combine the results. C) Association of functional HNF1A variants with cardiovascular disease risk factors in the UK Biobank rare variant carriers (European n=3,089, Multi-ancestry n=4,302, linear mixed effects regression). Effect size and the 95% confidence interval are shown for each phenotype. D) Association of liver specific predicted HNF1A gene expression with disease risk factors tested in C) among all UK Biobank European genotyped individuals (n=407,227, filled squares) and non-T2D European genotyped individuals (n=376,043, open squares).


European
Multi-ancestry
Non-T2D European
Non-T2D Multi-ancestry
A
T2D
0.4
0.6
0.8
1.0
1.2
T2D OR, per SD increase in HNF1A function
P-value Cases Controls
1.2e-5 299 2796
6.0e-7 483 3858
B
T2D Cohort
LOF
UKB
AMP–T2D
Meta
GOF
UKB
AMP–T2D
Meta
0
1
2
3
T2D OR
P-value Cases Controls
3.0e-5 50 212
0.43 19 12
4.1e-5 - -
0.017 119 1365
0.19 45 73
0.007 - -
C
HNF1A Protein Function
P-value
CRP 2.4e-4 6.3e-4
LDL 5.4e-3 0.054
ApoB 4.7e-3 0.024
GGT 0.054 0.051
TC 0.041 0.26
TG 0.058 0.17
ApoA–I 0.053 0.036
HDL 0.031 0.032
HbA1c 4.3e-3 3.6e-3
–0.1
0.0
0.1
0.2
0.3
unit change, per SD increase HNF1A function
D
Liver HNF1A Expression
P-value
CRP 9.1e-47 7.0e-43
LDL 8.6e-3 0.017
ApoB 0.082 0.15
GGT 4.45e-25 1.5e-22
TC 0.18 0.29
TG 0.67 0.56
ApoA–I 0.016 6.3e-3
HDL 0.051 0.034
HbA1c 0.094 0.022
0.0
0.2
0.4
unit change, per SD increase liver HNF1A expression

Given that individuals who carry pathogenic/MODY3 variants demonstrate alterations in hepatically secreted factors including hsCRP and lipoprotein profiles, we quantitatively examined the available biomarkers including hsCRP(*17*), lipoprotein levels(*45*), liver enzymes(*46*), and glycemic traits(*47*) in *HNF1A* rare variant carriers in the UK Biobank by regressing the normalized values for each measurement against the HNF1A function scores of the variant carriers (Figure 2.3C). A positive relationship was identified between HNF1A function and serum hsCRP, gamma glutamyltransferase (GGT), serum low-density lipoprotein (LDL) cholesterol, and apolipoprotein B (ApoB), while an inverse relationship was identified between HNF1A function and serum high-density lipoprotein (HDL) cholesterol and apolipoprotein A-I (ApoA-I). After excluding carriers with diagnosed diabetes, we examined the relationship between HNF1A function and glycemia, finding an inverse relationship between HNF1A function and glycosylated hemoglobin (HbA1c; Figure 2.3C). These findings were robust to correction for multiple hypothesis testing and suggest that genetic variants which increase HNF1A protein function increase serum hsCRP, GGT, and promote an atherogenic lipoprotein profile by raising serum ApoB/LDL cholesterol levels and lowering ApoA-I/HDL cholesterol. However, in concordance with conferring protection from T2D, HNF1A function increasing variants also decrease blood glucose in non-diabetic individuals.

In light of these observed pleiotropic and opposing effects of GOF in HNF1A on T2D and inflammatory/atherogenic factors, we examined the relationship between GOF HNF1A variants and CAD events in the UK Biobank and a cohort of individuals exome sequenced for early-onset myocardial infarction (MIGen(*48*); 24,337 cases, 28,922 controls). We observed no decrease in CAD risk in GOF *HNF1A* variant carriers

(OR=1.15, 95% CI=[0.93-1.43], p=0.21, logistic regression) suggesting that either the observed increase in hepatic pro-atherogenic factors is counterbalanced by the protection from T2D, or that the analysis lacked statistical power due to the limited number of CAD cases available. Notably, our combined CAD cohort analysis had over 90% power to detect a decrease in risk of similar magnitude (~25%) as observed with T2D and GOF variant carriers at the nominal type 1 error rate ($\alpha = 0.05$), but a smaller effect size may not have risen to the level of statistical significance.

To evaluate if these findings were the result of perturbed HNF1A activity in the liver we turned to liver-specific HNF1A gene expression reasoning that increasing the levels of intra-cellular HNF1A would produce a similar clinical effect as increasing intrinsic HNF1A protein function. To quantify the heritable component of HNF1A gene expression we selected common, non-coding SNPs identified in European liver samples from the Genotype-Tissue Expression (GTEx) Project(*49*, *50*) as expression quantitative trait loci (eQTLs) for HNF1A and combined them to produce a liver HNF1A gene expression score. After removing the carriers of rare protein-coding *HNF1A* variants shown in Figures 2.3A-C, we assigned liver-specific *HNF1A* expression scores to the remaining 407,227 European individuals using established methods (Functional Summary-based Imputation: FUSION)(*51*). These liver-specific *HNF1A* expression scores were then regressed against the same clinical phenotypes as the rare, protein-coding variant analysis (Figure 2.3C). This method confirmed that increased HNF1A gene expression encoded by liver eQTLs conferred increased levels of hsCRP ($p < 10^{-46}$, linear model; Figure 2.3D) as would be expected given the known biology of CRP being synthesized and secreted by hepatocytes(*52*). Similarly, serum GGT and LDL cholesterol were associated with

increased liver-specific *HNF1A* expression scores, whereas HDL and ApoA-I showed a nominal inverse association. Notably, glycemia measured by HbA1c ($p = 0.09$, linear model) did not associate with liver-specific HNF1A expression as would be expected from the known pancreatic role of HNF1A in regulating blood glucose(*53*).

To evaluate if these associations were being driven by or distinct from T2D, we repeated the above analyses excluding all 41,923 individuals (including 539 carriers of rare-protein coding variants) with diabetes from the UK Biobank population and recomputed the regression (Figure 2.3C and 2.3D) of clinical measurements against HNF1A function/expression. With regard to liver-specific HNF1A gene expression ($n = 376,043$ samples without T2D) the analyses were well powered and showed similar effect size, directionality, and significance of all quantitative trait associations suggesting a causal relationship with HNF1A that is independent of T2D (Figure 2.3D). Effect sizes in the rare, protein-coding variant analysis also remained directionally consistent, but several associations no longer passed the nominal threshold of statistical significance likely due to a loss of power ($n = 3,802$ rare-protein coding variant carriers without T2D). In summary, our data indicate that increased hepatic function/expression of HNF1A increases serum hsCRP, GGT, LDL while decreasing HDL/ApoA-I independently of its effects on T2D, and increased HNF1A function decreases blood glucose, but not through the liver.

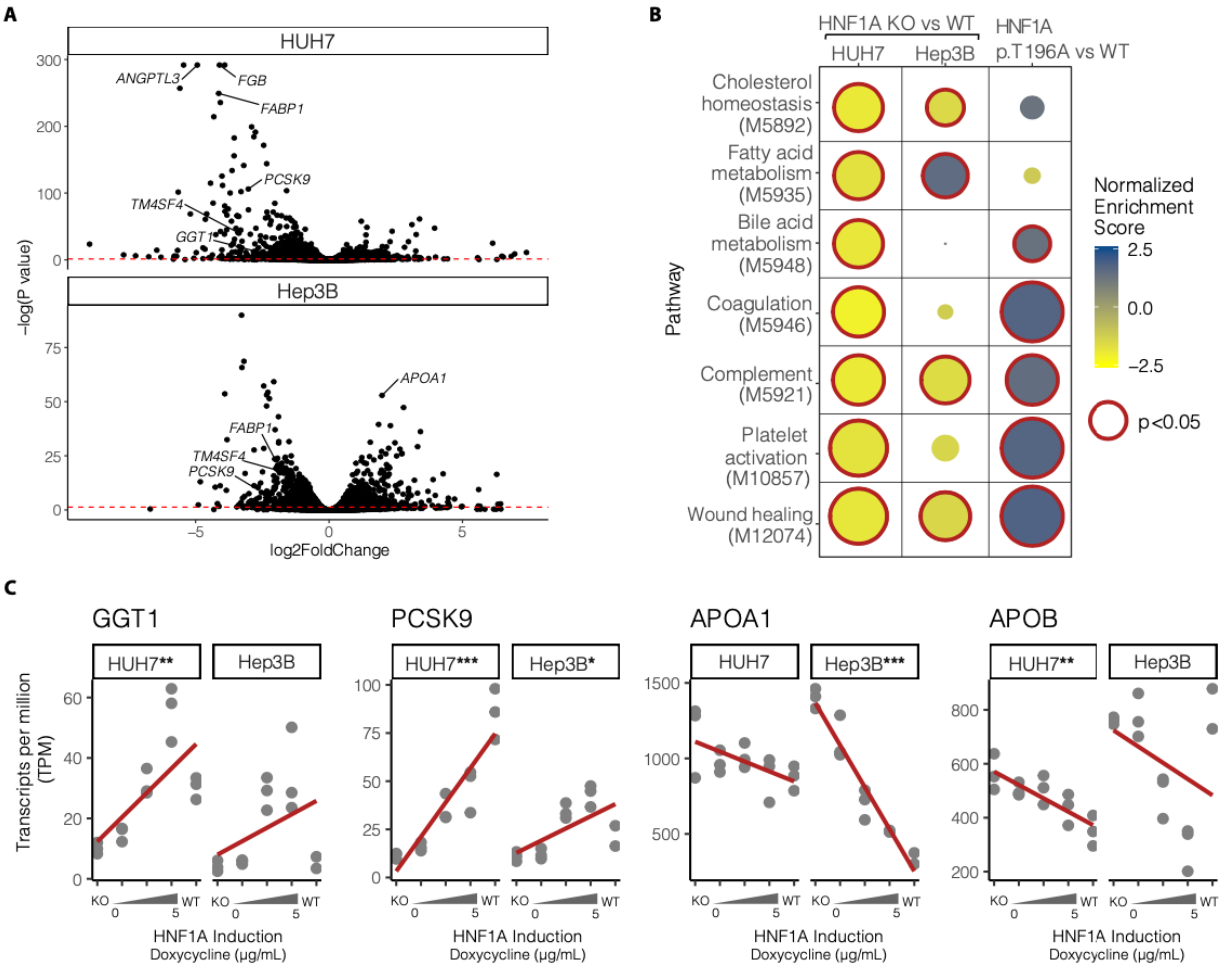
<u>HNF1A transcriptionally regulates complement activation, coagulation, and pro-atherogenic gene expression programs in hepatocytes</u>

To understand the molecular mechanisms by which functional genetic variants in

HNF1A regulate serum lipids, we deleted *HNF1A* in human hepatocytes using CRISPR/Cas9 and transcriptionally profiled the resulting cells using mRNA sequencing (Figure 2.4A). Experiments were carried out in both HUH7 and Hep3B cell lines to mitigate confounding cell line artifacts. Many of the top differentially expressed genes are known transcriptional targets of HNF1A including *FABP1(41)*, *FGB(54)*, and *TM4SF4(23)*. To examine which molecular pathways were perturbed by the loss of *HNF1A*, we performed gene set enrichment analysis (GSEA)(*55*, *56*) on the full list of differentially expressed genes in WT and *HNF1A* KO cells (Figure 2.4B). The GSEA analysis showed that *HNF1A*-null cells downregulated multiple lipid metabolism pathways including “fatty acid metabolism” (MsigDB #M5935), “cholesterol homeostasis” (MsigDB #M5892), and “bile acid metabolism” (MsigDB #5948), consistent with previous findings from murine *Hnf1a* LOF models(*9*). Additionally, pathways for “coagulation” (MsigDB #M5946) and “complement” (MsigDB #M5921) were also significantly downregulated in the absence of HNF1A and a more refined examination of sub-pathways underlying these further showed a downregulation in “platelet activation” (MsigDB #M10857) and “wound healing” (MsigDB #M12074).

Figure 2.4. HNF1A transcriptionally regulates complement activation, coagulation, and pro-atherogenic gene expression programs in hepatocytes.

A) Differentially expressed genes between wild-type and HNF1A knockout (KO) HUH7 and Hep3B cell lines quantified by mRNA sequencing (n=3 per group). The dotted red line indicates an adjusted p-value threshold of 0.05, with genes falling above the line meeting a global significance threshold for differential expression. B) Gene set enrichment analysis (GSEA) was performed on mRNA sequencing data shown in A) to identify pathways altered by deletion of HNF1A (HNF1A KO vs WT). Significant genesets in at least one cell line are shown. The effect of gain-of-function (GOF) in HNF1A on these pathways was examined by complementing HNF1A KO HUH7 cells with WT HNF1A and the GOF p.T196A variant (n=3 per group) followed by mRNA sequencing, differential expression, and GSEA (HNF1A p.T196A vs WT). Downregulated pathways (negative normalized enrichment scores) are shaded yellow and upregulated pathways are shaded blue. Size represents the -log10(P-value), and significantly enriched pathways ($p<0.05$) are outlined in red. C) HNF1A was reintroduced via doxycycline inducible transgenes into HNF1A KO hepatocyte cell lines and global gene expression was measured (TPM = transcripts per million, n=3 per group). In a dose dependent fashion, HNF1A regulates genes involved in oxidative stress and lipid regulation pathways. Significance levels from regressing TPM on HNF1A levels are shown for each gene and cell type if significant (*** $p < 10^{-6}$, ** $p < 10^{-3}$, * $p<0.05$, linear model). (GGT1, Gamma Glutamyltransferase 1; PCSK9, Proprotein Convertase Subtilisin/Kexin Type 9; APOA1, Apolipoprotein A1; APOB, Apolipoprotein B)


A
HUH7
Hep3B
ANGPTL3
FGB
FABP1
PCSK9
TM4SF4
GGT1
APOA1
−log(P value)
log2FoldChange
B
HNF1A KO vs WT
HUH7
Hep3B
HNF1A p.T196A vs WT
Cholesterol homeostasis (M5892)
Fatty acid metabolism (M5935)
Bile acid metabolism (M5948)
Coagulation (M5946)
Complement (M5921)
Platelet activation (M10857)
Wound healing (M12074)
Pathway
Normalized Enrichment Score
2.5
0.0
−2.5
p<0.05
C
GGT1
PCSK9
APOA1
APOB
HUH7**
Hep3B
HUH7***
Hep3B*
HUH7
Hep3B***
HUH7**
Hep3B
Transcripts per million (TPM)
KO
WT
HNF1A Induction
Doxycycline (µg/mL)

To evaluate global transcriptional changes from *HNF1A* GOF variants, we assessed the global transcriptional effects of expressing HNF1A WT and the most commonly found human GOF variant HNF1A p.T196A (Figure 2.2B). We performed short-term complementation experiments, expressing HNF1A WT and p.T196A in *HNF1A*-null HUH7 cells for 18 hours from in vitro-transcribed mRNA at levels similar to endogenous HNF1A in WT HUH7 cells. GSEA analysis of genes differentially expressed between the HNF1A WT- and p.T196A variant-complemented cells showed that relative to WT, the p.T196A variant more strongly upregulated several of the lipid metabolism and coagulation/complement activation pathways that were downregulated in *HNF1A* KO cells (Figure 2.4B), supporting a molecular mechanism of GOF that spans multiple transcriptional pathways.

To validate specific effector genes that could explain the observed inflammatory, oxidative stress, and atherogenic serum lipoprotein gene expression profiles in *HNF1A* variant carriers, we performed independent experiments evaluating HNF1A:effector gene dose responsiveness. Doxycycline-inducible *HNF1A* transgenes were introduced into *HNF1A*-null cells and exposed to varying doses of doxycycline to evaluate HNF1A:gene dose-response curves, quantified by regressing gene transcripts per million (TPM) onto HNF1A dose (Figure 2.4C). The most straightforward example was for serum GGT, a marker of oxidative stress(*57*) which is hepatically secreted, positively regulated by HNF1A (*GGT1* $p < 10^{-3}$, linear model, HUH7; Figure 2.4C), and whose serum values are positively correlated with genetically mediated increases in HNF1A function/liver-specific gene expression (Figure 2.3C and 2.3D). Effector genes that could account for the serum lipoprotein profile associated with GOF in HNF1A and showed dose response included

the LDL receptor recycling regulator Proprotein Convertase Subtilisin/Kexin Type 9(*58*) (*PCSK9* $p < 10^{-6}$, linear model, HUH7) and apolipoprotein A1, the major lipoprotein component of HDL (*APOA1* $p < 10^{-6}$, linear model, Hep3B). Conversely, a negative HNF1A:gene expression dose response was observed for apolipoprotein B, the major lipoprotein component of LDL (*APOB* $p < 10^{-3}$, linear model; HUH7) suggesting the positive correlation observed between serum ApoB levels with genetically mediated increases in HNF1A function/expression (Figure 2.3C and 2.3D) is not a hepatocyte cell autonomous effect of HNF1A.

To corroborate direct transcriptional regulation of these genes by HNF1A, we identified HNF1A consensus binding motifs within a 5 kilobase genomic window of the transcription start sites (TSS) and overlaid these with HNF1A ChIP-seq data from HepG2 cells made available through the ENCODE project(*59*). The *PCSK9 and GGT1* loci showed ChIP-seq peaks coincident with high-scoring HNF1A consensus motifs near the TSS strongly supporting direct transcriptional regulation. The *APOA1* and *APOB* loci showed more modest evidence of direct regulation having only ChIP-seq peaks near the TSS but lacking HNF1A consensus motifs.

Discussion

In this study we combine massively parallel functional characterization and saturation mutagenesis with genetic and clinical data from almost 600,000 individuals to evaluate a full allelic series of protein-coding variants in *HNF1A* and their impact on human health. By correlating variant, cellular function, and clinical phenotype in thousands of individuals, we find that the general human population contains carriers of

gain-of-function (GOF) variants that increase the transcriptional activity of HNF1A by 30-50%. These GOF variants have both beneficial and pathological clinical consequences in the people who carry them through pleiotropic actions in the pancreas and liver respectively. While GOF variants confer a ~30% decrease in T2D risk they, likely through liver-specific transcriptional effects, promote an atherogenic lipid profile (increased serum ApoB/LDL and decreased ApoA-I/HDL). Lastly, we demonstrate that the hepatic transcriptional regulation of key effector genes by HNF1A, including *APOA1*, *ANGPTL3* and *PCSK9*, mediates at least part of the proatherogenic lipid profile observed in GOF variant carriers.

Here we present a description of function increasing genetic variants in *HNF1A* and their presence in the human population, an unexpected finding given that thousands of individuals have undergone *HNF1A* gene sequencing and dozens of protein-coding variants have been functionally characterized over the past two decades(*60*, *61*). We note that in contrast to LOF variants, GOF HNF1A variants showed no evidence of being under evolutionary constraint and thus would escape discovery by computational prediction programs(*28*, *29*) which rely on evolutionary conservation, highlighting the importance of taking an experimental approach to pathogenicity prediction at clinically important genes. In retrospect, some variants we have demonstrated as GOF have been previously identified and experimentally tested showing transactivation signals greater than WT(*38*), but due to inconsistent results across different experiments or labs were labeled WT-like. For example, the p.T196A variant showed activity greater than WT in transformed COS-7 monkey fibroblasts but not in mouse insulinoma cells(*38*) whereas in a recent study(*62*) p.P291Q and p.T196A showed increased transcriptional activity over WT in both HeLa

and rat insulinoma cells but at only one of two expert labs testing the same variants. In our own experiments we were able to most clearly resolve variants with function greater than WT in cell systems without endogenous HNF1A (HUH7 KO and Hela cells; Figure 2.2C) using the standard transactivation assays, but the signal was more challenging to identify in MIN6 cells with a high level of basal HNF1A activity (Figure 2.2D). The variable function of GOF variants in MIN6 cells could also be attributed to differences in HNF1A function in human hepatocytes (in which GOF was called) versus the mouse beta cells from which MIN6 are derived. Moreover, such GOF variants, by virtue of conferring protection from diabetes, were less likely to be identified in previous studies that selected diabetic individuals for sequencing. By functionally characterizing all possible missense variants in a single experiment, our study revealed the full spectrum of possible functional variation including both loss and gain. We further benefited from ascertainment of *HNF1A* variants in the UK Biobank, a general population-based cohort not enriched for diabetes and therefore not depleted of GOF *HNF1A* variant carriers. In the context of another recent study that identified GOF variant carriers in MC4R in the UK Biobank(*63*), our study highlights the potential for discovering novel genetic mechanisms of disease by conducting large scale functional variant characterization and clinical correlation in population-based cohorts.

Through the identification of a full allelic series of HNF1A variants, our study advances previously reported genetic associations between HNF1A, clinical biomarkers and T2D(*42*) by providing tissue-specific, mechanistic insights and unexpected directional relationships. In the general population our data support that GOF in HNF1A reduces T2D risk and glycemia through enhanced beta cell function which complements the

mechanisms by which LOF HNF1A variants are known to cause T2D/MODY3(*3*, *36*). In the liver, however, we find that GOF in HNF1A is metabolically pathogenic causing increases in hepatically derived lipoprotein risk factors which have been shown causal for CAD as well as increases in correlative if not directly causal factors such as hsCRP and GGT(*17*, *46*, *64–66*). Despite a significant decrease in diabetes risk in GOF variant carriers, we found no signal for protection from CAD risk in GOF variant carriers and propose that the increase in hepatic pro-atherogenic factors is counterbalanced by the decrease in T2D risk (Figure 2.5). Furthermore, the identified genetic associations between GOF in HNF1A, T2D and atherogenic lipoproteins remained robust upon correction for common HNF1A protein-coding variants as well as the heritable component of hepatic HNF1A gene expression. An important medical implication of these findings is that HNF1A 'replacement' gene therapy for HNF1A MODY patients, which is technically feasible through mRNA therapeutics(*67*), should be approached with caution as hepatic increases in HNF1A would be clinically undesirable.

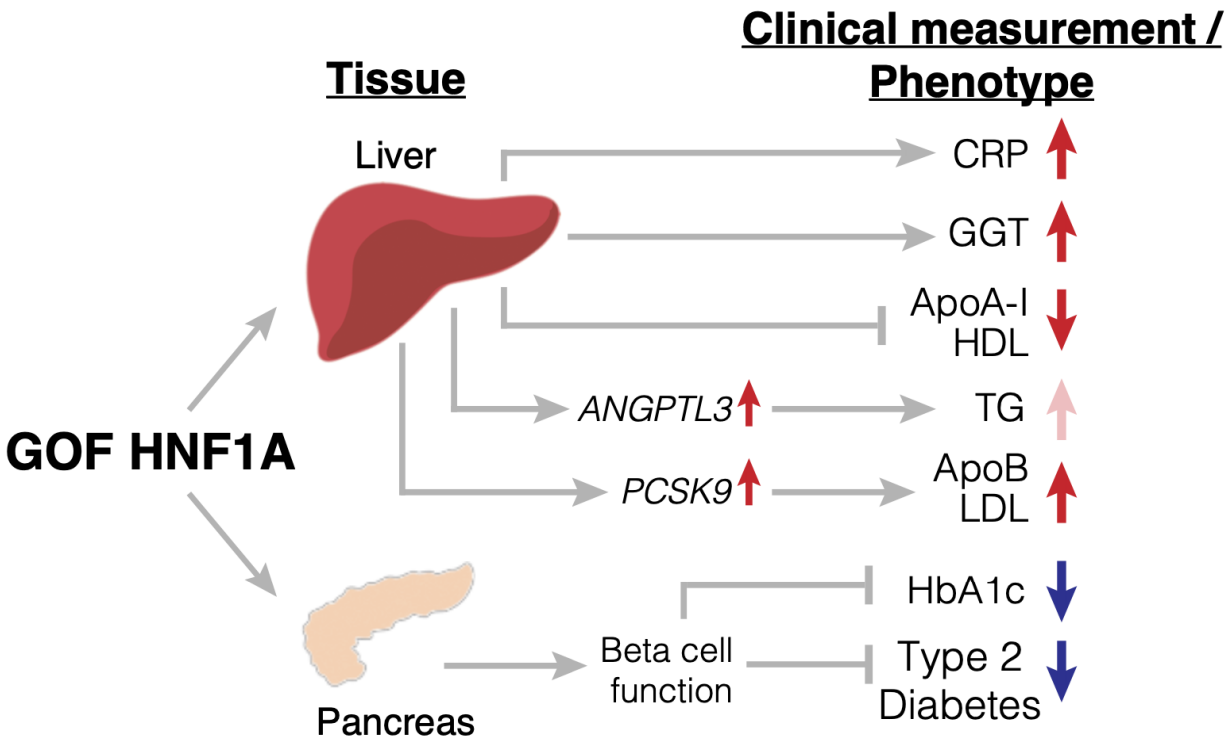


Figure 2.5. Graphical summary of proposed mechanism through which gain-of-function in HNF1A decreases diabetes risk through pancreas-specific mechanisms but increases the hepatic secretion of atherogenic risk factors.

GOF, gain-of-function; CRP, C-reactive protein; GGT, gamma-glutamyl transferase; ApoA-I, apolipoprotein A-I; HDL, high-density lipoprotein cholesterol; ANGPTL3, angiopoietin like 3; TG, triglycerides; PCSK9, proprotein convertase subtilisin/kexin type 9; ApoB, apolipoprotein B; LDL, low-density lipoprotein cholesterol; HbA1c, glycated hemoglobin.

Carriers of GOF variants in HNF1A have reduced T2D risk and HbA1c through enhanced beta cell function but have increases in CRP, GGT, and present a pro-atherogenic profile (suppression of ApoA-I/HDL levels and increased hepatic secretion of ANGPTL3 and ApoB/LDL levels through increased transcription of PCSK9). Red arrows indicate metabolically harmful outcomes, and blue arrows represent metabolically beneficial outcomes.

On the other hand, liver specific inhibition of HNF1A could be therapeutically attractive. At the molecular level, our data suggest that the suppression of ApoA-I/HDL levels by HNF1A is likely cell-autonomous as it occurred in cultured hepatocytes. Similarly, we find that HNF1A and GOF causing variants increase hepatocyte secretion of ANGPTL3 (Figure 2.2E), an emerging cardiovascular risk factor that inhibits endothelial lipase resulting in decreased clearance of triglyceride rich lipoproteins(*68*). Although we did not find significantly increased serum triglyceride levels in HNF1A GOF variant carriers (Figure 2.3C), ANGPTL3 has been proposed to promote atherosclerosis independent of triglycerides by increasing inflammation, angiogenesis, inhibiting reverse cholesterol transport(*69*). Additionally, we propose that the increase in serum LDL cholesterol in carriers of rare HNF1A GOF variants is at least partially mediated by a direct transcriptional activation by HNF1A of PCSK9 (Figures 2.4 and 2.5), a hepatic protease which regulates hepatic uptake of LDL(*70*, *71*) and is a pharmacological target for lipid lowering therapies(*72*–*74*). The lack of an HNF1A motif at the *APOB* promoter as well as the lack of positive HNF1A-*APOB* dose:response (Figure 2.4C) in cultured hepatocytes suggests the increase in serum ApoB levels from genetically mediated increase in HNF1A function/expression is indirect and is driven predominantly by increased serum LDL persistence from reduced LDLR recycling that would be caused by increasing PCSK9 transcription(*75*) (Figure 2.5). Concordantly, prior studies have found that liver specific HNF1A inhibition decreases PCSK9 levels, increases LDLR persistence and results in decreased serum total and LDL cholesterol(*76*, *77*). Intriguingly a recent phase 1 trial of a non-statin lipid lowering compound was demonstrated to prevent the dimerization and transactivation of HNF1A at its transcriptional targets ANGPTL3 and

PCSK9 resulting in lowered serum triglycerides and LDL cholesterol respectively(*78*). Our study predicts that GOF variant carriers in *HNF1A* would be highly responsive to this novel lipid lowering compound. Furthermore by virtue of having increased PCSK9 expression, HNF1A GOF carriers would also define a subpopulation that may respond to PCSK9 inhibitors more favorably for LDL reduction than conventional statin medications which are known to increase levels of PCSK9(*79*).

In summary, we find that ~1:300 individuals in the general population carry GOF variants in *HNF1A* which are not identifiable by computational prediction tools and pleiotropically decrease diabetes risk while increasing hepatic secretion of atherogenic lipids. In an era of rapidly deployable gene therapy(*80*), understanding the clinical consequences of increasing or decreasing gene function in humans bottlenecks therapeutic development. As biobanks continue to grow in size and phenotypic complexity, this study demonstrates the utility of comprehensive functional characterization to distill novel mechanisms of disease and identify adverse effects prior to undertaking drug development and putting clinical trial participants at risk.

Limitations of the Study

Limitations of our study include the precision of HNF1A function scores generated in our massively parallel assay and the cell models utilized to elucidate molecular mechanisms. Specifically, the use of TM4SF4 as the readout for HNF1A function would not capture the complexity of HNF1A transactivation functions across all of its targets. While our observation of a class of amino acid substitutions that increase the transcriptional activity of HNF1A (Figure 2.1D and 2.2A) is strongly supported by

experimental validation in multiple independent experiments (Figure 2.2C, 2.2D, 2.2E, and 2.4B), the precise level of increase in function conferred by any particular variants contains biological (different cellular contexts) and experimental (performed in different labs) uncertainty that limit their interpretation. For example p.P291Q was the top scoring GOF variant (Figure 2.2A), replicated in HeLa cells (Figure 2.2C) but did not show increase over WT in MIN6 cells (Figure 2.2D) or ANGPTL3 secretion (Figure 2.2E) assays. Conversely, p.T196A was the lowest scoring of the selected GOF variants but showed higher function than other variants in subsequent assays. Thus, we would caution against utilizing these function scores directly to provide an individualized risk estimate to a person carrying a protein-coding *HNF1A* variant. If used as a standalone classifier for GOF/LOF, our function scores would meet the 90% probability standard for clinical variant interpretation(*81*) for only 43 variants. Additionally, due to the design of our deep mutational scan, effects of splice variants are not detectable and functional interpretation of such variants should be approached with caution. Furthermore, our mechanistic evaluation in multiple human hepatocyte cell lines could miss important regulatory mechanisms due to the limitations of *in vitro* cell culture models. For example, our cellular data did not identify direct regulation of CRP by HNF1A as it is poorly expressed and lacks evidence for binding in the publicly available HNF1A ChIP-seq data(*59*), but previous investigators have reported HNF1A binding at the CRP promoter(*82*–*84*) supporting a direct regulation.

Acknowledgements

We would like to thank Joseph Witztum and Vikas Bansal for helpful discussions. We are

grateful to Xiaolan Zhang and Kyle Sanchez for assistance with preparing the HNF1A variant library, cellular screening/FACS sorting, and production of *HNF1A*-null hepatocyte cell lines. We would also like to thank Twist Bioscience for constructing the HNF1A variant library. Henrietta Lacks and the HeLa cell line that was established from her tumor cells without her knowledge or consent in 1951 have made significant contributions to scientific progress and advances in human health. We are grateful to Henrietta Lacks, now deceased, and to her surviving family members for their contributions to biomedical research.

Data availability

Function scores for HNF1A missense variants can be queried interactively via our custom data portal found at https://medschool.ucsd.edu/som/medicine/divisions/endocrinology/research/labs/majithia/data/Pages/default.aspx. All UK Biobank data used in this study are accessible through application to UK Biobank.

Funding

This work was supported by grants from the National Institute of Diabetes and Digestive and Kidney Diseases (1R03DK113328 and 1R01DK123422 to A.R.M, 1R01DK125490 to J.F and A.R.M), a UCSD/UCLA Pilot and Feasibility grant (P30 DK063491 to A.R.M), and a Ruth L. Kirschstein Institutional National Research Service Award T32 GM008666 from the National Institute of General Medical Sciences (to N.D.). S.H. received support from NIDDK grants P30DK063491 and P30DK120515, and NIGMS R01GM129523.

G.M.P is supported by NHLBI grants R01HL142711 and R01HL127564. This publication includes data generated at the UC San Diego IGM Genomics Center utilizing an Illumina NovaSeq 6000 that was purchased with funding from a National Institutes of Health SIG grant (#S10 OD026929).

Author Contributions

N.D and A.R.M. designed the study. N.D., B.K., S.Hu, R.I., L.K., M.W., J.G., X.D., C.D.A.S., and A.M.D. generated/acquired and analyzed the data. N.D., B.K., S.Hu, R.I., M.W., R.A., J.F., G.M.P., S.H., P.G., A.M.D., A.V.K., J.O., V.R., and A.R.M., were involved in data interpretation. N.D. and A.R.M. drafted the manuscript. N.D., B.K., R.I., L.K., M.W., X.D., V.M., G.M.P., P.G., A.M.D., A.V.K., J.O., V.R., S.H., and A.R.M. were involved in critical manuscript revision.

Declaration of Interests

A.M.D. and L.K. are employees and shareholders of Alnylam Pharmaceuticals.

Methods

EXPERIMENTAL MODEL AND STUDY PARTICIPANT DETAILS

Cell line generation

HUH7 and Hep3B cells were cultured in DMEM (Life Technologies #11995065) supplemented with 10% FBS (Sigma #F2442) and L-Glutamine and antibiotics (Omega PG-30). Polyclonal *HNF1A*-null cells were generated by introducing the endonuclease

Cas9 and sgRNA (CTGTCCCAACACCTCAACAA) targeting exon 2 of human *HNF1A* into cells by lentiviral transduction as we have previously described(*20*).

UK Biobank cohort

The UK Biobank is a prospective cohort study with genotypic and phenotypic data that enrolled approximately 500,000 individuals aged 40-69 from across the United Kingdom(*32*). The UK Biobank study was approved by the Research Ethics Committee and informed consent was obtained from all participants. Analysis of UK Biobank data was conducted under application numbers 51436 and 26041.

METHOD DETAILS

Synthesis of all possible HNF1A amino acid variants

Human *HNF1A* cDNA (CCDS9209.1) was recoded by selecting alternative codons to eliminate the CRISPR/CAS9 binding site and optimize expression. This cDNA corresponds to the longest amino acid coding sequence of HNF1A expressed in the liver and also the most abundant(*50*). Constructs encoding all 19 alternative amino acids at each position were synthesized (Twist Bioscience). An additional 613 constructs encoding synonymous variants spread across the coding sequence were also synthesized and included. The HNF1A construct library described above was cloned into a lenti-viral expression vector by simple restriction cloning, sequence verified for quality and completeness, and transfected into a packaging cell line to produce pooled lenti-virus(*20*).

Pooled experimental assay of 11,970 HNF1A variants

The lentiviral library of HNF1A amino acid variants was introduced into HUH7 *HNF1A*-null cells under a doxycycline responsive promoter to enable controlled protein expression(*85*). To minimize doubly infected cells, $1.2x10^7$ viral particles were combined with $4x10^7$ cells after which uninfected cells were eliminated by selection with 4 μg/mL puromycin. At least 10^7 cells were infected to ensure that each HNF1A variant was independently represented in 1000 cells. The resulting polyclonal population of HUH7 hepatocytes (each cell containing a different HNF1A variant) was stimulated with doxycycline (5 μg/mL) for 6 days and then immuno-stained for TM4SF4 (R&D systems FAB7998A). Stained cells were sorted using FACS (FACS Aria II and BD Influx) into two equal bins: high expression and low expression bins based on TM4SF4 expression with a ten-fold expression intensity difference between each bin (Figure 2.1A). Control experiments with *HNF1A*-null and WT cells were performed to identify the TM4SF4 expression distributions of cells with and without HNF1A (Figure 2.1A dashed red and dashed green lines). These distributions helped to establish TM4SF4 expression cutoffs for the FACS sorting. Three independent sorting experiments were performed, sorting ~500,000 cells per bin per replicate. To re-identify and quantitate the HNF1A variants in the TM4SF4 'high' and 'low' bins, genomic DNA (gDNA) was extracted (Qiagen #51304) from each sample of 500,000 cells and the integrated proviral HNF1A transgenes recovered by PCR. Each sample of purified gDNA was split into 14 separate 100 uL PCR reactions (~400ng gDNA per reaction) and amplified using Herculase II (Agilent #600675) according to the manufacturer protocol using a Tm of 57^{o}C. All 14 PCR reactions per sample were pooled together, and the transgenes were purified (Qiagen #28104) and prepared for sequencing (Illumina Nextera #FC-131-1096) according to the

manufacturer's protocol. The six samples were pooled and sequenced to obtain approximately 350 million paired-end 100-base reads (Illumina NovaSeq) resulting in a 12x base coverage assuming a target size of 6 x 10^9 bp (2000 bp transgene x 12,000 variants x 250 observations per variant). Raw sequencing reads were aligned to the reference HNF1A recoded cDNA sequence using ORFCall (https://github.com/tedsharpe/ORFCall)(*20*) and the number of occurrences of each amino acid at each position along the coding region were counted and tabulated.

Calculation of function scores

We constructed a likelihood function based on the log odds of an amino acid variant being in the fraction with high or low TM4SF4. The log odds for each amino acid variant was estimated by maximizing a likelihood function based on the observed counts of each amino acid variant in the fractions with high and low TM4SF4 normalized to the total read depth at that amino acid position. Data were combined across experimental replicates. To avoid spuriously high or low log-odds estimates for any given variant, we constrained the log-odds estimate with a Gaussian prior whose parameters were estimated from data combined across all variants. These procedures were analogous to those we have previously described(*20*). Variants which had total counts across all replicates and bins less than the known low count variants from the original library design were not assigned a function score (125 variants). A mutation tolerance score was calculated for each of the 631 amino acid positions along HNF1A by taking the mean of all assigned function scores at that position.

To calculate the posterior error probability (PEP) that any particular variant was

not neutral, we computed the probability that it belonged to the synonymous function score distribution instead of the nonsynonymous distribution using the formula PEP_nonWT = $P_{synonymous}/(P_{synonymous} + P_{nonsynonymous})$(*31*). The probability density functions for $P_{synonymous}$ and $P_{nonsynonymous}$ were generated by linearly interpolating the distribution of function scores for synonymous and nonsynonymous variants using an interpolation function as instantiated in R v3.5.1.

Genetic data

For all exome sequenced cohorts, variants within the genomic coordinates of *HNF1A* (chr12:120,977,683-121,002,512, GRCh38) were extracted, and variant annotation was performed using SnpEff v4.3(*86*). Nomenclature used for missense variants is for the canonical HNF1A transcript ENST00000257555.10; protein ENSP00000257555.4. Disruptive variants (i.e. splice region variants, stop gained variants, or frameshift variants) were assigned the function score equal to the lowest known pathogenic variant functional score, homozygous carriers of rare HNF1A protein-coding variants (n=5) were removed from all analyses, and compound heterozygotes (n=39) were included twice as single variant carriers.

Metabolic biomarkers and clinical phenotypes

Clinical measurements from all 502,489 UK Biobank participants were log transformed, if not normally distributed and residualized adjusting for age, age^2 and sex, then normalized using a rank based inverse normal transformation(*87*). Diabetes mellitus designation was based on a compilation of self-report or ICD-10 codes of E11 in hospitalization or primary care records(*88*, *89*).

Functional validation of GOF HNF1A variants

Each GOF variant selected for validation (Figure 2.2A) was recreated as a transgene based on the native sequence of the cDNA (CCDS9209.1) with the exact human carrier nucleotide change observed in the UK Biobank. The constructed variants were tested for transactivation potential in HeLa cells as we have previously published(*37*) using the rat albumin promoter. All transactivation experiments included three parallels performed on three independent experimental days (nine readings in total).

Transcriptional activity of the HNF1A GOF variants was also tested in MIN6 cells using the HNF4aP2 promoter(*38*). Cells were co-transfected with a DNA mixture consisting of 10 ng of pGL-Luc and pGL4.75, and 10 ng of variant or WT *HNF1A*. The level of the transcription factor activity was evaluated by measuring firefly luciferase activity relative to renilla luciferase activity using the Dual Luciferase Assay System (Promega #E1500), as described by the manufacturer.

Assessment of ANGPTL3 levels in cultured hepatocytes

HNF1A constructs (WT, p.T260M, and p.P447L) were introduced into WT and *HNF1A* KO human HUH7 cells by lentiviral induction under a doxycycline responsive promoter as above. Cells were stimulated with doxycycline (0, 0.5, 1, 5 µg/mL) for 5 days. Following doxycycline incubation, the spent media was collected for ANGPTL3 ELISA analysis (R&D Systems #DANL 30) and the cells for protein quantification (RIPA buffer + protease inhibitor) according to the manufacturer's protocol. GOF variant constructs were transfected into HUH7 hepatocytes using TransIT-LT1 transfection reagent (Mirus Bio),

puromycin selection was performed to remove untransfected cells, and cells were incubated in fresh media for 24 hours before harvest for ELISA and protein quantification as described above.

Functional genomic analysis of HNF1A deleted hepatocytes and complementation with GOF variants

HNF1A-null HUH7 and Hep3b cells were generated as described above, and WT HNF1A was reintroduced using a doxycycline-inducible promoter (0, 2, 5 μg/mL) in triplicate. mRNA was extracted (Qiagen #74104) and sequenced (Illumina TruSeq) according to the manufacturers' protocol to a depth of at least 30 million reads per sample. Raw reads were aligned using Kallisto (v0.44.0)(*90*) with default parameters to generate gene counts per cell. DESeq2 (R v3.5.1, package v1.22.1)(*91*) was used to perform differential expression analysis with effect sizes and p-values obtained from the Wald test as instantiated within DESeq2. Gene set enrichment analysis (GSEA)(*55*, *56*) was conducted using the fgsea package with 10000 permutations (package v1.8.0(*92*)) utilizing Hallmark gene sets(*93*) and Gene Ontology biological process gene sets(*94*, *95*) related to significantly enriched Hallmark gene sets (cholesterol homeostasis, fatty acid metabolism, bile acid metabolism, coagulation, and complement activation). Significance was determined by a two-sided enrichment p-value with Benjamini–Hochberg correction for multiple hypotheses as instantiated within the fgsea software package.

