From Gene Discovery to Understanding and Predicting Cardiometabolic Disease

SARA M. WILLEMS

From Gene Discovery to Understanding and Predicting Cardiometabolic Disease

Sara M. Willems

ACKNOWLEDGMENTS

From Gene Discovery to Understanding and Predicting Cardiometabolic Disease

Van het ontdekken van genen naar begrijpen en voorspellen
van cardiometabole aandoeningen

Proefschrift

ter verkrijging van de graad van doctor
aan de Erasmus Universiteit Rotterdam
op gezag van de rector magnificus
Prof.dr. H.A.P. Pols
en volgens besluit van het College voor Promoties.

De openbare verdediging zal plaatsvinden op
donderdag 18 december 2014 om 9:30 uur

door

Sara Marie Willems
geboren te Leiden

PROMOTIECOMMISSIE

Promotor: Prof.dr. C.M. van Duijn

Overige leden: Prof.dr. O.H. Franco Duran
                  Prof.dr. E.J.G. Sijbrands
                  Prof.dr. N.J. Wareham

Copromotor: Dr. A. Isaacs

"The known is finite, the unknown infinite; intellectually we stand on an island in the midst of an illimitable ocean of inexplicability. Our business in every generation is to reclaim a little more land."

Thomas Henry Huxley

Voor pappa en mamma

Voor Koen

TABLE OF CONTENTS

PUBLICATIONS AND MANUSCRIPTS BASED ON THE STUDIES DESCRIBED IN THIS THESIS

Chapter 2.1
Sara M. Willems, Raluca Mihaescu, Eric J.G. Sijbrands, Cornelia M. van Duijn, A. Cecile J.W. Janssens
A methodological perspective on genetic risk prediction studies in type 2 diabetes: recommendations for future research
Curr Diab Rep (2011) 11:511-518

Chapter 2.2
Sara M. Willems, Belinda K. Cornes, Jennifer A. Brody, Alanna C. Morrison, Leonard Lipovich, Marco Dauriz, Bianca Porneala, Yuning Chen, Ching-Ti Liu, Denis V. Rybin, Richard A. Gibbs, Donna Muzny, James S. Pankow, Bruce M. Psaty, Eric Boerwinkle, Jerome I. Rotter, David S. Siscovick, Ramachandran S. Vasan, Robert C. Kaplan, Aaron Isaacs, Josée Dupuis, Cornelia M. van Duijn, James B. Meigs
Association of the *IGF1* Gene with Fasting Insulin levels
Submitted

Chapter 2.3
Jennifer Wessel*, Audrey Y. Chu*, Sara M. Willems*, Shuai Wang*, Hanieh Yaghootkar, Jennifer A. Brody, Marco Dauriz, Marie-France Hivert, Sridharan Raghavan, Leonard Lipovich, Bertha Hidalgo, Keolu Fox, Jennifer E. Huffman, Ping An, Yingchang Lu, Laura J. Rasmussen-Torvik, Niels Grarup, Margaret G. Ehm, Li Li, Abigail S. Baldridge, Alena Stančáková, Ravinder Abrol, Céline Besse, Anne Boland, Jette Bork-Jensen, Myriam Fornage, Daniel F. Freitag, Melissa E. Garcia, Xiuqing Guo, Kazuo Hara, Aaron Isaacs, Johanna Jakobsdottir, Leslie A. Lange, Jill C. Layton, Man Li, Jing Hua Zhao, Karina Meidtner, Alanna C. Morrison, Mike A. Nalls, Marjolein J. Peters, Maria Sabater-Lleal, Claudia Schurmann, Angela Silveira, Albert V. Smith, Lorraine Southam, Marcus H. Stoiber, Rona J. Strawbridge, Kent D. Taylor, Tibor V. Varga, Kristine H. Allin, Najaf Amin, Jennifer L. Aponte, Tin Aung, Caterina Barbieri, Nathan A. Bihlmeyer, Michael Boehnke, Cristina Bombieri, Donald W. Bowden, Sean M. Burns, Yuning Chen, Yii-Der I. Chen, Ching-Yu Cheng, Adolfo Correa, Jacek Czajkowski, Abbas Dehghan, Georg B. Ehret, Gudny Eiriksdottir, Stefan A. Escher, Aliki-Eleni Farmaki, Mattias Frånberg, Giovanni Gambaro, Franco Giulianini, William A. Goddard III, Anuj Goel, Omri Gottesman, Megan L. Grove, Stefan Gustafsson, Yang Hai, Göran Hallmans, Jiyoung Heo, Per Hoffmann, Mohammad K. Ikram, Richard A. Jensen, Marit E. Jørgensen, Torben Jørgensen, Maria Karaleftheri, Chiea C. Khor, Andrea Kirkpatrick, Aldi T. Kraja, Johanna Kuusisto, Ethan M. Lange, I.T. Lee, Wen-Jane Lee, Aaron Leong, Jiemin Liao, Chunyu Liu, Yongmei Liu, Cecilia M. Lindgren, Allan Linneberg, Giovanni

Malerba, Vasiliki Mamakou, Eirini Marouli, Nisa M. Maruthur, Angela Matchan, Roberta McKean, Olga McLeod, Ginger A. Metcalf, Karen L. Mohlke, Donna M. Muzny, Ioanna Ntalla, Nicholette D. Palmer, Dorota Pasko, Andreas Peter, Nigel W. Rayner, Frida Renström, Ken Rice, Cinzia F. Sala, Bengt Sennblad, Ioannis Serafetinidis, Jennifer A. Smith, Nicole Soranzo, Elizabeth K. Speliotes, Eli A. Stahl, Kathleen Stirrups, Nikos Tentolouris, Anastasia Thanopoulou, Mina Torres, Michela Traglia, Emmanouil Tsafantakis, Sundas Javad, Lisa R. Yanek, Eleni Zengini, Diane M. Becker, Josh C. Bis, James B. Brown, L. Adrienne Cupples, Torben Hansen, Erik Ingelsson, Andrew J. Karter, Carlos Lorenzo, Rasika A. Mathias, Jill M. Norris, Gina M. Peloso, Wayne H.-H. Sheu, Daniela Toniolo, Dhananjay Vaidya, Rohit Varma, Lynne E. Wagenknecht, Heiner Boeing, Erwin P. Bottinger, George Dedoussis, Panos Deloukas, Ele Ferrannini, Oscar H. Franco, Paul W. Franks, Richard A. Gibbs, Vilmundur Gudnason, Anders Hamsten, Tamara B. Harris, Andrew T. Hattersley, Caroline Hayward, Albert Hofman, Jan-Håkan Jansson, Claudia Langenberg, Lenore J. Launer, Daniel Levy, Ben A. Oostra, Christopher J. O'Donnell, Stephen O'Rahilly, Sandosh Padmanabhan, James S. Pankow, Ozren Polasek, Michael A. Province, Stephen S. Rich, Paul M Ridker, Igor Rudan, Matthias B. Schulze, Blair Smith, André G. Uitterlinden, Mark Walker, Hugh Watkins, Tien Y. Wong, Eleftheria Zeggini, Generation Scotland, The EPIC-InterAct Consortium, Markku Laakso, Ingrid B. Borecki, Daniel I. Chasman, Oluf Pedersen, Bruce M. Psaty, E. Shyong Tai, Cornelia M. van Duijn, Nicholas J. Wareham, Dawn M. Waterworth, Eric Boerwinkle, WH Linda Kao, Jose C. Florez, Ruth J.F. Loos, James G. Wilson, Timothy M. Frayling, David S. Siscovick, Josée Dupuis, Jerome I. Rotter, James B. Meigs, Robert A. Scott, Mark O. Goodarzi
Low-frequency and rare exome chip variants associate with fasting glucose and type 2 diabetes susceptibility
Nat Commun in press

Chapter 3.1
Sara M. Willems, Aaron Isaacs, Abbas Dehghan, Maksim V. Struchalin, Elisabeth M. van Leeuwen, Ben A. Oostra, Albert Hofman, André G. Uitterlinden, Oscar H. Franco, Eric J. G. Sijbrands, Cornelia M. van Duijn
The role of common lipid-altering gene variants in the risk of dyslipidemia through old age
Submitted

Chapter 3.2
Aaron Isaacs*, Sara M. Willems*, Daniel Bos, Abbas Dehghan, Albert Hofman, M. Arfan Ikram, André G. Uitterlinden, Ben A. Oostra, Oscar H. Franco, Jacqueline C. Witteman and Cornelia M. van Duijn

Risk scores of common genetic variants for lipid levels influence atherosclerosis and incident coronary heart disease
Arterioscler Thromb Vasc Biol (2013) 33:2233-2239

Chapter 4.1
Sara M. Willems, Abbas Dehghan, Symen Ligthart, Jeannette M. Vergeer-Drop, Albert Hofman, Eric J. G. Sijbrands, André G. Uitterlinden, Ben A. Oostra, Jacqueline C. M. Witteman, Oscar H. Franco, Cornelia M. van Duijn, Aaron Isaacs
Risk scores comprised of common lipid-altering genetic variants are associated with lipid levels and suggest an altered role of common genetic variation in type 2 diabetes
Submitted

Chapter 1

General Introduction

Cardiovascular diseases (CVD) are the leading cause of morbidity and the number one cause of death worldwide.[1] An estimated 17.3 million people died from CVDs in 2008, including an estimated 7.3 million due to coronary heart disease (CHD) and 6.2 million due to stroke. The number of people dying from CVDs is expected to increase to 23.3 million in 2030 if no improvements in prevention and treatment will be implemented.[1] For many years, CVD was considered an inevitable consequence of aging and knowledge about the pathophysiology was limited. In 1961, Kannel et al[2] were the first to provide convincing evidence for age, sex (male), smoking, hypertension, diabetes and dyslipidemia as important risk factors for CVD, laying the foundation for the successful primary and secondary prevention programs that have been implemented. Despite these successes, CVDs impose a major burden on human health and healthcare systems.

An increasing portion of CVD cases can be prevented by addressing the modifiable risk factors, including type 2 diabetes (T2D) and dyslipidemia.[1] Several studies have shown that the relative risk for coronary deaths from diabetes is around 2.5 in women and around 2 in men[3] and a meta-analysis showed that a 1 mmol/L decrease in low-density lipoprotein cholesterol (LDL-C) level is associated with a 19% reduction in CHD mortality.[4] The high impact and frequency make T2D and dyslipidemia suitable candidates for targeting preventive interventions, such as medication, weight loss, and increased physical activity, which can prevent, slow, or even reverse the development of these risk factors and thus reduce the burden of CVDs.[5,6]

T2D is characterized by hyperglycemia due to insulin resistance in muscle and liver, causing impaired glucose uptake and impaired suppression of hepatic glucose production in response to insulin, and progressive impairment of insulin secretion by the pancreatic β-cells.[7,8] An estimated 285 million people worldwide have diabetes and this number is expected to increase by more than 50% in the next 20 years if no effective preventive strategies are implemented.[9] It is a multifactorial disease, caused by a complex interplay between genetic and nongenetic factors. Important nongenetic factors in the etiology of T2D are increasing age, higher body mass index, impaired fasting glucose, impaired glucose tolerance, higher glycated hemoglobin (HbA1c) level, and metabolic syndrome.[10–15] Heritability estimates are moderate to high, ranging from 26 to 69%.[16,17] Over the past few years, knowledge of the genetic variants underlying this heritability has rapidly increased through collaborations in large genetic consortia for gene discovery. These collaborative efforts have identified dozens of single nucleotide polymorphisms (SNPs) associated with T2D and the related quantitative traits fasting glucose (FG) and fasting insulin (FI) in the general population.[18–21] Together, the currently known SNPs explain about 10% of the heritability of T2D.

Dyslipidemia is a broad term encompassing several lipid disorders. In this thesis, the criteria used are those that have been implemented by the Dutch College of General Practitioners to identify individuals at increased CVD risk, based on both total cholesterol (TC) and high-density lipoprotein cholesterol (HDL-C) (TC ≥ 6.5 mmol/L when TC/HDL-C ratio ≥ 5, TC < 6.5 mmol/L when TC/HDL-C ratio > 8, or TC ≥ 9 mmol/L independent of TC/HDL-C ratio) and the criteria that are currently used based on TC alone (TC > 6.5 mmol/L or use of lipid lowering medication).[22] The prevalence of dyslipidemia in the general population is high. Data from nationally representative health examination surveys from England, Germany, Japan, Jordan, Mexico, Scotland, Thailand and the United States showed prevalences of elevated TC (TC ≥ 6.2 mmol/L) in adults aged 40-79 years from 19% in Mexico to 62% in Germany.[23] In individuals with T2D, levels of TC and LDL-C are similar to those in non-diabetic individuals; however, HDL-C levels are typically decreased and triglyceride (TG) levels increased compared to non-diabetic individuals.[24] As in the etiology of type 2 diabetes, genetic factors play an important role in the etiology of dyslipidemia with heritability estimates ranging from 24 to 56% for blood lipid levels (TC: h2=35%, LDL-C: h2=30%, HDL-C: h2=56%, TG: h2=24%).[25] Large numbers of SNPs associated with blood lipid levels in the general population have been identified in the past few years, explaining 25 to 30% of the heritability.[26,27]

The identification of large numbers of SNPs has raised the question of whether this genetic information can be used to identify individuals at high risk of T2D or dyslipidemia. Genetic information can be attractive for early risk assessment following the successes of phenylketonuria (PKU) prevention. Screening at birth for PKU has effectively prevented mental retardation.[28] Also, genetic screening of families with a history of familial hypercholesterolemia (FH), an autosomal co-dominant genetic disorder associated with increased levels of LDL-C, has proved useful in terms of sensitivity, quality adjusted life years and cost effectiveness[29,30]. Genotypes are invariant over time, which enables identification of individuals at high risk of developing morbidity at an early age and early interventions before the pathogenesis of disease leads to irreversible damage. However, the effect sizes of the variants associated with most complex traits and diseases, including T2D and dyslipidemia in the general population, are very modest. Per allele odds ratios (OR) for T2D are typically around 1.10.[18] Even the strongest susceptibility variant, rs7903146 in the *TCF7L2* gene (OR=1.39)[18], is a weaker predictor of T2D risk than most nongenetic risk factors. The variants associated with lipid levels typically result in a change of about 0.03 standard deviation (SD) in lipid value per copy of the effect allele.[27] The variants with the largest effects on TC, LDL-C, HDL-C and TG (these are variants in the *LDLR*, *APOE*, *CETP* and *APOA1* loci) are associated with changes of about 0.2 SD in lipid value per copy of the effect allele. Evidently, the low effect sizes make single genetic risk factors unsuitable for identifying individuals at high risk of T2D or dyslipidemia,

but there has been increasing interest in investigating the extent to which genetic risk factors combined can improve the prediction of disease. This approach was outlined by Fisher in the previous century, who predicted that multiple risk alleles, following a Gaussian distribution in the population, account for the phenotypic variability observed in the population.[29]

In addition to identification of high risk individuals, findings from large genome-wide association studies can improve the understanding of disease etiology, which can lead to identification of novel targets for therapeutic interventions. Several approaches have been applied to the currently well-established common T2D and lipid SNPs to identify the biological processes they might be involved in. These approaches have identified cell cycle regulation, adipocytokine signalling and CREBBP-related transcription factor activity as key processes involved in the pathogenesis of T2D.[18] For lipid levels, many of the identified loci harbor genes involved in lipid metabolism, which has been validated in mouse models.[26] In addition, large numbers of genes have been suggested on the basis of literature review, pathway analysis, regulation of mRNA expression levels, and protein altering variants, as interesting candidates to take forward to functional studies.[27]

Despite the large numbers of SNPs influencing lipid levels and T2D risk that have been identified, a large portion of the estimated heritability is still unexplained. Both for identification of individuals at increased risk of disease and to further improve the understanding of disease etiology, which can lead to better prevention, diagnosis and treatment, it is important to find the explanation for this "missing heritability". Because efforts have largely focused on common genetic variants (minor allele frequency (MAF) ≥ 5%), one hypothesis is that low frequency (defined in this thesis as MAF 1-5%) and rare (defined in this thesis as MAF <1%) variants could explain part of the missing heritability.[32,33] This is supported by recent large-scale sequencing studies that have reported that rapid expansions in the human population have introduced a substantial number of rare genetic variants[34,35], with purifying selection having had little time to act, which may harbor larger effects on complex traits than those observed for common variants.[35-39]

AIMS OF THIS THESIS

In this thesis I aim to improve our understanding of the etiology of T2D and dyslipidemia and to investigate the extent to which genetic risk factors combined can improve their prediction. Chapter 2 includes genetic studies of T2D and diabetes related quantitative traits. In Chapter 2.1, I give an overview of published genetic risk prediction studies for

T2D from a methodological perspective. In Chapter 2.2, I study the association of the *IGF1* gene with fasting insulin levels to obtain more insight into GWAS findings near this gene and to identify and characterize novel genetic variants at the locus. In Chaper 2.3, I describe exome-wide association analyses to identify rare and low-frequency variants, with potentially larger effect sizes, associated with FG, FI and T2D. Chapter 3 focuses on risk scores comprised of the known common genetic lipid variants. In Chapter 3.1, I investigate their ability to identify individuals at high risk of dyslipidemia through old age and, in Chapter 3.2, their association with subclinical atherosclerosis and incident coronary heart disease. In Chapter 4, I explore the influence of the lipid gene risk scores on lipid levels and dyslipidemia in the context of T2D. Ultimately, in Chapter 5, I discuss the findings of this thesis and their implications for future research.

REFERENCES

1. World Health Organization. Media centre, Fact Sheets: Cardiovascular diseases (CVDs). http://www.who.int/mediacentre/factsheets/fs317/en/ (July 2014)

2. Kannel, W.B., Dawber, T.R., Kagan, A., Revotskie, N. & Stokes, J., 3rd. Factors of risk in the development of coronary heart disease--six year follow-up experience. The Framingham Study. Ann Intern Med 55, 33-50 (1961).

3. Duckworth, W.C. Hyperglycemia and cardiovascular disease. Curr Atheroscler Rep 3, 383-91 (2001).

4. Baigent, C. et al. Efficacy and safety of cholesterol-lowering treatment: prospective meta-analysis of data from 90,056 participants in 14 randomised trials of statins. Lancet 366, 1267-78 (2005).

5. Gillies, C.L. et al. Pharmacological and lifestyle interventions to prevent or delay type 2 diabetes in people with impaired glucose tolerance: systematic review and meta-analysis. BMJ 334, 299 (2007).

6. Colagiuri, S. Epidemiology of prediabetes. Med Clin North Am 95, 299-307, vii (2011).

7. Defronzo, R.A. Banting Lecture. From the triumvirate to the ominous octet: a new paradigm for the treatment of type 2 diabetes mellitus. Diabetes 58, 773-95 (2009).

8. Wallia, A. & Molitch, M.E. Insulin therapy for type 2 diabetes mellitus. JAMA 311, 2315-25 (2014).

9. International Diabetes Federation: Diabetes and impaired glucose tolerance. Global burden: prevalence and projections, 2010 and 2030. http://www.idf.org/diabetesatlas/diabetes-and-impaired-glucose-tolerance (July 2014)

10. Flegal, K.M., Carroll, M.D., Ogden, C.L. & Curtin, L.R. Prevalence and trends in obesity among US adults, 1999-2008. JAMA 303, 235-41 (2010).

11. Nguyen, N.T., Nguyen, X.M., Lane, J. & Wang, P. Relationship between obesity and diabetes in a US adult population: findings from the National Health and Nutrition Examination Survey, 1999-2006. Obes Surg 21, 351-5 (2011).

12. Karve, A. & Hayward, R.A. Prevalence, diagnosis, and treatment of impaired fasting glucose and impaired glucose tolerance in nondiabetic U.S. adults. Diabetes Care 33, 2355-9 (2010).

13. Gerstein, H.C. et al. Annual incidence and relative risk of diabetes in people with various categories of dysglycemia: a systematic overview and meta-analysis of prospective studies. Diabetes Res Clin Pract 78, 305-12 (2007).

14. Selvin, E. et al. Glycated hemoglobin, diabetes, and cardiovascular risk in nondiabetic adults. N Engl J Med 362, 800-11 (2010).

15. Alexander, C.M. et al. NCEP-defined metabolic syndrome, diabetes, and prevalence of coronary heart disease among NHANES III participants age 50 years and older. Diabetes 52, 1210-4 (2003).

16. Poulsen, P., Kyvik, K.O., Vaag, A. & Beck-Nielsen, H. Heritability of type II (non-insulin-dependent) diabetes mellitus and abnormal glucose tolerance--a population-based twin study. Diabetologia 42, 139-45 (1999).

17. Almgren, P. et al. Heritability and familiality of type 2 diabetes and related quantitative traits in the Botnia Study. Diabetologia 54, 2811-9 (2011).

18. Morris, A.P. et al. Large-scale association analysis provides insights into the genetic architecture and pathophysiology of type 2 diabetes. Nat Genet 44, 981-90 (2012).

19. Dupuis, J. et al. New genetic loci implicated in fasting glucose homeostasis and their impact on type 2 diabetes risk. Nat Genet 42, 105-16 (2010).

20. Manning, A.K. et al. A genome-wide approach accounting for body mass index identifies genetic variants influencing fasting glycemic traits and insulin resistance. Nat Genet 44, 659-69 (2012).

21. Scott, R.A. et al. Large-scale association analyses identify new loci influencing glycemic traits and provide insight into the underlying biological pathways. Nat Genet 44, 991-1005 (2012).

22. NHG-Standaard cardiovasculair risicomanagement. https://www.nhg.org/standaarden/samen-vatting/cardiovasculair-risicomanagement (December 2013)

23. Gregory A Roth, S.D.F., Ali H Mokdad, Wichai Aekplakorn, Toshihiko Hasegawa, Stephen S Lim. High total serum cholesterol, medication coverage and therapeutic control: an analysis of national health examination survey data from eight countries. Bull World Health Organ, 92-101 (2011).

24. Kannel, W.B. Lipids, diabetes, and coronary heart disease: insights from the Framingham Study. Am Heart J 110, 1100-7 (1985).

25. Isaacs, A. et al. Heritabilities, apolipoprotein E, and effects of inbreeding on plasma lipids in a genetically isolated population: the Erasmus Rucphen Family Study. Eur J Epidemiol 22, 99-105 (2007).

26. Teslovich, T.M. et al. Biological, clinical and population relevance of 95 loci for blood lipids. Nature 466, 707-13 (2010).

27. Global Lipids Genetics Consortium, Willer, C.J. et al. Discovery and refinement of loci associated with lipid levels. Nat Genet 45, 1274-83 (2013).

28. Groselj, U., Tansek, M.Z. & Battelino, T. Fifty years of phenylketonuria newborn screening - A great success for many, but what about the rest? Mol Genet Metab 113, 8-10 (2014).

29. Hovingh, G.K., Davidson, M.H., Kastelein, J.J. & O'Connor, A.M. Diagnosis and treatment of familial hypercholesterolaemia. Eur Heart J 34, 962-71 (2013).

30. Elucigene FH20 and LIPOchip for the diagnosis of familial hypercholesterolaemia (DG2). National Institute for Health and Clinical Excellence. http://guidance.nice.org.uk/DG2 (December 2013).

31. Fisher, R.A. The correlation between relatives on the upposition of Mendelian inheritance. Phil Trans R Soc Edinburgh, 399-433 (1918).

32. Manolio, T.A. et al. Finding the missing heritability of complex diseases. Nature 461, 747-53 (2009).

33. Zuk, O. et al. Searching for missing heritability: designing rare variant association studies. Proc Natl Acad Sci U S A 111, E455-64 (2014).

34. Keinan, A. & Clark, A.G. Recent explosive human population growth has resulted in an excess of rare genetic variants. Science 336, 740-3 (2012).

35. Nelson, M.R. et al. An abundance of rare functional variants in 202 drug target genes sequenced in 14,002 people. Science 337, 100-4 (2012).

36. Coventry, A. et al. Deep resequencing reveals excess rare recent variants consistent with explosive population growth. Nat Commun 1, 131 (2010).

37. Fu, W. et al. Analysis of 6,515 exomes reveals the recent origin of most human protein-coding variants. Nature 493, 216-20 (2013).

38. Morrison, A.C. et al. Whole-genome sequence-based analysis of high-density lipoprotein cholesterol. Nat Genet 45, 899-901 (2013).

39. Tennessen, J.A. et al. Evolution and functional impact of rare coding variation from deep sequencing of human exomes. Science 337, 64-9 (2012).

Chapter 2

From gene discovery to understanding and predicting type 2 diabetes

```
                              ....TGACTACGT...
                      ...CTACTAGCACTGTACACG...              ..CTACGTACG
                  ..GTACGTAGTCTGATCAAAACGTA...         .GCTATACAGCTA
                ..CGTTGGTACCAGTACGTACGACTGCGA..     .GTACGTACGTACG
              .GTGTGTCACACAGACACGTGATATAGACCAGATGACACACGTAC
            ..GTACGACGGCGCGGCGCGATATATATAAAAGCACGGACACTACGA
           ..ACGATGACTGACTACGTGCGCGCGTAGCTAGCTACGGATAGGACTA
          ..TACTGACTACGTACTACTAAAGCTGCTTTTACGTACGTACGGGCATC
          .CATGTACTATCGCGGCGCGTAGACTGTACGACCAGACTGACTGACT
          .GACTGACTGACGCGCGTACGTAGCTACTGTAGTACGTACGTACGTA
           .TTCAGACTGACTGACTGAGATTACAGCTACGTAGTACTGACACGT
           ..TACGACGGCGCGGCGCGATATATATAAAAGCACGGACACTACGA
            .GATGACTGACTACGTGCGCGCGTAGCTAGCTACGGATAGGACTA
             .TGACTACGTACTACTAAAGCTGCTTTTACGTACGTACGGGCATC
              .CTGCGCATGTACTATCGCGGCGCGTAGACTGTACGACCAGAC
               .ACTGACGCGCGTACGTAGCTACTGTAGTACGTACGTACGTA
                .ACTGACTGACTGAGATTACAGCTACGTAGTACTGACACGT
                 .GCATACGAGTCGAAGCGCGCAATATATATTATATCGGCC
                  .GTGACTACATCATGCTGACTACGGTCGCGCTCAACG
                   .TGACTACGTACGTATACGACGTACTGACGGCGCGC
                    .GCACTGTACACGACTGACTTACTAGCTACGTACC
                     .TGATCAAAACGTACGCGCGGCTATACAGCTA
                       .CGTACGACTGCGATACGTACGTACGTACG
                        .CGTGATATAGACCAGATGACACACG
                          .TATATATAAAAGCACGGACAC
                            .GTAGCTAGCTACGGAT
                              .CTTTTACGTACGT
                                .ACGACCA
                                 .TAGT
```

A methodological perspective on genetic risk prediction studies in type 2 diabetes: recommendations for future research

ABSTRACT

Fueled by the successes of genome-wide association studies, numerous studies have investigated the predictive ability of genetic risk models in type 2 diabetes. In this paper, we review these studies from a methodological perspective, focusing on the variables included in the risk models as well as the study designs and populations investigated. We argue and show that differences in study design and characteristics of the study population have an impact on the observed predictive ability of risk models. This observation emphasizes that genetic risk prediction studies should be conducted in those populations in which the prediction models will ultimately be applied, if proven useful. Of all genetic risk prediction studies to date, only a few were conducted in populations that might be relevant for targeting preventive interventions.

INTRODUCTION

Type 2 diabetes (T2D) is a multifactorial disease, caused by a complex interplay between genetic and nongenetic risk factors. Compelling evidence has identified increasing age, higher body mass index (BMI), impaired fasting glucose, impaired glucose tolerance, higher glycated hemoglobin (HbA1c) level, and metabolic syndrome as important T2D risk factors (Table 1)[1–10]. These nongenetic factors have a substantial impact on disease risk and are frequent. For example, metabolic syndrome poses an eight times higher T2D risk and is present in more than 40% of the individuals over 50 years of age. The high impact and frequency make these risk factors suitable candidates for targeting preventive interventions, such as medication, weight loss, and increased physical activity that can slow down or even reverse the disease process[11,12].

In the past 5 years, genome-wide association studies have identified and replicated over 40 single nucleotide polymorphisms (SNPs) that predispose to T2D [13,14]. However, the effect sizes of the associated variants are very modest, with per allele odds ratios ranging from 1.05 to 1.35[13]. Even the strongest susceptibility variant, rs7903146 in the *TCF7L2* gene, is a weaker predictor of T2D risk than most nongenetic risk factors. Evidently, the low effect sizes make single genetic risk factors unsuitable for targeting preventive interventions, but there is increasing interest in investigating the extent to which genetic risk factors combined can improve the prediction of the disease.

An improvement in the early identification of high-risk groups is warranted because T2D imposes a great burden on human health and health care systems[15,16]. An estimated 285 million people worldwide have diabetes[15] and this number is expected to increase by more than 50% in the next 20 years if no preventive strategies are implemented[15]. To identify high-risk individuals, many risk prediction models have been proposed.

Guidelines for T2D prevention advocate the use of clinical risk scores as primary screening tools, followed by blood glucose measurements to detect individuals with impaired fasting glucose, impaired glucose tolerance, or metabolic syndrome[17]. Examples of commonly used risk scores include the FINDRISC (Finnish Diabetes Risk Score) and the Diabetes Risk Calculator[18,19]. The FINDRISC score is based on age, BMI, waist circumference, use of antihypertensive medication, history of elevated blood glucose, daily physical activity and daily intake of fruits or vegetables, and the Diabetes Risk Calculator on age, waist circumference, gestational diabetes, height, race/ethnicity, hypertension, family history of diabetes, and exercise.

The predictive ability of these clinical risk scores is modest, but satisfactory. The area under the receiver operating characteristic curve (AUC) is a commonly used measure to indicate the predictive ability; the AUC indicates the discriminative accuracy of a prediction model. To generate the curve, on the x-axis 1-specificity is plotted, and on the y-axis sensitivity is plotted. The AUC value represents the probability that the predicted risk of

Table 1. Risk factors for type 2 diabetes

Risk factor	Population	Frequency (%)	Diabetes risk (%)[+]	RR[#]
Age (years)				
0-44	General US population	61.3 [1]	1.7 [2]	1
45-64		25.9	12.2	7.2
65-74		6.8	19.9	11.7
75+		6.1	17.9	10.5
Sex				
Female	General US population	50.7 [1]	5.9 [3]	1
Male		49.3	6.6	1.1
BMI (kg/m^2)				
< 25	US adults aged ≥ 20 years	32.0 [4, 5]	8 [6]	1
25-<30		34.2	15	1.9
30-<35		19.5	23	2.9
35-<40		8.6	33	4.1
≥40		5.7	43	5.4
IFG/IGT				
Normoglycemic	Nondiabetic US adults aged ≥ 18 years (Frequency) Global cohorts (Diabetes risk and RR)	65.4 [7]	NA [8]	1[†] [8]
IGT only		5.4	4.4 – 6.4[*]	5.5[†]
IFG only		19.4	6.1 – 9.2[*]	7.5[†]
IFG + IGT		9.8	10 – 15[*]	12.1[†]
HbA1c (%)				
<5.0	Nondiabetic middle-aged adults from four US communities	8.6 [9]	6[◊] [9]	0.5[>] [9]
5.0-<5.5		44.6	12[◊]	1[>]
5.5-<6.0		33.2	21[◊]	1.9[>]
6.0-<6.5%		9.3	44[◊]	4.5[>]
≥6.5		4.3	79[◊]	16.5[>]
Metabolic Syndrome				
No	US adults aged ≥ 50 years	56.5 [10]	4.1 [10]	1
Yes		43.5	34.0	8.3

[†]values reported are prevalences unless otherwise indicated, [#]unless referenced, values are calculated from the values depicted in the column "Diabetes risk" [*]annualized incidence of diabetes, [†]annualized relative risk, [◊]cumulative 15-year incidence of diagnosed diabetes, [>]multivariable adjusted hazard ratio of 15-year risk for each absolute increase in 1 percentage point of glycated hemoglobin

BMI, body mass index; HbA1c, glycated hemoglobin; IFG, impaired fasting glucose; IGT, impaired glucose tolerance; NA, not available; RR, relative risk

a random "patient" is higher than that of a random "nonpatient." When predicted risks of individuals who will develop the disease are always higher than the risks of those who will not develop the disease, the AUC is 1.0. When their risks are higher for 50% of the random pairs, the AUC is 0.50, equaling the predictive performance of tossing a coin[20].

The AUC was 0.65 in men and 0.66 in women for the FINDRISC score predicting impaired fasting glucose, impaired glucose tolerance, or undiagnosed diabetes, and 0.72 and 0.75 for detecting metabolic syndrome[18]. The AUC of the Diabetes Risk Calculator was 0.70 for detecting impaired fasting glucose, impaired glucose tolerance, or undiagnosed diabetes[19]. These modest AUC values indicate that many people who will develop T2D are not identified as being at increased risk by these risk scores, and that many that will not develop the disease are labeled as increased risk. Although offering lifestyle modification programs to individuals who will not develop T2D may do no harm and may even provide other benefits by reducing the risk of other diseases, not recognizing the many who will develop diabetes would clearly be missed opportunities to reduce the serious burden of disease[12]. Some clinical risk models that include invasive measurements showed higher AUC values for detecting individuals who will develop T2D. An example is the Framingham Risk Score including age, sex, obesity, hypertension, parental history of diabetes, low levels of high-density lipoprotein cholesterol, elevated triglyceride levels, and impaired fasting glucose[21]. The AUC of this risk model was 0.85 for predicting T2D in middle-aged adults[21]. However, inclusion of invasive measurements that can change over time in clinical risk models might be inconvenient at the population level and these models still leave room for improvement.

Recent studies have investigated the predictive ability of risk models that include genetic variants only or genetic variants added to clinical risk factors. A study that investigated a genetic risk score based on 34 diabetes-associated variants showed a significant association of the risk score with risk of developing diabetes[22·]. This risk was attenuated by lifestyle interventions, also in individuals in the highest genetic risk quartile, suggesting that detecting individuals at high risk of developing T2D based on genetic variants and offering them lifestyle modification programs is useful. In this paper, we review genetic risk prediction studies from a methodologic perspective by focusing on factors in the choice of study design and population that may have impacted the observed predictive ability.

GENETIC RISK PREDICTION STUDIES

The number of studies that investigate the predictive ability of genetic variants in T2D has increased rapidly (Table 2; [23,24·,25–38,39·,40–42,43·]). These studies assessed risk models that were based on genetic variants only or on a combination of both genetic and nongenetic variants. The table shows that the number of SNPs included in the genetic models has increased from 3 in 2005 to 40 in 2011. The models show considerable overlap in the genetic variants that were considered, but there also are many differences. Since its discovery, all but one of the studies had included *TCF7L2* and the majority addition-

Table 2. Genetic risk prediction studies in type 2 diabetes

Study reference	No of poly-morphisms	Clinical risk factors	AUC genetic	AUC clinical	AUC combined	Design	Age (mean, years)*	Sex (% men)*	BMI (mean, kg/m²)*
European									
Balkau et al. [23]									
Men	2	FPG, smoking status, WC, GGT	NA	0.85	0.85	Prospective cohort	50 / 47	100 / 100	27.5 / 25.1
Women	2	FPG, BMI, FH, TG	NA	0.92	0.91	Prospective cohort	52 / 47	0 / 0	29.2 / 23.7
Lyssenko et al. [24•, 25]	3	BMI, FPG	NA	0.68	0.68	Prospective cohort	45.1#	51 / 46	25.3#
Weedon et al. [26]	3	NA	0.58	NA	NA	Case-control	48.7 / 31.8	58 / 50	31.4/ 27.2
Vaxillaire et al. [27]	3	Age, sex, BMI	0.56	0.82	0.83	Prospective cohort	47.7#	50	24.3#
Cornelis et al. [28]	10	Age, sex, BMI, FH, smoking, alcohol intake, PA	NA	0.78	0.79	Nested case-control	49.0 / 48.1	43 / 38	27.7 / 24.4
Lyssenko et al. [29]									
Malmo	11	Age, sex, BMI, FH, BP, TG, FPG	0.63	0.74	0.75	Prospective cohort	45.5	64.9	24.3
Botnia	11	Age, sex, BMI, FH, BP, TG, FPG, HDL, WC	0.68	0.79	0.80	Prospective cohort	44.9	45.5	25.6
Cauchi et al. [30]	15	Age, sex, BMI	NA	NA	0.86	Case-control	62.9 / 54.7	62 / 42	29.0 / 24.7
Lin et al. [31]	15	Age, sex, FH, PA, WHR, triacylglycerol/ HDL ratio	0.59	0.86	0.87	Cross-sectional	60.7 / 52.8	67 / 46	30.4 / 25.5
Fontaine-Bisson et al. [32]	17	Age, sex	NA	NA	0.59	Cross-sectional	53.6 / 53.1	58.4 / 50.2	29.5 / 25.8
van Hoek et al. [33]	18	Age, sex, BMI	0.60	0.66	0.68	Prospective cohort	68.2 / 69.0	44 / 40	28.0 / 26.0
Lango et al. [34]	18	Age, sex, BMI	0.60	0.78	0.80	Case-control	55.7 / NA	56/51	31.5 / 26.9
Meigs et al. [35]	18	Age, sex, BMI, FH, FPG, SBP, HDL, TG	0.58♦	0.90	0.90	Prospective cohort	42.1	47	25.6
Sparso et al. [36]	19	Age, sex, BMI	0.60	0.92	0.93	Case-control	60 / 47	59.3 / 46.3	30.6 / 25.6

Table 2. (continued)

Study reference	No of poly-morphisms	Clinical risk factors	AUC genetic	AUC clinical	AUC combined	Design	Age (mean, years)*	Sex (% men)*	BMI (mean, kg/m²)*
Wang et al. [37]									
FINDRISC	19	Age, BMI, WC, PA, FH, diet, antihypertensive medication, previously known high glucose, FINDRISC, TG, HDL, adiponectin, ALT	0.55	0.73	0.73	Cross-sectional	45-74†	100 / 100	NA
FIND-RISC+	19		0.55	0.77	0.77	Cross-sectional	45-74†	100 / 100	NA
Schulze et al. [38]	20	Age, WC, height, history of HT, PA, smoking, consumption of red meat, whole-grain bread, coffee and alcohol, glucose, HbA1c, TG, HDL, GGT, ALT, hs-CRP	NA	0.90	0.90	Prospective case-cohort	54.6 / 49.4	58.7 / 36.9	30.4 / 25.9
Talmud et al. [39-]									
Cambridge score	20	Age, sex, BMI, drug treatment, FH, smoking status	0.55	0.72	0.73	Prospective cohort	51.0 / 49.0	72.9 / 72.8	27.5 / 24.7
Framingham offspring score	20	Age, sex, BMI, parental history of T2D, HDL, TG, FPG	0.55	0.78	0.78	Prospective cohort	51.0 / 49.0	72.9 / 72.8	27.5 / 24.7
de Miguel-Yanes et al. [40]	40	Age, sex, FH, BMI, FPG, SBP, HDL, TG	0.61♦	0.90	0.91	Prospective cohort	46	47	26.0
Asian									
Miyake et al. [41]	11	Age, sex, BMI	0.63	0.68	0.72	Case-control	61.3 / 67.5	58 / 46	23.6 / 23.3
Hu et al. [42]	11	Age, sex, BMI	0.62	0.61	0.67	Case-control	61.2 / 57.4	52 / 41	24.0 / 23.6

Table adapted from Mihaescu et al. [43-]

*Values provided are for participants with and without T2D respectively when two values are reported and for the total population when one value is reported. For prospective cohort studies, descriptive data from baseline examinations are given. Values are means unless otherwise indicated. #median, ♦adjusted for sex, †range

ALT, alanine aminotransferase; BMI, body mass index; BP, blood pressure; FH, family history of T2D; FINDRISC, Finnish Diabetes risk score; FPG, fasting plasma glucose; GGT, γ-glutamyltransferase; HDL, high-density lipoprotein cholesterol; hs-CRP, high sensitivity c-reactive protein; HT, hypertension; NA, not available; PA, physical activity; SBP, systolic blood pressure; TG, triglycerides; WC, waist circumference; WHR, waist-hip ratio.

ally investigated *PPARG*, *CDKN2A/B*, *KCNJ11*, *IGF2BP2*, *SLC30A8*, and *HHEX-IDE-KIF11*. Yet, most other SNPs were included in one or two models only[43*]. The same was observed for the clinical models. Most clinical models included at least age, sex, and BMI, but they differed in the other factors that were added, such as blood pressure, family history of T2D, and fasting plasma glucose level.

Table 2 shows that, almost without exception, the genetic risk models had lower AUC values than the clinical models. The AUC values for the genetic models ranged from 0.55 to 0.68 and for the clinical models from 0.61 to 0.92. Table 2 also shows that the addition of genetic factors either did not or only marginally improved the AUC beyond that of the clinical risk models.

PREDICTIVE ABILITY OF CLINICAL RISK MODELS

The differences in the predictive ability of clinical risk models are explained by how many and which risk factors are included in the model and by differences in study design and study population. This is nicely illustrated by three studies that had investigated largely the same 18 genetic variants. The AUCs of the genetic risks models in these studies were similar (0.58–0.60), but the AUCs of the clinical models were 0.66, 0.78, and 0.90[33–35]. The clinical models with AUC values of 0.66 and 0.78 included age, sex, and BMI, but the model with an AUC value of 0.90 also included T2D family history, fasting plasma glucose, systolic blood pressure, high-density lipoprotein cholesterol, and triglycerides. The excellent predictive ability was likely due to the inclusion of fasting plasma glucose, as individuals with impaired fasting glucose have a very high risk of developing T2D (Table 1). Table 2 shows that AUC values tend to be higher when more risk factors are included in the model, particularly when fasting plasma glucose was included.

Yet, also the two studies that both investigated age, sex, and BMI in the clinical model had markedly different AUC values (0.66 and 0.78). The difference in these AUC values was likely explained by differences in the study design and population. The AUC of 0.66 was obtained in a prospective cohort study, the Rotterdam Study, and the AUC of 0.78 in a case–control study, consisting of case and control subjects from the GoDARTS (Genetics of Diabetes Audit and Research Tayside Study). Participants in the Rotterdam Study were older and less often men (Table 2), but the two populations predominantly differed in BMI. The mean BMI of the cases in the GoDARTS study was higher than the mean BMI of cases in the Rotterdam Study (31.5 vs 28.0 kg/m2). Also, the difference in mean BMI between cases and controls was much larger in the GoDARTS study compared with the Rotterdam Study (4.6 vs 2.0 kg/m^2). In general and by definition, the predictive ability of risk models is higher when there are larger differences between cases and controls on the risk factors included in the risk model. Along the same lines, study design and

population characteristics may have influenced the observed AUC values of the other clinical models, and also of AUC values of the genetic risk models.

PREDICTIVE ABILITY OF GENETIC RISK MODELS

The AUC values of the genetic risk models ranged from 0.55 to 0.68, a range that was much smaller than that of the clinical models. Similar as for the clinical risk models and given that all SNPs approximately have the same low effect size, one would expect better predictive ability for models that included a higher number of SNPs, but Figure 1 shows that this was not observed for the studies listed in Table 2. The differences in the AUC values of the genetic risk scores cannot be explained by the number of polymorphisms included in the risk models. In fact, the highest genetic AUC (0.68) was found for a model that included 11 SNPs, and the lowest for a model that included these exact 11 SNPs plus an additional 8 others. The explanation for the absence of this relationship is likely in the low effect sizes of the genetic variants. A higher number of SNPs only yields a slightly higher AUC, a combined effect that could easily be outweighed by the influence of other factors, such as study design and study population.

Genetic risk prediction models have been investigated in prospective cohort studies, in case–control studies and in cross-sectional studies, and in study populations that differed in age, sex, and BMI (Table 2). These methodologic aspects may have impacted the observed AUC values in a similar way as they impact the AUC values of the clinical models. First, clinical and demographic characteristics of the study population may
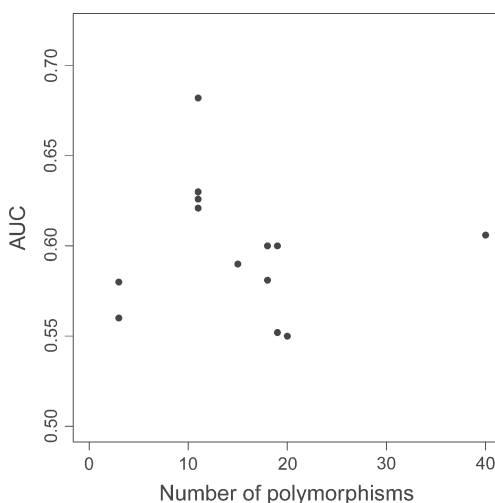

Figure 1. The area under the receiver operating characteristic curve (AUC) versus the number of single nucleotide polymorphisms included in the genetic risk models

have influenced the observed predictive ability of the genetic risk models. There are two ways in which these characteristics may impact the predictive ability: the clinical and demographic characteristics of the study population itself and the differences in these characteristics between patients and nonpatients.

Table 2 describes mean age and BMI and the percentage of men in published genetic risk prediction studies for T2D. Mean age varied from 42.1 to 68.9 years, mean BMI from 23.4 to 29.1 kg/m^2, and the percentage of men from 0% to 100%. It is often hypothesized that genetic risk factors may be more predictive in populations in which nongenetic T2D risk factors are not yet present (eg, in younger or normal weight cohorts), but AUC values of the genetic models were not markedly higher when populations were younger, had lower BMI, or had a lower percentage of men. However, because of the heterogeneity between the studies and their relatively small number, conclusions must be drawn with caution. Moreover, one study that had investigated the predictive performance in two age categories (< 50 years vs ≥ 50 years) did find higher AUC values for the genetic risk score in younger people (AUC 0.66 vs 0.59)[40]. The observation that a stratified analysis within a single study did show differences in predictive ability suggests that the absence of a clear relation of age, BMI, and sex with AUC values across studies is likely explained by the presence of other differences between the studies.

The other way in which clinical and demographic characteristics of the study population impact the predictive ability of risk models is through differences in these characteristics between patients and nonpatients. This specifically holds for characteristics that are included as risk factors in the prediction models, and for characteristics that are associated with these risk factors. Evidently and by definition, the presence of risk factors will differ between patients and nonpatients, but the difference can also be enlarged as a result of selection procedures. For example, patients who are recruited through hospitals may have more unfavorable risk profiles than patients randomly selected from the total patient population. Consequently, differences in risk factors between hospital-based cases and population-based controls will be larger and the impact of these risk factors on the predictive ability higher. For the studies listed in Table 2, differences in mean age ranged from −6.2 to 16.9 years, in mean BMI from 0.3 to 5.5 kg/m^2, and differences in the percentage of men from −0.1% to 21.8%. Figure 2 shows that larger differences in mean age and BMI between patients and nonpatients were associated with higher AUC values for the clinical risk models, and, although less apparent, lower AUC values for the genetic models. No relation was observed between clinical AUC values and the percentage of men included in the studies, but this may be because male sex only marginally increases T2D risk compared with age and BMI (Table 1).

A second methodological aspect that may impact the predictive ability of risk models is study design. Genetic risk prediction studies are preferably conducted in prospective follow-up studies, but cross-sectional and case–control studies have been used as well
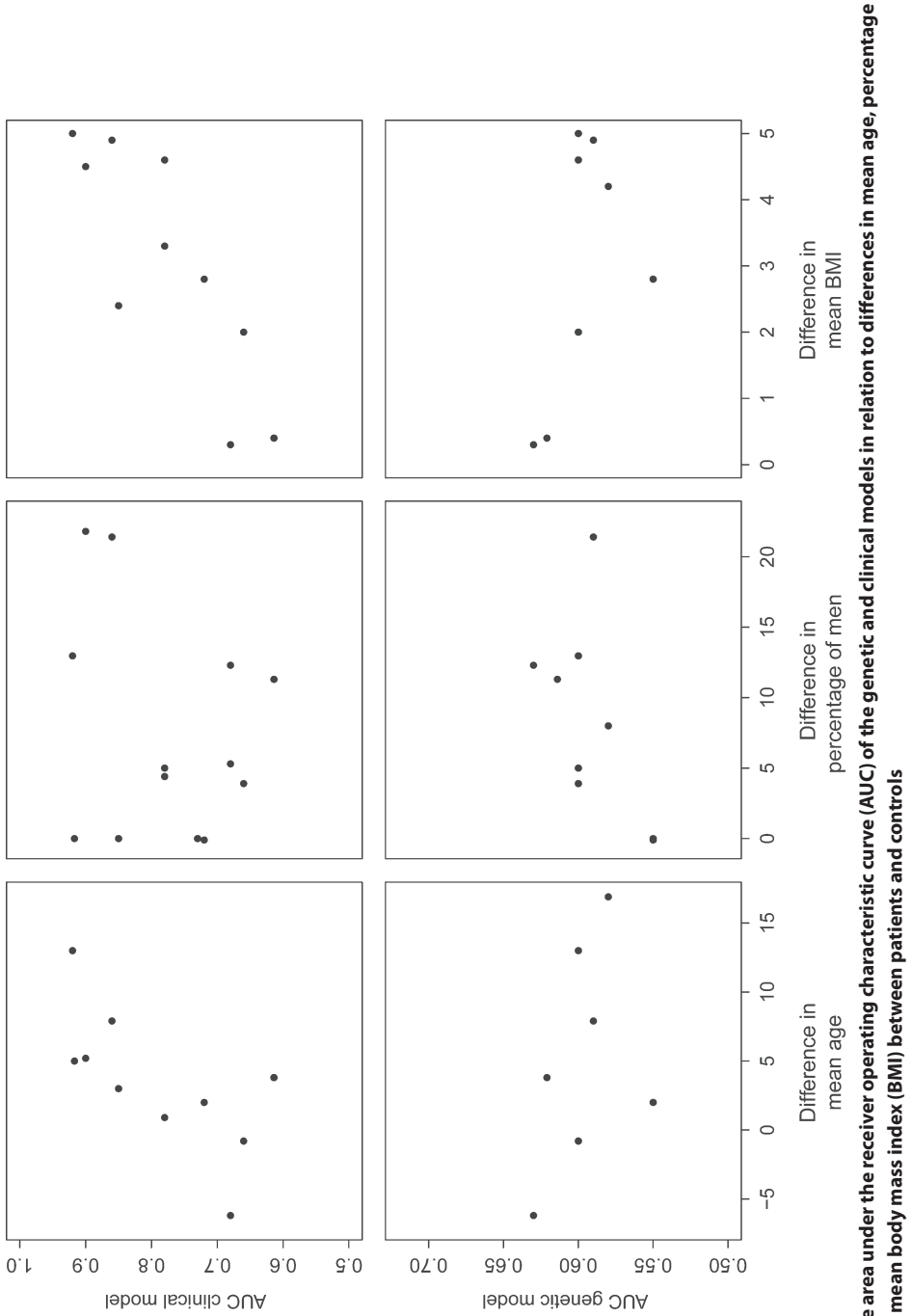
Figure 2. The area under the receiver operating characteristic curve (AUC) of the genetic and clinical models in relation to differences in mean age, percentage of men, and mean body mass index (BMI) between patients and controls

(Table 2). The impact of study design on AUC values of T2D risk prediction models is in part related to the impact of population characteristics. Selection procedures for cases and controls may affect differences in clinical and demographic characteristics between patients and nonpatients. Case–control studies may demonstrate AUC values that deviate from those observed in prospective cohort and cross-sectional studies when cases and controls are recruited from different sources.

Another way in which study design may impact the predictive ability of risk models is length of follow-up in prospective cohort studies. Longer follow-up increases the likelihood that clinical T2D risk factors change over time, and that as a result their baseline values will be less predictive for the development of disease, resulting in prediction models with lower AUC. The length of follow-up of the studies listed in Table 2 varied from 6 to 25 years. Again, the number of prospective cohort studies was too small to investigate the impact of follow-up duration, but one study investigated the predictive ability in quintiles of follow-up time. This study demonstrated that the AUC of the clinical risk model decreased with increasing duration of follow-up, whereas the AUC of the genetic risk model increased[29]. From the first to the fifth quintile, the clinical AUC value decreased from 0.75 to 0.67 and the genetic AUC value increased from 0.57 to 0.62[29].

CONCLUSIONS

In this review, we showed that study design and population characteristics may have affected the observed predictive performance of risk models. AUC values of the clinical risk models were higher and, although weaker, AUC values of the genetic risk models were lower when there were larger differences in age and BMI between cases and controls. This observation has important implications for the design and health care relevance of genetic risk prediction studies.

The predictive ability of risk models is preferably investigated in prospective cohort studies, but in practice often only case-control or cross-sectional designs are available. Because clinical risk factors, particularly the difference in risk factors between cases and controls, impact AUC values, it is expected that AUC values for genetic risk models obtained in case–control or cross-sectional studies may be valid when the distribution of these risk factors does not differ from prospective studies. For case–control studies, this means that the selection of cases and controls is not affected by these risk factors. In case of selection, transparency about the methods is important to enable a correct interpretation of the scientific and health care relevance of the results. For this reason, the GRIPS (Genetic Risk Prediction Studies) statement, a recently published guideline for the reporting of genetic risk prediction studies, recommends to describe eligibility criteria for participants, and sources and methods of selection of participants[44].

The observed impact of population characteristics implies that it is important to assess the predictive ability of risk scores in representative samples of the population in which the model is ultimately applied to get valid estimates of their performance in that population. The question then is: which populations do we want to target for the prevention of T2D? Evidently, these may include individuals with metabolic syndrome or overweight, but for genetic prediction this may particularly concern young individuals who have not developed clinical risk factors. To date none of the T2D risk prediction studies have been conducted in younger populations; all studies were conducted in populations who on average were older than 40 years of age, two even in populations over 60 years of age[33,41]. The study that best approximates the desired study population has been conducted in a population with a mean age of 42 years, a mean BMI of 25.6 kg/m^2, and an almost equal number of men and women[35]. Given the observed differences in AUC values, we must conclude that we do not know whether genetic variants are useful in predicting T2D risk in younger populations. None of the studies so far has started from a health care perspective when investigating the predictive ability of T2D risk models.

There is increasing interest in investigating the value of genetic risk factors in the prediction of T2D risk. In this review, we demonstrated that the choice of study design and predominantly the choice of study population impact the observed predictive ability of risk models. For this reason it is important that the planning of future genetic risk prediction studies in T2D starts from a health care perspective by asking in which population we want to predict T2D risk. It is the answer to this question that determines the population in which the predictive ability should be assessed and that determines whether the results of the study ultimately can be informative and change health care practice.

REFERENCES

Papers of particular interest, have been highlighted as: • Of importance

1. U.S. Census Bureau: General demographic characteristics, 2009 population estimates, United States. http://factfinder.census.gov/servlet/QTTable?-ds_name=PEP_2009_EST&-qr_name=PEP_2009_EST_DP1&-geo_id=01000US. (June 2011).

2. Centers for Disease Control and Prevention: Percentage of civilian, noninstitutionalized population with diagnosed diabetes, by age, United States, 1980–2009. http://www.cdc.gov/diabetes/statistics/prev/national/figbyage.htm. (June 2011).

3. Centers for Disease Control and Prevention: Age-adjusted percentage of civilian noninstitutionalized population with diagnosed diabetes, by sex, United States, 1980–2009. http://www.cdc.gov/diabetes/statstics/prev/national/figbysex.htm. (June 2011).

4. Centers for Disease Control and Prevention: Prevalence of overweight, obesity, and extreme obesity among adults: United States, trends 1976–1980 through 2007–2008. http://www.cdc.gov/nchs/data/hestat/obesity_adult_07_08/obesity_adult_07_08.pdf. (June 2011).

5. Flegal, K.M., Carroll, M.D., Ogden, C.L. et al. Prevalence and trends in obesity among US adults, 1999–2008. JAMA 303, 235–41 (2010).

6. Nguyen, N.T., Nguyen, X.T., Lane, J. et al. Relationship between obesity and diabetes in a US adult population: findings from the National Health and Nutrition Examinaton Survey, 1999–2006. Obes Surg 21, 351–5 (2011).

7. Karve, A., Hayward, R.A. Prevalence, diagnosis, and treatment of impaired fasting glucose and impaired glucose tolerance in nondiabetic U.S. adults. Diabetes Care 33, 2355–9 (2010).

8. Gerstein, H.C., Santaguida, P., Raina, P. et al. Annual incidence and relative risk of diabetes in people with various categories of dysglycemia: a systematic overview and meta-analysis of prospective studies. Diabetes Res Clin Pract 78, 305–12 (2007).

9. Selvin, E., Steffes, M.W., Zhu, H. et al. Glycated hemoglobin, diabetes, and cardiovascular risk in nondiabetic adults. N Engl J Med 362, 800–11 (2010).

10. Alexander, C.M., Landsman, P.B., Teutsch, S.M. et al. NCEP-defined metabolic syndrome, diabetes, and prevalence of coronary heart disease among NHANES III participants age 50 years and older. Diabetes 52, 1210–4 (2003).

11. Gillies, C.L., Abrams, K.R., Lambert, P.C. et al. Pharmacological and lifestyle interventions to prevent or delay type 2 diabetes in people with impaired glucose tolerance: systematic review and meta-analysis. BMJ 334, 299 (2007).

12. Colagiuri, S. Epidemiology of prediabetes. Med Clin North Am 95, 299–307 (2011).

13. McCarthy, M.I. From hype to hope? A journey through the genetics of type 2 diabetes. Diabet Med 28, 132–40 (2011).

14. McCarthy, M.I. Genomics, type 2 diabetes, and obesity. N Engl J Med 363, 2339–50 (2010).

15. International Diabetes Federation: Diabetes and impaired glucose tolerance. Global burden: prevalence and projections, 2010 and 2030. http://www.idf.org/diabetesatlas/diabetes-and-impaired-glucose-tolerance. (June 2011).

16. Stumvoll M, Goldstein BJ, van Haeften TW. Type 2 diabetes: principles of pathogenesis and therapy. Lancet. 2005;365:1333–46.

17. Paulweber, B., Valensi, P., Lindström, J. et al. A European evidence-based guideline for the prevention of type 2 diabetes. Horm Metab Res 42, Suppl 1:S3–36 (2010).

18. Saaristo, T., Peltonen, M., Lindström, J. et al. Cross-sectional evaluation of the Finnish Diabetes Risk Score: a tool to identify undetected type 2 diabetes, abnormal glucose tolerance and metabolic syndrome. Diab Vasc Dis Res 2, 67–72 (2005).

19. Heikes, K.E., Eddy, D.M., Arondekar, B., et al. Diabetes Risk Calculator: a simple tool for detecting undiagnosed diabetes and prediabetes. Diabetes Care 31, 1040–5 (2008).

20. Janssens, A.C.J.W., Khoury, M.J. Assessment of improved prediction beyond traditional risk factors – when does a difference make a difference? Circ Cardiovasc Genet 3, 3–5 (2010).

21. Wilson, P.W.F., Meigs, J.B., Sullivan, L. et al. Prediction of incident diabetes mellitus in middle-aged adults. Arch Intern Med 167, 1068–74 (2007).

22. • Hivert, M.F., Jablonski, K.A., Perreault, L. et al. Updated genetic score based on 34 confirmed type 2 diabetes loci is associated with diabetes incidence and regression to normoglycemia in the Diabetes Prevention Program. Diabetes 60, 1340–8 (2011). *This study investigated effectiveness of preventive interventions in individuals in all quartiles of genetic risk estimated by a genetic risk score. This is very interesting from a public health perspective.*

23. Balkau, B., Lange, C., Fezeu, L. et al. Predicting diabetes: clinical, biological, and genetic approaches: data from the Epidemiological Study on the Insulin Resistance Syndrome (DESIR). Diabetes Care 31, 2056–61 (2008).

24. • Lyssenko, V., Almgren, P., Anevski, D. et al. Genetic prediction of future type 2 diabetes. PLoS Med 2:e345 (2005). *This was the first paper in T2D that considered the combined effect of multiple susceptibility variants in the prediction of disease risk.*

25. Janssens, A.C.J.W., Gwinn, M., Subramonia-Iyer, S. et al. Does genetic testing really improve the prediction of future type 2 diabetes? PLoS Med 3:e114 (2006).

26. Weedon, M.N., McCarthy, M.I., Hitman, G. et al. Combining information from common type 2 diabetes risk polymorphisms improves disease prediction. PLoS Med 3:e374 (2006).

27. Vaxillaire, M., Veslot, J., Dina, C. et al. Impact of common type 2 diabetes risk polymorphisms in the DESIR prospective study. Diabetes 57, 244–54 (2008).

28. Cornelis, M.C., Qi, L., Zhang, C. et al. Joint effects of common genetic variants on the risk for type 2 diabetes in U.S. men and women of European ancestry. Ann Intern Med 150, 541–50 (2009).

29. Lyssenko, V., Jonsson, A., Almgren, P. et al. Clinical risk factors, DNA variants, and the development of type 2 diabetes. N Engl J Med 359, 2220–32 (2008).

30. Cauchi, S., Meyre, D., Durand, E. et al. Post genome-wide association studies of novel genes associated with type 2 diabetes show gene-gene interaction and high predictive value. PLoS One 3: e2031 (2008).

31. Lin, X., Song, K., Lim, N. et al. Risk prediction of prevalent diabetes in a Swiss population using a weighted genetic score–the CoLaus Study. Diabetologia 52, 600–8 (2009).

32. Fontaine-Bisson, B., Renström, F., Rolandsson, O. et al. Evaluating the discriminative power of multi-trait genetic risk scores for type 2 diabetes in a northern Swedish population. Diabetologia 53, 2155–62 (2010).

33. van Hoek, M., Dehghan, A., Witteman, J.C.M. et al. Predicting type 2 diabetes based on polymorphisms from genome-wide association studies: a population-based study. Diabetes 57, 3122–8 (2008).

34. Lango, H., Consortium UKTDG, Palmer, C.N.A. et al. Assessing the combined impact of 18 common genetic variants of modest effect sizes on type 2 diabetes risk. Diabetes 57, 3129–35 (2008).

35. Meigs, J.B., Shrader, P., Sullivan, L.M. et al. Genotype score in addition to common risk factors for prediction of type 2 diabetes. N Engl J Med 359, 2208–19 (2008).

36. Sparso, T., Grarup, N., Andreasen, C. et al. Combined analysis of 19 common validated type 2 diabetes susceptibility gene variants shows moderate discriminative value and no evidence of gene-gene interaction. Diabetologia 52, 1308–14 (2009).

37. Wang, J., Stancakova, A., Kuusisto, J. et al. Identification of undiagnosed type 2 diabetic individuals by the finnish diabetes risk score and biochemical and genetic markers: a population-based study of 7232 Finnish men. J Clin Endocrinol Metab 95, 3858–62 (2010).

38. Schulze, M.B., Weikert, C., Pischon, T. et al. Use of multiple metabolic and genetic markers to improve the prediction of type 2 diabetes: the EPIC-Potsdam Study. Diabetes Care 32, 2116–9 (2009).

39. • Talmud, P.J., Hingorani, A.D., Cooper, J.A. et al. Utility of genetic and non-genetic risk factors in prediction of type 2 diabetes: Whitehall II prospective cohort study. BMJ 340:b4838 (2010). *This study is a very good example of a genetic risk prediction study, with a very complete description of the study population and a thorough assessment of the predictive ability.*

40. de Miguel-Yanes, J.M., Shrader, P., Pencina, M.J. et al. Genetic risk reclassification for type 2 diabetes by age below or above 50 years using 40 type 2 diabetes risk single nucleotide polymorphisms. Diabetes Care 34, 121–5 (2011).

41. Miyake, K., Yang, W., Hara, K., et al. Construction of a prediction model for type 2 diabetes mellitus in the Japanese population based on 11 genes with strong evidence of the association. J Hum Genet 54, 236–41 (2009).

42. Hu, C., Zhang, R., Wang, C. et al. PPARG, KCNJ11, CDKAL1, CDKN2A-CDKN2B, IDE-KIF11-HHEX, IGF2BP2 and SLC30A8 are associated with type 2 diabetes in a Chinese population. PLoS One 4: e7643 (2009).

43. • Mihaescu, R., Meigs, J., Sijbrands, E. et al. Genetic risk profiling for type 2 diabetes. PloS Curr 3: RRN1208 (2011). *This paper is regularly updated and gives an overview of all genetic risk prediction studies conducted in T2D.*

44. • Janssens, A.C., Ioannidis, J.P., van Duijn, C.M. et al. Strengthening the reporting of genetic risk prediction studies: the GRIPS statement. Eur J Epidemiol 26, 255–9 (2011). *This paper describes reporting guidelines to enhance the transparency, quality, and completeness of the reporting of genetic risk prediction studies. A longer paper with examples and explanations was published simultaneously*

TGACTACGTA
CTACTAGCACTGTACACGA   GCTACGTACG
GTACGTAGTCTGATCAAAACGTA   GCTATACAGCTA
CGTTGGTACCAGTACGTACGACTGCGA GTACGTACGTACG
GTGTGTCACACAGACACGTGATATAGACCAGATGACACACGTAC
GTACGACGGCGCGGCGCGATATATATAAAAGCACGGACACTACGA
ACGATGACTGACTACGTGCGCGCGTAGCTAGCTACGGATAGGACTA
ACTGACTACGTACTACTAAAGCTGCTTTTACGTACGTACGGGCATC
CATGTACTATCGCGGCGCGTAGACTGTACGACCAGACTGACTGACT
GACTGACTGACGCGCGTACGTAGCTACTGTAGTACGTACGTACGTA
TTCAGACTGACTGACTGAGATTACAGCTACGTAGTACTGACACGT
TACGACGGCGCGGCGCGATATATATAAAAGCACGGACACTACGA
GATGACTGACTACGTGCGCGCGTAGCTAGCTACGGATAGGACTA
TGACTACGTACTACTAAAGCTGCTTTTACGTACGTACGGGCATC
CTGCGCATGTACTATCGCGGCGCGTAGACTGTACGACCAGAC
ACTGACGCGCGTACGTAGCTACTGTAGTACGTACGTACGTA
ACTGACTGACTGAGATTACAGCTACGTAGTACTGACACGT
GCATACGAGTCGAAGCGCGCAATATATATTATATCGGCC
GTGACTACATCATGCTGACTACGGTCGCGCTCAACG
TGACTACGTACGTATACGACGTACTGACGGCGCGC
GCACTGTACACGACTGACTTACTAGCTACGTACC
TGATCAAAACGTACGCGCGGCTATACAGCTA
CGTACGACTGCGATACGTACGTACGTACG
CGTGATATAGACCAGATGACACACGT
TATATATAAAAGCACGGACACT
GTAGCTAGCTACGGATA
CTTTTACGTACGT
ACGACCAG
TAGT

2.2

Association of the *IGF1* gene with fasting insulin levels

ABSTRACT

Objective: Insulin-like growth factor-1 (IGF-I) has been inversely associated with insulin resistance. Genome-wide association studies (GWAS) of fasting insulin (FI) identified single nucleotide variants (SNVs) near the *IGF1* gene, raising the hypotheses that associations of SNVs near *IGF1* with FI are mediated by IGF-I levels and that these non-coding GWAS variants either tag other functional variants in the *IGF1* region or are directly functional.

Methods: To test the first hypothesis, we performed mediation analyses using imputed genotyping array data in 5,141 non-diabetic individuals from three population-based cohort studies (CHS n=1,717; FHS n=3,293; RS n=140) with circulating IGF-I and FI levels available and to test the second hypothesis we performed single variant analyses and the Sequence Kernel Association Test (SKAT) using targeted sequencing data around *IGF1* in 3,539 non-diabetic individuals (ARIC n=1,761; CHS n=967; FHS n=811) that were part of the CHARGE Targeted Sequencing Study. In addition, we examined regulatory annotation using ENCODE data to generate hypotheses about a direct functional impact of non-coding FI associated GWAS variants.

Results: Mediation analyses suggest that GWAS associations of SNVs near *IGF1* with FI were not mediated by IGF-I levels. Targeted sequence data reveal a large number of novel rare variants at the *IGF1* locus. SKAT analyses show a significant FI association with a subset of rare nonsynonymous variants ($P = 5.7 \times 10^{-4}$). Conditioning on the GWAS variants suggested that the GWAS signal explains part of the rare variant signal and the presence of a residual independent rare variant effect ($P_{conditional} = 0.019$). Annotation of nearby non-coding genomic functional and regulatory elements suggest that the GWAS variants may have a direct functional role in insulin biology.

Conclusion: Our analyses suggest that association of SNVs near *IGF1* with FI is not mediated by circulating IGF-I levels. Our study provides insight into variation present at the IGF1 locus and into the genetic architecture underlying FI levels, suggesting a role for both rare non-synonymous and common functional variants in insulin biology.

INTRODUCTION

The *IGF1* gene encodes insulin-like growth factor-1 (IGF-I). This hormone has many biological functions involving cell growth, proliferation and apoptosis[1]. Circulating IGF-I concentrations have been associated with several human diseases, including cardiovascular mortality and cardiovascular risk factors such as age, body mass index, total cholesterol, the presence of diabetes, glomerular filtration rate, and alcohol consumption[2,3]. IGF-I levels are inversely correlated with insulin resistance[3], which may be explained by the insulin-like effects of IGF-I on glucose-uptake. IGF-I is structurally comparable to insulin and they both cross react with the other's receptor.

Genome wide association studies (GWAS) of fasting insulin (FI) levels revealed single nucleotide variants (SNVs) near the *IGF1* gene[4,5]. SNV rs35767 located 1.2 kb upstream of *IGF1*, was associated with a 0.010 pmol/L per (G) allele increase in FI level ($P = 3.3 \times 10^{-8}$) in a large GWAS meta-analysis[4]. Another large GWAS meta-analysis, in largely overlapping samples, revealed rs2114912 as the variant most strongly associated with FI in the IGF1 region[5]. This variant is located 54.7 kb upstream of the *IGF1* gene and is associated with a 0.024 pmol/L increase in FI per copy of the T allele. These findings have inspired further assessment of the role that the *IGF1* gene plays in insulin biology.

In this paper we hypothesize that the associations of SNVs near the *IGF1* gene with FI (hence insulin resistance) are mediated by circulating IGF-I levels, and that the GWAS variants tag other common or rare functional variants in the *IGF1* region associated with FI levels. To test the first hypothesis, we performed mediation analyses using imputed genotyping array data and to test the second hypothesis we performed association analyses using deep, high throughput next generation targeted sequencing data around *IGF1*. We also examined ENCODE Consortium datasets[6] of regulatory elements by viewing the *IGF1* region in the UCSC Genome Browser[7] in order to generate testable hypotheses about direct functional roles and mechanisms of the non-coding FI associated GWAS variants.

MATERIALS AND METHODS

Study Populations

Individuals of European ancestry from four cohorts of the Cohorts for Heart and Aging Research in Genomic Epidemiology (CHARGE) consortium were included in this study: the Atherosclerosis Risk in Communities (ARIC) study, Cardiovascular Health Study (CHS), Framingham Heart Study (FHS) and the Rotterdam Study (RS).[8]

Mediation Cohorts

5,141 non-diabetic individuals of CHS (n=1,717), FHS (n=3,293) and RS (n=140) were available to contribute to mediation analyses. Genotypic data and both FI and circulating IGF-I levels were available on the participants included in these analyses.

Sequencing Cohorts

3,539 non-diabetic individuals (ARIC n=1,761; CHS n=967; FHS n=811) that were part of the CHARGE Targeted Sequencing Study with successful sequencing and measured trait levels were available for analyses of targeted sequence data with the outcome FI. 567 of the CHS and 78 of the FHS participants included in these analyses were also included in the mediation analyses. The design of the CHARGE Targeted Sequencing Study, including the cohort sampling design, has been described in detail in Lumley et al[9] and Lin et al[10]. To set up the analytic sample a case-cohort design was used in which both a cohort random sample and participants with extreme phenotypes for each of 14 related cardiometabolic traits were included. This included a sample of 200 participants (100 ARIC, 50 CHS, 50 FHS) from the high tail of the FI (≥ 8 hour fast) distribution in individuals without diabetes, defined as either being diagnosed by a physician (ARIC), treated for diabetes or having a fasting glucose (FG) > 7 mmol/L (ARIC, FHS and CHS). Three FHS participants with type 1 diabetes were excluded from selection. Men and women were selected equally from each cohort.

Quantitative Trait Measurement

FI was measured from fasting plasma (FHS) or fasting serum (CHS, ARIC). In FHS, plasma was collected after a ≥ 8 hour overnight fast and FI was measured on frozen specimen using the DPC Coat-A-Count RIA (total immunoreactive insulin) assay (assay sensitivity 1.2 μU/mL). In CHS (≥ 12 hour fast), FI was measured using a competitive RIA (Diagnostic Products Corp., Malvern, PA). In ARIC (≥ 8 hour fast), FI was measured by radioimmunoassay (125Insulin kit; Cambridge Medical Diagnosis, Bilerica, MA) (assay sensitivity 2μU/mL). BMI, a covariate in the models that we analyzed, was measured using standard methods as previously described[5]. In CHS circulating IGF-I levels were measured by ELISA (Immunodiagnostic Systems Ltd , Boldon Business Park, Boldon, Tyne & Wear, England) and in RS by a radioimmunoassay (Medgenix Diagnostics, Brussels, Belgium).

Genotyping in Mediation Cohorts

In CHS, genotyping was performed at the General Clinical Research Center's Phenotyping/Genotyping Laboratory at Cedars-Sinai using the Illumina 370CNV BeadChip system. Genotypes were called using the Illumina BeadStudio software. The following exclusions were applied to identify a final set of 306,655 autosomal SNPs: call rate < 97%, HWE *P*-value < 10^{-5}, > 2 duplicate errors or Mendelian inconsistencies (for reference CEPH trios),

heterozygote frequency = 0, SNP not found in HapMap. Samples were excluded from analysis for sex mismatch, discordance with prior genotyping, or call rate < 95%. Imputation was performed using BIMBAM v0.99 with reference to HapMap CEU using release 22, build 36 using one round of imputations and the default expectation-maximization warm-ups and runs. In the FHS, genotyping was conducted using the Affymetrix 500K SNP arrays supplemented with the MIPS 50K array. Samples with call rate < 97%, excess Mendelian errors (≥ 1000) or average heterozygosity outside of 5 SD of mean (< 5.758% or > 29.958%) were excluded. A subset of 378,163 SNPs with minor allele frequency (MAF) ≥ 1%, call rate ≥ 97%, differential missingness P-value ≥ 10^{-9} and < 100 Mendelian errors were used for imputation based on the haplotypes of the HapMap CEU release 22 using the MACH software. In the Rotterdam Study, genomic DNA was extracted from venous blood samples obtained at baseline. DNA was extracted using the salting out method[11]. Genotyping was performed using 550 and 610K Illumina arrays. Exclusion criteria for individuals were excess autosomal heterozygosity, mismatches between called and phenotypic gender, and outliers identified by an IBS clustering analysis. SNVs were excluded for Hardy-Weinberg equilibrium P-value ≤ 10^{-6} or SNP call rate ≤ 98%. Genotypes with minor allele frequencies > 1% were used to impute about 2.5 million autosomal SNPs using HapMap CEU release 22 as a reference panel. Imputation was performed using MaCH[12]. Imputed genotypes were coded as dosages. These are values between 0 and 2 indicating the estimated number of copies of a given allele for each individual.

Targeted Next Generation Deep Sequencing

Target selection in the CHARGE Targeted Sequencing Study included regions that had been associated with one of 14 cardiometabolic traits by previous GWAS and regions that had been shown to exhibit pleiotropy, and included the *IGF1* gene[10]. Four regions in or near the *IGF1* gene were sequenced at a mean depth of 50X, including 1kb downstream, all 5 exons plus flanking regions, and 5 SNPs upstream that were associated with FI in GWAS[4,5]: rs35767, rs860598, rs855213, rs35747 and rs2114912 (Supplementary Figure 1). A total of 57.5kb per copy of the *IGF1* region was sequenced. Sequencing methods were described in detail in Lin et al.[10]. An extensive quality control (QC) pipeline was implemented, consisting of preliminary QC procedures in the sequencing laboratory followed by a series of variant-level filtering steps. These included the exclusion of variants that mapped more than 100 base pairs from the requested target capture region, exclusion of variants with a Phred-scaled base quality score[13] less than 30, with less than two reads of the alternate alleles, and variants with a depth of coverage of less than 10 total reads. Heterozygote genotypes were removed if their alternate to reference allele ratio was disproportionate (< 0.2 or > 0.8 for one allele). For strand bias, only variants with alternate allele reads obtained from both the positive and negative strands were

kept. Finally, SNPs that had > 20% missingness across all samples, more than 2 observed alleles, or were part of an overly dense SNP cluster (≥ 3 variants in a 10 bp window) were removed. Using only samples from the cohort random sample subjects, SNPs with HWE *P*-value < 1x10^{-5} were filtered. This criterion was not applied in the samples selected based on extreme phenotypes, potentially enriched for rare variants, to prevent filtering out interesting rare variants with a possible role in disease etiology. To validate sequence-based genotypes, cross-validation was performed with data from the Affymetrix Gene Chip 500K Array Set & 50K Human Gene Focused Panel in 1,096 FHS samples. A total of 558 SNPs were shared between the two platforms. After excluding missing genotypes, 98.0% of genotypes were concordant between the two platforms, suggesting high accuracy of the sequenced genotypes.

Variant Classification and Annotation

Variants identified by sequencing of the *IGF1* locus were classified as common if the MAF was ≥ 1% and rare if the MAF was < 1%. Novel variants were those not found in dbSNP, the 1000 Genomes Project or ESP 6500 (Exome Sequencing Project)[14,15]. Variants were annotated using several bioinformatics sources. ANNOVAR[16] was used to determine whether a variant was synonymous, non-synonymous, intergenic, upstream (within 1kb upstream of a transcription start site), downstream (within 1kb downstream of a transcription end site), intronic, in a three prime untranslated region (3'UTR) or in a 5'UTR. Non-coding variants were predicted to be functional if they were predicted to be highly conserved across species using phastCons[17], predicted to lie in transcription factor binding sites extracted from the HMR Conserved Transcription Factor Binding Site track of the UCSC Genome Browser[7], in DNAse I hypersensitive sites or transcription factor binding sites identified by the ENCODE Project[6] or predicted to be functional using the ORegAnno database[18]. In addition to this functional annotation of the variants present in the targeted sequencing data, we examined ENCODE Consortium regulatory element datasets (including DNAseI hypersensitive sites and histone modifications as well as TFBS Chip-seq) and public transcriptome data in the UCSC Genome Browser to determine whether the known common non-coding FI associated GWAS variants might be directly functional.

Follow-up Genotyping in FHS and lookup of Select Rare Variants

To verify the influence of variant rs151098426 on FI levels, the variant was genotyped in 1,745 FHS offspring and 3,372 FHS generation 3 participants with FI levels available that did not overlap with the FHS participants included in the targeted sequencing analyses. Genotyping was performed using TaqMan (ABI PRISM 7700 HT Sequence Detection System, Applied Biosystems, Foster City, California) at the Joslin Diabetes Center Advanced Genomics and Genetics Core. We also did a lookup of the variant in FI exome chip meta-

analysis results from the CHARGE diabetes-glycemia working group, including 38,528 samples.

Statistical Analyses

All analyses were adjusted for age, sex, BMI and study design variables (i.e. clinic site for CHS and ARIC and recruitment cohort for FHS). FI, in pmol/L, was natural log transformed prior to analyses to improve normality.

Mediation Analyses

To test whether association of FI with GWAS variants in the *IGF1* region (rs35767, rs860598, rs855213, rs35747 and rs2114912, pairwise r^2 0.272-1.00 in HapMap2 CEU (see Supplementary Table 1)) is mediated by IGF-I levels, in each cohort (CHS, FHS, RS) two linear regression models per SNP were fitted, assuming an additive allelic effect. In both models, ln(FI) was the outcome variable. Results from the three cohorts were combined using inverse variance weighted fixed effects meta-analysis as implemented in the R package rmeta[19]. In the first model, age, sex and BMI were included as covariates and in the second model IGF-I was added as a covariate. From the models a ratio β_{SNP_model2} / β_{SNP_model1} < 1 would suggest that IGF-I levels explained part of the SNP-FI association.

Analyses of Targeted Sequence Data

The analytic strategy of the targeted sequence data, described briefly below, followed the approach outlined in Lumley et al.[9] and Lin et al.[10].

Four subsets based on functional annotation of rare variants within the *IGF1* locus were tested for association with ln(FI) using the Sequence Kernel Association Test (SKAT)[20]. The subsets included 1) nonsynonymous variants, 2) novel nonsynonymous variants, 3) noncoding variants that were predicted to be functional and 4) novel noncoding variants that were predicted to be functional. FHS used a SKAT test that accounted for family structure[21]. SKAT tests were conducted within the three cohorts (CHS, FHS, ARIC) and meta-analyzed using a weighted sum of squares of z-statistics from single-variant score tests. These variant scores were squared, weighted based on combined allele frequencies across all studies, and summed to create a Q statistic. The significance of the Q statistics was determined using an asymptotic distribution, as described in Wu et al.[20]. The weighted squared z-score for each variant divided by the total Q statistic can be used to identify variants contributing most to the signal. To control type 1 error for this part of the analysis a *P*-value < 0.05/4 = 0.0125 (corrected for four tests: 1 trait x 4 subsets of variants) was used to define statistical significance for the SKAT tests.

To test whether rare variant associations were independent of the known FI GWAS hits near the *IGF1* gene, conditional analysis was performed by additionally adjusting for the two common variants rs35767 (FI top hit Dupuis et al[4]) and rs2114912 (FI top hit Manning et al[5]) ($r^2 = 0.272$ in HapMap2 CEU) in the rare variant analysis. Since these two variants were not present in the targeted sequence data, rs2162679 was used as a proxy for rs35767 ($r^2 = 0.915$ in HapMap2 CEU) and rs2607988 was used as a proxy for rs2114912 ($r^2 = 0.882$ in HapMap2 CEU). Conditional SKAT analyses were performed in each cohort seperately and then meta-analyzed. Similar *P*-values in unconditional and conditional analyses suggest that rare variant associations are independent of the known common variant signals.

Although tests of rare variation were the primary aim of the targeted regional sequencing study, we also tested association of all variants with minor allele count (MAC) ≥ 50 identified by sequencing with ln(FI). In ARIC and CHS standard additive genetic linear regression models were used, while in FHS mixed effects models were used to account for familial correlation. Results from each cohort were meta-analyzed using standard fixed-effect inverse-variance weighted meta-analysis[22]. *P*-values were obtained from unweighted regression models. Analyses weighted by the inverse of the sampling probability were used to obtain unbiased estimates of effect size[9]. The significance treshold for common variant analyses was set at *P*-value $< 1.0\text{x}10^{-3}$ (0.05/49 effective number of independent variants calculated using the Li and Ji approach[23])

For analyses of follow-up genotyping data in FHS, we used linear mixed effect model to compare the average trait values by genotype category. Since we performed two tests (offspring and generation 3 cohorts separately), we considered a *P*-value < 0.025 (0.05/2) as significant.

RESULTS

Descriptions of the CHARGE cohort characteristics are depicted in Table 1. Both in the individuals contributing GWAS data and in the targeted sequence samples, women were slightly overrepresented. The mean age ranged from 39 to 71 in the GWAS samples and from 54 to 72 in the targeted sequence samples. BMI was in the overweight range in all cohorts. As previously observed, FI values varied widely across studies[4]. The same was observed for the IGF-I levels in the GWAS samples.

Table 1. Descriptions of the study populations

	GWAS samples			Targeted sequence samples		
	CHS	FHS	RS	ARIC	CHS	FHS
n	1717	3293	140	1761	967	811
Men (n,%)	630 (36.7)	1558 (47.3)	68 (48.6)	875 (49.7)	432 (44.7)	392 (48.3)
Age (y)	71.6 (4.8)	39.9 (8.8)	66.2 (5.7)	54.7 (5.7)	72.5 (5.4)	54.1 (10.7)
BMI (kg/m^2)	26.1 (4.3)	27.0 (5.4)	26.4 (4.0)	27.2 (5.7)	26.4 (5.0)	27.9 (6.5)
FI (pmol/l)	72.2 (42.7)	30.9 (20.1)	90.1 (53.0)	83.1 (73.2)	103.1 (63.9)	32.6 (21.3)
IGF1 (ng/ml)	96 (32.7)	131.1 (42.8)	136.7 (53.3)	NA	NA	NA

Values are mean (SD) unless otherwise indicated. ARIC: Atherosclerosis Risk in Communities Study, CHS: Cardiovascular Health Study, FHS: Framingham Heart Study, RS: Rotterdam Study, n: number, BMI: body mass index, FI: fasting insulin, IGF1:insulin-like growth factor-1

Mediation Analyses

Mediation analyses results are depicted in Table 2. Meta-analyses *P*-values were nominal to borderline significant for each SNV in both models (*P* = 0.05-0.15). However, effect estimates were similar to the effect estimates in up to 51,750 samples in the discovery meta-analysis[5] and in FHS, the largest contributing cohort, *P*-values were significant for each SNV in both models (Table 2). Both in the meta-analysis and in FHS alone, effect estimates were similar between model 1 (ln(FI)~SNP+age+sex+BMI) and model 2

Table 2. Association of known fasting insulin GWAS SNPs in the *IGF1* region with fasting insulin levels without and with IGF1 levels as covariate in the model

	CHS			FHS			RS			Meta			Discovery paper*		
	β	S.E.	P	β	S.E.	P	β	S.E.	P	β	S.E.	P	β	S.E.	P
	Model1: ln(FI)~SNP+age+sex+BMI														
rs2114912	0.020	0.024	0.41	−0.039	0.015	0.011	0.002	0.093	0.98	−0.021	0.013	0.09	−0.024	0.004	3.4×10^{-11}
rs860598	0.007	0.020	0.72	−0.032	0.014	0.022	−0.072	0.076	0.34	−0.020	0.011	0.07	−0.021	0.003	6.9×10^{-10}
rs35747	0.005	0.019	0.81	−0.032	0.014	0.022	−0.079	0.079	0.32	−0.021	0.011	0.06	−0.021	0.004	8.9×10^{-10}
rs855213	0.005	0.020	0.81	−0.032	0.014	0.022	−0.072	0.076	0.34	−0.021	0.011	0.06	−0.021	0.004	1.0×10^{-9}
rs35767	0.013	0.020	0.50	−0.031	0.015	0.042	−0.127	0.080	0.11	−0.017	0.012	0.15	−0.022	0.004	2.4×10^{-9}
	Model2: ln(FI)~SNP+age+sex+BMI+IGF1														
rs2114912	0.018	0.024	0.45	−0.039	0.015	0.011	0.004	0.094	0.97	−0.022	0.013	0.08	NA	NA	NA
rs860598	0.004	0.020	0.85	−0.032	0.014	0.021	−0.071	0.077	0.36	−0.020	0.011	0.07	NA	NA	NA
rs35747	0.001	0.019	0.95	−0.033	0.014	0.020	−0.078	0.080	0.33	−0.022	0.011	0.05	NA	NA	NA
rs855213	0.002	0.020	0.94	−0.032	0.014	0.021	−0.071	0.077	0.36	−0.023	0.011	0.05	NA	NA	NA
rs35767	0.010	0.020	0.61	−0.031	0.015	0.041	−0.125	0.081	0.12	−0.018	0.012	0.13	NA	NA	NA

CHS: Cardiovascular Health Study (n=1,717), FHS: Framingham Heart Study (n=3,293), RS: Rotterdam Study (n=140), S.E.: standard error, *Manning et al.[5] (n up to 51,750)

(ln(FI)~SNP+age+sex+BMI+IGF-I). This is consistent with an effect of the variants near *IGF1* on FI levels that is not mediated by circulating IGF-I levels.

Analyses of Targeted Sequence Data

Table 3 and Supplementary Table 2 show descriptions of known and novel variants identified by targeted sequencing of the *IGF1* locus. Deep (mean read depth 50x) sequencing across the locus identified 1,393 variants, 1,143 (82.1%) of which were rare and novel. A total of 11 coding non-synonymous variants were present, including 6 that were novel. Of the 1,376 non-coding variants, 188 (14%) were predicted to be functional, including 156 that were novel. The large majority of the variants at the *IGF1* locus had MAF < 0.1% (Supplementary Figure 2). 64% of the variants were only observed one time in our samples.

Meta-analyzed SKAT results (Table 4) showed that the subset of 11 rare coding non-synonymous variants was significantly associated with ln(FI) ($P = 5.7 \times 10^{-4}$). One rare variant (chr12:101337467 (position hg18), rs151098426, MAF = 0.1%) accounted for 92.16% of the overall SKAT Q statistic (Supplementary Table 3 and Supplementary Figure 3). This variant resulted in an alanine to threonine substitution and was predicted to be damaging by PolyPhen-2[24], LRT[25] and MutationTaster[26]. In contrast to the positive effect estimate for the rare T allele of rs151098426 in the SKAT targeted sequencing analysis

Table 3. Descriptions of known and novel SNPs in the *IGF1* region in the CHARGE Targeted Sequencing Study cohorts combined

	known	novel*	total
no. SNPs	248	1145	1393
no. rare SNPs	133	1143	1276
coding variants (n=17)			
synonymous	2	4	6
nonsynonymous	5	6	11
non-coding variants (n=1376)			
intergenic	165	793	958
upstream	7	24	31
downstream	5	20	25
intronic	39	148	187
UTR3	24	146	170
UTR5	1	4	5
predicted functional#	32	156	188

Values are frequencies. *not known in dbSNP, 1000 genomes project or ESP 6500, #predicted transcription factor binding site (ENCODE ChipSeq, HMR) and/or DNAse hypersensitive site (ENCODE DHS) and/or ORegAnno regulatory variant and/or highly conserved (PhastCons)

Table 4. SKAT meta-analyses results for fasting insulin (BMI adjusted) from different subsets of rare (MAF < 1%) SNPs in the *IGF1* region

subset of rare SNVs	n SNVs in subset	P
coding nonsynonymous	11	5.7×10^{-4}
conditioned on GWAS variants[#]		0.019
coding nonsynonymous novel*	6	0.38
noncoding predicted functional**	188	0.38
noncoding predicted functional novel*	156	0.16

[#] conditioned on proxies of rs2114912 and rs35767, *not known in dbSNP, 1000 genomes project or ESP 6500, **predicted transcription factor binding site (ENCODE ChipSeq, HMR) or DNAse hypersensitive site (ENCODE DHS) or ORegAnno regulatory variant or highly conserved (PhastCons)

(Supplementary Table 3), 3 of the 1,745 FHS offspring participants and 11 of the 3,372 FHS generation 3 participants with follow-up genotyping of rs151098426 carrying the rare allele had lower FI levels than the non-carriers (offspring: $\beta = -0.05$; generation 3: $\beta = -0.15$). These differences between carriers and non-carriers were non-significant (offspring: $P = 0.734$; generation 3: $P = 0.313$). The geometric means and the corresponding confidence intervals in carriers and non-carriers are shown in Supplementary Figure 4. Lookup of the variant in CHARGE exome chip results revealed a positive, but also non-significant effect of rs151098426 on FI levels (MAF = 0.14%, $\beta = 0.02$, $P = 0.471$).

Conditioning on proxies of the known FI GWAS variants rs2114912 and rs35767 attenuated the significant SKAT result to a nominal significant *P*-value ($P_{conditioned} = 0.019$, Table 4), suggesting that the GWAS signal explains part of the rare variant signal and the presence of a residual independent rare variant effect. Examination of ENCODE Consortium regulatory element datasets and public transcriptome data in the UCSC Genome Browser suggested that GWAS variants in the vicinity of *IGF1* might have a direct functional role. In particular, rs35767 is approximately 1.2kb upstream of the *IGF1* promoter and merely a few bases away from a strong FOXA1 binding site that is observed in ENCODE ChIP-seq data across a variety of human cell lines. Similarly, rs2114912 is approximately 1.7kb away from a major multi cell line, including pancreatic islets, ENCODE DNAseI hypersensitive site that overlaps an ENCODE transcription factor binding site ChIP-seq cluster for several transcription factors, including FOXA1. This combination of open chromatin as delineated by the DNAse I hypersensitive site with transcription factor binding in ChIP-seq data constitutes a regulatory element signature that warrants experimental validation. Rs2607988, a SNP in high LD with rs2114912 ($r^2 = 0.882$ in HapMap2 CEU) is located in a ChIP-seq site for FOXA1 and alters a motif for FOXA.

Single variant analyses did not reveal significant associations with FI for any of the common variants present in the targeted sequence data (Supplementary Figure 5), includ-

ing the proxies of the known FI GWAS hits rs35767 (P_{meta} = 0.69) and rs2114912 (P_{meta} = 0.54) (Supplementary Table 4), most likely due to the much smaller sample size in these targeted sequence data compared with the original, very large discovery sample sizes.

DISCUSSION

This study suggests that previously observed associations between SNVs near *IGF1* with FI levels were not mediated by circulating IGF-I levels. Further investigation of the *IGF1* gene, using deep sequencing data, revealed a large number of rare variants at the locus that had not been previously described, the large majority of which was very rare. A subset of rare coding non-synonymous variants, including 6 novel variants and 5 variants that had been previously identified, was significantly associated with FI levels. Conditional analysis suggested that the common non-coding variants near *IGF1* that were identified in GWAS[4,5] explain part of the rare variant signal and the presence of a residual independent rare variant effect. Examination of regulatory element catalogs constructed through genome wide experimental efforts of the ENCODE Consortium showed that the GWAS variants were located in the proximity of FOXA1 binding sites and DNAseI hypersensitive sites, suggesting that they might have a direct functional role. This finding is noteworthy because FOXA1 is a key transcriptional regulator implicated in glucose metabolism and insulin secretion[27,28]. Studies in human cell culture and animal models will be needed to interrogate and validate the function of these non-coding variants in insulin biology.

One variant, rs151098426, resulting in an alanine to threonine substitution and predicted to be damaging by several annotation tools, seemed to drive the rare variant association. However, follow-up genotyping of rs151098426 in an independent set of samples and lookup of the variant in CHARGE exome chip results did not reveal significant differences in FI levels between carriers and non-carriers of the rare allele, suggesting the absence of a single variant effect for rs151098426 on FI levels. Several recently published studies have demonstrated the need for large sample sizes to robustly identify associations of low frequency variants with complex traits[29-35]. Because of the low MAF of rs151098426 and thus the relatively small number of carriers, analyzing the variant in large numbers of additional samples will be required to definitively conclude whether this variant is associated with FI levels.

We did not find a mediation effect of circulating IGF-I levels on the association of SNVs near *IGF1* with FI levels. However, measurement errors in IGF-I levels might be responsible for the absent observation of a mediation effect. Circulating IGF-I levels measured

with an imperfect assay and at a single point in time may not sufficiently characterize the biologically relevant levels. On the other hand, in 3,977 FHS participants circulating IGF-I levels correlated negatively with insulin resistance, diabetes and metabolic syndrome[3], suggesting that these measures do represent biologically relevant levels and thus making measurement errors a less likely cause for not observing a mediation effect of IGF-1 in our study.

The identification of variants at the *IGF1* locus that had not been previously described has increased our insight in the variation present at the locus. In line with previous sequencing studies[33,36,37], we identified a large number of very rare variants, the majority (64%) even observed only one time in our samples. The presence of large numbers of very rare variants in the human genome is likely explained by recent explosive human population growth[37,38]. It has been hypothesized that these variants might harbor larger effects than those observed for common variants, since selection can have influenced only the most deleterious variants[37]. However, even for rare variants with larger effects, large sample sizes are needed to definitely conclude whether they influence complex traits due to the low MAF.

Strengths of this study in the CHARGE Targeted Sequencing framework include the high average sequence depth combined with stringent QC applied across the three cohorts, increasing confidence that even the rarest observed variation is real variation and not a technical artifact. Further, we genotyped variant rs151098426 in non-overlapping samples serving as replication cohort and as further evidence that the variant is real. A limitation of this study is type 2 error, both in mediation and targeted sequence analyses, where limited sample sizes have limited power to detect common and rare variant associations. The targeted sequence samples included only 7 heterozygous carriers of variant of interest rs151098426. Further, because of the limited number of individuals with both targeted sequence data and IGF-I levels available in our study, it was not possible to test whether association of the subset of rare non-synonymous variants with FI was mediated by IGF-I levels. Mean BMI was in the overweight range in all cohorts. However, evidence exists that effect sizes of known glycemic trait associated variants do not differ between BMI strata[5]. As previously observed, FI values varied widely across studies, likely because of limited standardization across assays. Previous gene discovery studies, however, despite the same observation were successful in identifying FI-associated variants[4,5]. Finally, our study only included individuals of European ancestry, which might limit the generalizability to other ancestries of the observed *IGF1* variants and variant associations in this study.

In conclusion, our analyses suggest that association of SNVs near the *IGF1* gene with FI is not mediated by circulating IGF-I levels. Further, our study increased insight into variation present at the *IGF1* locus and thus into the specific local coding as well as non-coding genetic architecture underlying FI levels, showing a large number of novel rare variants present at the locus and suggesting association of both rare coding non-synonymous variants and a potential direct functional effect of common non-coding GWAS SNVs in the region on FI levels.

ACKNOWLEDGMENTS

REFERENCES

1. Jones, J.I. & Clemmons, D.R. Insulin-like growth factors and their binding proteins: biological actions. Endocr Rev 16, 3-34 (1995).
2. Laughlin, G.A., Barrett-Connor, E., Criqui, M.H. & Kritz-Silverstein, D. The prospective association of serum insulin-like growth factor I (IGF-I) and IGF-binding protein-1 levels with all cause and cardiovascular disease mortality in older adults: the Rancho Bernardo Study. J Clin Endocrinol Metab 89, 114-20 (2004).
3. Lam, C.S. et al. Circulating insulin-like growth factor-1 and its binding protein-3: metabolic and genetic correlates in the community. Arterioscler Thromb Vasc Biol 30, 1479-84 (2010).
4. Dupuis, J. et al. New genetic loci implicated in fasting glucose homeostasis and their impact on type 2 diabetes risk. Nat Genet 42, 105-16 (2010).
5. Manning, A.K. et al. A genome-wide approach accounting for body mass index identifies genetic variants influencing fasting glycemic traits and insulin resistance. Nat Genet 44, 659-69 (2012).
6. ENCODE Project Consortium et al. An integrated encyclopedia of DNA elements in the human genome. Nature 489, 57-74 (2012).
7. Karolchik, D. et al. The UCSC Genome Browser database: 2014 update. Nucleic Acids Res 42, D764-70 (2014).
8. Psaty, B.M. et al. Cohorts for Heart and Aging Research in Genomic Epidemiology (CHARGE) Consortium: Design of prospective meta-analyses of genome-wide association studies from 5 cohorts. Circ Cardiovasc Genet 2, 73-80 (2009).
9. Lumley, T. Two-phase subsampling designs for genomic resequencing studies. http://stattech.wordpress.fos.auckland.ac.nz/files/2012/05/design-paper.pdf
10. Lin, H. et al. Strategies to design and analyze targeted sequencing data: cohorts for Heart and Aging Research in Genomic Epidemiology (CHARGE) Consortium Targeted Sequencing Study. Circ Cardiovasc Genet 7, 335-43 (2014).
11. Miller, S.A., Dykes, D.D. & Polesky, H.F. A simple salting out procedure for extracting DNA from human nucleated cells. Nucleic Acids Res 16, 1215 (1988).
12. Li, Y., Willer, C.J., Ding, J., Scheet, P. & Abecasis, G.R. MaCH: using sequence and genotype data to estimate haplotypes and unobserved genotypes. Genet Epidemiol 34, 816-34 (2010).
13. Ewing, B. & Green, P. Base-calling of automated sequencer traces using phred. II. Error probabilities. Genome Res 8, 186-94 (1998).
14. 1000 Genomes Project Consortium, Abecasis, G.R. et al. A map of human genome variation from population-scale sequencing. Nature 467, 1061-73 (2010).
15. Exome Variant Server, NHLBI GO Exome Sequencing Project (ESP), Seattle, WA.
16. Wang, K., Li, M. & Hakonarson, H. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. Nucleic Acids Res 38, e164 (2010).
17. Siepel, A. et al. Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. Genome Res 15, 1034-50 (2005).
18. Griffith, O.L. et al. ORegAnno: an open-access community-driven resource for regulatory annotation. Nucleic Acids Res 36, D107-13 (2008).
19. Lumley, T. (2012) rmeta: Meta-analysis. R package version 2.16. http://CRAN.R-project.org/package=rmeta.
20. Wu, M.C. et al. Rare-variant association testing for sequencing data with the sequence kernel association test. Am J Hum Genet 89, 82-93 (2011).

21. Chen, H., Meigs, J.B. & Dupuis, J. Sequence kernel association test for quantitative traits in family samples. Genet Epidemiol 37, 196-204 (2013).

22. Willer, C.J., Li, Y. & Abecasis, G.R. METAL: fast and efficient meta-analysis of genomewide association scans. Bioinformatics 26, 2190-1 (2010).

23. Li, J. & Ji, L. Adjusting multiple testing in multilocus analyses using the eigenvalues of a correlation matrix. Heredity (Edinb) 95, 221-7 (2005).

24. Adzhubei, I.A. et al. A method and server for predicting damaging missense mutations. Nat Methods 7, 248-9 (2010).

25. Chun, S. & Fay, J.C. Identification of deleterious mutations within three human genomes. Genome Res 19, 1553-61 (2009).

26. Schwarz, J.M., Rodelsperger, C., Schuelke, M. & Seelow, D. MutationTaster evaluates disease-causing potential of sequence alterations. Nat Methods 7, 575-6 (2010).

27. Gao, N. et al. Foxa1 and Foxa2 maintain the metabolic and secretory features of the mature beta-cell. Mol Endocrinol 24, 1594-604 (2010).

28. Kaestner, K.H. The FoxA factors in organogenesis and differentiation. Curr Opin Genet Dev 20, 527-32 (2010).

29. Kozlitina, J. et al. Exome-wide association study identifies a TM6SF2 variant that confers susceptibility to nonalcoholic fatty liver disease. Nat Genet 46, 352-6 (2014).

30. Carty, C.L. et al. Replication of genetic loci for ages at menarche and menopause in the multi-ethnic Population Architecture using Genomics and Epidemiology (PAGE) study. Hum Reprod 28, 1695-706 (2013).

31. Peloso, G.M. et al. Association of low-frequency and rare coding-sequence variants with blood lipids and coronary heart disease in 56,000 whites and blacks. Am J Hum Genet 94, 223-32 (2014).

32. Huyghe, J.R. et al. Exome array analysis identifies new loci and low-frequency variants influencing insulin processing and secretion. Nat Genet 45, 197-201 (2013).

33. Flannick, J. et al. Loss-of-function mutations in SLC30A8 protect against type 2 diabetes. Nat Genet 46, 357-63 (2014).

34. Holmen, O.L. et al. Systematic evaluation of coding variation identifies a candidate causal variant in TM6SF2 influencing total cholesterol and myocardial infarction risk. Nat Genet 46, 345-51 (2014).

35. Zuk, O. et al. Searching for missing heritability: designing rare variant association studies. Proc Natl Acad Sci U S A 111, E455-64 (2014).

36. Morrison, A.C. et al. Whole-genome sequence-based analysis of high-density lipoprotein cholesterol. Nat Genet 45, 899-901 (2013).

37. Coventry, A. et al. Deep resequencing reveals excess rare recent variants consistent with explosive population growth. Nat Commun 1, 131 (2010).

38. Keinan, A. & Clark, A.G. Recent explosive human population growth has resulted in an excess of rare genetic variants. Science 336, 740-3 (2012).

```
          ...GACTACGTA...
       ..TACTAGCACTGTACACGA..        ..CTACGTACG
     ..GTACGTAGTCTGATCAAAACGTA..    ..CTATACAGCTA.
   ..CGTTGGTACCAGTACGTACGACTGCGA..  .GTACGTACGTACG.
  .GTGTGTCACACAGACACGTGATATAGACCAGATGACACACGTAC
  .GTACGACGGCGCGGCGCGATATATATAAAAGCACGGACACTACGA.
  ACGATGACTGACTACGTGCGCGCGTAGCTAGCTACGGATAGGACTA.
  TACTGACTACGTACTACTAAAGCTGCTTTTACGTACGTACGGGCATC
  CATGTACTATCGCGGCGCGTAGACTGTACGACCAGACTGACTGACT
  GACTGACTGACGCGCGTACGTAGCTACTGTAGTACGTACGTACGTA
  .TTCAGACTGACTGACTGAGATTACAGCTACGTAGTACTGACACGT
   .TACGACGGCGCGGCGCGATATATATAAAAGCACGGACACTACGA
    GATGACTGACTACGTGCGCGCGTAGCTAGCTACGGATAGGACTA
    TGACTACGTACTACTAAAGCTGCTTTTACGTACGTACGGGCATC
     CTGCGCATGTACTATCGCGGCGCGTAGACTGTACGACCAGAC
      ACTGACGCGCGTACGTAGCTACTGTAGTACGTACGTACGTA
      ACTGACTGACTGAGATTACAGCTACGTAGTACTGACACGT
      .GCATACGAGTCGAAGCGCGCAATATATATTATATCGGC
       .GTGACTACATCATGCTGACTACGGTCGCGCTCAACG
        .TGACTACGTACGTATACGACGTACTGACGGCGCG
        .CACTGTACACGACTGACTTACTAGCTACGTACG
         .TGATCAAAACGTACGCGCGGCTATACAGCTA
          .CGTACGACTGCGATACGTACGTACGTACG
           .CGTGATATAGACCAGATGACACACG
            .TATATATAAAAGCACGGACAC
             .GTAGCTAGCTACGGAT
              .CTTTTACGTACGT
               .ACGACCAG
                .TAGT
```

Low-frequency and rare exome chip variants associate with fasting glucose and type 2 diabetes susceptibility

ABSTRACT

Fasting glucose and insulin are intermediate traits for type 2 diabetes. Here we explore the role of coding variation on these traits by analysis of variants on the HumanExome BeadChip in 60,564 non-diabetic individuals and in 16,491 T2D cases and 81,877 controls. We identify a novel association of a low-frequency nonsynonymous SNV in *GLP1R* (A316T; rs10305492; MAF=1.4%) with lower FG (β=−0.09±0.01 mmol/L, *P*=3.4x10^{-12}), T2D risk (OR[95%CI]=0.86[0.76-0.96], *P*=0.010), early insulin secretion (β=−0.07±0.035 pmol$_{insulin}$/mmol$_{glucose}$, *P*=0.048), but higher 2-h glucose (β=0.16±0.05 mmol/L, *P*=4.3x10^{-4}). We identify a gene-based association with FG at *G6PC2* (P_{SKAT}=6.8x10^{-6}) driven by four rare protein-coding SNVs (H177Y, Y207S, R283X and S324P). We identify rs651007 (MAF=20%) in the first intron of *ABO* at the putative promoter of an antisense lncRNA, associating with higher FG (β=0.02±0.004 mmol/L, *P*=1.3x10^{-8}). Our approach identifies novel coding variant associations and extends the allelic spectrum of variation underlying diabetes-related quantitative traits and T2D susceptibility.