Complementation of *HNF1A*-null HUH7 cells with mRNA encoding WT HNF1A or the GOF p.T196A variant was essentially performed as described before(*96*), with modifications. Briefly, cDNA constructs were amplified from the expression plasmids

described above with primers that introduce a T7 promoter and AG as the first two transcribed nucleotides and that add a 179-nt-long poly(A) tail to the 3' end. PCR products were phenol/chloroform-extracted and in vitro-transcribed with T7 RNA Polymerase HiScribe Mix (NEB E2040S), CleanCap Reagent AG (TriLink N-7113) and 5-methoxyuridine triphosphate instead of UTP for 2 hours at 37°C to generate capped, polyadenylated mRNA. IVT mRNA was precipitated with 2.5 M LiCl final, washed twice with 80% EtOH and dissolved in water. For each construct, 10^6 *HNF1A*-null HUH7 cells were electroporated with 2 µg mRNA in 100 µl Buffer R with four 10-millisecond pulses at 1230 V using a Neon Transfection System (ThermoFisher MPK10025). Cells were cultured for 18 hours in 5 mL medium in 60 mm plates, and mRNA was extracted (Zymo R1050), sequenced (Illumina Stranded mRNA), and gene expression analyzed as above. Expression of HNF1A was verified by Western blot, and EGFP mRNA was used as electroporation control. Differential expression analysis between *HNF1A*-null HUH7 complemented with EGFP and WT HNF1A showed a profile of differentially expressed genes similar to *HNF1A*-null vs HNF1A WT cells (top genes in both experiments included FGB, FABP1, PCSK9, and ANGPTL3)

HNF1A binding motifs were scored using JASPAR(*97*) consensus matrix sequence matches 5000 bp before and after the genomic coordinates of the transcriptional start site for each gene of interest. HNF1A HepG2 ChIP-seq bigWig signal p-value and IDR ranked peaks result files from both replicates were downloaded from ENCODE experiment ENCSR633HRJ(*98*).

QUANTIFICATION AND STATISTICAL ANALYSIS

Genotype-function-phenotype association analyses

To examine the association between the full continuum of HNF1A function score and diabetes status in the UK Biobank, we performed logistic mixed effects regression (lme4 v1.1-17, lmerTest v3.1-2) with variant identity as a random effect in order to account for multiple carriers of the same variant, adjusting for age, age^2, sex, and the first 10 genetic principal components of ancestry (computed by UK Biobank for multi-ancestry analyses (Field 22009) and computed for European ancestry participants as previously described(*88*)). To account for related individuals, in the full multi-ancestry analysis we randomly removed one participant if a pair of them had a genetic relationship equal or closer than a second-degree relative, and in the European analysis, related individuals were identified and excluded as per Deaton et al(*99*). To examine the association between functional categories (i.e. GOF/LOF) and diabetes or CAD status in the UK Biobank, we performed logistic regression (glm function, R stats v3.5.1) adjusting for covariates previously mentioned. To test the effect of categorization threshold on T2D, GOF/LOF variants were re-categorized based on increasingly stringent cutoffs defined according to the spread of synonymous variant distributions (+/- 1SD, 1.25, 1.5, 1.75, 2) and the effect on T2D risk in HNF1A protein-coding variant carriers in the UK Biobank was recomputed for each recategorization threshold as above. To assess the effect of common HNF1A protein-coding variation on the identified associations with T2D, we first identified all independent HNF1A protein coding SNPs with MAF > 0.01. Two such SNPs were identified, p.I27L (rs1169288) and p.A98V (rs1800574), and incorporated as a covariates in the regressions of T2D on HNF1A function score / categorized HNF1A variants.

Analysis in the AMP-T2D-GENES study was conducted as previously described(*43*). Briefly, we conducted two burden tests -- one including gain of function variants and one including loss of function variants -- using the EPACTS software package, regressing sample phenotype on genotype dosage. We conducted each test across all unrelated individuals pooled together and included principal components of ancestry covariates (computed from common variants) as well as covariates for sample cohort of origin and sequencing technology. The effect size and significance of the burden test was evaluated using the two-sided Firth logistic regression test. Meta-analysis for the UK Biobank and AMP-T2D cohorts was conducted using fixed-effect inverse-variance meta-analysis as implemented by METAL(*100*).

To test the association of rare *HNF1A* variants in the MIGen ExSeq data set with the risk of coronary artery disease, a Firth logistic regression was applied with sex and top 10 principal components of ancestry as covariates and cleaned to second-degree of relationship. The Firth logistic regression is a test robust to association testing in the context of low carrier counts or case-control imbalance(*101*). For each person, the genetic score of the qualified rare variant carriers was coded as 1, and 0 otherwise. Meta-analysis for the UK Biobank and MIGen ExSeq cohorts was conducted using fixed-effect inverse-variance meta-analysis as implemented by METAL(*100*). Power calculations were performed using the Purcell et. al.(*102*) genetic power calculator with the following parameters: high risk / marker allele frequency = cumulative GOF allele frequency = 0.002, a 5.4% prevalence of CAD in the UK Biobank, genotype relative risk = GOF OR for T2D = 0.75, total number of CAD cases in MIGEN and UK Biobank = 49,666, D’ = 1, and control:cases ratio = 9.44. Power is reported for $\alpha = 0.05$.

Quantitative phenotype associations with function scores were performed within the UK Biobank using the normalized clinical measurements described above and a linear mixed model with function score as a fixed effect, adjusting for 10 genetic principal components of ancestry as described above, and with variant identity as a random effect in order to account for multiple carriers of the same variant (lme4 v1.1-17, lmerTest v3.1-2).

To quantify liver specific HNF1A gene expression, the Functional Summary-based Imputation (FUSION) framework for generating individual-level predicted gene expression(*51*) was applied to UK Biobank genotyped individuals. First, to ensure that this analysis was independent of our analysis of protein-coding variants, carriers of rare protein-coding HNF1A variants (n=2,203) in the UK Biobank were excluded. Next, all HNF1A (ENSG00000135100) expression reference weights computed in European liver RNA-sequenced samples from the Genotype-Tissue Expression (GTEx) Project v8 were extracted from the FUSION archive (http://gusevlab.org/projects/fusion/#gtex-v8-multi-tissue-expression) (n=466 SNPs with precomputed expression weights). These expression weights and their corresponding imputed genotypes in the UK Biobank European individuals (n=407,227) were combined into a predicted gene expression score for each individual by computing a linear combination across the HNF1A SNPs expression weights and genotypes, then dividing by the number of non-missing SNPs as implemented in the Plink score utility(*103*). Phenotypes were normalized as described above and regressed onto predicted HNF1A liver expression scores using a generalized linear model (glm function, R stats v3.5.1) adjusting for 10 genetic principal components of ancestry. Regressions additionally included as a covariate individual level genotype of

the common HNF1A variant rs1169288 (encoding p.I27L) which has previously been associated with metabolic phenotypes(*15*, *104*) and was significantly associated with the predicted HNF1A liver expression score ($p < 2 \times 10^{-16}$, linear model). To assess potential confounding of the HNF1A liver gene expression score with rare-variant phenotypic associations, the models for T2D and quantitative trait regression on HNF1A function score were adjusted for each individual's computed HNF1A liver gene expression. This analysis was performed in the subset of UK Biobank European participants for whom liver eQTLs are available.

P-values from the regression of categorical phenotypes against HNF1A category were adjusted for multiple comparisons using Bonferroni correction for two categories (LOF and GOF), and all p-values from the regression of quantitative traits against HNF1A function score / liver expression score were adjusted using Bonferroni correction for six phenotype groups (C-reactive protein, LDL/Apolipoprotein B, Cholesterol/Triglycerides, HDL/Apolipoprotein A, Gamma glutamyltransferase, and HbA1c).

Chapter 2, in full, is an edited reprint of the material as it appears in Cell Genomics 2023. DeForest, Natalie; Kavitha, Babu; Hu, Siqi; Isaac, Roi; Krohn, Lynne; Wang, Minxian; Du, Xiaomi; Saldanha, Camila De Arruda; Gylys, Jenny; Merli, Edoardo; Abagyan, Ruben; Najmi, Laeya; Mohan, Viswanathan; Alnylam Human Genetics; AMP-T2D Consortium; Flannick, Jason; Peloso, Gina M.; Gordts, Philip L.S.M.; Heinz, Sven; Deaton, Aimee M.; Khera, Amit V.; Olefsky, Jerrold; Radha, Venkatesan; Majithia, Amit R., Cell Press, 2023. The dissertation author was the primary investigator and author of this paper.

References

1. H. H. Lau, N. H. J. Ng, L. S. W. Loo, J. B. Jasmen, A. K. K. Teo, The molecular functions of hepatocyte nuclear factors - In and beyond the liver. *J. Hepatol.* **68**, 1033–1048 (2018).

2. S. S. Fajans, G. I. Bell, MODY: history, genetics, pathophysiology, and clinical decision making. *Diabetes Care*. **34**, 1878–1884 (2011).

3. K. Yamagata, N. Oda, P. J. Kaisaki, S. Menzel, H. Furuta, M. Vaxillaire, L. Southam, R. D. Cox, G. M. Lathrop, V. V. Boriraj, X. Chen, N. J. Cox, Y. Oda, H. Yano, M. M. Le Beau, S. Yamada, H. Nishigori, J. Takeda, S. S. Fajans, A. T. Hattersley, N. Iwasaki, T. Hansen, O. Pedersen, K. S. Polonsky, G. I. Bell, Mutations in the hepatocyte nuclear factor-1alpha gene in maturity-onset diabetes of the young (MODY3). *Nature*. **384**, 455–458 (1996).

4. M. Lehto, T. Tuomi, M. M. Mahtani, E. Widén, C. Forsblom, L. Sarelin, M. Gullström, B. Isomaa, M. Lehtovirta, A. Hyrkkö, T. Kanninen, M. Orho, S. Manley, R. C. Turner, T. Brettin, A. Kirby, J. Thomas, G. Duyk, E. Lander, M. R. Taskinen, L. Groop, Characterization of the MODY3 phenotype. Early-onset diabetes caused by an insulin secretion defect. *J. Clin. Invest.* **99**, 582–591 (1997).

5. T. J. McDonald, J. McEneny, E. R. Pearson, G. Thanabalasingham, M. Szopa, B. M. Shields, S. Ellard, K. R. Owen, M. T. Malecki, A. T. Hattersley, I. S. Young, Lipoprotein composition in HNF1A-MODY: differentiating between HNF1A-MODY and type 2 diabetes. *Clin. Chim. Acta*. **413**, 927–932 (2012).

6. T. J. McDonald, B. M. Shields, J. Lawry, K. R. Owen, A. L. Gloyn, S. Ellard, A. T. Hattersley, High-sensitivity CRP discriminates HNF1A-MODY from other subtypes of diabetes. *Diabetes Care*. **34**, 1860–1862 (2011).

7. I. J. Goldberg, Clinical review 124: Diabetic dyslipidemia: causes and consequences. *J. Clin. Endocrinol. Metab.* **86**, 965–971 (2001).

8. D. E. King, A. G. Mainous 3rd, T. A. Buchanan, W. S. Pearson, C-reactive protein and glycemic control in adults with diabetes. *Diabetes Care*. **26**, 1535–1539 (2003).

9. D. Q. Shih, M. Bussen, E. Sehayek, M. Ananthanarayanan, B. L. Shneider, F. J. Suchy, S. Shefer, J. S. Bollileni, F. J. Gonzalez, J. L. Breslow, M. Stoffel, Hepatocyte nuclear factor-1α is an essential regulator of bile acid and plasma cholesterol metabolism. *Nature Genetics*. **27** (2001), pp. 375–382.

10. B. F. Voight, L. J. Scott, V. Steinthorsdottir, A. P. Morris, C. Dina, R. P. Welch, E. Zeggini, C. Huth, Y. S. Aulchenko, G. Thorleifsson, L. J. McCulloch, T. Ferreira, H. Grallert, N. Amin, G. Wu, C. J. Willer, S. Raychaudhuri, S. A. McCarroll, C. Langenberg, O. M. Hofmann, J. Dupuis, L. Qi, A. V. Segrè, M. van Hoek, P. Navarro, K. Ardlie, B. Balkau, R. Benediktsson, A. J. Bennett, R. Blagieva, E.

Boerwinkle, L. L. Bonnycastle, K. Bengtsson Boström, B. Bravenboer, S. Bumpstead, N. P. Burtt, G. Charpentier, P. S. Chines, M. Cornelis, D. J. Couper, G. Crawford, A. S. F. Doney, K. S. Elliott, A. L. Elliott, M. R. Erdos, C. S. Fox, C. S. Franklin, M. Ganser, C. Gieger, N. Grarup, T. Green, S. Griffin, C. J. Groves, C. Guiducci, S. Hadjadj, N. Hassanali, C. Herder, B. Isomaa, A. U. Jackson, P. R. V. Johnson, T. Jørgensen, W. H. L. Kao, N. Klopp, A. Kong, P. Kraft, J. Kuusisto, T. Lauritzen, M. Li, A. Lieverse, C. M. Lindgren, V. Lyssenko, M. Marre, T. Meitinger, K. Midthjell, M. A. Morken, N. Narisu, P. Nilsson, K. R. Owen, F. Payne, J. R. B. Perry, A.-K. Petersen, C. Platou, C. Proença, I. Prokopenko, W. Rathmann, N. W. Rayner, N. R. Robertson, G. Rocheleau, M. Roden, M. J. Sampson, R. Saxena, B. M. Shields, P. Shrader, G. Sigurdsson, T. Sparsø, K. Strassburger, H. M. Stringham, Q. Sun, A. J. Swift, B. Thorand, J. Tichet, T. Tuomi, R. M. van Dam, T. W. van Haeften, T. van Herpt, J. V. van Vliet-Ostaptchouk, G. B. Walters, M. N. Weedon, C. Wijmenga, J. Witteman, R. N. Bergman, S. Cauchi, F. S. Collins, A. L. Gloyn, U. Gyllensten, T. Hansen, W. A. Hide, G. A. Hitman, A. Hofman, D. J. Hunter, K. Hveem, M. Laakso, K. L. Mohlke, A. D. Morris, C. N. A. Palmer, P. P. Pramstaller, I. Rudan, E. Sijbrands, L. D. Stein, J. Tuomilehto, A. Uitterlinden, M. Walker, N. J. Wareham, R. M. Watanabe, G. R. Abecasis, B. O. Boehm, H. Campbell, M. J. Daly, A. T. Hattersley, F. B. Hu, J. B. Meigs, J. S. Pankow, O. Pedersen, H.-E. Wichmann, I. Barroso, J. C. Florez, T. M. Frayling, L. Groop, R. Sladek, U. Thorsteinsdottir, J. F. Wilson, T. Illig, P. Froguel, C. M. van Duijn, K. Stefansson, D. Altshuler, M. Boehnke, M. I. McCarthy, MAGIC investigators, GIANT Consortium, Twelve type 2 diabetes susceptibility loci identified through large-scale association analysis. *Nat. Genet.* **42**, 579–589 (2010).

11. T. M. Teslovich, K. Musunuru, A. V. Smith, A. C. Edmondson, I. M. Stylianou, M. Koseki, J. P. Pirruccello, S. Ripatti, D. I. Chasman, C. J. Willer, C. T. Johansen, S. W. Fouchier, A. Isaacs, G. M. Peloso, M. Barbalic, S. L. Ricketts, J. C. Bis, Y. S. Aulchenko, G. Thorleifsson, M. F. Feitosa, J. Chambers, M. Orho-Melander, O. Melander, T. Johnson, X. Li, X. Guo, M. Li, Y. Shin Cho, M. Jin Go, Y. Jin Kim, J.-Y. Lee, T. Park, K. Kim, X. Sim, R. Twee-Hee Ong, D. C. Croteau-Chonka, L. A. Lange, J. D. Smith, K. Song, J. Hua Zhao, X. Yuan, J. 'an Luan, C. Lamina, A. Ziegler, W. Zhang, R. Y. L. Zee, A. F. Wright, J. C. M. Witteman, J. F. Wilson, G. Willemsen, H.-E. Wichmann, J. B. Whitfield, D. M. Waterworth, N. J. Wareham, G. Waeber, P. Vollenweider, B. F. Voight, V. Vitart, A. G. Uitterlinden, M. Uda, J. Tuomilehto, J. R. Thompson, T. Tanaka, I. Surakka, H. M. Stringham, T. D. Spector, N. Soranzo, J. H. Smit, J. Sinisalo, K. Silander, E. J. G. Sijbrands, A. Scuteri, J. Scott, D. Schlessinger, S. Sanna, V. Salomaa, J. Saharinen, C. Sabatti, A. Ruokonen, I. Rudan, L. M. Rose, R. Roberts, M. Rieder, B. M. Psaty, P. P. Pramstaller, I. Pichler, M. Perola, B. W. J. H. Penninx, N. L. Pedersen, C. Pattaro, A. N. Parker, G. Pare, B. A. Oostra, C. J. O'Donnell, M. S. Nieminen, D. A. Nickerson, G. W. Montgomery, T. Meitinger, R. McPherson, M. I. McCarthy, W. McArdle, D. Masson, N. G. Martin, F. Marroni, M. Mangino, P. K. E. Magnusson, G. Lucas, R. Luben, R. J. F. Loos, M.-L. Lokki, G. Lettre, C. Langenberg, L. J. Launer, E. G. Lakatta, R. Laaksonen, K. O. Kyvik, F. Kronenberg, I. R. König, K.-T. Khaw, J. Kaprio, L. M. Kaplan, A. Johansson, M.-R. Jarvelin, A. C. J. W. Janssens, E. Ingelsson, W. Igl, G. Kees Hovingh, J.-J. Hottenga, A. Hofman, A. A. Hicks, C.

Hengstenberg, I. M. Heid, C. Hayward, A. S. Havulinna, N. D. Hastie, T. B. Harris, T. Haritunians, A. S. Hall, U. Gyllensten, C. Guiducci, L. C. Groop, E. Gonzalez, C. Gieger, N. B. Freimer, L. Ferrucci, J. Erdmann, P. Elliott, K. G. Ejebe, A. Döring, A. F. Dominiczak, S. Demissie, P. Deloukas, E. J. C. de Geus, U. de Faire, G. Crawford, F. S. Collins, Y. I. Chen, M. J. Caulfield, H. Campbell, N. P. Burtt, L. L. Bonnycastle, D. I. Boomsma, S. M. Boekholdt, R. N. Bergman, I. Barroso, S. Bandinelli, C. M. Ballantyne, T. L. Assimes, T. Quertermous, D. Altshuler, M. Seielstad, T. Y. Wong, E.-S. Tai, A. B. Feranil, C. W. Kuzawa, L. S. Adair, H. A. Taylor Jr, I. B. Borecki, S. B. Gabriel, J. G. Wilson, H. Holm, U. Thorsteinsdottir, V. Gudnason, R. M. Krauss, K. L. Mohlke, J. M. Ordovas, P. B. Munroe, J. S. Kooner, A. R. Tall, R. A. Hegele, J. J. P. Kastelein, E. E. Schadt, J. I. Rotter, E. Boerwinkle, D. P. Strachan, V. Mooser, K. Stefansson, M. P. Reilly, N. J. Samani, H. Schunkert, L. A. Cupples, M. S. Sandhu, P. M. Ridker, D. J. Rader, C. M. van Duijn, L. Peltonen, G. R. Abecasis, M. Boehnke, S. Kathiresan, Biological, clinical and population relevance of 95 loci for blood lipids. *Nature*. **466**, 707–713 (2010).

12. Y. Wu, T. W. McDade, C. W. Kuzawa, J. Borja, Y. Li, L. S. Adair, K. L. Mohlke, L. A. Lange, Genome-wide association with C-reactive protein levels in CLHNS: evidence for the CRP and HNF1A loci and their interaction with exposure to a pathogenic environment. *Inflammation*. **35**, 574–583 (2012).

13. F. M. A. Giuffrida, G. K. Furuzawa, T. S. Kasamatsu, M. M. Oliveira, A. F. Reis, S. A. Dib, HNF1A gene polymorphisms and cardiovascular risk factors in individuals with late-onset autosomal dominant diabetes: a cross-sectional study. *Cardiovasc. Diabetol.* **8**, 28 (2009).

14. J. Erdmann, A. Grosshennig, P. S. Braund, I. R. König, C. Hengstenberg, A. S. Hall, P. Linsel-Nitschke, S. Kathiresan, B. Wright, D.-A. Trégouët, F. Cambien, P. Bruse, Z. Aherrahrou, A. K. Wagner, K. Stark, S. M. Schwartz, V. Salomaa, R. Elosua, O. Melander, B. F. Voight, C. J. O'Donnell, L. Peltonen, D. S. Siscovick, D. Altshuler, P. A. Merlini, F. Peyvandi, L. Bernardinelli, D. Ardissino, A. Schillert, S. Blankenberg, T. Zeller, P. Wild, D. F. Schwarz, L. Tiret, C. Perret, S. Schreiber, N. E. El Mokhtari, A. Schäfer, W. März, W. Renner, P. Bugert, H. Klüter, J. Schrezenmeir, D. Rubin, S. G. Ball, A. J. Balmforth, H.-E. Wichmann, T. Meitinger, M. Fischer, C. Meisinger, J. Baumert, A. Peters, W. H. Ouwehand, Italian Atherosclerosis, Thrombosis, and Vascular Biology Working Group, Myocardial Infarction Genetics Consortium, Wellcome Trust Case Control Consortium, Cardiogenics Consortium, P. Deloukas, J. R. Thompson, A. Ziegler, N. J. Samani, H. Schunkert, New susceptibility locus for coronary artery disease on chromosome 3q22.3. *Nat. Genet.* **41**, 280–282 (2009).

15. A. P. Reiner, M. D. Gross, C. S. Carlson, S. J. Bielinski, L. A. Lange, M. Fornage, N. S. Jenny, J. Walston, R. P. Tracy, O. D. Williams, D. R. Jacobs Jr, D. A. Nickerson, Common coding variants of the HNF1A gene are associated with multiple cardiovascular risk phenotypes in community-based samples of younger and older European-American adults: the Coronary Artery Risk Development in Young Adults Study and The Cardiovascular Health Study. *Circ. Cardiovasc.*

Genet. **2**, 244–254 (2009).

16. Ž. Reiner, Hypertriglyceridaemia and risk of coronary artery disease. *Nat. Rev. Cardiol.* **14**, 401–411 (2017).

17. W. K. Lagrand, C. A. Visser, W. T. Hermens, H. W. Niessen, F. W. Verheugt, G. J. Wolbink, C. E. Hack, C-reactive protein as a cardiovascular risk factor: more than an epiphenomenon? *Circulation*. **100**, 96–102 (1999).

18. J. Flannick, N. L. Beer, A. G. Bick, V. Agarwala, J. Molnes, N. Gupta, N. P. Burtt, J. C. Florez, J. B. Meigs, H. Taylor, V. Lyssenko, H. Irgens, E. Fox, F. Burslem, S. Johansson, M. J. Brosnan, J. K. Trimmer, C. Newton-Cheh, T. Tuomi, A. Molven, J. G. Wilson, C. J. O'Donnell, S. Kathiresan, J. N. Hirschhorn, P. R. Njølstad, T. Rolph, J. G. Seidman, S. Gabriel, D. R. Cox, C. E. Seidman, L. Groop, D. Altshuler, Assessing the phenotypic effects in the general population of rare variants in genes for a dominant Mendelian form of diabetes. *Nat. Genet.* **45**, 1380–1385 (2013).

19. D. M. Fowler, S. Fields, Deep mutational scanning: a new style of protein science. *Nat. Methods*. **11**, 801–807 (2014).

20. A. R. Majithia, B. Tsuda, M. Agostini, K. Gnanapradeepan, R. Rice, G. Peloso, K. A. Patel, X. Zhang, M. F. Broekema, N. Patterson, M. Duby, T. Sharpe, E. Kalkhoven, E. D. Rosen, I. Barroso, S. Ellard, UK Monogenic Diabetes Consortium, S. Kathiresan, Myocardial Infarction Genetics Consortium, S. O'Rahilly, UK Congenital Lipodystrophy Consortium, K. Chatterjee, J. C. Florez, T. Mikkelsen, D. B. Savage, D. Altshuler, Prospective functional classification of all possible missense variants in PPARG. *Nat. Genet.* **48**, 1570–1575 (2016).

21. G. M. Findlay, R. M. Daza, B. Martin, M. D. Zhang, A. P. Leith, M. Gasperini, J. D. Janizek, X. Huang, L. M. Starita, J. Shendure, Accurate classification of BRCA1 variants with saturation genome editing. *Nature*. **562**, 217–222 (2018).

22. E. V. Abel, M. Goto, B. Magnuson, S. Abraham, N. Ramanathan, E. Hotaling, A. A. Alaniz, C. Kumar-Sinha, M. L. Dziubinski, S. Urs, L. Wang, J. Shi, M. Waghray, M. Ljungman, H. C. Crawford, D. M. Simeone, HNF1A is a novel oncogene that regulates human pancreatic cancer stem cell properties. *Elife*. **7** (2018), doi:10.7554/eLife.33947.

23. J.-M. Servitja, M. Pignatelli, M. A. Maestro, C. Cardalda, S. F. Boj, J. Lozano, E. Blanco, A. Lafuente, M. I. McCarthy, L. Sumoy, R. Guigó, J. Ferrer, Hnf1alpha (MODY3) controls tissue-specific transcriptional programs and exerts opposed effects on cell growth in pancreatic islets and liver. *Mol. Cell. Biol.* **29**, 2945–2959 (2009).

24. A. O. Giacomelli, X. Yang, R. E. Lintner, J. M. McFarland, M. Duby, J. Kim, T. P. Howard, D. Y. Takeda, S. H. Ly, E. Kim, H. S. Gannon, B. Hurhula, T. Sharpe, A. Goodale, B. Fritchman, S. Steelman, F. Vazquez, A. Tsherniak, A. J. Aguirre, J. G. Doench, F. Piccioni, C. W. M. Roberts, M. Meyerson, G. Getz, C. M. Johannessen,

D. E. Root, W. C. Hahn, Mutational processes shape the landscape of TP53 mutations in human cancer. *Nat. Genet.* **50**, 1381–1387 (2018).

25. Y.-I. Chi, J. Daniel Frantz, B.-C. Oh, L. Hansen, S. Dhe-Paganon, S. E. Shoelson, Diabetes Mutations Delineate an Atypical POU Domain in HNF-1α. *Molecular Cell*. **10** (2002), pp. 1129–1137.

26. R. B. Rose, J. A. Endrizzi, J. D. Cronk, J. Holton, T. Alber, High-Resolution Structure of the HNF-1α Dimerization Domain. *Biochemistry*. **40** (2001), pp. 3242–3242.

27. G. M. Cooper, E. A. Stone, G. Asimenos, NISC Comparative Sequencing Program, E. D. Green, S. Batzoglou, A. Sidow, Distribution and intensity of constraint in mammalian genomic sequence. *Genome Res.* **15**, 901–913 (2005).

28. I. Adzhubei, D. M. Jordan, S. R. Sunyaev, Predicting functional effect of human missense mutations using PolyPhen-2. *Curr. Protoc. Hum. Genet.* **Chapter 7**, Unit7.20 (2013).

29. P. Rentzsch, D. Witten, G. M. Cooper, J. Shendure, M. Kircher, CADD: predicting the deleteriousness of variants throughout the human genome. *Nucleic Acids Res.* **47**, D886–D894 (2019).

30. M. J. Landrum, J. M. Lee, M. Benson, G. R. Brown, C. Chao, S. Chitipiralla, B. Gu, J. Hart, D. Hoffman, W. Jang, K. Karapetyan, K. Katz, C. Liu, Z. Maddipatla, A. Malheiro, K. McDaniel, M. Ovetsky, G. Riley, G. Zhou, J. B. Holmes, B. L. Kattman, D. R. Maglott, ClinVar: improving access to variant interpretations and supporting evidence. *Nucleic Acids Res.* **46**, D1062–D1067 (2018).

31. L. Käll, J. D. Storey, M. J. MacCoss, W. S. Noble, Posterior error probabilities and false discovery rates: two sides of the same coin. *J. Proteome Res.* **7**, 40–44 (2008).

32. C. Bycroft, C. Freeman, D. Petkova, G. Band, L. T. Elliott, K. Sharp, A. Motyer, D. Vukcevic, O. Delaneau, J. O'Connell, A. Cortes, S. Welsh, A. Young, M. Effingham, G. McVean, S. Leslie, N. Allen, P. Donnelly, J. Marchini, The UK Biobank resource with deep phenotyping and genomic data. *Nature*. **562**, 203–209 (2018).

33. J. D. Backman, A. H. Li, A. Marcketta, D. Sun, J. Mbatchou, M. D. Kessler, C. Benner, D. Liu, A. E. Locke, S. Balasubramanian, A. Yadav, N. Banerjee, C. E. Gillies, A. Damask, S. Liu, X. Bai, A. Hawes, E. Maxwell, L. Gurski, K. Watanabe, J. A. Kosmicki, V. Rajagopal, J. Mighty, Regeneron Genetics Center, DiscovEHR, M. Jones, L. Mitnaul, E. Stahl, G. Coppola, E. Jorgenson, L. Habegger, W. J. Salerno, A. R. Shuldiner, L. A. Lotta, J. D. Overton, M. N. Cantor, J. G. Reid, G. Yancopoulos, H. M. Kang, J. Marchini, A. Baras, G. R. Abecasis, M. A. R. Ferreira, Exome sequencing and analysis of 454,787 UK Biobank participants. *Nature*. **599**, 628–634 (2021).

34. J. A. Tennessen, A. W. Bigham, T. D. O'Connor, W. Fu, E. E. Kenny, S. Gravel, S. McGee, R. Do, X. Liu, G. Jun, H. M. Kang, D. Jordan, S. M. Leal, S. Gabriel, M. J. Rieder, G. Abecasis, D. Altshuler, D. A. Nickerson, E. Boerwinkle, S. Sunyaev, C. D. Bustamante, M. J. Bamshad, J. M. Akey, Broad GO, Seattle GO, NHLBI Exome Sequencing Project, Evolution and functional impact of rare coding variation from deep sequencing of human exomes. *Science*. **337**, 64–69 (2012).

35. B. M. Shields, S. Hicks, M. H. Shepherd, K. Colclough, A. T. Hattersley, S. Ellard, Maturity-onset diabetes of the young (MODY): how many cases are we missing? *Diabetologia*. **53**, 2504–2508 (2010).

36. L. A. Najmi, I. Aukrust, J. Flannick, J. Molnes, N. Burtt, A. Molven, L. Groop, D. Altshuler, S. Johansson, L. Bjørkhaug, P. R. Njølstad, Functional Investigations of HNF1A Identify Rare Variants as Risk Factors for Type 2 Diabetes in the General Population. *Diabetes*. **66** (2017), pp. 335–346.

37. K. Balamurugan, L. Bjørkhaug, S. Mahajan, S. Kanthimathi, P. R. Njølstad, N. Srinivasan, V. Mohan, V. Radha, Structure-function studies of HNF1A (MODY3) gene mutations in South Indian patients with monogenic diabetes. *Clin. Genet.* **90**, 486–495 (2016).

38. M. Galán, C.-M. García-Herrero, S. Azriel, M. Gargallo, M. Durán, J.-J. Gorgojo, V.-M. Andía, M.-A. Navas, Differential effects of HNF-1α mutations associated with familial young-onset diabetes on target gene regulation. *Mol. Med.* **17**, 256–265 (2011).

39. F. E. Dewey, V. Gusarova, R. L. Dunbar, C. O'Dushlaine, C. Schurmann, O. Gottesman, S. McCarthy, C. V. Van Hout, S. Bruse, H. M. Dansky, J. B. Leader, M. F. Murray, M. D. Ritchie, H. L. Kirchner, L. Habegger, A. Lopez, J. Penn, A. Zhao, W. Shao, N. Stahl, A. J. Murphy, S. Hamon, A. Bouzelmat, R. Zhang, B. Shumel, R. Pordy, D. Gipe, G. A. Herman, W. H. H. Sheu, I.-T. Lee, K.-W. Liang, X. Guo, J. I. Rotter, Y.-D. I. Chen, W. E. Kraus, S. H. Shah, S. Damrauer, A. Small, D. J. Rader, A. B. Wulff, B. G. Nordestgaard, A. Tybjærg-Hansen, A. M. van den Hoek, H. M. G. Princen, D. H. Ledbetter, D. J. Carey, J. D. Overton, J. G. Reid, W. J. Sasiela, P. Banerjee, A. R. Shuldiner, I. B. Borecki, T. M. Teslovich, G. D. Yancopoulos, S. J. Mellis, J. Gromada, A. Baras, Genetic and Pharmacologic Inactivation of ANGPTL3 and Cardiovascular Disease. *N. Engl. J. Med.* **377**, 211–221 (2017).

40. K. Musunuru, J. P. Pirruccello, R. Do, G. M. Peloso, C. Guiducci, C. Sougnez, K. V. Garimella, S. Fisher, J. Abreu, A. J. Barry, T. Fennell, E. Banks, L. Ambrogio, K. Cibulskis, A. Kernytsky, E. Gonzalez, N. Rudzicz, J. C. Engert, M. A. DePristo, M. J. Daly, J. C. Cohen, H. H. Hobbs, D. Altshuler, G. Schonfeld, S. B. Gabriel, P. Yue, S. Kathiresan, Exome sequencing, ANGPTL3 mutations, and familial combined hypolipidemia. *N. Engl. J. Med.* **363**, 2220–2227 (2010).

41. D. T. Odom, N. Zizlsperger, D. B. Gordon, G. W. Bell, N. J. Rinaldi, H. L. Murray, T.

L. Volkert, J. Schreiber, P. A. Rolfe, D. K. Gifford, E. Fraenkel, G. I. Bell, R. A. Young, Control of pancreas and liver gene expression by HNF transcription factors. *Science*. **303**, 1378–1381 (2004).

42. SIGMA Type 2 Diabetes Consortium, K. Estrada, I. Aukrust, L. Bjørkhaug, N. P. Burtt, J. M. Mercader, H. García-Ortiz, A. Huerta-Chagoya, H. Moreno-Macías, G. Walford, J. Flannick, A. L. Williams, M. J. Gómez-Vázquez, J. C. Fernandez-Lopez, A. Martínez-Hernández, S. Jiménez-Morales, F. Centeno-Cruz, E. Mendoza-Caamal, C. Revilla-Monsalve, S. Islas-Andrade, E. J. Córdova, X. Soberón, M. E. González-Villalpando, E. Henderson, L. R. Wilkens, L. Le Marchand, O. Arellano-Campos, M. L. Ordóñez-Sánchez, M. Rodríguez-Torres, R. Rodríguez-Guillén, L. Riba, L. A. Najmi, S. B. R. Jacobs, T. Fennell, S. Gabriel, P. Fontanillas, C. L. Hanis, D. M. Lehman, C. P. Jenkinson, H. E. Abboud, G. I. Bell, M. L. Cortes, M. Boehnke, C. González-Villalpando, L. Orozco, C. A. Haiman, T. Tusié-Luna, C. A. Aguilar-Salinas, D. Altshuler, P. R. Njølstad, J. C. Florez, D. G. MacArthur, Association of a low-frequency variant in HNF1A with type 2 diabetes in a Latino population. *JAMA*. **311**, 2305–2314 (2014).

43. J. Flannick, J. M. Mercader, C. Fuchsberger, M. S. Udler, A. Mahajan, J. Wessel, T. M. Teslovich, L. Caulkins, R. Koesterer, F. Barajas-Olmos, T. W. Blackwell, E. Boerwinkle, J. A. Brody, F. Centeno-Cruz, L. Chen, S. Chen, C. Contreras-Cubas, E. Córdova, A. Correa, M. Cortes, R. A. DeFronzo, L. Dolan, K. L. Drews, A. Elliott, J. S. Floyd, S. Gabriel, M. E. Garay-Sevilla, H. García-Ortiz, M. Gross, S. Han, N. L. Heard-Costa, A. U. Jackson, M. E. Jørgensen, H. M. Kang, M. Kelsey, B.-J. Kim, H. A. Koistinen, J. Kuusisto, J. B. Leader, A. Linneberg, C.-T. Liu, J. Liu, V. Lyssenko, A. K. Manning, A. Marcketta, J. M. Malacara-Hernandez, A. Martínez-Hernández, K. Matsuo, E. Mayer-Davis, E. Mendoza-Caamal, K. L. Mohlke, A. C. Morrison, A. Ndungu, M. C. Y. Ng, C. O'Dushlaine, A. J. Payne, C. Pihoker, Broad Genomics Platform, W. S. Post, M. Preuss, B. M. Psaty, R. S. Vasan, N. W. Rayner, A. P. Reiner, C. Revilla-Monsalve, N. R. Robertson, N. Santoro, C. Schurmann, W. Y. So, X. Soberón, H. M. Stringham, T. M. Strom, C. H. T. Tam, F. Thameem, B. Tomlinson, J. M. Torres, R. P. Tracy, R. M. van Dam, M. Vujkovic, S. Wang, R. P. Welch, D. R. Witte, T.-Y. Wong, G. Atzmon, N. Barzilai, J. Blangero, L. L. Bonnycastle, D. W. Bowden, J. C. Chambers, E. Chan, C.-Y. Cheng, Y. S. Cho, F. S. Collins, P. S. de Vries, R. Duggirala, B. Glaser, C. Gonzalez, M. E. Gonzalez, L. Groop, J. S. Kooner, S. H. Kwak, M. Laakso, D. M. Lehman, P. Nilsson, T. D. Spector, E. S. Tai, T. Tuomi, J. Tuomilehto, J. G. Wilson, C. A. Aguilar-Salinas, E. Bottinger, B. Burke, D. J. Carey, J. C. N. Chan, J. Dupuis, P. Frossard, S. R. Heckbert, M. Y. Hwang, Y. J. Kim, H. L. Kirchner, J.-Y. Lee, J. Lee, R. J. F. Loos, R. C. W. Ma, A. D. Morris, C. J. O'Donnell, C. N. A. Palmer, J. Pankow, K. S. Park, A. Rasheed, D. Saleheen, X. Sim, K. S. Small, Y. Y. Teo, C. Haiman, C. L. Hanis, B. E. Henderson, L. Orozco, T. Tusié-Luna, F. E. Dewey, A. Baras, C. Gieger, T. Meitinger, K. Strauch, L. Lange, N. Grarup, T. Hansen, O. Pedersen, P. Zeitler, D. Dabelea, G. Abecasis, G. I. Bell, N. J. Cox, M. Seielstad, R. Sladek, J. B. Meigs, S. S. Rich, J. I. Rotter, DiscovEHR Collaboration, CHARGE, LuCamp, ProDiGY, GoT2D, ESP, SIGMA-T2D, T2D-GENES, AMP-T2D-GENES, D. Altshuler, N. P. Burtt, L. J. Scott, A. P. Morris, J. C. Florez, M. I. McCarthy, M. Boehnke, Exome

sequencing of 20,791 cases of type 2 diabetes and 24,440 controls. *Nature*. **570**, 71–76 (2019).

44. K. J. Gaulton, T. Ferreira, Y. Lee, A. Raimondo, R. Mägi, M. E. Reschen, A. Mahajan, A. Locke, N. W. Rayner, N. Robertson, R. A. Scott, I. Prokopenko, L. J. Scott, T. Green, T. Sparso, D. Thuillier, L. Yengo, H. Grallert, S. Wahl, M. Frånberg, R. J. Strawbridge, H. Kestler, H. Chheda, L. Eisele, S. Gustafsson, V. Steinthorsdottir, G. Thorleifsson, L. Qi, L. C. Karssen, E. M. van Leeuwen, S. M. Willems, M. Li, H. Chen, C. Fuchsberger, P. Kwan, C. Ma, M. Linderman, Y. Lu, S. K. Thomsen, J. K. Rundle, N. L. Beer, M. van de Bunt, A. Chalisey, H. M. Kang, B. F. Voight, G. R. Abecasis, P. Almgren, D. Baldassarre, B. Balkau, R. Benediktsson, M. Blüher, H. Boeing, L. L. Bonnycastle, E. P. Bottinger, N. P. Burtt, J. Carey, G. Charpentier, P. S. Chines, M. C. Cornelis, D. J. Couper, A. T. Crenshaw, R. M. van Dam, A. S. F. Doney, M. Dorkhan, S. Edkins, J. G. Eriksson, T. Esko, E. Eury, J. Fadista, J. Flannick, P. Fontanillas, C. Fox, P. W. Franks, K. Gertow, C. Gieger, B. Gigante, O. Gottesman, G. B. Grant, N. Grarup, C. J. Groves, M. Hassinen, C. T. Have, C. Herder, O. L. Holmen, A. B. Hreidarsson, S. E. Humphries, D. J. Hunter, A. U. Jackson, A. Jonsson, M. E. Jørgensen, T. Jørgensen, W.-H. L. Kao, N. D. Kerrison, L. Kinnunen, N. Klopp, A. Kong, P. Kovacs, P. Kraft, J. Kravic, C. Langford, K. Leander, L. Liang, P. Lichtner, C. M. Lindgren, E. Lindholm, A. Linneberg, C.-T. Liu, S. Lobbens, J. ’an Luan, V. Lyssenko, S. Männistö, O. McLeod, J. Meyer, E. Mihailov, G. Mirza, T. W. Mühleisen, M. Müller-Nurasyid, C. Navarro, M. M. Nöthen, N. N. Oskolkov, K. R. Owen, D. Palli, S. Pechlivanis, L. Peltonen, J. R. B. Perry, C. G. P. Platou, M. Roden, D. Ruderfer, D. Rybin, Y. T. van der Schouw, B. Sennblad, G. Sigurðsson, A. Stančáková, G. Steinbach, P. Storm, K. Strauch, H. M. Stringham, Q. Sun, B. Thorand, E. Tikkanen, A. Tonjes, J. Trakalo, E. Tremoli, T. Tuomi, R. Wennauer, S. Wiltshire, A. R. Wood, E. Zeggini, I. Dunham, E. Birney, L. Pasquali, J. Ferrer, R. J. F. Loos, J. Dupuis, J. C. Florez, E. Boerwinkle, J. S. Pankow, C. van Duijn, E. Sijbrands, J. B. Meigs, F. B. Hu, U. Thorsteinsdottir, K. Stefansson, T. A. Lakka, R. Rauramaa, M. Stumvoll, N. L. Pedersen, L. Lind, S. M. Keinanen-Kiukaanniemi, E. Korpi-Hyövälti, T. E. Saaristo, J. Saltevo, J. Kuusisto, M. Laakso, A. Metspalu, R. Erbel, K.-H. Jöcke, S. Moebus, S. Ripatti, V. Salomaa, E. Ingelsson, B. O. Boehm, R. N. Bergman, F. S. Collins, K. L. Mohlke, H. Koistinen, J. Tuomilehto, K. Hveem, I. Njølstad, P. Deloukas, P. J. Donnelly, T. M. Frayling, A. T. Hattersley, U. de Faire, A. Hamsten, T. Illig, A. Peters, S. Cauchi, R. Sladek, P. Froguel, T. Hansen, O. Pedersen, A. D. Morris, C. N. A. Palmer, S. Kathiresan, O. Melander, P. M. Nilsson, L. C. Groop, I. Barroso, C. Langenberg, N. J. Wareham, C. A. O’Callaghan, A. L. Gloyn, D. Altshuler, M. Boehnke, T. M. Teslovich, M. I. McCarthy, A. P. Morris, DIAbetes Genetics Replication And Meta-analysis (DIAGRAM) Consortium, Genetic fine mapping and genomic annotation defines causal mechanisms at type 2 diabetes susceptibility loci. *Nat. Genet.* **47**, 1415–1425 (2015).