INTRODUCTION

Genome-wide association studies (GWAS) highlight the role of common genetic variation in quantitative glycemic traits and susceptibility to type 2 diabetes (T2D)[1,2]. However, recent large-scale sequencing studies report that rapid expansions in the human population have introduced a substantial number of rare genetic variants[3,4], with purifying selection having had little time to act, which may harbor larger effects on complex traits than those observed for common variants[3,5,6]. Recent efforts have identified the role of low frequency and rare coding variation in complex diseases and related traits[7-10], and highlight the need for large sample sizes to robustly identify such associations[11]. Thus, the Illumina HumanExome BeadChip (or exome chip) has been designed to allow the capture of rare (MAF<1%), low frequency (MAF=1-5%) and common (MAF≥5%) exonic single nucleotide variants (SNVs) in large sample sizes.

To identify novel coding SNVs and genes influencing quantitative glycemic traits and T2D, we perform meta-analyses of studies participating in the Cohorts for Heart and Aging Research in Genomic Epidemiology (CHARGE[12]) T2D-Glycemia Exome Consortium[13]. Our results show a novel association of a low frequency coding variant in *GLP1R*, a gene encoding a drug target in T2D therapy (the incretin mimetics), with FG and T2D. The minor allele is associated with lower FG, lower T2D risk, lower insulin response to a glucose challenge and higher 2-h glucose, pointing to physiological effects on the incretin system. Analyses of non-synonymous variants also enable us to identify particular genes likely to underlie previously identified associations at 6 loci associated with FG and/or FI (*G6PC2*, *GPSM1*, *SLC2A2*, *SLC30A8*, *RREB1*, and *COBLL1*) and 5 with T2D (*ARAP1, GIPR, KCNJ11, SLC30A8* and *WFS1*). Further, we found non-coding variants whose putative functions in epigenetic and post-transcriptional regulation of *ABO* and *G6PC2* are supported by experimental ENCODE Consortium, GTEx and transcriptome data from islets. In conclusion, our approach identifies novel coding and non-coding variants and extends the allelic and functional spectrum of genetic variation underlying diabetes-related quantitative traits and T2D susceptibility.

MATERIALS AND METHODS

Study cohorts

The CHARGE consortium was created to facilitate large-scale genomic meta-analyses and replication opportunities among multiple large population-based cohort studies[12]. The CHARGE T2D-Glycemia Exome Consortium was formed by cohorts within the CHARGE consortium as well as collaborating non-CHARGE studies to examine rare and common functional variation contributing to glycemic traits and T2D susceptibility.

Up to 23 cohorts participated in this effort representing a maximum total sample size of 60,564 (FG) and 48,118 (FI) participants without T2D for quantitative trait analyses. Individuals were of European (84%) and African (16%) ancestry. Full study characteristics are shown in Supplementary Data 1. Of the 23 studies contributing to quantitative trait analysis, 16 also contributed data on T2D status. These studies were combined with 6 additional cohorts with T2D case-control status for follow-up analyses of the variants observed to influence FG and FI and analysis of known T2D loci in up to 16,491 T2D cases and 81,877 controls across 4 ancestries combined (African, Asian, European and Hispanic; see Supplementary Data 2 for T2D case-control sample sizes by cohort and ancestry). All studies were approved by their local institutional review boards and written informed consent was obtained from all study participants.

Quantitative traits and phenotypes

FG (mmol/L) and FI (pmol/L) were analyzed in individuals free of T2D. FI was log transformed for genetic association tests. Study-specific sample exclusions and detailed descriptions of glycemic measurements are given in Supplementary Data 1. For consistency with previous glycemic genetic analyses, T2D was defined by cohort and included one or more of the following criteria: a physician diagnosis of diabetes, on anti-diabetic treatment, fasting plasma glucose ≥ 7 mmol/L, random plasma glucose ≥ 11.1 mmol/L, or hemoglobin A1C ≥ 6.5% (Supplementary Data 2).

Exome chip

The Illumina HumanExome BeadChip is a genotyping array containing 247,870 variants discovered through exome sequencing in ~12,000 individuals, with ~75% of the variants with a MAF < 0.5%. The main content of the chip is comprised of protein-altering variants (nonsynonymous coding, splice-site and stop gain or loss codons) seen at least 3 times in a study and in at least 2 studies providing information to the chip design. Additional variants on the chip included common variants found through GWAS, ancestry informative markers (for African and Native Americans), mitochondrial variants, randomly selected synonymous variants, HLA tag variants and Y chromosome variants. In the present study we analyzed association of the autosomal variants with glycemic traits and T2D. See Supplementary Fig. 1 for study design and analysis flow.

Exome array genotyping and quality control

Genotyping was performed with the Illumina HumanExome BeadChip v1.0 (N = 247,870 SNVs) or v1.1 (N = 242,901 SNVs). Illumina's GenTrain version 2.0 clustering algorithm in GenomeStudio or zCall[48] was used for genotype calling. Details regarding genotyping and QC for each study are summarized in Supplementary Data 1. To improve accurate calling of rare variants ten studies comprising N = 62,666 samples

participated in joint calling centrally, which has been described in detail elsewhere[13]. In brief, all samples were combined and genotypes were initially auto-called with the Illumina GenomeStudio v2011.1 software and the GenTrain2.0 clustering algorithm. SNVs meeting best practices criteria[13] based on call rates, genotyping quality score, reproducibility, heritability and sample statistics were then visually inspected and manually re-clustered when possible. The performance of the joint calling and best practices approach (CHARGE clustering method) was evaluated by comparing exome chip data to available whole exome sequencing data (N=530 in ARIC). The CHARGE clustering method performed better compared to other calling methods and showed 99.8% concordance between the exome chip and exome sequence data. 8,994 SNVs failed QC across joint calling of studies and were omitted from all analyses. Additional studies used the CHARGE cluster files to call genotypes or used a combination of gencall and zCall[48]. The quality control criteria performed by each study for filtering of poorly genotyped individuals and of low-quality SNVs included a call rate of <0.95, gender mismatch, excess autosomal heterozygosity, and SNV effect estimate standard error $>10^{-6}$. Concordance rates of genotyping across the exome chip and GWAS platforms was checked in ARIC and FHS and was > 99%. After SNV-level and sample-level quality control, 197,481 variants were available for analyses. The minor allele frequency spectrums of the exome chip SNVs by annotation category are depicted in Supplementary Table 22. Cluster plots of *GLP1R* and *ABO* variants are shown in Supplementary Fig. 9.

Whole exome sequencing

For exome sequencing analyses we had data from up to 14,118 individuals of European ancestry from 7 studies, including 4 studies contributing exome sequence samples that also participated in the exome chip analyses (Atherosclerosis Risk in Communities Study (ARIC, N = 2,905), Cardiovascular Health Study (CHS, N = 645), Framingham Heart Study (FHS, N = 666) and Rotterdam Study (RS, N = 702)) and three additional studies, Erasmus Rucphen Family Study (ERF, N = 1,196), the Exome Sequencing Project (ESP, N = 1,338), and the GlaxoSmithKline discovery sequence project[3] (GSK, N = 6,666). The GlaxoSmithKline (GSK) discovery sequence project provided summary level statistics combining data from GEMS, CoLaus and LOLIPOP collections that added additional exome sequence data at *GLP1R*, including N=3,602 samples with imputed genotypes. In all studies sequencing was performed using the Illumina HiSeq 2000 platform. The reads were mapped to the GRCh37 Human reference genome (http://www.ncbi.nlm. nih.gov/projects/genome/assembly/grc/human/) using the Burrows-Wheeler aligner (BWA[49], http://bio-bwa.sourceforge.net/), producing a BAM[50] (binary alignment/map) file. In ERF, the NARWHAL pipeline[51] was used for this purpose as well. In GSK paired-end short reads were aligned with SOAP[52]. GATK[53] (http://www.broadinstitute.org/gatk/)

and Picard (http://picard.sourceforge.net) were used to remove systematic biases and to do quality recalibration. In ARIC, CHS and FHS the Atlas2[54] suite (Atlas-SNP and Atlas-indel) was used to call variants and produce a variant call file (VCF[55]). In ERF and RS genetic variants were called using the Unified Genotyper Tool from GATK, for ESP the University of Michigan's multisample SNP calling pipeline UMAKE was used (H.M. Kang and G. Jun, unpublished data) and in GSK variants were called using SOAPsnp[56]. In ARIC, CHS and FHS variants were excluded if SNV posterior probability was < 0.95 (QUAL<22), number of variant reads were < 3, variant read ratio was < 0.1, > 99% variant reads were in a single strand direction, or total coverage was < 6. Samples that met a minimum of 70% of the targeted bases at 20X or greater coverage were submitted for subsequent analysis and QC in the three cohorts. SNVs with > 20% missingness, > 2 observed alleles, monomorphic, mean depth at the site of > 500-fold or HWE $P < 5 \times 10^{-6}$ were removed. After variant-level QC, a quality assessment of the final sequence data was performed in ARIC, CHS and FHS based on a number of measures, and all samples with a missingness rate of > 20% were removed. In RS, samples with low concordance to genotyping array (< 95%), low transition/transversion ratio (< 2.3) and high heterozygote to homozygote ratio (> 2.0) were removed from the data. In ERF, low quality variants were removed using a QUAL < 150 filter. Details of variant and sample exclusion criteria in ESP and GSK have been described before[3,57]. In brief, in ESP these were based on allelic balance (the proportional representation of each allele in likely heterozygotes), base quality distribution for sites supporting the reference and alternate alleles, relatedness between individuals and mismatch between called and phenotypic gender. In GSK these were based on sequence depth, consensus quality and concordance with genome-wide panel genotypes, amongst others.

Phenotyping glycemic physiologic traits in additional cohorts

We tested association of the lead signal rs10305492 at *GLP1R* with glycemic traits in the post absorptive state because it has a putative role in the incretin effect. Cohorts with measurements of glucose and/or insulin levels post 75g oral glucose tolerance test (OGTT) were included in the analysis (see Supplementary Table 2 for list of participating cohorts and sample sizes included for each trait). We used linear regression models under the assumption of an additive genetic effect for each physiologic trait tested.

Ten cohorts (ARIC, CoLaus, Ely, Fenland, FHS, GLACIER, Health2008, Inter99, METSIM, RISC, Supplementary Table 2) provided data for the 2-h glucose levels for a total sample size of 37,080 individuals. We collected results for 2-h insulin levels in a total of 19,362 individuals and for 30min-insulin levels in 16,601 individuals. Analyses of 2-h glucose, 2-h insulin, and 30min-insulin were adjusted using 3 models: 1) age, sex and center; 2) age, sex, center and BMI; and 3) age, sex, center, BMI, and FG. The main results in

the manuscript are presented using model 3. We opted for the model that included FG because these traits are dependent on baseline FG[1,58]. Adjusting for baseline FG assures the effect of a variant on these glycemic physiologic traits are independent of FG.

We calculated the insulinogenic index using the standard formula: [insulin 30 min – insulin baseline] / [glucose 30min – glucose baseline] and collected data from 5 cohorts with appropriate samples (total N = 16,203 individuals). Models were adjusted for age, sex, center, then additionally for BMI. In individuals with ≥ 3 points measured during OGTT, we calculated the area under the curve (AUC) for insulin and glucose excursion over the course of OGTT using the trapezoid method[59]. For the analysis of AUCins (N = 16,126 individuals) we used 3 models as discussed above. For the analysis of AUCins / AUCgluc (N = 16,015 individuals) we only used models 1 and 2 for adjustment.

To calculate the incretin effect, we used data derived from paired OGTT and intra-venous glucose tolerance test (IVGTT) performed in the same individuals using the formula: [AUCins OGTT-AUCins IVGTT] / [AUCins OGTT] in RISC (N = 738). We used models 1 and 2 (as discussed above) for adjustment.

We were also able to obtain lookups for estimates of insulin sensitivity from euglycemic-hyperinsulinemic clamps and from frequently sampled IVGTT from up to 2,170 and 1,208 individuals, respectively (Supplementary Table 3).

All outcome variables except 2-h glucose were log transformed. Effect sizes were reported as standard deviations using standard deviations of each trait in the Fenland study[60], the Ely study[61] for insulinogenic index and the RISC study[62] for incretin effects to allow for comparison of effect sizes across phenotypes.

Statistical analyses

The R package seqMeta was used for single variant, conditional and gene-based association analyses[63] (http://cran.r-project.org/web/packages/seqMeta/). We performed linear regression for the analysis of quantitative traits and logistic regression for the analysis of binary traits. For family-based cohorts linear mixed effects models were used for quantitative traits and related individuals were removed before logistic regression was performed. All studies used an additive coding of variants to the minor allele observed in the jointly called data set[13]. All analyses were adjusted for age, sex, principal components calculated from genome-wide or exome chip genotypes and study specific covariates (when applicable) (Supplementary Data 1). Models testing FI were further adjusted for BMI[32]. Each study analyzed ancestral groups separately. At the meta-analysis level ancestral groups were analyzed both separately and combined. Meta-analyses

were performed by two independent analysts and compared for consistency. Overall quantile-quantile plots are shown in Supplementary Fig. 10.

Bonferroni correction was used to determine the threshold of significance. In single variant analyses, for FG and FI, all variants with a MAF > 0.02% (equivalent to a MAC $\geq$ 20; N_{SNVs} = 150,558) were included in single variant association tests; the significance threshold was set to $P \leq 3 \times 10^{-7}$ ($P = 0.05/150,558$), corrected for the number of variants tested. For T2D, all variants with a MAF > 0.01% in T2D cases (equivalent to a MAC $\geq$ 20 in cases; N_{SNVs} = 111,347) were included in single variant tests; the significance threshold was set to $P \leq 4.5 \times 10^{-7}$ ($P = 0.05/111,347$).

We used two gene-based tests: the Sequence Kernel Association Test (SKAT) and the Weighted Sum Test (WST) using Madsen Browning weights to analyze variants with MAF < 1% in genes with a cumulative MAC $\geq$ 20 for quantitative traits and cumulative MAC $\geq$ 40 for binary traits. These analyses were limited to stop gain/loss, nonsynonymous, or splice-site variants as defined by dbNSFP v2.0[31]. We considered a Bonferroni corrected significance threshold of $P \leq 1.6 \times 10^{-6}$ (0.05/30,520 tests (15,260 genes x 2 gene-based tests)) in the analysis of FG and FI and $P \leq 1.7 \times 10^{-6}$ (0.05/29,732 tests (14,866 genes x 2 gene-based tests)) in the analysis of T2D. Due to the association of multiple rare variants with FG at *G6PC2* from both single and gene-based analyses, we removed 1 variant at a time and repeated the SKAT test to determine the impact of each variant on the gene-based association effects (Wu weight) and statistical significance.

We performed conditional analyses to control for the effects of known or newly dis-covered loci. The adjustment command in seqMeta was used to perform conditional analysis on SNVs within 500kb of the most significant SNV. For *ABO* we used the most significant SNV, rs651007. For *G6PC2* we used the previously reported GWAS variants, rs563694 and rs560887, which were also the most significant SNV(s) in the data analyzed here.

The threshold of significance for known FG and FI loci was set at $P_{FG} \leq 1.5 \times 10^{-3}$ and $P_{FI} < 2.9 \times 10^{-3}$ (= 0.05/34 known FG loci and = 0.05/17 known FI loci). For FG, FI and T2D functional variant analyses the threshold of significance was computed as $P = 1.1 \times 10^{-5}$ (= 0.05/4,513 protein affecting SNVs at 38 known FG susceptibility loci), $P = 3.9 \times 10^{-5}$ (= 0.05/1,281 protein affecting SNVs at 20 known FI susceptibility loci), $P = 1.3 \times 10^{-4}$ (= 0.05/412 protein affecting SNVs at 72 known T2D susceptibility loci); $P = 3.5 \times 10^{-4}$ (0.05/(72x2)) for the gene-based analysis of 72 known T2D susceptibility loci[2,34]. We assessed the associations of glycemic[1,32,64] and T2D[2,34] variants identified by previous GWAS in our population.

We developed a novel meta-analysis approach for haplotype results based on an extension of Zaykin's method[65]. We incorporated family structure into the basic model, making it applicable to both unrelated and related samples. All analyses were performed in R. We developed an R function to implement the association test at the cohort level. The general model formula for K observed haplotypes (with the most frequent haplotype used as the reference) is

$$Y = \mu + X\gamma + \beta_2 h_2 + \cdots + \beta_K + b + \varepsilon$$

where Y is the trait; X is the covariates matrix; h_m(m = 2,…, K) is the expected haplotype dosage: if the haplotype is observed, the value is 0 or 1; otherwise, the posterior probability is inferred from the genotypes; b is the random intercept accounting for the family structure (if it exists), and is 0 for unrelated samples; ε is the random error.

For meta-analysis, we adapted a multiple parameter meta-analysis method to summarize the findings from each cohort[66]. One primary advantage is that this approach allows variation in the haplotype set provided by each cohort. In other words, each cohort could contribute uniquely observed haplotypes in addition to those observed by multiple cohorts.

Associations of *ABO* variants with cardiometabolic traits

Variants in the *ABO* region have been associated with a number of cardiovascular and metabolic traits in other studies (Supplementary Table 8), suggesting a broad role for the locus in cardiometabolic risk. For significantly associated SNVs in this novel glycemic trait locus, we further investigated their association with other metabolic traits, including systolic blood pressure (SBP, in mmHg), diastolic blood pressure (DBP, in mmHg), body mass index (BMI, in kg/m^2), waist hip ratio (WHR) adjusted for BMI, high-density lipoprotein cholesterol (HDL-C, in mg/dl), low-density lipoprotein cholesterol (LDL-C, in mg/dl), triglycerides (TG, natural log transformed, in % change units) and total cholesterol (TC, in mg/dl). These traits were examined in single variant exome chip analysis results in collaboration with other CHARGE working groups. All analyses were conducted using the R packages skatMeta or seqMeta[63]. Analyses were either sex stratified (BMI and WHR analyses) or adjusted for sex. Other covariates in the models were age, principal components and study specific covariates. BMI, WHR, SBP and DBP analyses were additionally adjusted for age squared; WHR, SBP and DBP were BMI adjusted. For all individuals taking any blood pressure lowering medication, 15 mmHg was added to their measured SBP value and 10 mmHg to the measured DBP value. As described in detail previously[8] in selected individuals using lipid lowering medication, the untreated lipid levels were estimated and used in the analyses. All genetic variants were coded additively. Maximum sample sizes were 64,965 in adiposity analyses, 56,538 in lipid analyses and 92,615

in blood pressure analyses. Threshold of significance was $P = 6.2 \times 10^{-3}$ ($P = 0.05/8$, where 8 is the number of traits tested).

Pathway analyses of *GLP1R*

To examine whether biological pathways curated into gene sets in several publicly available databases harbored exome chip signals below the threshold of exome-wide significance for FG or FI, we applied the MAGENTA gene-set enrichment analysis (GSEA) software as previously described using all pathways in the Kyoto Encyclopedia of Genes and Genomes (KEGG), Gene Ontology (GO), Reactome, Panther, BioCarta, and Ingenuity pathway databases[67]. Genes in each pathway were scored based on unconditional meta-analysis *P*-values for SNVs falling within 40 kb upstream and 110 kb downstream of gene boundaries; we used a 95th percentile enrichment cutoff in MAGENTA, meaning pathways (gene sets) were evaluated for enrichment with genes harboring signals exceeding the 95th percentile of all genes. As we tested a total of 3,216 pathways in the analysis, we used a Bonferroni corrected significance threshold of $P < 1.6 \times 10^{-5}$ in this unbiased examination of pathways. To limit the GSEA analysis to pathways that might be implicated in glucose or insulin metabolism, we selected gene sets from the above databases whose names contained the terms "gluco," "glycol," "insulin," or "metabo." We ran MAGENTA with FG and FI datasets on these "glucometabolic" gene sets using the same gene boundary definitions and 95th percentile enrichment cutoff as described above; as this analysis involved 250 gene sets, we specified a Bonferroni corrected significance threshold of $P < 2.0 \times 10^{-4}$. Similarly, to examine whether genes associated with incretin signaling harbored exome chip signals, we applied MAGENTA software to a gene-set that we defined comprised of genes with putative biologic functions in pathways common to GLP1R activation and insulin secretion, using the same gene boundaries and 95th percentile enrichment cutoff described above (Supplementary Table 4). To select genes for inclusion in the incretin pathway gene set, we examined the "Insulin secretion" and "Glucagon-like peptide-1 regulates insulin secretion" pathways in KEGG and Reactome, respectively. From these two online resources, genes encoding proteins implicated in GLP1 production and degradation (namely glucagon and DPP4), acting in direct pathways common to GLP1R and insulin transcription, or involved in signaling pathways shared by GLP1R and other incretin family members were included in our incretin signaling pathway gene set; however, we did not include genes encoding proteins in the insulin secretory pathway or encoding cell membrane ion channels as these processes likely have broad implications for insulin secretion independent from GLP1R signaling. As this pathway included genes known to be associated with FG, we repeated the MAGENTA analysis excluding genes with known association from our gene set – *PDX1, ADCY5, GIPR* and *GLP1R* itself.

Protein conformation simulations

The A316T receptor mutant structure was modeled based on the WT receptor structure published previously[22]. First, the Threonine residue is introduced in place of Alanine at position 316. Then, this receptor structure is inserted back into the relaxed membrane-water system from the WT structure[22]. T316 residue and other residues within 5Å of itself are minimized using the CHARMM force field[68] in the NAMD[69] molecular dynamics (MD) program. This is followed by heating the full receptor-membrane-water to 310K and running MD simulation for 50 nanoseconds using the NAMD program. Electrostatics are treated by E-wald summation and a time step of 1 femtosecond is used during the simulation. The structure snapshots are saved every 1ps and the fluctuation analysis (Supplementary Fig. 3) used snapshots every 100ps. The final snapshot is shown in all the structural figures.

Annotation and functional prediction of variants

Variants were annotated using dbNSFP v2.0[31]. GTEx (Genotype-Tissue Expression Project) results were used to identify variants associated with gene expression levels using all available tissue types[16]. The Encyclopedia of DNA Elements (ENCODE) Consortium results[14] were used to identify non-coding regulatory regions, including but not limited to transcription factor binding sites (ChIP-seq), chromatin state signatures, DNAse I hypersensitive sites, and specific histone modifications (ChIP-seq) across the human cell lines and tissues profiled by ENCODE. We used the UCSC Genome Browser[15,70] to visualize these datasets, along with the public transcriptome data contained in the browser's "Genbank mRNA" (cDNA) and "Human ESTs" (Expressed Sequence Tags) tracks, on the hg19 human genome assembly. LncRNA and antisense transcription were inferred by manual annotation of these public transcriptome tracks at UCSC. All relevant track groups were displayed in Pack or Full mode and the Experimental Matrix for each subtrack was configured to display all extant intersections of these regulatory and transcriptional states with a selection of cell or tissue types comprised of ENCODE Tier 1 and Tier 2 human cell line panels, as well as all cells and tissues (including but not limited to pancreatic beta cells) of interest to glycemic regulation. We visually scanned large genomic regions containing genes and SNVs of interest and selected trends by manual annotation (this is a standard operating procedure in locus-specific in-depth analyses utilizing ENCODE and the UCSC Browser). Only a subset of tracks displaying gene structure, transcriptional and epigenetic datasets from or relevant to T2D, and SNVs in each region of interest was chosen for inclusion in each UCSC Genome Browser-based figure. Uninformative tracks (those not showing positional differences in signals relevant to SNVs or genes of interest) were not displayed in the figures. ENCODE and transcriptome datasets were accessed via UCSC in February and March 2014. In order to investigate the possible significant overlap between the *ABO* locus SNPs of interest

and ENCODE feature annotations we performed the following analysis. The following datasets were retrieved from the UCSC genome browser: wgEncodeRegTfbsClusteredV3 (TFBS); wgEncodeRegDnaseClusteredV2 (DNase); all H3K27ac peaks (all: wgEncodeBroa dHistone*H3k27acStdAln.bed files); and all H3K4me1 peaks (all: wgEncodeBroadHiston e*H3k4me1StdAln.bed files). The histone mark files were merged and the maximal score was taken at each base over all cell lines. These features were then overlapped with all SNPs on the exome chip from this study using bedtools (v2.20.1). GWAS SNPs were determined using the NHGRI GWAS catalog with $P < 5*10^{-8}$. LD values were obtained by the PLINK program based on the Rotterdam Study for SNPs within 100 kB with an r2 threshold of 0.7. Analysis of these files was completed with a custom R script to produce the fractions of non-GWAS SNPs with stronger feature overlap than the ABO SNPs as well as the supplementary figure.

RESULTS

An overview of the study design is shown in Supplementary Fig. 1, and participating studies and their characteristics are detailed in Supplementary Data 1. We conducted single variant and gene-based analyses for fasting glucose (FG) and fasting insulin (FI), by combining data from 23 studies comprising up to 60,564 (FG) and 48,118 (FI) non-diabetic individuals of European and African ancestry. We followed up associated variants at novel and known glycemic loci by tests of association with T2D, additional physiological quantitative traits (including post-absorptive glucose and insulin dynamic measures), pathway analyses, protein conformation modelling, comparison with whole exome sequence data, and interrogation of functional annotation resources including ENCODE[14,15] and GTEx[16]. We performed single variant analyses using additive genetic models of 150,558 SNVs (P-value for significance $\leq 3 \times 10^{-7}$) restricted to MAF > 0.02% (equivalent to a minor allele count (MAC) $\geq$ 20); and gene based tests using Sequence Kernel Association (SKAT) and Weighted Sum Tests (WST) restricted to variants with MAF < 1% in a total of 15,260 genes (P-value for significance $\leq 2 \times 10^{-6}$, based on number of gene tests performed). T2D case/control analyses included 16,491 individuals with T2D and 81,877 controls from 22 studies (Supplementary Data 2).
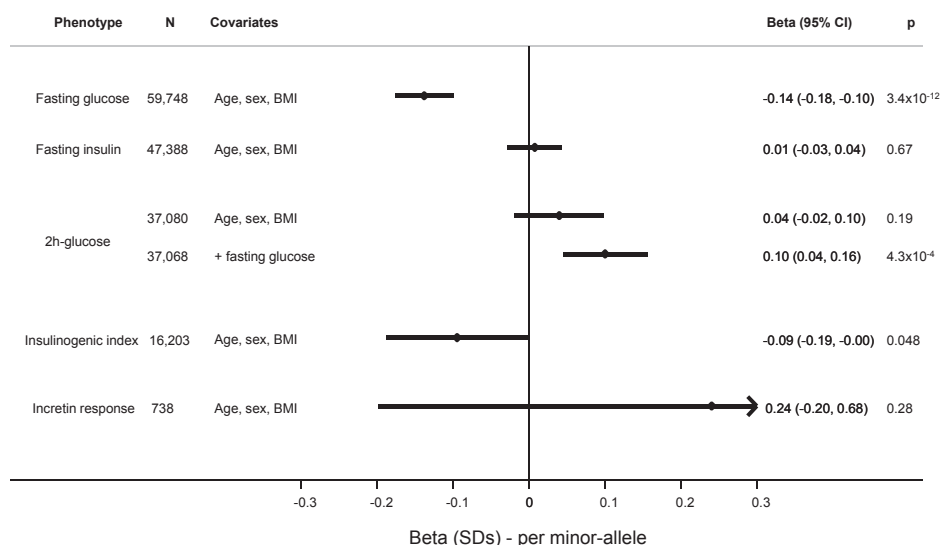
We identified a novel association of a nonsynonymous SNV (nsSNV) (A316T, rs10305492, MAF=1.4%) in the gene encoding the receptor for glucagon-like peptide 1 (*GLP1R*), with the minor (A) allele associated with lower FG ($\beta = -0.09\pm0.01$ mmol/L (equivalent to 0.14 SDs in FG), $P = 3.4 \times 10^{-12}$, variance explained = 0.03%, Table 1 and Fig. 1), but not with FI ($P = 0.67$, Supplementary Table 1). GLP-1 is secreted by intestinal L-cells in response to oral feeding and accounts for a major proportion of the so-called "incretin effect", i.e.

Table 1. Novel SNPs associated with fasting glucose in African and European ancestries combined

Gene	Variation type	Chr	Build 37 position	dbSNPID	Alleles		African and European				Proportion of trait variance explained
					Effect	Other	EAF	Beta	SE	p	
GLP1R	A316T	6	39046794	rs10305492	A	G	0.01	−0.09	0.013	3.4×10^{-12}	0.0003
ABO	intergenic	9	136153875	rs651007	A	G	0.20	0.02	0.004	1.3×10^{-8}	0.0002

Fasting glucose concentrations were adjusted for sex, age, cohort effects and up to 10 principal components in up to 60,564 (AF N=9,664 and EU N=50,900) non-diabetic individuals. Effects are reported per copy of the minor allele. Beta coefficient units are in mmol L^{-1}. EAF = effect allele frequency

the augmentation of insulin secretion following an oral glucose challenge relative to an intravenous glucose challenge. GLP-1 has a range of downstream actions including glucose-dependent stimulation of insulin release, inhibition of glucagon secretion from the islet alpha-cells, appetite suppression and slowing of gastrointestinal motility[17,18]. In follow-up analyses, the FG-lowering minor A allele was associated with lower T2D risk (OR [95%CI] = 0.86 [0.76-0.96], P = 0.010, Supplementary Data 3). Given the role of incretin hormones in post-prandial glucose regulation, we further investigated the association of A316T with measures of post-challenge glycemia, including 2-h glucose, and 30min-insulin and glucose responses expressed as the insulinogenic index[19] in up to

Phenotype	N	Covariates	Beta (95% CI)	p
Fasting glucose	59,748	Age, sex, BMI	-0.14 (-0.18, -0.10)	3.4×10^{-12}
Fasting insulin	47,388	Age, sex, BMI	0.01 (-0.03, 0.04)	0.67
2h-glucose	37,080	Age, sex, BMI	0.04 (-0.02, 0.10)	0.19
	37,068	+ fasting glucose	0.10 (0.04, 0.16)	4.3×10^{-4}
Insulinogenic index	16,203	Age, sex, BMI	-0.09 (-0.19, -0.00)	0.048
Incretin response	738	Age, sex, BMI	0.24 (-0.20, 0.68)	0.28

Beta (SDs) - per minor-allele

Figure 1. Glycemic associations with rs10305492 (*GLP1R* A316T).
Glycemic phenotypes were tested for association with rs10305492 in *GLP1R* (A316T). Each phenotype, sample size (N), covariates in each model, beta per standard deviation, 95% confidence interval (95%CI) and p-values (*p*) are reported. Analyses were performed on native distributions and scaled to SDs from the Fenland or Ely studies to allow comparisons of effect sizes across phenotypes.

37,080 individuals from 10 studies (Supplementary Table 2). The FG-lowering allele was associated with higher 2-h glucose levels (β in SDs per-minor allele [95%CI]: 0.10 [0.04, 0.16], P = 4.3x10^{-4}, N = 37,068) and lower insulinogenic index (−0.09 [−0.19, −0.00], P = 0.048, N = 16,203), indicating lower early insulin secretion (Fig. 1). Given the smaller sample size, these associations are less statistically compelling; however, the directions of effect indicated by their beta values are comparable to those observed for fasting glucose. We did not find a significant association between A316T and the measure of "incretin effect", but this was only available in a small sample size of 738 non-diabetic individuals with both oral and intravenous glucose tolerance test data (β in SDs per-minor allele [95%CI]: 0.24 [−0.20-0.68], P = 0.28, Fig. 1 and Supplementary Table 2). We did not see any association with insulin sensitivity estimated by euglycemic-hyperinsulinemic clamp or frequently sampled IVGTT (Supplementary Table 3). While stimulation of the GLP-1 receptor has been suggested to reduce appetite[20] and treatment with GLP1R agonists can result in reductions in BMI[21], these potential effects are unlikely to influence our results, which were adjusted for BMI.

In an effort to examine the potential functional consequence of the *GLP1R* A316T variant, we modeled the A316T receptor mutant structure based on the recently published[22] structural model of the full length human GLP-1 receptor bound to exendin-4 (an exogenous GLP-1 agonist). The mutant structural model was then relaxed in the membrane environment using molecular dynamics simulations. We found that the T316 variant (in transmembrane (TM) domain 5) disrupts hydrogen bonding between N320 (in TM5) and E364 (TM6) (Supplementary Fig. 2). In the mutant receptor, T316 displaces N320 and engages in a stable interaction with E364, resulting in slight shifts of TM5 towards the cytoplasm and TM6 away from the cytoplasm (Supplementary Figs. 3 and 4). This alters the conformation of the third intracellular loop, which connects TM5 and TM6 within the cell, potentially affecting downstream signaling through altered interaction with effectors such as G proteins.

A targeted Gene Set Enrichment Analysis (Supplementary Table 4) identified enrichment of genes biologically related to *GLP1R* in the incretin signaling pathway (P = 2x10-4); after excluding *GLP1R* and previously known loci *PDX1, GIPR* and *ADCY5*, the association was attenuated (P = 0.072). Gene-based tests at *GLP1R* did not identify significant associations with glycemic traits or T2D susceptibility, further supported by Fig. 2, which indicates only one variant in the *GLP1R* region on the exome chip showing association with FG.

To more fully characterize the extent of local sequence variation and its association with FG at *GLP1R*, we investigated 150 *GLP1R* SNVs identified from whole exome sequencing in up to 14,118 individuals available in CHARGE and the GlaxoSmithKline discovery sequence project (Supplementary Table 5). Single variant analysis identified association of 12 other SNVs with FG (P < 0.05; Supplementary Data 4) suggesting that
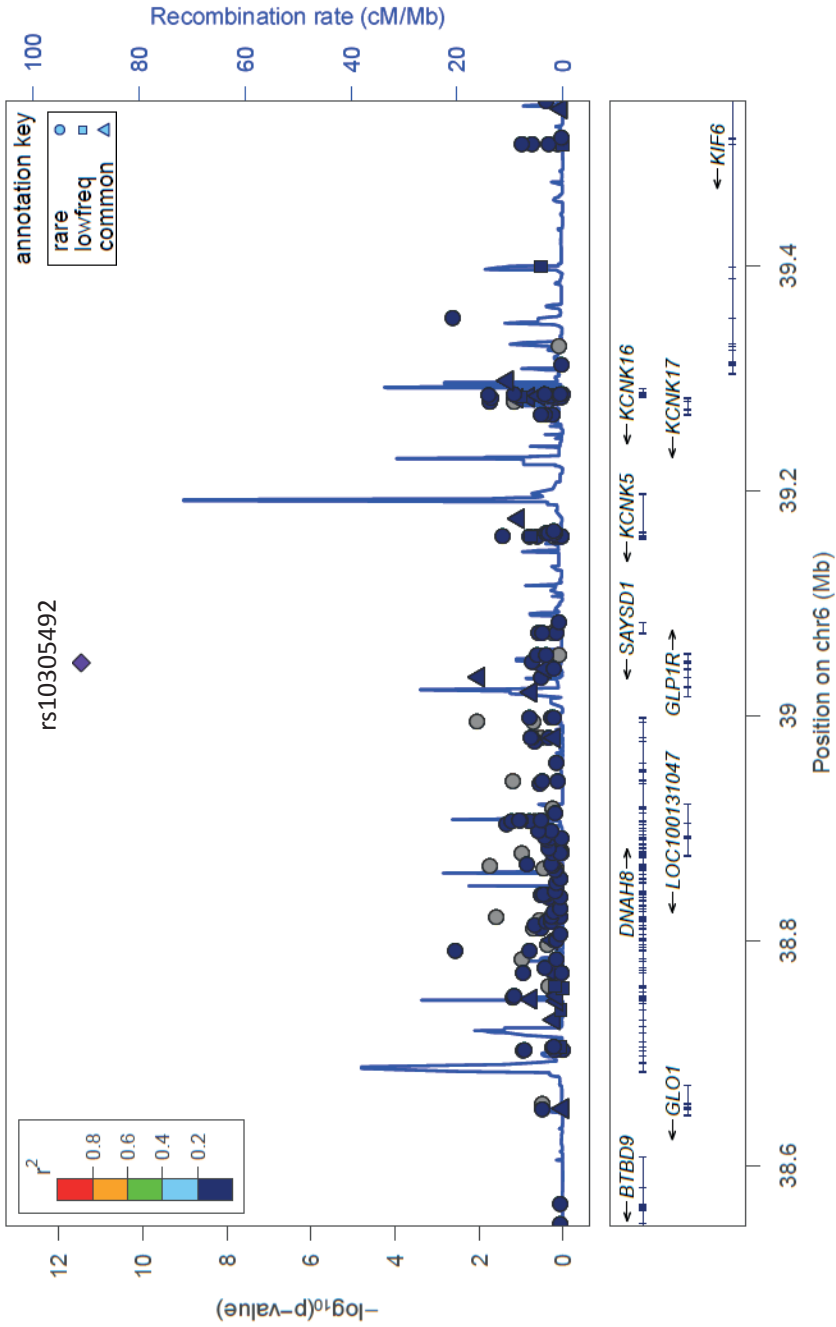
Figure 2. *GLP1R* regional association plot.

Regional association results ($-\log_{10}p$) for fasting glucose of *GLP1R* locus on chromosome 6. Linkage disequilibrium (r^2) indicated by color scale legend. Triangle symbols indicate variants with MAF>5%, square symbols indicate variants with MAF 1-5%, and circle symbols indicate variants with MAF <1%.

additional variants at this locus may influence FG, including two variants (rs10305457 and rs761386) in close proximity to splice sites that raise the possibility that their functional impact is exerted via effects on *GLP1R* pre-mRNA splicing. However, the smaller sample size of the sequence data limits power for firm conclusions.

We also newly identified that the minor allele A at rs651007 near the *ABO* gene was associated with higher FG (β = 0.02±0.004 mmol/L, MAF = 20%, P = 1.3x10^{-8}, variance explained = 0.02%, Table 1). Three other associated common variants in strong linkage disequilibrium (LD) (r^2 = 0.95-1) were also located in this region; conditional analyses suggested that these four variants reflect one association signal (Supplementary Table 6). The FG-raising allele of rs651007 was nominally associated with increased FI (β=0.008±0.003, P=0.02, Supplementary Table 1) and T2D risk (OR [95%CI] = 1.05 [1.01-1.08], P = 0.01, Supplementary Data 3). Further, we independently replicated the association at this locus with FG in non-overlapping data from MAGIC[1] using rs579459, a variant in LD with rs651007 and genotyped on the Illumina CardioMetabochip (β = 0.008±0.003 mmol/L, P = 5.0x10^{-3}; N_{MAGIC} = 88,287). The FG-associated SNV at *ABO* was in low LD with the three variants[23] that distinguish between the four major blood groups O, A1, A2 and B (rs8176719 r^2 = 0.18, rs8176749 r^2 = 0.01 and rs8176750 r^2 = 0.01). The blood group variants (or their proxies) were not associated with FG levels (Supplementary Table 7).

Variants in the *ABO* region have been associated with a number of cardiovascular and metabolic traits in other studies (Supplementary Table 8), suggesting a broad role for this locus in cardiometabolic risk. A search of the four FG-associated variants and their associations with metabolic traits using data available through other CHARGE working groups (Supplementary Table 9) revealed a significant association of rs651007 with BMI in women (β = 0.025±0.01 kg/m^2, P = 3.4x10^{-4}) but not in men. As previously reported[24,25], the FG increasing allele of rs651007 was associated with increased LDL and TC (LDL: β = 2.3±0.28 mg/dl, P = 6.1x10^{-16}; TC: β = 2.4±0.33 mg/dl, P = 3.4x10^{-13}). Because the FG-associated *ABO* variants were located in non-coding regions (intron 1 or intergenic) we interrogated public regulatory annotation datasets, GTEx[16] (http://www.gtexportal.org/home/) and the ENCODE Consortium resources[14] in the UCSC Genome Browser[15] (http://genome.ucsc.edu/) and identified a number of genomic features coincident with each of the four FG-associated variants. Three of these SNPs, upstream of the *ABO* promoter, reside in a DNase I hypersensitive site with canonical enhancer marks in ENCODE Consortium data: H3K4Me1 and H3K27Ac (Supplementary Fig. 5). We analyzed all SNPs with similar annotations, and find that these three are coincident with DNase, H3K4Me1 and H3K27Ac values each near the genome-wide mode of these assays (Supplementary Fig. 6). Indeed, in hematopoietic model K562 cells, the ENCODE Consortium has identified the region overlapping these SNPs as a putative enhancer[14]. Interrogating the GTEx database (N = 156), we found that rs651007 (P = 5.9x10^{-5}) and

rs579459 (P = 6.7 x10^{-5}) are eQTLs for *ABO*, and rs635634 (P = 1.1x10^{-4}) is an eQTL for SLC2A6 in whole blood (Supplementary Table 10). The fourth SNP, rs507666, resides near the transcription start site of a long non-coding RNA that is antisense to exon 1 of *ABO* and expressed in pancreatic islets (Supplementary Fig. 5). rs507666 was also an eQTL for the glucose transporter *SLC2A6* (P = 1.1x10^{-4}) (Supplementary Fig. 5 and Supplementary Table 10). *SLC2A6* codes for a glucose transporter whose relevance to glycemia and T2D is largely unknown, but expression is increased in rodent models of diabetes[26]. Gene-based analyses did not reveal significant quantitative trait associations with rare coding variation in *ABO*.

At the known glycemic locus *G6PC2*, gene-based analyses of 15 rare predicted protein-altering variants (MAF < 1%) present on the exome chip revealed a significant associa-tion of this gene with FG (cumulative MAF of 1.6%, P_{SKAT}=8.2x10^{-18}, P_{WST}=4.1x10^{-9}; Table 2). The combination of 15 rare SNVs remained associated with FG after conditioning on two known common SNVs in LD[27] with each other (rs560887 in intron 1 of *G6PC2* and rs563694 located in the intergenic region between *G6PC2* and *ABCB11*) (conditional P_{SKAT} = 5.2x10^{-9}, P_{WST} = 3.1x10^{-5}; Table 2 and Fig. 3), suggesting that the observed rare vari-ant associations were distinct from known common variant signals. While *ABCB11* has been proposed to be the causal gene at this locus[28], identification of rare and putatively functional variants implicates *G6PC2* as the much more likely causal candidate. Since rare alleles that increase risk for common disease may be obscured by rare, neutral mutations[4], we tested the contribution of each *G6PC2* variant by removing one SNV at a time and re-calculating the evidence for association across the gene. Four SNVs, rs138726309 (H177Y), rs2232323 (Y207S), rs146779637 (R283X) and rs2232326 (S324P), each contributed to the association with FG (Fig. 3c and Supplementary Table 11). Each

Table 2. Gene-based associations of *G6PC2* with fasting glucose in African and European ancestries combined

Gene	Chr: Build 37 position	cMAF[a]	SNVs (n)[b]	Weighted Sum Test (WST)				Sequence Kernel Association Test (SKAT)			
				p	p^c	p^d	p^e	p	p^c	p^d	p^e
G6PC2	2:169757930-169764491	0.016	15	4.1×10^{-9}	2.6×10^{-5}	2.3×10^{-4}	3.1×10^{-5}	8.2×10^{-18}	4.8×10^{-9}	6.8×10^{-6}	5.2×10^{-9}

Fasting glucose concentrations were adjusted for sex, age, cohort effects and up to 10 principal compo-nents in up to 60,564 non-diabetic individuals.
[a]cMAF=combined minor allele frequency of all variants included in the analysis.
[b]SNVs(n)=number of variants included in the analysis; variants were restricted to those with MAF<0.01 and annotated as nonsynonymous, splice-site, or stop loss/gain variants.
[c]p value for gene-based test after conditioning on rs563694.
[d]p value for gene-based test after conditioning on rs560887.
[e]p value for gene-based test after conditioning on rs563694 and rs560887.
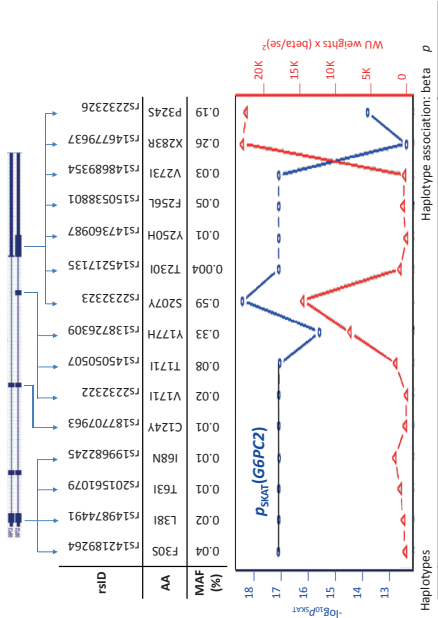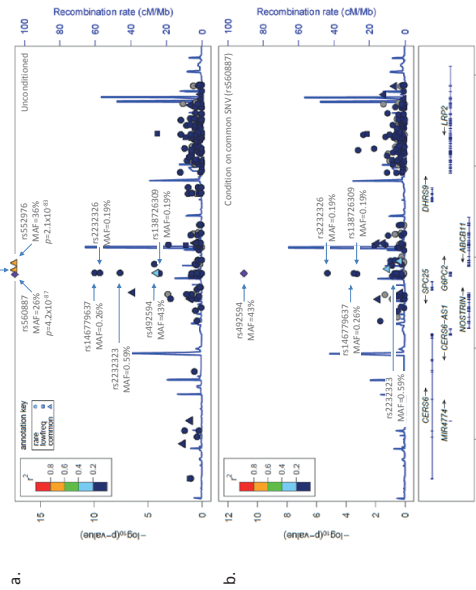
Figure 3. *G6PC2*

(a) Regional association results (-$\log_{10}p$) for fasting glucose of the *G6PC2* locus on chromosome 2. Minor allele frequencies (MAF) of common and rare *G6PC2* SNVs from single variant analyses are shown. *P* values for rs560887, rs563694 and rs552976 were artificially trimmed for the figure. Linkage disequilibrium (r^2) indicated by color scale legend. Y-axis scaled to show associations for variant rs560887 (purple dot, MAF=43%, p=4.2x10^{-87}). Triangle symbols indicate variants with MAF>5%, square symbols indicate variants with MAF 1-5%, and circle symbols indicate variants with MAF <1%.

(b) Regional association results (-$\log_{10}p$) for fasting glucose conditioned on rs560887 of *G6PC2*. After adjustment for rs560887, both rare SNVs rs2232326 (S324P) and rs146779637 (R283X), and common SNV rs492594 remain significantly associated with FG indicating the presence of multiple independent associations with FG at the *G6PC2* locus.

(c) Inset of *G6PC2* gene with depiction of exon locations, amino acid substitutions, and MAFs of the 15 SNVs included in gene-based analysis (MAF<1% and nonsynonymous, splice-site and gain/loss-of-function variation types as annotated by dbNSFPv2.0).

(d) The contribution of each variant on significance and effect on the SKAT test when one variant is removed the test. Gene-based SKAT p-values (blue line) and test statistic (red line) of *G6PC2* after removing one SNV at a time and re-calculating the association.

(e) Haplotypes and haplotype association statistics and p-values generated from the 15 rare SNVs from gene-based analysis of *G6PC2* from 18 cohorts and listed in panel (c). Global haplotype association, p=1.1x10^{-17}. Haplotypes ordered by decreasing frequency with haplotype 1 as the reference. Orange highlighting indicates the minor allele of the SNV on the haplotype.

of these SNVs also showed association with FG of larger effect size in unconditional single variant analyses (Supplementary Data 5), consistent with a recent report in which H177Y was associated with lower FG levels in Finnish cohorts[29]. We developed a novel haplotype meta-analysis method to examine the opposing direction of effects of each SNV. Meta-analysis of haplotypes with the 15 rare SNVs showed a significant global test of association with FG ($P_{\text{global test}}$ = 1.1x10^{-17}) (Supplementary Table 12), and supported the findings from the gene-based tests. Individual haplotype tests showed that the most significantly associated haplotypes were those carrying a single rare allele at R283X (P = 2.8x10^{-10}), S324P (P = 1.4x10^{-7}) or Y207S (P = 1.5x10^{-6}) compared to the most common haplotype. Addition of the known common intronic variant (rs560887) resulted in a stronger global haplotype association test (Pglobal test=1.5x10^{-81}), with the most strongly associated haplotype carrying the minor allele at rs560887 (Supplementary Table 13). Evaluation of regulatory annotation found that this intronic SNV is near the splice acceptor of intron 3 (RefSeq: NM_021176.2) and has been implicated in *G6PC2* pre-mRNA splicing[30]; it is also near the transcription start site of the expressed sequence tag (EST) DB031634, a potential cryptic minor isoform of *G6PC2* mRNA (Supplementary Fig. 7). No associations were observed in gene-based analysis of *G6PC2* with FI or T2D (Supplementary Tables 14 and 15).

Further characterization of exonic variation in *G6PC2* by exome sequencing in up to 7,452 individuals identified 68 SNVs (Supplementary Table 5), of which 4 were individually associated with FG levels and are on the exome chip (H177Y, MAF = 0.3%, P = 9.6 x10^{-5};

R283X, MAF = 0.2%, $P = 8.4 \times 10^{-3}$; S324P, MAF = 0.1%, $P = 1.7 \times 10^{-2}$; rs560887, intronic, MAF = 40%; $P = 7 \times 10^{-9}$) (Supplementary Data 6). 36 SNVs met criteria for entering into gene-based analyses (each MAF < 1%). This combination of 36 coding variants was associated with FG (cumulative MAF = 2.7%, $P_{SKAT} = 1.4 \times 10^{-3}$, $P_{WST} = 5.4 \times 10^{-4}$, Supplementary Table 16). Ten of these SNVs had been included in the exome chip gene-based analyses. Analyses indicated that the 10 variants included on the exome chip data had a stronger association with FG ($P_{SKAT} = 1.3 \times 10^{-3}$, $P_{WST} = 3.2 \times 10^{-3}$ vs. $P_{SKAT} = 0.6$, $P_{WST} = 0.04$ using the 10 exome chip or the 26 variants not captured on the chip, respectively, Supplementary Table 16).

In agnostic pathway analysis applying MAGENTA (http://www.broadinstitute.org/mpg/magenta/) to all curated biological pathways in KEGG (http://www.genome.jp/kegg/), GO (http://www.geneontology.org), Reactome (http://www.reactome.org), Panther (http://www.pantherdb.org), Biocarta (http://www.biocarta.com), and Ingenuity (http://www.ingenuity.com/) databases, no pathways achieved our Bonferroni-corrected threshold for significance of $P < 1.6 \times 10^{-6}$ for gene set enrichment in either FI or FG datasets (Supplementary Tables 17 and 18). The pathway P-values were further attenuated when loci known to be associated with either trait were excluded from the analysis. Similarly, even after narrowing the MAGENTA analysis to gene sets in curated databases with names suggestive of roles in glucose, insulin, or broader metabolic pathways, we did not identify any pathways that met our Bonferroni-corrected threshold for significance of $P < 2 \times 10^{-4}$ (Supplementary Table 19).