45. Prospective Studies Collaboration, S. Lewington, G. Whitlock, R. Clarke, P. Sherliker, J. Emberson, J. Halsey, N. Qizilbash, R. Peto, R. Collins, Blood cholesterol and vascular mortality by age, sex, and blood pressure: a meta-analysis of individual data from 61 prospective studies with 55,000 vascular deaths. *Lancet*.

370, 1829–1839 (2007).

46. G. Targher, C. D. Byrne, Circulating Markers of Liver Function and Cardiovascular Disease Risk. *Arterioscler. Thromb. Vasc. Biol.* **35**, 2290–2296 (2015).

47. C. Nielson, T. Lange, N. Hadjokas, Blood Glucose and Coronary Artery Disease in Nondiabetic Patients. *Diabetes Care*. **29**, 998–1001 (2006).

48. Myocardial Infarction Genetics Consortium Investigators, N. O. Stitziel, H.-H. Won, A. C. Morrison, G. M. Peloso, R. Do, L. A. Lange, P. Fontanillas, N. Gupta, S. Duga, A. Goel, M. Farrall, D. Saleheen, P. Ferrario, I. König, R. Asselta, P. A. Merlini, N. Marziliano, M. F. Notarangelo, U. Schick, P. Auer, T. L. Assimes, M. Reilly, R. Wilensky, D. J. Rader, G. K. Hovingh, T. Meitinger, T. Kessler, A. Kastrati, K.-L. Laugwitz, D. Siscovick, J. I. Rotter, S. L. Hazen, R. Tracy, S. Cresci, J. Spertus, R. Jackson, S. M. Schwartz, P. Natarajan, J. Crosby, D. Muzny, C. Ballantyne, S. S. Rich, C. J. O'Donnell, G. Abecasis, S. Sunaev, D. A. Nickerson, J. E. Buring, P. M. Ridker, D. I. Chasman, E. Austin, I. J. Kullo, P. E. Weeke, C. M. Shaffer, L. A. Bastarache, J. C. Denny, D. M. Roden, C. Palmer, P. Deloukas, D.-Y. Lin, Z.-Z. Tang, J. Erdmann, H. Schunkert, J. Danesh, J. Marrugat, R. Elosua, D. Ardissino, R. McPherson, H. Watkins, A. P. Reiner, J. G. Wilson, D. Altshuler, R. A. Gibbs, E. S. Lander, E. Boerwinkle, S. Gabriel, S. Kathiresan, Inactivating mutations in NPC1L1 and protection from coronary heart disease. *N. Engl. J. Med.* **371**, 2072–2082 (2014).

49. Genetic effects on gene expression across human tissues. *Nature*. **550**, 204–213 (2017).

50. GTEx Consortium, The Genotype-Tissue Expression (GTEx) project. *Nat. Genet.* **45**, 580–585 (2013).

51. A. Gusev, A. Ko, H. Shi, G. Bhatia, W. Chung, B. W. Penninx, R. Jansen, E. J. de Geus, D. I. Boomsma, F. A. Wright, P. F. Sullivan, E. Nikkola, M. Alvarez, M. Civelek, A. J. Lusis, T. Lehtimäki, E. Raitoharju, M. Kähönen, I. Seppälä, O. T. Raitakari, J. Kuusisto, M. Laakso, A. L. Price, P. Pajukanta, B. Pasaniuc, Integrative approaches for large-scale transcriptome-wide association studies. *Nat. Genet.* **48** (2016), doi:10.1038/ng.3506.

52. M. B. Pepys, G. M. Hirschfield, C-reactive protein: a critical update. *J. Clin. Invest.* **111**, 1805–1812 (2003).

53. K. Yamagata, Roles of HNF1α and HNF4α in pancreatic β-cells: lessons from a monogenic form of diabetes (MODY). *Vitam. Horm.* **95**, 407–423 (2014).

54. J. Dalmon, M. Laurent, G. Courtois, The human beta fibrinogen promoter contains a hepatocyte nuclear factor 1-dependent interleukin-6-responsive element. *Mol. Cell. Biol.* **13**, 1183–1193 (1993).

55. A. Subramanian, P. Tamayo, V. K. Mootha, S. Mukherjee, B. L. Ebert, M. A.

Gillette, A. Paulovich, S. L. Pomeroy, T. R. Golub, E. S. Lander, J. P. Mesirov, Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl. Acad. Sci. U. S. A.* **102**, 15545–15550 (2005).

56. V. K. Mootha, C. M. Lindgren, K.-F. Eriksson, A. Subramanian, S. Sihag, J. Lehar, P. Puigserver, E. Carlsson, M. Ridderstråle, E. Laurila, N. Houstis, M. J. Daly, N. Patterson, J. P. Mesirov, T. R. Golub, P. Tamayo, B. Spiegelman, E. S. Lander, J. N. Hirschhorn, D. Altshuler, L. C. Groop, PGC-1alpha-responsive genes involved in oxidative phosphorylation are coordinately downregulated in human diabetes. *Nat. Genet.* **34**, 267–273 (2003).

57. D.-H. Lee, R. Blomhoff, D. R. Jacobs Jr, Is serum gamma glutamyltransferase a marker of oxidative stress? *Free Radic. Res.* **38**, 535–539 (2004).

58. S. Kathiresan, D. Srivastava, Genetics of human cardiovascular disease. *Cell*. **148**, 1242–1257 (2012).

59. E. C. Partridge, S. B. Chhetri, J. W. Prokop, R. C. Ramaker, C. S. Jansen, S.-T. Goh, M. Mackiewicz, K. M. Newberry, L. A. Brandsmeier, S. K. Meadows, C. L. Messer, A. A. Hardigan, C. J. Coppola, E. C. Dean, S. Jiang, D. Savic, A. Mortazavi, B. J. Wold, R. M. Myers, E. M. Mendenhall, Occupancy maps of 208 chromatin-associated proteins in one human cell type. *Nature*. **583**, 720–728 (2020).

60. C. Bellanné-Chantelot, C. Carette, J.-P. Riveline, R. Valéro, J.-F. Gautier, E. Larger, Y. Reznik, P.-H. Ducluzeau, A. Sola, A. Hartemann-Heurtier, P. Lecomte, L. Chaillous, M. Laloi-Michelin, J.-M. Wilhem, P. Cuny, F. Duron, B. Guerci, N. Jeandidier, H. Mosnier-Pudar, M. Assayag, D. Dubois-Laforgue, G. Velho, J. Timsit, The type and the position of HNF1A mutation modulate age at diagnosis of diabetes in patients with maturity-onset diabetes of the young (MODY)-3. *Diabetes*. **57**, 503–508 (2008).

61. S. Ellard, Hepatocyte nuclear factor 1 alpha (HNF-1α) mutations in maturity-onset diabetes of the young. *Hum. Mutat.* **16**, 377–385 (2000).

62. S. Althari, L. A. Najmi, A. J. Bennett, I. Aukrust, J. K. Rundle, K. Colclough, J. Molnes, A. Kaci, S. Nawaz, T. van der Lugt, N. Hassanali, A. Mahajan, A. Molven, S. Ellard, M. I. McCarthy, L. Bjørkhaug, P. R. Njølstad, A. L. Gloyn, Unsupervised Clustering of Missense Variants in HNF1A Using Multidimensional Functional Data Aids Clinical Interpretation. *Am. J. Hum. Genet.* **107**, 670–682 (2020).

63. L. A. Lotta, J. Mokrosiński, E. Mendes de Oliveira, C. Li, S. J. Sharp, J. 'an Luan, B. Brouwers, V. Ayinampudi, N. Bowker, N. Kerrison, V. Kaimakis, D. Hoult, I. D. Stewart, E. Wheeler, F. R. Day, J. R. B. Perry, C. Langenberg, N. J. Wareham, I. S. Farooqi, Human Gain-of-Function MC4R Variants Show Signaling Bias and Protect against Obesity. *Cell*. **177**, 597–607.e9 (2019).

64. Ž. Reiner, Hypertriglyceridaemia and risk of coronary artery disease. *Nat. Rev. Cardiol.* **14**, 401–411 (2017).

65. P. P. Toth, High-density lipoprotein and cardiovascular risk. *Circulation*. **109**, 1809–1812 (2004).

66. W. B. Kannel, K. Anderson, P. W. Wilson, White blood cell count and cardiovascular disease. Insights from the Framingham Study. *JAMA*. **267**, 1253–1256 (1992).

67. U. Sahin, K. Karikó, Ö. Türeci, mRNA-based therapeutics—developing a new class of drugs. *Nat. Rev. Drug Discov.* **13**, 759–780 (2014).

68. S. Kersten, Angiopoietin-like 3 in lipoprotein metabolism (2017), doi:10.1038/nrendo.2017.119.

69. M. Lupo, N. Ferri, Angiopoietin-Like 3 (ANGPTL3) and Atherosclerosis: Lipid and Non-Lipid Related Effects. *Journal of Cardiovascular Development and Disease*. **5**, 39 (2018).

70. J. D. Horton, J. C. Cohen, H. H. Hobbs, Molecular biology of PCSK9: its role in LDL metabolism. *Trends Biochem. Sci.* **32**, 71–77 (2007).

71. H. Sun, A. Samarghandi, N. Zhang, Z. Yao, M. Xiong, B.-B. Teng, Proprotein convertase subtilisin/kexin type 9 interacts with apolipoprotein B and prevents its intracellular degradation, irrespective of the low-density lipoprotein receptor. *Arterioscler. Thromb. Vasc. Biol.* **32**, 1585–1595 (2012).

72. M. J. Graham, K. M. Lemonidis, C. P. Whipple, A. Subramaniam, B. P. Monia, S. T. Crooke, R. M. Crooke, Antisense inhibition of proprotein convertase subtilisin/kexin type 9 reduces serum LDL in hyperlipidemic mice. *J. Lipid Res.* **48** (2007), doi:10.1194/jlr.C600025-JLR200.

73. M. Frank-Kamenetsky, A. Grefhorst, N. N. Anderson, T. S. Racie, B. Bramlage, A. Akinc, D. Butler, K. Charisse, R. Dorkin, Y. Fan, C. Gamba-Vitalo, P. Hadwiger, M. Jayaraman, M. John, K. N. Jayaprakash, M. Maier, L. Nechev, K. G. Rajeev, T. Read, I. Röhl, J. Soutschek, P. Tan, J. Wong, G. Wang, T. Zimmermann, A. de Fougerolles, H. P. Vornlocher, R. Langer, D. G. Anderson, M. Manoharan, V. Koteliansky, J. D. Horton, K. Fitzgerald, Therapeutic RNAi targeting PCSK9 acutely lowers plasma cholesterol in rodents and LDL cholesterol in nonhuman primates. *Proc. Natl. Acad. Sci. U. S. A.* **105** (2008), doi:10.1073/pnas.0805434105.

74. C. J. Duff, M. J. Scott, I. T. Kirby, S. E. Hutchinson, S. L. Martin, N. M. Hooper, Antibody-mediated disruption of the interaction between PCSK9 and the low-density lipoprotein receptor. *Biochem. J.* **419** (2009), doi:10.1042/BJ20082407.

75. I. K. Kotowski, A. Pertsemlidis, A. Luke, R. S. Cooper, G. L. Vega, J. C. Cohen, H. H. Hobbs, A spectrum of PCSK9 alleles contributes to plasma levels of low-density

lipoprotein cholesterol. *Am. J. Hum. Genet.* **78**, 410–422 (2006).

76. V. R. Shende, M. Wu, A. B. Singh, B. Dong, C. F. K. Kan, J. Liu, Reduction of circulating PCSK9 and LDL-C levels by liver-specific knockdown of HNF1α in normolipidemic mice. *J. Lipid Res.* **56**, 801–809 (2015).

77. B. Dong, A. B. Singh, V. R. Shende, J. Liu, Hepatic HNF1 transcription factors control the induction of PCSK9 mediated by rosuvastatin in normolipidemic hamsters. *Int. J. Mol. Med.* **39**, 749–756 (2017).

78. J. Wang, J. Zhao, C. Yan, C. Xi, C. Wu, J. Zhao, F. Li, Y. Ding, R. Zhang, S. Qi, X. Li, C. Liu, W. Hou, H. Chen, Y. Wang, D. Wu, K. Chen, H. Jiang, H. Huang, H. Liu, Identification and evaluation of a lipid-lowering small compound in preclinical models and in a Phase I trial. *Cell Metab.* **34**, 667–680.e6 (2022).

79. G. Dubuc, A. Chamberland, H. Wassef, J. Davignon, N. G. Seidah, L. Bernier, A. Prat, Statins upregulate PCSK9, the gene encoding the proprotein convertase neural apoptosis-regulated convertase-1 implicated in familial hypercholesterolemia. *Arterioscler. Thromb. Vasc. Biol.* **24**, 1454–1459 (2004).

80. J. Kim, C. Hu, C. Moufawad El Achkar, L. E. Black, J. Douville, A. Larson, M. K. Pendergast, S. F. Goldkind, E. A. Lee, A. Kuniholm, A. Soucy, J. Vaze, N. R. Belur, K. Fredriksen, I. Stojkovska, A. Tsytsykova, M. Armant, R. L. DiDonato, J. Choi, L. Cornelissen, L. M. Pereira, E. F. Augustine, C. A. Genetti, K. Dies, B. Barton, L. Williams, B. D. Goodlett, B. L. Riley, A. Pasternak, E. R. Berry, K. A. Pflock, S. Chu, C. Reed, K. Tyndall, P. B. Agrawal, A. H. Beggs, P. E. Grant, D. K. Urion, R. O. Snyder, S. E. Waisbren, A. Poduri, P. J. Park, A. Patterson, A. Biffi, J. R. Mazzulli, O. Bodamer, C. B. Berde, T. W. Yu, Patient-Customized Oligonucleotide Therapy for a Rare Genetic Disease. *N. Engl. J. Med.* **381**, 1644–1652 (2019).

81. S. Richards, N. Aziz, S. Bale, D. Bick, S. Das, J. Gastier-Foster, W. W. Grody, M. Hegde, E. Lyon, E. Spector, K. Voelkerding, H. L. Rehm, ACMG Laboratory Quality Assurance Committee, Standards and guidelines for the interpretation of sequence variants: a joint consensus recommendation of the American College of Medical Genetics and Genomics and the Association for Molecular Pathology. *Genet. Med.* **17**, 405–424 (2015).

82. T. Nishikawa, K. Hagihara, S. Serada, T. Isobe, A. Matsumura, J. Song, T. Tanaka, I. Kawase, T. Naka, K. Yoshizaki, Transcriptional complex formation of c-Fos, STAT3, and hepatocyte NF-1 alpha is essential for cytokine-driven C-reactive protein gene expression. *J. Immunol.* **180**, 3492–3501 (2008).

83. C. Toniatti, A. Demartis, P. Monaci, A. Nicosia, G. Ciliberto, Synergistic trans-activation of the human C-reactive protein promoter by transcription factor HNF-1 binding at two distinct sites. *EMBO J.* **9**, 4467–4475 (1990).

84. S. P. Li, N. D. Goldman, Regulation of human C-reactive protein gene expression by two synergistic IL-6 responsive elements. *Biochemistry*. **35**, 9060–9068 (1996).

85. A. R. Majithia, J. Flannick, P. Shahinian, M. Guo, M.-A. Bray, P. Fontanillas, S. B. Gabriel, GoT2D Consortium, NHGRI JHS/FHS Allelic Spectrum Project, SIGMA T2D Consortium, T2D-GENES Consortium, E. D. Rosen, D. Altshuler, Rare variants in PPARG with decreased activity in adipocyte differentiation are associated with increased risk of type 2 diabetes. *Proc. Natl. Acad. Sci. U. S. A.* **111**, 13127–13132 (2014).

86. P. Cingolani, A. Platts, L. L. Wang, M. Coon, T. Nguyen, L. Wang, S. J. Land, X. Lu, D. M. Ruden, A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster strain w1118; iso-2; iso-3. *Fly* . **6**, 80–92 (2012).

87. D. Klarin, Global Lipids Genetics Consortium, S. M. Damrauer, K. Cho, Y. V. Sun, T. M. Teslovich, J. Honerlaw, D. R. Gagnon, S. L. DuVall, J. Li, G. M. Peloso, M. Chaffin, A. M. Small, J. Huang, H. Tang, J. A. Lynch, Y.-L. Ho, D. J. Liu, C. A. Emdin, A. H. Li, J. E. Huffman, J. S. Lee, P. Natarajan, R. Chowdhury, D. Saleheen, M. Vujkovic, A. Baras, S. Pyarajan, E. Di Angelantonio, B. M. Neale, A. Naheed, A. V. Khera, J. Danesh, K.-M. Chang, G. Abecasis, C. Willer, F. E. Dewey, D. J. Carey, J. Concato, J. Michael Gaziano, C. J. O'Donnell, P. S. Tsao, S. Kathiresan, D. J. Rader, P. W. F. Wilson, T. L. Assimes, Myocardial Infarction Genetics (MIGen) Consortium, The Geisinger-Regeneron DiscovEHR Collaboration, The VA Million Veteran Program, Genetics of blood lipids among ~300,000 multi-ethnic participants of the Million Veteran Program. *Nature Genetics*. **50** (2018), pp. 1514–1523.

88. A. V. Khera, M. Chaffin, K. G. Aragam, M. E. Haas, C. Roselli, S. H. Choi, P. Natarajan, E. S. Lander, S. A. Lubitz, P. T. Ellinor, S. Kathiresan, Genome-wide polygenic scores for common diseases identify individuals with risk equivalent to monogenic mutations. *Nat. Genet.* **50**, 1219–1224 (2018).

89. X. Tan, C. Benedict, Increased Risk of Myocardial Infarction Among Patients With Type 2 Diabetes Who Carry the Common rs10830963 Variant in the MTNR1B Gene. *Diabetes Care*. **43** (2020), pp. 2289–2292.

90. N. L. Bray, H. Pimentel, P. Melsted, L. Pachter, Near-optimal probabilistic RNA-seq quantification. *Nat. Biotechnol.* **34**, 525–527 (2016).

91. M. I. Love, W. Huber, S. Anders, Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.* **15**, 1–21 (2014).

92. G. Korotkevich, V. Sukhov, A. Sergushichev, Fast gene set enrichment analysis. *Cold Spring Harbor Laboratory* (2019), p. 060012.

93. A. Liberzon, C. Birger, H. Thorvaldsdóttir, M. Ghandi, J. P. Mesirov, P. Tamayo, The Molecular Signatures Database Hallmark Gene Set Collection. *Cell Systems*. **1** (2015), pp. 417–425.

94. M. Ashburner, C. A. Ball, J. A. Blake, D. Botstein, H. Butler, J. M. Cherry, A. P.

Davis, K. Dolinski, S. S. Dwight, J. T. Eppig, M. A. Harris, D. P. Hill, L. Issel-Tarver, A. Kasarskis, S. Lewis, J. C. Matese, J. E. Richardson, M. Ringwald, G. M. Rubin, G. Sherlock, Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat. Genet.* **25**, 25–29 (2000).

95. Gene Ontology Consortium, The Gene Ontology resource: enriching a GOld mine. *Nucleic Acids Res.* **49**, D325–D334 (2021).

96. S. Heinz, L. Texari, M. G. B. Hayes, M. Urbanowski, M. W. Chang, N. Givarkes, A. Rialdi, K. M. White, R. A. Albrecht, L. Pache, I. Marazzi, A. García-Sastre, M. L. Shaw, C. Benner, Transcription Elongation Can Affect Genome 3D Structure. *Cell*. **174**, 1522–1536.e22 (2018).

97. O. Fornes, J. A. Castro-Mondragon, A. Khan, R. van der Lee, X. Zhang, P. A. Richmond, B. P. Modi, S. Correard, M. Gheorghe, D. Baranašić, W. Santana-Garcia, G. Tan, J. Chèneby, B. Ballester, F. Parcy, A. Sandelin, B. Lenhard, W. W. Wasserman, A. Mathelier, JASPAR 2020: update of the open-access database of transcription factor binding profiles. *Nucleic Acids Res.* **48**, D87–D92 (2019).

98. ENCODE Project Consortium, An integrated encyclopedia of DNA elements in the human genome. *Nature*. **489**, 57–74 (2012).

99. A. M. Deaton, M. M. Parker, L. D. Ward, A. O. Flynn-Carroll, L. BonDurant, G. Hinkle, P. Akbari, L. A. Lotta, A. Baras, P. Nioi, Gene-level analysis of rare variants in 379,066 whole exome sequences identifies an association of GIGYF1 loss of function with type 2 diabetes. *Sci. Rep.* **11**, 1–16 (2021).

100. C. J. Willer, Y. Li, G. R. Abecasis, METAL: fast and efficient meta-analysis of genomewide association scans. *Bioinformatics*. **26**, 2190–2191 (2010).

101. R. Dey, E. M. Schmidt, G. R. Abecasis, S. Lee, A Fast and Accurate Algorithm to Test for Binary Phenotypes and Its Application to PheWAS. *Am. J. Hum. Genet.* **101**, 37–49 (2017).

102. S. Purcell, S. S. Cherny, P. C. Sham, Genetic Power Calculator: design of linkage and association genetic mapping studies of complex traits. *Bioinformatics*. **19**, 149–150 (2003).

103. S. Purcell, B. Neale, K. Todd-Brown, L. Thomas, M. A. R. Ferreira, D. Bender, J. Maller, P. Sklar, P. I. W. de Bakker, M. J. Daly, P. C. Sham, PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am. J. Hum. Genet.* **81**, 559–575 (2007).

104. A. P. Reiner, M. J. Barber, Y. Guan, P. M. Ridker, L. A. Lange, D. I. Chasman, J. D. Walston, G. M. Cooper, N. S. Jenny, M. J. Rieder, J. Peter Durda, J. D. Smith, J. Novembre, R. P. Tracy, J. I. Rotter, M. Stephens, D. A. Nickerson, R. M. Krauss, Polymorphisms of the HNF1A Gene Encoding Hepatocyte Nuclear Factor-1α are Associated with C-Reactive Protein. *The American Journal of Human Genetics*. **82**

(2008), pp. 1193–1201.

CHAPTER 3

Plasma Proteome-wide Mendelian Randomization identifies RSPO3 as a mediator of NASH pathogenesis

Abstract

Nonalcoholic steatohepatitis (NASH) is a metabolic disorder characterized by fat accumulation, inflammation, and hepatocellular injury which currently lacks approved therapeutic interventions. To identify key causal proteins in NASH pathogenesis, we performed plasma proteome-wide Mendelian randomization (MR) using genetic determinants of plasma protein levels (n=4,907 proteins) identified in the deCODE genetics programs (n=35,559) and NASH biomarkers alanine aminotransferase (ALT), aspartate aminotransferase (AST), and liver fat accumulation measured in the UK Biobank cohort (n=488,377). We identified 36 proteins to have proteome-wide significant and directionally consistent causal effects on ALT and AST, 8 of which demonstrated consistent causal effects on liver fat. Among these identified proteins include R-spondin 3 (RSPO3), a secreted modulator of Wnt signaling pathways which causally increases ALT ($p=2.3 \times 10^{-18}$), AST ($p=3.5 \times 10^{-6}$) and liver fat ($p=5.3 \times 10^{-4}$). We further investigated the impact of RSPO3 on adiposity distribution, a recognized NASH risk factor, and found circulating RSPO3 causally increases waist-to-hip ratio adjusted for body mass index (WHRadjBMI, $p=4 \times 10^{-14}$), and subsequently increases visceral adipose tissue (VAT, $p=1 \times 10^{-18}$) while diminishing gluteofemoral adipose tissue (GFAT, $p=2 \times 10^{-10}$). Mediation analysis unveiled roughly half of RSPO3's effect on ALT and AST and 70% of its influence on liver fat is mediated through WHRadjBMI. This evidence suggests RSPO3 plays a role in NASH pathogenesis independently of its effect on adiposity, inviting further

investigations into the molecular mechanisms underpinning RSPO3's role in NASH particularly with regard to its tissue-specific effects in metabolically relevant tissues including adipose tissue and liver.

Introduction

Non-alcoholic fatty liver disease (NAFLD), defined by liver steatosis (fat accumulation), affects approximately 25% of the population globally(*1*). The more severe form of NAFLD, termed nonalcoholic steatohepatitis (NASH), is characterized by fat accumulation, inflammation, and hepatocellular injury(*2*), and cirrhosis resulting from NASH is the second leading cause of liver transplantation in America (*3*). At present, there are no approved pharmacological treatments for any stage of NAFLD, including NASH(*2*), underscoring the critical need to identify biomarkers for monitoring NAFLD disease progression and effective therapeutic targets.

Genetic studies in human populations have become a prominent resource for identifying and validating therapeutic targets(*4*). Association studies have yielded valuable insights into the genetic basis of NAFLD and NASH including nominating several key effector genes such as PNPLA3(*5*), TM6SF2(*6*), and HSD17B13(*7*). However the majority of genetic associations involve variants within noncoding regions(*8*), making identification of causal genes challenging. Furthermore, the directionality and causality of nominated effector genes on phenotype remains unknown without additional functional experimentation which is difficult to perform at scale.

In addition to genetic sequencing of large population cohorts, in recent years biobank studies have harnessed proteomic methods to assess protein levels in the blood of large numbers of individuals which has enabled association studies to identify genetic variants associated with altered protein levels (protein quantitative trait loci, pQTLs)(*9*). Examining the genetic component of circulating protein levels provides a direct method to assess the relationship between putative effector proteins and clinical outcomes.

Furthermore, proteome-wide Mendelian randomization (MR), which utilizes genetic predictors of protein levels to test the putative causal effects of protein change on phenotypes, is able to unveil directional and causal associations which are often unattainable from correlational analyses.

To identify and characterize proteins causally implicated in NASH pathogenesis, we systematically estimated the putative causal role of 1,831 plasma proteins measured in 35,559 individuals on NASH biomarkers by employing a comprehensive MR approach utilizing genetic determinants of circulating protein levels (pQTLs). From this proteome-wide scan, we prioritize RSPO3 as a NASH effector protein and find RSPO3 to be causally associated with increased NASH biomarkers ALT, AST, and liver fat as well as metabolically harmful adiposity distribution profiles. We further apply statistical mediation techniques to disentangle the direct effects of RSPO3 on NASH from the effects on fat distribution and provide insights into the molecular underpinnings pertaining to the role of RSPO3 in NASH pathogenesis.

Results

Proteome-wide Mendelian randomization implicates RSPO3 as a causal factor for NASH

To prioritize circulating proteins which may have a causal role in NASH pathogenesis, we first identified plasma proteins measured in the population-based deCODE genetics study(*10*) which have valid genetic instruments predicting circulating levels. In brief, we scanned within a 1 megabase window around each protein-coding gene, selecting cis-acting pQTLs as genetic instruments for MR analysis as such are more likely to have effects specific to the selected protein compared to trans-acting

pQTLs(*11*). From the 4,907 measured proteins among 35,559 individuals in deCODE, we identified 39,735 independent pQTLs across 1,831 proteins.

In order to estimate the putative causal effects of these 1,831 proteins on NASH, we undertook two-sample MR quantifying the effects of each protein's genetic instruments on NASH biomarkers ALT, AST, and liver fat using genome-wide association results from the UK Biobank population (n=488,377)(*12*). We first tested each protein with ALT and AST as these phenotypes were more powered for association testing given the vastly greater samples size (n=420,449) compared to participants with liver fat quantification (n=36,703)(*13*). Of the 1,831 proteins, 1,804 had valid instruments to perform MR for ALT and AST (meaning at least one genetic instrument per protein was present upon harmonization with the outcome genetic association study, Methods), and 1,150 (66%) of the assessed proteins have three or more pQTLs found in the cis window. Upon conducting two-sample MR using a inverse variance-weighted (IVW) model to estimate effects, we identified 36 proteins which were significantly associated in the same direction with ALT and AST upon adjustment for the number of proteins assessed ($p < 0.05 / 1804$, Figure 1A, Table 3.1).

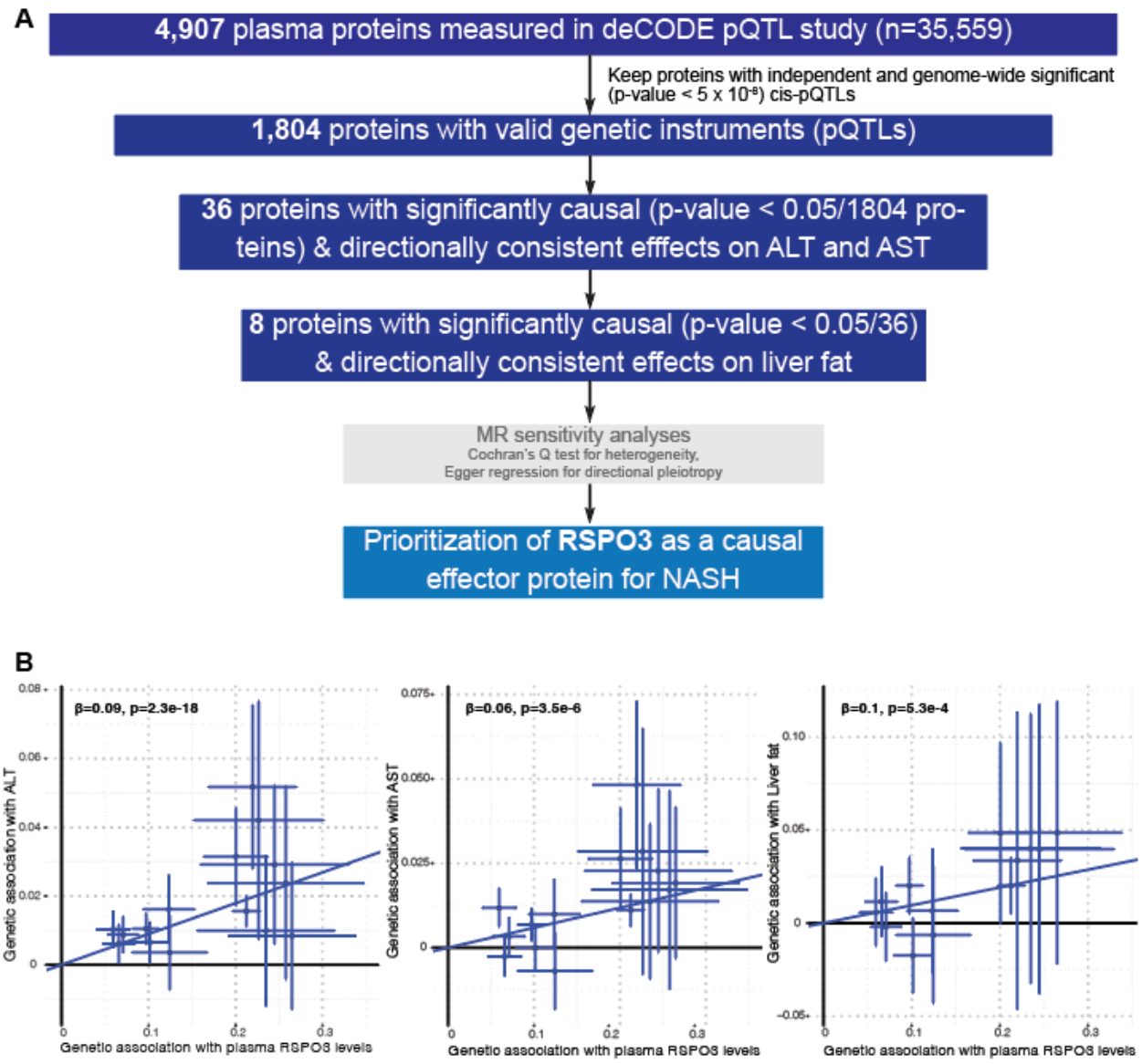


Figure 3.1. Proteome-wide Mendelian randomization implicates RSPO3 as a causal factor for NASH.

pQTL, protein quantitative trait loci; NASH, nonalcoholic steatohepatitis; ALT, alanine aminotransferase; AST, aspartate aminotransferase.

A) Overview of proteome-wide Mendelian randomization approach to identify causal NASH effector proteins.

B) MR plots displaying the genetic associations of instrumental variables determined from the deCODE proteomics study (n=15 variants) with RSPO3 levels and NASH biomarkers ALT (left), AST (middle), and liver fat (right) in the UK Biobank. Error bars represent the standard error of each variant's genetic association with either the exposure (RSPO3 levels) or outcome (NASH phenotype). Overlaid upon each plot is a line representing the inverse variance-weighted estimate along with numerical estimates (beta) and p-values.

Table 3.1. Secreted proteins with significant causal effects on NASH biomarkers.
36 proteins significantly associated in concordant directions with alanine aminotransferase (ALT) and aspartate aminotransferase (AST) upon adjustment for the number of proteins assessed ($p < 0.05 / 1804$). Protein name and description are given along with the number of SNPs used for each protein in the Mendelian randomization (MR) analyses. MR effect size and P-values for ALT, AST, and liver fat are given. Protein names in bold represent the 8 proteins also significantly associated with liver fat ($p < 0.05 / 36$) with consistent directionality of effect as with liver enzymes.

			Protein level's effect on ALT		Protein level's effect on AST		Protein level's effect on liver fat	
Protein	Description	nSNP	Effect size	P-value	Effect size	P-value	Effect size	P-value
MCL1	myeloid cell leukemia sequence 1	5	0.086	3.1E-53	0.067	4.5E-32	0.0821	1.8E-06
RSPO3	R-spondin 3	15	0.088	2.3E-18	0.057	3.5E-06	0.0953	5.3E-04
GUSB	glucuronidase, beta	19	0.056	3.1E-12	0.064	3.0E-11	0.1830	4.9E-13
APOH	apolipoprotein H (beta-2-glycoprotein I)	15	-0.035	1.3E-16	-0.029	1.7E-10	-0.0535	1.5E-04
BTN3A3	butyrophilin, subfamily 3, member A3	102	0.012	1.1E-12	0.034	1.7E-42	0.0150	3.7E-04
F13B	coagulation factor XIII, B polypeptide	58	-0.027	6.6E-11	-0.023	1.4E-07	-0.0380	9.9E-04
NCAN	neurocan	10	-0.193	7.5E-07	-0.150	7.6E-08	-0.5649	1.2E-06
RAB21	RAB21, member RAS oncogene family	1	-0.129	4.6E-06	-0.224	6.6E-11	-0.3178	5.3E-04
HP	haptoglobin	132	-0.025	2.5E-23	-0.029	4.2E-34	-0.0173	3.6E-03
PLA2G12B	phospholipase A2, group XIIB	6	0.136	8.1E-18	0.086	4.3E-08	0.0353	4.6E-01
APOL3	apolipoprotein L, 3	67	0.033	2.3E-17	0.026	3.4E-14	-0.0036	7.0E-01
CCBL2	cysteine conjugate-beta lyase 2	29	0.045	2.7E-17	0.065	1.6E-17	0.0166	3.3E-01
HGFAC	HGF activator	87	-0.020	4.0E-15	-0.019	7.7E-11	0.0016	8.2E-01
ECM1	extracellular matrix protein 1	78	0.029	1.9E-14	0.038	2.3E-19	-0.0025	7.4E-01
CFHR5	complement factor H-related 5	55	-0.022	2.2E-14	-0.016	1.1E-05	-0.0128	1.3E-01
SERPINA1	serpin peptidase inhibitor, A member 1	51	-0.035	5.1E-14	-0.027	1.8E-07	0.0312	3.8E-03

Table 3.1 (continued). Secreted proteins with significant causal effects on NASH biomarkers.

			Protein level's effect on ALT		Protein level's effect on AST		Protein level's effect on liver fat	
Protein	**Description**	**nSNP**	**Effect size**	**P-value**	**Effect size**	**P-value**	**Effect size**	**P-value**
FLRT3	fibronectin leucine rich transmembrane protein 3	100	-0.014	8.2E-14	-0.013	2.8E-12	-0.0019	8.0E-01
HAPLN4	hyaluronan and proteoglycan link protein 4	2	-0.166	3.3E-11	-0.160	5.5E-10	-0.1214	9.9E-02
OGN	osteoglycin	49	-0.024	6.0E-09	-0.030	4.2E-16	-0.0212	4.6E-02
APOC1	apolipoprotein C-I	17	0.053	1.3E-07	0.063	4.7E-19	-0.0461	1.2E-01
LYZ	lysozyme	64	0.020	1.7E-07	0.025	1.1E-07	-0.0249	4.7E-02
GAS6	growth arrest-specific 6	17	-0.038	2.0E-07	-0.071	8.3E-09	-0.0006	9.8E-01
LCT	lactase	52	0.007	3.1E-07	0.012	2.0E-16	-0.0003	9.6E-01
CFHR1	complement factor H-related 1	81	0.011	3.7E-07	0.023	5.8E-27	0.0013	8.4E-01
NEK7	NIMA-related kinase 7	5	-0.051	5.0E-07	-0.081	4.8E-13	0.1853	6.7E-04
CFL2	cofilin 2 (muscle)	1	-0.109	1.0E-06	-0.125	1.6E-07	-0.0649	3.6E-01
BTN3A1	butyrophilin, subfamily 3, member A1	2	0.191	2.2E-06	0.541	1.2E-12	0.2720	1.7E-02
PTPRJ	protein tyrosine phosphatase, receptor type, J	10	0.074	3.4E-06	0.137	1.2E-06	-0.0039	9.3E-01
MANEA	mannosidase, endo-alpha	66	-0.009	3.9E-06	-0.011	1.9E-07	0.0011	9.0E-01
COL10A1	collagen, type X, a1	24	0.039	4.0E-06	0.046	1.6E-05	-0.0006	9.8E-01
CD248	CD248 molecule, endosialin	3	0.229	5.5E-06	0.251	2.7E-06	-0.2440	1.1E-01
NEO1	neogenin 1	10	-0.071	7.7E-06	-0.101	3.2E-07	-0.0607	2.3E-02
ASPN	asporin	54	0.020	1.3E-05	0.029	1.8E-11	0.0129	2.1E-01
IL1RN	interleukin 1 receptor antagonist	9	0.067	1.5E-05	0.098	1.2E-10	-0.0323	3.0E-01
HYAL1	hyaluronoglucosaminidase 1	14	-0.041	1.8E-05	-0.046	5.0E-06	-0.0325	3.1E-01
FCRLB	Fc receptor-like B	30	0.029	2.2E-05	0.054	1.8E-08	0.0079	5.9E-01

To provide additional evidence to prioritize NASH effector proteins, we next examined the causal effect of these 36 ALT/AST associated proteins on liver fat quantified by MRI in the UK Biobank(*13*). Using consistent methodology as before, MR analysis identified 8 proteins which were significantly associated ($p < 0.05 / 36$ proteins tested) with liver fat in the concordant direction as ALT and AST (bold genes, Table 3.1). Given the strong and directionally consistent causal associations, these 8 proteins were prioritized as putative causal effector proteins for NASH.

We next performed sensitivity analyses to confirm the robustness of the MR analyses (Figure 3.1A). First, we assessed for indications of heterogeneity quantified by the Cochran's Q test(*14*). Among the 8 prioritized proteins, for the following three proteins we found no indication of heterogeneity in the analyses of either ALT or AST and liver fat: MCL1 (Cochran's Q P-values: $P_{ALT} = 0.51$, $P_{liver\ fat} = 0.88$), RSPO3 ($P_{ALT} = 0.08$, $P_{liver\ fat} = 0.36$), and GUSB ($P_{ALT} = 0.16$, $P_{liver\ fat} = 0.58$). Genetic instruments can have pleiotropic effects on outcome phenotypes and subsequently introduce bias into causal estimates determined by MR analysis(*15*). To correct for potential directional pleiotropy, we next assessed the causal effects of these 3 proteins without evidence of heterogeneity on NASH biomarkers using Egger regression MR, a method which has lower power of detection but can provide a consistent estimate of the causal effect(*15*). Of the 3 proteins, only for RSPO3 did the determined effects on all NASH biomarkers remain significant through Egger regression (ALT: b=0.07, p=0.003; AST: b=0.07, p=0.02; Liver fat: b=0.13, p=0.03). From this comprehensive proteome-wide analysis, we prioritized RSPO3 as a causal effector protein for NASH (Figure 3.1A).

RSPO3 is a secreted protein of the R-spondin family which potentiates Wnt ligand activity as either a coactivator or inhibitor of Wnt signaling receptor degradation(*16*). From our MR analyses, we determine circulating levels of RSPO3 causally increase NASH biomarkers as evidenced by significantly positive MR estimates of RSPO3 on ALT ($b=0.09$, $p=2.3 \times 10^{-18}$), AST ($b=0.06$, $p=3.5 \times 10^{-6}$), and liver fat ($b=0.1$, $p=5.3 \times 10^{-4}$, Figure 3.1B). To eliminate the possibility of spurious results due to reverse causation, we performed bi-directional MR using genetic determinants of NASH outcomes on RSPO3 and found no evidence of reverse causation for ALT ($p=0.1$) and liver fat ($p=0.5$), however AST showed nominal significance for causal effects on RSPO3 levels ($p=0.02$) though this does not meet statistical significance upon multiple hypothesis correction to account for the three MR analyses conducted. Furthermore, we found no indication that the genetic determinants of RSPO3 levels used as instrumental variables in the discovery MR analyses had horizontal pleiotropy as assessed by the MR-Egger intercept ($P_{Egger\text{-}intercept} > 0.05$ for ALT, AST, and liver fat).

RSPO3 is causally associated with metabolically harmful fat distribution profiles

The RSPO3 locus has been found to be significantly associated with waist-to-hip ratio adjusted for body mass index (WHRadjBMI) and has been shown to regulate adipocyte biology in vitro(*17*). Given this implication of RSPO3 in adipose tissue and in light of the well-established connection between central obesity and NASH(*18*), we next sought to characterize the causal effects of RSPO3 on adiposity distribution in humans. Using consistent methodology as previous two-sample MR analyses, we examined the causal effects of RSPO3 on WHRadjBMI using the genetic determinants of plasma

RSPO3 levels and results from the largest WHRadjBMI genome-wide association study to date (GIANT consortium, n=694,649)(*19*). We found circulating RSPO3 levels were causally associated with increased WHRadjBMI (b=0.19, $p=3.9 \times 10^{-14}$, Figure 3.2A). The determined effect remained consistent when analyzing waist-hip ratio without adjustment for body mass index (b=0.12, $p=1.3 \times 10^{-13}$). We found no evidence of horizontal pleiotropy ($P_{Egger\text{-}intercept} = 0.75$) as well as no evidence of reverse causation by bi-directional MR of WHRadjBMI on RSPO3 levels (p=0.31).

Though WHR is a simple and widely available estimate of metabolic risk, it cannot distinguish between subcutaneous and visceral abdominal fat. On the other hand, direct measurements of adiposity via visualization provides a more accurate assessment of regional fat depots. In order to provide additional support for the causal effects of RSPO3 on body fat distribution, we conducted MR analyses assessing the effect of RSPO3 levels on MRI-derived visceral (VAT), gluteofemoral (GFAT), and abdominal subcutaneous (ASAT) adipose tissue volumes measured in 38,965 UK Biobank participants(*20*). MR analysis revealed circulating RSPO3 levels causally increased VAT (b=0.1, $p=6 \times 10^{-4}$) and decreased GFAT (b=-0.15, $p=1.4 \times 10^{-4}$) but had no significant effect on ASAT (p=0.7). These findings corroborate our previous result regarding the positive effect of RSPO3 on WHRadjBMI, as elevated WHRadjBMI typically indicates a higher distribution of abdominal fat (waist circumference) and less GFAT (hip circumference)(*21*).

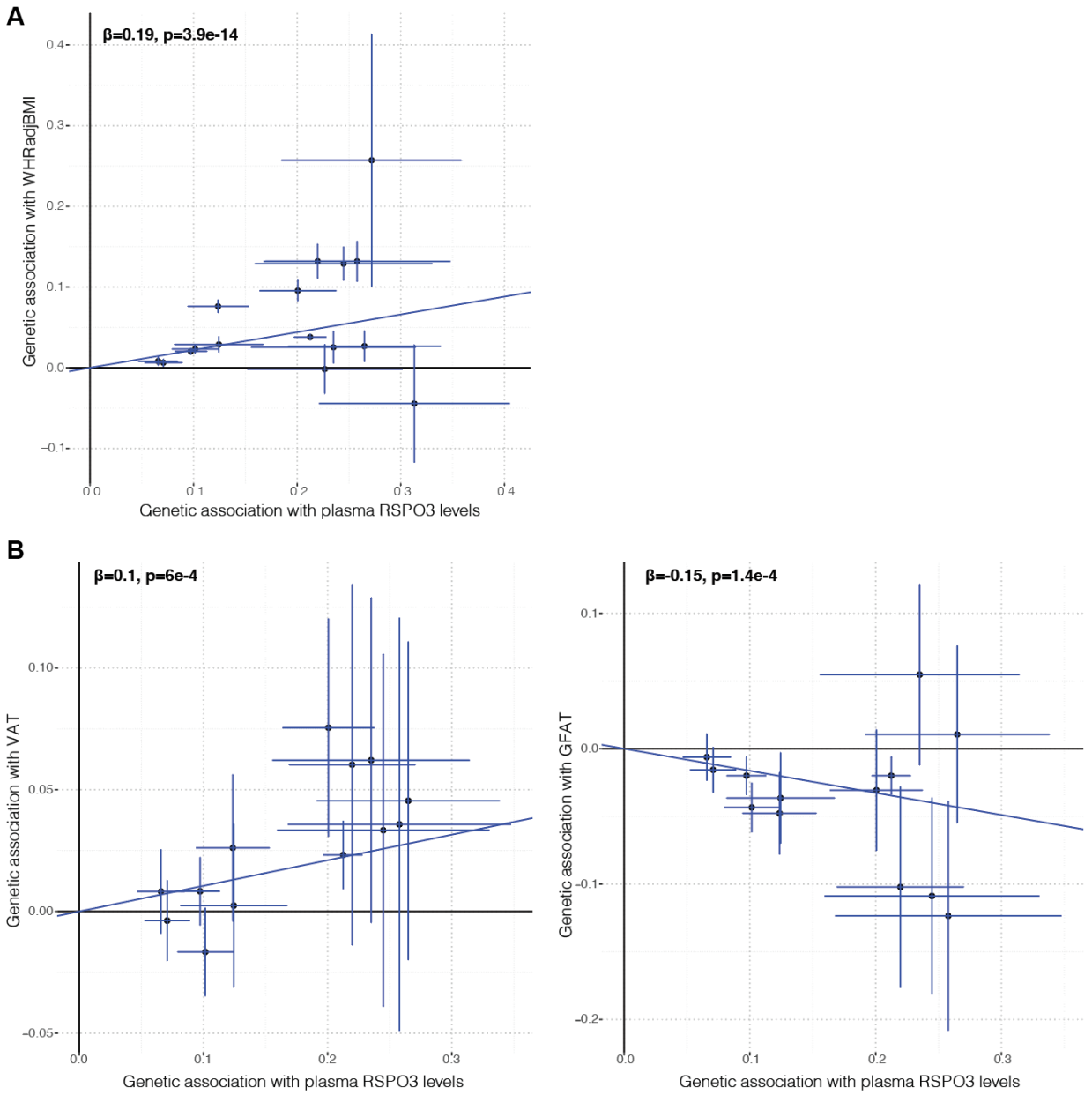


Figure 3.2. RSPO3 is causally associated with metabolically harmful fat distribution profiles.

WHRadjBMI, waist-to-hip ratio adjusted for body mass index; VAT, visceral adipose tissue; GFAT, gluteofemoral adipose tissue.

MR plots displaying the genetic associations of instrumental variables determined from the deCODE proteomics study with RSPO3 levels and A) WHRadjBMI in the GIANT consortium study (n=694,649) and B) VAT (left) and GFAT (right) measured by MRI in the UK Biobank (n=38,965). Error bars represent the standard error of each variant's genetic association with either the exposure (RSPO3 levels) or outcome (adiposity measurement). Overlaid upon each plot is a line representing the inverse variance-weighted estimate along with numerical estimates (beta) and p-values.

Effect of RSPO3 on NASH is mediated in part through waist-hip ratio

In light of our finding that RSPO3 has causal effects on adiposity distribution, a known risk factor for metabolic syndrome(*22*), we next investigated the direct and indirect effects via WHRadjBMI of RSPO3 on NASH. To assess this, we performed mediation analysis using two-step MR(*23*), quantifying the effects of RSPO3 (exposure) on NASH biomarkers (outcomes) via WHRadjBMI (mediator, Figure 3.3A). We employed the product method to estimate the indirect effect of RSPO3 on outcomes via WHRadjBMI and the delta method to estimate standard error and confidence intervals (Methods). The proportion of mediation effect of RSPO3 via WHRadjBMI on ALT and AST is about one-half (ALT = 53%, AST = 54%), while the indirect effect of RSPO3 on liver fat via WHRadjBMI is 70%.

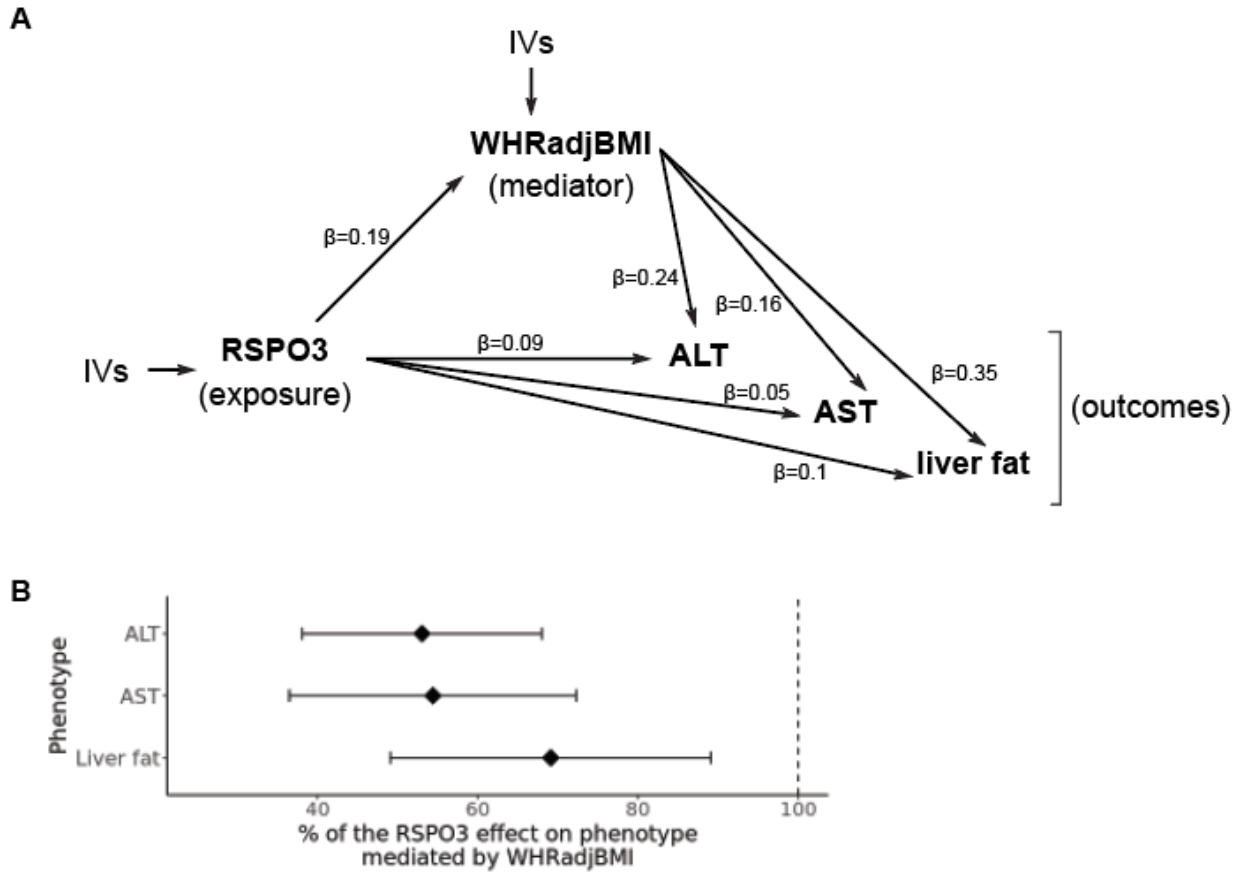


Figure 3.3. Effect of RSPO3 on NASH is mediated in part through waist-hip ratio.
WHRadjBMI, waist-to-hip ratio adjusted for body mass index; ALT, alanine aminotransferase; AST, aspartate aminotransferase
A) Mediation analysis to quantify the effects of RSPO3 on NASH phenotypes via WHRadjBMI. Two-sample MR analyses were conducted to quantify the effects of exposure on outcomes, exposure on mediator, and mediator on outcomes. Effect for each respective MR analysis are shown (beta). B) Indirect effect of RSPO3 effect on NASH phenotypes mediated by WHRadjBMI. Shown are the proportions of effect via mediator and 95% confidence intervals determined by the product method and delta method respectively.

Discussion

In recent years, biobank studies have employed high-throughput proteomic techniques enabling association studies to identify genetic determinants of circulating protein levels and providing new opportunities to identify novel disease effector proteins. In this study we systematically estimated the putative causal role of 1,831 proteins measured in the deCODE genetics consortium on NASH phenotypes (ALT, AST, and liver fat) measured in the UK Biobank. Through proteome-wide Mendelian randomization we prioritized RSPO3 as a causal effector protein which elevates biomarkers of NASH and confers a metabolically harmful adipose distribution profile. By conducting mediation analysis we find that approximately half of the effect of RSPO3 on liver enzymes ALT and AST and approximately 70% of the effect on liver fat is mediated by WHRadjBMI.

Here we present the first large-scale Mendelian randomization analysis using the human plasma proteome to identify effector proteins for NASH. Though previous studies have adopted similar approaches harnessing genetic determinants of circulating protein levels to discover putative effector proteins and therapeutic targets(*24–26*), this study is the first to our knowledge to investigate the causal effects of proteins specifically on NASH phenotypes using human genetics at the proteome-wide scale. From this systematic investigation we prioritized the Wnt signaling activator RSPO3 as the top causal effector protein for NASH. The RSPO3 locus has previously been implicated in WHRadjBMI from human genetic association studies however followup investigations have been limited to single variant characterization(*17*, *27*). Rather than investigate individual variants which provide low statistical power and often exert weak effects(*28*), in this work we leverage genetic variation across the whole RSPO3 locus to determine directional effects (all cis

RSPO3 pQTLs). Furthermore, functional characterization of RSPO3 to date has only been conducted *in vitro* or in animal models(*17*, *27*) however in this study we determine causal effects of RSPO3 on NASH phenotypes directly in humans.

By employing robust MR analyses, we determine RSPO3 causally increases NASH biomarkers suggesting circulating RSPO3 could be a potential NASH biomarker for early detection or inhibition of RSPO3 may be a promising therapeutic strategy for NASH. We further demonstrate RSPO3 causally increases WHRadjBMI and concordantly increases VAT and diminishes GFAT. We did not observe any significant effects of RSPO3 on ASAT demonstrating the increasing WHRadjBMI is driven predominantly by effects on VAT and GFAT congruent with WHR being a surrogate measure of enhanced visceral adiposity(*29*). Prior functional characterization *in vitro* involving RSPO3 knockdown and treatment of recombinant human RSPO3 treatment in visceral and gluteal adipose progenitor cells has shown RSPO3 predominantly inhibits gluteal adipose progenitor proliferation while enhancing proliferation of abdominal adipose progenitors(*17*). In this work, we corroborate these *in vitro* findings directly in humans via a population genetics approach and further implicate RSPO3 in regulating fat distribution.

Central obesity has been recognized as a metabolic risk factor for cardiometabolic and fatty liver disease(*30*). Employing statistical mediation techniques we determine approximately half of the effect of RSPO3 on liver enzymes is mediated by effects on adiposity indicating RSPO3 in part has a direct effect on NASH pathogenesis. On the other hand for liver fat accumulation RSPO3 exhibits a smaller direct effect, suggesting RSPO3 through regulation of adipose specific mechanisms plays a relatively greater role

on liver fat accumulation compared to ALT/AST. Taken together, the evidence suggests RSPO3 may act in part through adipose tissue but also acts through independent mechanisms and tissues to elevate liver enzymes and potentially accelerate the progression of fatty liver to NASH.

Liver enzymes are well-known to correlate with NASH and fibrosis(*31*). The greater direct effect of RSPO3 on ALT and AST compared to liver fat alone suggests RSPO3 may play a causal role in developing hallmarks of NASH besides liver fat accumulation. For instance, liver inflammation and fibrosis are often measured via liver enzymes, particularly ALT and AST, and non-invasive clinical measures of fibrosis are predominantly derived from ALT and AST levels(*32*). RSPO3 may play a role in fibrosis via amplification of Wnt/β-catenin signaling, a process that has been implicated in the development of fibrosis(*33*). Inhibition of RSPO3 in mice via commercially available antibody has been shown to attenuate liver fibrosis as a result of activated β-catenin(*27*) however these results have not yet been replicated nor assessed in humans.

One limitation of our study is the availability of large-scale population cohorts with direct diagnosis of metabolic liver disease. The gold standard diagnosis of NASH is via liver biopsy however such are difficult to obtain at scale, therefore population-based studies predominantly rely on surrogates such as liver fat content and liver enzyme markers. While these proxies have been found to be correlated with NASH(*34*), the development of noninvasive NASH biomarkers still has great need for improvement. As larger biopsied cohorts are curated and next generation noninvasive NASH diagnosis metrics are developed, we expect to see better-powered and more accurate

NAFLD/NASH GWAS which will enhance our ability to investigate the effects of RSPO3 on these conditions more conclusively.

In summary, we find RSPO3 plays a pivotal role in NASH pathogenesis in part via adipose specific mechanisms. This work lays the foundation for future research to investigate the unknown tissue and cell-type specific mechanisms by which RSPO3 directly affects NASH.

Future directions

Taken together, our results and prior findings suggest that RSPO3 may play a critical role in NASH pathogenesis however it is currently unclear if RSPO3 exerts effects in an endocrine, paracrine, or autocrine manner. Though RSPO3 circulates systemically, it is currently unclear which tissue is the main source of the secreted RSPO3 which we find causally increases NASH. Among the tissues known to be involved in NASH pathogenesis (adipose, liver, and muscle), RSPO3 expression is the highest in visceral adipose tissue (median transcripts per million (TPM) = 40.57, GTEx(*35*)) compared to subcutaneous adipose (TPM = 5.4), liver (TPM = 2.2), and skeletal muscle (TPM = 8.5). In light of our adiposity related findings, we hypothesize that RSPO3 may be secreted into the blood primarily from visceral adipose tissue. To test this hypothesis, ongoing efforts are underway to quantify the levels of RSPO3 in conditioned media from human visceral adipose tissue biopsies.

The molecular mechanisms by which RSPO3 acts on the liver to exacerbate NASH have yet to be understood. Although RSPO3 may be secreted from another tissue to exert effects on the liver, it is also possible cells within the liver may produce RSPO3 itself.

Inspection of single cell RNA-sequencing of human liver(*36*) reveals RSPO3 is expressed primarily in liver sinusoidal endothelial cells (LSECs) and hepatic stellate cells (HSCs) (Figure 3.4A). RSPO3 has also been found to be an endothelial specific marker in mice specifically in the central vein compartment of the liver(*37*). LSECs play a critical role in liver homeostasis, forming the interface between blood and intra hepatic cells, regulating hepatic vascular tone, and maintaining HSCs quiescence(*38*). As the main effector cell type of fibrosis development, activation of HSCs results in a downstream cascade involving secretion of fibrogenic factors(*39*).

Though RSPO3 may reach the liver either via secretion from other tissues or production within LSECs or HSCs, it is still unclear how RSPO3 exert effects on the liver. Leucine Rich Repeat Containing G Protein-Coupled Receptor 4 and 5 (LGR4 and LGR5), the identified receptors of RSPO3, bind to RSPO3 and potentiate Wnt/β-catenin signaling(*40*). In bulk transcriptomic profiling across tissues, LGR5 is barely expressed in the liver (TPM=0.3, GTEx(*35*)) while LGR4 is ubiquitously expressed across tissues. Single cell RNA-sequencing of human liver(*36*) reveals LGR4 is predominantly expressed in hepatocytes as well as cholangiocytes, macrophages, HSCs and LSECs (Figure 3.4B). Given multiple cell types may be involved, ongoing research is focused on examining the effect of altered RSPO3 in a physiologically represented liver system which encompasses the dynamic relationships between different cell types. To accomplish this, research is currently underway to develop liver spheroids, a hetero-cellular *in vitro* system comprised of primary human liver cell types including hepatocytes and non-parenchymal cells, to understand the complex cell-type specific interplay of RSPO3 in liver to cause NASH.

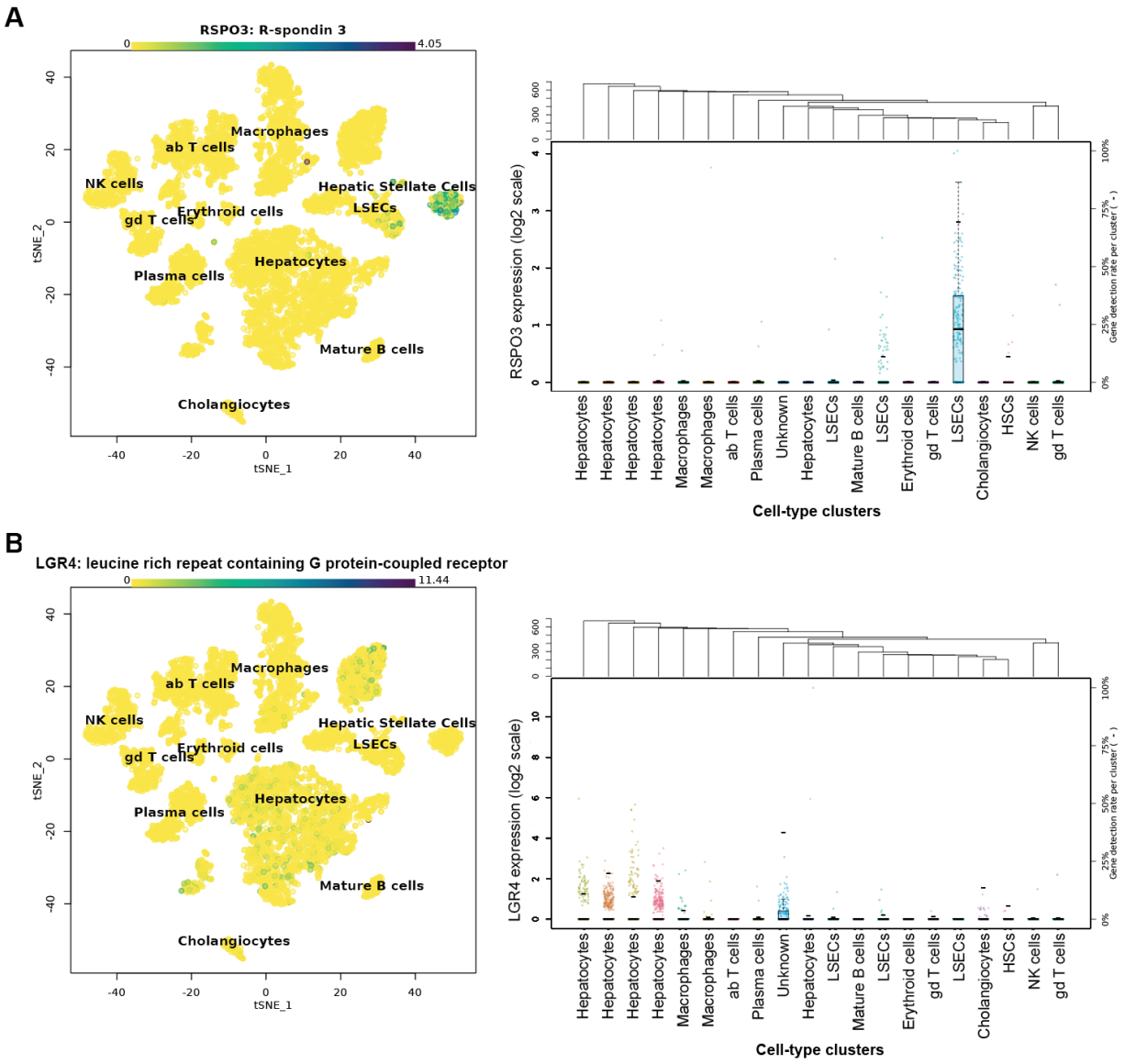


Figure 3.4 RSPO3 and LGR4 expression in single cell RNA-sequencing of human liver.
Cell-type specific expression of A) RSPO3 and its receptor B) LGR4 determined by single cell RNA-sequencing of human livers (n=5). (left) tSNE plots of cells with cluster assignments. Color of each cell represents gene expression, with yellow representing none/low expression and blue to purple representing higher expression. (right) Gene expression distributions per cluster shown in tSNE plots. Cell type clusters are ordered by heatmap dendrogram of clustering cell types (top). Scatter plots represent gene expression distribution of individual cells with each cluster colored differently. Detection rate (n cells in which gene is detected / n cells in each cluster) per cluster is indicated as a black bar with corresponding proportions listed along the right y-axis.

Methods

Mendelian Randomization of NASH biomarkers

Instrument selection and outcomes

Valid instrumental variables to represent altered plasma protein abundance from the deCODE proteomics association study(*10*) were defined using variants associated in cis (within 1Mb of the protein coding region) at genome-wide significance ($p<5\times10^{-8}$). LD clumping ($r^2 < 0.01$) was performed to ensure pQTLs were independent and SNPs with F-statistics < 10 were removed to avoid weak instrument bias.

Outcomes of interest included NASH biomarkers (ALT, AST, liver fat, WHRadjBMI, VAT, GFAT, and ASAT). Data for liver enzyme outcomes ALT and AST were derived from genome-wide association studies in the UK Biobank. The UK Biobank is a prospective cohort study with genotypic and phenotypic data that enrolled approximately 500,000 individuals aged 40-69 from across the United Kingdom(*12*). For quality control of SNPs from the UK Biobank genotypes, only variants with MAF > 0.001, HWE $p > 1 \times 10^{-10}$ were included, and for imputed SNPs only those with INFO scores > 0.8 were included. Linear regressions correcting for covariates age, age^2, sex, and the first 20 principal components of ancestry were run to determine genome-wide associations for ALT and AST. Data for the remaining NASH biomarkers were extracted from the following studies: liver fat(*13*), WHRadjBMI(*19*), and adipose depots (VAT, GFAT, and ASAT)(*20*). Notably, only ALT and AST were analyzed at the proteome-wide levels, while 38 proteins were examined for effects on liver fat and only RSPO3 was analyzed for effects on adiposity phenotypes (WHRadjBMI and adipose depot measurements).

Mendelian randomization analysis

pQTL effects were harmonized with outcome association statistics to ensure the effect allele was consistent in all studies. If a genetic instrument was not present in the outcome association results, a proxy SNP (r^2>0.8) was used as a replacement if available. We performed two-sample MR using the R package MendelianRandomization v.0.6.0(*41*) and the delta weighted inverse variance weighted (IVM, mr_ivm function) model to meta-analyze the individual variant effects and estimate the causal relationship of genetically determined protein abundance on genetic associations with outcomes of interest.

Sensitivity analyses

Cochran's Q statistics to assess heterogeneity and Egger regressions to correct for horizontal pleiotropy were computed using the mr_ivm and mr_egger functionality as instantiated in MendelianRandomization v.0.6.0(*41*). Reverse causation was assessed using bi-directional MR adopting congruent methodology as the discovery MR analyses but instead using NASH phenotypes as the exposure and plasma RSPO3 levels as the outcome. In brief, valid instrumental variables to represent the NASH phenotypes were defined using genome-wide significant variants ($p<5\times10^{-8}$) within each respective phenotype's association study. As done previously, we then performed LD clumping, removed SNPs with F-statistics < 10, harmonized variants with RSPO3 pQTL association statistics, identified proxy SNPs if available, and conducted two-sample MR.

Mediation analyses

MR analyses of WHRadjBMI on outcomes ALT, AST, and liver fat for subsequent mediation analyses were conducted using consistent methodology as described. In brief, valid instrumental variables to represent the mediator WHRadjBMI were defined using variants associated with WHRadjBMI at genome-wide significance ($p<5\times10^{-8}$) within the GIANT consortium(*19*). As done previously, we then performed LD clumping, removed SNPs with F-statistics < 10, harmonized variants with outcome association statistics, identified proxy SNPs if available, and conducted two-sample MR.

To determine the indirect effect of the mediator (WHRadjBMI) on the outcome (ALT, AST, liver fat), the effect estimate of the exposure (RSPO3) on the mediator and the effect estimate of the mediator on the outcome were multiplied together (product method)(*23*). Standard errors and 95% confidence intervals were derived using the delta method as done previously(*25*, *42*).

References

1. Z. M. Younossi, A. B. Koenig, D. Abdelatif, Y. Fazel, L. Henry, M. Wymer, Global epidemiology of nonalcoholic fatty liver disease-Meta-analytic assessment of prevalence, incidence, and outcomes. *Hepatology*. **64**, 73–84 (2016).

2. N. Chalasani, Z. Younossi, J. E. Lavine, M. Charlton, K. Cusi, M. Rinella, S. A. Harrison, E. M. Brunt, A. J. Sanyal, The diagnosis and management of nonalcoholic fatty liver disease: Practice guidance from the American Association for the Study of Liver Diseases. *Hepatology*. **67**, 328–357 (2018).

3. R. J. Wong, A. K. Singal, Trends in Liver Disease Etiology Among Adults Awaiting Liver Transplantation in the United States, 2014-2019. *JAMA Netw Open*. **3**, e1920294 (2020).

4. C. Finan, A. Gaulton, F. A. Kruger, R. T. Lumbers, T. Shah, J. Engmann, L. Galver, R. Kelley, A. Karlsson, R. Santos, J. P. Overington, A. D. Hingorani, J. P. Casas, The druggable genome and support for target identification and validation in drug development. *Sci. Transl. Med.* **9** (2017), doi:10.1126/scitranslmed.aag1166.

5. S. Romeo, J. Kozlitina, C. Xing, A. Pertsemlidis, D. Cox, L. A. Pennacchio, E. Boerwinkle, J. C. Cohen, H. H. Hobbs, Genetic variation in PNPLA3 confers susceptibility to nonalcoholic fatty liver disease. *Nat. Genet.* **40**, 1461–1465 (2008).

6. J. Kozlitina, E. Smagris, S. Stender, B. G. Nordestgaard, H. H. Zhou, A. Tybjærg-Hansen, T. F. Vogt, H. H. Hobbs, J. C. Cohen, Exome-wide association study identifies a TM6SF2 variant that confers susceptibility to nonalcoholic fatty liver disease. *Nat. Genet.* **46**, 352–356 (2014).

7. N. S. Abul-Husn, X. Cheng, A. H. Li, Y. Xin, C. Schurmann, P. Stevis, Y. Liu, J. Kozlitina, S. Stender, G. C. Wood, A. N. Stepanchick, M. D. Still, S. McCarthy, C. O'Dushlaine, J. S. Packer, S. Balasubramanian, N. Gosalia, D. Esopi, S. Y. Kim, S. Mukherjee, A. E. Lopez, E. D. Fuller, J. Penn, X. Chu, J. Z. Luo, U. L. Mirshahi, D. J. Carey, C. D. Still, M. D. Feldman, A. Small, S. M. Damrauer, D. J. Rader, B. Zambrowicz, W. Olson, A. J. Murphy, I. B. Borecki, A. R. Shuldiner, J. G. Reid, J. D. Overton, G. D. Yancopoulos, H. H. Hobbs, J. C. Cohen, O. Gottesman, T. M. Teslovich, A. Baras, T. Mirshahi, J. Gromada, F. E. Dewey, A Protein-Truncating HSD17B13 Variant and Protection from Chronic Liver Disease. *N. Engl. J. Med.* **378**, 1096–1106 (2018).

8. F. Zhang, J. R. Lupski, Non-coding genetic variants in human disease. *Hum. Mol. Genet.* **24**, R102–R110 (2015).

9. K. Suhre, M. I. McCarthy, J. M. Schwenk, Genetics meets proteomics: perspectives for large population-based studies. *Nat. Rev. Genet.* **22**, 19–37 (2021).

10. E. Ferkingstad, P. Sulem, B. A. Atlason, G. Sveinbjornsson, M. I. Magnusson, E. L. Styrmisdottir, K. Gunnarsdottir, A. Helgason, A. Oddsson, B. V. Halldorsson, B. O. Jensson, F. Zink, G. H. Halldorsson, G. Masson, G. A. Arnadottir, H. Katrinardottir, K. Juliusson, M. K. Magnusson, O. T. Magnusson, R. Fridriksdottir, S. Saevarsdottir, S. A. Gudjonsson, S. N. Stacey, S. Rognvaldsson, T. Eiriksdottir, T. A. Olafsdottir, V. Steinthorsdottir, V. Tragante, M. O. Ulfarsson, H. Stefansson, I. Jonsdottir, H. Holm, T. Rafnar, P. Melsted, J. Saemundsdottir, G. L. Norddahl, S. H. Lund, D. F. Gudbjartsson, U. Thorsteinsdottir, K. Stefansson, Large-scale integration of the plasma proteome with genetics and disease. *Nat. Genet.* **53**, 1712–1721 (2021).

11. B. B. Sun, J. C. Maranville, J. E. Peters, D. Stacey, J. R. Staley, J. Blackshaw, S. Burgess, T. Jiang, E. Paige, P. Surendran, C. Oliver-Williams, M. A. Kamat, B. P. Prins, S. K. Wilcox, E. S. Zimmerman, A. Chi, N. Bansal, S. L. Spain, A. M. Wood, N. W. Morrell, J. R. Bradley, N. Janjic, D. J. Roberts, W. H. Ouwehand, J. A. Todd, N. Soranzo, K. Suhre, D. S. Paul, C. S. Fox, R. M. Plenge, J. Danesh, H. Runz, A. S. Butterworth, Genomic atlas of the human plasma proteome. *Nature*. **558**, 73–79 (2018).

12. C. Bycroft, C. Freeman, D. Petkova, G. Band, L. T. Elliott, K. Sharp, A. Motyer, D. Vukcevic, O. Delaneau, J. O'Connell, A. Cortes, S. Welsh, A. Young, M. Effingham, G. McVean, S. Leslie, N. Allen, P. Donnelly, J. Marchini, The UK Biobank resource with deep phenotyping and genomic data. *Nature*. **562**, 203–209 (2018).

13. M. E. Haas, J. P. Pirruccello, S. N. Friedman, M. Wang, C. A. Emdin, V. H. Ajmera, T. G. Simon, J. R. Homburger, X. Guo, M. Budoff, K. E. Corey, A. Y. Zhou, A. Philippakis, P. T. Ellinor, R. Loomba, P. Batra, A. V. Khera, Machine learning enables new insights into genetic contributions to liver fat accumulation. *Cell Genom*. **1** (2021), doi:10.1016/j.xgen.2021.100066.

14. J. Bowden, F. Del Greco M, C. Minelli, Q. Zhao, D. A. Lawlor, N. A. Sheehan, J. Thompson, G. Davey Smith, Improving the accuracy of two-sample summary-data Mendelian randomization: moving beyond the NOME assumption. *Int. J. Epidemiol.* **48**, 728–742 (2019).

15. J. Bowden, G. Davey Smith, S. Burgess, Mendelian randomization with invalid instruments: effect estimation and bias detection through Egger regression. *Int. J. Epidemiol.* **44**, 512–525 (2015).

16. K. S. Carmon, X. Gong, J. Yi, A. Thomas, Q. Liu, RSPO-LGR4 functions via IQGAP1 to potentiate Wnt signaling. *Proc. Natl. Acad. Sci. U. S. A.* **111**, E1221–9 (2014).

17. N. Y. Loh, J. E. N. Minchin, K. E. Pinnick, M. Verma, M. Todorčević, N. Denton, J. E.-S. Moustafa, J. P. Kemp, C. L. Gregson, D. M. Evans, M. J. Neville, K. S. Small, M. I. McCarthy, A. Mahajan, J. F. Rawls, F. Karpe, C. Christodoulides, RSPO3 impacts body fat distribution and regulates adipose cell biology in vitro. *Nat.*

Commun. **11**, 2797 (2020).

18. E. Fabbrini, S. Sullivan, S. Klein, Obesity and nonalcoholic fatty liver disease: biochemical, metabolic, and clinical implications. *Hepatology*. **51**, 679–689 (2010).

19. S. L. Pulit, C. Stoneman, A. P. Morris, A. R. Wood, C. A. Glastonbury, J. Tyrrell, L. Yengo, T. Ferreira, E. Marouli, Y. Ji, J. Yang, S. Jones, R. Beaumont, D. C. Croteau-Chonka, T. W. Winkler, GIANT Consortium, A. T. Hattersley, R. J. F. Loos, J. N. Hirschhorn, P. M. Visscher, T. M. Frayling, H. Yaghootkar, C. M. Lindgren, Meta-analysis of genome-wide association studies for body fat distribution in 694 649 individuals of European ancestry. *Hum. Mol. Genet.* **28**, 166–174 (2019).

20. S. Agrawal, M. Wang, M. D. R. Klarqvist, K. Smith, J. Shin, H. Dashti, N. Diamant, S. H. Choi, S. J. Jurgens, P. T. Ellinor, A. Philippakis, M. Claussnitzer, K. Ng, M. S. Udler, P. Batra, A. V. Khera, Inherited basis of visceral, abdominal subcutaneous and gluteofemoral fat depots. *Nat. Commun.* **13**, 3771 (2022).

21. I. J. Neeland, R. Ross, J.-P. Després, Y. Matsuzawa, S. Yamashita, I. Shai, J. Seidell, P. Magni, R. D. Santos, B. Arsenault, A. Cuevas, F. B. Hu, B. Griffin, A. Zambon, P. Barter, J.-C. Fruchart, R. H. Eckel, International Atherosclerosis Society, International Chair on Cardiometabolic Risk Working Group on Visceral Obesity, Visceral and ectopic fat, atherosclerosis, and cardiometabolic disease: a position statement. *Lancet Diabetes Endocrinol*. **7**, 715–725 (2019).

22. D. Kim, G. E. Chung, M.-S. Kwak, H. B. Seo, J. H. Kang, W. Kim, Y. J. Kim, J.-H. Yoon, H.-S. Lee, C. Y. Kim, Body Fat Distribution and Risk of Incident and Regressed Nonalcoholic Fatty Liver Disease. *Clin. Gastroenterol. Hepatol.* **14**, 132–8.e4 (2016).

23. A. R. Carter, E. Sanderson, G. Hammerton, R. C. Richmond, G. Davey Smith, J. Heron, A. E. Taylor, N. M. Davies, L. D. Howe, Mendelian randomisation for mediation analysis: current methods and challenges for implementation. *Eur. J. Epidemiol.* **36**, 465–478 (2021).

24. H. Zhao, H. Rasheed, T. H. Nøst, Y. Cho, Y. Liu, L. Bhatta, A. Bhattacharya, Global Biobank Meta-analysis Initiative, G. Hemani, G. Davey Smith, B. M. Brumpton, W. Zhou, B. M. Neale, T. R. Gaunt, J. Zheng, *Cell Genom*, in press.

25. L. Chen, J. E. Peters, B. Prins, E. Persyn, M. Traylor, P. Surendran, S. Karthikeyan, E. Yonova-Doing, E. Di Angelantonio, D. J. Roberts, N. A. Watkins, W. H. Ouwehand, J. Danesh, C. M. Lewis, P. G. Bronson, H. S. Markus, S. Burgess, A. S. Butterworth, J. M. M. Howson, Systematic Mendelian randomization using the human plasma proteome to discover potential therapeutic targets for stroke. *Nat. Commun.* **13**, 6143 (2022).

26. J. Zheng, V. Haberland, D. Baird, V. Walker, P. C. Haycock, M. R. Hurle, A. Gutteridge, P. Erola, Y. Liu, S. Luo, J. Robinson, T. G. Richardson, J. R. Staley, B. Elsworth, S. Burgess, B. B. Sun, J. Danesh, H. Runz, J. C. Maranville, H. M.

Martin, J. Yarmolinsky, C. Laurin, M. V. Holmes, J. Z. Liu, K. Estrada, R. Santos, L. McCarthy, D. Waterworth, M. R. Nelson, G. D. Smith, A. S. Butterworth, G. Hemani, R. A. Scott, T. R. Gaunt, Phenome-wide Mendelian randomization mapping the influence of the plasma proteome on complex diseases. *Nat. Genet.* **52**, 1122–1131 (2020).

27. M. Zhang, M. Haughey, N.-Y. Wang, K. Blease, A. M. Kapoun, S. Couto, I. Belka, T. Hoey, M. Groza, J. Hartke, B. Bennett, J. Cain, A. Gurney, B. Benish, P. Castiglioni, C. Drew, J. Lachowicz, L. Carayannopoulos, S. D. Nathan, J. Distler, D. A. Brenner, K. Hariharan, H. Cho, W. Xie, Targeting the Wnt signaling pathway through R-spondin 3 identifies an anti-fibrosis treatment strategy for multiple organs. *PLoS One*. **15**, e0229445 (2020).

28. S. Lee, G. R. Abecasis, M. Boehnke, X. Lin, Rare-variant association analysis: study designs and statistical tests. *Am. J. Hum. Genet.* **95**, 5–23 (2014).

29. I. J. Neeland, R. Ross, J.-P. Després, Y. Matsuzawa, S. Yamashita, I. Shai, J. Seidell, P. Magni, R. D. Santos, B. Arsenault, A. Cuevas, F. B. Hu, B. Griffin, A. Zambon, P. Barter, J.-C. Fruchart, R. H. Eckel, International Atherosclerosis Society, International Chair on Cardiometabolic Risk Working Group on Visceral Obesity, Visceral and ectopic fat, atherosclerosis, and cardiometabolic disease: a position statement. *Lancet Diabetes Endocrinol*. **7**, 715–725 (2019).

30. M. Eslam, P. N. Newsome, S. K. Sarin, Q. M. Anstee, G. Targher, M. Romero-Gomez, S. Zelber-Sagi, V. Wai-Sun Wong, J.-F. Dufour, J. M. Schattenberg, T. Kawaguchi, M. Arrese, L. Valenti, G. Shiha, C. Tiribelli, H. Yki-Järvinen, J.-G. Fan, H. Grønbæk, Y. Yilmaz, H. Cortez-Pinto, C. P. Oliveira, P. Bedossa, L. A. Adams, M.-H. Zheng, Y. Fouad, W.-K. Chan, N. Mendez-Sanchez, S. H. Ahn, L. Castera, E. Bugianesi, V. Ratziu, J. George, A new definition for metabolic dysfunction-associated fatty liver disease: An international expert consensus statement. *J. Hepatol.* **73**, 202–209 (2020).

31. X. Du, N. DeForest, A. R. Majithia, Human Genetics to Identify Therapeutic Targets for NAFLD: Challenges and Opportunities. *Front. Endocrinol.* . **12**, 777075 (2021).

32. A. G. Shah, A. Lydecker, K. Murray, B. N. Tetri, M. J. Contos, A. J. Sanyal, Nash Clinical Research Network, Comparison of noninvasive markers of fibrosis in patients with nonalcoholic fatty liver disease. *Clin. Gastroenterol. Hepatol.* **7**, 1104–1112 (2009).

33. A. Akhmetshina, K. Palumbo, C. Dees, C. Bergmann, P. Venalis, P. Zerr, A. Horn, T. Kireva, C. Beyer, J. Zwerina, H. Schneider, A. Sadowski, M.-O. Riener, O. A. MacDougald, O. Distler, G. Schett, J. H. W. Distler, Activation of canonical Wnt signalling is required for TGF-β-mediated fibrosis. *Nat. Commun.* **3**, 735 (2012).

34. V. W.-S. Wong, L. A. Adams, V. de Lédinghen, G. L.-H. Wong, S. Sookoian, Noninvasive biomarkers in NAFLD and NASH - current progress and future promise. *Nat. Rev. Gastroenterol. Hepatol.* **15**, 461–478 (2018).

35. GTEx Consortium, The Genotype-Tissue Expression (GTEx) project. *Nat. Genet.* **45**, 580–585 (2013).

36. S. A. MacParland, J. C. Liu, X.-Z. Ma, B. T. Innes, A. M. Bartczak, B. K. Gage, J. Manuel, N. Khuu, J. Echeverri, I. Linares, R. Gupta, M. L. Cheng, L. Y. Liu, D. Camat, S. W. Chung, R. K. Seliga, Z. Shao, E. Lee, S. Ogawa, M. Ogawa, M. D. Wilson, J. E. Fish, M. Selzner, A. Ghanekar, D. Grant, P. Greig, G. Sapisochin, N. Selzner, N. Winegarden, O. Adeyi, G. Keller, G. D. Bader, I. D. McGilvray, Single cell RNA sequencing of human liver reveals distinct intrahepatic macrophage populations. *Nat. Commun.* **9**, 4383 (2018).