Due to the expected functional effects of protein-altering variants, we tested SNVs (4,513 for FG and 1,281 for FI) annotated as nonsynonymous, splice-site or stop gain/loss by dbNSFP[31] in genes within 500kb of known glycemic variants[1,27,32] for association with FG and FI to identify associated coding variants which may implicate causal genes at these loci (Supplementary Table 20). At the *DNLZ-GPSM1* locus, a common nsSNV (rs60980157; S391L) in the *GPSM1* gene was significantly associated with FG (Bonferroni corrected $P < 1.1 \times 10^{-5} = 0.05/4513$ SNVs for FG), and had previously been associated with insulinogenic index[9]. The *GPSM1* variant is common and in LD with the intronic index variant in the *DNLZ* gene (rs3829109) from previous FG GWAS[1] ($r^2_{EU} = 0.68$; 1000 Genomes EU). The association of rs3829109 with FG was previously identified using data from the Illumina CardioMetabochip, which poorly captured exonic variation in the region[1]. Our results implicate *GPSM1* as the most likely causal gene at this locus (Supplementary Fig. 8a). We also observed significant associations with FG for eight other potentially protein-altering variants in five known FG loci, implicating three genes (*SLC30A8*, *SLC2A2*, and *RREB1*) as potentially causal, but still undetermined for two loci (*MADD* and *IKBKAP*) (Supplementary Figs. 8b-f). At the *GRB14/COBLL1* locus, the known

GWAS[1,32] nsSNV rs7607980 in the *COBLL1* gene was significantly associated with FI (Bonferroni corrected $P < 3.9 \times 10^{-5} = 0.05/1281$ SNVs for FI), further suggesting *COBLL1* as the causal gene, despite prior functional evidence that *GRB14* may represent the causal gene at the locus[33] (Supplementary Fig. 8g).

Similarly, we performed analyses for loci previously identified by GWAS of T2D, but only focusing on the 412 protein-altering variants within the exonic coding region of the annotated gene(s) at 72 known T2D loci[2,34] on the exome chip. In combined ancestry analysis, three nsSNVs were associated with T2D (Bonferroni corrected p value threshold ($P < 0.05/412 = 1.3 \times 10^{-4}$) (Supplementary Data 8). At *WFS1*, *SLC30A8* and *KCNJ11*, the associated exome chip variants were all common and in LD with the index variant from previous T2D GWAS in our population (r^2_{EU}: 0.6-1.0; 1000 Genomes), indicating these coding variants might be the functional variants that were tagged by GWAS SNVs. In ancestry stratified analysis, three additional nsSNVs in *SLC30A8*, *ARAP1* and *GIPR* were significantly associated with T2D exclusively in African ancestry cohorts among the same 412 protein-altering variants (Supplementary Data 9), all with MAF > 0.5% in the African ancestry cohorts, but MAF < 0.02% in the European ancestry cohorts. The three nsSNVs were in incomplete LD with the index variants at each locus ($r^2_{AF} = 0$, $D'_{AF} = 1$; 1000 Genomes). SNV rs1552224 at *ARAP1* was recently shown to increase *ARAP1* mRNA expression in pancreatic islets[35] which further supports *ARAP1* as the causal gene underlying the common GWAS signal[36]. The association for nsSNV rs73317647 in *SLC30A8* (OR_{AF}[95%CI]: 0.45[0.31-0.65], $P_{AF} = 2.4 \times 10^{-5}$, $MAF_{AF} = 0.6\%$) is consistent with the recent report that rare or low frequency protein-altering variants at this locus are associated with protection against T2D[10]. The protein-coding effects of the identified variants indicate all five genes are excellent causal candidates for T2D risk. We did not observe any other single variant nor gene-based associations with T2D that met chip-wide Bonferroni significance thresholds ($P < 4.5 \times 10^{-7}$ and $P < 1.7 \times 10^{-6}$, respectively).

For the previous reported GWAS loci we tested the known FG and FI SNVs on the exome chip. Overall, 34 of the 38 known FG GWAS index SNVs and 17 of the 20 known FI GWAS SNVs (or proxies, $r^2 \geq 0.8$ 1000 Genomes) were present on the exome chip. 26 of the FG and 15 of the FI SNVs met the threshold for significance ($P_{FG} < 1.5 \times 10^{-3}$ (0.05/34 FG SNVs), $P_{FI} < 2.9 \times 10^{-3}$ (0.05/17 FI SNVs)) and were in the direction consistent with previous GWAS publications. In total, the direction of effect was consistent with previous GWAS publications for 33 of the 34 FG SNVs and for 16 of the 17 FI SNVs (binomial probability: $P_{FG} = 2.0 \times 10^{-9}$, $P_{FI} = 1.4 \times 10^{-4}$, Supplementary Data 10). Of the known 72 T2D susceptibility loci, we identified 59 index variants (or proxies $r^2 \geq 0.8$ 1000 Genomes) on the exome chip; 57 were in the direction consistent with previous publications (binomial probability: $P = 3.1 \times 10^{-15}$, see Supplementary Data 11). Additionally, two of the known MODY vari-

ants were on the exome chip. Only HNF4A showed nominal significance with FG levels (rs139591750, $P = 3 \times 10^{-3}$, Supplementary Table 21).

DISCUSSION

Our large-scale exome chip-wide analyses identified a novel association of a low frequency coding variant in *GLP1R* with FG and T2D. The minor allele, which lowered FG and T2D risk, was associated with a lower early insulin response to a glucose challenge and higher 2-h glucose. While the effect size on FG is slightly larger than for most loci reported to date, our findings suggest that few low frequency variants have a very large effect on glycemic traits and further demonstrate the need for large sample sizes to identify associations of low frequency variation with complex traits. However, by directly genotyping low frequency coding variants that are poorly captured through imputation, we were able to identify particular genes likely to underlie previously identified associations. Using this approach, we implicate causal genes at 6 loci associated with FG and/or FI (*G6PC2*, *GPSM1*, *SLC2A2*, *SLC30A8*, *RREB1*, and *COBLL1*) and 5 with T2D (*ARAP1*, *GIPR*, *KCNJ11*, *SLC30A8* and *WFS1*). For example, via gene-based analyses, we identified 15 rare variants in *G6PC2* ($P_{SKAT} = 8.2 \times 10^{-18}$), which are independent of the common non-coding signals at this locus and implicate this gene as underlying previously identified associations. We also revealed non-coding variants whose putative functions in epigenetic and post-transcriptional regulation of *ABO* and *G6PC2* are supported by experimental ENCODE Consortium, GTEx and transcriptome data from islets and for which future focused investigations using human cell culture and animal models will be needed to clarify their functional influence on glycemic regulation.

The seemingly paradoxical observation that the minor allele at *GLP1R* is associated with opposite effects on FG and 2-h glucose is not unique to this locus, and is also observed at the *GIPR* locus, which encodes the receptor for gastric inhibitory peptide (GIP), the other major incretin hormone. However, for *GLP1R*, we observe that the FG-lowering allele is associated with lower risk of T2D, while at *GIPR*, the FG-lowering allele is associated with higher risk of T2D (and higher 2-h glucose)[1]. The observation that variation in both major incretin receptors is associated with opposite effects on FG and 2-h glucose is a finding whose functional elucidation will yield new insights into incretin biology. An example where apparently paradoxical findings prompted cellular physiologic experimentation that yielded new knowledge is the *GCKR* variant P446L associated with opposing effects on FG and triglycerides[37,38]. The *GCKR* variant was found to increase active cytosolic GCK, promoting glycolysis and hepatic glucose uptake while increasing substrate for lipid synthesis[39,40].

Two studies have characterized the *GLP1R* A316T variant in vitro. The first study found no effect of this variant on cAMP response to full length GLP-1 or exendin-4 (endogenous and exogenous agonists)[41]. The second study corroborated these findings, but documented as much as 75% reduced cell surface expression of T316 compared to wild-type, with no alteration in agonist binding affinity. While this reduced expression had little impact on agonist-induced cAMP response or ERK1/2 activation, receptors with T316 had greatly reduced intracellular calcium mobilization in response to GLP-1(7-36NH$_2$) and exendin-4[42]. Given that GLP-1 induced calcium mobilization is a key factor in the incretin response, the in vitro functional data on T316 is consistent with the reduced early insulin response we observed for this variant, further supported by the Glp1r knockout mouse, which shows lower early insulin secretion relative to wild type mice43.

The associations of *GLP1R* variation with lower FG and T2D risk are more challenging to explain, and highlight the diverse and complex roles of *GLP1R* in glycemic regulation. While future experiments will be needed, here we offer the following hypothesis. Given fasting hyperglycemia observed in Glp1r knockout mice[43], A316T may be a gain-of-function allele that activates the receptor in a constitutive fashion, causing beta cells to secrete insulin at a lower ambient glucose level, thereby maintaining a lower FG; this could in turn cause down-regulation of GLP1 receptors over time, causing incretin resistance and a higher 2-h glucose after an oral carbohydrate load. Other variants in G protein-coupled receptors central to endocrine function such as the TSH receptor (TSHR), often in the transmembrane domains[44] (like A316T, which is in a transmembrane helix (TM5) of the receptor peptide), have been associated with increased constitutive activity alongside reduced cell surface expression[45,46], but blunted or lost ligand-dependent signaling[46,47].

The association of variation in *GLP1R* with FG and T2D represents another instance wherein genetic epidemiology has identified a gene that codes for a direct drug target in T2D therapy (incretin mimetics), other examples including *ABCC8/KCNJ11* (encoding the targets of sulfonylureas) and *PPARG* (encoding the target of thiazolidinediones). In these examples, the drug preceded the genetic discovery. Today, there are over 100 loci showing association with T2D and glycemic traits. Given that at least three of these loci code for potent antihyperglycemic targets, these genetic discoveries represent a promising long-term source of potential targets for future diabetes therapies.

In conclusion, our study has shown the use of analyzing the variants present on the exome chip, followed-up with exome sequencing, regulatory annotation and additional phenotypic characterization, in revealing novel genetic effects on glycemic homeostasis and has extended the allelic and functional spectrum of genetic variation underlying diabetes-related quantitative traits and T2D susceptibility.

ACKNOWLEDGMENTS

REFERENCES

1. Scott, R.A. et al. Large-scale association analyses identify new loci influencing glycemic traits and provide insight into the underlying biological pathways. Nat Genet 44, 991-1005 (2012).
2. Genome-wide trans-ancestry meta-analysis provides insight into the genetic architecture of type 2 diabetes susceptibility. Nat Genet 46, 234-44 (2014).
3. Nelson, M.R. et al. An abundance of rare functional variants in 202 drug target genes sequenced in 14,002 people. Science 337, 100-4 (2012).
4. Keinan, A. & Clark, A.G. Recent explosive human population growth has resulted in an excess of rare genetic variants. Science 336, 740-3 (2012).
5. Tennessen, J.A. et al. Evolution and functional impact of rare coding variation from deep sequencing of human exomes. Science 337, 64-9 (2012).
6. Fu, W. et al. Analysis of 6,515 exomes reveals the recent origin of most human protein-coding variants. Nature 493, 216-20 (2013).
7. Morrison, A.C. et al. Whole-genome sequence-based analysis of high-density lipoprotein cholesterol. Nature genetics 45, 899-901 (2013).
8. Peloso, G.M. et al. Association of low-frequency and rare coding-sequence variants with blood lipids and coronary heart disease in 56,000 whites and blacks. Am J Hum Genet 94, 223-32 (2014).
9. Huyghe, J.R. et al. Exome array analysis identifies new loci and low-frequency variants influencing insulin processing and secretion. Nat Genet 45, 197-201 (2013).
10. Flannick, J. et al. Loss-of-function mutations in SLC30A8 protect against type 2 diabetes. Nat Genet 46, 357-63 (2014).
11. Zuk, O. et al. Searching for missing heritability: designing rare variant association studies. Proc Natl Acad Sci U S A 111, E455-64 (2014).
12. Psaty, B.M. et al. Cohorts for Heart and Aging Research in Genomic Epidemiology (CHARGE) Consortium: Design of prospective meta-analyses of genome-wide association studies from 5 cohorts. Circ Cardiovasc Genet 2, 73-80 (2009).
13. Grove, M.L. et al. Best practices and joint calling of the HumanExome BeadChip: the CHARGE Consortium. PLoS One 8, e68095 (2013).
14. Bernstein, B.E. et al. An integrated encyclopedia of DNA elements in the human genome. Nature 489, 57-74 (2012).
15. Rosenbloom, K.R. et al. ENCODE data in the UCSC Genome Browser: year 5 update. Nucleic Acids Res 41, D56-63 (2013).
16. The Genotype-Tissue Expression (GTEx) project. Nat Genet 45, 580-5 (2013).
17. Drucker, D.J. & Nauck, M.A. The incretin system: glucagon-like peptide-1 receptor agonists and dipeptidyl peptidase-4 inhibitors in type 2 diabetes. Lancet 368, 1696-705 (2006).
18. Garber, A.J. Incretin therapy--present and future. Rev Diabet Stud 8, 307-22 (2011).
19. Seltzer, H.S., Allen, E.W., Herron, A.L., Jr. & Brennan, M.T. Insulin secretion in response to glycemic stimulus: relation of delayed initial release to carbohydrate intolerance in mild diabetes mellitus. J Clin Invest 46, 323-35 (1967).
20. Dailey, M.J. & Moran, T.H. Glucagon-like peptide 1 and appetite. Trends Endocrinol Metab 24, 85-91 (2013).
21. Astrup, A. et al. Safety, tolerability and sustained weight loss over 2 years with the once-daily human GLP-1 analog, liraglutide. Int J Obes (Lond) 36, 843-54 (2012).

22. Kirkpatrick, A., Heo, J., Abrol, R. & Goddard, W.A., 3rd. Predicted structure of agonist-bound glucagon-like peptide 1 receptor, a class B G protein-coupled receptor. Proc Natl Acad Sci U S A 109, 19988-93 (2012).

23. Olsson, M.L. & Chester, M.A. Polymorphism and recombination events at the ABO locus: a major challenge for genomic ABO blood grouping strategies. Transfus Med 11, 295-313 (2001).

24. Schunkert, H. et al. Large-scale association analysis identifies 13 new susceptibility loci for coronary artery disease. Nat Genet 43, 333-8 (2011).

25. Teslovich, T.M. et al. Biological, clinical and population relevance of 95 loci for blood lipids. Nature 466, 707-13 (2010).

26. Keembiyehetty, C. et al. Mouse glucose transporter 9 splice variants are expressed in adult liver and kidney and are up-regulated in diabetes. Mol Endocrinol 20, 686-97 (2006).

27. Dupuis, J. et al. New genetic loci implicated in fasting glucose homeostasis and their impact on type 2 diabetes risk. Nat Genet 42, 105-16 (2010).

28. Chen, W.M. et al. Variations in the G6PC2/ABCB11 genomic region are associated with fasting glucose levels. J Clin Invest 118, 2620-8 (2008).

29. Service, S.K. et al. Re-sequencing expands our understanding of the phenotypic impact of variants at GWAS loci. PLoS Genet 10, e1004147 (2014).

30. Baerenwald, D.A. et al. Multiple functional polymorphisms in the G6PC2 gene contribute to the association with higher fasting plasma glucose levels. Diabetologia 56, 1306-16 (2013).

31. Liu, X., Jian, X. & Boerwinkle, E. dbNSFP v2.0: a database of human non-synonymous SNVs and their functional predictions and annotations. Hum Mutat 34, E2393-402 (2013).

32. Manning, A.K. et al. A genome-wide approach accounting for body mass index identifies genetic variants influencing fasting glycemic traits and insulin resistance. Nat Genet 44, 659-69 (2012).

33. Hemming, R. et al. Human growth factor receptor bound 14 binds the activated insulin receptor and alters the insulin-stimulated tyrosine phosphorylation levels of multiple proteins. Biochem Cell Biol 79, 21-32 (2001).

34. Morris, A.P. et al. Large-scale association analysis provides insights into the genetic architecture and pathophysiology of type 2 diabetes. Nat Genet 44, 981-90 (2012).

35. Kulzer, J.R. et al. A common functional regulatory variant at a type 2 diabetes locus upregulates ARAP1 expression in the pancreatic beta cell. Am J Hum Genet 94, 186-97 (2014).

36. Voight, B.F. et al. Twelve type 2 diabetes susceptibility loci identified through large-scale association analysis. Nat Genet 42, 579-89 (2010).

37. Diabetes Genetics Initiative of Broad Institute of Harvard and MIT, L.U.a.N.I.o.B.R. et al. Genome-wide association analysis identifies loci for type 2 diabetes and triglyceride levels. Science 316, 1331-6 (2007).

38. Orho-Melander, M. et al. Common missense variant in the glucokinase regulatory protein gene is associated with increased plasma triglyceride and C-reactive protein but lower fasting glucose concentrations. Diabetes 57, 3112-21 (2008).

39. Rees, M.G. et al. Cellular characterisation of the GCKR P446L variant associated with type 2 diabetes risk. Diabetologia 55, 114-22 (2012).

40. Beer, N.L. et al. The P446L variant in GCKR associated with fasting plasma glucose and triglyceride levels exerts its effect through increased glucokinase activity in liver. Hum Mol Genet 18, 4081-8 (2009).

41. Fortin, J.P., Schroeder, J.C., Zhu, Y., Beinborn, M. & Kopin, A.S. Pharmacological characterization of human incretin receptor missense variants. J Pharmacol Exp Ther 332, 274-80 (2010).

42. Koole, C. et al. Polymorphism and ligand dependent changes in human glucagon-like peptide-1 receptor (GLP-1R) function: allosteric rescue of loss of function mutation. Mol Pharmacol 80, 486-97 (2011).

43. Scrocchi, L.A. et al. Glucose intolerance but normal satiety in mice with a null mutation in the glucagon-like peptide 1 receptor gene. Nat Med 2, 1254-8 (1996).

44. Gozu, H.I., Lublinghoff, J., Bircan, R. & Paschke, R. Genetics and phenomics of inherited and sporadic non-autoimmune hyperthyroidism. Molecular and cellular endocrinology 322, 125-34 (2010).

45. Vassart, G. & Costagliola, S. G protein-coupled receptors: mutations and endocrine diseases. Nat Rev Endocrinol 7, 362-72 (2011).

46. Van Sande, J. et al. Somatic and germline mutations of the TSH receptor gene in thyroid diseases. J Clin Endocrinol Metab 80, 2577-85 (1995).

47. Tonacchera, M. et al. Functional characteristics of three new germline mutations of the thyrotropin receptor gene causing autosomal dominant toxic thyroid hyperplasia. J Clin Endocrinol Metab 81, 547-54 (1996).

48. Goldstein, J.I. et al. zCall: a rare variant caller for array-based genotyping: genetics and population analysis. Bioinformatics 28, 2543-5 (2012).

49. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics 25, 1754-60 (2009).

50. Li, H. et al. The Sequence Alignment/Map format and SAMtools. Bioinformatics 25, 2078-9 (2009).

51. Brouwer, R.W., van den Hout, M.C., Grosveld, F.G. & van Ijcken, W.F. NARWHAL, a primary analysis pipeline for NGS data. Bioinformatics 28, 284-5 (2012).

52. Li, R., Li, Y., Kristiansen, K. & Wang, J. SOAP: short oligonucleotide alignment program. Bioinformatics 24, 713-4 (2008).

53. DePristo, M.A. et al. A framework for variation discovery and genotyping using next-generation DNA sequencing data. Nat Genet 43, 491-8 (2011).

54. Challis, D. et al. An integrative variant analysis suite for whole exome next-generation sequencing data. BMC Bioinformatics 13, 8 (2012).

55. Danecek, P. et al. The variant call format and VCFtools. Bioinformatics 27, 2156-8 (2011).

56. Li, R. et al. SNP detection for massively parallel whole-genome resequencing. Genome Res 19, 1124-32 (2009).

57. Lange, L.A. et al. Whole-exome sequencing identifies rare and low-frequency coding variants associated with LDL cholesterol. Am J Hum Genet 94, 233-45 (2014).

58. Saxena, R. et al. Genetic variation in GIPR influences the glucose and insulin responses to an oral glucose challenge. Nat Genet 42, 142-8 (2010).

59. Matthews, J.N., Altman, D.G., Campbell, M.J. & Royston, P. Analysis of serial measurements in medical research. BMJ 300, 230-5 (1990).

60. Rolfe Ede, L. et al. Association between birth weight and visceral fat in adults. Am J Clin Nutr 92, 347-52 (2010).

61. Forouhi, N.G., Luan, J., Hennings, S. & Wareham, N.J. Incidence of Type 2 diabetes in England and its association with baseline impaired fasting glucose: the Ely study 1990-2000. Diabet Med 24, 200-7 (2007).

62. Hills, S.A. et al. The EGIR-RISC STUDY (The European group for the study of insulin resistance: relationship between insulin sensitivity and cardiovascular disease risk): I. Methodology and objectives. Diabetologia 47, 566-70 (2004).

63.   Voorman, A., Brody, J., Chen, H. & Lumley, T. seqMeta: An R package for meta-analyzing region-based tests of rare DNA variants. R package version 1.3. (2013).

64.   Holmen, O.L. et al. Systematic evaluation of coding variation identifies a candidate causal variant in TM6SF2 influencing total cholesterol and myocardial infarction risk. Nat Genet 46, 345-51 (2014).

65.   Zaykin, D.V. et al. Testing association of statistically inferred haplotypes with discrete and continuous traits in samples of unrelated individuals. Hum Hered 53, 79-91 (2002).

66.   Becker, B.J. & Wu, M.J. The synthesis of regression slopes in meta-analysis. Statistical Science 22, 414-429 (2007).

67.   Segre, A.V., Groop, L., Mootha, V.K., Daly, M.J. & Altshuler, D. Common inherited variation in mitochondrial genes is not enriched for associations with type 2 diabetes or related glycemic traits. PLoS Genet 6(2010).

68.   Brooks, B.R. et al. CHARMM: the biomolecular simulation program. J Comput Chem 30, 1545-614 (2009).

69.   Phillips, J.C. et al. Scalable molecular dynamics with NAMD. J Comput Chem 26, 1781-802 (2005).

70.   Karolchik, D., Hinrichs, A.S. & Kent, W.J. The UCSC Genome Browser. Curr Protoc Bioinformatics Chapter 1, Unit 1.4 (2012).

Chapter 3

The role of lipid-altering gene variants in dyslipidemia and cardiovascular disease

3.1

The role of common lipid-altering gene variants in the risk of dyslipidemia through old age

ABSTRACT

Objective: In recent decades, there has been major progress in elucidating the genetics of lipid metabolism. Currently, genetic screening targets early life and is limited to rare Mendelian forms of dyslipidemia. A question that remains is the extent to which common variants can be effective in the identification of individuals at increased risk of dyslipidemia in the general population across the age range, including old age.

Methods: A risk score was computed for each individual from total cholesterol (TC) altering single nucleotide polymorphisms (SNPs) in the Rotterdam Study (n=10,072) and Erasmus Rucphen Family Study (n=2,715). Association of the risk score with prevalent dyslipidemia was analyzed using regression models. In the Rotterdam Study, Kaplan Meier survival analyses were performed to assess age-specific penetrance of incident dyslipidemia stratified by TC gene risk score quartiles. To test the ability of the risk scores to predict incident dyslipidemia, areas under the receiver operating characteristic curves (AUCs) were calculated.

Results: TC gene risk score quartiles were strongly associated with dyslipidemia. Overall, odds ratios increased from 1.61[1.44-1.80] in quartile 2 to 3.55[3.18-3.97] in the highest quartile. In normal weight and overweight individuals, age-specific penetrance of dyslipidemia increased per risk score quartile. The TC gene risk score discriminated incident dyslipidemia significantly better (AUC=0.61[0.58-0.64]) than a model including age, sex and BMI (AUC=0.53[0.50-0.56]) and combining both models did not improve the AUC compared to the genetic risk score alone.

Conclusions: Our results suggest that common genetic variants play a strong role in determining the development of dyslipidemia throughout the age range. These findings suggest that common genetic variants can be effective in the identification of individuals at increased risk of dyslipidemia in the general population across ages, including old age.

INTRODUCTION

Over the past decade, there has been major progress in elucidating the genetics of lipid metabolism. Genome-wide association studies (GWAS) have identified many common genetic variants contributing to inter-individual differences in circulating lipid levels in the general population[1-4]. Two large meta-analyses performed by the Global Lipid Genetics Consortium (GLGC) reported 157 genetic loci containing common single nucleotide polymorphisms associated with circulating levels of total cholesterol (TC), low-density lipoprotein cholesterol (LDL-C), high-density lipoprotein cholesterol (HDL-C) or triglycerides (TG)[5,6]. The currently known common TC variants jointly explain ~15% of the variance in TC. In addition to these common polymorphisms, large numbers of rare variants underlying familial hypercholesterolemia (FH), an autosomal co-dominant genetic disorder associated with increased levels of LDL-C, have been identified in patients with extreme hypercholesterolemia[7].

In a number of countries, including the Netherlands, genetic screening of families with a family history of FH is performed to diagnose FH at a young age[7]. Comprehensive genetic testing of rare variants with large effects is useful in terms of sensitivity, quality adjusted life years and cost-effectiveness[8]. Thus far, genetic screening targets early life and is limited to rare Mendelian forms of dyslipidemia. In contrast to the rapid translation from discovery to implementation in the clinic of the rare variants, common variants have not been taken to prevention or clinical care despite the fact that the currently discovered lipid loci jointly explain substantial percentages of the variance in lipid traits[5,6]. Whereas the effects of the rare Mendelian variants are seen at early age[7,9], the common variants were identified in a heterogeneous population spanning the full age range. A question that remains is the extent to which common variants can be effective in the identification of individuals at increased risk of dyslipidemia in the general population over all ages.

In this study, we assessed the combined effect of the 75 genetic loci associated with TC levels from the GLGC meta-analyses on dyslipidemia risk[5,6]. We calculated a genetic risk score based on the top SNPs from these 75 loci and tested association of this risk score with prevalent dyslipidemia; analyzed age-specific penetrance of incident dyslipidemia stratified by risk score quartiles; and assessed the discriminative ability of the risk score for incident dyslipidemia. To evaluate the added value of the genetic risk score compared to age, sex and BMI, we analyzed the discriminative ability of the genetic risk score, a score with these non-genetic factors and a combination of both.

MATERIALS AND METHODS

Study Populations

This study was embedded in the Rotterdam Study and the Erasmus Rucphen Family Study (ERF)[10-12]. The Rotterdam Study is a single-center prospective cohort study comprised of elderly individuals living in Ommoord, a district in the city of Rotterdam, the Netherlands. Participants were recruited into the study at three points in time: 1990 – 1993 (RS-I, n = 7,983, age at baseline ≥ 55 years), 2000 – 2001 (RS-II, n = 3,011, age at baseline ≥ 55 years) and 2006 – 2008 (RS-III, n = 3,932, age at baseline ≥ 45 years). ERF is a family-based cohort study composed of individuals living in a contiguous geographic region in the southwest of the Netherlands. All living descendants aged 18 years or above, of twenty-two couples that had a large number of children baptized in the community church between 1850 and 1900 were invited to participate in the study. Their spouses were invited as well. Approximately 3,200 individuals participated. Examinations took place between June 2002 and February 2005. Participants in both the Rotterdam Study and ERF filled out questionnaires and underwent extensive interviews and examinations at dedicated research centers. In the current analyses, 10,072 individuals from RS-I (n = 5,866), RS-II (n = 2,134) and RS-III (n = 2,072) for whom both clinical and genotypic information was available were included. In ERF, 2,715 individuals for whom clinical and genotypic information was available were included. Participants from both studies, or their legal guardians, provided written informed consent. Both studies were approved by the Medical Ethics board of the Erasmus Medical Center Rotterdam, the Netherlands.

Clinical and laboratory assessment

For both the Rotterdam Study and ERF, a broad range of examinations were conducted according to a standardized research protocol. At the research centers, height and weight were assessed and from these BMI was defined as weight in kilograms divided by the square of height in meters. In both studies, venous blood samples were obtained from study participants. In the RS-I cohort, TC was measured using enzymatic colorimetric methods (Kone Specific Analyzer, Kone Instruments). In RS-II and RS-III, TC was measured using comparable enzymatic procedures (Hitachi Analyzer, Roche Diagnostics). In ERF, TC was measured using a Synchron LX 20 Systems analyzer (Beckman Coulter, Fullerton, CA, USA). In both the Rotterdam Study and ERF, participants were asked to present the medications they used, including lipid-lowering medications, during their visit to the research center. Dyslipidemia was defined as TC > 6.5 mmol/L or use of lipid lowering medication13. In the Dutch cardiovascular risk management guidelines for general practitioners, this criterion is used to identify individuals at increased cardiovascular disease (CVD) risk. In RS-I and RS-II, incident dyslipidemia was studied, which was defined as

free of dyslipidemia at baseline and development of dyslipidemia during follow-up. This included three follow-up visits for RS-I and two follow-up visits for RS-II; the last follow-up visit for the cohorts took place between 2009-2011 and 2011-2012, respectively.

Genotyping

Genomic DNA was extracted from venous blood samples obtained at baseline in both the Rotterdam Study and ERF. DNA was extracted using the salting out method[14]. Genotyping in the Rotterdam Study was performed using the 550 and 610 K Illumina arrays and in the ERF cohort using Illumina 318 and 370K arrays. Exclusion criteria for individuals were excess autosomal heterozygosity, mismatches between called and phenotypic gender and outliers identified by an IBS clustering analysis. Single nucleotide polymorphisms (SNPs) were excluded for Hardy-Weinberg equilibrium P-value ≤ 10-6 or SNP call rate ≤ 98%. Genotypes with minor allele frequencies > 1% were used to impute about 2.5 million autosomal SNPs using HapMap CEU release 22 samples as a reference panel. Imputation was performed using MaCH15. Imputed genotypes were coded as dosages. These are values between 0 and 2 indicating the estimated number of copies of a given allele for each individual.

Genotype Scores

A genetic risk score for TC was calculated, per individual, based on the lead SNPs in the 75 loci from the large lipid GWAS meta-analyses by Teslovich et al. and Willer et al.[5,6]. The risk score was calculated per individual as:

$$\left(\sum_{i=1}^{n} (G_i * \beta_i) \right) / n$$

where n is the number of SNPs comprising the score, Gi is the number of TC increasing alleles at the ith genotype, and βi is the per allele effect estimate for the ith SNP as obtained in the GLGC lipid GWAS meta-analyses[5,6].

Statistical Analyses

TC outliers of more than four standard deviations were excluded from the analyses. The percentage of the heritability explained by the TC genes combined in the family-based ERF cohort was assessed by calculating the polygenic heritability as implemented in the SOLAR software package[16]. To assess the relationship of quartiles of the genetic risk score with prevalent dyslipidemia in the Rotterdam Study, logistic regression adjusting for age, sex and BMI was performed using R[17]. A variable indicating the sub-cohort (RS-I, RS-II or RS-III) was added to all models. In ERF, to account for relatedness in the family-based cohort, generalized estimating equations, as implemented in the R package gee, were used[18]. Results from the Rotterdam Study and ERF were combined using inverse

variance weighted random effects meta-analysis as implemented in the R package rmeta[19]. Age-specific penetrance of dyslipidemia in the Rotterdam Study, overall and in BMI subgroups, was estimated using Kaplan Meier survival analyses as implemented in the R package survival[20]. Age, sex and a variable indicating the sub-cohort were included as covariates in these analyses and individuals were censored at age 90 years because of the small number of individuals per TC gene risk score quartile above that age. Age-specific penetrance of dyslipidemia and mortality stratified by BMI group were assessed in the same manner as described above. The R package PredictABEL[21] was used to calculate the area under the receiver operating characteristic curves (AUC) for prediction of incident dyslipidemia in the RS-I and RS-II cohorts. The predictive ability of the genetic risk score, an epidemiological model including age, sex and BMI, and a combination of both was assessed. Results from RS-I and RS-II were combined using random effects meta-analysis as implemented in the R package rmeta[19].

RESULTS

A total of 2,715 ERF participants and 10,072 Rotterdam Study participants were included in the study. Descriptions of the cohorts' baseline characteristics are presented in Table 1. A total of 870 ERF participants (32.0%) and 4,641 Rotterdam Study participants (46.1%) were dyslipidemic upon inclusion, which included those receiving treatment and those with a total cholesterol > 6.5 mmol/L. As expected, the ERF participants were on average younger, spanning an age range from 17 to 86 years. In the RS-I and RS-II sub-cohorts, the elderly were oversampled (age range 55 – 99 years), offering the opportunity to obtain reliable estimates in the oldest people.

Table 1. Description of the study populations

	ERF	RS
n	2,715	10,072
Age (years)	48.9 (14.3)	65.7 (9.8)
Male (n, %)	1,216 (44.8)	4,280 (42.3)
BMI (kg/m^2)	26.9 (4.7)	26.8 (4.0)
TC (mmol/l)	5.6 (1.1)	6.2 (1.2)
LLT (n, %)	347 (12.8)	875 (8.7)
Prevalent dyslipidemia (n, %)	870 (32.0)	4,641 (46.1)
Incident dyslipidemia (n, %)*	NA	483 (6.0)

Mean (SD) unless otherwise indicated, ERF: Erasmus Rucphen Family Study, RS: Rotterdam Study, n: number, BMI: body mass index, TC: total cholesterol, LLT: lipid lowering therapy
*The percentage is calculated based on the 8000 individuals from the RS-I and RS-II cohorts, because follow-up data was only available for those cohorts.

Jointly, the TC genes explained 27.3% of the heritability of TC levels in ERF. The TC gene risk score was divided into quartiles and these were tested for association with dyslipidemia both in the total sample and stratified according to age (Table 2), using those in the lowest risk score quartile as reference. The risk score was strongly associated with dyslipidemia in the total sample in both the Rotterdam Study and ERF. The meta-analysis of the cohorts yielded odds ratios increasing from 1.61 [95% CI 1.44 – 1.80; $P = 2.2 \times 10^{-17}$] in quartile 2 to 2.26 [2.02 – 2.52; $P = 7.5 \times 10^{-48}$] in quartile 3 and 3.55 [3.18 – 3.97; $P = 4.4 \times 10^{-112}$] in quartile 4. A similar significant effect was seen in all age groups (Table 2), except for quartile 3 in the age 65-74 years subgroup ($P = 0.110$) and quartile 2 in the

Table 2. Association of TC genes risk score quartiles with prevalent dyslipidemia in the total sample and stratified according to age subgroups

Risk score quartile	ERF		RS		Meta	
	OR [95% CI]	*P*	OR [95% CI]	*P*	OR [95% CI]	*P*
	all (n=12,215)					
2	1.68 [1.25-2.26]	5.1×10^{-4}	1.60 [1.42-1.80]	9.4×10^{-15}	1.61 [1.44-1.80]	2.2×10^{-17}
3	2.55 [1.90-3.41]	3.3×10^{-10}	2.21 [1.97-2.49]	2.13×10^{-39}	2.26 [2.02-2.52]	7.5×10^{-48}
4	4.02 [3.04-5.33]	2.9×10^{-22}	3.47 [3.08-3.91]	8.6×10^{-92}	3.55 [3.18-3.97]	4.4×10^{-112}
	subgroup age < 45 years (n=949)					
2	1.36 [0.68-2.71]	0.380	NA	NA	NA	NA
3	3.00 [1.61,5.58]	5.1×10^{-4}	NA	NA	NA	NA
4	3.28 [1.79-6.01]	1.2×10^{-4}	NA	NA	NA	NA
	subgroup age 45-54 years (n=1,416)					
2	2.34 [1.29-4.22]	0.005	1.82 [1.16-2.85]	0.009	1.99 [1.39-2.85]	1.6×10^{-4}
3	3.80 [2.13-6.79]	6.2×10^{-6}	2.22 [1.43-3.45]	4.0×10^{-4}	2.81 [1.66-4.74]	1.2×10^{-4}
4	6.11 [3.43-10.88]	7.6×10^{-10}	2.43 [1.57-3.77]	7.3×10^{-5}	3.78 [1.53-9.32]	0.004
	subgroup age 55-64 years (n=5,212)					
2	1.85 [1.13-3.03]	0.014	1.71 [1.44-2.02]	8.7×10^{-10}	1.72 [1.47-2.02]	4.2×10^{-11}
3	2.64 [1.58-4.42]	2.1×10^{-4}	2.19 [1.84-2.60]	3.1×10^{-19}	2.23 [1.90-2.62]	3.7×10^{-22}
4	4.43 [2.62-7.49]	2.7×10^{-8}	3.65 [3.06-4.34]	2.7×10^{-48}	3.72 [3.15-4.38]	6.2×10^{-55}
	subgroup age 65-74 years (n=2,798)					
2	1.09 [0.53-2.22]	0.817	1.75 [1.39-2.21]	2.2×10^{-6}	1.56 [1.05-2.33]	0.027
3	1.13 [0.56-2.28]	0.731	2.33 [1.84-2.94]	1.1×10^{-12}	1.76 [0.88-3.50]	0.110
4	2.65 [1.29-5.44]	0.008	4.02 [3.16-5.11]	1.0×10^{-29}	3.77 [2.80-5.06]	1.3×10^{-18}
	subgroup age ≥ 75 years (n=1,840)					
2	0.76 [0.12-4.86]	0.768	1.12 [0.84-1.50]	0.423	1.11 [0.84-1.48]	0.455
3	1.33 [0.29-6.05]	0.713	2.21 [1.66-2.93]	4.3×10^{-8}	2.17 [1.64-2.87]	4.9×10^{-8}
4	2.13 [0.41-11.07]	0.367	2.93 [2.21-3.90]	1.2×10^{-13}	2.91 [2.20-3.85]	8.7×10^{-14}

ERF: Erasmus Rucphen Family Study, RS: Rotterdam Study, OR: odds ratio, CI: confidence interval, *P*: *P*-value, Q: quartile

A. ERF
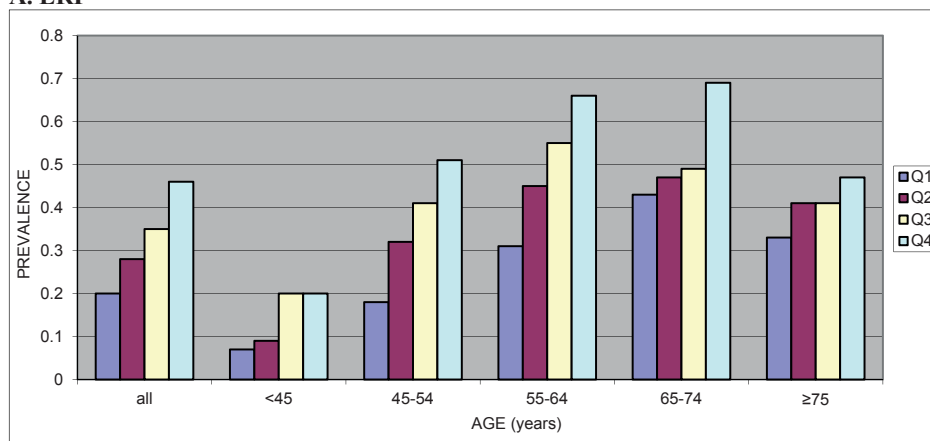
B. Rotterdam Study
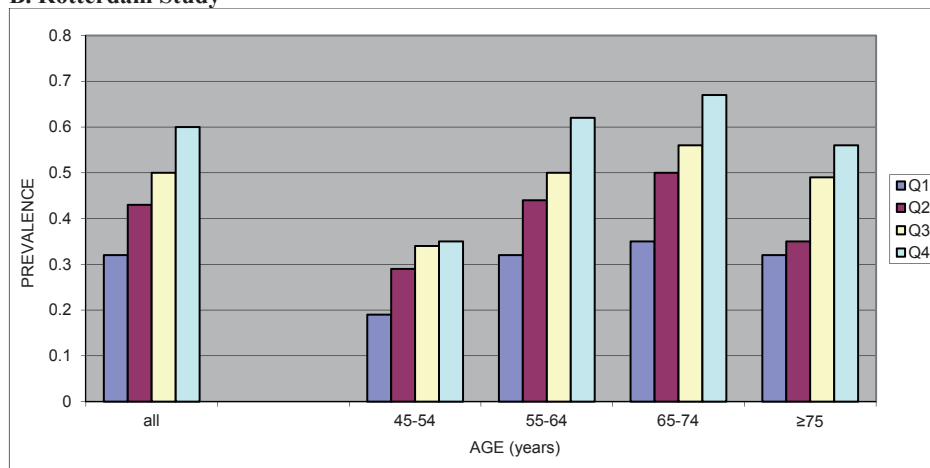
Figure 1. Prevalence of dyslipidemia in the total sample and in different age subgroups stratified according to TC genes risk score quartile
Q: TC gene risk score quartile

highest age subgroup ($\geq$ 75 years; $P = 0.455$). Figure 1 plots the prevalence of dyslipidemia according to TC gene risk score quartiles. The prevalence increased from 20% to 46% in the overall analyses in ERF and from 32% to 60% in the Rotterdam Study. P-values for trend across the risk score quartiles were significant for all age subgroups in both studies ($P = 1.3 \times 10^{-96} - 0.002$), except for the highest age subgroup in ERF ($P = 0.431$).

Data on incident dyslipidemia were available for the RS-I and RS-II cohorts. 483 of the 4,149 individuals in these cohorts who were free of dyslipidemia at baseline developed this outcome. Age-specific penetrance curves of incident dyslipidemia, stratified by TC
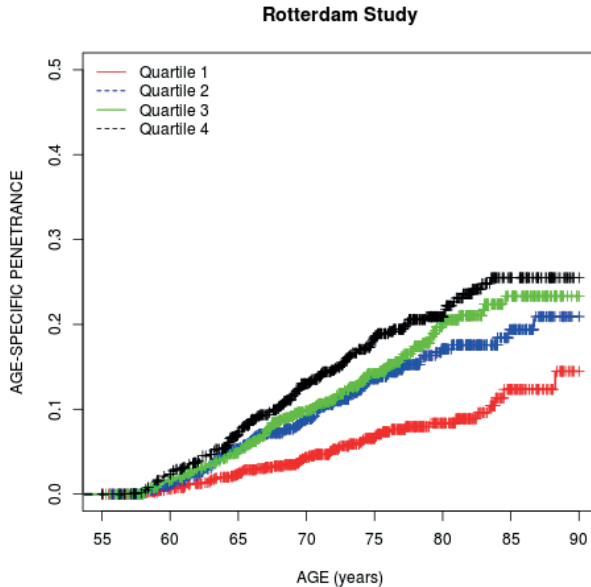
Figure 2. Age-specific penetrance of dyslipidemia stratified by TC gene risk score quartile

gene risk score are shown in Figure 2. The risk increased according to risk score quartile up until age 85 years. By this age, 12% of the individuals in the lowest risk score quartile and 25% in the highest quartile developed dyslipidemia. Of note is the low risk in the lowest quartile as compared to the other three quartiles until age 90. Figure 3 shows age-specific penetrance curves for incident dyslipidemia stratified by BMI (normal, over-weight and obese). There was not a consistent increased risk according to BMI. Although overweight individuals were at increased risk of dyslipidemia at all ages compared to those with a normal weight, obese participants were not at the highest risk compared to these two weight groups over all ages. Between the ages of 55 and 83 years, dyslip-idemia risk was similar in the obese and normal weight groups. After age 83 years, there was a steep increase in dyslipidemia risk in the obese group and, at age 90 years, the risk was highest in the obese compared to the normal weight and overweight individuals. In line with these findings, the TC gene risk score (meta-analysis AUC = 0.61 [0.58-0.64]) discriminated incident dyslipidemia significantly better than an epidemiological risk model including age, sex and BMI (AUC = 0.53 [0.50 – 0.56]) (Table 3). The discriminative ability of the TC gene risk score alone was similar to that of the risk score, age, sex and BMI combined (AUC = 0.61 [0.59 – 0.64]) suggesting that the addition of BMI to the ge-netic risk score has no additive value. Figure 4 shows that, at least in those with normal or overweight, the genetic risk score does have utility over BMI. In the normal weight group, by the age of 85 years, 8% of the individuals in the lowest TC gene risk score quartile and 25% in the highest quartile developed dyslipidemia. These percentages
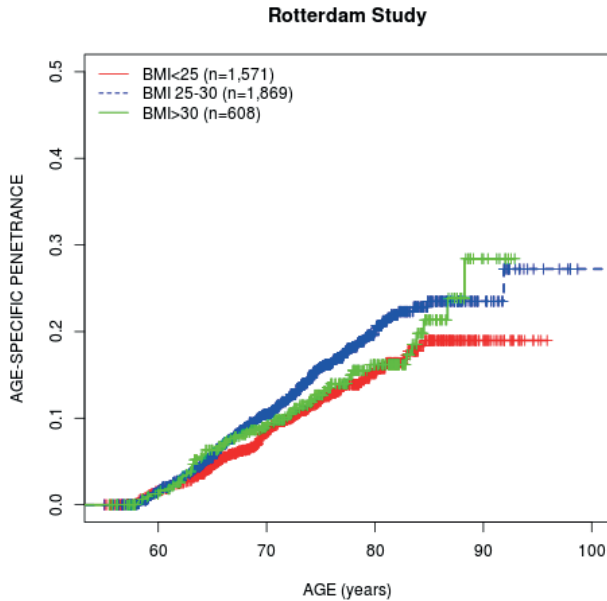
Figure 3. Age-specific penetrance of dyslipidemia stratified by BMI
BMI: body mass index

Table 3. Discriminative ability of clinical and genetic models for incident dyslipidemia

Predictor	AUC [95% CI]		
	RS-I	RS-II	Meta
Age, sex, BMI	0.52 [0.48-0.56]	0.54 [0.50-0.58]	0.53 [0.50-0.56]
TC gene risk score	0.60 [0.56-0.64]	0.62 [0.58-0.66]	0.61 [0.58-0.64]
TC gene risk score, age, sex, BMI	0.60 [0.56-0.64]	0.62 [0.59-0.66]	0.61 [0.59-0.64]

AUC: area under the receiver operating characteristic curve, CI: confidence interval, RS: Rotterdam Study,
BMI: body mass index, TC: total cholesterol

were 12% and 32%, respectively, in the overweight individuals. In obese individuals, the genetic risk score was not informative. Up to the age of 85 years, those in the lowest risk score quartile still seemed to have a protective effect, but being in either one of the other three quartiles was not informative. Above the age of 85 years, the genetic risk score was not informative in determining dyslipidemia risk in obese individuals.
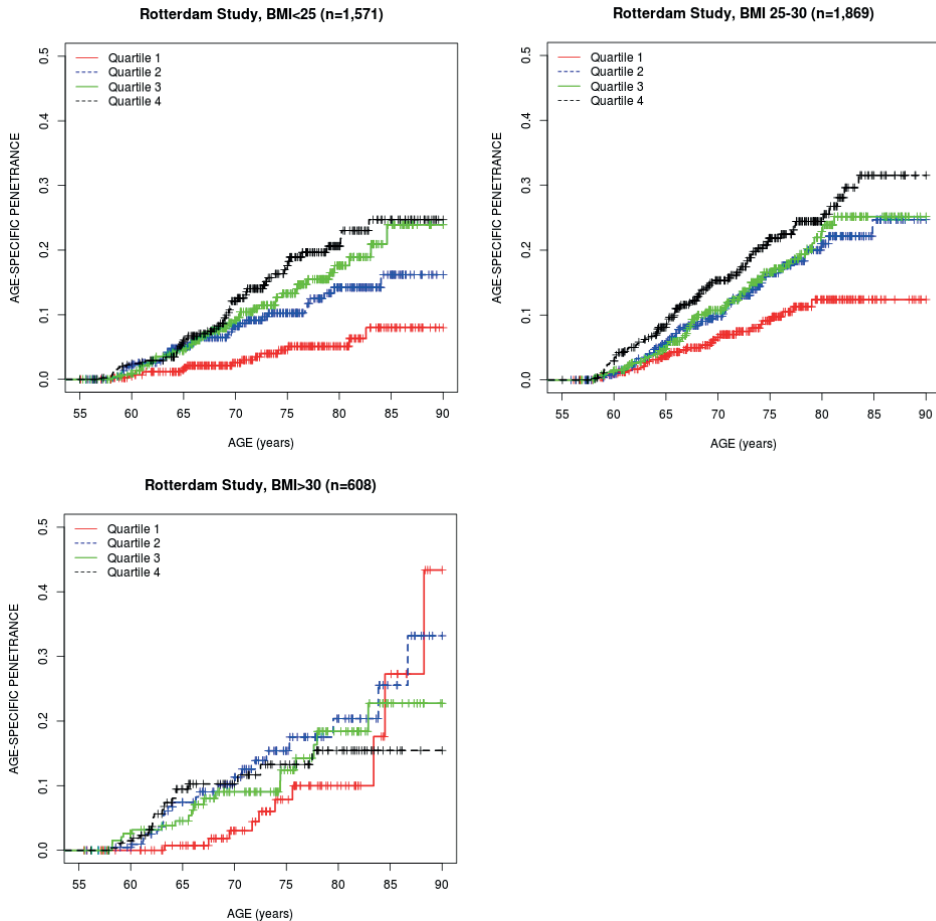
Figure 4. Age-specific penetrance of dyslipidemia in different BMI groups stratified by TC gene risk score quartile

TC: total cholesterol, BMI: body mass index

DISCUSSION

This study shows strong association of a TC gene risk score with dyslipidemia in the general population. Comparing those in the lowest and highest quartile of the genetic risk score, the prevalence of dyslipidemia doubled in two independent studies. This trend was seen over a wide age range. In the Rotterdam Study, the follow-up data showed that age-specific incidence of dyslipidemia increased through old age (90+ years). Further, in this study, age, sex and BMI did not improve the discrimination of the genetic model. In contrast, common genetic variants played an important role in determining

who develops dyslipidemia between the ages of 55 and 90 years in normal weight and overweight individuals.

Associations of genetic risk scores comprised of common lipid-altering gene variants with their corresponding lipid levels, extreme lipid values and intervention thresholds for blood lipids have been previously described[5,22,23,24]. To our knowledge, however, this is the first study extensively assessing to what extent all currently known common TC-altering gene variants can be effective in the identification of individuals at increased risk of dyslipidemia in the general population across ages, including old age, and to what extent the genetic risk score has additive utility over BMI and vice versa. The present study includes all 75 currently known common TC-altering gene variants, including recently discovered loci[6]. A further strength of our study is that it jointly analyses the data of two large population-based studies that were not selected on the basis of the phenotypes. Since both ERF and the Rotterdam Study were part of the discovery GWAS meta-analyses, a possible limitation of this study might be that the risk score is not completely independent from the GWAS results. Because of the very large number of individuals included in the discovery meta-analyses (>188,000), this effect should be limited. Despite the large difference in mean age between the Rotterdam Study and ERF, study heterogeneity was limited.

Our results show that common genetic variants play an important role in determining who develops dyslipidemia from the age of 55 years through age 90 in normal weight and overweight individuals, underscoring the value of the genome wide association studies for age-related diseases. Despite the small effects of the genes identified to date, when combined into a risk score the effects are substantial and discriminate future patients with dyslipidemia better than an established epidemiological risk factor such as BMI. Although we used the latest common variants identified, the genetic risk score is far from complete, explaining 27% of the heritability of TC levels. Further gene discovery efforts could improve the identification of those individuals likely to develop dyslipidemia.

The age-specific penetrance of dyslipidemia stratified by BMI and age-specific penetrance of dyslipidemia stratified by TC gene risk score in different BMI groups showed inconsistent patterns in those with obesity (BMI > 30 kg/m^2). Through 83 years of age, the risk of dyslipidemia for the obese participants of the Rotterdam Study is similar to that of normal weight individuals. There may be several explanations for these findings. First, selection may have taken place: those with obesity may have developed secondary dyslipidemia early as a consequence of dietary habits and resistance to insulin and therefore are in the analyses of prevalent but not of incident dyslipidemia or obese participants with comorbidities are least likely to participate in our study resulting in

a selection of relatively healthy obese participants. Second, there could be differential mortality in those with obesity. However, our results did not provide evidence for this explanation (Supplemental Figure 1).

A next step after finding that common genetic variants play an important role in determining who develops dyslipidemia through old age is to determine whether screening for these variants to identify individuals at increased risk of dyslipidemia in the general population would be useful. Genetic screening has proven useful in FH[8], but despite the large role of genes in both conditions, there are important differences between FH and dyslipidemia in the general population. FH is an autosomal co-dominant disorder with nearly complete penetrance[9], with a prevalence of 1:500 in most Western countries. Left untreated, men that are heterozygous for the mutation experience clinical symptoms of CVD typically in their fourth decade and women in their fifth decade of life[7]. Statin therapy lowers CVD risk substantially in these individuals[25]. Although the prevalence of dyslipidemia in the general population is high, even in individuals in the highest TC gene risk score quartile penetrance is not complete. Yet in this high risk group prevalence approaches 70% between the ages of 65 and 74 years. However, controversy exists about the effect of total cholesterol on total- and cardiovascular mortality in elderly people, questioning the validity of identifying individuals with dyslipidemia in high age groups[26,27]. To determine whether screening would be useful, and in which age groups, age-specific penetrance of dyslipidemia over a wider age range than assessed in our study should be investigated, including the clinical consequences. Corresponding morbidity and mortality might be prevented by early identification and treatment of individuals at high polygenic risk of dyslipidemia. Health economic evaluation is required, as extrapolations from the quickly deleterious FH to a late onset polygenic situation may not be accurate.

In conclusion, our results show a strong role of common genetic variants in determining who develops dyslipidemia throughout the age range. These findings suggest that common genetic variants can be effective in the identification of individuals at increased risk of dyslipidemia in the general population across ages, including old age.