37. A. S. Rocha, V. Vidal, M. Mertz, T. J. Kendall, A. Charlet, H. Okamoto, A. Schedl, The Angiocrine Factor Rspondin3 Is a Key Determinant of Liver Zonation. *Cell Rep.* **13**, 1757–1764 (2015).

38. J. Poisson, S. Lemoinne, C. Boulanger, F. Durand, R. Moreau, D. Valla, P.-E. Rautou, Liver sinusoidal endothelial cells: Physiology and role in liver diseases. *J. Hepatol.* **66**, 212–227 (2017).

39. U. E. Lee, S. L. Friedman, Mechanisms of hepatic fibrogenesis. *Best Pract. Res. Clin. Gastroenterol.* **25**, 195–206 (2011).

40. A. Glinka, C. Dolde, N. Kirsch, Y.-L. Huang, O. Kazanskaya, D. Ingelfinger, M. Boutros, C.-M. Cruciat, C. Niehrs, LGR4 and LGR5 are R-spondin receptors mediating Wnt/β-catenin and Wnt/PCP signalling. *EMBO Rep.* **12**, 1055–1061 (2011).

41. O. O. Yavorska, S. Burgess, MendelianRandomization: an R package for performing Mendelian randomization analyses using summarized data. *Int. J. Epidemiol.* **46**, 1734–1739 (2017).

42. C. T. Saunders, J. D. Blume, A classical regression framework for mediation analysis: fitting one model to estimate mediation effects. *Biostatistics*. **19**, 514–528 (2018).

CHAPTER 4

Genome-wide discovery and integrative genomic characterization of insulin resistance loci using serum triglycerides to HDL-cholesterol ratio

Abstract

Insulin resistance causes multiple epidemic metabolic diseases including type 2 diabetes, cardiovascular disease and fatty liver, but is not routinely measured in epidemiological studies. To discover novel insulin resistance genes in the general population, we conducted genome-wide association studies in 382,129 individuals for triglyceride to HDL-cholesterol ratio (TG/HDL), a surrogate marker of insulin resistance calculable from commonly measured serum lipid profiles. We identified 251 independent loci, of which 62 were more strongly associated with TG/HDL compared to TG or HDL alone, suggesting them as insulin resistance loci. Causal genes at these loci were prioritized by fine mapping with directions-of-effect and tissue specificity annotated through analysis of protein-coding and expression quantitative trait genetic variation. Annotations were independently corroborated by relating gene expression in biopsied tissues of individuals undergoing glycemic clamp to directly measure insulin resistance. We highlight two phospholipase encoding genes *PLA2G12A* and *PLA2G6* which liberate arachidonic acid and improve insulin sensitivity, and *VGLL3*, a pro-fibrotic, transcriptional co-factor that increases insulin resistance partially through enhanced adiposity. Finally, we implicate the anti-apoptotic factor *TNFAIP8* for the first time in insulin sensitization specifically in females, which acts by increasing visceral adiposity. In summary our study identifies several new modulators of insulin resistance with potential to serve as biomarkers or pharmacological targets.

Introduction

Insulin resistance is a major cause of multiple chronic and epidemic diseases including type 2 diabetes (T2D), non-alcoholic fatty liver disease (NAFLD), cardiovascular disease (CVD), and cancer(*1*, *2*). Though environment and lifestyle factors play a role in insulin resistance pathogenesis, an individual's risk of developing insulin resistance and its consequential disorders is strongly determined by genetic factors(*3*), with heritability estimates ranging from 24-73%(*4*, *5*). Family-based studies have implicated several insulin resistance related genes(*6*–*8*), but modern genome-wide association studies (GWAS) with sequenced population-based cohorts accommodate larger sample sizes and offer more statistical power to identify novel effector genes(*9*). GWAS for insulin resistance however are challenging to perform in large population cohorts as the protocols for directly measuring whole-body insulin sensitivity, such as the gold standard hyperinsulinemic-euglycemic clamp, are difficult to perform at scale. To circumvent this, GWAS have been conducted using surrogate markers of insulin resistance such as glycemic traits(*10*, *11*) and anthropomorphic measurements(*12*) which have successfully identified numerous risk loci enriched in insulin-dependent tissues such as muscle, adipose, and liver. Larger sample sizes have further enabled investigation into the sex-dimorphism of insulin resistance and have identified numerous sex-dimorphic risk loci(*12*, *13*). Though these studies have revealed valuable insights, each surrogate phenotype only captures a portion of insulin resistance pathophysiology. For instance, GWAS for glycemic traits have dissected the genetic basis related to glucose metabolism dysfunction and genetic findings from examining anthropomorphic measurements have been limited to explain adipose tissue related insulin resistance. The variation in

loci/genes identified by these studies suggest that genetic mapping of other complementary insulin resistance biomarkers would deepen our knowledge of genetic factors modulating insulin resistance.

The triglyceride (TG) to HDL-cholesterol (HDL) ratio (TG/HDL) is a widely available, surrogate marker of insulin resistance that is calculable from commonly measured clinical tests and is highly correlated with hyperinsulinemic-euglycemic clamp based measurements(*14*). To characterize the genetic basis of TG/HDL and systematically identify novel insulin resistance genes, we conducted a GWAS for TG/HDL in 382,129 UK Biobank (UKBB) participants and prioritized target genes at top genomic risk loci through statistical fine-mapping and combined single nucleotide polymorphisms (SNP)-to-gene linking strategies. Prioritized genes were further interrogated in multiple orthogonal genetic and genomic analyses to determine directional effects and tissue-specificity.

Results

GWAS for TG/HDL in 382,129 UK Biobank participants identifies 251 associated loci.

To confirm the relationship between TG/HDL and insulin resistance measurements, we compared the TG/HDL value concurrently obtained from a previous study of 45 individuals undergoing hyperinsulinemic-euglycemic clamp, the gold standard of insulin sensitivity quantification(*15*) (Figure 4.1). We observed a strong and significantly negative phenotypic correlation between TG/HDL and clamp-based glucose disposal rate (Rd) (ρ=-0.48, p=0.003, Pearson correlation) as well as with other commonly used surrogate markers such as fasting insulin (FI) (ρ=0.42, p=0.009).

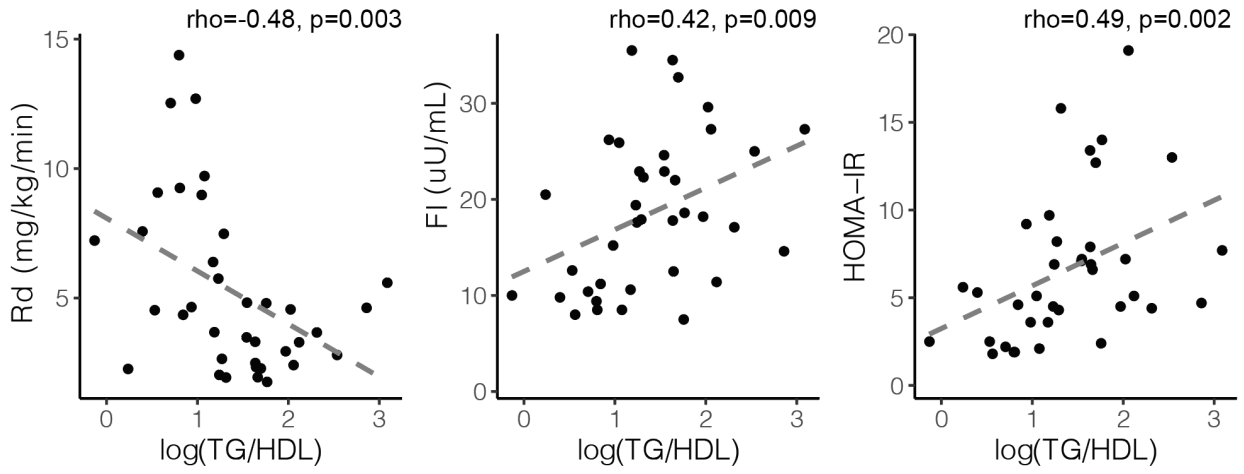


Figure 4.1. Correlation of TG/HDL with other measures of insulin resistance.
Spearman correlations among 45 individuals between TG/HDL and measures of insulin resistance including glucose disposal rate (Rd) measured by undergoing hyperinsulinemic-euglycemic clamp (left), fasting insulin (FI, middle), and Homeostatic Model Assessment for Insulin Resistance (HOMA-IR, right). Spearman rank correlation coefficients (rho) and p-values are given.

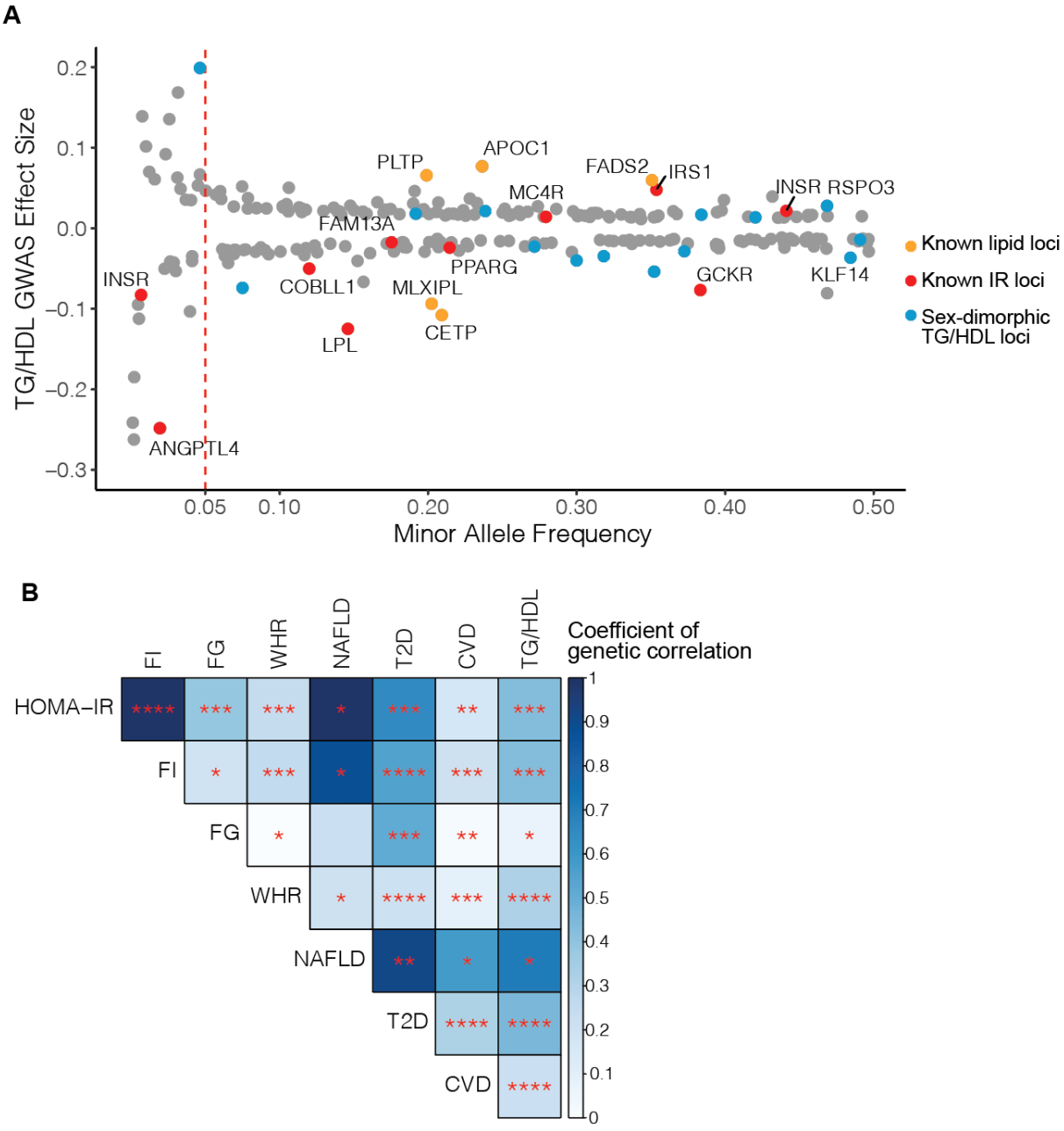
Aiming to identify genomic risk loci for TG/HDL, we conducted a GWAS on serum TG/HDL levels in 382,129 individuals from the UK Biobank who had both TG and HDL measurements available. Median age of the included individuals was 58 years, median BMI was 26.7, 54% were female, and 93% were of European ancestry. After quality control and filtering based on the Global Lipids Genetics Consortium standards(*16*), approximately 12 million SNPs were tested for association with TG/HDL using whole-genome linear regression as instantiated by REGENIE(*17*). The raw SNP associations were clustered into 251 TG/HDL genomic risk loci using the FUMA platform(*18*). The genetic architecture of the 251 identified TG/HDL loci was consistent with a highly polygenic trait with the majority of identified loci tagged by common SNPs (223/251 lead SNPs with minor allele frequency (MAF) > 0.05) of relatively smaller effect sizes and fewer low frequency variants (MAF < 0.01) with effect sizes 2-3x larger than those of lead

common variants (Figure 4.2A). Known insulin resistance signals from prior GWAS were re-identified including *GCKR(11), IRS1(19), PPARG(20), FAM13A(21)*, *ANGPTL4(22),* and *LPL(23)*. Lipid-associated loci for TG and HDL alone were also re-identified including *MLXIPL, FADS2, CETP,* and *PLTP(24)*. Overlapping the lead SNPs marking these loci with tissue specific regulatory elements (as instantiated in GREGOR(*25*)) we found enrichment in adipose tissue, liver, pancreatic islets, and skeletal muscle, all of which have been implicated in insulin sensitivity and insulin action(*26*).

Given the known sexual dimorphism in insulin resistance and metabolic disease risk(*27*), we also conducted a sex-stratified GWAS for TG/HDL in the same UKBB cohort (176,117 males vs. 206,012 females) assessing each of the 251 genomic risk loci from the sex-combined TG/HDL GWAS (Methods) for evidence of sexually dimorphic signals. Of the 251 loci, 17 showed significant evidence of sex dimorphic genetic association ($p < 0.05/251$ t-test, Figure 4.2A) with six of these having stronger association in females (positive t-statistic). Among these were known sex-dimorphic insulin resistance genes *KLF14(28)* and *RSPO3(29)*.

Figure 4.2. GWAS for TG/HDL in 382,129 UK Biobank participants identifies 251 associated loci.

A) Genetic architecture of 251 TG/HDL associated loci identified in the UK Biobank (n=382,129 participants). Shown are the effect sizes of the lead SNPs plotted according to minor allele frequency. Previously identified lipid and insulin resistance (IR) loci are colored orange and red respectively. Loci with sex-dimorphic association signals are highlighted in blue (n = 17/251). B) Genetic correlations of TG/HDL with other surrogate measures of insulin resistance: fasting insulin (FI), fasting glucose (FG), waist-hip ratio (WHR), and metabolic diseases: non-alcoholic fatty liver disease (NAFLD), type 2 diabetes (T2D), and cardiovascular disease (CVD). Color intensity represents the magnitude of the coefficient of genetic correlation and statistical significance is indicated in the following bins: ****p<1e-20, ***p<1e-5, **p<0.001, *p<0.05. TG/HDL shows strong positive genetic correlation with surrogate markers of insulin resistance and metabolic diseases.


A
TG/HDL GWAS Effect Size
0.2
0.1
0.0
−0.1
−0.2
−0.3
0.05
0.10
0.20
0.30
0.40
0.50
Minor Allele Frequency
PLTP
APOC1
FADS2
IRS1
INSR RSPO3
MC4R
FAM13A
PPARG
INSR
COBLL1
MLXIPL
GCKR
KLF14
LPL
CETP
ANGPTL4
Known lipid loci
Known IR loci
Sex-dimorphic TG/HDL loci
B
FI
FG
WHR
NAFLD
T2D
CVD
TG/HDL
Coefficient of genetic correlation
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
HOMA−IR
FI
FG
WHR
NAFLD
T2D
CVD

We attempted to corroborate these associations in the Mass General Biobank (MGBB, n=37,545)(*30*), an independent dataset from the UKBB. Given the MGBB cohort was ten percent the size of the UKBB, we did not attempt to corroborate individual loci for lack of statistical power (only 11 loci had greater than 80% power for replication). Instead we compared all loci in which UKBB discovery analysis that had proxy SNPs available in MGBB (n=240) finding 40 loci robustly replicated (p-value < 0.05 / 240), 130 loci nominally replicated (p-value < 0.05), and 226 loci (94%) had the same direction of effect between UKBB and MGBB.

Having observed a phenotypic correlation between TG/HDL, clamp-based glucose infusion rate, and surrogate markers of insulin resistance (Figure 4.1), we sought to quantify the degree to which TG/HDL, other surrogate insulin resistance markers, and metabolic disease outcomes shared similar genetic influences. We computed the genetic correlations(*31*) between our TG/HDL GWAS and previous association studies of Homeostatic Model Assessment for Insulin Resistance (HOMA-IR)(*11*), fasting insulin (FI)(*13*), fasting glucose (FG)(*13*), waist-hip ratio adjusted for BMI (WHR)(*12*), as well as metabolic disease sequelae including NAFLD(*32*), T2D(*33*), CVD(*34*) (Figure 4.2B). Positive genetic correlations were observed between TG/HDL and T2D (r_g=0.51, p=1.9 x 10^{-36}), CVD (r_g=0.31, p=2.8 x 10^{-28}), and NAFLD (r_g=0.74, p=0.001). Notably, TG/HDL had comparable genetic correlation with T2D as compared to glycemic traits (FI r_g=0.6, p=5.2 x10^{-29}; FG r_g=0.57, p=3.8 x 10^{-18}) but greater genetic correlation with CVD (FI r_g=0.29, p=3 x 10^{-9}; FG r_g=0.12, p=1 x 10^{-4}) and NAFLD (FI r_g=0.9, p=0.01; FG r_g=0.3, p > 0.05).

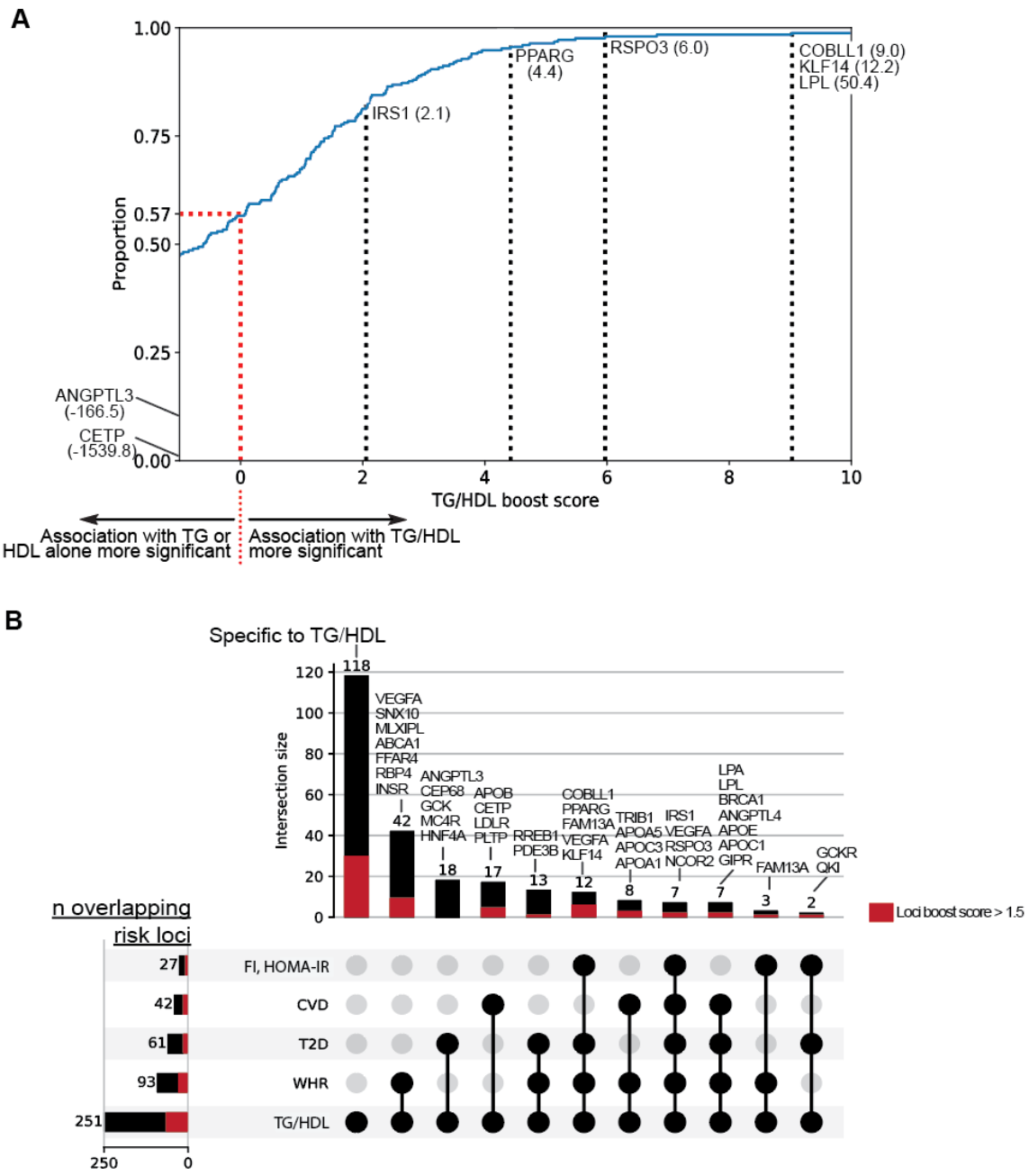
Prioritizing “boosted” TG/HDL associations to identify insulin resistance loci

Given that TG/HDL as a surrogate measure of insulin resistance is derived from serum TG and HDL, we next sought to identify signals that were “boosted” for association with TG/HDL versus TG or HDL alone under the model that boosted loci would be more likely to contain genes related to insulin resistance than lipid regulation. Using the lead SNP of each TG/HDL risk locus, we computed a “boost score” (TG/HDL_{BS}) utilizing the association p-values for TG/HDL, TG, and HDL for that SNP (Methods). Of the 251 loci identified in our study, 109 (43%) had boost score > 0 indicating increased significance in GWAS for TG/HDL compared to TG or HDL alone (Figure 4.3A, red line). Focusing on known loci, we highlight that the *LPL* locus met the threshold for genome-wide significance for TG, HDL, and TG/HDL and had the highest boost score (TG/HDL_{BS} = 50.4; TG/HDL p-value < 10^{-411}, TG p-value < 10^{-324}, HDL p-value = 10^{-362}). These results are highly concordant with the well-characterized role of LPL in regulating both insulin sensitivity and serum lipid levels(*35*). Other previously known and recently characterized insulin resistance loci including *IRS1(19)* (TG/HDL_{BS} = 2.1)*, PPARG(20)* (4.4)*, RSPO3(29)* (6.0)*, KLF14(28)* (12.2), and *COBLL1(36)* (9.0) demonstrated highly positive boost values as well (Figure 4.3A). On the other hand, genes known to regulate lipids without relation to insulin sensitivity showed strongly negative boost scores. For example, the locus containing *CETP* had the lowest boost score (TG/HDL_{BS} = -1539.8) driven by its substantially stronger association with HDL (p < 10^{-1952}) compared to TG/HDL (p < 10^{-412}); this is consistent with the known role of CETP as a lipid transfer protein responsible for transferring cholesterol esters from HDL to other lipoproteins(*37*). *ANGPTL3* also had a negative boost score (TG/HDL_{BS} = -166.5) due to a stronger association with TG (p <

10^{-309}) compared to TG/HDL ($p < 10^{-143}$) as would be expected from its role in inhibiting TG hydrolysis through the inhibition of lipoprotein lipase(*38*). We proceeded to rank the 251 associated loci according to their boost score and focused further investigation on the top quartile of loci (n = 62, $TG/HDL_{BS} > 1.5$).

Given the shared genetic basis of TG/HDL, other surrogate insulin resistance biomarkers, and metabolic disease outcomes, we performed a locus level comparison with previous biomarker/metabolic disease association studies to survey which genetic signals were common and distinct with our TG/HDL GWAS and top-quartile boosted loci (Figure 4.3B). We scored the overlap of SNPs identified from studies of WHR (n=694,649)(*12*), FI (n=105,056)(*13*) or HOMA-IR (n=46,186)(*11*), T2D (n=180,834 cases vs. 1,159,055 controls)(*33*), and CVD (n=181,522 cases vs. 1,165,690 controls)(*39*) with regard to their positional overlap with any of 251 TG/HDL loci identified in our study (Figure 4.3B). WHR shared the largest number of overlapping associated loci with TG/HDL (n=93 loci) followed by T2D (n=61 loci), CVD (n=42 loci), and loci associated with either FI or HOMA-IR (n=27 loci). Of the 251 loci identified in our TG/HDL study, 118 (containing 29 boosted loci) had not previously been implicated as insulin resistance or metabolic disease risk loci (Figure 4.3B).

Figure 4.3. Prioritizing "boosted" TG/HDL associations to identify insulin resistance loci.
A) Cumulative distribution plot of 251 TG/HDL loci according to TG/HDL Boost score. Boost scores greater than 0 signify a stronger association with TG/HDL than TG or HDL alone. Boost scores less than 0 signify the opposite. Among the 251 TG/HDL loci 109 (43%) have a boost score > 0 (dotted red line). Notable lipid regulation genes (TG: ANGPTL3, HDL: CETP) have highly negative negative boost scores, whereas known insulin resistance genes show positive boost scores. B) Upset plot showing intersection of the 251 identified TG/HDL loci with previous GWAS for surrogate insulin resistance markers and metabolic diseases: fasting insulin (FI), HOMA-IR, coronary artery disease (CVD), type 2 diabetes (T2D), waist hip ratio (WHR). (left) Total number of loci which overlap from each study with the 251 TG/HDL loci. (top) Number of loci within each intersecting group of studies denoted by black circles with selected known insulin resistance loci labeled above. The proportion of each bar composed of loci with top-quartile boost scores (> 1.5) is highlighted in red. For example, TG/HDL, T2D, and FI share 2 loci, GCKR and QKI, in common.


A
1.00
0.75
0.57
0.50
0.25
0.00
Proportion
IRS1 (2.1)
PPARG (4.4)
RSPO3 (6.0)
COBLL1 (9.0)
KLF14 (12.2)
LPL (50.4)
ANGPTL3 (-166.5)
CETP (-1539.8)
0
2
4
6
8
10
TG/HDL boost score
Association with TG or HDL alone more significant
Association with TG/HDL more significant
B
Specific to TG/HDL
Intersection size
118
VEGFA SNX10 MLXIPL ABCA1 FFAR4 RBP4 INSR
42
ANGPTL3 CEP68 GCK MC4R HNF4A
18
APOB CETP LDLR PLTP
17
RREB1 PDE3B
13
COBLL1 PPARG FAM13A VEGFA KLF14
12
TRIB1 APOA5 APOC3 APOA1
8
IRS1 VEGFA RSPO3 NCOR2
7
LPA LPL BRCA1 ANGPTL4 APOE APOC1 GIPR
7
FAM13A
3
GCKR QKI
2
Loci boost score > 1.5
n overlapping risk loci
27
42
61
93
251
250
0
FI, HOMA-IR
CVD
T2D
WHR
TG/HDL

Putative causal gene identification and integrative genomics to infer direction-of-effect of "boosted" TG/HDL loci

For the top quartile of TG/HDL associated loci with positive boost scores (i.e. TG/HDL_{BS} > 1.5, n = 62 loci) we employed an integrative statistical fine-mapping approach instantiated in the "Sum of Single Effects" (SuSiE) model(*40*) (Methods) to identify likely causal variants and genes. The SuSiE fine-mapping algorithm converged for 57 of the 62 loci resulting in credible sets of SNPs with 95% posterior probability of containing a causal variant. Across the 57 mappable loci, the defined credible sets contained a median of 66 variants (range 1-196); at 25 of these loci, at least one credible set consisted of only a single, high-confidence variant (posterior inclusion probability (PIP) > 95%).

Having defined the sets of potentially causal variants, we then used this information to nominate the most likely causal gene at each locus applying commonly used practices(*41*). Fifty fine-mapped loci contained at least one putative "causal" variant (defined as a variant with PIP > 0.1) (Table 4.1). These putative causal variants were cross-referenced with genomic and functional annotations (e.g. exons, promoters, expression quantitative trait associations) to assign a gene at each locus(*41*). If the putative causal variant did not overlap with any known functional/genomic association, the nearest gene was assigned to be causal(*41*) (n = 27/50). This data-driven approach to identify causal genes recovered several well-established insulin resistance genes at their respective risk loci including *LPL*, *IRS1*, and *PPARG* (Table 4.1).

Table 4.1. Nominated causal genes at top TG/HDL risk loci.

50 nominated genes at top quartile boosted TG/HDL risk loci. Each locus is defined by a lead SNP, and the corresponding Beta, standard error (SE), and GWAS P-value are listed. The boost score is listed which quantifies the difference in association of each locus between TG/HDL and TG or HDL alone, with boost score > 0 indicating increased significance in GWAS for TG/HDL compared to TG or HDL alone. The sex-dimorphic effects of each locus are shown, with a positive t-statistic indicating greater significance in female-specific GWAS for TG/HDL compared to male, and the sex-dimorphic p-value quantifying if the difference between sexes is significant. Top causal SNPs are given if not the locus lead SNP, and the annotation of the lead SNP / top causal SNP for gene nomination at each locus is listed.

Nominated Gene	Lead SNP	Beta (SE)	P-value	TG/HDL Boost Score	Sex-dimorphic t.statistic	Sex-dimorphic p-value	Top causal SNP if not lead SNP	SNP-gene annotation
LPL	rs271	-0.125 (0.003)	3.58e-307	50.353	-2.189	2.86e-02	rs268	Exonic (p.Asn318Ser)
ANGPTL4	rs116843064	-0.248 (0.007)	1.03e-189	49.807	-3.802	1.43e-04		Exonic (p.Glu40Lys); Regulatory variant
IRS1	rs2943645	0.048 (0.002)	2.48e-85	2.056	3.240	1.20e-03		Fine-mapped eQTL
APOB	rs533617	-0.103 (0.005)	5.14e-67	5.480	-0.803	4.22e-01		Exonic (p.His1923Arg)
VEGFA	rs6905288	0.039 (0.002)	8.54e-63	6.813	0.106	9.16e-01		Regulatory variant
KLF14	7:130438531_CTTTTTT_C	-0.037 (0.002)	2.50e-55	12.202	7.959	1.73e-15		Nearest gene
COBLL1	rs79953491	-0.05 (0.003)	2.78e-44	9.028	3.251	1.15e-03		Promoter
PLCB3	rs71468663	0.067 (0.005)	1.23e-32	3.241	-0.994	3.20e-01		Nearest gene
RSPO3	rs6916318	0.028 (0.002)	1.42e-32	5.971	-5.274	1.34e-07	rs577721086	Exonic (5_prime_UTR_variant)
ARID1A	rs114165349	0.092 (0.007)	3.13e-32	4.273	-1.461	1.44e-01		Nearest gene
NTAN1	rs11075253	-0.03 (0.002)	1.00e-30	3.781	-0.415	6.78e-01		Promoter; Fine-mapped eQTL
PCCB	rs684773	0.031 (0.002)	6.27e-30	2.487	-0.741	4.59e-01		Regulatory variant
NEK4	rs6800707	0.031 (0.003)	5.37e-25	3.775	-2.656	7.90e-03	rs11235	Exonic (3_prime_UTR_variant)
PLA2G12A	rs114816312	0.139 (0.012)	6.43e-25	3.737	0.681	4.96e-01	rs41278045	Exonic (p.Cys131Arg)

Table 4.1 (continued). Nominated causal genes at top TG/HDL risk loci.

Nominated Gene	Lead SNP	Beta (SE)	P-value	TG/HDL Boost Score	Sex-dimorphic t.statistic	Sex-dimorphic p-value	Top causal SNP if not lead SNP	SNP-gene annotation
PLA2G6	rs200725415	-0.024 (0.002)	9.26e-25	3.950	-2.334	1.96e-02		Nearest gene
RBPJ	4:26050450_AC_A	0.032 (0.003)	1.01e-24	4.731	-3.646	2.66e-04		Fine-mapped eQTL; Regulatory variant
PDE3A	rs11045171	-0.029 (0.003)	4.56e-23	4.616	1.173	2.41e-01	rs4762753	Fine-mapped eQTL
ACACB	rs149793040	-0.262 (0.023)	5.87e-23	1.906	0.336	7.37e-01		Exonic (p.Tyr1282Cys)
SNX10	rs1534696	-0.023 (0.002)	3.93e-22	5.134	2.542	1.10e-02		Fine-mapped eQTL; Regulatory variant
TSC22D2	rs62271373	0.046 (0.004)	2.56e-20	5.194	-0.016	9.88e-01		Nearest gene
AC022431.2	rs60803019	-0.043 (0.004)	4.90e-20	3.422	-0.853	3.94e-01	rs455660	Nearest gene
INSR	rs3890483	0.022 (0.002)	5.41e-20	3.064	-2.326	2.00e-02	rs1799816	Exonic (p.Val1012Met)
TTLL4	rs148358468	0.047 (0.005)	5.50e-18	3.488	-0.348	7.28e-01	rs116204487	Promoter; Regulatory variant
TNFAIP8	rs1045241	-0.023 (0.002)	5.53e-18	1.758	4.014	5.97e-05		Exonic (3_prime_UTR_variant)
RP11-1102P16.1	rs13269725	0.037 (0.004)	1.21e-17	2.402	-3.241	1.19e-03		Exonic (5_prime_UTR_variant)
ADRB1	rs2773469	-0.022 (0.002)	1.92e-17	2.904	-0.204	8.38e-01		Nearest gene
PPARG	rs2067819	-0.024 (0.002)	2.04e-17	4.421	-2.385	1.71e-02	rs59447614	Nearest gene
PEMT	rs11658944	0.042 (0.004)	1.07e-16	3.618	-3.067	2.16e-03		Nearest gene
USP3	rs7175602	0.024 (0.003)	5.84e-16	1.546	0.016	9.87e-01	rs17184382	Fine-mapped eQTL; Regulatory variant
SIPA1	rs2306363	-0.023 (0.003)	7.64e-16	1.966	-1.857	6.34e-02		Exonic (5_prime_UTR_variant)
KCNH2	7:150288766_GA_G	0.022 (0.002)	2.39e-15	2.383	-0.857	3.91e-01	rs2968864	Fine-mapped eQTL
CRTAC1	rs563296	0.018 (0.002)	2.71e-14	2.073	-0.256	7.98e-01		Nearest gene
HNRNPK	rs296886	-0.021 (0.002)	7.79e-14	2.882	1.784	7.44e-02		Nearest gene

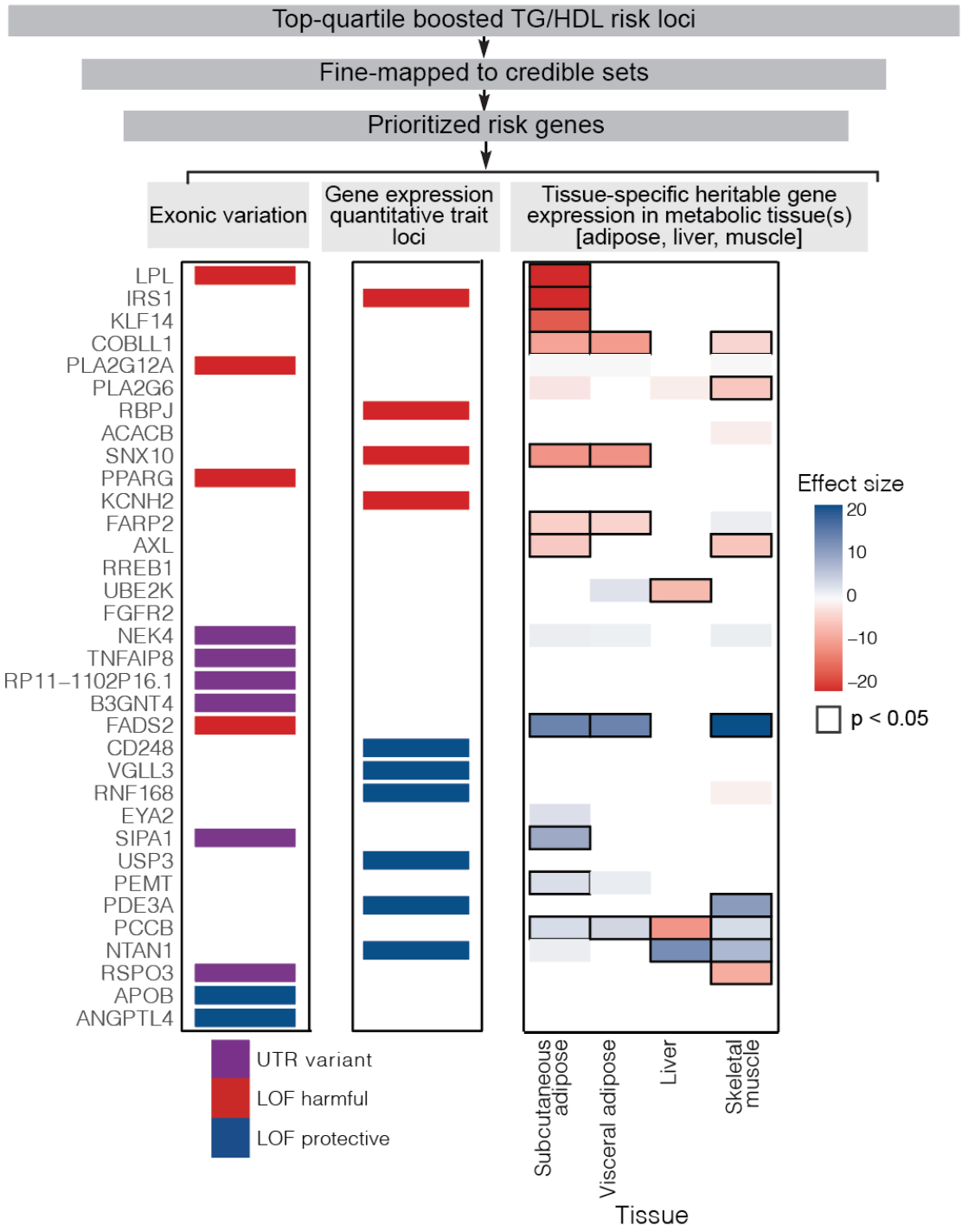
Table 4.1 (continued). Nominated causal genes at top TG/HDL risk loci.

Nominated Gene	Lead SNP	Beta (SE)	P-value	TG/HDL Boost Score	Sex-dimorphic t.statistic	Sex-dimorphic p-value	Top causal SNP if not lead SNP	SNP-gene annotation
AHR	rs4410790	0.018 (0.002)	3.71e-13	2.820	-1.691	9.08e-02		Nearest gene
RP11-399J13.3	rs678614	-0.019 (0.002)	7.56e-13	3.879	-0.635	5.26e-01		Nearest gene
FARP2	rs4675812	-0.017 (0.002)	1.53e-12	2.106	1.097	2.73e-01		Nearest gene
GADD45G	rs10797115	0.017 (0.002)	1.59e-12	1.648	-1.267	2.05e-01		Nearest gene
EYA2	rs55966194	-0.018 (0.002)	3.59e-12	1.690	1.087	2.77e-01		Nearest gene
RNF168	rs13094241	-0.018 (0.002)	9.51e-12	3.532	0.595	5.52e-01		Fine-mapped eQTL
VGLL3	rs13066793	-0.027 (0.004)	1.98e-11	2.121	0.239	8.11e-01		Fine-mapped eQTL; Regulatory variant
CD248	rs490972	0.015 (0.002)	1.91e-10	2.400	-0.621	5.35e-01	rs1625595	Fine-mapped eQTL
B3GNT4	12:122515547_ATTTTTC_A	0.018 (0.003)	1.06e-09	1.533	-4.281	1.86e-05	rs12827843	Exonic (5_prime_UTR_variant)
ZIM2	rs8102873	0.014 (0.002)	1.43e-09	1.927	-2.175	2.96e-02		Regulatory variant
AXL	rs4802113	-0.014 (0.002)	1.45e-09	1.884	-1.125	2.60e-01		Nearest gene
RREB1	rs35742417	-0.018 (0.003)	2.91e-09	1.986	-0.358	7.20e-01		Exonic (p.Ser1554Tyr); Regulatory variant
UBE2K	rs145766974	-0.013 (0.002)	1.41e-08	2.346	2.409	1.60e-02	rs34258469	Nearest gene
QKI	rs4709746	-0.019 (0.003)	2.77e-08	2.129	0.675	5.00e-01	rs7761309	Regulatory variant
FGFR2	rs878409	-0.013 (0.002)	3.39e-08	2.376	0.982	3.26e-01		Nearest gene
THRB	rs7640648	0.021 (0.003)	3.63e-08	1.506	0.344	7.31e-01		Nearest gene
FADS2	rs117186302	-0.041 (0.007)	4.36e-08	1.545	-0.130	8.96e-01	rs174566	Nearest gene

For each nominated causal gene, we integrated diverse sources of gene-level evidence to identify the direction-of-effect on insulin resistance and evaluate tissue-specificity (Figure 4.4). To determine directionality with respect to gene function and TG/HDL levels in carriers, we utilized two complementary approaches leveraging rare and common genetic variation, respectively: 1) aggregation of rare, loss-of-function coding variants within each gene identified from exome sequencing of the UKBB participants (i.e. burden tests)(*42*) and 2) functional annotation of individual putative causal SNPs with sufficiently large MAF to permit individual association with predicted protein function or modification of gene expression (eQTLs). We further extended analysis of eQTLs to all 50 putatively causal genes using established methods (Functional Summary-based Imputation (FUSION)(*43*) to correlate genetically predicted tissue-specific gene expression in metabolically relevant tissues (subcutaneous and visceral adipose, liver, and skeletal muscle) with TG/HDL in carriers.

Figure 4.4. Causal gene identification and integrative genomics to infer direction-of-effect of “boosted” TG/HDL loci.

Overview of the gene prioritization strategy for 251 TG/HDL associated loci. The top-quartile of TG/HDL boosted loci (n = 62) were selected for fine mapping to obtain 95% credible sets of causal SNPs (n = 57). Of the loci with at least one causal variant identified from fine-mapping (n = 50), causal genes were assigned by incorporating genomic and functional annotations including exonic variation and expression quantitative trait loci (eQTLs). Direction-of-effect on insulin resistance and tissue-specificity of causal genes was assigned by examination of the following gene level evidence: (left) causal variants in exons were annotated as as UTR (untranslated region; purple) or protein coding with loss-of-function (LOF harmful: LOF increases TG/HDL (red); LOF protective: LOF decreases TG/HDL (blue)). (middle) Causal variants that were eQTLs are assigned as LOF harmful or LOF protective if decreasing expression increases (red) or decreases TG/HDL (blue). (right) Association of tissue-specific predicted gene expression with TG/HDL in the UK Biobank. Direction and magnitude of effect is represented by color, with positive and negative associations of gene expression with TG/HDL shown as blue and red respectively, and darker shades demonstrating stronger effect sizes. Significant associations are outlined in black.


Top-quartile boosted TG/HDL risk loci
Fine-mapped to credible sets
Prioritized risk genes
Exonic variation
Gene expression quantitative trait loci
Tissue-specific heritable gene expression in metabolic tissue(s) [adipose, liver, muscle]
LPL
IRS1
KLF14
COBLL1
PLA2G12A
PLA2G6
RBPJ
ACACB
SNX10
PPARG
KCNH2
FARP2
AXL
RREB1
UBE2K
FGFR2
NEK4
TNFAIP8
RP11–1102P16.1
B3GNT4
FADS2
CD248
VGLL3
RNF168
EYA2
SIPA1
USP3
PEMT
PDE3A
PCCB
NTAN1
RSPO3
APOB
ANGPTL4
UTR variant
LOF harmful
LOF protective
Effect size
20
10
0
–10
–20
p < 0.05
Subcutaneous adipose
Visceral adipose
Liver
Skeletal muscle
Tissue

This gene-level integrative genomics approach successfully re-captured the direction-of-effect and tissue specificity of several well-characterized insulin resistance effector genes including *LPL*, *PPARG* and *ANGPTL4*, as well as a more recently characterized gene *COBLL1*. Gene-level burden tests showed that loss-of-function (LOF) protein-coding variation in *LPL* and *PPARG* is associated with an increase in TG/HDL (LPL: effect size=0.04, p=3.1 x 10^{-23}; PPARG: effect size=0.04, p=5 x 10^{-6}), consistent with the well-characterized roles of LPL and PPARG in insulin resistance(*35*, *44*). At the *ANGPTL4* locus, fine-mapping identified causal SNPs rs116843064 (PIP=1) and rs140744493 (PIP=0.98) which encode missense variants in *ANGPTL4* (p.E40K and p.R336C respectively) that are computationally predicted by Ensembl Variant Effect Predictor (VEP)(*45*) to be deleterious. These SNPs were strongly associated with decreases in TG/HDL (rs116843064: effect size=-0.25, p=1.0 x 10^{-189}; rs140744493: effect size=-0.01, p=1.2 x 10^{-5}) in our data suggesting that LOF in *ANGPTL4* increases insulin sensitivity, a finding congruent with LOF mouse models of *ANGPTL4* resulting in reduced serum triglycerides and improved glucose homeostasis(*22*). At the *COBLL1* locus, our genetically predicted gene expression analysis indicated that increased gene expression of *COBLL1* primarily in adipocytes (subcutaneous: t-statistic=-8.8, p=2 x 10^{-16}; visceral: t-statistic=-9.7, p=4.8 x 10^{-20}; liver: t-statistic=3.0, p=0.03; muscle: t-statistic=-3.5, p=0.007) is associated with decreased TG/HDL which aligns with recent work showing that LOF in *COBLL1* results in insulin resistance through adipose-specific mechanisms(*36*).

Evaluation of TG/HDL associated loci nominate *PLA2G12A, PLA2G6,* and *VGLL3* as novel insulin resistance genes

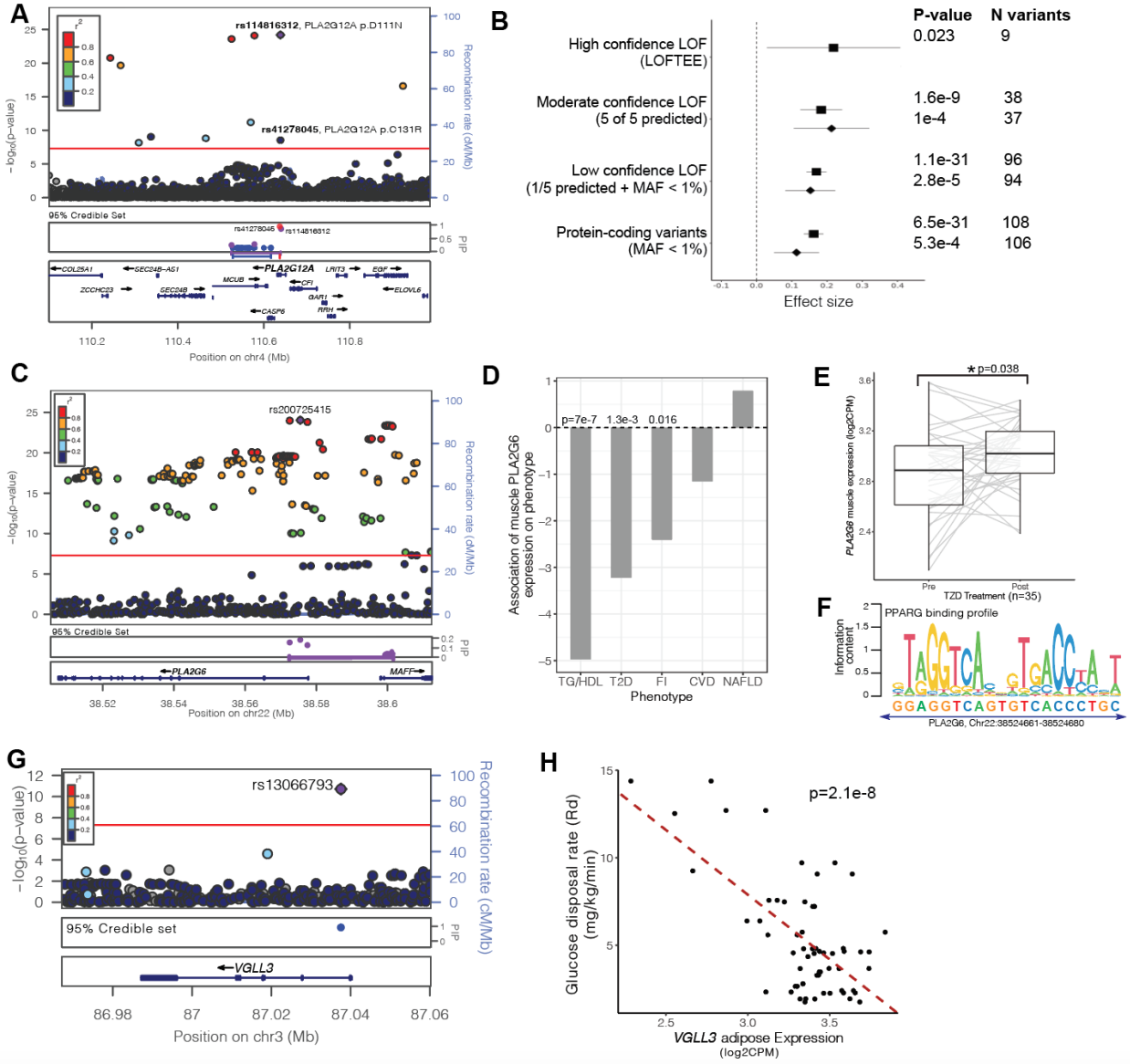
In an effort to identify novel insulin resistance genes we focused our attention on loci that had not previously been associated with insulin resistance and metabolic diseases (Figure 4.3B).

PLA2G12A: One such locus on chromosome 4, defined by lead SNP rs114816312 (effect size=0.1, $p=6 \times 10^{-25}$), was strongly boosted ($TG/HDL_{BS} = 3.7$) and contained many genes (Figure 4.5A). Fine mapping narrowed the credible set of SNPs to overlap with three genes: *PLA2G12A*, *MCUB* and *CASP6*. The lead variants with greatest probability of being causal both encoded missense variants in the *PLA2G12A* gene (rs114816312: p.D111N, PIP=0.7; rs41278045: p.C131R, PIP=1.0) strongly suggesting this as the causal gene (Figure 4.5A). *PLA2G12A* encodes a ubiquitously expressed, secreted member of the family of phospholipase A2 enzymes that hydrolyze phospholipids to release arachidonic acid(*46*). Computational variant effect prediction with Ensembl Variant Effect Predictor (VEP)(*45*) predicted both causal missense variants to be deleterious to PLA2G12A protein function, and both variants were associated with increased TG/HDL levels suggesting that PLA2G12A promotes insulin sensitivity. As *PLA2G12A* was ubiquitously expressed but was not estimated to have significantly heritable gene expression, we therefore could not associate genetically predicted gene expression with TG/HDL values. Rather, we sought to independently corroborate the direction-of-effect with insulin sensitivity by investigating rare coding variation in the gene obtained from the UKBB exomes. Hypothesizing that LOF in *PLA2G12A* increases insulin

resistance, we performed a series of genetic burden tests(*47*) aggregating LOF variants with decreasing stringency and computing association with TG/HDL. The most stringent filter for LOF variants (frameshift, splice, and premature stop variants) as defined by LOFTEE(*48*); n=9) was associated with increased TG/HDL levels (effect size=0.2, p=0.02, Figure 4.5B) supporting our hypothesis. As the stringency was relaxed (Methods), missense variants computationally predicted to be deleterious were included in the LOF group increasing the strength of association between LOF in *PLA2G12A* and increased TG/HDL. Even upon removal of the missense variants rs114816312 and rs41278045 described above, a significant signal for association between LOF in *PLA2G12A* and increased TG/HDL was identified (Figure 4.5B).

Figure 4.5. Evaluation of TG/HDL associated loci nominate PLA2G12A, PLA2G6, and VGLL3 as novel insulin resistance genes.

A) LocusZoom plot of the PLA2G12A TG/HDL locus. The purple diamond represents the lead SNP, and color represents the linkage disequilibrium between each variant with the lead SNP. The PLA2G12A locus contains three 95% credible sets (depicted as red, purple, blue), with the top two causal SNPs rs114816312 (PIP=0.7) and rs41278045 (PIP=1) encoding PLA2G12A missense variants. B) PLA2G12A loss-of-function (LOF) burden tests in the UK Biobank exome sequenced individuals with increasing confidence (stringency) to define LOF variants. Effect sizes and 95% confidence intervals are shown. High confidence LOF variants were defined by LOFTEE (stop, splice, and frameshift), moderate confidence LOF variants were called deleterious by 5 out of 5 bioinformatics prediction tools (see Methods), low confidence LOF variants were those called deleterious by at least 1 bioinformatic prediction tool and had minor allele frequency (MAF) < 1%. Standard burden test results are indicated by squares, and diamonds represent sensitivity analyses removing the top causal variants rs41278045 and rs114816312. P-values and number of variants for each analysis are listed. C) LocusZoom plot and 95% credible set for the PLA2G6 TG/HDL locus. D) Associations of predicted muscle PLA2G6 expression on TG/HDL, type 2 diabetes (T2D), fasting insulin (FI), coronary artery disease (CVD), and non-alcoholic fatty liver disease (NAFLD). Effect sizes are shown, and p-values are given if significant. E) Upregulation of skeletal muscle PLA2G6 upon TZD treatment in 35 biopsied clinical trial participants (fold-change=1.11, *p=0.038). F) PPARG consensus motif overlaid over a high-scoring matching sequence 2231 bp 5' of the PLA2G6 transcription start site. G) LocusZoom plot and 95% credible set for the VGLL3 TG/HDL locus. H) Association of adipose tissue VGLL3 expression with glucose disposal rate (Rd) measured by hyperinsulinemic-euglycemic clamp among 32 biopsied participants (b=-7.4, p=2.1e-8; linear model).


A
rs114816312, PLA2G12A p.D111N
rs41278045, PLA2G12A p.C131R
95% Credible Set
PLA2G12A
Position on chr4 (Mb)
B
P-value
N variants
High confidence LOF (LOFTEE)
0.023
9
Moderate confidence LOF (5 of 5 predicted)
1.6e-9
38
1e-4
37
Low confidence LOF (1/5 predicted + MAF < 1%)
1.1e-31
96
2.8e-5
94
Protein-coding variants (MAF < 1%)
6.5e-31
108
5.3e-4
106
Effect size
C
rs200725415
95% Credible Set
PLA2G6
MAFF
Position on chr22 (Mb)
D
p=7e-7
1.3e-3
0.016
TG/HDL
T2D
FI
CVD
NAFLD
Phenotype
Association of muscle PLA2G6 expression on phenotype
E
*p=0.038
PLA2G6 muscle expression (log2CPM)
Pre
Post
TZD Treatment (n=35)
F
PPARG binding profile
Information content
PLA2G6, Chr22:38524661-38524680
G
rs13066793
95% Credible set
VGLL3
Position on chr3 (Mb)
H
p=2.1e-8
Glucose disposal rate (Rd) (mg/kg/min)
VGLL3 adipose Expression (log2CPM)

PLA2G6: Another novel locus located on chromosome 22 marked by lead SNP rs200725415 (effect size=-0.02 p=9.3 x 10^{-25}) was also strongly boosted (TG/HDL_{BS} = 4.0) (Figure 4.5C). Fine-mapping of this locus produced only a single credible set containing variants with PIP > 0.1 which overlapped two genes *PLA2G6* and *MAFF.* However, all of the variants in the credible set with PIP > 0.1 were located within *PLA2G6* nominating it as the causal gene. *PLA2G6* encodes another member of the family of phospholipase A2 enzymes which hydrolyze phospholipids to release arachidonic acid but functions intracellularly(*49*).

In contrast to *PLA2G12A*, *PLA2G6* was estimated to have significantly heritable gene expression in skeletal muscle (heritability h^2 = 0.04, p=9.9 x 10^{-4}, see web resources(*43*)). Predicted skeletal muscle expression of *PLA2G6* was negatively associated with TG/HDL (t-statistic=-5, p=6.7 x 10^{-7}), suggesting increasing muscle expression of *PLA2G6* would potentially enhance insulin sensitivity. Concordantly, we found that genetically predicted muscle-specific *PLA2G6* expression was negatively associated with FI (t-statistic=-2.4, p=0.016) in the MAGIC study (n=105,056, Figure 4.5D)(*13*). To evaluate the potential consequence of muscle-specific *PLA2G6* expression on metabolic disease we performed a similar association analyses with T2D (n = 180,834 cases vs. 1,159,055 controls)(*33*), CVD (n = 181,522 cases vs. 1,165,690 controls)(*39*), and NAFLD (n = 1,483 cases vs. 17,781 controls)(*32*). We found a significantly negative association of predicted gene expression with T2D only (t-statistic=-3.2, p=0.001, Figure 4.5D). These genetic analyses suggest that the heritable component of *PLA2G6* gene expression in muscle improves insulin sensitivity and reduces T2D risk.

To corroborate these genetic association analyses we examined measured *PLA2G6* gene expression in biopsied skeletal muscle of 36 individuals undergoing insulin sensitizing treatment with PPARG agonists, or thiazolidinediones (TZDs)(*15*). Skeletal muscle *PLA2G6* expression was increased in paired samples after three months of TZD treatment relative to baseline (fold-change=1.11, p=0.038; Figure 4.5E). We also identified a strong PPARG binding motif (JASPAR(*50*): MA0066.1; chr22:38524661-38524680) in the intron of *PLA2G6* located 2231 bp upstream of the transcription start site (Figure 4.5F). Taken together, these data suggest that PLA2G6 may function through muscle to increase insulin sensitivity and decrease T2D risk.

***VGLL3*:** A third novel TG/HDL locus identified on chromosome 3 was defined by lead SNP rs13066793 (effect size=-0.03, p=2 x 10^{-11}, Figure 4.5G) which was located within an intron of the gene *VGLL3* and annotated as an eQTL for *VGLL3* expression in muscle (effect size=-0.25, p=8 x 10^{-7}) and nominally in subcutaneous adipose (effect size=-0.1, p=0.02). VGLL3 e. These data suggest that increased expression of *VGLL3,* which encodes a transcriptional co-activator for TEAD family transcription factors(*51*), increases insulin resistance. We note that despite the highly significant muscle eQTL, *VGLL3* was lowly expressed in human skeletal muscle (median transcripts per million (TPM): muscle = 0.9 TPM) whereas it was highly expressed in human adipose tissues (subcutaneous adipose: 24.8 TPM, visceral adipose: 17.4 TPM, GTEx(*52*)). Thus, we attempted to corroborate adipose *VGLL3* expression with insulin sensitivity in transcriptomically profiled adipose and skeletal muscle tissue biopsies obtained from 32 individuals undergoing hyperinsulinemic-euglycemic glycemic clamps(*15*). We identified a

significantly negative correlation between glucose disposal rate (Rd) and adipose tissue *VGLL3* expression (estimate=-7.3, p=2.1 x 10^{-8}, Figure 4.5H), but no significant correlation with skeletal muscle *VGLL3* expression (estimate=1.5, p=0.33). We additionally adjusted gene expression for subject-specific covariates and found no effect for age or sex but a strong effect for BMI (estimate=-0.26, p=2.8 x 10^{-8}). Formal mediation analysis comparing regression models of Rd against adipose *VGLL3* expression with and without BMI revealed approximately 50% of the effect of *VGLL3* adipose expression on glucose disposal is mediated by BMI. Taken together, these data suggest that *VGLL3* increases insulin resistance by increasing adiposity.

TNFAIP8 increases insulin resistance through alteration of adipose tissue depots in a female-specific manner

Among the TG/HDL "boosted" association signals, a locus on chromosome 5 marked by lead SNP rs1045241 stood out as demonstrating strong evidence for sexual dimorphism indicating a stronger effect in females (t_{sex} = 4.0, p_{sex} = 6.0 x 10^{-5}). Sex-stratified analysis in the female (n=206,012) and male (n=176,117) UKBB participants showed that the association with TG/HDL was highly female-specific with little evidence of association in the subgroup of male participants (effect $size_{female}$= -0.03, p_{female}=1.6 x 10^{-19}; effect $size_{male}$=-0.01, p_{male}=3.5 x10^{-4}; Figure 4.6A). In fact, despite half the sample size, the association was stronger in the subgroup of female participants than in the sex-combined discovery cohort (n=382,129). Fine-mapping at the locus produced a single credible set with six causal variants, all of which were located within the gene *TNFAIP8* (Figure 4.6A, inset panel). The lead variant rs1045241 (chr5:118729286:C:T) had the highest probability of being causal (PIP = 0.24) and was located in the 3' UTR of *TNFAIP8*.

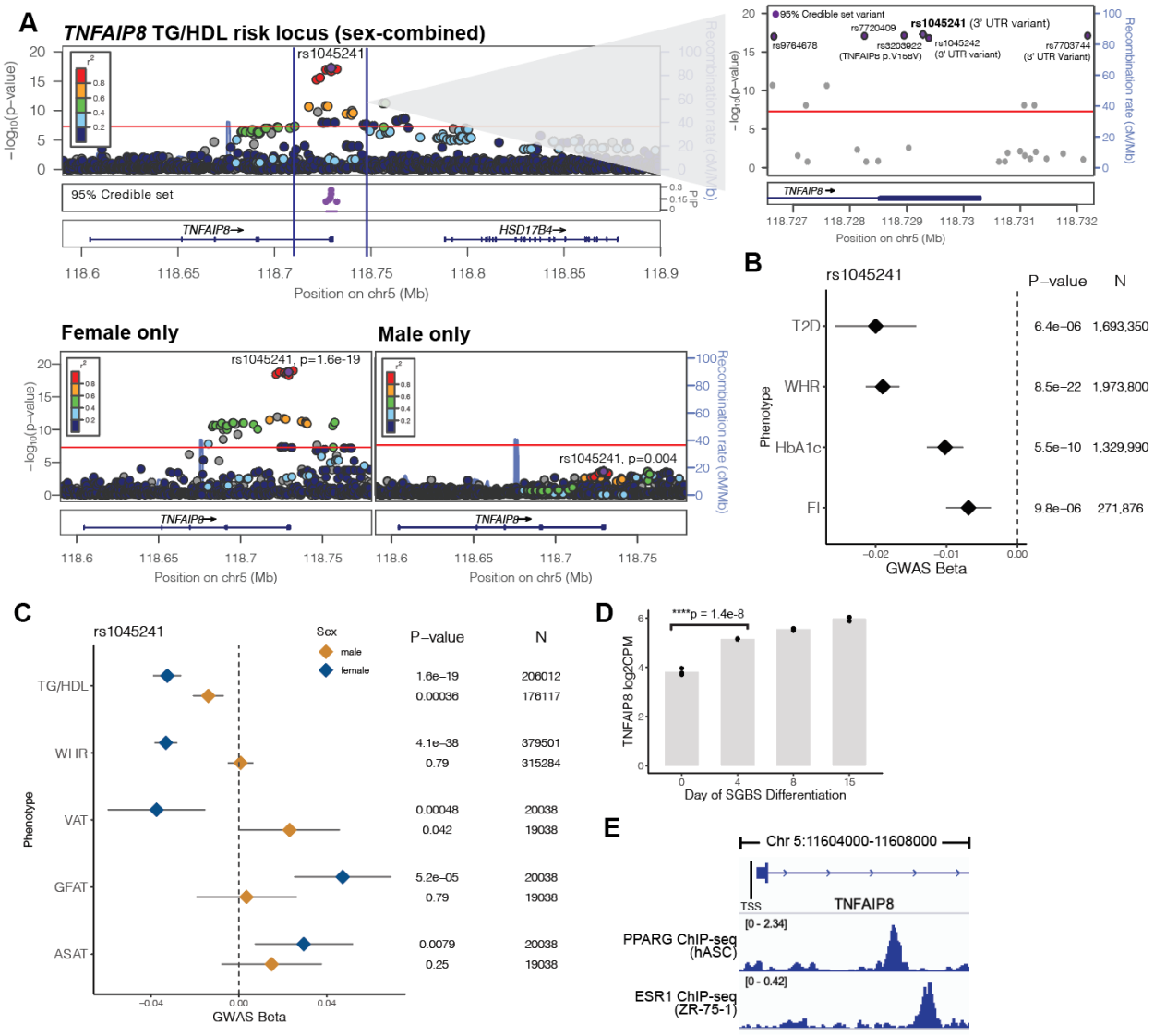
Overall, four of the six variants were located within this last exon of *TNFAIP8*– one coding and three in the 3' UTR (Figure 4.6A) strongly nominating it as the causal gene. *TNFAIP8* encodes the tumor necrosis factor accessory protein 8 which has been found to play an anti-apoptotic role for TNFalpha induced tumor apoptosis(*53*), but has no known role in insulin resistance or metabolic disease.

In a serum proteomics study of 35,559 individuals(*54*), the effect allele (chr5:118729286:C:T) of the lead causal variant rs1045241 which was associated with decreased TG/HDL also was significantly associated with decreased plasma levels of TNFAIP8 (effect size=-0.05, p=8.7 $x10^{-10}$), suggesting TNFAIP8 promotes insulin resistance. To corroborate this direction-of-effect we examined other metabolic disease/trait associations for rs1045241 (meta-analyzed within the Common Metabolic Diseases Knowledge Portal (see Web Resources)) and found that the effect allele (chr5:118729286:C:T) was also associated with decreased waist-to-hip ratio (WHR effect size=-0.02, p=8.5 x 10^{-22}, n=1,973,800), decreased FI (effect size=-0.01, p=9.8 x 10^{-6}, n = 271,876), decreased glycated hemoglobin (HbA1c; effect size=-0.01, p=2.7 x 10^{-9}, n = 1,329,990) and decreased risk for T2D (odds ratio=0.98, p=6.4 x 10^{-6}, n = 1,693,350) (Figure 4.6B). These complementary associations support an insulin sensitizing effect from decreasing serum TNFAIP8 levels.

Figure 4.6. TNFAIP8 increases insulin resistance through alteration of adipose tissue depots in a female-specific manner.

WHR, waist-hip ratio; ASAT, abdominal subcutaneous adipose; VAT, visceral adipose; GFAT, gluteofemoral adipose; T2D, type 2 diabetes; FI, fasting insulin; TG, triglycerides; HDL, HDL cholesterol; GGT, gamma glutamyl transferase; HbA1c, glycated hemoglobin; ApoA, apolipoprotein A.

A) (top) LocusZoom plot of the TG/HDL associated locus containing TNFAIP8 from the sex-combined GWAS. The lead SNP (rs1045241, shown as purple diamond) and GWAS p-value is labeled, and color represents the LD (r2) between each variant with the lead SNP. (inset panel) Zoom in of the 95% credible set region consisting of 6 SNPs (purple). (bottom) LocusZoom plots of the TNFAIP8 TG/HDL locus from the sex-stratified GWAS. The association is only significant in females. B) Association of TNFAIP8 lead SNP rs1045241 in meta-analyzed metabolic disease/trait association studies. T2D, WHR, FI and HbA1c are significantly decreased by the TG/HDL lowering effect allele. Effect estimates (beta) and 95% confidence intervals are plotted along with association p-values and sample sizes. C) Sex-stratified associations of rs1045241 with metabolic phenotypes. Effect estimates (beta) and 95% confidence intervals are shown for female (blue) and male (orange) stratified GWAS, along with association p-values and sample sizes. The TG/HDL lowering effect allele decreases VAT and increases GFAT and ASAT in a female specific manner. D) Upregulation of TNFAIP8 expression in human pre-adipocytes during in vitro differentiation. E) Gene model of TNFAIP8 highlighting the transcriptional start site (TSS) and first intron (top). (second track) PPARG ChIP-seq profiling in human adipocytes (human adipose stem cells, hASC). (bottom) ESR1 ChIP-seq profiling in human breast cell line ZR-75-1.


A
TNFAIP8 TG/HDL risk locus (sex-combined)
rs1045241
95% Credible set
TNFAIP8
HSD17B4
Position on chr5 (Mb)
95% Credible set variant
rs1045241 (3' UTR variant)
rs9764678
rs7720409
rs3203922 (TNFAIP8 p.V156V)
rs1045242 (3' UTR variant)
rs7703744 (3' UTR Variant)
Female only
Male only
rs1045241, p=1.6e-19
rs1045241, p=0.004
B
rs1045241
P–value
N
T2D
WHR
HbA1c
FI
6.4e–06 1,693,350
8.5e–22 1,973,800
5.5e–10 1,329,990
9.8e–06 271,876
Phenotype
GWAS Beta
C
rs1045241
Sex
male
female
P–value
N
TG/HDL
WHR
VAT
GFAT
ASAT
1.6e–19 206012
0.00036 176117
4.1e–38 379501
0.79 315284
0.00048 20038
0.042 19038
5.2e–05 20038
0.79 19038
0.0079 20038
0.25 19038
Phenotype
GWAS Beta
D
****p = 1.4e-8
TNFAIP8 log2CPM
Day of SGBS Differentiation
E
Chr 5:11604000-11608000
TSS
TNFAIP8
PPARG ChIP-seq (hASC)
[0 - 2.34]
ESR1 ChIP-seq (ZR-75-1)
[0 - 0.42]

We further investigated the WHR association where the UK Biobank-GIANT consortium(*12*) had made available sex-stratified summary statistics. rs1045241 had a much stronger association with WHR in females ($p_{female} = 4.1 \times 10^{-38}$) and no evidence of association in males ($p_{male} = 0.8$) (Figure 4.6C). As WHR is a surrogate measure for enhanced visceral adiposity(*55*), we examined the association of rs1045241 with MRI quantified visceral adipose tissue (VAT), gluteofemoral adipose tissue (GFAT), and abdominal subcutaneous adipose tissue (ASAT) from a prior study of 19,038 male and 20,038 female participants in the UKBB(*56*). We found that the effect allele (chr5:118729286:C:T) that lowered both TG/HDL and WHR with a female-biased effect concordantly decreased VAT and increased GFAT in females (effect $size_{VAT} = -0.4$, $p_{VAT} = 4.8 \times 10^{-4}$; effect $size_{GFAT} = 0.05$, $p_{GFAT} = 5.2 \times 10^{-5}$; n = 20,038) but not in men (Figure 4.6C). ASAT was also increased in women (effect $size_{ASAT} = 0.03$, $p_{ASAT} = 7.9 \times 10^{-3}$) but not in men. The sexual dimorphism in these associations was so strong that the association signal was lost in a sex-combined analysis of rs1045241 with VAT (p=0.2, n=39,076) despite doubling the number of samples. These data suggest that TNFAIP8 may operate in adipose tissue in a sex-specific manner to promote insulin resistance.

To corroborate these findings at a cellular and molecular level, we analyzed *TNFAIP8* expression levels during the course of human adipocyte differentiation and found it to be significantly upregulated (fold-change=2.5, $p=1.4 \times 10^{-8}$) during the conversion of pre-adipocytes into mature adipocytes *in vitro(57)* (Figure 4.6D). We analyzed data of chromatin immunoprecipitation and DNA sequencing (ChIP-seq) of PPARG(*58*), a master regulator of adipocyte differentiation(*59*), and found a strong peak of binding (chr5:118606764-118607039) 2377 base pairs downstream of the *TNFAIP8*

transcriptional start site in the first intron (Figure 4.6E). Given the female-specific association of TNFAIP8, we next examined publicly available estrogen receptor (ESR1) ChIP-seq data at the *TNFAIP8* locus and identified a strong binding peak of binding (chr5:118607354-118607660) in human breast cell line ZR-75-1(*60*) approximately 600 base pairs downstream of the PPARG binding site. This points to a potential molecular mechanism for the female-specific signals of association observed between *TNFAIP8* and metabolic traits.

Discussion

In this study we conducted the largest GWAS to date for TG/HDL, a surrogate marker of insulin resistance, identifying 251 associated loci, of which 62 associated more strongly with TG/HDL compared to TG or HDL alone, and 118 which had not been previously reported in prior insulin resistance GWAS studies(*11–13*, *33*, *39*). At 50 of these loci we identified causal variants and genes performing integrative genomic analysis to propose directions-of-effect and tissue of action. For selected genes we independently corroborate these genetic findings by interrogating gene expression in biopsied skeletal muscle and adipose tissue in individuals with directly measured insulin resistance by glycemic clamp. We highlight two phospholipase encoding genes *PLA2G12A* and *PLA2G6* which enhance insulin sensitivity, a transcriptional co-activator VGLL3 which decreases insulin sensitivity, and a TNFα accessory protein TNFAIP8 which decreases insulin sensitivity specifically in females.

The re-identification of loci known to encode insulin signaling components (e.g. *INSR*, *IRS1*) as well as those identified by previous surrogate biomarker studies confirms the fidelity of TG/HDL as a surrogate marker of insulin resistance providing confidence

that the novel loci identified are true mediators of insulin resistance. Compared to prior surrogate marker studies, the new signals identified in our study could arise for two reasons– 1) sample size and 2) physiological/biological heterogeneity. Regarding sample size, our study (n=382,129) is the largest among serum derived biomarkers (FI: n=105,056(*13*)), but is exceeded in sample by anthropometric biomarker studies (WHR: n=694,649(*12*)). If sample size related statistical power were a major determinant of locus identification, we would expect that association strength for known causal insulin resistance genes (e.g. *PPARG*, *IRS1*) would be proportional to study size, but in fact found no such relationship for *PPARG* ($p_{PPARG\text{-}FI} = 1.5 \times 10^{-21}$, $p_{PPARG\text{-}TG/HDL} = 2 \times 10^{-17}$, $p_{PPARG\text{-}WHR} = 3.8 \times 10^{-25}$) or *IRS1* ($p_{IRS1\text{-}FI} = 8.5 \times 10^{-39}$, $p_{IRS1\text{-}TG/HDL} = 2.5 \times 10^{-85}$, $p_{IRS1\text{-}WHR} = 1.6 \times 10^{-9}$). Regarding biological/physiological heterogeneity, insulin is known to signal in a tissue specific manner– e.g. the level of insulin required physiologically to suppress ketogenesis or lipolysis is orders of magnitude less than required to induce glucose uptake(*61*). Thus, each surrogate biomarker may capture overlapping but distinct aspects of insulin resistance physiology. Surrogate biomarker GWAS demonstrating differing patterns of tissue enrichments of associated loci support this idea. The loci associated with TG/HDL in our study are most highly enriched in adipose tissue followed by liver and skeletal muscle whereas for FI (skeletal muscle enrichment = adipose > liver) and for FG (liver enrichment > skeletal muscle > adipose)(*10*). Taken together these observations suggest surrogate biomarker studies of insulin resistance capture complementary aspects of physiology and distinct loci from these studies could point to new mechanisms of molecular pathogenesis.

Among the associated loci uniquely associated with TG/HDL, we prioritized two phospholipase A2 enzymes, PLA2G12A and PLA2G6, that function in the release of arachidonic acid, a precursor of bioactive lipid molecules eicosanoids(*62*). Eicosanoids are known to be involved in inflammation and insulin sensitivity(*63*), suggesting dysregulation of PLA2s may increase insulin resistance and related metabolic disorders via disruption of eicosanoid signaling. Through examination of rare protein-coding variation from exome sequencing, we find that loss-of-function in PLA2G12A increases insulin resistance as measured by TG/HDL. PLA2G12A is secreted to the blood, ubiquitously expressed across tissues, and does not have any tissue-specific associations, indicating that it may function across several tissues to affect whole body insulin sensitization. On the other hand, our analyses propose a tissue-specific role for PLA2G6 in muscle, under the regulation of PPARG to increase insulin sensitivity and decrease T2D risk (Figure 4 D,E,F). The metabolic role of PLA2G6 in pancreatic islets has previously been examined *in vitro* and in murine models(*64*), but to our knowledge this is the first study implicating it in human insulin resistance.

Another novel insulin resistance effector gene prioritized is *VGLL3*, a transcriptional cofactor(*65*) whose adipose expression is strongly correlated with glucose disposal rate in humans suggesting *VGLL3* promotes insulin resistance. Prior characterization in mice has suggested VGLL3 protein plays a role in adipogenesis(*51*), a result further corroborated by our finding that BMI mediates a substantial portion of the effect of VGLL3 on insulin sensitivity. Taken together, these findings suggest that VGLL3 influences metabolic disease risk via adiposity regulation and nominate VGLL3 as a potential therapeutic target for pharmacological inhibition to treat metabolic syndrome.

Investigation into the sex-dimorphism of TG/HDL through sex-stratified GWAS revealed *TNFAIP8* as a novel female-specific insulin resistance gene. TNFAIP8 plays a crucial role in regulating cell survival and its transcription is induced by the TNFα signaling cascade(*66*). Previous studies have found *TNFAIP8* is a regulator of insulin-mediated inhibition of autophagy and has binding affinity for fatty acids including oleic and palmitic acids(*53*). Though *TNFAIP8* has been implicated previously in WHR GWAS(*12*), it has not been nominated as a sex-specific gene nor has its role in insulin resistance been fully characterized. We find *TNFAIP8* increases insulin resistance through alteration of adipose tissue depots in a female-specific manner however further investigation is warranted to elucidate specific molecular mechanisms.

Our study has several important limitations that should be considered. First, we used the TG/HDL ratio as a surrogate marker of insulin resistance, which may not reflect the true causal relationship between genetic variants and insulin resistance as is the case with other surrogate marker GWAS(*11–13*). We believe this trade off is reasonable given the vast increase in sample size possible through the use of TG/HDL and its strong phenotypic and genetic correlations with directly measured insulin resistance and metabolic diseases (Figure 4.1 and Figure 4.2). Moreover, the TG/HDL ratio is influenced by the individual levels of TG and HDL, which share loci with the TG/HDL ratio. Therefore, we prioritized “boosted” genes that showed stronger association with the TG/HDL ratio than with TG or HDL alone, but we cannot exclude the possibility that some of these genes are also involved in lipid metabolism (e.g. APOB, TG/HDL_{BS} = 5.5). Second, we inferred the tissue-specificity of the genes based on their heritable expression patterns in metabolic tissues, which have some limitations. For example, gene expression can be

correlated across different tissues, making it difficult to identify the primary tissue of action for each gene(*67*). Also, due to linkage disequilibrium the genetic determinants of gene expression (eQTLs) may not be specific to the gene of interest, but may tag other nearby or distant genes that have functional effects on insulin resistance. Furthermore, the heritability of gene expression is sensitive to sample size and may be low or nonsignificant for many genes in relevant metabolic tissues. In our analysis we find nonsignificant heritability estimates for over half (52.5%) of our 50 genes of interest in relevant metabolic tissues, not a surprising proportion given less than half of all genes have significantly heritable gene expression in blood at existing sample sizes(*68*). Therefore, while we were able to corroborate tissue specificity for several genes highlighted in this study (*PLA2G6*, *VGLL3*, *TNFAIP8*) our tissue specificity analysis cannot be considered comprehensive for the nominated genes. Further study of these genes is needed to elucidate the causal mechanisms and target tissues underlying modulation of insulin resistance.

In conclusion, our study reports on GWAS of TG/HDL ratio at population scale expanding our knowledge of the genetic determinants of insulin resistance to include dozens of new loci/genes not previously implicated. At four loci *PLA2G12A, PLA2G6, VGLL3, and TNFAIP8,* we validate direction-of-effect, potential regulatory mechanisms, and sex-specific effects. These findings provide specific, testable hypotheses for future investigation to credential these genes as potential new therapeutic targets and/or biomarkers for insulin resistance.

Web resources

- Common Metabolic Diseases Knowledge Portal (cmdkp.org) https://hugeamp.org/variant.html?variant=rs1045241

- Functional Summary-based Imputation (FUSION) GTEx v8 multi-tissue expression http://gusevlab.org/projects/fusion/#gtex-v8-multi- tissue-expression

Acknowledgements

We would like to thank Jerrold M. Olefsky for helpful discussions and assistance with data acquisition of clinical trial participants. Analysis of UK Biobank data was conducted under application number 51436.

Author Contributions

N.D., G.M.P., and A.R.M. designed the study. N.D., Z.Z., J.S.D., P.N., J.F., generated/acquired and analyzed the data. N.D., T.A., G.M.P., and A.R.M., were involved in data interpretation. N.D. and A.R.M. drafted the manuscript. N.D., J.S.D., P.N., J.F., T.A., G.M.P., and A.R.M. were involved in critical manuscript revision.

Methods

Phenotypic correlations of TG/HDL with insulin resistance biomarkers in a clinical cohort

To assess the relationship between TG/HDL and other surrogate insulin resistance phenotypes, we examined a clinical cohort of 45 human participants who underwent hyperinsulinemic-euglycemic clamp and additionally had other metabolic phenotypes measured(*15*). Pairwise correlations of TG/HDL with glucose disposal rate measured by hyperinsulinemic-euglycemic clamp, fasting insulin, and HOMA-IR were assessed using Spearman rank correlation.

Genome-wide association analyses

UK Biobank data

The UK Biobank is a prospective cohort study with genotypic and phenotypic data that enrolled approximately 500,000 individuals aged 40-69 from across the United Kingdom(*69*). To remove potential outliers and individuals with poor data quality, we applied several sample level filters to the UK Biobank participants following the GLGC standards(*16*). These included retaining only those individuals with a sample call rate greater than 95% and with heterozygosity values less than the median + 3x the interquartile range (IQR), and excluding those with reported gender-genetic sex mismatch. After filtering, 382,129 UK Biobank participants with TG and HDL measurements available remained of which 206,012 were female and 176,117 were male.

For quality control of SNPs from the UK Biobank genotypes, we included only variants with SNP call rate > 98%, MAF > 0.001, HWE $p > 5 \times 10^{-8}$, and for imputed SNPs included those with INFO scores > 0.4. We excluded duplicated SNPs and removed any monomorphic markers. After filtering, 489,897 genotype array SNPs and 12,024,213 imputed SNPs remained.

Association testing

REGENIE was run following the published recommendations for UK Biobank analysis (https://rgcgithub.github.io/regenie/recommendations/) to conduct genome-wide association analyses for TG/HDL, TG, and HDL using the 382,129 participants and 12,024,213 SNPs which passed QC. In brief, TG and TG/HDL were natural log transformed, and all phenotypes (TG, HDL, and TG/HDL) were inverse rank normalized

using the REGENIE --apply-rint parameter. In the statistical models, the covariates included were age, age^2, sex, and the first 20 principal components of ancestry to correct for population stratification. As recommended, Step 1 of REGENIE was first run using the 489,897 genotype array SNPs which passed QC to fit the whole genome regression model to the phenotypes and produce genomic predictions, then Step 2 was run to perform association testing of the 12,024,213 imputed SNPs. We next estimated the genomic inflation for the TG/HDL GWAS (λ=1.35) and subsequently performed correction for inflation using the genomic control method(*70*). Lead SNPs and genomic risk loci from the conducted GWAS were then defined using the FUMA platform using the default parameters(*18*, *70*).

Replication GWAS for TG/HDL in Mass General Biobank (MGBB)

GWAS for TG/HDL was conducted in the MGBB (n=37,545) using consistent methodology as conducted in the UK Biobank. Lead SNPs of each of the 251 TG/HDL genomic risk loci were examined in the MGBB, and if the lead SNP was not present in the MGBB, the next most significant SNP within the locus clump was used as a proxy SNP for replication. Statistical power for replication in the MGBB was calculated using the power calculator for genetic association studies (R genpwr v1.0.4)(*71*).

Sex differences in TG/HDL genetic effects

Sex-specific GWAS for TG/HDL were conducted following the same methods above except with sex removed as a covariate and the --sex-specific female/male REGENIE parameter added. To quantify sex differences in genetic effects of each

TG/HDL risk loci defined in the sex-combined GWAS, we calculated the t-statistic and p-value (by two-tailed Student's t-test) for the lead SNP of each locus using the sex-stratified GWAS beta estimates and standard errors as previously described(*72*).

Tissue enrichment of TG/HDL genomic risk loci

We used GREGOR (v.1.4.0)(*25*) to determine the enrichment of the identified TG/HDL genomic risk loci that overlapped tissue-specific stretch enhancers, using the lead SNP at each locus as input. Genomic regions of stretch enhancers in each tissue were defined as in Chen et al. 2021(*10*). The GREGOR default parameters were used to run enrichment analyses (r^2 threshold= 0.8, LD window size = 1 Mb, and minimum neighbor number = 500).