ACKNOWLEDGMENTS

REFERENCES

1.   Willer, C.J. et al. Newly identified loci that influence lipid concentrations and risk of coronary artery disease. Nat Genet 40, 161-9 (2008).
2.   Kathiresan, S. et al. Six new loci associated with blood low-density lipoprotein cholesterol, high-density lipoprotein cholesterol or triglycerides in humans. Nat Genet 40, 189-97 (2008).
3.   Kathiresan, S. et al. Common variants at 30 loci contribute to polygenic dyslipidemia. Nat Genet 41, 56-65 (2009).
4.   Aulchenko, Y.S. et al. Loci influencing lipid levels and coronary heart disease risk in 16 European population cohorts. Nat Genet 41, 47-55 (2009).
5.   Teslovich, T.M. et al. Biological, clinical and population relevance of 95 loci for blood lipids. Nature 466, 707-13 (2010).
6.   Global Lipids Genetics Consortium, Willer, C.J. et al. Discovery and refinement of loci associated with lipid levels. Nat Genet 45, 1274-83 (2013).
7.   Hovingh, G.K., Davidson, M.H., Kastelein, J.J. & O'Connor, A.M. Diagnosis and treatment of familial hypercholesterolaemia. Eur Heart J 34, 962-71 (2013).
8.   Elucigene FH20 and LIPOchip for the diagnosis of familial hypercholesterolaemia (DG2). National Institute for Health and Clinical Excellence. http://guidance.nice.org.uk/DG2 (December 2013)
9.   Defesche, J.C. Defining the challenges of FH screening for familial hypercholesterolemia. J Clin Lipidol 4, 338-41 (2010).
10.   Hofman, A., Grobbee, D.E., de Jong, P.T. & van den Ouweland, F.A. Determinants of disease and disability in the elderly: the Rotterdam Elderly Study. Eur J Epidemiol 7, 403-22 (1991).
11.   Pardo, L.M., MacKay, I., Oostra, B., van Duijn, C.M. & Aulchenko, Y.S. The effect of genetic drift in a young genetically isolated population. Ann Hum Genet 69, 288-95 (2005).
12.   Hofman, A. et al. The Rotterdam Study: 2014 objectives and design update. Eur J Epidemiol 28, 889-926 (2013).
13.   NHG-Standaard cardiovasculair risicomanagement. https://www.nhg.org/standaarden/samenvatting/cardiovasculair-risicomanagement (December 2013)
14.   Miller, S.A., Dykes, D.D. & Polesky, H.F. A simple salting out procedure for extracting DNA from human nucleated cells. Nucleic Acids Res 16, 1215 (1988).
15.   Li, Y., Willer, C.J., Ding, J., Scheet, P. & Abecasis, G.R. MaCH: using sequence and genotype data to estimate haplotypes and unobserved genotypes. Genet Epidemiol 34, 816-34 (2010).
16.   Almasy, L. & Blangero, J. Multipoint quantitative-trait linkage analysis in general pedigrees. Am J Hum Genet 62, 1198-211 (1998).
17.   R Core Team (2013). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL http://www.R-project.org
18.   Carey VJ, Ported to R by Lumley T and Ripley B (2012). gee: Generalized Estimation Equation solver. R package version 4.13-18. http://CRAN.R-project.org/package=gee
19.   Lumley T (2009). rmeta: Meta-analysis. R package version 2.16. http://CRAN.R-project.org/package=rmeta.
20.   Therneau T (2013). A Package for Survival Analysis in S. R package version 2.37-4, http://CRAN.R-project.org/package=survival
21.   Kundu S, Aulchenko YS, Janssens AC (2011). PredictABEL: Assessment of risk prediction models. R package version 1.2. http://CRAN.R-project.org/package=PredictABEL
22.   Kathiresan, S. et al. Polymorphisms associated with cholesterol and risk of cardiovascular events. N Engl J Med 358, 1240-9 (2008).

23. Murray, A. et al. Common lipid-altering gene variants are associated with therapeutic intervention thresholds of lipid levels in older people. Eur Heart J 30, 1711-9 (2009).

24. Isaacs, A. et al. Risk scores of common genetic variants for lipid levels influence atherosclerosis and incident coronary heart disease. Arterioscler Thromb Vasc Biol 33, 2233-9 (2013).

25. Versmissen, J. et al. Efficacy of statins in familial hypercholesterolaemia: a long term cohort study. BMJ 337, a2423 (2008).

26. Weverling-Rijnsburger, A.W. et al. Total cholesterol and risk of mortality in the oldest old. Lancet 350, 1119-23 (1997).

27. Prospective Studies Collaboration, Lewington, S. et al. 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, 1829-39 (2007).

Risk scores of common genetic variants for lipid levels influence atherosclerosis and incident coronary heart disease

ABSTRACT

Objective: Circulating levels of total cholesterol, low-density lipoprotein cholesterol (LDL-C), high-density lipoprotein cholesterol, and triglycerides are recognized risk factors for cardiovascular disease. We tested the hypothesis that the cumulative effects of common genetic variants for lipids are collectively associated with subclinical atherosclerosis and incident coronary heart disease.

Methods and Results: Participants were drawn from the Erasmus Rucphen Family Study (n = 2,269) and the Rotterdam Study (n = 8,130). Linear regression and Cox proportional hazards models were applied to assess the influence of 4 risk scores derived from common genetic variants for lipids (total cholesterol, LDL-C, high-density lipoprotein cholesterol, and triglycerides) on carotid plaque, intima-media thickness, incident myocardial infarction, and coronary heart disease. Adjusted for age and sex, all 4 risk scores were associated with carotid plaque. This relationship was the strongest for the LDL-C score, which increased plaque score by 0.102 per SD increase in genetic risk score ($P = 3.2 \times 10^{-8}$). The LDL-C score was also nominally associated with intima-media thickness, which increased 0.006 mm per SD increase in score ($P = 0.05$). Both the total cholesterol and LDL-C scores were associated with incident myocardial infarction and coronary heart disease with hazard ratios between 1.10 and 1.13 per SD increase in score. Inclusion of additional risk factors as covariates minimally affected these results.

Conclusions: Common genetic variants with small effects on lipid levels are, in combination, significantly associated with subclinical and clinical cardiovascular outcomes. As knowledge of genetic variation increases, preclinical genetic screening tools might enhance the prediction and prevention of clinical events.

INTRODUCTION

Cardiovascular disease (CVD) is a leading cause of morbidity and mortality worldwide.[1] Increased serum levels of total cholesterol (TC) and low-density lipoprotein cholesterol (LDL-C) are among the most important risk factors for CVD. Increased levels of triglycerides (TG) and decreased levels of high-density lipoprotein cholesterol (HDL-C) are also associated with increased CVD risk.[2–4] Heritability estimates for lipid levels are moderate to high,[5,6] and genome-wide association studies (GWAS) have identified many novel genetic variants contributing to interindividual differences in circulating lipid levels.[7–10]

A meta-analysis performed by the Global Lipid Genetics Consortium (GLGC) of 46 lipid GWAS, involving measurements from >100 000 individuals, reported 95 genetic loci containing single nucleotide polymorphisms (SNPs) associated with one or more of the blood lipid measurements (TC, LDL-C, HDL-C, and TG) at a genome-wide significance level ($P < 5 \times 10^{-8}$).[11] LDL-C, HDL-C, and TG genetic risk scores, computed from the SNPs with the lowest P values in the loci associated with those phenotypes, successfully discriminated individuals with extreme lipid values from low lipid controls. Individuals in the top quartiles of the risk scores were 4× to 44× more likely to have extreme lipid values than individuals in the bottom quartiles.[11] For complex phenotypes, the classical theory of Fisher already predicted that multiple risk alleles, following a Gaussian distribution in the population, account for the phenotypic variability observed in the population.[12] This implies that susceptibility to complex diseases increases with the number of risk alleles a person carries.

The important contribution of lipid levels to CVD risk notwithstanding, only a limited number of the 95 loci were associated with coronary artery disease.[11] A possible explanation is the typically small effect sizes of these common genetic variants individually. To obtain further insight into the genetic architecture underlying CVD, and to obtain more insight into the possibility of detecting individuals at increased risk for CVD based on their genetic profile, we investigated the same genetic risk scores used in the GLGC publication[11] to test the hypothesis that the cumulative effects of such common genetic lipid variants are associated with subclinical atherosclerosis and incident coronary heart disease (CHD).

MATERIALS AND METHODS

Study Populations

Study participants were drawn from two previously described population-based cohorts, the Erasmus Rucphen Family Study (ERF)[13] and the Rotterdam Study (RS)[14,15]. Briefly, ERF is a family-based cohort ascertained from a contiguous geographic region in the southwest of the Netherlands. The aim of this study is to identify genetic risk factors for the development of complex disorders. In ERF, twenty-two families that had a minimum of five children baptized in the community church between 1850 and 1900 were identified with the help of detailed genealogical records. All living descendants of these couples, and their spouses, were invited to take part in the study. Comprehensive interviews, questionnaires, and examinations were completed at a research center in the area; approximately 3,200 individuals participated. Data collection started in June 2002 and was completed in February 2005. In the current analyses, 2,269 participants for whom phenotypic, genotypic and genealogical information was available were studied.

The Rotterdam Study is a single-center prospective cohort comprised of 10,994 individuals (7,983 from the RS1 cohort and 3,011 from the RS2 cohort) aged 55 and older at study entry. Baseline examinations took place between 1990 - 1993 for RS1 and 2000 – 2001 for RS2. Participants underwent extensive physical examinations and completed both questionnaires and interviews with trained research assistants. In the current analyses, 8,130 Rotterdam Study participants for whom phenotypic and genotypic information was available were studied.

All participants in both studies completed written informed consents and the Medical Ethics Committee at Erasmus University approved the protocols for the ascertainment and examination of human subjects.

Genotyping

Genomic DNA was extracted from whole blood samples drawn at the baseline examination, utilizing the salting out method[16]. Genotyping in the Rotterdam Study was performed using 550 and 610 K Illumina arrays; in the ERF cohort, genotypes were measured using Illumina 318 and 370K arrays. Individuals were excluded for excess autosomal heterozygosity, mismatches between called and phenotypic gender, and if there were outliers identified by an IBS clustering analysis. The exclusion criteria for SNPs were Hardy-Weinberg equilibrium (HWE) $P \leq 10^{-6}$ or SNP call rate $\leq 98\%$. After this quality control, measured genotypes which had minor allele frequencies > 1% were used to impute ~2.5 million autosomal SNPs with the CEU samples from HapMap release 22 (build 36) as a reference panel using MaCH[17]. Imputed genotypes were coded as dosages, fractional values between 0 and 2 reflecting the estimated number of copies

of a given allele for a given SNP for each individual. The use of dosages allows for the incorporation of imputation uncertainty in subsequent analysis.

Genotype scores

For each lipid outcome (TC, LDL-C, HDL-C, and TG), significant SNPs were drawn from Supplementary Table 2 in the GLGC paper[11]; effect estimates were extracted from the GLGC inverse-variance weighted meta-analysis results. Scores were calculated, per individual, as:

$$\left(\sum_{i=1}^{n} (G_i * \beta_i) \right) / n$$

where n is the number of SNPs contributing to the score for the jth individual, G_i is the number of lipid increasing alleles (lipid decreasing alleles for HDL) at the ith genotype, and β_i is the per allele effect estimate for the ith SNP. These scores were calculated individually for each population. To aid in the interpretation of analysis results, the scores were standardized to yield a mean of zero and a standard deviation of one (such that regression coefficients equal the change in value, or risk, per standard deviation increase, or decrease, in score).

Additionally, to aid in the interpretation of the results, risk scores were calculated for LDL-C, HDL-C and TG after the exclusion of SNPs that were associated with one or more of the other outcomes ("pure" risk scores). SNPs that were also associated with TC were not excluded from these scores, because of the large number of TC SNPs associated with the other outcomes and because SNPs with a large effect on LDL-C or HDL-C also influence the composite TC measurements.

Phenotype Determination

Venous blood samples were obtained from Rotterdam Study and ERF participants. In the RS1 cohort, TC and HDL-C were measured using enzymatic colorimetric methods (Kone Specific Analyzer, Kone Instruments). In RS2, TC, HDL-C and TG were measured using comparable enzymatic procedures (Hitachi Analyzer, Roche Diagnostics) and LDL-C was calculated using the Friedewald formula: LDL-C = TC - HDL-C - 0.45 * TG when TG ≤ 4.52 mmol/L. In ERF TC, HDL-C, LDL-C and TG were measured using a Synchron LX 20 Systems analyzer (Beckman Coulter, Fullerton, CA, USA). In both the Rotterdam Study and ERF, participants were asked to present the medications they used, including lipid-lowering medication.

In both studies, high-resolution B-mode ultrasonography of the left and right common carotid arteries was performed with a 7.5-MHz linear-array transducer (ATL UltraMark IV).

The maximum carotid intima-media thickness, summarized as the mean of the maximal measurements from the near and far walls on both the left and right sides, was used as an atherosclerosis measure for these analyses, as previously described[18]. As an additional atherosclerosis measure, carotid plaque score was determined by the number of sites (common carotid, internal carotid, and bifurcation on both the left and right sides) that showed visible focal widening relative to adjacent segments, with protrusion into the lumen composed of either only calcified deposits or a combination of calcified and non-calcified material, and scored from zero to six[19].

In the Rotterdam Study, follow-up data collection between baseline and January 1, 2007 included data on the incidence of myocardial infarction (MI) and coronary heart disease (a composite endpoint defined as occurrence of MI, heart failure, percutaneous transluminal coronary angioplasty, or coronary artery bypass graft). Information on fatal and non-fatal events was obtained from general practitioners through a computerized reporting system. Two research physicians examined the patients' medical records and verified events. When these physicians disagreed, a medical expert in the field determined the diagnosis. In the case of multiple events, the first event was used for this analysis. Incident MI and CHD were only studied in the Rotterdam Study because of absence of follow-up data on these outcomes in the ERF study.

Information on covariates, obtained during the baseline examinations for both studies, included age, gender, current and former smoking status, and alcohol consumption. Body mass index was defined as weight in kilograms divided by the square of height in meters. Hypertension was defined as systolic blood pressure ≥ 140 mmHg, diastolic blood pressure ≥ 90 mmHg, or use of medications indicated for the treatment of hypertension. Diabetes was defined as fasting plasma glucose levels ≥ 7 mmol/L, random plasma glucose ≥ 11.1 mmol/L, or use of medications indicated for the treatment of diabetes.

Statistical Analyses
All statistical analyses were performed using R[20]. Means and standard deviations were calculated for continuous variables, while absolute numbers and percentages were computed for dichotomous variables.

To improve residual normality, IMT values were natural-log transformed prior to analysis. Linear regression models were used to assess the relationships between risk scores and their corresponding lipid levels and between risk scores and atherosclerosis measures (IMT and plaque) in the Rotterdam Study. To properly account for relatedness in the family-based ERF population, analysis was conducted using variance component meth-

odology as realized in the SOLAR software package[21]. Cox proportional hazards models, as implemented in the survival package for R[22], were utilized to assess the relationship between the risk scores and incidence of MI and CHD, using follow-up time on the independent axis. Proportionality assumptions were tested by analyzing weighted Schoenfeld residuals[23].

The models for the analyses of association between the risk scores and their corresponding lipid levels were adjusted for age and sex. Individuals on lipid lowering medication were excluded from these analyses. Two models were fitted for the analyses of atherosclerosis. The first was adjusted only for age and sex. The second was adjusted for age, sex, hypertension, body-mass index, diabetes, current and former smoking status and alcohol consumption. For the Rotterdam Study, an additional variable indicating the cohort was added to all models to account for possible differences between RS1 and RS2. The same models were also fitted for Cox analysis of incident disease; these models were run both including and excluding individuals who were prevalent cases at baseline. To assess the additive value of the genetic risk scores above the corresponding lipid levels, the models that provided evidence of association were re-analyzed with the corresponding lipid level included as a covariate.

Results from the two populations were combined with inverse-variance weighted random effects meta-analyses, as implemented in the rmeta package for R[24].

In the Rotterdam Study, the R package PredictABEL[25] was used to calculate the area under the receiver operating characteristic curve (AUC) for prediction of incident MI and CHD. The Framingham Risk Score (FRS) for 10 year CHD risk prediction[3] was calculated for the Rotterdam Study participants. The predictive ability of the genotype scores, FRS and a combination of both was assessed. A variable indicating the cohort (RS1 or RS2) was added to all models.

RESULTS

Descriptive statistics for the 2 populations are provided in Table 1. Except for sex and systolic blood pressure, all tested characteristics exhibited significant differences between the populations, which differ by 2 decades in age. Among the 8,130 Rotterdam Study (RS) participants, 499 myocardial infarction (MI) cases (mean follow-up 9.76 years) and 1,194 CHD cases (mean follow-up 9.54 years) were present. Exclusion of prevalent cases resulted in 398 MI and 924 CHD cases.

Table 1. Description of the study populations

	ERF	RS
n	2269	8130
Age (years)	48.3 (14.5)	68.2 (9.1)
Male n (%)	991 (43.7)	3411 (42.0)
Current Smoker n (%)	889 (39.2)	2464 (30.3)
Former Smoker n (%)	679 (29.9)	3503 (43.1)
Hypertension n (%)	641 (28.3)	2750 (33.8)
BMI (kg/m^2)	26.8 (4.7)	26.6 (3.8)
Diabetes n (%)	135 (6.0)	845 (10.4)
Alcohol Use n (%)	1519 (67.0)	5477 (67.4)
SBP (mmHg)	139.8 (20.2)	140.3 (22.1)
DBP (mmHg)	80.2 (9.9)	75.1 (11.6)
TC (mmol/L)	5.6 (1.1)	6.4 (1.2)
LDL-C (mmol/L)	3.71 (0.97)	3.73 (0.88)
HDL-C (mmol/L)	1.3 (0.4)	1.4 (0.4)
TG (mmol/L)	1.32 (0.68)	1.52 (0.69)
LLT n (%)	297 (13.1)	426 (5.2)
IMT (mm)	0.8 (0.2)	0.9 (0.2)
Plaque	2.6 (2.1)	1.5 (1.6)
CHD n (%)	NA	1194 (14.7)
MI n (%)	NA	499 (6.1)
CHD prev. excl. n (%)	NA	924 (11.4)
MI prev. excl. n (%)	NA	398 (4.9)
Follow-up CHD (years)	NA	9.54 (4.7)
Follow-up MI (years)	NA	9.76 (4.7)
FRS points	NA	10.13 (3.6)

Continuous variables are presented as mean (standard deviation). Dichotomous variables are presented as number (%). ERF: Erasmus Rucphen Family Study, RS: Rotterdam Study, n: number, BMI: body mass index, SBP: systolic blood pressure, DBP: diastolic blood pressure, TC: total cholesterol, HDL-C: high-density lipo-protein cholesterol, TG: triglycerides, LLT: lipid lowering therapy, IMT: carotid intima media thickness, CHD: coronary heart disease, NA: not applicable, MI: myocardial infarction, prev. excl.: prevalent cases excluded, FRS: Framingham Risk Score

Table 2 provides means of the risk scores, and the mean number of risk alleles, for the various sets of lipid SNPs as well as the total number of SNPs used to compute the scores. The means and SDs of the TC, LDL-C, HDL-C, and TG risk scores were similar in Erasmus Rucphen Family (ERF) and the RS, as were the means, SDs, and overall distributions of risk alleles (Figure 1). Numbers of overlapping SNPs and loci between each pair of risk scores are depicted in Table 3. Detailed information about the exact SNPs used to compute the 4 risk scores is depicted in Table I in the online-only Data Supplement, which was adapted from Table II in the online-only Data Supplement from the GLGC article.[11]
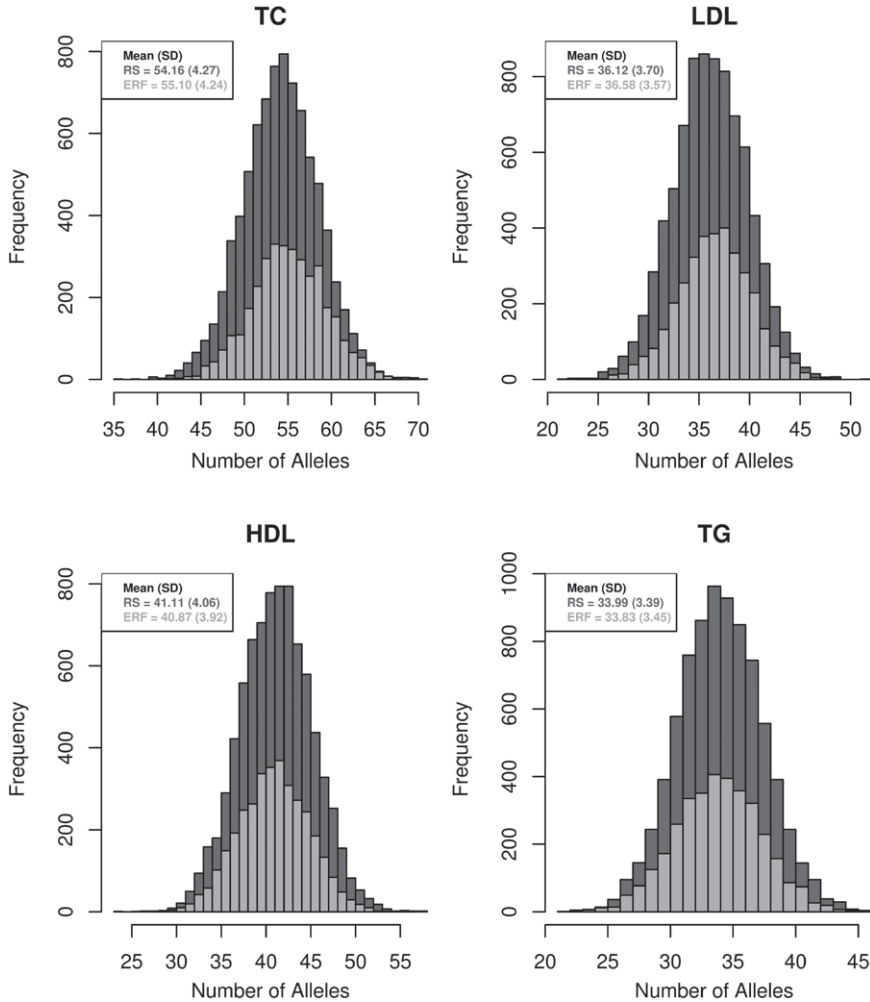
Figure 1. Distribution of lipid increasing (high-density lipoprotein cholesterol
[HDL] decreasing) alleles in the Rotterdam Study (RS) and the Erasmus Rucphen Family Study (ERF). LDL indicates low-density lipoprotein cholesterol; TC, total cholesterol; and TG, triglycerides.

Table 2. Genetic risk score means and number of SNPs

	Weighted score		Allele count		Number of SNPs
	ERF	**RS**	**ERF**	**RS**	
TC Risk Score	0.061 (0.005)	0.061 (0.005)	55.1 (4.24)	54.16 (4.27)	52
LDL-C Risk Score	0.054 (0.006)	0.054 (0.006)	36.58 (3.57)	36.12 (3.7)	37
HDL-C Risk Score	0.018 (0.002)	0.019 (0.002)	40.87 (3.92)	41.11 (4.06)	47
TG Risk Score	0.035 (0.004)	0.035 (0.004)	33.83 (3.45)	33.99 (3.39)	32

Presented as mean (standard deviation). ERF: Erasmus Rucphen Family Study, RS: Rotterdam Study, SNP: single nucleotide polymorphism, TC: total cholesterol, LDL-C: low-density lipoprotein cholesterol, HDL-C: high-density lipoprotein cholesterol, TG: triglycerides

Table 3. Overlapping SNPs and loci in the genetic risk scores

	TC Risk Score	LDL Risk Score	HDL Risk Score	TG Risk Score
TC Risk Score	**52**	25	6	3
LDL Risk Score	36	**37**	2	2
HDL Risk Score	16	8	**47**	3
TG Risk Score	13	10	15	**32**

In bold the numbers of SNPs used to compute the genetic risk scores, above the diagonal the numbers of SNPs shared by the pairs of risk scores, below the diagonal the numbers of loci shared by the pairs of risk scores, TC: total cholesterol, LDL-C: low-density lipoprotein cholesterol HDL-C: high-density lipoprotein cholesterol, TG: triglycerides

All 4 risk scores were robustly associated with their corresponding lipid levels in both cohorts separately, as well as in the meta-analysis (Table II in the online-only Data Supplement). Effect estimates were similar in both cohorts. The proportion of variance of the lipid levels explained by the genetic risk scores ranged from 6% for TG to 8% for TC for 1 SD increase in score above the mean.

The associations between the genetic risk scores and carotid atherosclerosis measures are described in Table 4. The TC risk score (Table 4) was nominally associated with carotid intima-media thickness (IMT) in both the age and sex-adjusted and full models in ERF ($P = 0.021$ and $P = 0.044$). This risk score was not associated with IMT in the RS. In the meta-analysis of the 2 populations, there was only a borderline association in the full model ($P = 0.071$). In terms of carotid plaque score, the TC score was robustly associated in both cohorts individually and in meta-analysis. The meta-analysis results demonstrated an increase in plaque score of 0.094 per SD increase in risk score in the age and sex-adjusted model ($P = 3.97 \times 10^{-7}$) and 0.106 per SD increase in the fully adjusted model ($P = 4.91 \times 10^{-8}$). The effect size estimates were consistent between each of the 2 populations. The LDL-C risk score (Table 4) was marginally associated with IMT in both the age and sex-adjusted model ($\beta = 0.006$; $P = 0.051$) and the full model ($\beta = 0.005$; $P = 0.039$). The magnitude of the association was slightly higher than for the TC risk score. With respect to plaques, the LDL-C score was strongly associated. Each SD increase in risk score resulted in an ≈ 0.1 increase in plaque score ($P = 3.15 \times 10^{-8}$ in the age and sex model; and $P = 3.18 \times 10^{-8}$ in the full model). The regression coefficients for the association with plaque score were very similar to those in the TC risk score models. As was the case for the TC risk score, the associations between the LDL risk score and plaque would survive any correction for multiple-testing. The genetic HDL-C risk score was not associated with IMT (Table 4), but it was modestly associated with plaque. An SD increase in the genetic score increased plaque by ≈ 0.04 ($P \approx 0.02$), irrespective of the model. However, as with the associations between the TC and LDL-C scores and IMT, this association would not survive a reasonable multiple-testing correction. Similarly, the TG

Table 4. Genetic risk scores for lipid levels and atherosclerosis

TC

		Age & Sex			Full Model		
		β	S.E.	*P*-value	β	S.E.	*P*-value
IMT	RS	0.002	0.002	0.338	0.003	0.002	0.218
	ERF	0.009	0.004	0.021	0.008	0.004	0.044
	Meta	0.005	0.003	0.156	0.004	0.002	0.071
Plaque	RS	0.087	0.021	4.90×10^{-5}	0.109	0.023	1.47×10^{-6}
	ERF	0.114	0.037	0.002	0.097	0.038	0.010
	Meta	0.094	0.019	3.97×10^{-7}	0.106	0.019	4.91×10^{-8}

LDL-C

		Age & Sex			Full Model		
		β	S.E.	*P*-value	β	S.E.	*P*-value
IMT	RS	0.004	0.002	0.096	0.004	0.002	0.109
	ERF	0.010	0.004	0.012	0.009	0.004	0.022
	Meta	0.006	0.003	0.051	0.005	0.003	0.039
Plaque	RS	0.100	0.021	2.88×10^{-6}	0.113	0.023	5.15×10^{-7}
	ERF	0.108	0.037	0.003	0.090	0.038	0.018
	Meta	0.102	0.019	3.15×10^{-8}	0.107	0.019	3.18×10^{-8}

HDL-C

		Age & Sex			Full Model		
		β	S.E.	*P*-value	β	S.E.	*P*-value
IMT	RS	0.004	0.002	0.047	0.004	0.002	0.099
	ERF	−0.002	0.004	0.526	−0.000	0.004	0.968
	Meta	0.002	0.003	0.608	0.003	0.002	0.156
Plaque	RS	0.048	0.021	0.023	0.055	0.023	0.014
	ERF	0.024	0.036	0.509	0.030	0.037	0.416
	Meta	0.042	0.018	0.022	0.049	0.019	0.012

TG

		Age & Sex			Full Model		
		β	S.E.	*P*-value	β	S.E.	*P*-value
IMT	RS	0.001	0.002	0.550	0.002	0.002	0.480
	ERF	−0.007	0.004	0.064	−0.006	0.004	0.123
	Meta	−0.002	0.004	0.575	−0.002	0.004	0.677
Plaque	RS	0.035	0.022	0.106	0.056	0.023	0.016
	ERF	0.043	0.037	0.248	0.044	0.038	0.242
	Meta	0.037	0.019	0.048	0.052	0.020	0.008

Full model adjusted for age, sex, current and former smoking, hypertension, body-mass index, diabetes and alcohol consumption. S.E.: standard error, IMT: intima media thickness, RS: Rotterdam Study, ERF: Erasmus Rucphen Family Study, TC: total cholesterol, LDL-C: low-density lipoprotein cholesterol, HDL-C: high-density lipoprotein cholesterol, TG: triglycerides

risk score (Table 4) was modestly associated with plaque in the full model ($\beta = 0.052$; $P = 0.008$). The effect estimate was slightly lower, and the P value substantially higher, in the model that only adjusted for age and sex ($\beta = 0.037$; $P = 0.048$).

In the models that showed significant association between the genetic risk scores and IMT or plaque, the lipid levels corresponding to the genetic risk scores were included in the model. After inclusion of these lipid levels, effect estimates for the genetic risk scores

Table 5. Genetic risk scores for lipid levels and incident MI and CHD in the Rotterdam Study

TC

		Age & Sex		Full Model	
		HR [95% C.I.]	**P-value**	**HR [95% C.I.]**	**P-value**
All	MI	1.12 [1.03, 1.22]	0.012	1.11 [1.00, 1.22]	0.040
	CHD	1.10 [1.04, 1.17]	7.02×10^{-4}	1.11 [1.04, 1.18]	0.003
Prevalent Excluded	MI	1.13 [1.03, 1.25]	0.012	1.11 [0.99, 1.24]	0.063
	CHD	1.10 [1.03, 1.18]	0.003	1.09 [1.01, 1.18]	0.021

LDL-C

		Age & Sex		Full Model	
		HR [95% C.I.]	**P-value**	**HR [95% C.I.]**	**P-value**
All	MI	1.12 [1.03, 1.23]	0.011	1.11 [1.01, 1.23]	0.033
	CHD	1.10 [1.04, 1.17]	7.00×10^{-4}	1.11 [1.03, 1.18]	0.003
Prevalent Excluded	MI	1.13 [1.02, 1.25]	0.015	1.11 [0.99, 1.23]	0.077
	CHD	1.10 [1.03, 1.17]	0.006	1.08 [1.00, 1.16]	0.045

HDL-C

		Age & Sex		Full Model	
		HR [95% C.I.]	**P-value**	**HR [95% C.I.]**	**P-value**
All	MI	1.01 [0.93, 1.10]	0.776	0.99 [0.90, 1.09]	0.790
	CHD	1.03 [0.97, 1.09]	0.362	1.01 [0.95, 1.08]	0.687
Prevalent Excluded	MI	0.97 [0.88, 1.06]	0.489	0.94 [0.84, 1.05]	0.255
	CHD	1.01 [0.94, 1.07]	0.879	0.99 [0.92, 1.07]	0.845

TG

		Age & Sex		Full Model	
		HR [95% C.I.]	**P-value**	**HR [95% C.I.]**	**P-value**
All	MI	1.02 [0.93, 1.11]	0.688	0.99 [0.90, 1.09]	0.830
	CHD	1.05 [1.00, 1.11]	0.071	1.05 [0.99, 1.12]	0.113
Prevalent Excluded	MI	1.00 [0.91, 1.10]	0.974	0.98 [0.88, 1.09]	0.686
	CHD	1.04 [0.98, 1.11]	0.185	1.04 [0.96, 1.12]	0.310

Full model adjusted for age, sex, current and former smoking, hypertension, body-mass index, diabetes and alcohol consumption. TC: total cholesterol, HR: hazard ratio, MI: myocardial infarction, CHD: coronary heart disease, LDL-C: low-density lipoprotein cholesterol, HDL-C: high-density lipoprotein cholesterol, TG: triglycerides

were lower and *P*-values higher, but there were still significant associations: the TC and LDL-C scores were still associated with plaque in the full model (TC: $\beta = 0.054$; $P = 0.010$; LDL-C: $\beta = 0.047$; $P = 0.040$) and marginally associated in the age and sex-adjusted model (TC: $\beta = 0.038$; $P = 0.053$; LDL-C: $\beta = 0.042$; $P = 0.060$).

The analyses of the relationships between the genetic risk scores and incident MI and CHD are presented in Table 5. For all of the models tested, proportionality assumptions were met (minimum $P = 0.18$). The TC and LDL-C scores (Table 5) were associated with both MI and CHD. The effect estimates were consistent across models and, after the exclusion of prevalent cases, were similar for MI and CHD (hazard ratios [HRs] ≈1.10). The *P*-values increased in the full models, and with prevalent cases excluded, as would be expected because of reductions in sample size. The findings for CHD were particularly strong; all models achieved at least nominal significance for both the TC and LDL-C risk scores (P between 0.045 and 7.0×10^{-4}). There were no significant associations between the HDL-C and TG risk scores (Table 5) with either incident MI or CHD, irrespective of the inclusion or exclusion of prevalent cases, although there was a borderline association between the TG score and CHD (HR [95% confidence interval] = 1.05 [1.00–1.11]; $P = 0.071$).

Inclusion of TC levels in the models that assessed the association of the TC genetic risk score with MI and CHD resulted in lower HRs and higher *P*-values, but associations with CHD were still borderline significant in the age and sex-adjusted models (HR [95% confidence interval] all: 1.06 [1.00–1.13], prevalent excluded: 1.07 [0.99–1.14]). Adding LDL-C to the models that showed significant association of the LDL-C genetic risk score with MI or CHD resulted in HRs similar to those in the original models. *P*-values were higher, but still (borderline) significant for the associations with CHD ($P = 0.004$–0.068). For the MI outcome, only a borderline significant association remained in the age and sex-adjusted model without exclusion of the prevalent cases ($P = 0.071$).

The associations between the pure genetic risk scores for LDL-C, HDL-C, and TG with atherosclerosis, after exclusion of the variants that were also associated with at least 1 of the other lipid measures, are described in Table III in the online-only Data Supplement. Although *P*-values were higher and effect estimates slightly lower, the LDL-C risk score was still associated with plaque score. Additionally, the LDL-C score was borderline associated with IMT, whereas the pure HDL-C and TG risk scores still showed marginal association with plaques.

Table IV in the online-only Data Supplement describes the associations of the 3 pure genetic risk scores (for LDL-C, HDL-C, and TG) with MI and CHD in the RS. Effect estimates

for the associations of the LDL-C score with MI and CHD were similar to those for the original risk score and remained significantly associated with both outcomes in the age and sex-adjusted model. In the full model, the score was still at least borderline significantly associated with both outcomes when prevalent cases were not excluded. After exclusion of prevalent cases, only a borderline significant association with CHD was left.

Table V in the online-only Data Supplement describes area under the curves (AUCs) for the prediction of incident MI and CHD in the RS. Framingham Risk Score (FRS) discriminated MI better than the genetic risk scores combined (AUC 0.65 versus AUC 0.62) and combining both slightly and only borderline significantly improved the results compared with the FRS AUC alone (AUC 0.66; $P = 0.069$). For the CHD outcome, results were similar: the AUC for the FRS was 0.65, 0.61 for the genetic risk scores combined, and 0.65 for the FRS and genetic risk scores combined with a slightly narrower confidence interval than for the FRS alone. However, this improvement was not significant. Figure I in the online-only Data Supplement shows the receiver-operating characteristic curves for the prediction of MI and CHD.

DISCUSSION

This study shows that aggregations of common genetic variants influencing lipid levels play a significant role in the development of atherosclerosis and the subsequent occurrence of CVD despite generally small effects on lipid levels individually. All genetic risk scores (for TC, LDL-C, HDL-C, and TG) were associated with carotid plaque with the effects of the TC and LDL-C scores being ≈ 2.5× as large as the effects of the HDL-C and TG scores. IMT was marginally associated with the LDL-C score. The TC and LDL-C scores were robustly associated with incident CVD, especially with a composite CHD phenotype.

Associations of the TC and LDL-C risk scores with plaque were still at least marginally significant after inclusion of the lipid levels corresponding to the genetic risk score in the models. The TC and LDL-C scores were also still (borderline) significantly associated with MI and CHD. These findings suggest some added value of these genetic risk scores beyond the lipid levels themselves. This may be because they reflect lifelong exposure to higher lipid values (in contrast to fluctuating occasional lipid measurements), which would make genes relevant for early prediction and prevention purposes. A recent Framingham Heart Study article described a significant association of an LDL-C genetic risk score with coronary artery calcium that vanished after inclusion of LDL-C levels in the model.[26] This study, however, found a stronger association of early and long-term

average lipid levels compared with current measures. This supports the idea that risk scores reflecting long-term elevated lipid levels may be of use compared with current lipid measurements.

After exclusion of the variants from the genetic risk scores for LDL-C, HDL-C, and TG that were also associated with at least 1 of the other outcomes, most *P*-values were higher and, for the atherosclerosis outcomes, effect estimates slightly lower. However, these risk scores were still associated with plaque, and the LDL-C score was still associated with MI and CHD in most models, showing that these associations did not result solely from the influence of variants associated with one of the other lipid outcomes.

Although the HDL-C and TG risk scores were associated with plaque, the evidence for association with subclinical atherosclerosis and incident CVD obtained for these scores was clearly weaker compared with the TC and LDL-C scores. This is in line with evidence from 2 recent studies[27,28]. In one study, we analyzed 1,987 genotyped RS participants who underwent computed tomography of the aortic arch and carotid arteries to quantify atherosclerotic calcification[27]. In that study, we demonstrated that genetic risk scores comprised of TC and LDL-C SNPs were more predominantly associated with larger calcification volumes than HDL-C and TG risk scores in all vessel beds. In the other study, association between risk scores based on previously reported lipid SNPs and CHD was assessed in 2 UK cohorts of middle-aged men and women. Similar to our findings, the TC and LDL-C risk scores were associated with higher CHD risk in both cohorts (odds ratios, 1.30–1.42 for individuals in the highest quintiles of the risk scores compared with individuals in the lowest quintiles). In these cohorts, the HDL-C score was not associated with the outcome, and the TG score was only associated in 1 of the 2.

Our findings are also in line with the evidence for causal association of HDL-C and TG levels themselves with CVD, which has not been as solid as the evidence for TC and LDL-C levels.[4,29] In a recent large Mendelian randomization study, a single polymorphism in the endothelial lipase gene and a risk score based on 14 SNPs that are exclusively associated with HDL-C were investigated for association with HDL-C levels and with MI.[30] In this study, both the single SNP and the risk score were associated with HDL-C levels, but not with MI. These findings suggest that at least some genetic mechanisms that raise HDL-C do not lower risk of MI. TG has been considered a biomarker for CVD because of its association with other atherogenic particles rather than being directly atherogenic.[29] However, this was challenged by the identification of a genetic variant consistently associated with TG that was also associated with incident CHD.[31]

It is remarkable that all genetic risk scores were associated with carotid plaque and only the LDL-C score with IMT. An explanation might be that the plaque score, in contrast to IMT, is based on a direct measurement of the presence of atherosclerotic plaques and, thus, better represents atherosclerosis. In the RS, carotid plaques were more strongly associated with coronary calcification than with common carotid IMT.[32] An association between IMT and coronary calcification was present in that study, suggesting that carotid IMT may be regarded as a measure of generalized atherosclerosis. The finding that this association was weaker, however, supports the idea that carotid plaques better represent the presence of atherosclerosis.

In a previous study, we observed that genetic risk scores for lipids, based on 26 loci, did not improve prediction of incident CHD.[10] Although the number of lipid loci increased to 95, the FRS still discriminated MI and CHD better than the genetic risk scores in our study and combining both only marginally improved the FRS AUC. However, because the genetic risk scores are invariant over time, they may prove useful for risk stratification at younger ages, where the clinical models may not be as effective. Moreover, despite the large number of lipid loci that have been discovered, there is still substantial missing heritability. The 95 GLGC loci, plus an additional 26 identified by conditional association analyses, explained 9.6% to 12.4% of the total variance in the 4 lipid traits in the Framingham Heart Study. This corresponds to $\approx$ 25% to 30% of the estimated heritability.[11] As we further elucidate the genetic basis for lipid levels, the predictive ability of genetic risk scores should improve.

Different hypotheses have been developed about the genetic architecture underlying common complex traits and diseases. The common disease, common variant hypothesis argues that common variants with modest effects underlie many complex traits.[33,34] However, for most complex phenotypes, the currently known common variants only explain a small portion of the estimated trait heritability. This is also the case for the common genetic CHD variants identified by a large GWAS meta-analysis comprising 22,233 cases and 64,762 controls. Those variants, in addition to the lead SNPs of previously established loci, explain $\approx$ 10% of the heritability of CHD.[35] The common disease, rare variant hypothesis, by contrast, posits that less frequent variants with larger effect sizes underlie genetic susceptibility to many common complex diseases.[33,34] One cause of CVD, for example, is familial hypercholesterolemia, in which rare LDL receptor mutations exert a large influence on lipid levels and subsequent premature CVD.[36,37] Our results show that the combined effects of numerous common variants associated with small effects on lipid levels are associated with the risk of subclinical (plaque) and, in the case of the genetic risk scores for TC and LDL-C, clinical outcomes, including CHD.

In a recent study, common and rare variants identified by the 1000 Genomes Project and by resequencing 7 genes in loci associated with LDL-C doubled the estimated heritability of LDL-C accounted for by these genes, indicating that both additional common and rare variants influence blood lipid levels.[38] With ongoing efforts in gene discovery, such as 1000 Genomes–based imputations and whole exome and genome sequencing, it is likely that many more variants, spanning a broad frequency range, will be discovered. The identification of these might improve the ability to detect individuals at increased risk for CVD. Finally, it may be that the cumulative effects of genetic variants associated with other intermediate phenotypes, such as blood pressure and obesity, will also be associated with CVD. Genetic risk scores based on these endophenotypes, in combination with those derived from the lipids, might lead to improved detection of individuals at increased risk for CVD.

Major strengths of our study are the use of population-based cohorts for the assessment of the risk scores and the weighting of the SNPs by the effect estimates obtained in the discovery cohorts. This gives an indication of the performance of risk scores, based on common variants and their effect estimates resulting from large GWAS, in the general population. Because both ERF and the RS were part of the discovery GWAS meta-analysis, a limitation of this study might be that results are slightly overestimated. However, because of the very large number of individuals that were included in the GLGC meta-analysis and that were not part of the ERF or RS cohorts, this effect is expected to be limited. The small number of individuals in this study compared with the discovery GWAS might even have led to false-negative results.

In conclusion, our results show that the cumulative effects of common genetic variants associated with TC, LDL-C, HDL-C, and TG are associated with subclinical outcomes, such as carotid plaque. The genetic risk scores for TC and LDL-C are also associated with incident CHD. Although genetic risk scores did not improve clinical AUCs, our study provides evidence for the added value of genetic risk scores above lipid levels themselves when studying subclinical atherosclerosis and CHD. As our knowledge of genetic variation increases, preclinical genetic screening tools might detect individuals at increased risk for CVD and, thus, enhance prediction and prevention of clinical events.

ACKNOWLEDGMENTS

REFERENCES

1. World health Organization. Media centre, Fact Sheets: Cardiovascular diseases (CVDs). http://www.who.int/mediacentre/factsheets/fs317/en/index.html. (August 2012)

2. Kannel, W.B., Dawber, T.R., Kagan, A., Revotskie, N. & Stokes, J., 3rd. Factors of risk in the development of coronary heart disease--six year follow-up experience. The Framingham Study. Ann Intern Med 55, 33-50 (1961).

3. Wilson, P.W. et al. Prediction of coronary heart disease using risk factor categories. Circulation 97, 1837-47 (1998).

4. Arsenault, B.J., Boekholdt, S.M. & Kastelein, J.J. Lipid parameters for measuring risk of cardiovascular disease. Nat Rev Cardiol 8, 197-206 (2011).

5. Isaacs, A. et al. Heritabilities, apolipoprotein E, and effects of inbreeding on plasma lipids in a genetically isolated population: the Erasmus Rucphen Family Study. Eur J Epidemiol 22, 99-105 (2007).

6. Weiss, L.A., Pan, L., Abney, M. & Ober, C. The sex-specific genetic architecture of quantitative traits in humans. Nat Genet 38, 218-22 (2006).

7. Willer, C.J. et al. Newly identified loci that influence lipid concentrations and risk of coronary artery disease. Nat Genet 40, 161-9 (2008).

8. Kathiresan, S. et al. Six new loci associated with blood low-density lipoprotein cholesterol, high-density lipoprotein cholesterol or triglycerides in humans. Nat Genet 40, 189-97 (2008).

9. Kathiresan, S. et al. Common variants at 30 loci contribute to polygenic dyslipidemia. Nat Genet 41, 56-65 (2009).

10. Aulchenko, Y.S. et al. Loci influencing lipid levels and coronary heart disease risk in 16 European population cohorts. Nat Genet 41, 47-55 (2009).

11. Teslovich, T.M. et al. Biological, clinical and population relevance of 95 loci for blood lipids. Nature 466, 707-13 (2010).

12. Fisher, R.A. The correlation between relatives on the supposition of Mendelian inheritance. Philosophical Transactions of the Royal Society of Edinburgh, 399-433 (1918).

13. Pardo, L.M., MacKay, I., Oostra, B., van Duijn, C.M. & Aulchenko, Y.S. The effect of genetic drift in a young genetically isolated population. Ann Hum Genet 69, 288-95 (2005).

14. Hofman, A., Grobbee, D.E., de Jong, P.T. & van den Ouweland, F.A. Determinants of disease and disability in the elderly: the Rotterdam Elderly Study. Eur J Epidemiol 7, 403-22 (1991).

15. Hofman, A. et al. The Rotterdam Study: 2010 objectives and design update. Eur J Epidemiol 24, 553-72 (2009).

16. Miller, S.A., Dykes, D.D. & Polesky, H.F. A simple salting out procedure for extracting DNA from human nucleated cells. Nucleic Acids Res 16, 1215 (1988).

17. Li, Y., Willer, C.J., Ding, J., Scheet, P. & Abecasis, G.R. MaCH: using sequence and genotype data to estimate haplotypes and unobserved genotypes. Genet Epidemiol 34, 816-34 (2010).

18. Bots, M.L., Hoes, A.W., Koudstaal, P.J., Hofman, A. & Grobbee, D.E. Common carotid intima-media thickness and risk of stroke and myocardial infarction: the Rotterdam Study. Circulation 96, 1432-7 (1997).

19. Hollander, M. et al. Carotid plaques increase the risk of stroke and subtypes of cerebral infarction in asymptomatic elderly: the Rotterdam study. Circulation 105, 2872-7 (2002).

20. R Core Team (2012). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. http://www.R-project.org/

21. Almasy, L. & Blangero, J. Multipoint quantitative-trait linkage analysis in general pedigrees. Am J Hum Genet 62, 1198-211 (1998).

22. Therneau, T. (2012). A Package for Survival Analysis in S. R package version 2.36-14 A Package for Survival Analysis in S.

23. Grambsch, P., Therneau, T. Proportional hazards tests and diagnostics based on weighted residuals. Biometrika 81, 515-26 (1994).

24. Lumley, T. (2009). rmeta: Meta-analysis. R package vesion 2.16. http://CRAN.R-project.org/package=rmeta

25. Kundu, S., Aulchenko, Y.S., Janssen, A.C.J.W. (2011). PredictABEL: Assessment of risk prediction models. R package version 1.2. CRAN.R-project.org/package=PredictABEL

26. Tsao, C.W. et al. Relations of long-term and contemporary lipid levels and lipid genetic risk scores with coronary artery calcium in the framingham heart study. J Am Coll Cardiol 60, 2364-71 (2012).

27. Bos, D. et al. Genetic Loci for Coronary Calcification and Serum Lipids Relate to Aortic and Carotid Calcification. Circ Cardiovasc Genet (2012).

28. Shah, S. et al. Influence of common genetic variation on blood lipid levels, cardiovascular risk, and coronary events in two British prospective cohort studies. Eur Heart J (2012).

29. Talayero, B.G. & Sacks, F.M. The role of triglycerides in atherosclerosis. Curr Cardiol Rep 13, 544-52 (2011).

30. Voight, B.F. et al. Plasma HDL cholesterol and risk of myocardial infarction: a mendelian randomisation study. Lancet 380, 572-80 (2012).

31. Sarwar, N. et al. Triglyceride-mediated pathways and coronary disease: collaborative analysis of 101 studies. Lancet 375, 1634–39 (2010).Shah, S. et al. Influence of common genetic variation on blood lipid levels, cardiovascular risk, and coronary events in two British prospective cohort studies. Eur Heart J 34, 972-81 (2013).

32. Oei, H.H. et al. The association between coronary calcification assessed by electron beam computed tomography and measures of extracoronary atherosclerosis: the Rotterdam Coronary Calcification Study. J Am Coll Cardiol 39, 1745-51 (2002).

33. Schork, N.J., Murray, S.S., Frazer, K.A. & Topol, E.J. Common vs. rare allele hypotheses for complex diseases. Curr Opin Genet Dev 19, 212-9 (2009).

34. Marian, A.J. Molecular genetic studies of complex phenotypes. Transl Res 159, 64-79 (2012).

35. Schunkert, H. et al. Large-scale association analysis identifies 13 new susceptibility loci for coronary artery disease. Nat Genet. 43, 333–8 (2011).

36. van der Graaf, A. et al. Molecular basis of autosomal dominant hypercholesterolemia: assessment in a large cohort of hypercholesterolemic children. Circulation 123, 1167-73 (2011).

37. Watts, G.F. et al. Familial hypercholesterolaemia: a review with emphasis on evidence for treatment, new models of care and health economic evaluations. Int J Evid Based Healthc 10, 211-221 (2012).

38. Sanna, S. et al. Fine mapping of five loci associated with low-density lipoprotein cholesterol detects variants that double the explained heritability. PLoS Genet 7, e1002198 (2011).

Chapter 4

The role of lipid-altering gene variants in the context of type 2 diabetes

CTGACTACGTA
CTACTAGCACTGTACACGA   GCTACGTACG
CGTACGTAGTCTGATCAAAACGTA   GCTATACAGCTA
CGTTGGTACCAGTACGTACGACTGCGA   GTACGTACGTACG
TGTGTGTCACACAGACACGTGATATAGACCAGATGACACACGTAC
GTACGACGGCGCGGCGCGATATATATAAAAGCACGGACACTACGA
ACGATGACTGACTACGTGCGCGCGTAGCTAGCTACGGATAGGACTA
TACTGACTACGTACTACTAAAGCTGCTTTTACGTACGTACGGGCATC
CATGTACTATCGCGGCGCGTAGACTGTACGACCAGACTGACTGACT
GACTGACTGACGCGCGTACGTAGCTACTGTAGTACGTACGTACGTA
TTCAGACTGACTGACTGAGATTACAGCTACGTAGTACTGACACGT
TACGACGGCGCGGCGCGATATATATAAAAGCACGGACACTACGA
GATGACTGACTACGTGCGCGCGTAGCTAGCTACGGATAGGACTA
TGACTACGTACTACTAAAGCTGCTTTTACGTACGTACGGGCATC
CTGCGCATGTACTATCGCGGCGCGTAGACTGTACGACCAGAC
ACTGACGCGCGTACGTAGCTACTGTAGTACGTACGTACGTA
ACTGACTGACTGAGATTACAGCTACGTAGTACTGACACGT
GCATACGAGTCGAAGCGCGCAATATATATTATATCGGCC
CGTGACTACATCATGCTGACTACGGTCGCGCTCAACG
TGACTACGTACGTATACGACGTACTGACGGCGCGC
GCACTGTACACGACTGACTTACTAGCTACGTACG
TGATCAAAACGTACGCGCGGCTATACAGCTA
CGTACGACTGCGATACGTACGTACGTACG
CGTGATATAGACCAGATGACACACGT
TATATATAAAAGCACGGACACT
GTAGCTAGCTACGGATA
CTTTTACGTACGT
ACGACCAG
TAGT

Risk scores comprised of common lipid-altering genetic variants are associated with lipid levels and suggest an altered role of common genetic variation in type 2 diabetes

ABSTRACT

Objective: A large proportion of type 2 diabetes patients have dyslipidemia, which is a major risk factor for cardiovascular complications. To obtain more insight into the role of common genetic variation in diabetic dyslipidemia, we studied the role of lipid-altering single nucleotide polymorphisms (SNPs) in patients with type 2 diabetes.

Methods: Risk scores were computed from total cholesterol (TC), low-density lipoprotein cholesterol (LDL-C), high-density lipoprotein cholesterol (HDL-C) and triglyceride (TG) altering SNPs in individuals with and without type 2 diabetes in the Rotterdam Study (n=9,791) and Erasmus Rucphen Family Study (n=2,313). Associations of these scores with lipid levels and prevalent dyslipidemia were assessed using regression models. Areas under the receiver operating characteristic curves (AUCs) were calculated to test the ability of the risk scores to predict incident dyslipidemia.

Results: In individuals with diabetes, the LDL-C, HDL-C and TG scores were associated with their corresponding lipids (P=0.006 - 4.2×10^{-15}), while only the TG score was associated with dyslipidemia (P=8.1×10^{-6}). Effect estimates were not significantly different between diabetic and non-diabetic individuals. The AUC of the genetic risk scores for prediction of incident dyslipidemia was significantly higher in diabetes cases (AUC=0.76) than in controls (AUC=0.62) (P=1.8×10^{-7}).

Conclusions: Risk scores comprised of common lipid-altering SNPs were associated with lipid levels in patients with type 2 diabetes. The AUCs of the genetic risk scores suggest an altered role of common genetic variation in type 2 diabetes and raise the question of whether genetic scores could be used for prognostic purposes in the future.

INTRODUCTION

Type 2 diabetes poses a major global health problem. The number of people living with diabetes is expected to increase from 366 million in 2011 to 552 million in 2030 without urgent preventive interventions[1]. A large proportion of diabetes patients have dyslipidemia, which is a major risk factor for cardiovascular complications[2,3]. In diabetic dyslipidemia, triglyceride (TG) levels are typically increased, while high-density lipoprotein (HDL-C) is decreased. Additionally, low-density lipoprotein (LDL-C) is converted to small, dense LDL[4,5]. Total cholesterol (TC) and LDL-C levels may be similar, or slightly lower, in individuals with type 2 diabetes compared to those without[6,7], but these differences are minor compared to those for HDL-C and TG.

A number of processes are likely to be involved in diabetic dyslipidemia, including the effects of insulin on hepatic production of apolipoprotein B (apoB). ApoB is the major apolipoprotein component of very-low-density lipoprotein (VLDL) and is elevated in patients with type 2 diabetes, resulting in increased plasma VLDL[4]. Raised plasma VLDL results in an increased production of precursors of small dense LDL[5]. The fact that LDL-C is not typically increased in type 2 diabetes may be explained by a balance of factors that affect LDL-C production and catabolism[4]. Further, alterations in the production of biologically active lipoprotein lipase (LpL), increased transfer of cholesteryl esters in the core of LDL and HDL by cholesteryl ester transfer protein (CETP) in the presence of elevated VLDL, and the effects of insulin on adipose tissue and muscle seem to play a role in diabetic dyslipidemia[4,8].

In two large genome-wide association studies (GWAS) of circulating lipids in the general population, genetic variants in 157 loci were found to be associated with TC, HDL-C, LDL-C or TG[9,10]. Risk scores comprised of lipid-altering genetic variants in 95 loci from the first of these meta-analyses were previously associated with lipid levels, atherosclerotic calcification and coronary heart disease[11,12]. To our knowledge, however, association of these variants in individuals with type 2 diabetes has not been studied. We hypothesized that, in addition to the metabolic alterations described above, these variants also influence lipid levels in individuals with type 2 diabetes. To test our hypothesis, we computed risk scores from the lead SNPs as reported in the two discovery GWAS meta-analyses[9,10] in each of the 157 loci to assess whether the combined effects of these variants are associated with lipid levels and prevalent dyslipidemia in patients with type 2 diabetes. Additionally, we investigated whether dyslipidemia is influenced by an interaction between the risk scores and diabetes status and we assessed the ability of the risk scores to predict incident dyslipidemia in individuals with and without type 2 diabetes.

MATERIALS AND METHODS

Study Populations

Participants from the Rotterdam Study and the Erasmus Rucphen Family Study (ERF) were included in these analyses[13-15]. The Rotterdam Study is a single-center prospective cohort study comprised of elderly individuals living in Ommoord, a district in the city of Rotterdam, the Netherlands. At three points in time, participants were recruited into the cohort. These are the RS-I (n = 7,983, age at baseline ≥ 55 years), RS-II (n = 3,011, age at baseline ≥ 55 years) and RS-III (n = 3,932, age at baseline ≥ 45 years) sub-cohorts. Baseline examinations took place between 1990 - 1993, 2000 - 2001 and 2006 - 2008. Participants underwent extensive physical examinations and interviews. In the current analyses, 9,791 individuals from RS-I (n = 5,598, 614 diabetic), RS-II (n = 2,137, 247 diabetic) and RS-III (n = 2,056, 156 diabetic) for whom both phenotypic and genotypic information was available were included. ERF is a family-based cohort study composed of individuals living in a contiguous geographic region in the southwest of the Netherlands. All living descendants aged 18 years or above, of twenty-two couples that had a large number of children baptized in the community church between 1850 and 1900 were invited to participate in the study. Their spouses were invited as well. Approximately 3,200 individuals participated. Examinations took place between June 2002 and February 2005. Participants filled out questionnaires and underwent extensive interviews and physical examinations at a research center in the area. 2,313 individuals (148 diabetic) for whom phenotypic, genotypic and genealogical information was available were included in the current analyses. All participants, or their legal guardians, provided written informed consent. Both studies were approved by the Medical Ethics board of the Erasmus Medical Center Rotterdam, the Netherlands.

Clinical and laboratory assessment

In both studies, venous blood samples were obtained from study participants. Participants in the RS-I cohort were non-fasting during baseline examinations, however they were fasting at their third visit to the research center (RS-I-3). Participants in the RS-II, RS-III and ERF cohorts were fasting at the time of blood sampling. In RS-I, glucose was measured by the glucose hexokinase method and TC and HDL-C were measured using enzymatic colorimetric methods (Kone Specific Analyzer, Kone Instruments). TG was not measured in RS-I at baseline, but it was measured in RS-I-3. In RS-I-3, RS-II and RS-III, TC, HDL-C and TG were measured using comparable enzymatic procedures (Hitachi Analyzer, Roche Diagnostics) and LDL-C was calculated using the Friedewald formula: LDL-C = TC - HDL-C - 0.45*TG when TG ≤ 4.52 mmol/L[16]. In ERF, glucose, TC, HDL-C, LDL-C and TG were measured using a Synchron LX 20 Systems analyzer (Beckman Coulter, Fullerton, CA, USA). In both the Rotterdam Study and ERF, participants were asked to present

the medications they used, including lipid-lowering and anti-diabetic medications. In individuals using statins, to account for the effect of statins on lipids, TC was divided by 0.797, HDL-C was divided by 1.056, TG by 0.868 and, if directly measured, LDL-C was divided by 0.722. These adjustments are based on the sample-size weighted mean proportional differences in a large prospective meta-analysis including 14 randomized trials of statins[17]. In individuals using nicotinic acid derivatives, HDL-C was divided by 1.157 and TG by 0.800, and to adjust for use of fibric acid derivatives, HDL-C was divided by 1.100 and TG was divided by 0.637. These adjustments are the meta-analysis proportional differences from a large meta-analysis of those compounds[18]. If individuals used more than one type of lipid lowering medication, the lipid levels were adjusted for the one with the largest influence on each specific lipid. If use of lipid-lowering medication was not further specified, the adjustments for statin use were applied. In RS-I medication was not specified, however blood sampling was done before the publication of primary results of the Scandinavian Simvastatin Survival Study[19]. Therefore, in RS-I, no adjustments of lipid levels were applied. Diabetes was defined as fasting plasma glucose levels ≥ 7 mmol/L, 2-h plasma glucose ≥ 11.1 mmol/L, random plasma glucose ≥ 11.1 mmol/L, or use of medications indicated for the treatment of diabetes. Dyslipidemia was defined based on guidelines used by the Dutch college of general practitioners: TC ≥ 6.5 mmol/L when TC/HDL-C ratio ≥ 5, TC < 6.5 mmol/L when TC/HDL-C ratio > 8, or TC ≥ 9 mmol/L independent of TC/HDL-C ratio. Individuals using lipid-lowering medication were also considered to have dyslipidemia. In RS-I and RS-II, incident dyslipidemia was studied, which was defined as free of dyslipidemia at baseline and development of dyslipidemia during follow-up. This included three follow-up visits for RS-I and two follow-up visits for RS-II; the last follow-up visit for the cohorts took place between 2009-2011 and 2011-2012 respectively. Information on age, sex and BMI, covariates in the analyses, was obtained during baseline examinations. BMI was defined as weight in kilograms divided by the square of height in meters.

Genotyping

Genomic DNA was extracted from venous blood samples obtained at baseline in both the Rotterdam Study and ERF. DNA was extracted using the salting out method[20]. Genotyping in the Rotterdam Study was performed using 550 and 610K Illumina arrays and in ERF using 318 and 370K Illumina arrays. Exclusion criteria for individuals were excess autosomal heterozygosity, mismatches between called and phenotypic gender, and outliers identified by an IBS clustering analysis. Single nucleotide polymorphisms (SNPs) were excluded for Hardy-Weinberg equilibrium P-value ≤ 10-6 or SNP call rate ≤ 98%. Genotypes with minor allele frequencies > 1% were used to impute about 2.5 million autosomal SNPs using HapMap CEU release 22 as a reference panel. Imputation was performed using MaCH[21]. Imputed genotypes were coded as dosages. These are

values between 0 and 2 indicating the estimated number of copies of a given allele for each individual.

Genotype Scores

For each major lipid class (TC, LDL-C, HDL-C and TG), genetic risk scores were calculated, per individual, based on the lead SNPs in the 157 loci from the large lipid GWAS meta-analyses by Teslovich et al. and Willer et al.[9,10]. The risk scores were calculated per individual as:

$$\left(\sum_{i=1}^{n} (G_i * \beta_i)\right) / n$$

where n is the number of SNPs comprising the score, G_i is the number of lipid increasing alleles (lipid decreasing alleles for HDL) at the ith genotype, and β_i is the per allele effect estimate for the ith SNP as obtained in the GWAS meta-analyses[9,10]. To aid in the interpretation of the analysis results, the scores were standardized to yield a mean of zero and a standard deviation of one (such that regression coefficients equal the change in value, or risk, per standard deviation increase, or decrease, in score).

Statistical Analyses

TC, LDL-C, HDL-C and TG values more than five standard deviations from the mean were excluded from the analyses. To improve normality, TG values were natural-log transformed prior to analyses.

To assess the associations of the genetic risk scores and risk score*diabetes interactions with lipid levels in the Rotterdam Study, linear regression was performed using R[22]. In ERF, to account for relatedness in the family-based cohort, variance components methodology, as implemented in the SOLAR software package, was used[23]. To test the associations of the risk scores and risk score*diabetes interactions with prevalent dyslipidemia in the Rotterdam Study, R was used to perform logistic regression. In ERF, to take relatedness into account, generalized estimating equations were used as implemented in the R package gee[24]. All models were adjusted for age, sex and BMI. In the Rotterdam Study, a variable indicating the sub-cohort (RS-I, RS-II or RS-III) was added to all models that were analyzed in this study.

Results from the Rotterdam Study and ERF were combined using inverse variance weighted random effects meta-analysis as implemented in the R package rmeta[25]. The relationships between single SNP*diabetes interactions and dyslipidemia were assessed in the same manner as described above. Kolmogorov-Smirnov tests, as implemented in R, were used to compare the distributions of interaction P-values for each SNP set with the null distribution[22].

In RS-I and RS-II, the Rotterdam Study sub-cohorts with follow-up data, the R package PredictABEL[26] was used to calculate the areas under the receiver operating characteristic curves (AUCs) for prediction of incident dyslipidemia. DeLong's test, as implemented in the R package pROC[27], was used to compare nested AUCs. To compare independent AUCs, the chi-square statistic was obtained to calculate the *P*-value.

RESULTS

Descriptions of the cohorts' characteristics are presented in Table 1. All variables, except for the percentage of male participants, BMI and LDL levels, differed significantly between the non-diabetic participants from the Rotterdam Study and ERF. In diabetes cases, all variables, except for the percentage of male participants, LDL and TG levels differed significantly between the two cohorts. The difference in mean age between the cohorts was approximately two decades in controls and one decade in cases.

Risk scores were comprised of 44 to 75 SNPs: the TG score included 44 SNPs, the LDL score 58, the HDL score 73 and the TC score 75. Risk score means and standard deviations were very similar in the Rotterdam Study and ERF. The mean TC scores (SD) were 0.042 (0.003) and 0.043 (0.003) in the Rotterdam Study and ERF, respectively. These values were 0.034 (0.004) for the LDL-C risk score and 0.013 (0.002) for the HDL-C risk score in both cohorts.

Table 1. Descriptions of the study populations

	RS		ERF	
	Controls	Cases	Controls	Cases
n	8774	1017	2165	148
Age (years)	65.0 (9.6)	69.3 (10.2)	47.7 (14.3)	59.6 (10.4)
Male n (%)	3690 (42.1)	476 (46.8)	944 (43.6)	76 (51.4)
BMI (kg/m^2)	26.6 (3.9)	28.2 (4.9)	26.7 (4.6)	30.6 (5.9)
TC (mmol/L)	6.3 (1.2)	6.2 (1.3)	5.7 (1.1)	5.6 (1.4)
LDL (mmol/L)	3.9 (0.9)	3.6 (1.0)	3.9 (1.0)	3.7 (1.1)
HDL (mmol/L)	1.4 (0.4)	1.2 (0.4)	1.3 (0.4)	1.1 (0.3)
TG (mmol/L)	1.5 (0.7)	2.1 (1.2)	1.3 (0.7)	2.0 (1.3)
LLT n (%)	722 (8.2)	144 (14.2)	220 (10.2)	70 (47.3)
Dyslipidemia (n%)	2776 (31.6)	411 (40.4)	497(23.0)	82 (55.4)
Incident Dyslipidemia (n%)*	713 (10.4)	93 (10.8)	NA	NA

Mean (SD) unless otherwise indicated, RS: Rotterdam Study, ERF: Erasmus Rucphen Family Study, n: number, TC: total cholesterol, TG: triglycerides, LLT: lipid lowering therapy *The percentages are calculated based on the 6874 non-diabetic and 861 diabetic individuals from the RS-I and RS-II cohorts, because follow-up data was only available for those cohorts.

The mean (SD) for the TG risk score was 0.026 (0.003) in the Rotterdam Study and 0.025 (0.003) in ERF.

Associations of the genetic risk scores with lipid levels and dyslipidemia are described in Table 2. All four risk scores were strongly associated with lipid levels in non-diabetic individuals ($P = 7.5 \times 10^{-168} - 1.8 \times 10^{-24}$). In individuals with diabetes, the TC score was not significantly associated with TC levels ($P = 0.138$). The LDL-C ($\beta = 0.189$, $P = 0.006$), HDL-C ($\beta = -0.041$, $P = 2.1 \times 10^{-5}$) and TG ($\beta = 0.121$, $P = 4.2 \times 10^{-15}$) risk scores, however, were associated with their corresponding lipid levels. In diabetic compared to non-diabetic individuals, directions of effect were consistent for all risk scores and effect estimates were not significantly different. In individuals without diabetes, the TC, LDL-C and TG risk scores were strongly associated with dyslipidemia (OR = 1.54 [1.34-1.77], 1.59 [1.39-1.83] and 1.33 [1.28-1.39], $P = 8.9 \times 10^{-10}$, 3.9×10^{-11} and 3.9×10^{-39}), but the HDL-C score was not ($P = 0.140$). In diabetes cases, only the TG score was associated with dyslipidemia ($P = 8.0 \times 10^{-5}$) with an OR of 1.40 per standard deviation increase in risk score. Odds

Table 2. Association of genetic lipid risk scores with lipid levels and dyslipidemia

		RS			ERF			Meta-analysis		
		β	S.E.	P-value	β	S.E.	P-value	β	S.E.	P-value
Lipids										
Controls	TC	0.287	0.012	3.7×10^{-130}	0.275	0.022	3.8×10^{-35}	0.284	0.010	7.5×10^{-168}
	LDL	0.261	0.011	9.8×10^{-111}	0.244	0.020	1.8×10^{-33}	0.257	0.010	1.2×10^{-146}
	HDL	−0.060	0.004	4.8×10^{-59}	−0.048	0.007	4.3×10^{-11}	−0.056	0.005	1.8×10^{-24}
	TG	0.116	0.005	4.3×10^{-112}	0.122	0.01	2.6×10^{-33}	0.117	0.005	5.7×10^{-148}
Cases	TC	0.376	0.036	9.9×10^{-25}	0.055	0.119	0.642	0.236	0.159	0.138
	LDL	0.234	0.032	5.8×10^{-13}	0.084	0.100	0.397	0.189	0.069	0.006
	HDL	−0.037	0.010	3.6×10^{-4}	−0.060	0.025	0.014	−0.041	0.010	2.1×10^{-5}
	TG	0.118	0.016	9.0×10^{-13}	0.145	0.048	0.003	0.121	0.015	4.2×10^{-15}
		OR	95% CI	P-value	OR	95% CI	P-value	OR	95% CI	P-value
Dyslipidemia										
Controls	TC	1.45	1.38-1.52	5.6×10^{-51}	1.66	1.49-1.86	2.9×10^{-20}	1.54	1.34-1.77	8.9×10^{-10}
	LDL	1.50	1.43-1.57	1.7×10^{-59}	1.73	1.55-1.93	1.1×10^{-22}	1.59	1.39-1.83	3.9×10^{-11}
	HDL	1.15	1.09-1.20	6.9×10^{-9}	1.02	0.92-1.13	0.726	1.09	0.97-1.22	0.140
	TG	1.32	1.26-1.39	2.05×10^{-31}	1.37	1.24-1.53	2.2×10^{-9}	1.33	1.28-1.39	3.9×10^{-39}
Cases	TC	1.89	1.63-2.19	1.7×10^{-17}	0.86	0.60-1.24	0.423	1.30	0.60-2.80	0.503
	LDL	1.85	1.60-2.14	1.5×10^{-16}	0.79	0.54-1.15	0.223	1.23	0.54-2.83	0.663
	HDL	1.07	0.94-1.22	0.285	1.07	0.76-1.54	0.679	1.07	0.95-1.21	0.251
	TG	1.46	1.28-1.67	2.4×10^{-8}	1.18	0.83-1.68	0.367	1.40	1.19-1.66	8.0×10^{-5}

RS: Rotterdam Study, ERF: Erasmus Rucphen Family Study, TC: total cholesterol, TG: triglycerides

ratios were not significantly different between individuals with and without diabetes after meta-analysis of the Rotterdam Study and ERF cohorts. In the Rotterdam Study alone, the largest cohort, the odds ratios for the TC and LDL-C risk score were significantly higher in diabetes cases compared to non-diabetic controls. Only in this cohort, the interaction between the genetic risk score and diabetes status was significant for TC ($\beta = 0.094$, $P = 0.009$) and marginally for HDL-C ($\beta = 0.022$, $P = 0.051$) and the LDL-C risk score*diabetes interaction term was associated with increased dyslipidemia risk (OR = 1.23 [1.06-1.44], $P = 0.008$).

Associations of single SNP*diabetes interactions with dyslipidemia are shown in Supplemental Table S1. After Bonferroni correction for multiple testing (significance thresholds: TC score $P = 6.7 \times 10^{-4}$, LDL score $P = 8.6 \times 10^{-4}$, HDL score $P = 6.8 \times 10^{-4}$, TG score $P = 1.1 \times 10^{-3}$), none of the interaction terms were associated with dyslipidemia. However, Kolmogorov-Smirnov tests showed significant deviation of the interaction P-values from the null distribution for all four sets of SNPs (TC SNPs: $P = 7.0 \times 10^{-7}$, LDL-C SNPs: $P = 2.8 \times 10^{-8}$, HDL-C SNPs: $P = 4.4 \times 10^{-4}$, TG SNPs: $P = 1.7 \times 10^{-5}$). SNPs in, or near GCKR, LIPG, ABCB11, NYNRIN, CTF1, APOB and FRK showed nominal P-values below 0.05 ($P = 0.004 – 0.038$).

Receiver operating characteristic curves for prediction of incident dyslipidemia in the Rotterdam Study are shown in Figure 1 and the corresponding areas under the receiver operating characteristic curves (AUCs) are depicted in Table 3. In both controls and diabetes cases, the clinical model including age, sex and BMI was a better predictor than
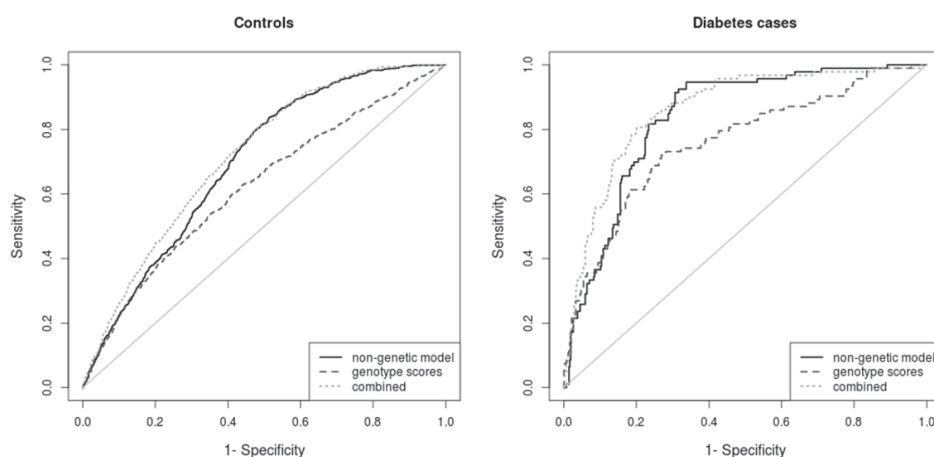
Figure 1. Receiver operating characteristic curves for prediction of dyslipidemia in the Rotterdam Study

Table 3. Prediction of dyslipidemia in the Rotterdam Study

Predictors	AUC [95% CI]		P-value*
	Controls	Cases	
age+sex+BMI	0.70 [0.68-0.72]	0.84 [0.80-0.88]	2.4×10^{-14}
LDL, HDL, TG genes	0.62 [0.60-0.65]	0.76 [0.71-0.82]	1.8×10^{-7}
combined†	0.72 [0.70-0.74]	0.86 [0.82-0.90]	2.8×10^{-15}

AUC: area under the receiver operating characteristic curve
*P-value for the difference in AUC between controls and cases
†age+sex+BMI+lipid genes scores

the gene based LDL-C, HDL-C and TG risk scores, while adding the risk scores to the non-genetic risk factors improved the AUCs (controls: $\Delta AUC = 0.02$, $P = 3.0 \times 10^{-5}$, cases: $\Delta AUC = 0.02$, $P = 0.012$). For all tested risk models, the AUCs in individuals with diabetes were significantly higher than the AUCs in non-diabetic individuals (clinical risk model: $\Delta AUC = 0.14$, $P = 2.4 \times 10^{-14}$, genetic risk model: $\Delta AUC = 0.14$, $P = 1.8 \times 10^{-7}$, combined model: $\Delta AUC = 0.14$, $P = 2.8 \times 10^{-15}$).

DISCUSSION

In patients with diabetes, the LDL-C, HDL-C and TG genetic risk scores were associated with their corresponding lipid levels, while only the TG score was associated with dyslipidemia. Effect estimates were not significantly different between diabetes cases and non-diabetic controls. In both controls and diabetes cases, adding the LDL-C, HDL-C and TG genotype scores to a non-genetic risk model significantly improved the AUCs for prediction of incident dyslipidemia. All risk models (genetic, non-genetic and both combined) performed significantly better in diabetic than in non-diabetic individuals.

The differences in the ability of the genetic risk scores to predict incident dyslipidemia in individuals with and without diabetes in our study suggest that the role of lipid-altering gene variants might be modified in the context of diabetes. This could be based on perturbations in ApoB and subsequent VLDL production caused by insulin effects on the liver. Results of analyses to test whether the differences might be explained by genetic risk score*diabetes interactions were inconclusive. However, some findings suggest that it might to be of interest to further investigate these interactions in other cohorts. These findings include significant associations of the interaction terms with lipid levels and dyslipidemia in the Rotterdam Study and, in terms of single SNP*diabetes interaction analyses, significant enrichment for low P-values. Further, the lowest single SNP*diabetes interaction P-values were found for TC, HDL-C and TG SNPs in the GCKR and LIPG loci. Interestingly, associations of these loci with type 2 diabetes and type 2 diabetes related

traits have been described. SNPs in *GCKR* were previously associated with type 2 diabetes, fasting glucose and fasting insulin[28] and *LIPG* was found under a type 2 diabetes linkage peak[29]. Additionally, association of *LIPG**overweight/obesity interaction with ApoB, an important component in diabetic dyslipidemia, has been reported[30].

In addition to an altered role of the SNPs, or lack of power in diabetes cases due to the relatively small sample size, a possible explanation for the lack of association between the TC gene risk score and TC levels and the LDL-C gene risk score with dyslipidemia in diabetes cases might be the cumulative SNP effects over time in these individuals. The Rotterdam Study and ERF differ substantially in mean age, which is much lower in ERF. Therefore, differences in SNP effects over time might explain heterogeneity between the analysis results of the two cohorts and subsequent non-significant meta-analysis results.

The high AUCs for prediction of dyslipidemia in individuals with type 2 diabetes compared to non-diabetic individuals is not only interesting from a biological point of view, but also from a clinical one; it raises the question of whether the genetic risk scores could be used for prognostic purposes in diabetic individuals. Currently, the clinical model including the easy to obtain variables age, sex and BMI predicts dyslipidemia better than the genetic risk scores and the small added value of the genetic risk scores to the clinical model (ΔAUC=0.02), although significant, won't be useful in clinical practice. However, if more lipid-altering genetic variants are discovered, explaining more of the heritability of these traits, it is likely that the predictive ability of genetic lipid risk scores will further increase.

Associations of genetic risk scores comprised of common lipid-altering variants with their corresponding lipid levels, extreme lipid values and intervention thresholds for blood lipids have been previously described[31,32]. In the Rotterdam Study and ERF, we previously showed association of the genetic risk scores based on the 95 lipid loci reported by Teslovich et al with their respective lipid levels[12]. To our knowledge, however, this is the first study assessing the influence of genetic risk scores for lipid levels in individuals with type 2 diabetes.

A strength of our study is the inclusion of all 157 currently known common lipid-altering gene variants, including 62 recently discovered loci[10]. Further strengths are the use of population-based cohorts for the assessment of the genetic risk scores and the weighting of the SNPs included in the risk scores by the effect estimates obtained in the discovery analyses. This gives an indication of the performance of risk scores, based on common variants and their effect estimates resulting from large GWAS, in the general population. Since both ERF and the Rotterdam Study were part of the discovery GWAS

meta-analyses, a limitation of this study might be that the risk scores are not completely independent from the GWAS results. Because of the very large number of individuals included in the discovery meta-analyses (>188,000), however, this effect should be limited. Another shortcoming in our study is the comparatively low number of individuals with type 2 diabetes, particularly in the ERF cohort, which likely explains the wider confidence intervals resulting from the analyses in these individuals. Furthermore, since longitudinal data is not available for ERF, the discriminative ability of the risk scores for incident dyslipidemia was only assessed in the elderly of the Rotterdam Study. This may limit the generalizability of these findings.

In conclusion, our results show the cumulative influence of common lipid-altering genetic variants on lipid levels in patients with type 2 diabetes. The AUC values for prediction of dyslipidemia suggest that the role of common genetic variation may be modified in the context of diabetes. The high AUCs of the lipid gene risk scores for prediction of dyslipidemia in individuals with type 2 diabetes raise the possibility that, if more genetic variants, including those with lower frequencies and commensurately larger effects, are discovered, these scores could be used for prognostic purposes.

ACKNOWLEDGMENTS

REFERENCES

1. IDF Diabetes Atlas. The Global Burden. http://www.idf.org/diabetesatlas/5e/the-global-burden (July 2012)

2. European Association for Cardiovascular Prevention & Rehabilitation, Reiner, Z. et al. ESC/EAS Guidelines for the management of dyslipidaemias: the Task Force for the management of dyslipidaemias of the European Society of Cardiology (ESC) and the European Atherosclerosis Society (EAS). Eur Heart J 32, 1769-818 (2011).

3. Scott, R. et al. Effects of fenofibrate treatment on cardiovascular disease risk in 9,795 individuals with type 2 diabetes and various components of the metabolic syndrome: the Fenofibrate Intervention and Event Lowering in Diabetes (FIELD) study. Diabetes Care 32, 493-8 (2009).

4. Goldberg, I.J. Clinical review 124: Diabetic dyslipidemia: causes and consequences. J Clin Endocrinol Metab 86, 965-71 (2001).

5. Krauss, R.M. Lipids and lipoproteins in patients with type 2 diabetes. Diabetes Care 27, 1496-504 (2004).

6. Haffner, S.M., Lehto, S., Ronnemaa, T., Pyorala, K. & Laakso, M. Mortality from coronary heart disease in subjects with type 2 diabetes and in nondiabetic subjects with and without prior myocardial infarction. N Engl J Med 339, 229-34 (1998).

7. Howard, B.V. et al. LDL cholesterol as a strong predictor of coronary heart disease in diabetic individuals with insulin resistance and low LDL: The Strong Heart Study. Arterioscler Thromb Vasc Biol 20, 830-5 (2000).

8. Chapman, M.J. Metabolic syndrome and type 2 diabetes: lipid and physiological consequences. Diab Vasc Dis Res 4 Suppl 3, S5-8 (2007).

9. Teslovich, T.M. et al. Biological, clinical and population relevance of 95 loci for blood lipids. Nature 466, 707-13 (2010).

10. Global Lipids Genetics Consortium, Willer, C.J. et al. Discovery and refinement of loci associated with lipid levels. Nat Genet 45,1274-83 (2013).

11. Bos, D. et al. Genetic loci for coronary calcification and serum lipids relate to aortic and carotid calcification. Circ Cardiovasc Genet 6, 47-53 (2013).

12. Isaacs, A. et al. Risk Scores of Common Genetic Variants for Lipid Levels Influence Atherosclerosis and Incident Coronary Heart Disease. Arterioscler Thromb Vasc Biol 33, 2233-39 (2013).

13. Hofman, A. et al. The Rotterdam Study: 2010 objectives and design update. Eur J Epidemiol 24, 553-72 (2009).

14. Hofman, A., Grobbee, D.E., de Jong, P.T. & van den Ouweland, F.A. Determinants of disease and disability in the elderly: the Rotterdam Elderly Study. Eur J Epidemiol 7, 403-22 (1991).

15. Pardo, L.M., MacKay, I., Oostra, B., van Duijn, C.M. & Aulchenko, Y.S. The effect of genetic drift in a young genetically isolated population. Ann Hum Genet 69, 288-95 (2005).

16. Friedewald, W.T., Levy, R.I. & Fredrickson, D.S. Estimation of the concentration of low-density lipoprotein cholesterol in plasma, without use of the preparative ultracentrifuge. Clin Chem 18, 499-502 (1972).

17. Baigent, C. et al. Efficacy and safety of cholesterol-lowering treatment: prospective meta-analysis of data from 90,056 participants in 14 randomised trials of statins. Lancet 366, 1267-78 (2005).

18. Birjmohun, R.S., Hutten, B.A., Kastelein, J.J. & Stroes, E.S. Efficacy and safety of high-density lipoprotein cholesterol-increasing compounds: a meta-analysis of randomized controlled trials. J Am Coll Cardiol 45, 185-97 (2005).

19. Randomised trial of cholesterol lowering in 4444 patients with coronary heart disease: the Scandinavian Simvastatin Survival Study (4S). Lancet 344, 1383-9 (1994).

20. Miller, S.A., Dykes, D.D. & Polesky, H.F. A simple salting out procedure for extracting DNA from human nucleated cells. Nucleic Acids Res 16, 1215 (1988).

21. Li, Y., Willer, C.J., Ding, J., Scheet, P. & Abecasis, G.R. MaCH: using sequence and genotype data to estimate haplotypes and unobserved genotypes. Genet Epidemiol 34, 816-34 (2010).

22. R Core Team (2012). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. http://www.R-project.org/

23. Almasy, L. & Blangero, J. Multipoint quantitative-trait linkage analysis in general pedigrees. Am J Hum Genet 62, 1198-211 (1998).

24. Carey, V.J., Ported to R by Lumley, T. and Ripley, B. (2012). gee: Generalized Estimation Equation solver. R package version 4.13-18. http://CRAN.R-project.org/package=gee

25. Lumley, T. (2009) rmeta: Meta-analysis. R package version 2.16. http://CRAN.R-project.org/package=rmeta.

26. Kundu, S., Aulchenko, Y.S., Janssens, A.C. (2011). PredictABEL: Assessment of risk prediction models. R package version 1.2. http://CRAN.R-project.org/package=PredictABEL

27. Robin, X. et al. pROC: an open-source package for R and S+ to analyze and compare ROC curves. BMC Bioinformatics 12, 77 (2011).

28. Dupuis, J. et al. New genetic loci implicated in fasting glucose homeostasis and their impact on type 2 diabetes risk. Nat Genet 42, 105-16 (2010).

29. Hasstedt, S.J. et al. Five linkage regions each harbor multiple type 2 diabetes genes in the African American subset of the GENNID Study. J Hum Genet 58, 378-83 (2013).

30. Yin, R.X. et al. Several genetic polymorphisms interact with overweight/obesity to influence serum lipid levels. Cardiovasc Diabetol 11, 123 (2012).

31. Kathiresan, S. et al. Polymorphisms associated with cholesterol and risk of cardiovascular events. N Engl J Med 358, 1240-9 (2008).

32. Murray, A. et al. Common lipid-altering gene variants are associated with therapeutic intervention thresholds of lipid levels in older people. Eur Heart J 30, 1711-9 (2009).

Chapter 5

General discussion

In this thesis, I sought to improve our understanding of the etiology of type 2 diabetes (T2D) and dyslipidemia and to investigate the extent to which genetic risk factors, combined, can improve their prediction. The thesis describes a search for genetic variants associated with T2D, and the related quantitative traits fasting glucose (FG) and insulin (FI), in the general population and the performance of genetic risk scores in the identification of individuals at increased risk for T2D and dyslipidemia. This chapter summarizes the findings of the thesis and describes their implications for future research.

FINDINGS OF THIS THESIS

From gene discovery to understanding and predicting type 2 diabetes

Chapter 2 starts with a review of published genetic risk prediction studies for T2D, assessed from a methodological perspective. In this review I showed that, almost without exception, the genetic risk models had lower AUC values than the clinical models and that addition of genetic factors either did not or only marginally improved the AUC beyond that of clinical risk models. I also showed that study design and population characteristics may have affected the observed predictive performance of the risk models. AUC values of the clinical risk models were higher and, although weaker, AUC values of the genetic risk models were lower when there were larger differences in age and BMI between cases and controls. This observation has important implications for the design and health care relevance of genetic risk prediction studies. The predictive ability of risk models is preferably investigated in prospective cohort studies, but often, in practice, only case-control or cross-sectional designs are available. Because clinical risk factors, particularly the difference in risk factors between cases and controls, impact AUC values, it is expected that AUC values for genetic risk models obtained in case-control or cross-sectional studies may be valid when the distribution of these risk factors does not differ from prospective studies. For case-control studies, this means that the selection of cases and controls is not affected by these risk factors. In case of selection, transparency about the methods is important to enable a correct interpretation of the scientific and health care relevance of the results. For this reason, the GRIPS (Genetic Risk Prediction Studies) statement, a guideline for the reporting of genetic risk prediction studies, recommends description of the eligibility criteria for participants, and the sources and methods of participant selection[1]. The observed impact of population characteristics implies that it is important to assess the predictive ability of risk scores in representative samples of the population in which the model is ultimately going to be applied to get valid estimates of their performance in that population. The question then is: which populations do we want to target for the prevention of T2D? For genetic prediction, this might concern young individuals who have not yet developed clinical risk factors and that

could be referred to a personalized prevention program. None of the T2D risk prediction studies included in our review were conducted in younger populations; all studies were conducted in populations who on average were older than 40 years of age[2-20], including two in populations over 60 years of age[13,17].

Two studies, published after the publication of our review, assessed whether genetic risk scores improve clinical prediction models for incident diabetes from adolescence[21] and young adulthood[22]. Similar to the studies included in our review, genetic factors either did not or only marginally improved prediction beyond clinical risk models. However, variants included in the risk scores only explained a small portion of the heritability of type 2 diabetes. If sequencing efforts identify the causal variants for which several of the GWAS SNPs are imperfect proxies, and if further gene discovery efforts identify additional common, low frequency and rare variants associated with T2D, inclusion of those variants might improve the ability of genetic risk scores to detect individuals at increased risk of T2D.

In this chapter, we further focused on identifying potential causal rare and low frequency variants associated with diabetes risk and the related quantitative traits fasting glucose and fasting insulin. In chapter 2.2, we sought to obtain more insight into GWAS findings near the *IGF1* gene and to identify and characterize novel genetic variants at the locus. This study suggested that previously observed associations between SNPs near *IGF1* with FI levels were not mediated by circulating IGF-I levels. Investigation of the *IGF1* gene using deep sequencing data, revealed a large number of rare variants at the locus that had not been previously described, most of which were very rare. A subset of rare non-synonymous variants, including six novel variants and five variants that had been previously identified, were significantly associated with FI levels. Conditional analysis suggested that the common non-coding variants near *IGF1* that were identified through GWAS[23,24] explain part of the rare variant signal and the presence of a residual independent rare variant effect. Examination of regulatory annotation using ENCODE data showed that the GWAS variants were located in the proximity of FOXA1 binding and DNaseI hypersensitive sites, suggesting that they might have a direct functional role. This finding is noteworthy because FOXA1 is a key transcriptional regulator implicated in glucose metabolism and insulin secretion[25,26]. Studies of human cell cultures and animal models will be needed to interrogate and validate the function of these non-coding variants in insulin biology.

One variant, rs151098426, resulting in an alanine to threonine substitution and predicted to be damaging by several annotation tools, seemed to drive the rare variant association.

However, follow-up genotyping of rs151098426 in an independent set of samples, and lookup of the variant in CHARGE exome chip results did not reveal significant differences in FI levels between carriers and non-carriers of the rare allele, suggesting the absence of a single variant effect for rs151098426 on FI levels. Several recently published studies have demonstrated the need for large sample sizes to robustly identify associations of low frequency variants with complex traits[27-33]. Because of the low minor allele frequency (MAF) of rs151098426 and thus the relatively small number of carriers, analyzing the variant in large numbers of additional samples will be required to definitively conclude whether it is associated with FI levels.

The identification of variants at the *IGF1* locus that had not been previously described has increased our insight into the variation present at the locus. In line with previous sequencing studies[28,34,35], we identified a large number of very rare variants, the major-ity (64%) observed only one time in our samples. The presence of large numbers of very rare variants in the human genome is likely explained by recent explosive human population growth[34,36]. It has been hypothesized that these variants might harbor larger effects than those observed for common variants, since selection could have influenced only the most deleterious variants[34]. However, even for rare variants with larger effects, large sample sizes are needed to definitely conclude whether they influence complex traits due to the low MAF.

In the CHARGE consortium, we further sought rare and low-frequency variants associ-ated with FG, FI and T2D using the HumanExome BeadChip (chapter 2.3). The FG meta-analysis in up to 60,564 individuals and follow up in up to 14,884 T2D cases and 78,097 controls identified a novel low-frequency (minor allele frequency 1.4%) nonsynonymous SNP in the gene encoding the receptor for glucagon-like peptide 1 (*GLP1R*) associated with FG and T2D. The minor allele was associated with lower FG, lower T2D risk, lower insulin response during a glucose challenge and higher 2-h glucose, pointing to physi-ological effects on the incretin system. This association represents another instance in which genetic epidemiology identified a gene that codes for a direct drug target in T2D; other examples include ABCC8/KCNJ11 (encoding the target for sulfonylureas) and PPARG (encoding the target of thiazolidinediones). In these examples, the drug preceded the genetic discovery. Given these findings, the large number of variants asso-ciated with T2D and glycemic traits today likely represent a promising long-term source of potential targets for future diabetes therapies. Further, at known FG, FI and T2D loci, we found associations with nonsynonymous SNPs in 6 FG and/or FI loci (G6PC2, *COBLL1*, *GPSM1*, *SLC2A2*, *SLC30A8* and *RREB1*) and in 5 T2D loci (*ARAP1*, *GIPR*, *KCNJ11*, *SLC30A8* and *WFS1*) as potential causal candidates underlying previously identified associations. For example, gene-based analyses revealed a set of rare nonsynonymous variants in

G6PC2 which are independent of the common non-coding FG variants at this locus and implicate this gene as underlying previously identified associations. We also revealed non-coding variants whose putative functions in epigenetic and post-transcriptional regulation of *ABO* and *G6PC2* are supported by experimental ENCODE Consortium, GTEx and transcriptome data from islets and for which future focused investigations using human cell culture and animal models will be needed to clarify their functional influence on glycemic regulation.

This effort has shown the utility of analyzing the variants present on the exome chip, followed by regulatory annotation and additional phenotypic characterization, in revealing novel genetic effects on glycemic homeostasis. Furthermore, analyzing coding variation in known loci assisted in identifying potential causal genes influencing glucose metabolism and T2D risk, and increased understanding of the genetic architecture of diabetes-related quantitative traits and T2D. While the effect size of the novel low-frequency variant in *GLP1R* on fasting glucose is slightly larger than for most loci reported to date[37], our findings suggest that few low frequency variants exert a very large influence on glycemic traits and further demonstrate the need for large sample sizes to identify associations of low frequency variation with complex traits[27,30,31,38-40,33]. Therefore, for several cardiometabolic traits, exome chip efforts are underway combining large numbers of samples within and across consortia to increase power for detecting rare variant associations. This includes an effort in the MAGIC consortium, which will include an estimated 120,000 individuals with data on quantitative glycemic traits.

The role of lipid-altering gene variants in dyslipidemia and cardiovascular disease

Chapter 3 focuses on the ability of risk scores comprised of known common genetic variants influencing lipid levels to identify individuals at high risk of dyslipidemia and the association of those risk scores with subclinical atherosclerosis and incident coronary heart disease. In chapter 3.1, we showed that common genetic variants play an important role in determining who develops dyslipidemia between the ages of 55 and 90 years in normal weight and overweight individuals. Despite the small effects of the common genetic variants on lipid levels individually, when combined into a risk score the effects were substantial and discriminated future patients with dyslipidemia better than an established epidemiological risk factor such as BMI. This underscores the value of genome wide association studies for age-related diseases.

A next step, after the finding that common genetic variants play an important role in determining who develops dyslipidemia through old age, is to determine whether screening for these variants to identify individuals at increased risk of dyslipidemia in

the general population would be valuable. Genetic screening has proven useful in familial hypercholesterolemia (FH)[41], but, despite the large role of genes in both conditions, there are important differences between FH and dyslipidemia in the general population. FH is an autosomal co-dominant disorder with nearly complete penetrance[42], with a prevalence of 1:500 in most Western countries. Left untreated, men experience clinical symptoms of CVD typically in their fourth decade and women in their fifth decade of life. Statin therapy lowers CVD risk significantly in these individuals[43]. Although the prevalence of dyslipidemia in the general population is much higher than 1:500, even in individuals in the highest TC gene risk score quartile penetrance is not complete. Yet in this high risk group prevalence approaches 70% between the ages of 65 and 74 years. To be able to answer the question whether screening would be useful, and from what age onwards, age specific penetrance of dyslipidemia over a wider age range than assessed in our study should be investigated. Further, corresponding morbidity and mortality which could be prevented by early identification of individuals at high genetic risk should be assessed.

In Chapter 3.2, we showed that risk scores comprised of common genetic lipid influencing variants were associated with the more distant, but clinically very relevant, atherosclerosis and incident coronary heart disease outcomes. All genetic risk scores that we analyzed (for TC, LDL-C, HDL-C, and TG) were associated with carotid plaque and the TC and LDL-C scores were robustly associated with incident coronary heart disease. This supports the hypothesis that, in addition to common and rare variants found to be associated with cardiovascular disease individually, aggregations of common genetic variants influencing lipid levels play a significant role in the development of atherosclerosis and the subsequent occurrence of cardiovascular disease. The risk scores did not improve clinical AUCs for prediction of CHD. However, associations of the TC and LDL-C risk scores with plaque and CHD were still (borderline) significant after inclusion of the lipid levels corresponding to the genetic risk score in the models, suggesting added value of the risk scores beyond the lipid levels themselves. This may reflect lifelong exposure to higher lipid values, in contrast to fluctuating occasional lipid measurements, which would make genes relevant for early prediction and prevention purposes. With ongoing efforts in gene discovery, it is likely that many more variants will be discovered, both for the outcome cardiovascular disease itself and for intermediate phenotypes such as lipid levels. The identification of these might improve the ability to detect individuals at increased risk for cardiovascular disease.

The role of lipid-altering gene variants in the context of type 2 diabetes
In chapter 4, we tested the hypothesis that in individuals with type 2 diabetes, common genetic lipid variants identified in the general population also influence their lipid levels

despite an altered metabolic milieu. Meta-analysis results from the Rotterdam Study and ERF showed that in patients with diabetes, the LDL-C, HDL-C and TG genetic risk scores were significantly associated with their corresponding lipid levels and that the TG score was associated with prevalent dyslipidemia. Effect estimates were not significantly different between diabetes cases and non-diabetic controls. However, the ability of the genetic risk scores to predict incident dyslipidemia in the Rotterdam Study was significantly higher in cases than in controls, suggesting that the role of lipid-altering genetic variants might be modified in the context of diabetes. Results of analyses to test whether the differences might be explained by genetic risk score*diabetes interactions were inconclusive. Our findings suggest that it might be of interest to further investigate these interactions in other cohorts. These findings include significant associations of interaction terms with lipid levels and dyslipidemia in the Rotterdam Study and, in single SNP*diabetes interaction analyses, significant enrichment for low p-values. To further increase our understanding of the role of genetic variation in diabetic dyslipidemia, GWAS on lipid levels in large numbers of individuals with type 2 diabetes will be another interesting follow-up step. This can further increase insight into the overlap, and differences, in genetic risk factors for dyslipidemia in individuals with and without type 2 diabetes and thereby further increase our understanding of this important CVD risk factor.

FUTURE RESEARCH

Gene Finding

For CVD and T2D in particular, but also for dyslipidemia, a large portion of the estimated heritability remains unexplained. Both to improve the understanding of disease etiology, which can lead to identification of novel targets for therapeutic interventions, and to improve the identification of individuals at increased risk of disease, it is important to find the explanation for the "missing heritability".

Results from several studies analyzing complex traits suggest that the largest portion of the heritability is attributable to common genetic variation[35,44,45]. The largest T2D GWAS meta-analysis published to date, including 34,840 cases and 114,981 controls, brought the total number of independent T2D associated loci to 65, together explaining ~5.7% of the variance in disease susceptibility[44]. Additional analyses suggested that 63% of T2D susceptibility might be explained by common genetic variants, most of very modest effect. A recent study by the CHARGE consortium estimated the contributions of variants within different MAF bins to heritability for HDL-C[35] . In this study, using whole-genome sequence data, 62% of the variance in HDL-C levels was estimated to be explained by

variants with MAF>1% and rare variants (MAF < 1%) explained an additional 8% of the variance. It is therefore likely that larger sample sizes will identify more common genetic variants, most of very modest effect. As further discussed below, adding these variants to genetic risk scores might improve the identification of individuals at increased risk of disease.

An important question, as the number of disease susceptibility loci increases, will be whether disease etiology implicated by genetic variants will coalesce around a limited set of core pathways. The currently identified T2D associated variants suggest that this may be the case, pointing to beta cell function, cell cycle regulation, adipocytokine signalling and CREBBP-related transcription factor activity as key processes involved in the pathogenesis of T2D[44].

Other possible explanations for missing heritability are gene*gene and gene*environment interactions[46,47]. The findings in chapter 4 suggest the potential influence of SNP*diabetes interactions on dyslipidemia. Another recent study, including 12,403 incident T2D cases and a representative sub-cohort of 16,154 European individuals, investigated interactions between an additive genetic T2D risk score and modifiable and non-modifiable risk factors on T2D risk [48]. Significant interactions were identified between the genetic score and age, BMI and WHR: the relative genetic risk was larger in younger participants and in participants who were leaner. In this study, similar to our study in chapter 4, interactions with genetic variants that were known to be individually associated with the outcome were investigated. Genome-wide gene*gene and gene*environment interaction studies could reveal novel genetic variants that only exert a substantial effect on disease in interaction with other genetic or environmental factors. However, especially for genome wide gene*gene interaction studies, the increased multiple testing burden will reduce statistical power. Detecting an influence of these interactions on disease risk, therefore, will be challenging.

Despite evidence for a limited contribution of rare variants to complex trait heritability[35,45], the identification of these variants can increase our understanding of disease etiology. This is illustrated in Chapter 2.3 where we identified a novel low-frequency (MAF 1.4%) nonsynonymous SNP associated with FG and T2D in the gene encoding the receptor for GLP1. This variant only explained 0.03% of the variance in FG, but is of large relevance: it is located in a gene that codes for a direct drug target in T2D. Our findings in chapter 2.3 also suggested that few low frequency variants exert a very large influence on glycemic traits and further demonstrated the need for large sample sizes to identify associations of low frequency variation with complex traits[27-32]. A recently described conceptual framework for the design of rare variant association studies also highlighted

the need for large sample sizes to detect association[33]. As described above, for several cardiometabolic traits, exome chip efforts are underway combining large numbers of samples within and across consortia to increase power for detecting low frequency and rare variant associations.

Because rare variant association studies are still in their early stages, several questions about optimal analysis methods remain unanswered. Zuk et al.[33] address some fundamental questions in the design of rare variant association studies, including the choice of variants to include in gene-based analyses. In this study a binary trait is considered, analyzed with burden tests, where all alleles are either null or neutral. Although this method is simplified, it provides insight into many alternative methods for rare variant association studies. Three different sets of criteria for inclusion of rare variants in gene based tests are considered: 1) inclusion of disruptive alleles only, 2) adding missense alleles filtered by frequency, 3) adding missense alleles filtered by severity. In the CHARGE consortium, an ongoing effort is comparing results from gene-based analyses applying these criteria for inclusion of variants in the tests for the quantitative traits FG and FI. Preliminary results show clear differences in *P*-values for several genes depending on the choice of variants included in the gene-based tests, suggesting that a more comprehensive SNP annotation may lead to the discovery of additional genes and rare variants influencing complex traits.

Genetic risk prediction

The large proportion of variance of common complex traits that might be explained by common genetic variants, most of very modest effect, suggests that the predictive ability of genetic risk models could substantially improve when these variants are included in the model. A recent study comprehensively assessing the performance of risk prediction based on polygenic models found that several factors influence their predictive ability[41]. These include the total heritability of the trait, the underlying effect size distributions, the sample size for the training data set, and the *P*-value threshold for inclusion of the SNPs in the model. For T2D, as described in chapter 2.1, current genetic risk prediction models typically have AUCs around 0.60. Chatterjee et al[41] estimated that a polygenic model based on a training data set including ~200.000 samples, using the optimal *P*-value threshold for inclusion of SNPs in the model, will increase the AUC for prediction of T2D to 0.79. Using this model, the estimated proportion of T2D cases among the 20% of subjects with the highest polygenic risk was 0.63[41]. In chapter 3.1, we showed the substantial role of genetic risk scores, comprised of variants influencing lipid levels, in determining who develops dyslipidemia. The findings of Chatterjee et al[41] suggest that genetic risk scores for T2D could also have a substantial role in identifying individuals at increased disease risk.

When more variants influencing complex traits are discovered, studies assessing the predictive ability of the updated risk scores will be needed. The observation that study design and population characteristics may affect the observed predictive performance of risk models, as described in more detail in chapter 2, implies that it is important to assess the predictive ability in representative samples of the population in which the model is ultimately to be applied to get valid estimates of their performance in that population. Because clinical risk factors, particularly the difference in risk factors between cases and controls, impact AUC values, prospective cohort studies are the preferred design. In these studies, selection of cases and controls will not be affected by these risk factors.

As described in chapter 3.1, a next step after the finding that genetic variants play an important role in determining who develops disease, is to determine whether screening for these variants to identify individuals at increased disease risk in the general population would be useful. This requires a medical economics evaluation, investigating both health benefits and costs related to implementation of a genetic screening program.