Examining the genetic overlap between TG/HDL and insulin resistance phenotypes

We assessed the genetic correlation between TG/HDL and insulin resistance phenotypes using cross-trait LD Score regression with the default settings as instantiated in LD SCore (LDSC v1.0.1)(*31*, *73*). All association statistics were harmonized prior to computing genetic correlations, ensuring consistent genome build and defined effect alleles across studies.

We further examined the positional overlap between genomic risk loci associated with TG/HDL with other previously identified insulin resistance related genetic risk loci. Overlap of genomic risk loci was defined if a lead SNP defining an insulin resistance related locus fell within the locus boundaries as defined by FUMA of a TG/HDL genomic risk locus identified in our GWAS.

Computing boost score

To quantify the difference in association between TG/HDL compared to TG or HDL alone a SNP-wise "boost score" (TG/HDL_{BS}) was computed from GWAS performed in the UK Biobank for TG, HDL and TG/HDL. For each lead SNP in the TG/HDL study, the more significant association p-value (-log10(p)) between TG or HDL was subtracted from the -log10(p-value) of the TG/HDL association.

$$TG/HDL_{BS} = -\log10(p_{TG/HDL}) - \max\{-\log10(p_{TG}), -\log10(p_{HDL})\}$$

Fine-mapping and nomination of top causal gene at top quartile boosted loci

To determine causal SNPs at each TG/HDL genomic risk locus, statistical fine-mapping was performed using the "Sum of Single Effects" (SuSie) model(*40*) using default parameters. UK Biobank genotypes were down-sampled to 10,000 individuals to be used as an LD reference, and regions for fine-mapping risk loci were defined by taking a 1 Mb window around the lead SNP of each locus. Upon fine-mapping each region, 95% credible sets were defined and posterior inclusion probabilities (PIP) per SNP in the fine-mapping region were computed to evaluate variant causality.

To nominate the top causal gene at each locus we employed the combined SNP-to-gene linking strategy elaborated by Gazal et al.(*41*), integrating genome annotations (i.e. exonic, promoter regions) and functional annotations (i.e. fine-mapped cis-eQTL, regulatory annotations) with the causal variants in each locus determined by fine-mapping. We were able to apply this SNP-to-gene method on the 50 fine-mapped loci which contained a credible set with at least one SNP with PIP > 0.1. 23 of these 50 loci

resulted in a predicted causal SNP-gene pair and the remaining 27 loci were assigned the nearest gene to the SNP with the highest PIP.

LOF burden tests

Results from gene-level burden tests of predicted LOF protein-coding variants (pLOF mask) in nominated genes of interest with TG/HDL among the exome-sequenced individuals in the UK Biobank(*42*) (n=343,470 with TG/HDL measurements available) was extracted from https://app.genebass.org/ (Accessed: June 16, 2022).

To perform LOF burden tests for *PLA2G12A* in the UK Biobank, we used different definitions for LOF variants (masks) with increasing stringency to classify variants, with the most stringent mask being LOF variants as defined by LOFTEE(*48*) ("High confidence LOF variants"). In subsequent masks of decreasing LOF stringency, the following five computational prediction tools were used: Polyphen2_HDIV(*74*), Polyphen2_HVAR(*74*), SIFT(*75*), LRT(*76*), and MutationTaster(*77*). "Moderate confidence LOF variants" included variants predicted deleterious by all five tools, "Low confidence LOF variants" included rare (MAF < 1%) variants predicted deleterious by at least 1 tool, and the least stringent mask included all rare variants (MAF < 1%).

Assessment of directional effects using functional genomic annotations of causal SNPs

To review the consequences on protein function and gene expression of prioritized genes on TG/HDL, we examined causal fine-mapped variants (PIP > 0.1) which were annotated to be exonic or fine-mapped eQTLs. For causal eQTLs, the direction of effect was determined based on the variant's effect on gene expression (eQTL beta) and the

effect on TG/HDL (GWAS beta) in the respective study which the eQTL was fine-mapped for (GTEx, eQTLGen). Similarly, the functional consequence of each causal coding variant on protein function was predicted using established computational tools PolyPhen-2(*74*) and SIFT(*75*) as determined by Ensembl Variant Effect Predictor (VEP)(*45*) and then compared with the direction of effect of the variant on TG/HDL (GWAS beta).

Associations of heritable gene expression with TG/HDL in the UK Biobank

To examine the effect of heritable gene expression on TG/HDL for the top nominated genes of interest in relevant tissues (subcutaneous and visceral adipose, liver, and skeletal muscle), we first assigned per-tissue gene expression scores to each UK Biobank participant using established methods (Functional Summary-based Imputation [FUSION])(*43*). In brief, gene expression reference weights from the Genotype-Tissue Expression (GTEx) Project v8 for tissues of interest were extracted from the FUSION archive (http://gusevlab.org/projects/fusion/#gtex-v8-multi- tissue-expression). The expression weights per gene and their corresponding imputed genotypes in the UK Biobank genotyped population were combined into a predicted gene expression score for each individual by computing a linear combination across the SNPs expression weights and genotypes, then dividing by the number of non-missing SNPs as implemented in the Plink score utility(*78*). TG/HDL measurements were normalized as described above and regressed onto predicted gene expression scores as done previously(*79*).

Association of heritable *PLA2G6* muscle expression on metabolic outcomes

Using the same *PLA2G6* muscle expression weights as described previously, we investigated the role of predicted *PLA2G6* muscle expression on metabolic outcomes T2D(*33*), FI(*13*), CVD(*39*), and NAFLD(*32*) using the FUSION association test function (FUSION.assoc_test) with default parameters(*43*).

Differential expression of *PLA2G6* in skeletal muscle upon TZD treatment

Skeletal muscle biopsies were obtained from individuals before and after three months of TZD treatment (n=35) and RNA-sequenced(*15*). Raw reads were aligned using Kallisto(*80*) with default parameters to generate gene counts. Differential expression analysis for *PLA2G6* upon TZD treatment was performed in dream (variancePartition v1.20.0, edgeR v3.32.1)(*81*) with pre/post TZD treatment as the fixed effect and patient ID as a random effect to account for paired samples.

Correlations between adipose tissue and skeletal muscle *VGLL3* expression and glucose disposal rate in human participants

Subcutaneous adipose tissue and skeletal muscle biopsies were obtained and RNA-sequenced, and clinical measurements of insulin sensitivity along with other metabolic health metrics were assessed in a clinical cohort (n=35) as previously described(*15*). Raw reads were aligned using Kallisto(*80*) with default parameters to generate gene counts. *VGLL3* adipose expression levels were regressed on glucose disposal rate measured by hyperinsulinemic-euglycemic clamp in the 35 individuals using linear regression (lm function, R stats v3.5.1). To avoid confounding effects as a result of insulin sensitizing treatment tested in this clinical trial, only data from tissue biopsies

obtained before treatment were used in this analysis. Mediation analysis to quantify the proportion of effect of *VGLL3* expression (exposure) on Rd (outcome) mediated by BMI (mediator) was conducted using the mediate function as implemented in R package mediation v4.5.0(*82*) using linear models of mediator regressed onto exposure and outcome regressed on exposure + mediator.

RNA-sequencing of human adipocytes through differentiation

Human Simpson-Golabi-Behmel syndrome (SGBS) preadipocyte cells were differentiated according to published protocols(*57*). RNA was extracted from cells at days 0, 4, 8, and 15 and RNA-sequenced. Raw reads were aligned using Kallisto(*80*) with default parameters to generate gene counts. Differential expression analysis for TNFAIP8 between differentiation days 0 and 4 was performed in dream as described previously.

Chapter 4, in full, is currently being prepared for submission for publication of the material. DeForest, Natalie; Zhu, Zhiyi; Dron, Jacqueline S.; Natarajan, Pradeep; Flannick, Jason; Amariuta, Tiffany; Peloso, Gina M.; Majithia, Amit R. The dissertation author was the primary investigator and author of this paper.

References

1. E. Orgel, S. D. Mittelman, The links between insulin resistance, diabetes, and cancer. *Curr. Diab. Rep.* **13**, 213–222 (2013).

2. G. Reaven, Insulin resistance and coronary heart disease in nondiabetic individuals. *Arterioscler. Thromb. Vasc. Biol.* **32**, 1754–1759 (2012).

3. A. E. Brown, M. Walker, Genetics of Insulin Resistance and the Metabolic Syndrome. *Curr. Cardiol. Rep.* **18**, 75 (2016).

4. L. J. Rasmussen-Torvik, J. S. Pankow, D. R. Jacobs, L. M. Steffen, A. M. Moran, J. Steinberger, A. R. Sinaiko, Heritability and genetic correlations of insulin sensitivity measured by the euglycaemic clamp. *Diabet. Med.* **24**, 1286–1289 (2007).

5. X. Guo, J. Cui, M. R. Jones, T. Haritunians, A. H. Xiang, Y.-D. I. Chen, K. D. Taylor, T. A. Buchanan, R. C. Davis, W. A. Hsueh, L. J. Raffel, J. I. Rotter, M. O. Goodarzi, Insulin clearance: confirmation as a highly heritable trait, and genome-wide linkage analysis. *Diabetologia*. **55**, 2183–2192 (2012).

6. S. George, J. J. Rochford, C. Wolfrum, S. L. Gray, S. Schinner, J. C. Wilson, M. A. Soos, P. R. Murgatroyd, R. M. Williams, C. L. Acerini, D. B. Dunger, D. Barford, A. M. Umpleby, N. J. Wareham, H. A. Davies, A. J. Schafer, M. Stoffel, S. O'Rahilly, I. Barroso, A family with severe insulin resistance and diabetes due to a mutation in AKT2. *Science*. **304**, 1325–1328 (2004).

7. B. D. Mitchell, D. Zaccaro, L. E. Wagenknecht, A. L. Scherzinger, R. N. Bergman, S. M. Haffner, J. Hokanson, J. M. Norris, J. I. Rotter, M. F. Saad, Insulin sensitivity, body fat distribution, and family diabetes history: the IRAS Family Study. *Obes. Res.* **12**, 831–839 (2004).

8. B. C. Martin, J. H. Warram, B. Rosner, S. S. Rich, J. S. Soeldner, A. S. Krolewski, Familial clustering of insulin sensitivity. *Diabetes*. **41**, 850–854 (1992).

9. E. Uffelmann, Q. Q. Huang, N. S. Munung, J. de Vries, Y. Okada, A. R. Martin, H. C. Martin, T. Lappalainen, D. Posthuma, Genome-wide association studies. *Nat. Rev. Methods Primers*. **1** (2021), doi:10.1038/s43586-021-00056-9.

10. J. Chen, C. N. Spracklen, G. Marenne, A. Varshney, L. J. Corbin, J. 'an Luan, S. M. Willems, Y. Wu, X. Zhang, M. Horikoshi, T. S. Boutin, R. Mägi, J. Waage, R. Li-Gao, K. H. K. Chan, J. Yao, M. D. Anasanti, A. Y. Chu, A. Claringbould, J. Heikkinen, J. Hong, J.-J. Hottenga, S. Huo, M. A. Kaakinen, T. Louie, W. März, H. Moreno-Macias, A. Ndungu, S. C. Nelson, I. M. Nolte, K. E. North, C. K. Raulerson, D. Ray, R. Rohde, D. Rybin, C. Schurmann, X. Sim, L. Southam, I. D. Stewart, C. A. Wang, Y. Wang, P. Wu, W. Zhang, T. S. Ahluwalia, E. V. R. Appel, L. F. Bielak, J. A. Brody, N. P. Burtt, C. P. Cabrera, B. E. Cade, J. F. Chai, X. Chai, L.-C. Chang, C.-H. Chen, B. H. Chen, K. N. Chitrala, Y.-F. Chiu, H. G. de Haan, G. E.

Delgado, A. Demirkan, Q. Duan, J. Engmann, S. A. Fatumo, J. Gayán, F. Giulianini, J. H. Gong, S. Gustafsson, Y. Hai, F. P. Hartwig, J. He, Y. Heianza, T. Huang, A. Huerta-Chagoya, M. Y. Hwang, R. A. Jensen, T. Kawaguchi, K. A. Kentistou, Y. J. Kim, M. E. Kleber, I. K. Kooner, S. Lai, L. A. Lange, C. D. Langefeld, M. Lauzon, M. Li, S. Ligthart, J. Liu, M. Loh, J. Long, V. Lyssenko, M. Mangino, C. Marzi, M. E. Montasser, A. Nag, M. Nakatochi, D. Noce, R. Noordam, G. Pistis, M. Preuss, L. Raffield, L. J. Rasmussen-Torvik, S. S. Rich, N. R. Robertson, R. Rueedi, K. Ryan, S. Sanna, R. Saxena, K. E. Schraut, B. Sennblad, K. Setoh, A. V. Smith, T. Sparsø, R. J. Strawbridge, F. Takeuchi, J. Tan, S. Trompet, E. van den Akker, P. J. van der Most, N. Verweij, M. Vogel, H. Wang, C. Wang, N. Wang, H. R. Warren, W. Wen, T. Wilsgaard, A. Wong, A. R. Wood, T. Xie, M. H. Zafarmand, J.-H. Zhao, W. Zhao, N. Amin, Z. Arzumanyan, A. Astrup, S. J. L. Bakker, D. Baldassarre, M. Beekman, R. N. Bergman, A. Bertoni, M. Blüher, L. L. Bonnycastle, S. R. Bornstein, D. W. Bowden, Q. Cai, A. Campbell, H. Campbell, Y. C. Chang, E. J. C. de Geus, A. Dehghan, S. Du, G. Eiriksdottir, A. E. Farmaki, M. Frånberg, C. Fuchsberger, Y. Gao, A. P. Gjesing, A. Goel, S. Han, C. A. Hartman, C. Herder, A. A. Hicks, C.-H. Hsieh, W. A. Hsueh, S. Ichihara, M. Igase, M. A. Ikram, W. C. Johnson, M. E. Jørgensen, P. K. Joshi, R. R. Kalyani, F. R. Kandeel, T. Katsuya, C. C. Khor, W. Kiess, I. Kolcic, T. Kuulasmaa, J. Kuusisto, K. Läll, K. Lam, D. A. Lawlor, N. R. Lee, R. N. Lemaitre, H. Li, Lifelines Cohort Study, S.-Y. Lin, J. Lindström, A. Linneberg, J. Liu, C. Lorenzo, T. Matsubara, F. Matsuda, G. Mingrone, S. Mooijaart, S. Moon, T. Nabika, G. N. Nadkarni, J. L. Nadler, M. Nelis, M. J. Neville, J. M. Norris, Y. Ohyagi, A. Peters, P. A. Peyser, O. Polasek, Q. Qi, D. Raven, D. F. Reilly, A. Reiner, F. Rivideneira, K. Roll, I. Rudan, C. Sabanayagam, K. Sandow, N. Sattar, A. Schürmann, J. Shi, H. M. Stringham, K. D. Taylor, T. M. Teslovich, B. Thuesen, P. R. H. J. Timmers, E. Tremoli, M. Y. Tsai, A. Uitterlinden, R. M. van Dam, D. van Heemst, A. van Hylckama Vlieg, J. V. van Vliet-Ostaptchouk, J. Vangipurapu, H. Vestergaard, T. Wang, K. Willems van Dijk, T. Zemunik, G. R. Abecasis, L. S. Adair, C. A. Aguilar-Salinas, M. E. Alarcón-Riquelme, P. An, L. Aviles-Santa, D. M. Becker, L. J. Beilin, S. Bergmann, H. Bisgaard, C. Black, M. Boehnke, E. Boerwinkle, B. O. Böhm, K. Bønnelykke, D. I. Boomsma, E. P. Bottinger, T. A. Buchanan, M. Canouil, M. J. Caulfield, J. C. Chambers, D. I. Chasman, Y.-D. I. Chen, C.-Y. Cheng, F. S. Collins, A. Correa, F. Cucca, H. J. de Silva, G. Dedoussis, S. Elmståhl, M. K. Evans, E. Ferrannini, L. Ferrucci, J. C. Florez, P. W. Franks, T. M. Frayling, P. Froguel, B. Gigante, M. O. Goodarzi, P. Gordon-Larsen, H. Grallert, N. Grarup, S. Grimsgaard, L. Groop, V. Gudnason, X. Guo, A. Hamsten, T. Hansen, C. Hayward, S. R. Heckbert, B. L. Horta, W. Huang, E. Ingelsson, P. S. James, M.-R. Jarvelin, J. B. Jonas, J. W. Jukema, P. Kaleebu, R. Kaplan, S. L. R. Kardia, N. Kato, S. M. Keinanen-Kiukaanniemi, B.-J. Kim, M. Kivimaki, H. A. Koistinen, J. S. Kooner, A. Körner, P. Kovacs, D. Kuh, M. Kumari, Z. Kutalik, M. Laakso, T. A. Lakka, L. J. Launer, K. Leander, H. Li, X. Lin, L. Lind, C. Lindgren, S. Liu, R. J. F. Loos, P. K. E. Magnusson, A. Mahajan, A. Metspalu, D. O. Mook-Kanamori, T. A. Mori, P. B. Munroe, I. Njølstad, J. R. O'Connell, A. J. Oldehinkel, K. K. Ong, S. Padmanabhan, C. N. A. Palmer, N. D. Palmer, O. Pedersen, C. E. Pennell, D. J. Porteous, P. P. Pramstaller, M. A. Province, B. M. Psaty, L. Qi, L. J. Raffel, R. Rauramaa, S.

Redline, P. M. Ridker, F. R. Rosendaal, T. E. Saaristo, M. Sandhu, J. Saramies, N. Schneiderman, P. Schwarz, L. J. Scott, E. Selvin, P. Sever, X.-O. Shu, P. E. Slagboom, K. S. Small, B. H. Smith, H. Snieder, T. Sofer, T. I. A. Sørensen, T. D. Spector, A. Stanton, C. J. Steves, M. Stumvoll, L. Sun, Y. Tabara, E. S. Tai, N. J. Timpson, A. Tönjes, J. Tuomilehto, T. Tusie, M. Uusitupa, P. van der Harst, C. van Duijn, V. Vitart, P. Vollenweider, T. G. M. Vrijkotte, L. E. Wagenknecht, M. Walker, Y. X. Wang, N. J. Wareham, R. M. Watanabe, H. Watkins, W. B. Wei, A. R. Wickremasinghe, G. Willemsen, J. F. Wilson, T.-Y. Wong, J.-Y. Wu, A. H. Xiang, L. R. Yanek, L. Yengo, M. Yokota, E. Zeggini, W. Zheng, A. B. Zonderman, J. I. Rotter, A. L. Gloyn, M. I. McCarthy, J. Dupuis, J. B. Meigs, R. A. Scott, I. Prokopenko, A. Leong, C.-T. Liu, S. C. J. Parker, K. L. Mohlke, C. Langenberg, E. Wheeler, A. P. Morris, I. Barroso, Meta-Analysis of Glucose and Insulin-related Traits Consortium (MAGIC), The trans-ancestral genomic architecture of glycemic traits. *Nat. Genet.* **53**, 840–860 (2021).

11. J. Dupuis, C. Langenberg, I. Prokopenko, R. Saxena, N. Soranzo, A. U. Jackson, E. Wheeler, N. L. Glazer, N. Bouatia-Naji, A. L. Gloyn, C. M. Lindgren, R. Mägi, A. P. Morris, J. Randall, T. Johnson, P. Elliott, D. Rybin, G. Thorleifsson, V. Steinthorsdottir, P. Henneman, H. Grallert, A. Dehghan, J. J. Hottenga, C. S. Franklin, P. Navarro, K. Song, A. Goel, J. R. B. Perry, J. M. Egan, T. Lajunen, N. Grarup, T. Sparsø, A. Doney, B. F. Voight, H. M. Stringham, M. Li, S. Kanoni, P. Shrader, C. Cavalcanti-Proença, M. Kumari, L. Qi, N. J. Timpson, C. Gieger, C. Zabena, G. Rocheleau, E. Ingelsson, P. An, J. O'Connell, J. 'an Luan, A. Elliott, S. A. McCarroll, F. Payne, R. M. Roccasecca, F. Pattou, P. Sethupathy, K. Ardlie, Y. Ariyurek, B. Balkau, P. Barter, J. P. Beilby, Y. Ben-Shlomo, R. Benediktsson, A. J. Bennett, S. Bergmann, M. Bochud, E. Boerwinkle, A. Bonnefond, L. L. Bonnycastle, K. Borch-Johnsen, Y. Böttcher, E. Brunner, S. J. Bumpstead, G. Charpentier, Y.-D. I. Chen, P. Chines, R. Clarke, L. J. M. Coin, M. N. Cooper, M. Cornelis, G. Crawford, L. Crisponi, I. N. M. Day, E. J. C. de Geus, J. Delplanque, C. Dina, M. R. Erdos, A. C. Fedson, A. Fischer-Rosinsky, N. G. Forouhi, C. S. Fox, R. Frants, M. G. Franzosi, P. Galan, M. O. Goodarzi, J. Graessler, C. J. Groves, S. Grundy, R. Gwilliam, U. Gyllensten, S. Hadjadj, G. Hallmans, N. Hammond, X. Han, A.-L. Hartikainen, N. Hassanali, C. Hayward, S. C. Heath, S. Hercberg, C. Herder, A. A. Hicks, D. R. Hillman, A. D. Hingorani, A. Hofman, J. Hui, J. Hung, B. Isomaa, P. R. V. Johnson, T. Jørgensen, A. Jula, M. Kaakinen, J. Kaprio, Y. A. Kesaniemi, M. Kivimaki, B. Knight, S. Koskinen, P. Kovacs, K. O. Kyvik, G. M. Lathrop, D. A. Lawlor, O. Le Bacquer, C. Lecoeur, Y. Li, V. Lyssenko, R. Mahley, M. Mangino, A. K. Manning, M. T. Martínez-Larrad, J. B. McAteer, L. J. McCulloch, R. McPherson, C. Meisinger, D. Melzer, D. Meyre, B. D. Mitchell, M. A. Morken, S. Mukherjee, S. Naitza, N. Narisu, M. J. Neville, B. A. Oostra, M. Orrù, R. Pakyz, C. N. A. Palmer, G. Paolisso, C. Pattaro, D. Pearson, J. F. Peden, N. L. Pedersen, M. Perola, A. F. H. Pfeiffer, I. Pichler, O. Polasek, D. Posthuma, S. C. Potter, A. Pouta, M. A. Province, B. M. Psaty, W. Rathmann, N. W. Rayner, K. Rice, S. Ripatti, F. Rivadeneira, M. Roden, O. Rolandsson, A. Sandbaek, M. Sandhu, S. Sanna, A. A. Sayer, P. Scheet, L. J. Scott, U. Seedorf, S. J. Sharp, B. Shields, G. Sigurethsson, E. J. G. Sijbrands, A. Silveira, L. Simpson, A. Singleton, N. L. Smith, U. Sovio, A. Swift, H. Syddall, A.-C. Syvänen, T. Tanaka, B. Thorand, J. Tichet, A. Tönjes, T.

Tuomi, A. G. Uitterlinden, K. W. van Dijk, M. van Hoek, D. Varma, S. Visvikis-Siest, V. Vitart, N. Vogelzangs, G. Waeber, P. J. Wagner, A. Walley, G. B. Walters, K. L. Ward, H. Watkins, M. N. Weedon, S. H. Wild, G. Willemsen, J. C. M. Witteman, J. W. G. Yarnell, E. Zeggini, D. Zelenika, B. Zethelius, G. Zhai, J. H. Zhao, M. C. Zillikens, DIAGRAM Consortium, GIANT Consortium, Global BPgen Consortium, I. B. Borecki, R. J. F. Loos, P. Meneton, P. K. E. Magnusson, D. M. Nathan, G. H. Williams, A. T. Hattersley, K. Silander, V. Salomaa, G. D. Smith, S. R. Bornstein, P. Schwarz, J. Spranger, F. Karpe, A. R. Shuldiner, C. Cooper, G. V. Dedoussis, M. Serrano-Ríos, A. D. Morris, L. Lind, L. J. Palmer, F. B. Hu, P. W. Franks, S. Ebrahim, M. Marmot, W. H. L. Kao, J. S. Pankow, M. J. Sampson, J. Kuusisto, M. Laakso, T. Hansen, O. Pedersen, P. P. Pramstaller, H. E. Wichmann, T. Illig, I. Rudan, A. F. Wright, M. Stumvoll, H. Campbell, J. F. Wilson, Anders Hamsten on behalf of Procardis Consortium, MAGIC investigators, R. N. Bergman, T. A. Buchanan, F. S. Collins, K. L. Mohlke, J. Tuomilehto, T. T. Valle, D. Altshuler, J. I. Rotter, D. S. Siscovick, B. W. J. H. Penninx, D. I. Boomsma, P. Deloukas, T. D. Spector, T. M. Frayling, L. Ferrucci, A. Kong, U. Thorsteinsdottir, K. Stefansson, C. M. van Duijn, Y. S. Aulchenko, A. Cao, A. Scuteri, D. Schlessinger, M. Uda, A. Ruokonen, M.-R. Jarvelin, D. M. Waterworth, P. Vollenweider, L. Peltonen, V. Mooser, G. R. Abecasis, N. J. Wareham, R. Sladek, P. Froguel, R. M. Watanabe, J. B. Meigs, L. Groop, M. Boehnke, M. I. McCarthy, J. C. Florez, I. Barroso, New genetic loci implicated in fasting glucose homeostasis and their impact on type 2 diabetes risk. *Nat. Genet.* **42**, 105–116 (2010).

12. S. L. Pulit, C. Stoneman, A. P. Morris, A. R. Wood, C. A. Glastonbury, J. Tyrrell, L. Yengo, T. Ferreira, E. Marouli, Y. Ji, J. Yang, S. Jones, R. Beaumont, D. C. Croteau-Chonka, T. W. Winkler, GIANT Consortium, A. T. Hattersley, R. J. F. Loos, J. N. Hirschhorn, P. M. Visscher, T. M. Frayling, H. Yaghootkar, C. M. Lindgren, Meta-analysis of genome-wide association studies for body fat distribution in 694 649 individuals of European ancestry. *Hum. Mol. Genet.* **28**, 166–174 (2019).

13. V. Lagou, R. Mägi, J.-J. Hottenga, H. Grallert, J. R. B. Perry, N. Bouatia-Naji, L. Marullo, D. Rybin, R. Jansen, J. L. Min, A. S. Dimas, A. Ulrich, L. Zudina, J. R. Gådin, L. Jiang, A. Faggian, A. Bonnefond, J. Fadista, M. G. Stathopoulou, A. Isaacs, S. M. Willems, P. Navarro, T. Tanaka, A. U. Jackson, M. E. Montasser, J. R. O'Connell, L. F. Bielak, R. J. Webster, R. Saxena, J. M. Stafford, B. S. Pourcain, N. J. Timpson, P. Salo, S.-Y. Shin, N. Amin, A. V. Smith, G. Li, N. Verweij, A. Goel, I. Ford, P. C. D. Johnson, T. Johnson, K. Kapur, G. Thorleifsson, R. J. Strawbridge, L. J. Rasmussen-Torvik, T. Esko, E. Mihailov, T. Fall, R. M. Fraser, A. Mahajan, S. Kanoni, V. Giedraitis, M. E. Kleber, G. Silbernagel, J. Meyer, M. Müller-Nurasyid, A. Ganna, A.-P. Sarin, L. Yengo, D. Shungin, J. 'an Luan, M. Horikoshi, P. An, S. Sanna, Y. Boettcher, N. W. Rayner, I. M. Nolte, T. Zemunik, E. van Iperen, P. Kovacs, N. D. Hastie, S. H. Wild, S. McLachlan, S. Campbell, O. Polasek, O. Carlson, J. Egan, W. Kiess, G. Willemsen, J. Kuusisto, M. Laakso, M. Dimitriou, A. A. Hicks, R. Rauramaa, S. Bandinelli, B. Thorand, Y. Liu, I. Miljkovic, L. Lind, A. Doney, M. Perola, A. Hingorani, M. Kivimaki, M. Kumari, A. J. Bennett, C. J. Groves, C. Herder, H. A. Koistinen, L. Kinnunen, U. de Faire, S. J. L. Bakker, M. Uusitupa, C. N. A. Palmer, J. W. Jukema, N. Sattar, A. Pouta, H. Snieder, E.

Boerwinkle, J. S. Pankow, P. K. Magnusson, U. Krus, C. Scapoli, E. J. C. N. de Geus, M. Blüher, B. H. R. Wolffenbuttel, M. A. Province, G. R. Abecasis, J. B. Meigs, G. K. Hovingh, J. Lindström, J. F. Wilson, A. F. Wright, G. V. Dedoussis, S. R. Bornstein, P. E. H. Schwarz, A. Tönjes, B. R. Winkelmann, B. O. Boehm, W. März, A. Metspalu, J. F. Price, P. Deloukas, A. Körner, T. A. Lakka, S. M. Keinanen-Kiukaanniemi, T. E. Saaristo, R. N. Bergman, J. Tuomilehto, N. J. Wareham, C. Langenberg, S. Männistö, P. W. Franks, C. Hayward, V. Vitart, J. Kaprio, S. Visvikis-Siest, B. Balkau, D. Altshuler, I. Rudan, M. Stumvoll, H. Campbell, C. M. van Duijn, C. Gieger, T. Illig, L. Ferrucci, N. L. Pedersen, P. P. Pramstaller, M. Boehnke, T. M. Frayling, A. R. Shuldiner, P. A. Peyser, S. L. R. Kardia, L. J. Palmer, B. W. Penninx, P. Meneton, T. B. Harris, G. Navis, P. van der Harst, G. D. Smith, N. G. Forouhi, R. J. F. Loos, V. Salomaa, N. Soranzo, D. I. Boomsma, L. Groop, T. Tuomi, A. Hofman, P. B. Munroe, V. Gudnason, D. S. Siscovick, H. Watkins, C. Lecoeur, P. Vollenweider, A. Franco-Cereceda, P. Eriksson, M.-R. Jarvelin, K. Stefansson, A. Hamsten, G. Nicholson, F. Karpe, E. T. Dermitzakis, C. M. Lindgren, M. I. McCarthy, P. Froguel, M. A. Kaakinen, V. Lyssenko, R. M. Watanabe, E. Ingelsson, J. C. Florez, J. Dupuis, I. Barroso, A. P. Morris, I. Prokopenko, Meta-Analyses of Glucose and Insulin-related traits Consortium (MAGIC), Sex-dimorphic genetic effects and novel loci for fasting glucose and insulin variability. *Nat. Commun.* **12**, 24 (2021).

14. T. McLaughlin, G. Reaven, F. Abbasi, C. Lamendola, M. Saad, D. Waters, J. Simon, R. M. Krauss, Is there a simple way to identify insulin-resistant individuals at increased risk of cardiovascular disease? *Am. J. Cardiol.* **96**, 399–404 (2005).

15. D. D. Sears, G. Hsiao, A. Hsiao, J. G. Yu, C. H. Courtney, J. M. Ofrecio, J. Chapman, S. Subramaniam, Mechanisms of human insulin resistance and thiazolidinedione-mediated insulin sensitization. *Proc. Natl. Acad. Sci. U. S. A.* **106**, 18745–18750 (2009).

16. C. Willer, Analysis plan for primary cohort GWAS for blood lipid levels for the Global Lipids Genetics Consortium. *Research Square* (2021), , doi:10.21203/rs.3.pex-1687/v1.

17. J. Mbatchou, L. Barnard, J. Backman, A. Marcketta, J. A. Kosmicki, A. Ziyatdinov, C. Benner, C. O'Dushlaine, M. Barber, B. Boutkov, L. Habegger, M. Ferreira, A. Baras, J. Reid, G. Abecasis, E. Maxwell, J. Marchini, Computationally efficient whole-genome regression for quantitative and binary traits. *Nat. Genet.* **53**, 1097–1103 (2021).

18. K. Watanabe, E. Taskesen, A. van Bochoven, D. Posthuma, Functional mapping and annotation of genetic associations with FUMA. *Nat. Commun.* **8**, 1826 (2017).

19. J. Rung, S. Cauchi, A. Albrechtsen, L. Shen, G. Rocheleau, C. Cavalcanti-Proença, F. Bacot, B. Balkau, A. Belisle, K. Borch-Johnsen, G. Charpentier, C. Dina, E. Durand, P. Elliott, S. Hadjadj, M.-R. Järvelin, J. Laitinen, T. Lauritzen, M. Marre, A. Mazur, D. Meyre, A. Montpetit, C. Pisinger, B. Posner, P. Poulsen, A. Pouta, M.

Prentki, R. Ribel-Madsen, A. Ruokonen, A. Sandbaek, D. Serre, J. Tichet, M. Vaxillaire, J. F. P. Wojtaszewski, A. Vaag, T. Hansen, C. Polychronakos, O. Pedersen, P. Froguel, R. Sladek, Genetic variant near IRS1 is associated with type 2 diabetes, insulin resistance and hyperinsulinemia. *Nat. Genet.* **41**, 1110–1115 (2009).

20. A. R. Majithia, J. Flannick, P. Shahinian, M. Guo, M.-A. Bray, P. Fontanillas, S. B. Gabriel, GoT2D Consortium, NHGRI JHS/FHS Allelic Spectrum Project, SIGMA T2D Consortium, T2D-GENES Consortium, E. D. Rosen, D. Altshuler, Rare variants in PPARG with decreased activity in adipocyte differentiation are associated with increased risk of type 2 diabetes. *Proc. Natl. Acad. Sci. U. S. A.* **111**, 13127–13132 (2014).

21. L. A. Lotta, P. Gulati, F. R. Day, F. Payne, H. Ongen, M. van de Bunt, K. J. Gaulton, J. D. Eicher, S. J. Sharp, J. 'an Luan, E. De Lucia Rolfe, I. D. Stewart, E. Wheeler, S. M. Willems, C. Adams, H. Yaghootkar, EPIC-InterAct Consortium, Cambridge FPLD1 Consortium, N. G. Forouhi, K.-T. Khaw, A. D. Johnson, R. K. Semple, T. Frayling, J. R. B. Perry, E. Dermitzakis, M. I. McCarthy, I. Barroso, N. J. Wareham, D. B. Savage, C. Langenberg, S. O'Rahilly, R. A. Scott, Integrative genomic analysis implicates limited peripheral adipose storage capacity in the pathogenesis of human insulin resistance. *Nat. Genet.* **49**, 17–26 (2017).

22. V. Gusarova, C. O'Dushlaine, T. M. Teslovich, P. N. Benotti, T. Mirshahi, O. Gottesman, C. V. Van Hout, M. F. Murray, A. Mahajan, J. B. Nielsen, L. Fritsche, A. B. Wulff, D. F. Gudbjartsson, M. Sjögren, C. A. Emdin, R. A. Scott, W.-J. Lee, A. Small, L. C. Kwee, O. P. Dwivedi, R. B. Prasad, S. Bruse, A. E. Lopez, J. Penn, A. Marcketta, J. B. Leader, C. D. Still, H. L. Kirchner, U. L. Mirshahi, A. H. Wardeh, C. M. Hartle, L. Habegger, S. N. Fetterolf, T. Tusie-Luna, A. P. Morris, H. Holm, V. Steinthorsdottir, P. Sulem, U. Thorsteinsdottir, J. I. Rotter, L.-M. Chuang, S. Damrauer, D. Birtwell, C. M. Brummett, A. V. Khera, P. Natarajan, M. Orho-Melander, J. Flannick, L. A. Lotta, C. J. Willer, O. L. Holmen, M. D. Ritchie, D. H. Ledbetter, A. J. Murphy, I. B. Borecki, J. G. Reid, J. D. Overton, O. Hansson, L. Groop, S. H. Shah, W. E. Kraus, D. J. Rader, Y.-D. I. Chen, K. Hveem, N. J. Wareham, S. Kathiresan, O. Melander, K. Stefansson, B. G. Nordestgaard, A. Tybjærg-Hansen, G. R. Abecasis, D. Altshuler, J. C. Florez, M. Boehnke, M. I. McCarthy, G. D. Yancopoulos, D. J. Carey, A. R. Shuldiner, A. Baras, F. E. Dewey, J. Gromada, Genetic inactivation of ANGPTL4 improves glucose homeostasis and is associated with reduced risk of diabetes. *Nat. Commun.* **9**, 2252 (2018).

23. M. O. Goodarzi, X. Guo, K. D. Taylor, M. J. Quiñones, M. F. Saad, H. Yang, W. A. Hsueh, J. I. Rotter, Lipoprotein lipase is a gene for insulin resistance in Mexican Americans. *Diabetes*. **53**, 214–220 (2004).

24. S. E. Graham, S. L. Clarke, K.-H. H. Wu, S. Kanoni, G. J. M. Zajac, S. Ramdas, I. Surakka, I. Ntalla, S. Vedantam, T. W. Winkler, A. E. Locke, E. Marouli, M. Y. Hwang, S. Han, A. Narita, A. Choudhury, A. R. Bentley, K. Ekoru, A. Verma, B. Trivedi, H. C. Martin, K. A. Hunt, Q. Hui, D. Klarin, X. Zhu, G. Thorleifsson, A.

Helgadottir, D. F. Gudbjartsson, H. Holm, I. Olafsson, M. Akiyama, S. Sakaue, C. Terao, M. Kanai, W. Zhou, B. M. Brumpton, H. Rasheed, S. E. Ruotsalainen, A. S. Havulinna, Y. Veturi, Q. Feng, E. A. Rosenthal, T. Lingren, J. A. Pacheco, S. A. Pendergrass, J. Haessler, F. Giulianini, Y. Bradford, J. E. Miller, A. Campbell, K. Lin, I. Y. Millwood, G. Hindy, A. Rasheed, J. D. Faul, W. Zhao, D. R. Weir, C. Turman, H. Huang, M. Graff, A. Mahajan, M. R. Brown, W. Zhang, K. Yu, E. M. Schmidt, A. Pandit, S. Gustafsson, X. Yin, J. 'an Luan, J.-H. Zhao, F. Matsuda, H.-M. Jang, K. Yoon, C. Medina-Gomez, A. Pitsillides, J. J. Hottenga, G. Willemsen, A. R. Wood, Y. Ji, Z. Gao, S. Haworth, R. E. Mitchell, J. F. Chai, M. Aadahl, J. Yao, A. Manichaikul, H. R. Warren, J. Ramirez, J. Bork-Jensen, L. L. Kårhus, A. Goel, M. Sabater-Lleal, R. Noordam, C. Sidore, E. Fiorillo, A. F. McDaid, P. Marques-Vidal, M. Wielscher, S. Trompet, N. Sattar, L. T. Møllehave, B. H. Thuesen, M. Munz, L. Zeng, J. Huang, B. Yang, A. Poveda, A. Kurbasic, C. Lamina, L. Forer, M. Scholz, T. E. Galesloot, J. P. Bradfield, E. W. Daw, J. M. Zmuda, J. S. Mitchell, C. Fuchsberger, H. Christensen, J. A. Brody, M. F. Feitosa, M. K. Wojczynski, M. Preuss, M. Mangino, P. Christofidou, N. Verweij, J. W. Benjamins, J. Engmann, R. L. Kember, R. C. Slieker, K. S. Lo, N. R. Zilhao, P. Le, M. E. Kleber, G. E. Delgado, S. Huo, D. D. Ikeda, H. Iha, J. Yang, J. Liu, H. L. Leonard, J. Marten, B. Schmidt, M. Arendt, L. J. Smyth, M. Cañadas-Garre, C. Wang, M. Nakatochi, A. Wong, N. Hutri-Kähönen, X. Sim, R. Xia, A. Huerta-Chagoya, J. C. Fernandez-Lopez, V. Lyssenko, M. Ahmed, A. U. Jackson, N. A. Yousri, M. R. Irvin, C. Oldmeadow, H.-N. Kim, S. Ryu, P. R. H. J. Timmers, L. Arbeeva, R. Dorajoo, L. A. Lange, X. Chai, G. Prasad, L. Lorés-Motta, M. Pauper, J. Long, X. Li, E. Theusch, F. Takeuchi, C. N. Spracklen, A. Loukola, S. Bollepalli, S. C. Warner, Y. X. Wang, W. B. Wei, T. Nutile, D. Ruggiero, Y. J. Sung, Y.-J. Hung, S. Chen, F. Liu, J. Yang, K. A. Kentistou, M. Gorski, M. Brumat, K. Meidtner, L. F. Bielak, J. A. Smith, P. Hebbar, A.-E. Farmaki, E. Hofer, M. Lin, C. Xue, J. Zhang, M. P. Concas, S. Vaccargiu, P. J. van der Most, N. Pitkänen, B. E. Cade, J. Lee, S. W. van der Laan, K. N. Chitrala, S. Weiss, M. E. Zimmermann, J. Y. Lee, H. S. Choi, M. Nethander, S. Freitag-Wolf, L. Southam, N. W. Rayner, C. A. Wang, S.-Y. Lin, J.-S. Wang, C. Couture, L.-P. Lyytikäinen, K. Nikus, G. Cuellar-Partida, H. Vestergaard, B. Hildalgo, O. Giannakopoulou, Q. Cai, M. O. Obura, J. van Setten, X. Li, K. Schwander, N. Terzikhan, J. H. Shin, R. D. Jackson, A. P. Reiner, L. W. Martin, Z. Chen, L. Li, H. M. Highland, K. L. Young, T. Kawaguchi, J. Thiery, J. C. Bis, G. N. Nadkarni, L. J. Launer, H. Li, M. A. Nalls, O. T. Raitakari, S. Ichihara, S. H. Wild, C. P. Nelson, H. Campbell, S. Jäger, T. Nabika, F. Al-Mulla, H. Niinikoski, P. S. Braund, I. Kolcic, P. Kovacs, T. Giardoglou, T. Katsuya, K. F. Bhatti, D. de Kleijn, G. J. de Borst, E. K. Kim, H. H. H. Adams, M. A. Ikram, X. Zhu, F. W. Asselbergs, A. O. Kraaijeveld, J. W. J. Beulens, X.-O. Shu, L. S. Rallidis, O. Pedersen, T. Hansen, P. Mitchell, A. W. Hewitt, M. Kähönen, L. Pérusse, C. Bouchard, A. Tönjes, Y.-D. I. Chen, C. E. Pennell, T. A. Mori, W. Lieb, A. Franke, C. Ohlsson, D. Mellström, Y. S. Cho, H. Lee, J.-M. Yuan, W.-P. Koh, S. Y. Rhee, J.-T. Woo, I. M. Heid, K. J. Stark, H. Völzke, G. Homuth, M. K. Evans, A. B. Zonderman, O. Polasek, G. Pasterkamp, I. E. Hoefer, S. Redline, K. Pahkala, A. J. Oldehinkel, H. Snieder, G. Biino, R. Schmidt, H. Schmidt, Y. E. Chen, S. Bandinelli, G. Dedoussis, T. A. Thanaraj, S. L. R. Kardia, N. Kato, M. B. Schulze, G. Girotto, B. Jung, C. A.

Böger, P. K. Joshi, D. A. Bennett, P. L. De Jager, X. Lu, V. Mamakou, M. Brown, M. J. Caulfield, P. B. Munroe, X. Guo, M. Ciullo, J. B. Jonas, N. J. Samani, J. Kaprio, P. Pajukanta, L. S. Adair, S. A. Bechayda, H. J. de Silva, A. R. Wickremasinghe, R. M. Krauss, J.-Y. Wu, W. Zheng, A. I. den Hollander, D. Bharadwaj, A. Correa, J. G. Wilson, L. Lind, C.-K. Heng, A. E. Nelson, Y. M. Golightly, J. F. Wilson, B. Penninx, H.-L. Kim, J. Attia, R. J. Scott, D. C. Rao, D. K. Arnett, S. C. Hunt, M. Walker, H. A. Koistinen, G. R. Chandak, C. S. Yajnik, J. M. Mercader, T. Tusié-Luna, C. A. Aguilar-Salinas, C. G. Villalpando, L. Orozco, M. Fornage, E. S. Tai, R. M. van Dam, T. Lehtimäki, N. Chaturvedi, M. Yokota, J. Liu, D. F. Reilly, A. J. McKnight, F. Kee, K.-H. Jöckel, M. I. McCarthy, C. N. A. Palmer, V. Vitart, C. Hayward, E. Simonsick, C. M. van Duijn, F. Lu, J. Qu, H. Hishigaki, X. Lin, W. März, E. J. Parra, M. Cruz, V. Gudnason, J.-C. Tardif, G. Lettre, L. M. 't Hart, P. J. M. Elders, S. M. Damrauer, M. Kumari, M. Kivimaki, P. van der Harst, T. D. Spector, R. J. F. Loos, M. A. Province, B. M. Psaty, I. Brandslund, P. P. Pramstaller, K. Christensen, S. Ripatti, E. Widén, H. Hakonarson, S. F. A. Grant, L. A. L. M. Kiemeney, J. de Graaf, M. Loeffler, F. Kronenberg, D. Gu, J. Erdmann, H. Schunkert, P. W. Franks, A. Linneberg, J. W. Jukema, A. V. Khera, M. Männikkö, M.-R. Jarvelin, Z. Kutalik, F. Cucca, D. O. Mook-Kanamori, K. W. van Dijk, H. Watkins, D. P. Strachan, N. Grarup, P. Sever, N. Poulter, J. I. Rotter, T. M. Dantoft, F. Karpe, M. J. Neville, N. J. Timpson, C.-Y. Cheng, T.-Y. Wong, C. C. Khor, C. Sabanayagam, A. Peters, C. Gieger, A. T. Hattersley, N. L. Pedersen, P. K. E. Magnusson, D. I. Boomsma, E. J. C. de Geus, L. A. Cupples, J. B. J. van Meurs, M. Ghanbari, P. Gordon-Larsen, W. Huang, Y. J. Kim, Y. Tabara, N. J. Wareham, C. Langenberg, E. Zeggini, J. Kuusisto, M. Laakso, E. Ingelsson, G. Abecasis, J. C. Chambers, J. S. Kooner, P. S. de Vries, A. C. Morrison, K. E. North, M. Daviglus, P. Kraft, N. G. Martin, J. B. Whitfield, S. Abbas, D. Saleheen, R. G. Walters, M. V. Holmes, C. Black, B. H. Smith, A. E. Justice, A. Baras, J. E. Buring, P. M. Ridker, D. I. Chasman, C. Kooperberg, W.-Q. Wei, G. P. Jarvik, B. Namjou, M. G. Hayes, M. D. Ritchie, P. Jousilahti, V. Salomaa, K. Hveem, B. O. Åsvold, M. Kubo, Y. Kamatani, Y. Okada, Y. Murakami, U. Thorsteinsdottir, K. Stefansson, Y.-L. Ho, J. A. Lynch, D. J. Rader, P. S. Tsao, K.-M. Chang, K. Cho, C. J. O'Donnell, J. M. Gaziano, P. Wilson, C. N. Rotimi, S. Hazelhurst, M. Ramsay, R. C. Trembath, D. A. van Heel, G. Tamiya, M. Yamamoto, B.-J. Kim, K. L. Mohlke, T. M. Frayling, J. N. Hirschhorn, S. Kathiresan, VA Million Veteran Program, Global Lipids Genetics Consortium*, M. Boehnke, P. Natarajan, G. M. Peloso, C. D. Brown, A. P. Morris, T. L. Assimes, P. Deloukas, Y. V. Sun, C. J. Willer, The power of genetic diversity in genome-wide association studies of lipids. *Nature*. **600**, 675–679 (2021).

25. E. M. Schmidt, J. Zhang, W. Zhou, J. Chen, K. L. Mohlke, Y. E. Chen, C. J. Willer, GREGOR: evaluating global enrichment of trait-associated variants in epigenomic features using a systematic, data-driven approach. *Bioinformatics*. **31**, 2601–2606 (2015).

26. R. A. Defronzo, Banting Lecture. From the triumvirate to the ominous octet: a new paradigm for the treatment of type 2 diabetes mellitus. *Diabetes*. **58**, 773–795 (2009).

27. B. Tramunt, S. Smati, N. Grandgeorge, F. Lenfant, J.-F. Arnal, A. Montagner, P. Gourdy, Sex differences in metabolic regulation and diabetes susceptibility. *Diabetologia*. **63**, 453–461 (2020).

28. K. S. Small, M. Todorčević, M. Civelek, J. S. El-Sayed Moustafa, X. Wang, M. M. Simon, J. Fernandez-Tajes, A. Mahajan, M. Horikoshi, A. Hugill, C. A. Glastonbury, L. Quaye, M. J. Neville, S. Sethi, M. Yon, C. Pan, N. Che, A. Viñuela, P.-C. Tsai, A. Nag, A. Buil, G. Thorleifsson, A. Raghavan, Q. Ding, A. P. Morris, J. T. Bell, U. Thorsteinsdottir, K. Stefansson, M. Laakso, I. Dahlman, P. Arner, A. L. Gloyn, K. Musunuru, A. J. Lusis, R. D. Cox, F. Karpe, M. I. McCarthy, Regulatory variants at KLF14 influence type 2 diabetes risk via a female-specific effect on adipocyte size and body composition. *Nat. Genet.* **50**, 572–580 (2018).

29. N. Y. Loh, J. E. N. Minchin, K. E. Pinnick, M. Verma, M. Todorčević, N. Denton, J. E.-S. Moustafa, J. P. Kemp, C. L. Gregson, D. M. Evans, M. J. Neville, K. S. Small, M. I. McCarthy, A. Mahajan, J. F. Rawls, F. Karpe, C. Christodoulides, RSPO3 impacts body fat distribution and regulates adipose cell biology in vitro. *Nat. Commun.* **11**, 2797 (2020).

30. N. T. Boutin, S. B. Schecter, E. F. Perez, N. S. Tchamitchian, X. R. Cerretani, V. S. Gainer, M. S. Lebo, L. M. Mahanta, E. W. Karlson, J. W. Smoller, The Evolution of a Large Biobank at Mass General Brigham. *J Pers Med*. **12** (2022), doi:10.3390/jpm12081323.

31. B. Bulik-Sullivan, H. K. Finucane, V. Anttila, A. Gusev, F. R. Day, P.-R. Loh, ReproGen Consortium, Psychiatric Genomics Consortium, Genetic Consortium for Anorexia Nervosa of the Wellcome Trust Case Control Consortium 3, L. Duncan, J. R. B. Perry, N. Patterson, E. B. Robinson, M. J. Daly, A. L. Price, B. M. Neale, An atlas of genetic correlations across human diseases and traits. *Nat. Genet.* **47**, 1236–1241 (2015).

32. Q. M. Anstee, R. Darlay, S. Cockell, M. Meroni, O. Govaere, D. Tiniakos, A. D. Burt, P. Bedossa, J. Palmer, Y.-L. Liu, G. P. Aithal, M. Allison, H. Yki-Järvinen, M. Vacca, J.-F. Dufour, P. Invernizzi, D. Prati, M. Ekstedt, S. Kechagias, S. Francque, S. Petta, E. Bugianesi, K. Clement, V. Ratziu, J. M. Schattenberg, L. Valenti, C. P. Day, H. J. Cordell, A. K. Daly, EPoS Consortium Investigators, Genome-wide association study of non-alcoholic fatty liver and steatohepatitis in a histologically characterised cohort☆. *J. Hepatol.* **73**, 505–515 (2020).

33. A. Mahajan, C. N. Spracklen, W. Zhang, M. C. Y. Ng, L. E. Petty, H. Kitajima, G. Z. Yu, S. Rüeger, L. Speidel, Y. J. Kim, M. Horikoshi, J. M. Mercader, D. Taliun, S. Moon, S.-H. Kwak, N. R. Robertson, N. W. Rayner, M. Loh, B.-J. Kim, J. Chiou, I. Miguel-Escalada, P. Della Briotta Parolo, K. Lin, F. Bragg, M. H. Preuss, F. Takeuchi, J. Nano, X. Guo, A. Lamri, M. Nakatochi, R. A. Scott, J.-J. Lee, A. Huerta-Chagoya, M. Graff, J.-F. Chai, E. J. Parra, J. Yao, L. F. Bielak, Y. Tabara, Y. Hai, V. Steinthorsdottir, J. P. Cook, M. Kals, N. Grarup, E. M. Schmidt, I. Pan, T. Sofer, M. Wuttke, C. Sarnowski, C. Gieger, D. Nousome, S. Trompet, J. Long, M.

Sun, L. Tong, W.-M. Chen, M. Ahmad, R. Noordam, V. J. Y. Lim, C. H. T. Tam, Y. Y. Joo, C.-H. Chen, L. M. Raffield, C. Lecoeur, B. P. Prins, A. Nicolas, L. R. Yanek, G. Chen, R. A. Jensen, S. Tajuddin, E. K. Kabagambe, P. An, A. H. Xiang, H. S. Choi, B. E. Cade, J. Tan, J. Flanagan, F. Abaitua, L. S. Adair, A. Adeyemo, C. A. Aguilar-Salinas, M. Akiyama, S. S. Anand, A. Bertoni, Z. Bian, J. Bork-Jensen, I. Brandslund, J. A. Brody, C. M. Brummett, T. A. Buchanan, M. Canouil, J. C. N. Chan, L.-C. Chang, M.-L. Chee, J. Chen, S.-H. Chen, Y.-T. Chen, Z. Chen, L.-M. Chuang, M. Cushman, S. K. Das, H. J. de Silva, G. Dedoussis, L. Dimitrov, A. P. Doumatey, S. Du, Q. Duan, K.-U. Eckardt, L. S. Emery, D. S. Evans, M. K. Evans, K. Fischer, J. S. Floyd, I. Ford, M. Fornage, O. H. Franco, T. M. Frayling, B. I. Freedman, C. Fuchsberger, P. Genter, H. C. Gerstein, V. Giedraitis, C. González-Villalpando, M. E. González-Villalpando, M. O. Goodarzi, P. Gordon-Larsen, D. Gorkin, M. Gross, Y. Guo, S. Hackinger, S. Han, A. T. Hattersley, C. Herder, A.-G. Howard, W. Hsueh, M. Huang, W. Huang, Y.-J. Hung, M. Y. Hwang, C.-M. Hwu, S. Ichihara, M. A. Ikram, M. Ingelsson, M. T. Islam, M. Isono, H.-M. Jang, F. Jasmine, G. Jiang, J. B. Jonas, M. E. Jørgensen, T. Jørgensen, Y. Kamatani, F. R. Kandeel, A. Kasturiratne, T. Katsuya, V. Kaur, T. Kawaguchi, J. M. Keaton, A. N. Kho, C.-C. Khor, M. G. Kibriya, D.-H. Kim, K. Kohara, J. Kriebel, F. Kronenberg, J. Kuusisto, K. Läll, L. A. Lange, M.-S. Lee, N. R. Lee, A. Leong, L. Li, Y. Li, R. Li-Gao, S. Ligthart, C. M. Lindgren, A. Linneberg, C.-T. Liu, J. Liu, A. E. Locke, T. Louie, J. ’an Luan, A. O. Luk, X. Luo, J. Lv, V. Lyssenko, V. Mamakou, K. R. Mani, T. Meitinger, A. Metspalu, A. D. Morris, G. N. Nadkarni, J. L. Nadler, M. A. Nalls, U. Nayak, S. S. Nongmaithem, I. Ntalla, Y. Okada, L. Orozco, S. R. Patel, M. A. Pereira, A. Peters, F. J. Pirie, B. Porneala, G. Prasad, S. Preissl, L. J. Rasmussen-Torvik, A. P. Reiner, M. Roden, R. Rohde, K. Roll, C. Sabanayagam, M. Sander, K. Sandow, N. Sattar, S. Schönherr, C. Schurmann, M. Shahriar, J. Shi, D. M. Shin, D. Shriner, J. A. Smith, W. Y. So, A. Stančáková, A. M. Stilp, K. Strauch, K. Suzuki, A. Takahashi, K. D. Taylor, B. Thorand, G. Thorleifsson, U. Thorsteinsdottir, B. Tomlinson, J. M. Torres, F.-J. Tsai, J. Tuomilehto, T. Tusie-Luna, M. S. Udler, A. Valladares-Salgado, R. M. van Dam, J. B. van Klinken, R. Varma, M. Vujkovic, N. Wacher-Rodarte, E. Wheeler, E. A. Whitsel, A. R. Wickremasinghe, K. W. van Dijk, D. R. Witte, C. S. Yajnik, K. Yamamoto, T. Yamauchi, L. Yengo, K. Yoon, C. Yu, J.-M. Yuan, S. Yusuf, L. Zhang, W. Zheng, FinnGen, eMERGE Consortium, L. J. Raffel, M. Igase, E. Ipp, S. Redline, Y. S. Cho, L. Lind, M. A. Province, C. L. Hanis, P. A. Peyser, E. Ingelsson, A. B. Zonderman, B. M. Psaty, Y.-X. Wang, C. N. Rotimi, D. M. Becker, F. Matsuda, Y. Liu, E. Zeggini, M. Yokota, S. S. Rich, C. Kooperberg, J. S. Pankow, J. C. Engert, Y.-D. I. Chen, P. Froguel, J. G. Wilson, W. H. H. Sheu, S. L. R. Kardia, J.-Y. Wu, M. G. Hayes, R. C. W. Ma, T.-Y. Wong, L. Groop, D. O. Mook-Kanamori, G. R. Chandak, F. S. Collins, D. Bharadwaj, G. Paré, M. M. Sale, H. Ahsan, A. A. Motala, X.-O. Shu, K.-S. Park, J. W. Jukema, M. Cruz, R. McKean-Cowdin, H. Grallert, C.-Y. Cheng, E. P. Bottinger, A. Dehghan, E.-S. Tai, J. Dupuis, N. Kato, M. Laakso, A. Köttgen, W.-P. Koh, C. N. A. Palmer, S. Liu, G. Abecasis, J. S. Kooner, R. J. F. Loos, K. E. North, C. A. Haiman, J. C. Florez, D. Saleheen, T. Hansen, O. Pedersen, R. Mägi, C. Langenberg, N. J. Wareham, S. Maeda, T. Kadowaki, J. Lee, I. Y. Millwood, R. G. Walters, K. Stefansson, S. R. Myers, J. Ferrer, K. J. Gaulton, J. B. Meigs, K. L. Mohlke, A. L. Gloyn, D. W.

Bowden, J. E. Below, J. C. Chambers, X. Sim, M. Boehnke, J. I. Rotter, M. I. McCarthy, A. P. Morris, Multi-ancestry genetic study of type 2 diabetes highlights the power of diverse populations for discovery and translation. *Nat. Genet.* **54**, 560–572 (2022).

34. C. Tcheandjieu, X. Zhu, A. T. Hilliard, S. L. Clarke, V. Napolioni, S. Ma, K. M. Lee, H. Fang, F. Chen, Y. Lu, N. L. Tsao, S. Raghavan, S. Koyama, B. R. Gorman, M. Vujkovic, D. Klarin, M. G. Levin, N. Sinnott-Armstrong, G. L. Wojcik, M. E. Plomondon, T. M. Maddox, S. W. Waldo, A. G. Bick, S. Pyarajan, J. Huang, R. Song, Y.-L. Ho, S. Buyske, C. Kooperberg, J. Haessler, R. J. F. Loos, R. Do, M. Verbanck, K. Chaudhary, K. E. North, C. L. Avery, M. Graff, C. A. Haiman, L. Le Marchand, L. R. Wilkens, J. C. Bis, H. Leonard, B. Shen, L. A. Lange, A. Giri, O. Dikilitas, I. J. Kullo, I. B. Stanaway, G. P. Jarvik, A. S. Gordon, S. Hebbring, B. Namjou, K. M. Kaufman, K. Ito, K. Ishigaki, Y. Kamatani, S. S. Verma, M. D. Ritchie, R. L. Kember, A. Baras, L. A. Lotta, S. Kathiresan, E. R. Hauser, D. R. Miller, J. S. Lee, D. Saleheen, P. D. Reaven, K. Cho, J. M. Gaziano, P. Natarajan, J. E. Huffman, B. F. Voight, D. J. Rader, K.-M. Chang, J. A. Lynch, S. M. Damrauer, P. W. F. Wilson, H. Tang, Y. V. Sun, P. S. Tsao, C. J. O'Donnell, T. L. Assimes, Large-scale genome-wide association study of coronary artery disease in genetically diverse populations. *Nat. Med.* **28**, 1679–1692 (2022).

35. D. Panarotto, P. Rémillard, L. Bouffard, P. Maheux, Insulin resistance affects the regulation of lipoprotein lipase in the postprandial period and in an adipose tissue-specific manner. *Eur. J. Clin. Invest.* **32**, 84–92 (2002).

36. V. Glunk, S. Laber, N. Sinnott-Armstrong, D. R. Sobreira, S. M. Strobel, T. M. Batista, P. Kubitz, B. N. Moud, H. Ebert, Y. Huang, B. Brandl, G. Garbo, J. Honecker, D. R. Stirling, N. Abdennur, V. Calabuig-Navarro, T. Skurk, S. Ocvirk, K. Stemmer, B. A. Cimini, A. E. Carpenter, S. N. Dankel, C. M. Lindgren, H. Hauner, M. A. Nobrega, M. Claussnitzer, A non-coding variant linked to metabolic obesity with normal weight affects actin remodelling in subcutaneous adipocytes. *Nat Metab*. **5**, 861–879 (2023).

37. P. J. Barter, H. B. Brewer Jr, M. J. Chapman, C. H. Hennekens, D. J. Rader, A. R. Tall, Cholesteryl ester transfer protein: a novel target for raising HDL and inhibiting atherosclerosis. *Arterioscler. Thromb. Vasc. Biol.* **23**, 160–167 (2003).

38. K. Musunuru, J. P. Pirruccello, R. Do, G. M. Peloso, C. Guiducci, C. Sougnez, K. V. Garimella, S. Fisher, J. Abreu, A. J. Barry, T. Fennell, E. Banks, L. Ambrogio, K. Cibulskis, A. Kernytsky, E. Gonzalez, N. Rudzicz, J. C. Engert, M. A. DePristo, M. J. Daly, J. C. Cohen, H. H. Hobbs, D. Altshuler, G. Schonfeld, S. B. Gabriel, P. Yue, S. Kathiresan, Exome sequencing, ANGPTL3 mutations, and familial combined hypolipidemia. *N. Engl. J. Med.* **363**, 2220–2227 (2010).

39. K. G. Aragam, T. Jiang, A. Goel, S. Kanoni, B. N. Wolford, D. S. Atri, E. M. Weeks, M. Wang, G. Hindy, W. Zhou, C. Grace, C. Roselli, N. A. Marston, F. K. Kamanu, I. Surakka, L. M. Venegas, P. Sherliker, S. Koyama, K. Ishigaki, B. O. Åsvold, M. R.

Brown, B. Brumpton, P. S. de Vries, O. Giannakopoulou, P. Giardoglou, D. F. Gudbjartsson, U. Güldener, S. M. I. Haider, A. Helgadottir, M. Ibrahim, A. Kastrati, T. Kessler, T. Kyriakou, T. Konopka, L. Li, L. Ma, T. Meitinger, S. Mucha, M. Munz, F. Murgia, J. B. Nielsen, M. M. Nöthen, S. Pang, T. Reinberger, G. Schnitzler, D. Smedley, G. Thorleifsson, M. von Scheidt, J. C. Ulirsch, Biobank Japan, EPIC-CVD, D. O. Arnar, N. P. Burtt, M. C. Costanzo, J. Flannick, K. Ito, D.-K. Jang, Y. Kamatani, A. V. Khera, I. Komuro, I. J. Kullo, L. A. Lotta, C. P. Nelson, R. Roberts, G. Thorgeirsson, U. Thorsteinsdottir, T. R. Webb, A. Baras, J. L. M. Björkegren, E. Boerwinkle, G. Dedoussis, H. Holm, K. Hveem, O. Melander, A. C. Morrison, M. Orho-Melander, L. S. Rallidis, A. Ruusalepp, M. S. Sabatine, K. Stefansson, P. Zalloua, P. T. Ellinor, M. Farrall, J. Danesh, C. T. Ruff, H. K. Finucane, J. C. Hopewell, R. Clarke, R. M. Gupta, J. Erdmann, N. J. Samani, H. Schunkert, H. Watkins, C. J. Willer, P. Deloukas, S. Kathiresan, A. S. Butterworth, CARDIoGRAMplusC4D Consortium, Discovery and systematic characterization of risk variants and genes for coronary artery disease in over a million participants. *Nat. Genet.* **54**, 1803–1815 (2022).

40. G. Wang, A. Sarkar, P. Carbonetto, M. Stephens, A simple new approach to variable selection in regression, with application to genetic fine mapping. *J. R. Stat. Soc. Series B Stat. Methodol.* **82**, 1273–1300 (2020).

41. S. Gazal, O. Weissbrod, F. Hormozdiari, K. K. Dey, J. Nasser, K. A. Jagadeesh, D. J. Weiner, H. Shi, C. P. Fulco, L. J. O'Connor, B. Pasaniuc, J. M. Engreitz, A. L. Price, Combining SNP-to-gene linking strategies to identify disease genes and assess disease omnigenicity. *Nat. Genet.* **54**, 827–836 (2022).

42. K. J. Karczewski, M. Solomonson, K. R. Chao, J. K. Goodrich, G. Tiao, W. Lu, B. M. Riley-Gillis, E. A. Tsai, H. I. Kim, X. Zheng, F. Rahimov, S. Esmaeeli, A. J. Grundstad, M. Reppell, J. Waring, H. Jacob, D. Sexton, P. G. Bronson, X. Chen, X. Hu, J. I. Goldstein, D. King, C. Vittal, T. Poterba, D. S. Palmer, C. Churchhouse, D. P. Howrigan, W. Zhou, N. A. Watts, K. Nguyen, H. Nguyen, C. Mason, C. Farnham, C. Tolonen, L. D. Gauthier, N. Gupta, D. G. MacArthur, H. L. Rehm, C. Seed, A. A. Philippakis, M. J. Daly, J. W. Davis, H. Runz, M. R. Miller, B. M. Neale, Systematic single-variant and gene-based association testing of thousands of phenotypes in 394,841 UK Biobank exomes. *Cell Genom*. **2**, 100168 (2022).

43. A. Gusev, A. Ko, H. Shi, G. Bhatia, W. Chung, B. W. J. H. Penninx, R. Jansen, E. J. C. de Geus, D. I. Boomsma, F. A. Wright, P. F. Sullivan, E. Nikkola, M. Alvarez, M. Civelek, A. J. Lusis, T. Lehtimäki, E. Raitoharju, M. Kähönen, I. Seppälä, O. T. Raitakari, J. Kuusisto, M. Laakso, A. L. Price, P. Pajukanta, B. Pasaniuc, Integrative approaches for large-scale transcriptome-wide association studies. *Nat. Genet.* **48**, 245–252 (2016).

44. I. Barroso, M. Gurnell, V. E. Crowley, M. Agostini, J. W. Schwabe, M. A. Soos, G. L. Maslen, T. D. Williams, H. Lewis, A. J. Schafer, V. K. Chatterjee, S. O'Rahilly, Dominant negative mutations in human PPARgamma associated with severe insulin resistance, diabetes mellitus and hypertension. *Nature*. **402**, 880–883

(1999).

45. W. McLaren, L. Gil, S. E. Hunt, H. S. Riat, G. R. S. Ritchie, A. Thormann, P. Flicek, F. Cunningham, The Ensembl Variant Effect Predictor. *Genome Biol.* **17**, 122 (2016).

46. J. E. Burke, E. A. Dennis, Phospholipase A2 structure/function, mechanism, and signaling. *J. Lipid Res.* **50 Suppl**, S237–42 (2009).

47. S. Lee, G. R. Abecasis, M. Boehnke, X. Lin, Rare-variant association analysis: study designs and statistical tests. *Am. J. Hum. Genet.* **95**, 5–23 (2014).

48. K. J. Karczewski, L. C. Francioli, G. Tiao, B. B. Cummings, J. Alföldi, Q. Wang, R. L. Collins, K. M. Laricchia, A. Ganna, D. P. Birnbaum, L. D. Gauthier, H. Brand, M. Solomonson, N. A. Watts, D. Rhodes, M. Singer-Berk, E. M. England, E. G. Seaby, J. A. Kosmicki, R. K. Walters, K. Tashman, Y. Farjoun, E. Banks, T. Poterba, A. Wang, C. Seed, N. Whiffin, J. X. Chong, K. E. Samocha, E. Pierce-Hoffman, Z. Zappala, A. H. O'Donnell-Luria, E. V. Minikel, B. Weisburd, M. Lek, J. S. Ware, C. Vittal, I. M. Armean, L. Bergelson, K. Cibulskis, K. M. Connolly, M. Covarrubias, S. Donnelly, S. Ferriera, S. Gabriel, J. Gentry, N. Gupta, T. Jeandet, D. Kaplan, C. Llanwarne, R. Munshi, S. Novod, N. Petrillo, D. Roazen, V. Ruano-Rubio, A. Saltzman, M. Schleicher, J. Soto, K. Tibbetts, C. Tolonen, G. Wade, M. E. Talkowski, Genome Aggregation Database Consortium, B. M. Neale, M. J. Daly, D. G. MacArthur, The mutational constraint spectrum quantified from variation in 141,456 humans. *Nature*. **581**, 434–443 (2020).

49. P. K. Larsson, H. E. Claesson, B. P. Kennedy, Multiple splice variants of the human calcium-independent phospholipase A2 and their effect on enzyme activity. *J. Biol. Chem.* **273**, 207–214 (1998).

50. O. Fornes, J. A. Castro-Mondragon, A. Khan, R. van der Lee, X. Zhang, P. A. Richmond, B. P. Modi, S. Correard, M. Gheorghe, D. Baranašić, W. Santana-Garcia, G. Tan, J. Chèneby, B. Ballester, F. Parcy, A. Sandelin, B. Lenhard, W. W. Wasserman, A. Mathelier, JASPAR 2020: update of the open-access database of transcription factor binding profiles. *Nucleic Acids Res.* **48** (2020), doi:10.1093/nar/gkz1001.

51. D. S. Halperin, C. Pan, A. J. Lusis, P. Tontonoz, Vestigial-like 3 is an inhibitor of adipocyte differentiation. *J. Lipid Res.* **54**, 473–481 (2013).

52. GTEx Consortium, The Genotype-Tissue Expression (GTEx) project. *Nat. Genet.* **45**, 580–585 (2013).

53. S. Niture, M. A. Gyamfi, M. Lin, U. Chimeh, X. Dong, W. Zheng, J. Moore, D. Kumar, TNFAIP8 regulates autophagy, cell steatosis, and promotes hepatocellular carcinoma cell proliferation. *Cell Death Dis.* **11**, 178 (2020).

54. E. Ferkingstad, P. Sulem, B. A. Atlason, G. Sveinbjornsson, M. I. Magnusson, E. L.

Styrmisdottir, K. Gunnarsdottir, A. Helgason, A. Oddsson, B. V. Halldorsson, B. O. Jensson, F. Zink, G. H. Halldorsson, G. Masson, G. A. Arnadottir, H. Katrinardottir, K. Juliusson, M. K. Magnusson, O. T. Magnusson, R. Fridriksdottir, S. Saevarsdottir, S. A. Gudjonsson, S. N. Stacey, S. Rognvaldsson, T. Eiriksdottir, T. A. Olafsdottir, V. Steinthorsdottir, V. Tragante, M. O. Ulfarsson, H. Stefansson, I. Jonsdottir, H. Holm, T. Rafnar, P. Melsted, J. Saemundsdottir, G. L. Norddahl, S. H. Lund, D. F. Gudbjartsson, U. Thorsteinsdottir, K. Stefansson, Large-scale integration of the plasma proteome with genetics and disease. *Nat. Genet.* **53**, 1712–1721 (2021).

55. I. J. Neeland, R. Ross, J.-P. Després, Y. Matsuzawa, S. Yamashita, I. Shai, J. Seidell, P. Magni, R. D. Santos, B. Arsenault, A. Cuevas, F. B. Hu, B. Griffin, A. Zambon, P. Barter, J.-C. Fruchart, R. H. Eckel, International Atherosclerosis Society, International Chair on Cardiometabolic Risk Working Group on Visceral Obesity, Visceral and ectopic fat, atherosclerosis, and cardiometabolic disease: a position statement. *Lancet Diabetes Endocrinol*. **7**, 715–725 (2019).

56. S. Agrawal, M. Wang, M. D. R. Klarqvist, K. Smith, J. Shin, H. Dashti, N. Diamant, S. H. Choi, S. J. Jurgens, P. T. Ellinor, A. Philippakis, M. Claussnitzer, K. Ng, M. S. Udler, P. Batra, A. V. Khera, Inherited basis of visceral, abdominal subcutaneous and gluteofemoral fat depots. *Nat. Commun.* **13**, 3771 (2022).

57. Y. Jiao, U. Ahmed, M. F. M. Sim, A. Bejar, X. Zhang, M. M. U. Talukder, R. Rice, J. Flannick, A. I. Podgornaia, D. F. Reilly, J. M. Engreitz, M. Kost-Alimova, K. Hartland, J.-M. Mercader, S. Georges, V. Wagh, M. Tadin-Strapps, J. G. Doench, J. M. Edwardson, J. J. Rochford, E. D. Rosen, A. R. Majithia, Discovering metabolic disease gene interactions by correlated effects on cellular morphology. *Mol Metab*. **24**, 108–119 (2019).

58. T. S. Mikkelsen, Z. Xu, X. Zhang, L. Wang, J. M. Gimble, E. S. Lander, E. D. Rosen, Comparative epigenomic analysis of murine and human adipogenesis. *Cell*. **143**, 156–169 (2010).

59. M. Ahmadian, J. M. Suh, N. Hah, C. Liddle, A. R. Atkins, M. Downes, R. M. Evans, PPARγ signaling and metabolism: the good, the bad and the future. *Nat. Med.* **19**, 557–566 (2013).

60. C. S. Ross-Innes, R. Stark, A. E. Teschendorff, K. A. Holmes, H. R. Ali, M. J. Dunning, G. D. Brown, O. Gojis, I. O. Ellis, A. R. Green, S. Ali, S.-F. Chin, C. Palmieri, C. Caldas, J. S. Carroll, Differential oestrogen receptor binding is associated with clinical outcome in breast cancer. *Nature*. **481**, 389–393 (2012).

61. M. C. Petersen, G. I. Shulman, Mechanisms of Insulin Action and Insulin Resistance. *Physiol. Rev.* **98**, 2133–2223 (2018).

62. I. Kudo, M. Murakami, Phospholipase A2 enzymes. *Prostaglandins Other Lipid Mediat.* **68-69**, 3–58 (2002).

63. J. P. Hardwick, K. Eckman, Y. K. Lee, M. A. Abdelmegeed, A. Esterle, W. M. Chilian, J. Y. Chiang, B.-J. Song, Eicosanoids in metabolic syndrome. *Adv. Pharmacol.* **66**, 157–266 (2013).

64. K. Song, X. Zhang, C. Zhao, N. T. Ang, Z. A. Ma, Inhibition of Ca2+-independent phospholipase A2 results in insufficient insulin secretion and impaired glucose tolerance. *Mol. Endocrinol.* **19**, 504–515 (2005).

65. E. Simon, C. Faucheux, A. Zider, N. Thézé, P. Thiébaud, From vestigial to vestigial-like: the Drosophila gene that has taken wing. *Dev. Genes Evol.* **226**, 297–315 (2016).

66. S. Niture, X. Dong, E. Arthur, U. Chimeh, S. S. Niture, W. Zheng, D. Kumar, Oncogenic Role of Tumor Necrosis Factor α-Induced Protein 8 (TNFAIP8). *Cells*. **8** (2018), doi:10.3390/cells8010009.

67. I. Kosti, N. Jain, D. Aran, A. J. Butte, M. Sirota, Cross-tissue Analysis of Gene and Protein Expression in Normal and Cancer Tissues. *Sci. Rep.* **6**, 24799 (2016).

68. A. L. Price, A. Helgason, G. Thorleifsson, S. A. McCarroll, A. Kong, K. Stefansson, Single-tissue and cross-tissue heritability of gene expression via identity-by-descent in related or unrelated individuals. *PLoS Genet.* **7**, e1001317 (2011).

69. C. Bycroft, C. Freeman, D. Petkova, G. Band, L. T. Elliott, K. Sharp, A. Motyer, D. Vukcevic, O. Delaneau, J. O'Connell, A. Cortes, S. Welsh, A. Young, M. Effingham, G. McVean, S. Leslie, N. Allen, P. Donnelly, J. Marchini, The UK Biobank resource with deep phenotyping and genomic data. *Nature*. **562**, 203–209 (2018).

70. B. Devlin, K. Roeder, Genomic control for association studies. *Biometrics*. **55**, 997–1004 (1999).

71. C. M. Moore, S. A. Jacobson, T. E. Fingerlin, Power and Sample Size Calculations for Genetic Association Studies in the Presence of Genetic Model Misspecification. *Hum. Hered.* **84**, 256–271 (2019).

72. E. Bernabeu, O. Canela-Xandri, K. Rawlik, A. Talenti, J. Prendergast, A. Tenesa, Sex differences in genetic architecture in the UK Biobank. *Nat. Genet.* **53**, 1283–1289 (2021).

73. B. K. Bulik-Sullivan, P.-R. Loh, H. K. Finucane, S. Ripke, J. Yang, Schizophrenia Working Group of the Psychiatric Genomics Consortium, N. Patterson, M. J. Daly, A. L. Price, B. M. Neale, LD Score regression distinguishes confounding from polygenicity in genome-wide association studies. *Nat. Genet.* **47**, 291–295 (2015).

74. I. Adzhubei, D. M. Jordan, S. R. Sunyaev, Predicting functional effect of human missense mutations using PolyPhen-2. *Curr. Protoc. Hum. Genet.* **Chapter 7**, Unit7.20 (2013).

75. P. C. Ng, S. Henikoff, SIFT: Predicting amino acid changes that affect protein function. *Nucleic Acids Res.* **31**, 3812–3814 (2003).

76. S. Chun, J. C. Fay, Identification of deleterious mutations within three human genomes. *Genome Res.* **19** (2009), doi:10.1101/gr.092619.109.

77. J. M. Schwarz, D. N. Cooper, M. Schuelke, D. Seelow, MutationTaster2: mutation prediction for the deep-sequencing age. *Nat. Methods*. **11** (2014), doi:10.1038/nmeth.2890.

78. S. Purcell, B. Neale, K. Todd-Brown, L. Thomas, M. A. R. Ferreira, D. Bender, J. Maller, P. Sklar, P. I. W. de Bakker, M. J. Daly, P. C. Sham, PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am. J. Hum. Genet.* **81**, 559–575 (2007).

79. N. DeForest, B. Kavitha, S. Hu, R. Isaac, L. Krohn, M. Wang, X. Du, C. De Arruda Saldanha, J. Gylys, E. Merli, R. Abagyan, L. Najmi, V. Mohan, Alnylam Human Genetics, AMP-T2D Consortium, J. Flannick, G. M. Peloso, P. L. S. M. Gordts, S. Heinz, A. M. Deaton, A. V. Khera, J. Olefsky, V. Radha, A. R. Majithia, Human gain-of-function variants in HNF1A confer protection from diabetes but independently increase hepatic secretion of atherogenic lipoproteins. *Cell Genom*. **3**, 100339 (2023).

80. N. L. Bray, H. Pimentel, P. Melsted, L. Pachter, Near-optimal probabilistic RNA-seq quantification. *Nat. Biotechnol.* **34**, 525–527 (2016).

81. G. E. Hoffman, P. Roussos, Dream: powerful differential expression analysis for repeated measures designs. *Bioinformatics*. **37**, 192–201 (2021).

82. D. Tingley, T. Yamamoto, K. Hirose, L. Keele, K. Imai, Mediation:RPackage for causal mediation analysis. *J. Stat. Softw.* **59** (2014), doi:10.18637/jss.v059.i05.

CONCLUSION

Insulin resistance poses a substantial public health concern, as insulin resistance often remains undetected but is responsible for causing major epidemic diseases globally, including type 2 diabetes, cardiovascular disease, and fatty liver disease. In this work, I present a framework to investigate the molecular mechanisms of insulin resistance related disorders and discover novel insulin resistance effector genes using human population genetics in combination with high-throughput functional genomics.

Starting in Chapter 1, I detail how insulin resistance and more specifically, type 2 diabetes, is a multifactorial and highly heritable syndrome. Through a comprehensive review of recent genetic association studies for type 2 diabetes, I demonstrate how human genetic studies have and continue to show great potential to identify therapeutic targets for insulin resistance and uncover novel mechanisms of disease progression.

In Chapter 2, I exemplify how high-throughput functional genomics can be employed in combination with large-scale sequenced biobanks to characterize novel molecular mechanisms of insulin resistance related disorders. Specifically, by performing a deep mutational scan of the insulin resistance gene HNF1A and conducting clinical correlations with over half a million exome-sequenced individuals, we identify and characterize a novel class of function increasing variants which pleiotropically protect from diabetes through pancreatic mechanisms but promote a pro-atherogenic serum profile. In this work, we not only present an efficient strategy, rooted in human genetics, for elucidating molecular mechanisms of implicated metabolic disease genes, but also advance the field by characterizing novel gain-of-function mechanisms of HNF1A in driving metabolic disease.

In Chapter 3 I employ plasma proteome-wide Mendelian randomization to prioritize proteins casual for NASH, specifically identifying and characterizing RSPO3. In this work, I find RSPO3 causally increases NASH in part by causing a metabolically harmful fat distribution, and nominate potential tissues and cell types of action for ongoing functional characterization of RSPO3. This work advances the field of biomarker and target discovery for NASH by not only prioritizing RSPO3 as a potential NASH mediator, but also nominating 7 other secreted proteins with putatively causal effects on NASH from evidence directly from human participants. These proteins can be investigated further in downstream experiments to understand how each alters NASH pathogenesis and additionally can be evaluated for therapeutic potential as a biomarker or treatment for NASH.

Furthermore in Chapter 4, I harness genome-wide association studies in combination with diverse genomic data resources to prioritize insulin resistance effector genes. By integrating independent genetic and genomic datasets, I further characterize directions-of-effect of prioritized genes on metabolic disease as well as provide insights into tissues of action. From this integrative workflow, I highlight two phospholipase encoding genes PLA2G12A and PLA2G6 which improve insulin sensitivity along with VGLL3, a transcriptional co-factor that increases insulin resistance partially through enhanced adiposity. From sex-specific dissection of insulin resistance, I additionally implicate the anti-apoptotic factor TNFAIP8 in insulin-sensitization specifically in females, which acts by increasing visceral adiposity. In summary, the work described in this chapter reports on GWAS of TG/HDL ratio at population scale, expanding knowledge of the genetic determinants of insulin resistance to include dozens of new loci/genes not

previously implicated. Further, I provide specific, testable hypotheses for future investigation to credential these nominated genes as potential new therapeutic targets and/or biomarkers for insulin resistance.

In Chapters 3 and 4, I employ two different human genetic approaches to prioritize new modulators of insulin resistance which present strong evidence directly in humans for potential to serve as biomarkers or pharmacological targets of insulin resistance and its related disorders (Table 3.1, 4.1). These prioritized effectors identified through our comprehensive analyses not only advance the field of human genetic investigation of insulin resistance and its related disorders, but also provide a solid foundation for further investigation of top candidates in downstream experiments. For instance, building upon the findings presented in Chapter 3, research is currently underway to examine the effect of altered RSPO3 in liver spheroids, a physiologically represented liver system which encompasses the dynamic relationships between different cell types. This ongoing work seeks to understand the complex cell-type specific interplay of RSPO3 in the liver and aims to provide further insights into the direct role of RSPO3 in causing NASH.

Additionally, to build upon the findings in Chapter 4, ongoing research is currently underway to develop disease-calibrated, high-throughput functional assays in liver (hepatocytes) and adipose (adipocytes) which can be used to further validate and investigate the effects of the prioritized insulin resistance genes on metabolic disease. For instance, prioritized insulin resistance genes with evidence suggesting liver as a tissue of action will be further investigated via gene ablation in an internally developed high-throughput image-based assay in human hepatocytes. This approach will facilitate understanding of the effect of nominated genes on key liver-specific metabolic processes,

specifically liver fat accumulation quantified by high-throughput microscopy. Additionally, for nominated insulin resistance genes which evidence points to adipose as a primary tissue of action, current research is underway to develop a disease-relevant, high-throughput functional assay which will test the metabolic effect of gene knockout in adipocytes. Several pathophysiologically relevant readouts are currently being explored, such as quantifying the effects of gene knockout on adipocyte differentiation or lipid accumulation. Taken together, the work detailed in Chapter 3 and 4 sets the foundation for downstream research to further investigate the role of the prioritized proteins and genes in insulin resistance and metabolic disease pathogenesis, some of which may be promising candidates for future investigation using deep mutational scanning integrated with large-scale sequenced cohorts, as detailed in Chapter 2.

In summary, in this work I demonstrate how this integrated approach, rooted in human genetics, offers promising avenues for the identification of novel biomarkers and therapeutic targets of metabolic disease, and ultimately advances the field of research to effectively manage and treat insulin resistance and its related disorders.