CONCLUSION

In this thesis, I sought to improve our understanding of the etiology of T2D and dyslipidemia and to investigate the extent to which genetic risk factors combined can improve their prediction. The main findings are the identification of a large number of novel rare variants in the *IGF1* gene, which has improved insight into the genetic variation present at this locus, the finding that both rare nonsynonymous variants and common non-coding variants with putative function in and near the *IGF1* gene may have a role in insulin biology (Chapter 2.2), and the identification of a low-frequency variant associated with FG and T2D in *GLP1R*, a gene encoding a direct drug target for T2D (Chapter 2.3). Further, while the currently identified common T2D associated genetic variants do not yet improve prediction of T2D compared to clinical prediction models, Chapter 3.1 provides evidence that genetic risk scores for lipid levels play a large role in determining who develops dyslipidemia through old age. Finally, in Chapter 4, I show that the role of common genetic variants associated with lipid levels may be modified in the context of diabetes. Several efforts are underway which may further increase our understanding of the etiology of T2D and dyslipidemia and the identification of individuals at high risk of developing disease. These include common and rare variant analyses using larger sample sizes, efforts to improve insight into methods for rare variant analyses which may lead to better identification of genes and rare variants associated with disease, and gene*gene and gene*environment studies. These efforts may bring us closer to the translation of complex disease genetics into clinical practice by identifying potential

drug targets and improving genetic risk prediction, which could reduce the prevalence of the CVD risk factors T2D and dyslipidemia and thus the burden of CVD.

REFERENCES

1. Janssens, A.C. et al. Strengthening the reporting of genetic risk prediction studies: the GRIPS statement. Eur J Epidemiol 26, 255-9 (2011).

2. Balkau, B. et al. Predicting diabetes: clinical, biological, and genetic approaches: data from the Epidemiological Study on the Insulin Resistance Syndrome (DESIR). Diabetes Care 31, 2056-61 (2008).

3. Cauchi, S. et al. Post genome-wide association studies of novel genes associated with type 2 diabetes show gene-gene interaction and high predictive value. PLoS One 3, e2031 (2008).

4. Cornelis, M.C. et al. Joint effects of common genetic variants on the risk for type 2 diabetes in U.S. men and women of European ancestry. Ann Intern Med 150, 541-50 (2009).

5. de Miguel-Yanes, J.M. et al. Genetic risk reclassification for type 2 diabetes by age below or above 50 years using 40 type 2 diabetes risk single nucleotide polymorphisms. Diabetes Care 34, 121-5 (2011).

6. Fontaine-Bisson, B. et al. Evaluating the discriminative power of multi-trait genetic risk scores for type 2 diabetes in a northern Swedish population. Diabetologia 53, 2155-62 (2010).

7. Hu, C. et al. PPARG, KCNJ11, CDKAL1, CDKN2A-CDKN2B, IDE-KIF11-HHEX, IGF2BP2 and SLC30A8 are associated with type 2 diabetes in a Chinese population. PLoS One 4, e7643 (2009).

8. Lango, H. et al. Assessing the combined impact of 18 common genetic variants of modest effect sizes on type 2 diabetes risk. Diabetes 57, 3129-35 (2008).

9. Lin, X. et al. Risk prediction of prevalent diabetes in a Swiss population using a weighted genetic score--the CoLaus Study. Diabetologia 52, 600-8 (2009).

10. Lyssenko, V. et al. Genetic prediction of future type 2 diabetes. PLoS Med 2, e345 (2005).

11. Lyssenko, V. et al. Clinical risk factors, DNA variants, and the development of type 2 diabetes. N Engl J Med 359, 2220-32 (2008).

12. Meigs, J.B. et al. Genotype score in addition to common risk factors for prediction of type 2 diabetes. N Engl J Med 359, 2208-19 (2008).

13. Miyake, K. et al. Construction of a prediction model for type 2 diabetes mellitus in the Japanese population based on 11 genes with strong evidence of the association. J Hum Genet 54, 236-41 (2009).

14. Schulze, M.B. et al. Use of multiple metabolic and genetic markers to improve the prediction of type 2 diabetes: the EPIC-Potsdam Study. Diabetes Care 32, 2116-9 (2009).

15. Sparso, T. et al. Combined analysis of 19 common validated type 2 diabetes susceptibility gene variants shows moderate discriminative value and no evidence of gene-gene interaction. Diabetologia 52, 1308-14 (2009).

16. Talmud, P.J. et al. Utility of genetic and non-genetic risk factors in prediction of type 2 diabetes: Whitehall II prospective cohort study. BMJ 340, b4838 (2010).

17. van Hoek, M. et al. Predicting type 2 diabetes based on polymorphisms from genome-wide association studies: a population-based study. Diabetes 57, 3122-8 (2008).

18. Vaxillaire, M. et al. Impact of common type 2 diabetes risk polymorphisms in the DESIR prospective study. Diabetes 57, 244-54 (2008).

19. Wang, J., Stancakova, A., Kuusisto, J. & Laakso, M. Identification of undiagnosed type 2 diabetic individuals by the finnish diabetes risk score and biochemical and genetic markers: a population-based study of 7232 Finnish men. J Clin Endocrinol Metab 95, 3858-62 (2010).

20. Weedon, M.N. et al. Combining information from common type 2 diabetes risk polymorphisms improves disease prediction. PLoS Med 3, e374 (2006).

21. Vassy, J.L. et al. Genotype prediction of adult type 2 diabetes from adolescence in a multiracial population. Pediatrics 130, e1235-42 (2012).

22. Vassy, J.L. et al. A genotype risk score predicts type 2 diabetes from young adulthood: the CARDIA study. Diabetologia 55, 2604-12 (2012).

23. Dupuis, J. et al. New genetic loci implicated in fasting glucose homeostasis and their impact on type 2 diabetes risk. Nat Genet 42, 105-16 (2010).

24. Manning, A.K. et al. A genome-wide approach accounting for body mass index identifies genetic variants influencing fasting glycemic traits and insulin resistance. Nat Genet 44, 659-69 (2012).

25. Gao, N. et al. Foxa1 and Foxa2 maintain the metabolic and secretory features of the mature beta-cell. Mol Endocrinol 24, 1594-604 (2010).

26. Kaestner, K.H. The FoxA factors in organogenesis and differentiation. Curr Opin Genet Dev 20, 527-32 (2010).

27. Carty, C.L. et al. Replication of genetic loci for ages at menarche and menopause in the multi-ethnic Population Architecture using Genomics and Epidemiology (PAGE) study. Hum Reprod 28, 1695-706 (2013).

28. Flannick, J. et al. Loss-of-function mutations in SLC30A8 protect against type 2 diabetes. Nat Genet 46, 357-63 (2014).

29. Holmen, O.L. et al. Systematic evaluation of coding variation identifies a candidate causal variant in TM6SF2 influencing total cholesterol and myocardial infarction risk. Nat Genet 46, 345-51 (2014).

30. Huyghe, J.R. et al. Exome array analysis identifies new loci and low-frequency variants influencing insulin processing and secretion. Nat Genet 45, 197-201 (2013).

31. Kozlitina, J. et al. Exome-wide association study identifies a TM6SF2 variant that confers susceptibility to nonalcoholic fatty liver disease. Nat Genet 46, 352-6 (2014).

32. Peloso, G.M. et al. Association of low-frequency and rare coding-sequence variants with blood lipids and coronary heart disease in 56,000 whites and blacks. Am J Hum Genet 94, 223-32 (2014).

33. Zuk, O. et al. Searching for missing heritability: designing rare variant association studies. Proc Natl Acad Sci U S A 111, E455-64 (2014).

34. Coventry, A. et al. Deep resequencing reveals excess rare recent variants consistent with explosive population growth. Nat Commun 1, 131 (2010).

35. Morrison, A.C. et al. Whole-genome sequence-based analysis of high-density lipoprotein cholesterol. Nat Genet 45, 899-901 (2013).

36. Keinan, A. & Clark, A.G. Recent explosive human population growth has resulted in an excess of rare genetic variants. Science 336, 740-3 (2012).

37. Scott, R.A. et al. Large-scale association analyses identify new loci influencing glycemic traits and provide insight into the underlying biological pathways. Nat Genet 44, 991-1005 (2012).

38. Peloso, G.M. et al. Association of low-frequency and rare coding-sequence variants with blood lipids and coronary heart disease in 56,000 whites and blacks. Am J Hum Genet 94, 223-32 (2014).

39. Flannick, J. et al. Loss-of-function mutations in SLC30A8 protect against type 2 diabetes. Nat Genet 46, 357-63 (2014).

40. Holmen, O.L. et al. Systematic evaluation of coding variation identifies a candidate causal variant in TM6SF2 influencing total cholesterol and myocardial infarction risk. Nat Genet 46, 345-51 (2014).

41. Chatterjee, N. et al. Projecting the performance of risk prediction based on polygenic analyses of genome-wide association studies. Nat Genet 45, 400-5, 405e1-3 (2013).

42. Defesche, J.C. Defining the challenges of FH screening for familial hypercholesterolemia. J Clin Lipidol 4, 338-41 (2010).
43. Versmissen, J. et al. Efficacy of statins in familial hypercholesterolaemia: a long term cohort study. BMJ 337, a2423 (2008).
44. Morris, A.P. et al. Large-scale association analysis provides insights into the genetic architecture and pathophysiology of type 2 diabetes. Nat Genet 44, 981-90 (2012).
45. Hunt, K.A. et al. Negligible impact of rare autoimmune-locus coding-region variants on missing heritability. Nature 498, 232-5 (2013).
46. Manolio, T.A. et al. Finding the missing heritability of complex diseases. Nature 461, 747-53 (2009).
47. Thomas, D. Gene--environment-wide association studies: emerging approaches. Nat Rev Genet 11, 259-72 (2010).
48. Langenberg, C. et al. Gene-lifestyle interaction and type 2 diabetes: the EPIC interact case-cohort study. PLoS Med 11, e1001647 (2014).

Chapter 6

Summary/Samenvatting

Summary

Cardiovascular diseases (CVDs) are the leading cause of morbidity and the number one cause of death worldwide. Despite the successes of primary and secondary prevention programs, CVDs impose a major burden on human health and healthcare systems. An increasing portion of CVD cases can be prevented by addressing the modifiable risk factors, including type 2 diabetes (T2D) and dyslipidemia. The high impact and frequency make T2D and dyslipidemia suitable candidates for targeting preventive interventions, such as medication, weight loss, and increased physical activity, which can prevent, slow or even reverse the development of these risk factors and thus reduce the burden of CVDs.

Both T2D and dyslipidemia are caused by a complex interplay between genetic and nongenetic factors, with heritability estimates ranging from 26 to 69% for T2D and from 24 to 56% for total cholesterol (TC), low-density lipoprotein cholesterol (LDL-C), high-density lipoprotein cholesterol (HDL-C) and triglyceride (TG) levels. Over the past few years, knowledge of genetic variants driving this heritability has rapidly increased through collaborations in large genetic consortia for gene discovery. These collaborative efforts have revealed dozens of single nucleotide polymorphisms (SNPs) associated with lipid levels and with T2D and the related quantitative traits fasting glucose (FG) and fasting insulin (FI) in the general population. Together, the currently know SNPs explain about 10% of the heritability of type 2 diabetes and 25 to 30% of the heritability of the blood lipid levels.

The identification of large numbers of SNPs has improved understanding of disease etiology, which can lead to identification of novel targets for therapeutic interventions, and has raised the question whether the genetic information can be used to identify individuals at high risk of T2D or dyslipidemia. In this thesis I aimed to further improve our understanding of the etiology of T2D and dyslipidemia and to investigate the extent to which genetic risk factors combined can improve their prediction. The thesis describes a search for genetic variants associated with T2D and the related quantitative traits fasting glucose and insulin in the general population and the performance of genetic risk scores in the identification of individuals at increased risk for T2D and dyslipidemia.

In chapter 2.1, I present an overview of published genetic risk prediction studies for T2D from a methodological perspective, focusing on the variables included in the risk models as well as the study designs and populations investigated. We argued and showed that differences in study design and characteristics of the study population have an impact on the observed predictive ability of risk models. This observation emphasizes that genetic risk prediction studies should be conducted in those populations in which the prediction models will ultimately be applied, if proven useful.

Chapter 2 further focuses on identifying potential causal rare and low frequency variants associated with diabetes risk and the related quantitative traits FG and FI. In chapter 2.2, we sought to obtain more insight in genome-wide association study (GWAS) findings near the *IGF1* gene and to identify and characterize novel genetic variants at the locus. Mediation analyses in 5,141 non-diabetic individuals of the Cardiovascular Health Study (CHS), Framingham Heart Study (FHS) and Rotterdam Study suggested that previously observed associations between SNPs near *IGF1* with FI levels were not mediated by circulating IGF-I levels. Further investigation of the *IGF1* gene, using deep sequencing data in 3,539 non-diabetic individuals from the Atherosclerosis Risk in Communities (ARIC) study, CHS and FHS that were part of the Cohorts for Heart and Aging Research in Genomic Epidemiology (CHARGE) targeted sequencing study, revealed a large number of rare variants at the locus that had not been previously described, the large majority of which was very rare. A subset of rare non-synonymous variants, including 6 novel variants and 5 variants that had been previously identified, was significantly associated with FI levels. Conditional analysis suggested that the common non-coding variants near *IGF1* that were identified in GWAS explain part of the rare variant signal and the presence of a residual independent rare variant effect. Examination of regulatory element catalogs constructed through genome wide experimental efforts of the ENCODE Consortium showed that the GWAS variants were located in the proximity of FOXA1 binding sites and DNAseI hypersensitive sites, suggesting that they might have a direct functional role. This finding is noteworthy because FOXA1 is a key transcriptional regulator implicated in glucose metabolism and insulin secretion.

In the CHARGE consortium we further sought for rare and low-frequency variants associated with fasting glucose, fasting insulin and T2D using the HumanExome BeadChip (chapter 2.3). The FG meta-analysis in up to 60,564 individuals and follow up in up to 14,884 T2D cases and 78,097 controls identified a novel low-frequency (minor allele frequency 1.4%) non-synonymous SNP in the gene encoding the receptor for glucagon-like peptide 1 (*GLP1R*) associated with FG and T2D. The minor allele was associated with lower FG, lower T2D risk, lower insulin response during a glucose challenge and higher 2-h glucose, pointing to physiological effects on the incretin system. This association represents another instance wherein genetic epidemiology has identified a gene that codes for a direct drug target in T2D. Further, at known FG, FI and T2D loci we found associations with non-synonymous SNPs in 6 FG and/or FI loci (G6PC2, *COBLL1*, *GPSM1*, *SLC2A2*, *SLC30A8* and *RREB1*) and in 5 T2D loci (*ARAP1*, *GIPR*, *KCNJ11*, *SLC30A8* and *WFS1*) as potential causal candidates underlying previously identified associations. We also revealed non-coding variants whose putative functions in epigenetic and post-transcriptional regulation of *ABO* and *G6PC2* are supported by experimental ENCODE

Consortium, GTEx and transcriptome data from islets and for which future focused investigations using human cell culture and animal models will be needed to clarify their functional influence on glycemic regulation.

Chapter 3 focuses on the ability of risk scores comprised of known common genetic lipid variants to identify individuals at high risk of dyslipidemia and association of the risk scores with subclinical atherosclerosis and incident coronary heart disease. In chapter 3.1, we computed risk scores from known common genetic variants for lipid levels per individual in 2,715 Erasmus Rucphen Family (ERF) and 10,072 Rotterdam Study participants and showed that these variants play an important role in determining who develops dyslipidemia after the age of 55 years up until age 90 years in normal weight and overweight individuals. Despite the small effects of the common genetic lipid variants on lipid levels individually, when joint into a risk score the effects were substantial and discriminated future patients with dyslipidemia better than an established epidemiological risk factor such as body mass index (BMI). This underscores the value of the GWAS for age-related diseases.

In chapter 3.2 we showed in ERF (n = 2,269) and the Rotterdam Study (n = 8,130) that risk scores comprised of common genetic lipid variants were associated with the more distant, but clinically very relevant outcomes subclinical atherosclerosis and incident coronary heart disease (CHD). All genetic risk scores that we analyzed (for TC, LDL-C, HDL-C, and TG) were associated with carotid plaque and the TC and LDL-C scores were robustly associated with incident CHD. This supports the hypothesis that besides common and rare variants found to be associated with cardiovascular disease individually, aggregations of common genetic variants influencing lipid levels play a significant role in the development of atherosclerosis and the subsequent occurrence of cardiovascular disease. The risk scores did not improve clinical areas under the curve (AUCs) for prediction of CHD. However, associations of the TC and LDL-C risk scores with plaque and CHD were still (borderline) significant after inclusion of the lipid levels corresponding to the genetic risk score in the models, suggesting added value of the risk scores beyond the lipid levels themselves.

In chapter 4 we tested the hypothesis that in individuals with T2D, besides metabolic alterations, common genetic lipid variants identified in the general population also influence their lipid levels. Meta-analysis results from the Rotterdam Study (n = 9,791) and ERF (n = 2,313) showed that in patients with diabetes LDL-C, HDL-C and TG genetic risk scores were significantly associated with their corresponding lipid levels and the TG score was associated with prevalent dyslipidemia. Effect estimates were not significantly different between diabetes cases and non-diabetic controls. However, the ability

of the genetic risk scores to predict incident dyslipidemia in the Rotterdam Study was significantly higher in cases than in controls, suggesting that the role of lipid-altering genetic variants might be modified in the context of diabetes. Results of analyses to test whether the differences might be explained by genetic risk score*diabetes interactions were inconclusive. Our findings suggest that it might be of interest to further investigate these interactions in other cohorts. These findings include significant associations of interaction terms with lipid levels and dyslipidemia in the Rotterdam Study and, in single SNP*diabetes interaction analyses, significant enrichment for low p-values.

In conclusion, the main findings of this thesis are the identification of a large number of novel rare variants in the *IGF1* gene, which has improved insight in the genetic variation present at the locus, the finding that both rare non-synonymous variants and common non-coding variants with putative function in and near the *IGF1* gene may have a role in insulin biology (chapter 2.2), and the identification of a low-frequency variant associated with FG and T2D in *GLP1R*, a gene encoding a direct drug target for T2D (chapter 2.3). Further, while the currently identified common type 2 diabetes associated genetic variants do not yet improve prediction of type 2 diabetes compared to clinical prediction models, chapter 3.1 provides evidence that genetic risk scores for lipid levels play a large role in determining who develops dyslipidemia through old age. Finally, in Chapter 4, I show that the role of common genetic variants associated with lipid levels may be modified in the context of diabetes. Several efforts are underway which may further increase our understanding of the etiology of T2D and dyslipidemia and the identification of individuals at high risk of developing disease. These include common and rare variant analyses using larger sample sizes, improved insight in methods for rare variant analyses which may lead to better identification of genes and rare variants associated with disease, and gene*gene and gene*environment interaction studies. These efforts may bring us closer to translation of complex disease genetics in clinical practice by identifying potential targets for therapeutic interventions and improving genetic risk prediction. Ultimately this may reduce the prevalence of the CVD risk factors T2D and dyslipidemia and thus the burden of morbidity.

Samenvatting

GSTCG
GACTGACTGC
CAACGACTGATC
GACTGTACGCGCTA
GATATAGACAGACGA
ACTGACTGACTGACTG
ACTCAGACTCGATGACT
GACATATAGCTACGTAC
TGACGTACGTACGTACG
GTACTACTACGGTACT/
AAACGGGTGTGTGTC
ACTGACTGACTGACT(
ACTCAGACTCGATGA
GACATATAGCTACG
TGACTGACTGAC
GTACTACTACG(
AAACGGGTG
GCATGATGA
TACGTAC
GSTCG
GACT
CA

Cardiovasculaire aandoeningen vormen wereldwijd de belangrijkste oorzaak van morbiditeit en mortaliteit. Ondanks de successen van primaire en secundaire preventieprogramma's, leggen deze aandoeningen een groot beslag op de volksgezondheid en gezondheidszorg. Het aantal nieuwe patienten kan in toenemende mate worden teruggedrongen door verbeteren van preventie en behandeling van risicofactoren, waaronder type 2 diabetes (T2D) en dyslipidemie. De grote impact en hoge frequentie van T2D en dyslipidemie maken het geschikte kandidaten voor gerichte preventieve interventies zoals medicatie, afvallen en verhogen van de lichaamsbeweging. Deze interventies kunnen T2D en dyslipidemie voorkomen of zelfs genezen en daarmee de prevalentie en incidentie van cardiovasculaire aandoeningen terugdringen.

Zowel T2D als dyslipidemie worden veroorzaakt door een complex samenspel van genetische en niet-genetische factoren. Schattingen van de erfelijkheid variëren van 26 tot 69% voor T2D en van 24 tot 56% voor totaal cholesterol (TC), low-density lipoprotein cholesterol (LDL-C), high-density lipoprotein cholesterol (HDL-C) en triglyceride (TG) niveaus. De afgelopen jaren is de kennis van de onderliggende genetische varianten snel toegenomen door samenwerkingen in grote consortia gericht op het ontdekken van genen. Deze samenwerkingen hebben tientallen single nucleotide polymorphisms (SNPs) ontdekt die geassocieerd zijn met lipide niveaus en met T2D en de gerelateerde kwantitatieve uitkomsten nuchter glucose en nuchter insuline. De tot nu toe bekende SNPs verklaren samen ongeveer 10% van de erfelijkheid van T2D en 25 tot 30% van de erfelijkheid van lipide niveaus in het bloed.

De identificatie van grote aantallen SNPs heeft ons begrip van de etiologie van T2D en dyslipidemie vergroot, wat kan leiden tot ontdekking van nieuwe aangrijpingspunten voor therapeutische interventies. Daarnaast heeft het de vraag opgeworpen of de genetische informatie gebruikt kan worden voor het identificeren van individuen met hoog ziekterisico. In dit proefschrift was mijn doel ons begrip van de etiologie van T2D en dyslipidemie verder te vergroten en te onderzoeken in hoeverre genetische risicofactoren samen predictie kunnen verbeteren. Het proefschrift beschrijft een zoektocht naar genetische varianten geassocieerd met T2D, nuchter glucose en nuchter insuline in de algemene bevolking en het vermogen van genetische risicoscores om individuen met verhoogd risico op T2D en dyslipidemie te identificeren.

In hoofdstuk 2.1 laat ik een overzicht zien van gepubliceerde genetisch risico predictie onderzoeken voor T2D vanuit methodologisch perspectief, gericht op de variabelen die geïncludeerd zijn in de predictiemodellen, het ontwerp van de studies en de onderzochte populaties. We hebben beargumenteerd en laten zien dat verschillen in studieontwerp en karakteristieken van de studiepopulatie het geobserveerde voorspellende

vermogen van risicomodellen beïnvloeden. Deze observatie benadrukt dat genetisch risico predictie studies uitgevoerd zouden moeten worden in de populaties waarin, als bewezen is dat genetisch risico predictie zinvol is, we de modellen uiteindelijk willen toepassen.

Hoofdstuk 2 is verder gericht op identificatie van potentieel causale zeldzame en laag frequente varianten geassocieerd met risico op diabetes en met de gerelateerde kwantitatieve uitkomsten nuchter glucose en nuchter insuline. In hoofdstuk 2.2 hebben we getracht meer inzicht te krijgen in bevindingen van genoomwijde associatiestudies (GWAS) in het *IGF1* locus en nieuwe genetische varianten in het locus te identificeren en karakteriseren. Mediatie-analyses in 5,141 niet-diabetische individuen van de Cardiovascular Health Study (CHS), Framingham Heart Study (FHS) en Rotterdam Study suggereerden dat voorheen geobserveerde associaties tussen SNPs dichtbij *IGF1* en nuchter insuline niet waren gemedieerd door IGF-I niveaus. Verder onderzoek van het *IGF1* gen, gebruikmakend van gedetailleerde sequentie data in 3,539 niet-diabetische individuen van de Atherosclerosis Risk in Communities (ARIC) studie, CHS en FHS die onderdeel waren van de Cohorts for Heart and Aging Research in Genomic Epidemiology (CHARGE) targeted sequencing study, heeft grote aantallen varianten geïdentificeerd die niet eerder waren beschreven. De grote meerderheid van deze varianten was zeer zeldzaam. Een set van 6 niet eerder beschreven en 5 al bekende zeldzame niet-synonieme varianten was significant geassocieerd met nuchter insuline niveaus. Conditionele analyses suggereerden dat de veel voorkomende niet-coderende varianten dichtbij het *IGF1* gen, geïdentificeerd in GWAS, een deel van de associatie verklaren. Daarnaast suggereerden deze analyses aanwezigheid van een residu onafhankelijk effect van de zeldzame varianten. Onderzoek van regulatoire regios in het ENCODE consortium, liet zien dat de GWAS varianten waren gelegen in de buurt van FOXA1 bindingsplaatsen en DNAseI hypersensitieve plaatsen. Dit suggereert een mogelijke direct functionele rol voor de GWAS varianten. Deze bevinding is noemenswaardig omdat FOXA1 een belangrijke transcriptieregulator is, geïmpliceerd in glucose metabolisme en insuline secretie.

In het CHARGE consortium hebben we verder gezocht naar zeldzame en laag frequente varianten geassocieerd met nuchter glucose, nuchter insuline en T2D gebruikmakend van de HumanExome BeadChip (hoofdstuk 2.3). De nuchter glucose meta-analyse in 60,564 individuen en follow-up in 14,884 individuen met T2D en 78,097 controles identificeerde een nieuwe laag frequente (frequentie zeldzame allel: 1.4%) niet-synonieme SNP geassocieerd met nuchter glucose en T2D in het gen dat codeert voor de glucagon-like peptide 1 receptor (*GLP1R*). Het laag frequente allel was geassocieerd met lager nuchter glucose, lager T2D risico, verminderde insulinereactie tijdens een glucose challenge en hoger 2-uurs glucose, wijzend op fysiologische effecten op het incretinesysteem.

Deze associatie vormt een nieuw voorbeeld waarin genetische epidemiologie een gen heeft geïdentificeerd dat codeert voor een direct aangrijpingspunt voor T2D medicatie. Verder hebben we associaties gevonden met niet-synonieme SNPs in 6 bekende FG en/of FI loci (*G6PC2*, *COBLL1*, *GPSM1*, *SLC2A2*, *SLC30A8* en *RREB1*) en in 5 bekende T2D loci (*ARAP1*, *GIPR*, *KCNJ11*, *SLC30A8* en *WFS1*) als potentieel causale kandidaten die de voorheen geïdentificeerde associaties zouden kunnen verklaren. We hebben ook associaties gevonden met niet-coderende varianten, waarvan vermeende functie in epigenetische en post-transcriptie regulatie van *ABO* en *G6PC2* werd ondersteund door experimentele ENCODE consortium, GTEx en transcriptoom data. Onderzoek van humane celculturen en diermodellen zullen nodig zijn om de vermeende functionele rol in glycemische regulatie verder te verhelderen.

Hoofdstuk 3 gaat in op het vermogen van risicoscores, bestaande uit bekende veel voorkomende varianten die lipide niveaus beïnvloeden, om individuen met hoog risico op dyslipidemie te identificeren. Dit hoofdstuk gaat verder in op de associatie van deze risicoscores met subklinische atherosclerose en incidente cardiovasculaire ziekte. In hoofdstuk 3.1 hebben we de risicoscores berekend per individu in 2,715 Erasmus Rucphen Familieonderzoek (ERF) en 10,072 Rotterdam Studie deelnemers. We hebben laten zien dat de genetische varianten samen een belangrijke rol spelen in het bepalen welke individuen dyslipidemie ontwikkelen in individuen met normaal gewicht en overgewicht in de leeftijd tussen 55 en 90 jaar. Ondanks de kleine effecten van de veel voorkomende genetische varianten op lipide niveaus individueel, zijn de effecten substantieel als ze worden gecombineerd in een risicoscore en kunnen ze toekomstige patiënten met dyslipidemie beter identificeren dan een gevestigde epidemiologische risicofactor zoals body mass index (BMI). Dit onderstreept de waarde van GWAS voor leeftijdsgerelateerde aandoeningen.

In hoofdstuk 3.2 hebben we laten zien in ERF (n = 2,269) en de Rotterdam Studie (n = 8,130) dat de risicoscores samengesteld uit veelvoorkomende varianten die lipiden niveaus beïnvloeden, waren geassocieerd met de verder weg gelegen, maar klinisch zeer relevante uitkomsten subklinische atherosclerose en incidente coronaire hartziekte. Alle genetische risicoscores die we hebben geanalyseerd (voor TC, LDL-C, HDL-C en TG) waren geassocieerd met carotis plaque en de TC en LDL-C scores waren daarnaast geassocieerd met coronaire hartziekte. Dit ondersteunt de hypothese dat naast veel voorkomende en zeldzame varianten waarvoor individuele associaties met cardiovasculaire aandoeningen zijn gevonden, ook aggregaties van veel voorkomende varianten die lipide niveaus beïnvloeden een significante rol spelen in de ontwikkeling van atherosclerose en vervolgens het ontstaan van coronaire hartziekte. De genetische risicoscores verbeterden niet de klinische areas under the curve (AUCs) voor predictie van coronaire hartziekte. Echter,

associaties van de TC- en LDL-C risicoscores met plaque en coronaire hartziekte waren nog steeds (borderline) significant na inclusie van de lipide niveaus corresponderend met de genetische risicoscore in de modellen. Dit suggereert toegevoegde waarde van de risicoscores boven de lipide niveaus zelf.

In hoofdstuk 4 hebben we de hypothese getest dat in individuen met T2D, naast metabole veranderingen, de hierboven beschreven genetische risicoscores samengesteld uit varianten geïdentificeerd in de algemene bevolking ook de lipide niveaus beïnvloeden. Meta-analyse resultaten van de Rotterdam Studie (n = 9,791) en ERF (n = 2,313) hebben laten zien dat in patiënten met diabetes LDL-C, HDL-C en TG genetische risicoscores significant geassocieerd waren met de corresponderende lipide niveaus en de TG score was geassocieerd met prevalente dyslipidemie. Effectmaten waren niet significant verschillend tussen diabetici en niet-diabetici. Het vermogen echter van de genetische risicoscores om incidente dyslipidemie te voorspellen in de Rotterdam Studie was significant beter in individuen met dan in controles vrij van diabetes. Dit suggereert dat de rol van lipide-beïnvloedende varianten anders kan zijn in de context van diabetes. Resultaten van analyses om te onderzoeken of de verschillen verklaard kunnen worden door interacties tussen de genetische risicoscores en diabetes waren niet conclusief. Een aantal van onze bevindingen suggereert dat het interessant zou zijn de interacties verder te onderzoeken in andere cohorten. Deze bevindingen omvatten significante associaties van risicoscore*diabetes interactie met lipide niveaus en dyslipidemie in de Rotterdam Studie en, in analyse van interacties tussen individuele SNPs en diabetes, significante verrijking voor lage p-waarden.

Concluderend zijn de belangrijkste bevindingen van dit proefschrift: 1) de identificatie van een groot aantal niet eerder beschreven zeldzame varianten in het *IGF1* gen, wat inzicht geeft in het locus, 2) de bevinding dat zowel zeldzame coderende varianten als frequente functionele niet-coderende varianten in en rond het *IGF1* gen mogelijk insuline niveaus beïnvloeden (hoofdstuk 2.2) en 3) de identificatie van een laag frequente variant geassocieerd met nuchter glucose en T2D in *GLP1R*, een gen dat codeert voor een direct aangrijpingspunt van T2D medicatie (hoofdstuk 2.3). Verder laat hoofdstuk 3.1 zien dat genetische risicoscores voor lipide niveaus tot op hoge leeftijd een grote rol spelen in het bepalen welke individuen dyslipidemie ontwikkelen. Ten slotte laat ik in hoofdstuk 4.1 zien dat de rol van veel voorkomende varianten geassocieerd met het vetgehalte in het bloed mogelijk anders is in de context van diabetes.

In de nabije toekomst richt het onderzoek naar T2D en dyslipidemie zich verder op de vraag hoe individuen met een hoog ziekterisico vroeg op te sporen en op identificatie van potentiele aangrijpingspunten voor therapeutische interventie. Onderzoek in grote aantallen personen is nodig om zeldzame genetische varianten met potentieel hoog

risico op ziekte te vinden, de frequente varianten met een laag risico verder in kaart te brengen en gen*gen en gen*omgeving interacties te ontdekken. Deze onderzoeken kunnen ons dichter bij de vertaling van de genetica van complexe aandoeningen naar de klinische praktijk brengen en daarmee naar het verder terugdringen van de ingrijpende gevolgen van cardiovasculaire aandoeningen.

Chapter 7

Supplemental material

```
                    TGACTACGT
            CTACTAGCACTGTACACG         CTACGTAC
        CGTACGTAGTCTGATCAAAACGTA     GCTATACAGCTA
      CGTTGGTACCAGTACGTACGACTGCGA   GTACGTACGTACG
     GTGTGTCACACAGACACGTGATATAGACCAGATGACACACGTAC
    GTACGACGGCGCGGCGCGATATATATAAAAGCACGGACACTACGA
   ACGATGACTGACTACGTGCGCGCGTAGCTAGCTACGGATAGGACTA
   ACTGACTACGTACTACTAAAGCTGCTTTTACGTACGTACGGGCATG
   CATGTACTATCGCGGCGCGTAGACTGTACGACCAGACTGACTGACT
   GACTGACTGACGCGCGTACGTAGCTACTGTAGTACGTACGTACGTA
   TTCAGACTGACTGACTGAGATTACAGCTACGTAGTACTGACACGT
    TACGACGGCGCGGCGCGATATATATAAAAGCACGGACACTACGA
    GATGACTGACTACGTGCGCGCGTAGCTAGCTACGGATAGGACTA
     TGACTACGTACTACTAAAGCTGCTTTTACGTACGTACGGGCATG
      CTGCGCATGTACTATCGCGGCGCGTAGACTGTACGACCAGAC
      ACTGACGCGCGTACGTAGCTACTGTAGTACGTACGTACGTA
       ACTGACTGACTGAGATTACAGCTACGTAGTACTGACACGT
        GCATACGAGTCGAAGCGCGCAATATATATTATATCGGC
         GTGACTACATCATGCTGACTACGGTCGCGCTCAACG
          TGACTACGTACGTATACGACGTACTGACGGCGCG
          GCACTGTACACGACTGACTTACTAGCTACGTACG
           TGATCAAAACGTACGCGCGGCTATACAGCTA
            CGTACGACTGCGATACGTACGTACGTACG
             CGTGATATAGACCAGATGACACACG
              TATATATAAAAGCACGGACAC
               GTAGCTAGCTACGGATA
                CTTTTACGTACGT
                 ACGACCAG
                  TAGT
```

Authors and affiliations

Chapter 2.1 A methodological perspective on genetic risk prediction studies in type 2 diabetes: recommendations for future research

Sara M. Willems[1], Raluca Mihaescu[2], Eric J.G. Sijbrands[3], Cornelia M. van Duijn[1], A. Cecile J.W. Janssens[2]

1. Genetic Epidemiology Unit, Department of Epidemiology, Erasmus University Medical Center, Rotterdam, the Netherlands
2. Department of Epidemiology, Erasmus University Medical Center, Rotterdam, the Netherlands
3. Department of Internal Medicine, Erasmus University Medical Center, Rotterdam, the Netherlands

Chapter 2.2 Association of the *IGF1* gene with fasting insulin levels

Sara M. Willems[1], Belinda K. Cornes[2,3], Jennifer A. Brody[4], Alanna C. Morrison[5], Leonard Lipovich[6,7], Marco Dauriz[2,3,8], Bianca Porneala[2], Yuning Chen[9], Ching-Ti Liu[9], Denis V. Rybin[10], Richard A Gibbs[11], Donna Muzny[11], James S. Pankow[12], Bruce M. Psaty[13], Eric Boerwinkle[5,11], Jerome I. Rotter[14], David S. Siscovick[15], Ramachandran S. Vasan[16,17], Robert C. Kaplan[18], Aaron Isaacs[1], Josée Dupuis[9,17], Cornelia M. van Duijn[1], James B. Meigs[2,3]

1. Genetic Epidemiology Unit, Department of Epidemiology, Erasmus Medical Center, Rotterdam, the Netherlands
2. Division of General Internal Medicine, Massachusetts General Hospital, Boston, Massachusetts, USA
3. Department of Medicine, Harvard Medical School, Boston, Massachusetts, USA
4. Cardiovascular Health Research Unit, Department of Medicine, University of Washington, Seattle, Washington, USA
5. School of Public Health, University of Texas Health Science Center at Houston, Texas, USA
6. Center for Molecular Medicine and Genetics, Wayne State University, Detroit, Michigan, USA
7. Department of Neurology, Wayne State University School of Medicine, Detroit, Michigan, USA
8. Division of Endocrinology, Diabetes and Metabolism, Department of Medicine, University of Verona Medical School and Hospital Trust of Verona, Verona, Italy
9. Department of Biostatistics, Boston University School of Public Health, Boston, Massachusetts, USA
10. Boston University Data Coordinating Center, Boston, Massachusetts, USA

11. Human Genome Sequencing Center, Baylor College of Medicine, University of Texas Health Science Center, Houston, TX

12. Division of Epidemiology and Community Health (J.S.P.), University of Minnesota, MN, USA

13. Cardiovascular Health Research Unit, Departments of Medicine, Epidemiology, and Health Services, University of Washington, Seattle, WA; Group Health Research Institute, Group Health Cooperative, Seattle, WA

14. Institute for Translational Genomics and Population Sciences, Los Angeles Biomedical Reasearch Institute and Department of Pediatrics, Harbor-UCLA Medical Center Torrance, California, USA

15. New York Academy of Medicine, New York, New York, USA

16. Cardiology Section, Department of Preventive Medicine and Epidemiology, Boston University School of Medicine, Boston, Massachusetts, USA

17. National Heart, Lung, and Blood Institute's Framingham Heart Study, Framingham, Massachusetts, USA

18. Department of Epidemiology and Population Health, Albert Einstein College of Medicine, New York, USA

Chapter 2.3 Low-frequency and rare exome chip variants associate with fasting glucose and type 2 diabetes susceptibility

Jennifer Wessel*[1,2], Audrey Y. Chu*[3], Sara M. Willems*[4,5], Shuai Wang*[6], Hanieh Yaghootkar[7], Jennifer A. Brody[8,9], Marco Dauriz[10-12], Marie-France Hivert[13-15], Sridharan Raghavan[10,11], Leonard Lipovich[16,17], Bertha Hidalgo[18], Keolu Fox[9,19], Jennifer E. Huffman[20], Ping An[21], Yingchang Lu[22,23], Laura J. Rasmussen-Torvik[24], Niels Grarup[25], Margaret G. Ehm[26], Li Li[26], Abigail S. Baldridge[24], Alena Stančáková[27], Ravinder Abrol[28,29], Céline Besse[30], Anne Boland[30], Jette Bork-Jensen[25], Myriam Fornage[31], Daniel F. Freitag[32,33], Melissa E. Garcia[34], Xiuqing Guo[35], Kazuo Hara[22,23], Aaron Isaacs[4], Johanna Jakobsdottir[36], Leslie A. Lange[37], Jill C. Layton[38], Man Li[39], Jing Hua Zhao[5], Karina Meidtner[40], Alanna C. Morrison[41], Mike A. Nalls[42], Marjolein J. Peters[43,44], Maria Sabater-Lleal[45], Claudia Schurmann[22,23], Angela Silveira[45], Albert V. Smith[36,46], Lorraine Southam[32,47], Marcus H. Stoiber[48], Rona J. Strawbridge[45], Kent D. Taylor[35], Tibor V. Varga[49], Kristine H. Allin[25], Najaf Amin[4], Jennifer L. Aponte[50], Tin Aung[51,52], Caterina Barbieri[53], Nathan A. Bihlmeyer[54,55], Michael Boehnke[56], Cristina Bombieri[57], Donald W. Bowden[58], Sean M. Burns[59], Yuning Chen[6], Yii-Der I. Chen[35], Ching-Yu Cheng[51,52,60,61], Adolfo Correa[62], Jacek Czajkowski[21], Abbas Dehghan[63], Georg B. Ehret[64,65], Gudny Eiriksdottir[36], Stefan A. Escher[66], Aliki-Eleni Farmaki[67], Mattias Frånberg[45,68], Giovanni Gambaro[69], Franco Giulianini[3], William A. Goddard III[29], Anuj Goel[70], Omri Gottesman[22], Megan L. Grove[41], Stefan Gustafsson[71], Yang Hai[35], Göran Hallmans[72], Jiyoung Heo[73], Per Hoffmann[74-76], Mohammad K. Ikram[51,61,77], Richard

A. Jensen[8,9], Marit E. Jørgensen[78], Torben Jørgensen [79,80], Maria Karaleftheri[81], Chiea C. Khor[52,60,82], Andrea Kirkpatrick[29], Aldi T. Kraja[21], Johanna Kuusisto[83], Ethan M. Lange[84], I.T. Lee[85], Wen-Jane Lee[86], Aaron Leong[10,11], Jiemin Liao[51,52], Chunyu Liu[87,88], Yongmei Liu[89], Cecilia M. Lindgren[90], Allan Linneberg[79], Giovanni Malerba[57], Vasiliki Mamakou[91], Eirini Marouli[92], Nisa M. Maruthur[93,94], Angela Matchan[32], Roberta McKean[95], Olga McLeod[45], Ginger A. Metcalf[96], Karen L. Mohlke[37], Donna M. Muzny[96], Ioanna Ntalla[97], Nicholette D. Palmer[58,98], Dorota Pasko[7], Andreas Peter[99,100], Nigel W. Rayner[32,47,101], Frida Renström[66], Ken Rice[8,102], Cinzia F. Sala[53], Bengt Sennblad[45,103], Ioannis Serafetinidis[104], Jennifer A. Smith[105], Nicole Soranzo[32,106], Elizabeth K. Speliotes[107], Eli A. Stahl[108], Kathleen Stir- rups[32,109], Nikos Tentolouris[110], Anastasia Thanopoulou[111], Mina Torres[95], Michela Traglia[53], Emmanouil Tsafantakis[112], Sundas Javad[5], Lisa R. Yanek[113], Eleni Zengini[114,115], Diane M. Becker[113], Josh C. Bis[8,9], James B. Brown[48,116], L. Adrienne Cupples[6,88], Torben Hansen[25], Erik Ingelsson[71,90], Andrew J. Karter[117], Carlos Lorenzo[118], Rasika A. Mathias[113], Jill M. Norris[119], Gina M. Peloso[43,59,120], Wayne H.-H. Sheu[85,121,122], Daniela Toniolo[53], Dhananjay Vaidya[113], Rohit Varma[95], Lynne E. Wagenknecht[123], Heiner Boeing[124], Erwin P. Bottinger[22], George Dedoussis[125], Panos Deloukas[109,126,127], Ele Ferrannini[128], Oscar H. Franco[63], Paul W. Franks[66,129,130], Richard A. Gibbs[96], Vilmundur Gudnason[36,46], Anders Hamsten[45], Tamara B. Harris[34], Andrew T. Hattersley[131], Caroline Hayward[20], Albert Hofman[63], Jan-Håkan Jansson[120,129], Claudia Langenberg[5], Lenore J. Launer[34], Daniel Levy[87,132], Ben A. Oostra[4], Christopher J. O'Donnell[11,88,133], Stephen O'Rahilly[134], Sandosh Padmanabhan[135], James S. Pankow[136], Ozren Polasek[137], Michael A. Province[21], Stephen S. Rich[138], Paul M Rid- ker[3,139], Igor Rudan[140], Matthias B. Schulze[40,100], Blair Smith[141], André G. Uitterlinden[43,63], Mark Walker[142], Hugh Watkins[70], Tien Y. Wong[51,52,61], Eleftheria Zeggini[32], Generation Scotland [143], The EPIC-InterAct Consortium, Markku Laakso[83], Ingrid B. Borecki[21], Daniel I. Chasman[3], Oluf Pedersen[25], Bruce M. Psaty[8,9,15,144,145], E. Shyong Tai[60,146], Cornelia M. van Duijn[4,147], Nicholas J. Wareham[5], Dawn M. Waterworth[148], Eric Boerwinkle[41,96], WH Linda Kao[39,94,149], Jose C. Florez[11,59,150,151], Ruth J.F. Loos[22,23,152], James G. Wilson[153], Timothy M. Frayling[7], David S. Siscovick[154,155], Josée Dupuis[6,88], Jerome I. Rotter[35], James B. Meigs[10,11], Robert A. Scott[5], Mark O. Goodarzi[28,156].

1. Fairbanks School of Public Health, Department of Epidemiology, Indianapolis, IN, USA.

2. Indiana University School of Medicine, Department of Medicine, Indianapolis, IN, USA.

3. Division of Preventive Medicine, Brigham and Women's Hospital, Boston MA, USA.

4. Genetic Epidemiology Unit, Department of Epidemiology, Erasmus University Medical Center, Rotterdam, The Netherlands.

5. MRC Epidemiology Unit, University of Cambridge School of Clinical Medicine, Institute of Metabolic Science, Cambridge Biomedical Campus, Cambridge, UK.

6. Department of Biostatistics, Boston University School of Public Health, Boston, MA, USA.

7. Genetics of Complex Traits, University of Exeter Medical School, University of Exeter, Exeter, UK.

8. Cardiovascular Health Research Unit, University of Washington, Seattle, WA, USA.

9. Department of Medicine, University of Washington, Seattle, WA, USA.

10. Massachusetts General Hospital, General Medicine Division, Boston, MA, USA.

11. Department of Medicine, Harvard Medical School, Boston, MA, USA.

12. Division of Endocrinology, Diabetes and Metabolism, Department of Medicine, University of Verona Medical School and Hospital Trust of Verona, Verona, Italy.

13. Harvard Pilgrim Health Care Institute, Department of Population Medicine, Harvard Medical School, Boston, MA, USA.

14. Division of Endocrinology and Metabolism, Department of Medicine, Université de Sherbrooke, Sherbrooke, Québec, Canada.

15. Department of Epidemiology, University of Washington, Seattle, Washington, USA.

16. Center for Molecular Medicine and Genetics, Wayne State University, Detroit, MI, USA.

17. Department of Neurology, Wayne State University School of Medicine, Detroit, MI, USA.

18. Section on Statistical Genetics, Department of Biostatistics, University of Alabama at Birmingham, Birmingham, AL.

19. Department of Genome Sciences, University of Washington, Seattle, WA, USA.

20. MRC Human Genetics Unit, MRC IGMM, University of Edinburgh, Edinburgh, Scotland, UK.

21. Division of Statistical Genomics and Department of Genetics, Washington University School of Medicine, St. Louis, MO, USA.

22. The Charles Bronfman Institute for Personalized Medicine, The Icahn School of Medicine at Mount Sinai, New York, NY, USA.

23. The Genetics of Obesity and Related Metabolic Traits Program, The Icahn School of Medicine at Mount Sinai, New York, NY, USA.

24. Department of Preventive Medicine, Northwestern University Feinberg School of Medicine, Chicago, IL, USA.

25. The Novo Nordisk Foundation Center for Basic Metabolic Research, Faculty of Health and Medical Sciences, University of Copenhagen, Copenhagen, Denmark.

26. Statistical Genetics, PCPS, GlaxoSmithKline, RTP, NC, USA.

27. Institute of Clinical Medicine, Internal Medicine, University of Eastern Finland, Kuopio, Finland.

28. Department of Medicine and Department of Biomedical Sciences, Cedars-Sinai Medical Center, Los Angeles, CA, USA.

29. Materials and Process Simulation Center, California Institute of Technology, Pasadena, CA, USA.

30. CEA, Institut de Génomique, Centre National de Génotypage, 2 Rue Gaston Crémieux, 91057 EVRY Cedex, France.

31. Brown Foundation Institute of Molecular Medicine, University of Texas Health Science Center, Houston, TX, USA.

32. The Wellcome Trust Sanger Institute, Hinxton, UK.

33. Department of Public Health and Primary Care, Strangeways Research Laboratory, University of Cambridge, Cambridge, UK.

34. National Institute on Aging, Bethesda, MD, USA.

35. Institute for Translational Genomics and Population Sciences, Los Angeles Biomedical Research Institute at Harbor-UCLA Medical Center, Torrance, CA, USA.

36. Icelandic Heart Association, Hotlasmari 1, 201 Kopavogur, Iceland.

37. Department of Genetics, University of North Carolina, Chapel Hill, NC, USA.

38. Indiana University, Fairbanks School of Public Health, Indianapolis, IN, USA.

39. Department of Epidemiology, Johns Hopkins University, Baltimore, Maryland, USA.

40. Department of Molecular Epidemiology, German Institute of Human Nutrition Potsdam-Rehbrücke, Nuthetal, Germany.

41. Human Genetics Center, School of Public Health, The University of Texas Health Science Center at Houston, Houston, Texas, USA.

42. Laboratory of Neurogenetics, National Institute on Aging, Bethesda, MD, USA.

43. Department of Internal Medicine, Erasmus University Medical Center, Rotterdam, The Netherlands

44. The Netherlands Genomics Initiative-sponsored Netherlands Consortium for Healthy Aging (NGI-NCHA), Leiden / Rotterdam, the Netherlands.

45. Atherosclerosis Research Unit, Department of Medicine Solna, Karolinska Institutet, Stockholm, Sweden.

46. University of Iceland, 101 Reykjavik, Iceland.

47. Wellcome Trust Centre for Human Genetics, Oxford, UK.

48. Department of Genome Dynamics, Lawrence Berkeley National Laboratory, Berkeley, CA, USA.

49. Department of Clinical Sciences, Genetic and Molecular Epidemiology Unit, Lund University, Skåne University Hospital, Malmö, Sweden.

50. Genetics, PCPS, GlaxoSmithKline, RTP, NC, USA.

51. Singapore Eye Research Institute, Singapore National Eye Centre, Singapore.

52. Department of Ophthalmology, National University of Singapore and National University Health System, Singapore.

53. Division of Genetics and Cell Biology, San Raffaele Research institute, Milano, Italy.

54. Predoctoral Training Program in Human Genetics, McKusick-Nathans Institute of Genetic Medicine, Johns Hopkins University School of Medicine, MD, USA.

55. McKusick-Nathans Institute of Genetic Medicine, Johns Hopkins University School of Medicine, Baltimore, MD, USA.

56. Department of Biostatistics and Center for Statistical Genetics, University of Michigan, Ann Arbor, Michigan, USA.

57. Section of Biology and Genetics, Department of Life and Reproduction Sciences, University of Verona, Verona, Italy.

58. Department of Biochemistry, Wake Forest School of Medicine, Winston-Salem, NC, USA.

59. Diabetes Unit, Department of Medicine, Massachusetts General, Boston, MA, USA.

60. Saw Swee Hock School of Public Health, National University of Singapore and National University Health System, Singapore.

61. Office of Clinical Sciences, Duke-NUS Graduate Medical School, National University of Singapore, Singapore.

62. Department of Medicine, University of Mississippi Medical Center, Jackson, MS, USA.

63. Department of Epidemiology, Erasmus University Medical Center, Rotterdam, The Netherlands.

64. McKusick-Nathans Institute of Genetic Medicine, Johns Hopkins University, Baltimore, MD.

65. Division of Cardiology, Geneva University Hospital, Switzerland.

66. Department of Clinical Sciences, Genetic and Molecular Epidemiology Unit, Skåne University, Hospital, Malmö, Sweden.

67. Department of Nutrition and Dietetics, School of Health Science and Education, Harokopio University, Athens, Greece.

68. Department of Numerical Analysis and Computer Science, SciLifeLab, Stockholm University, Stockholm, Sweden.

69. Division of Nephrology, Department of Internal Medicine and Medical Specialties, Columbus-Gemelli University Hospital, Catholic University, Rome, Italy.

70. Department of Cardiovascular Medicine, The Wellcome Trust Centre for Human Genetics, University of Oxford, Oxford, UK.

71. Department of Medical Sciences, Molecular Epidemiology and Science for Life Laboratory, Uppsala University, Uppsala, Sweden.

72. Department of Biobank Research, Umeå University, Umeå, Sweden.

73. Department of Biomedical Technology, Sangmyung University, Chungnam, Korea.

74. Institute of Human Genetics, Department of Genomics, Life & Brain Center, University of Bonn, Bonn, Germany.

75. Human Genomics Research Group, Division of Medical Genetics, University Hospital Basel Department of Biomedicine, Basel, Switzerland.

76. Institute of Neuroscience and Medicine (INM-1) Genomic Imaging Research Center Juelich, Juelich, Germany.

77. Memory Aging & Cognition Centre (MACC), National University Health System, Singapore.

78. Steno Diabetes Center, Gentofte, Denmark.

79. Research Centre for Prevention and Health, Glostrup University Hospital, Glostrup, Denmark.

80. Faculty of Medicine, University of Aalborg, Aalborg, Denmark.

81. Echinos Medical Centre, Echinos, Greece.

82. Division of Human Genetics, Genome Institute of Singapore, Singapore.

83. Institute of Clinical Medicine, Internal Medicine, University of Eastern Finland and Kuopio University Hospital, Kuopio, Finland.

84. Department of Genetics and Department of Biostatistics, University of North Carolina, Chapel Hill, NC, USA.

85. Division of Endocrine and Metabolism, Department of Internal Medicine, Taichung Veterans General Hospital, Taichung 40705, Taiwan.

86. Department of Medical Research, Taichung Veterans General Hospital, Taichung 407, Taiwan.

87. Framingham Heart Study, Framingham, MA, USA.

88. National Heart, Lung, and Blood Institute (NHLBI) Framingham Heart Study, Framingham, MA, USA.

89. Department of Epidemiology & Prevention, Division of Public Health Sciences, Wake Forest University, Winston-Salem, NC, USA.

90. Wellcome Trust Centre for Human Genetics, University of Oxford, Oxford, UK.

91. National and Kapodistrian University of Athens, Dromokaiteio Psychiatric Hospital, Athens, Greece

92. University of Athens, Department of Dietetics and Nutritional Science, Harokopio University of Athens, Athens, Greece.

93. Division of General Internal Medicine, Johns Hopkins University School of Medicine, Baltimore, MD, USA.

94. Welch Center for Prevention, Epidemiology, and Clinical Research, Johns Hopkins University, Baltimore, MD, USA.

95. USC Eye Institute, Department of Ophthalmology, Keck School of Medicine of the University of Southern California, USA.

96. Human Genome Sequencing Center, Baylor College of Medicine, Houston, TX, USA.

97. Harokopio University of Athens, Department of Nutrition and Dietetics, Athens, Greece.

98. Center for Genomics and Personalized Medicine Research, Wake Forest School of Medicine, Winston-Salem, NC, USA.

99. Department of Internal Medicine, Division of Endocrinology, Metabolism, Patho-biochemistry and Clinical Chemistry and Institute of Diabetes Research and Metabolic Diseases, University of Tübingen, Tübingen, Germany.

100. German Center for Diabetes Research (DZD), Germany.

101. The Oxford Centre for Diabetes, Endocrinology and Metabolism, University of Oxford, Oxford, UK.

102. Department of Biostatistics, University of Washington, Seattle, WA, USA.

103. Science for Life Laboratory, Karolinska Institutet, Stockholm, Sweden.

104. Department of Gastroenterology, Gennimatas General Hospital, Athens, Greece.

105. Department of Epidemiology, School of Public Health, University of Michigan, Ann Arbor, MI, USA.

106. Department of Hematology, Long Road, Cambridge, UK.

107. Department of Internal Medicine, Division of Gastroenterology and Department of Computational Medicine and Bioinformatics, University of Michigan, Ann Arbor, MI.

108. Division of Psychiatric Genomics, The Icahn School of Medicine at Mount Sinai, New York, NY, USA.

109. William Harvey Research Institute, Barts and The London School of Medicine and Dentistry, Queen Mary University of London, UK.

110. First Department of Propaedeutic and Internal Medicine, Athens University Medical School, Laiko General Hospital, Athens, Greece.

111. Diabetes Centre, 2nd Department of Internal Medicine, National University of Athens, Hippokration General Hospital, Athens, Greece.

112. Anogia Medical Centre, Anogia, Greece.

113. The GeneSTAR Research Program, Division of General Internal Medicine, Department of Medicine, The Johns Hopkins University School of Medicine, Baltimore, MD.

114. University of Sheffield, Sheffield, UK.

115. Dromokaiteio Psychyatric Hospital, Athens, Greece.

116. Department of Statistics, University of California at Berkeley, Berkeley, CA, USA.

117. Division of Research, Kaiser Permanente, Northern California Region, Oakland, CA, USA.

118. Department of Medicine, University of Texas Health Science Center, San Antonio, TX, USA.

119. Department of Epidemiology, Colorado School of Public Health, University of Colorado Denver, Aurora, CO, USA.

120. Research Unit, Skellefteå, Sweden.

121. College of Medicine, National Defense Medical Center, Taipei, Taiwan.

122. School of Medicine, National Yang-Ming University, Taipei, Taiwan.

123. Division of Public Health Sciences, Wake Forest School of Medicine, Winston-Salem, NC, USA.

124. Department of Epidemiology, German Institute of Human Nutrition Potsdam Rehbrücke, Nuthetal, Germany.

125. Harokopio University Athens, Athens, Greece.

126. Wellcome Trust Sanger Institute, Hinxton, Cambridge, UK.

127. Princess Al-Jawhara Al-Brahim Centre of Excellence in Research of Hereditary Disorders (PACER-HD), King Abdulaziz University, Jeddah, Saudi Arabia.

128. Department of Clinical and Experimental Medicine, University of Pisa School of Medicine, Pisa, Italy.

129. Department of Public Health & Clinical Medicine, Umeå University, Umeå, Sweden.

130. Department of Nutrition, Harvard School of Public Health, Boston, USA.

131. Genetics of Diabetes, University of Exeter Medical School, University of Exeter, Exeter, UK.

132. Population Sciences Branch, National Heart, Lung, and Blood Institute, National Institutes of Health, Bethesda, MD, USA.

133. Cardiology Division, Department of Medicine, Massachusetts General Hospital and Harvard Medical School, Boston, MA, USA.

134. University of Cambridge Metabolic Research Laboratories, MRC Metabolic Diseases Unit and NIHR Cambridge Biomedical Research Centre, Wellcome Trust-MRC Institute of Metabolic Science, Addenbrooke's Hospital, Cambridge, UK.

135. Institute of Cardiovascular and Medical Sciences, University of Glasgow, Glasgow, UK.

136. Division of Epidemiology and Community Health, School of Public Health, University of Minnesota, Minneapolis, MN, USA.

137. Department of Public Health, Faculty of Medicine, University of Split, Split, Croatia.

138. Center for Public Health Genomics, Department of Public Health Sciences, University of Virginia, Charlottesville, VA, USA.

139. Division of Cardiology, Brigham and Women's Hospital and Harvard Medical School, Boston MA, USA.

140. Centre for Population Health Sciences, Medical School, University of Edinburgh, Edinburgh, Scotland, UK.

141. Medical Research Institute, University of Dundee, Dundee, UK.

142. Institute of Cellular Medicine, Newcastle University, Newcastle-upon-Tyne, UK.

143. Generation Scotland, A Collaboration between the University Medical Schools and NHS, Aberdeen, Dundee, Edinburgh and Glasgow, UK.

144. Department of Health Services, University of Washington, Seattle, WA, USA.

145. Group Health Research Institute, Group Health Cooperative, Seattle, WA, USA.

146. Department of Medicine, Yong Loo Lin School of Medicine, National University of Singapore, Singapore.

147. Center for Medical Systems Biology, Leiden, The Netherlands.

148. Genetics, PCPS, GlaxoSmithKline, Philadelphia, PA, USA.

149. Department of Medicine, Johns Hopkins University, Baltimore, MD, USA.

150. Center for Human Genetic Research, Massachusetts General Hospital, Boston, MA, USA.

151. Program in Medical and Population Genetics, Broad Institute, Cambridge, MA, USA.

152. The Mindich Child Health and Development Institute, The Icahn School of Medicine at Mount Sinai, New York, NY, USA.

153. Department of Physiology and Biophysics, University of Mississippi Medical Center, Jackson, MS, USA.

154. New York Academy of Medicine, New York, New York, USA.

155. Cardiovascular Health Research Unit, Departments of Medicine and Epidemiology, University of Washington, Seattle, WA, USA.

156. Division of Endocrinology, Diabetes and Metabolism, Cedars-Sinai Medical Center, Los Angeles, CA, USA.

Chapter 3.1 The role of common lipid-altering gene variants in the risk of dyslipidemia through old age

Sara M. Willems[1], Aaron Isaacs[1], Abbas Dehghan[2], Maksim V. Struchalin[1], Elisabeth M. van Leeuwen, Ben A. Oostra, Albert Hofman[2], André G. Uitterlinden[2,3], Oscar H. Franc[2], Eric J. G. Sijbrands[3], Cornelia M. van Duijn[1]

1. Genetic Epidemiology Unit, Department of Epidemiology, Erasmus Medical Center, Rotterdam, the Netherlands
2. Department of Epidemiology, Erasmus Medical Center, Rotterdam, the Netherlands
3. Department of Internal Medicine, Erasmus Medical Center, Rotterdam, the Netherlands

Chapter 3.2 Risk scores of common genetic variants for lipid levels influence atherosclerosis and incident coronary heart disease

Aaron Isaacs[*1,2], Sara M. Willems[*1], Daniel Bos[3,4], Abbas Dehghan[3], Albert Hofman[3], M. Arfan Ikram[3,4], André G. Uitterlinden[3,5], Ben A. Oostra[1], Oscar H. Franco[1], Jacqueline C. Witteman[3], Cornelia M. van Duijn[1,2]

1. Genetic Epidemiology Unit, Department of Epidemiology, Erasmus Medical Center, Rotterdam, the Netherlands
2. Centre for Medical Systems Biology, Leiden, the Netherlands
3. Department of Epidemiology, Erasmus Medical Center, Rotterdam, the Netherlands

4. Department of Radiology, Erasmus Medical Center, Rotterdam, the Netherlands
5. Department of Internal Medicine, Erasmus Medical Center, Rotterdam, the Netherlands

Chapter 4.1 Risk scores comprised of common lipid-altering genetic variants are associated with lipid levels and suggest an altered role of common genetic variation in type 2 diabetes

Sara M. Willems[1], Abbas Dehghan[2], Symen Ligthart[2], Jeannette M. Vergeer-Drop[1], Albert Hofman[2], Eric J. G. Sijbrands[3], André G. Uitterlinden[2,3], Ben A. Oostra[1], Jacqueline C. M. Witteman[2], Oscar H. Franco[2], Cornelia M. van Duijn[1], Aaron Isaacs[1]

1. Genetic Epidemiology Unit, Department of Epidemiology, Erasmus University Medical Center, Rotterdam, the Netherlands
2. Department of Epidemiology, Erasmus University Medical Center, Rotterdam, the Netherlands
3. Department of Internal Medicine, Erasmus University Medical Center, Rotterdam, the Netherlands

```
          .TGACTACGT.
      .CTACTAGCACTGTACACG.        .CTACGTACG
    .CGTACGTAGTCTGATCAAAACGTA.   .GCTATACAGCTA
   .CGTTGGTACCAGTACGTACGACTGCGA.  .GTACGTACGTACG
  TGTGTGTCACACAGACACGTGATATAGACCAGATGACACACGTAC
 .GTACGACGGCGCGGCGCGATATATATAAAAGCACGGACACTACGA
 ACGATGACTGACTACGTGCGCGCGTAGCTAGCTACGGATAGGACTA
 TACTGACTACGTACTACTAAAGCTGCTTTTACGTACGTACGGGCATC
 CATGTACTATCGCGGCGCGTAGACTGTACGACCAGACTGACTGACT
 GACTGACTGACGCGCGTACGTAGCTACTGTAGTACGTACGTACGTA
 TTCAGACTGACTGACTGAGATTACAGCTACGTAGTACTGACACGT
  TACGACGGCGCGGCGCGATATATATAAAAGCACGGACACTACGA
  GATGACTGACTACGTGCGCGCGTAGCTAGCTACGGATAGGACTA
   TGACTACGTACTACTAAAGCTGCTTTTACGTACGTACGGGCATC
   CTGCGCATGTACTATCGCGGCGCGTAGACTGTACGACCAGAC
    ACTGACGCGCGTACGTAGCTACTGTAGTACGTACGTACGTA
    ACTGACTGACTGAGATTACAGCTACGTAGTACTGACACGT
     GCATACGAGTCGAAGCGCGCAATATATATTATATCGGCC
      CGTGACTACATCATGCTGACTACGGTCGCGCTCAACG
       TGACTACGTACGTATACGACGTACTGACGGCGCGC
        GCACTGTACACGACTGACTTACTAGCTACGTACG
         TGATCAAAACGTACGCGCGGCTATACAGCTA
          CGTACGACTGCGATACGTACGTACGTACG
           CGTGATATAGACCAGATGACACACGT
            TATATATAAAAGCACGGACACT
              GTAGCTAGCTACGGATA
                CTTTTACGTACGT
                  ACGACCAG
                   TAGT
```

Supplemental material Chapter 2.2

Association of the *IGF1* gene with fasting insulin levels

Supplementary Table 1. Correlation between FI associated *IGF1* GWAS variants

	rs35767	rs855213	rs35747	rs860598	rs2114912
rs2114912	0.27	0.41	0.41	0.41	1.0
rs860598	0.78	1.0	1.0	1.0	
rs35747	0.78	1.0	1.0		
rs855213	0.78	1.0			
rs35767	1.0				

Values are r^2 between the SNVs in HapMap2 CEU

Supplementary table 2. Descriptions of known and novel SNPs in the *IGF1* region per cohort

	ARIC		CHS		FHS	
	known	novel*	known	novel*	known	novel*
no. SNPs	209	637	204	356	207	405
no. rare SNPs	99	637	98	355	97	403
		coding variants				
synonymous	1	2	2	0	2	2
nonsynonymous	1	5	3	0	3	1
		non-coding variants				
intergenic	146	432	137	252	146	294
upstream	6	13	7	7	7	7
downstream	5	16	4	4	2	9
intronic	31	88	31	42	27	43
UTR3	19	79	20	49	19	47
UTR5	0	2	0	2	1	2
predicted TFBS[#]	10	34	11	16	10	20
predicted DHS[†]	10	30	10	13	12	26
highly conserved[‡]	7	27	7	18	10	17
ORegAnno regulatory variant	3	16	3	6	4	11
total predicted functional**	25	88	25	41	28	60

Values are frequencies. ARIC: Atherosclerosis Risk in Communities Study, CHS: Cardiovascular Health Study, FHS: Framingham Heart Study *not known in dbSNP, 1000 genomes project or ESP 6500, [#] predicted transcription factor binding site (ENCODE ChipSeq, HMR), [†] DNAse hypersensitive site (ENCODE DHS) , [‡] highly conserved (PhastCons) **predicted transcription factor binding site (ENCODE ChipSeq, HMR) and/or DNAse hypersensitive site (ENCODE DHS) and/or ORegAnno regulatory variant and/or highly conserved (PhastCons)

Supplementary table 3. Contribution of all exonic non-synonymous variants at the *IGF1* locus to the SKAT result

Position (hg18)	β	S.E.	P	MAC	Weighted Score	% Test Contribution
101320456	1.220	0.612	0.046	1	0.557	4.249
101335739	0.605	0.612	0.323	1	0.137	1.046
101335754	0.417	0.411	0.310	1	0.145	1.105
101335775	−0.288	0.612	0.637	1	0.031	0.237
101335779	0.090	0.411	0.827	1	0.007	0.051
101335790	−0.297	0.612	0.628	1	0.033	0.252
101337467	0.550	0.154	3.6×10^{-4}	7	12.077	92.163
101337526	0.060	0.323	0.853	1	0.005	0.037
101393674	0.154	0.612	0.801	1	0.009	0.068
101396405	−0.172	0.323	0.596	1	0.039	0.301
101398245	−0.110	0.230	0.631	2	0.064	0.490

S.E.: standard error, MAC: minor allele count

Supplementary table 4. Results GWAS top hits for fasting insulin from common variant analyses of the *IGF1* targeted sequence data

	SNP rs35767			SNP rs2114912		
	Proxy rs2162679			Proxy rs2607988		
	β	S.E.	P	β	S.E.	P
ARIC	0.057	0.027	0.039	0.075	0.029	0.009
CHS	−0.005	0.028	0.873	−0.004	0.029	0.902
FHS	−0.019	0.021	0.364	−0.022	0.022	0.317
Meta	0.006	0.015	0.694	0.009	0.015	0.536

S.E.: standard error, ARIC: Atherosclerosis Risk in Communities Study, CHS: Cardiovascular Health Study, FHS: Framingham Heart Study
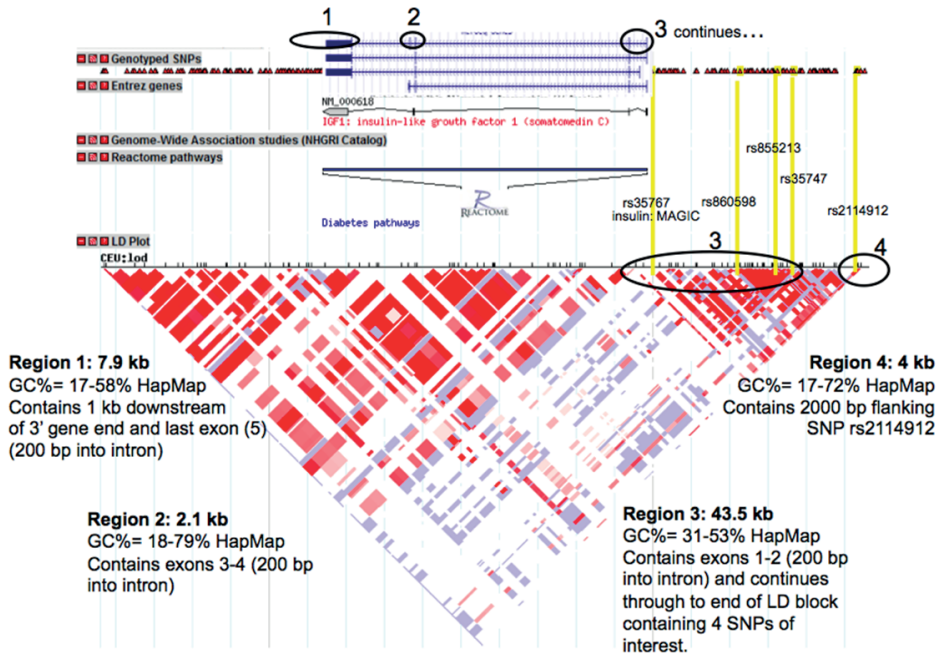
Supplementary Figure 1. *IGF1* targeted sequencing regions
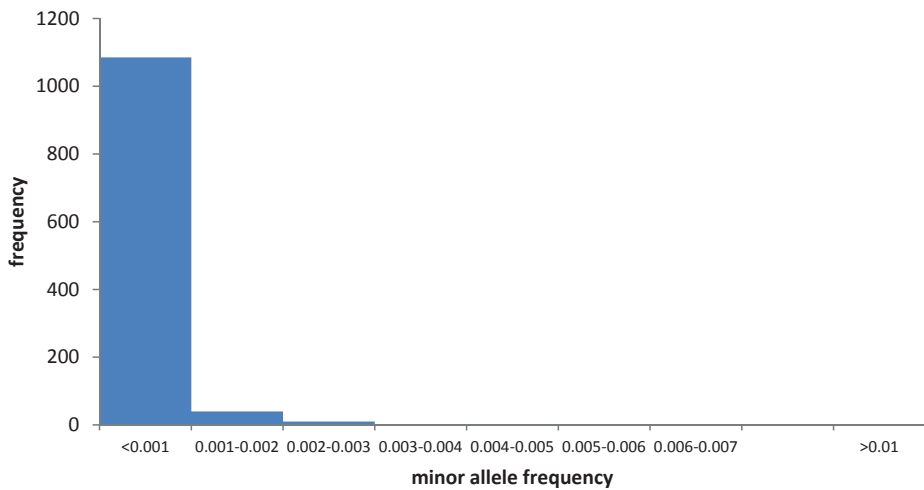
Supplementary Figure 2. MAF distribution of novel SNPs in the *IGF1* region
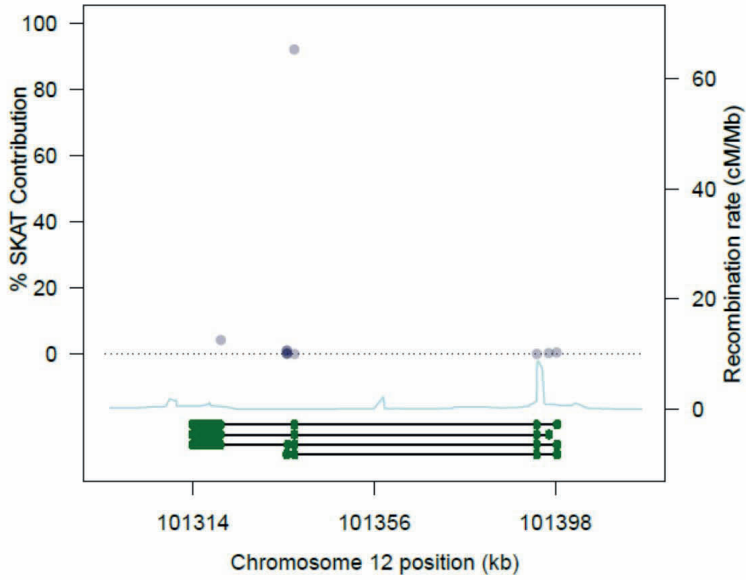
Supplementary Figure 3. Regional plot for association of rare coding non-synonymous *IGF1* variants with fasting insulin (BMI-adjusted)
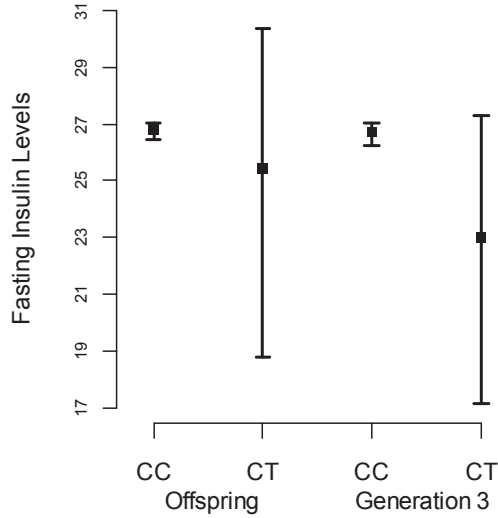
Supplementary Figure 4. Distribution of fasting insulin levels in carriers of the rare T allele at rs151098426
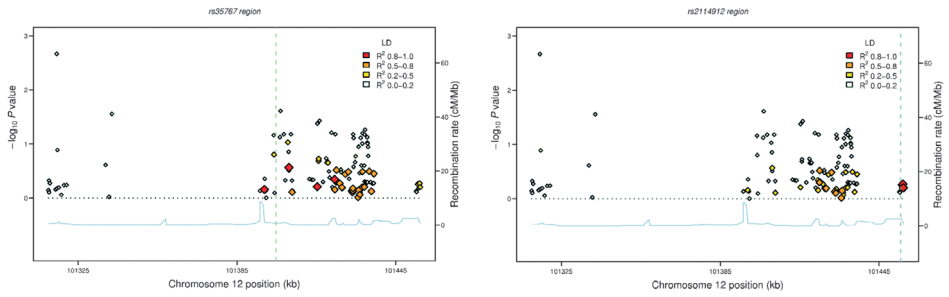
Supplementary Figure 5. Region plots for the common variant meta-analysis results
The green dotted line indicates the position of the GWAS SNP that corresponds to the plot title. LD indicates the LD, based on HapMap2 CEU, with that SNP

CTGACTACGTA
CTACTAGCACTGTACACGA   CTACGTAC
CGTACGTAGTCTGATCAAAACGTA   GCTATACAGCTA
CGTTGGTACCAGTACGTACGACTGCGA   GTACGTACGTACG
GTGTGTCACACAGACACGTGATATAGACCAGATGACACACGTAC
GTACGACGGCGCGGCGCGATATATATAAAAGCACGGACACTACGA
ACGATGACTGACTACGTGCGCGCGTAGCTAGCTACGGATAGGACTA
TACTGACTACGTACTACTAAAGCTGCTTTTACGTACGTACGGGCATC
CATGTACTATCGCGGCGCGTAGACTGTACGACCAGACTGACTGACT
GACTGACTGACGCGCGTACGTAGCTACTGTAGTACGTACGTACGTA
TTCAGACTGACTGACTGAGATTACAGCTACGTAGTACTGACACGT
TACGACGGCGCGGCGCGATATATATAAAAGCACGGACACTACGA
GATGACTGACTACGTGCGCGCGTAGCTAGCTACGGATAGGACTA
TGACTACGTACTACTAAAGCTGCTTTTACGTACGTACGGGCATC
CTGCGCATGTACTATCGCGGCGCGTAGACTGTACGACCAGAC
ACTGACGCGCGTACGTAGCTACTGTAGTACGTACGTACGTA
ACTGACTGACTGAGATTACAGCTACGTAGTACTGACACGT
GCATACGAGTCGAAGCGCGCAATATATATTATATCGGCC
CGTGACTACATCATGCTGACTACGGTCGCGCTCAACG
TGACTACGTACGTATACGACGTACTGACGGCGCGC
GCACTGTACACGACTGACTTACTAGCTACGTACC
TGATCAAAACGTACGCGCGGCTATACAGCTA
CGTACGACTGCGATACGTACGTACGTACG
CGTGATATAGACCAGATGACACACGT
TATATATAAAAGCACGGACACT
GTAGCTAGCTACGGATA
CTTTTACGTACGT
ACGACCAG
TAGT

Supplemental material Chapter 2.3

Low-frequency and rare exome chip variants associate with fasting glucose and type 2 diabetes susceptibility

Supplementary Table 1. Association of novel fasting glucose loci with fasting insulin

Gene	rsID	Chr	Build 37 position	Variation type	Alleles		EAF	Beta	SE	p	N
					Effect	Other					
GLP1R	rs10305492	6	39046794	nonsynonymous	A	G	0.015	0.005	0.011	0.67	47388
ABO	rs507666	9	136149399	intronic	A	G	0.177	0.007	0.004	0.05	47388
ABO	rs651007	9	136153875	intergenic	A	G	0.201	0.008	0.003	0.02	47148
ABO	rs579459	9	136154168	intergenic	C	T	0.201	0.008	0.003	0.02	47148
ABO	rs635634	9	136155000	intergenic	A	G	0.177	0.008	0.004	0.03	47148

Fasting insulin concentrations were log transformed and adjusted for sex, age, BMI, cohort effects and up to 10 principal components in up to 48,118 non-diabetic individuals. Effects are reported per copy of the minor allele. EAF: Effect allele frequency; N: sample size

Supplementary Table 2. Sample sizes from cohorts participating in the glycemic physiologic trait analyses

Cohorts	Sample sizes for traits derived from oral glucose tolerance test (OGTT)						
	2h-glucose	2h-insulin	30min insulin	insulino-genic index	AUC insulin	ratio of AUC ins/ AUC gluc	paired IVGTT and OGTT
Ely	1392	1377	1361	1345	1303	1217	NA
Fenland	6319	NA	NA	NA	NA	NA	NA
CoLaus	498	NA	NA	NA	NA	NA	NA
FHS	5716	2625	NA	NA	NA	NA	NA
ARIC	6707	NA	NA	NA	NA	NA	NA
GLACIER	916	NA	NA	NA	NA	NA	NA
Health2008	608	611	594	573	593	589	NA
Inter99	5419	5268	5210	4979	4872	4854	NA
METSIM	8230	8221	8189	8113	8182	8181	NA
RISC	1275	1260	1247	1193	1176	1174	738
total N	37080	19362	16601	16203	16126	16015	738

Glycemic physiologic traits were tested for association with *GLP1R* A316T rs10305492. See methods for estimation of glycemic trait measurements.

Supplementary Table 3. Association of *GLP1R* A316T with insulin sensitivity

SI from frequently sampled IV glucose tolerance test*				Standardized M/I from clamp	
IRAS study (n=184)**		IRAS Family study (n=1,024)***		RISC & ULSAM studies (n=2,170)	
Beta (SE)	p	Beta (SE)	p	Beta (SD)	p
0.51 (0.3)	0.13	−0.23 (0.2)	0.11	−0.04 (−0.25, 0.17)	0.71

Results are presented per minor allele. *Log transformed. **African-Americans. ***Hispanic-Americans

Supplementary Table 4. Genes included in the *GLP1R* pathway for MAGENTA gene-set enrichment analysis

				Gene				
ADCY1	*ADCY5*	*ADCY9*	*CALM2*	*GCG*	*GLP2R*	*PCLO*	*PRKX*	*VIP*
ADCY2	*ADCY6*	*ADCYAP1*	*CALM3*	*GIP*	*GNAS*	*PDX1*	*RAB3A*	*VIPR1*
ADCY3	*ADCY7*	*ADCYAP1R1*	*CREB1*	*GIPR*	*ITPR3*	*PRKACA*	*RAPGEF3*	*VIPR2*
ADCY4	*ADCY8*	*CALM1*	*DPP4*	*GLP1R*	*MZB1*	*PRKACB*	*RIMS2*	*WFS1*

Set of 36 genes were defined as having putative biological functions in a pathway between GLP1R activation and insulin secretion

Supplementary Table 5. Annotation descriptions of exome sequence SNVs for *GLP1R* and *G6PC2* in all SNVs and stratified by MAF<1% and MAF≥1%

	GLP1R			G6PC2		
Variation type	All MAF	MAF<1%	MAF≥1%	All MAF	MAF<1%	MAF≥1%
Nonsynonymous	34	30	4	33	30	3
Splicing	0	0	0	1	1	0
Stopgain	2	2	0	4	4	0
Synonymous	29	24	5	9	9	0
Intronic	81	61	20	19	16	3
3′ UTR	3	3	0	2	1	1
5′ UTR	1	1	0	0	0	0
TOTAL	150	121	29	68	61	7

MAF: minor allele frequency; UTR: untranslated region

Supplementary Table 6. Novel *ABO* SNPs associated with fasting glucose in African and European ancestries combined and stratified by ancestry

rsID	Alleles		Combined ancestry analysis					European ancestry analysis					African ancestry analysis				
	Effect	Other	EAF	Beta	SE	p	$p_{cond}{}^{a}$	EAF	Beta	SE	p	$p_{cond}{}^{a}$	EAF	Beta	SE	p	$p_{cond}{}^{a}$
rs507666	A	G	0.17	0.02	0.004	7.4E−08	0.28	0.19	0.02	0.004	4.0E−07	0.60	0.10	0.03	0.015	0.02	0.11
rs651007	A	G	0.20	0.02	0.004	1.3E−08	NA	0.21	0.02	0.004	5.8E−08	NA	0.14	0.02	0.013	0.09	NA
rs579459	C	T	0.20	0.02	0.004	2.6E−08	0.15	0.21	0.02	0.004	6.5E−08	0.25	0.14	0.02	0.013	0.10	0.38
rs635634	A	G	0.17	0.02	0.004	1.4E−08	0.17	0.19	0.02	0.004	1.9E−07	0.46	0.11	0.03	0.014	0.02	0.07

Fasting glucose concentrations were adjusted for sex, age, cohort effects and up to 10 principal components in up to 60,564 (African ancestry n=9664 and European ancestry n=50,900) non-diabetic individuals. Effects are reported per copy of the minor allele. Beta coefficient units are in mmol/L. EAF: Effect allele frequency. a Conditional p-value; variants near the ABO locus were conditioned on the most significant SNP in the region (rs651007)

Supplementary Table 7. Associations of *ABO* variants genotyped on the HumanExome BeadChip with FG in combined European and African ancestries

Gene	rsID	Chr	Build 37 Position	Variation type	Alleles		EAF	N	Beta	SE	p	p_{cond}[a]	Proxy[b]
					Effect	Other							
ABO	rs7466899	9	136131069	nonsynonymous	A	G	0.001	59748	0.077	0.047	1.0E−01	0.07	
ABO	rs201604341	9	136131119	nonsynonymous	A	G	0.000	59748	0.085	0.250	7.3E−01	0.85	
ABO	rs8176749	9	136131188	nonsynonymous	T	C	0.089	59748	−0.001	0.006	9.2E−01	0.03	Y
ABO	rs8176746	9	136131322	synonymous	A	C	0.089	59748	0.000	0.006	9.3E−01	0.03	Y
ABO	rs8176745	9	136131347	nonsynonymous	A	G	0.246	59016	−0.007	0.004	4.4E−02	0.28	
ABO	rs35494115	9	136131389	nonsynonymous	A	G	0.001	59748	−0.075	0.058	1.9E−01	0.24	
ABO	rs201439325	9	136131407	nonsynonymous	A	G	0.000	59748	−0.103	0.131	4.3E−01	0.20	
ABO	rs8176741	9	136131461	nonsynonymous	A	G	0.089	59748	0.000	0.006	1.0E+00	0.02	Y
ABO	rs8176740	9	136131472	nonsynonymous	T	A	0.246	59748	−0.007	0.004	4.0E−02	0.31	Y
ABO	rs55764262	9	136131539	nonsynonymous	G	A	0.000	57931	0.056	0.131	6.7E−01	0.92	
ABO	rs55727303	9	136131576	nonsynonymous	T	C	0.017	57615	−0.003	0.012	7.8E−01	0.68	
ABO	rs7853989	9	136131592	nonsynonymous	C	G	0.109	49364	0.004	0.006	4.6E−01	0.47	Y
ABO	rs201567722	9	136131629	nonsynonymous	T	C	0.000	59748	0.034	0.090	7.0E−01	0.91	
ABO	rs55756402	9	136131630	nonsynonymous	A	G	0.001	59748	0.079	0.057	1.6E−01	0.92	
ABO	rs200932155	9	136131635	nonsynonymous	A	G	0.001	57615	0.067	0.053	2.0E−01	0.36	
ABO	rs8176738	9	136131636	nonsynonymous	T	C	0.002	59748	0.003	0.037	9.3E−01	0.31	
ABO	rs1053878	9	136131651	nonsynonymous	A	G	0.094	59748	−0.005	0.006	4.1E−01	0.34	Y
ABO	rs201105186	9	136131740	nonsynonymous	A	G	0.000	59748	0.023	0.075	7.6E−01	0.91	
ABO	rs8176721	9	136132852	nonsynonymous	A	G	0.023	59748	−0.018	0.012	1.4E−01	0.25	
ABO	rs8176720	9	136132873	nonsynonymous	C	T	0.363	59748	−0.005	0.003	1.1E−01	0.92	Y
ABO	rs512770	9	136133506	nonsynonymous	A	G	0.217	59748	−0.004	0.004	2.9E−01	0.30	Y
ABO	rs56335272	9	136135236	nonsynonymous	T	C	0.002	59748	0.057	0.043	1.9E−01	0.14	
ABO	rs549446	9	136135238	nonsynonymous	T	C	0.257	59748	−0.007	0.004	3.6E−02	0.27	
ABO	rs688976	9	136136770	nonsynonymous	A	C	0.258	59748	−0.007	0.004	3.6E−02	0.26	
ABO	rs8176696	9	136136773	nonsynonymous	T	C	0.022	59748	−0.013	0.011	2.3E−01	0.20	
ABO	rs687621	9	136137065	intronic	C	T	0.348	59748	0.012	0.003	2.0E−04	0.10	
ABO	rs55876802	9	136137547	nonsynonymous	A	C	0.019	59748	0.013	0.011	2.2E−01	0.38	
ABO	rs55917063	9	136137554	nonsynonymous	T	C	0.003	59748	−0.081	0.034	1.5E−02	0.02	
ABO	rs657152	9	136139265	intronic	T	G	0.366	59748	0.012	0.003	2.1E−04	0.39	
ABO	rs514659	9	136142203	intronic	C	A	0.351	59748	0.010	0.003	1.5E−03	0.39	
ABO	rs505922	9	136149229	intronic	C	T	0.334	59748	0.011	0.003	1.2E−03	0.39	
ABO	rs507666	9	136149399	intronic	A	G	0.173	59748	0.022	0.004	**7.4E−08**	0.28	
ABO-SURF6	rs651007	9	136153875	intergenic	A	G	0.196	59502	0.022	0.004	**1.3E−08**	NA	
ABO-SURF6	rs579459	9	136154168	intergenic	C	T	0.196	59502	0.022	0.004	**1.6E−08**	0.15	
ABO-SURF6	rs635634	9	136155000	intergenic	A	G	0.172	59502	0.023	0.004	**1.4E−08**	0.17	

Fasting glucose concentrations were adjusted for sex, age, cohort effects and up to 10 principal components in up to 60,564 non-diabetic individuals of African and European ancestry. Effects are reported per copy of the minor allele. Beta coefficient units are in mmol/L. Bolded p-values meet significance threshold for single variant analysis. EAF: effect allele frequency; N: sample size. [a] Conditional p-value; variants near the ABO locus were conditioned on the most significant SNP in the region (rs651007). [b] Variant is a proxy for one of the major blood group alleles A_1, A_2, B or O

Supplementary Table 8. Pleiotropic associations at the *ABO* locus from previously reported studies

rsID	Trait	Effect Allele	Reported effect (Beta or OR)	Reported *p*	r^2	Reference
rs651007*	sE-selectin	T	−17.23	1.2E−44	1	1
rs579459*	CAD (OR)	C	1.1	4.1E−14	1	2
rs651007*	TC	T	2.3	8.7E−21	1	2
rs579459*	TC	C	1.72	3.8E−03	1	2
rs651007	LDL-C	A	2.2026	9.8E−09	1	3
rs579459*	LDL-C	C	1.54	4.9E−03	1	3
rs649129	LDL-C	T	2.24	6.0E−13	1	4
rs495828	RBC	T	−0.091	3.3E−12	1	4
rs495828	Hb	T	−0.089	1.2E−11	1	4
rs495828	Ht	T	−0.081	6.1E−10	1	4
rs8176746	MCHC	T	0.084	4.3E−08	0.01	5
rs612169	FAaP	G	NR	9.1E−40	0.51	6
rs507666*	sICAM-1	A	−17.3	3.0E−91	0.96	6
rs514659	disposition index	C	−0.09	3.8E−09	0.53	7

CAD: coronary artery disease, OR: odds ratio, TC: total cholesterol, LDL-C: low-density lipoprotein cholesterol, RBC: red blood cell, Hb:hemoglobin concentration, Ht: hematocrit, MCHC: mean corpuscular hemoglobin concentration, FAaP: fibrinogen A-alpha phosphorylation, NR: not reported, sICAM-1: soluable ICAM1. *Variants associated with fasting glucose in Table 1 and Supplementary Table 10. r^2 is between rs651007 and each SNP listed the first column

References:

1. Qi, L. et al. Genetic variants in ABO blood group region, plasma soluble E-selectin levels and risk of type 2 diabetes. Hum Mol Genet 19, 1856-62 (2010).

2. Schunkert, H. et al. Large-scale association analysis identifies 13 new susceptibility loci for coronary artery disease. Nat Genet 43, 333-8 (2011).

3. Teslovich, T.M. et al. Biological, clinical and population relevance of 95 loci for blood lipids. Nature 466, 707-13 (2010). 4. Kamatani, Y. et al. Genome-wide association study of hematological and biochemical traits in a Japanese population. Nat Genet 42, 210-5 (2010).

5. Kim, Y.J. et al. Large-scale genome-wide association studies in East Asians identify new genetic loci influencing metabolic traits. Nat Genet 43, 990-5 (2011).

6. Suhre, K. et al. Human metabolic individuality in biomedical and pharmaceutical research. Nature 477, 54-60 (2011). 7. Huyghe, J.R. et al. Exome array analysis identifies new loci and low-frequency variants influencing insulin processing and secretion. Nat Genet 45, 197-201 (2013).

Supplementary Table 9. Lookups of *ABO* top hits in adiposity, lipid, and blood pressure traits

Adiposity (N_{max}=64,965)

rsID	Alleles		BMI women			BMI men			WHR women			WHR men		
	Effect	Other	Beta	SE	*p*	Beta	SE	*p*	Beta	SE	*p*	Beta	SE	*p*
rs507666	A	G	0.021	0.01	4.0E−03	−0.006	0.01	0.53	−0.011	0.01	0.18	0.001	0.01	0.93
rs651007	A	G	0.025	0.01	3.1E−04	−0.004	0.01	0.62	−0.010	0.01	0.23	0.003	0.01	0.77
rs579459	C	T	0.025	0.01	3.5E−04	−0.004	0.01	0.61	−0.010	0.01	0.24	0.003	0.01	0.80
rs635634	A	G	0.020	0.01	6.0E−03	−0.005	0.01	0.60	−0.010	0.01	0.22	0.002	0.01	0.86

Lipids (N$_{max}$=56,538)

rsID	Alleles		HDL-C			LDL-C			TG			TC		
	Effect	Other	Beta	SE	p	Beta	SE	p	Beta*	SE	p	Beta	SE	p
rs507666	A	G	0.103	0.11	0.36	2.594	0.30	1.9E−18	−0.001	0.00	0.73	2.748	0.33	5.0E−17
rs651007	A	G	0.052	0.11	0.63	2.276	0.28	6.1E−16	−0.001	0.00	0.88	2.397	0.33	3.4E−13
rs579459	C	T	0.052	0.11	0.63	2.274	0.28	6.5E−16	−0.001	0.00	0.88	2.394	0.33	3.6E−13
rs635634	A	G	0.097	0.11	0.40	2.594	0.30	1.9E−18	−0.002	0.00	0.61	2.681	0.35	1.0E−14

Blood pressure (N$_{max}$=92,615)

rsID	Alleles		DBP			SBP		
	Effect	Other	Beta	SE	p	Beta	SE	p
rs507666	A	G	−0.153	0.07	2.5E−02	−0.011	0.11	0.92
rs651007	A	G	−0.096	0.07	0.14	0.016	0.11	0.88
rs579459	C	T	−0.095	0.07	0.15	0.022	0.11	0.84
rs635634	A	G	−0.144	0.07	3.5E−02	0.003	0.11	0.98

Fasting lipid concentrations were used. Individuals on lipid or blood pressure lowering medication had their individual values adjusted, see Methods for details. Analyses were adjusted for sex (adiposity was stratified by sex), BMI (for WHR), age, cohort effects and up to 10 principal components. Effects are reported per copy of the minor allele. Beta coefficient units are in kg/m^2 for BMI, mg/dL for lipids and mmHg for blood pressure. BMI: body mass index, WHR: waist-hip ratio, HDL-C: high-density lipoprotein cholesterol, LDL-C: low-density lipoprotein cholesterol, TG: triglycerides, TC: total cholesterol, DBP: diastolic blood pressure, SBP: systolic blood pressure. *Triglycerides are log transformed.

Supplementary Table 10. Associations with ABO SNVs and eQTLs from the GTEx database

Gene Id	Gene Symbol	rsID	p	Tissue
ENSG00000160326.9	*SLC2A6*	rs507666	1.1E−04	Whole_Blood
ENSG00000175164.8	*ABO*	rs651007	5.9E−05	Whole_Blood
ENSG00000175164.8	*ABO*	rs579459	6.7E−05	Whole_Blood
ENSG00000160326.9	*SLC2A6*	rs635634	1.1E−04	Whole_Blood

Supplementary Table 11. Gene based association results for *G6PC2* removing sets of rare SNVs

cMAF	SNVs(n)	p$_{SKAT}$	Variant in SKAT gene-based test
0.014	4	8.3E−18	Y177H, S207Y, R283X, P324S
0.003	11	0.36	Removing 4 above

SKAT gene-based test with and without the 4 significant variants identified in single variant analyses. Initital gene-based tests (Table 2) used 15 rare SNVs (MAF<0.01) and annotated as nonsynonymous, splice-site, or loss/gain-of-function variants.

cMAF: cumulative minor allele frequency; SNVs(n): number of SNVs in gene-based SKAT test; p$_{SKAT}$: p-value from gene-based SKAT analysis

Supplementary Table 12. Association of fasting glucose and *G6PC2* haplotypes of 15 rare SNVs

| rs142189264 | rs149874491 | rs201561079 | rs199682245 | rs187707963 | rs2232322 | rs145050507 | rs138726309 | rs2232323 | rs145217135 | rs147360987 | rs150538801 | rs148689354 | rs146779637 | rs2232326 | 15 rare SNVs (Overall p-value=1.1e-17) | | |
S30F	I38I	I63T	N68I	Y124C	I171V	I171T	H177Y	Y207S	I230T	H250Y	F256L	I273V	R283X	S324P	N study	Beta	*p*
C	A	T	A	A	A	T	C	A	T	C	T	A	**T**	T	17	−0.22	2.84E−10
C	A	T	A	A	A	T	C	A	T	C	T	A	C	**C**	13	−0.26	1.40E−07
C	A	T	A	A	A	T	C	**C**	T	C	T	A	C	T	18	−0.11	1.45E−06
C	A	T	A	A	A	T	**T**	**C**	T	C	T	A	C	T	3	−0.89	0.005
C	A	T	NA	A	A	T	**T**	A	T	C	T	A	**T**	T	1	1.31	0.005
C	A	T	A	A	A	T	**T**	A	T	C	T	A	C	T	16	−0.09	0.021
T	A	T	A	A	A	T	C	A	T	C	T	A	C	T	10	−0.22	0.029
C	A	T	A	A	**G**	T	C	A	T	C	T	A	C	T	7	0.22	0.134
C	A	T	A	A	A	T	**T**	A	T	C	T	A	C	T	1	−0.52	0.139
C	**C**	T	A	A	A	T	C	A	T	C	T	A	C	T	3	−0.19	0.140
C	A	**C**	A	A	A	T	C	A	T	C	T	A	C	T	2	0.57	0.216
C	A	T	A	A	A	T	C	A	T	C	**C**	A	C	T	11	−0.13	0.220
C	A	T	A	**G**	A	T	C	A	T	C	T	A	C	T	1	−0.48	0.407
C	A	T	A	NA	A	T	C	A	**C**	C	T	A	C	T	1	0.91	0.417
C	A	T	A	A	A	**C**	C	A	T	C	T	A	C	T	11	−0.07	0.435
C	A	T	A	A	A	T	C	**C**	T	C	T	A	C	**C**	1	−1.10	0.438
C	A	T	A	A	A	T	**T**	A	T	C	T	A	C	**C**	1	−0.73	0.592
T	A	T	A	A	A	T	C	A	T	C	T	A	C	**C**	1	0.21	0.645
C	A	T	**T**	A	A	T	C	A	T	C	T	A	C	**C**	3	−0.21	0.700
C	A	T	NA	A	A	T	C	A	**C**	C	T	A	C	T	1	0.10	0.833
C	A	T	A	A	A	T	C	A	T	C	T	A	C	T	18	NA	NA

18 cohorts contributed data. NA is the reference haplotype. Yellow highlighted alleles are the minor allele. N study: number of studies contributing the haplotype observed.

Supplementary Table 13. Association of *G6PC2* haplotypes of 15 rare SNVs and one common SNV (rs560887) with fasting glucose

rs142189264	rs149874491	rs201561079	rs199682245	rs187707963	rs560887	rs2232322	rs145050507	rs138726309	rs2232323	rs145217135	rs147360987	rs150538801	rs148689354	rs146779637	rs2232326	15 rare SNVs plus rs560887 (Overall p-value=1.5e-81)		
S30F	I38I	I63T	N68I	Y124C	NA	I171V	I171T	H177Y	Y207S	I230T	H250Y	F256L	I273V	R283X	S324P	N study	Beta	p
C	A	T	A	A	T	A	T	C	A	T	C	T	A	C	T	18	−0.08	8.92E−77
C	A	T	A	A	T	A	T	C	A	T	C	T	A	T	T	17	−0.24	5.63E−12
C	A	T	A	A	T	A	T	C	C	T	C	T	A	C	T	18	−0.13	7.70E−09
C	A	T	A	A	C	A	T	C	A	T	C	T	A	C	C	13	−0.28	2.02E−08
C	A	T	A	A	T	A	T	T	C	T	C	T	A	C	T	3	−0.95	0.003
C	A	T	NA	A	T	A	T	T	A	T	C	T	A	T	T	1	1.29	0.006
C	A	T	A	A	C	A	T	T	A	T	C	T	A	C	T	16	−0.11	0.007
T	A	T	A	A	C	A	T	C	A	T	C	T	A	C	T	10	−0.28	0.010
C	A	T	A	A	C	A	T	C	A	T	C	C	A	C	T	11	−0.19	0.10
C	A	T	A	A	C	G	T	C	A	T	C	T	A	C	T	6	0.32	0.17
C	C	T	A	A	C	A	T	C	A	T	C	T	A	C	T	3	−0.18	0.23
C	A	C	A	A	C	A	T	C	A	T	C	T	A	C	T	2	0.51	0.26
C	A	T	A	A	T	A	T	T	A	T	C	T	A	C	T	9	−2.51	0.30
C	A	T	A	G	C	A	T	C	A	T	C	T	A	C	T	1	−0.53	0.36
C	A	T	T	A	C	A	T	C	A	T	C	T	A	C	T	1	−0.45	0.37
C	A	T	A	A	T	A	C	C	A	T	C	T	A	C	T	7	−0.38	0.40
C	A	T	A	A	T	A	T	C	C	T	C	T	A	C	C	1	−1.13	0.43
C	A	T	A	A	C	A	C	C	A	T	C	T	A	C	T	9	−0.08	0.43
C	A	T	A	NA	C	A	T	C	A	C	C	T	A	C	T	1	0.86	0.44
C	A	T	T	A	T	A	T	C	A	T	C	T	A	C	T	1	−0.96	0.53
C	A	T	A	A	C	A	T	T	A	T	C	T	A	C	C	1	−0.78	0.57
C	C	T	A	A	T	A	T	C	A	T	C	T	A	C	T	1	−200	0.57
C	A	T	A	A	C	A	T	C	C	T	C	T	A	C	T	4	−0.32	0.58
C	A	T	T	A	C	A	T	C	A	T	C	T	A	C	C	3	−0.25	0.64
C	A	T	A	A	T	A	T	C	A	T	C	T	A	C	C	8	−0.22	0.66
T	A	T	A	A	C	A	T	C	A	T	C	T	A	C	C	1	0.16	0.71
C	A	T	A	A	T	A	T	C	A	T	C	C	A	C	T	4	0.13	0.74
C	A	T	A	A	C	A	T	C	A	T	C	T	A	T	T	2	−619118	0.76
C	A	T	A	A	T	G	T	C	A	T	C	T	A	C	T	5	0.06	0.79
T	A	T	A	A	T	A	T	C	A	T	C	T	A	C	T	6	0.11	0.81
C	A	T	NA	A	C	A	T	C	A	C	C	T	A	C	T	1	0.05	0.91
C	A	T	A	A	C	A	T	C	A	T	C	T	A	C	T	18	NA	NA

18 cohorts contributed data. NA is the reference haplotype. Yellow highlighted alleles are the minor allele. N study: number of studies contributing the haplotype observed.

Supplementary Table 14. Gene based association results for *G6PC2* and fasting insulin

Ancestry	cMAF	SNVs(n)	p_{WST}	p_{SKAT}	N
Combined ancestry	0.02	17	0.11	0.53	47388
European ancestry only	0.02	14	0.45	0.55	38528
African ancestry only	0.01	13	0.75	0.76	8860

Analyses adjusted for sex, age, cohort effects and up to 10 principal components in up to 47,388 in the combined ancestry analysis, 38,528 in the European ancestry analysis, and 8860 in the African ancestry analysis. SNVs(n), number of variants included in the analysis; variants were restricted to those with MAF<0.01 and annotated as nonsynonymous splice-site, or loss/gain-of-function variants. cMAF, cumulative MAF: combined minor allele frequency of all variants included in the analysis; pWST: p-value from weighted sum test (WST); pSKAT: p-value from sequence kernal association test (SKAT); N: sample size.

Supplementary Table 15. Gene based association results for *G6PC2* and type 2 diabetes

Ancestry	cMAF	SNVs(n)	p_{WST}	p_{SKAT}	N
Combined ancestry	0.019	18	0.75	0.68	34984
European ancestry only	0.019	16	0.49	0.52	17651
African ancestry only	0.010	15	0.60	0.60	3814

Analyses adjusted for sex, age, cohort effects and up to 10 principal components in up to 16,491 T2D cases and 81,877 controls in the combined ancestry analysis, 10,240 T2D cases and 63,105 controls in the European ancestry analysis, and 3,097 T2D cases and 10,326 controls in the African ancestry analysis. cMAF, cumulative MAF: combined minor allele frequency of all variants included in the analysis. SNVs(n), number of variants included in the analysis; variants were restricted to those with MAF<0.01 and annotated as nonsynonymous, splice-site, or loss/gain-of-function variants. pWST: p-value from weighted sum test (WST); pSKAT: p-value from sequence kernal association test (SKAT); N: sample size

Supplementary Table 16. Gene based association results for *G6PC2* and fasting glucose from exome sequence analyses in up to 7,452 individuals of European ancestry

Gene	cMAF	SNVs(n)	p_{WST}	p_{SKAT}
G6PC2	0.027	36	5.4E−04	1.4E−03
G6PC2 (exome chip variants)	0.018	10	3.2E−03	1.3E−03
G6PC2 (excluding exome chip variants)	0.009	26	4.0E−02	6.1E−01

cMAF, cumulative MAF: combined minor allele frequency of all variants included in the analysis; SNVs(n): number of variants included in the analysis; pWST: p-value from weighted sum test (WST). pSKAT: p-value from sequence kernal association test (SKAT). Variants were restricted to those with MAF<0.01 and annotated as nonsynonymous splice-site, or loss/gain-of-function variants.SNVs(n)=36 variants met criteria for inclusion in gene based tests. SNVs(n)=10 are the same variants available on the exome chip. SNVs(n)=26 are the the variants available for analyses after excluding the 10 above.

Supplementary Table 17. Top ten pathways with lowest p-values in MAGENTA analysis of FG, analyzing all genes (A) and excluding those with known associations with FG (B)

A.

Database	Pathway name	GSEA p-value	FDR
GOTERM	Positive regulation of DNA replication	9.80E−05	1.33E−01
KEGG	Glioma	3.00E−04	4.80E−02
GOTERM	Pancreas development	5.00E−04	2.24E−01
PANTHER	Insulin/IGF pathway, protein kinase B signaling cascade	1.00E−03	1.16E−01
REACTOME	Signal attenuation	1.00E−03	1.43E−01
KEGG	Citrate cycle, TCA cycle	1.70E−03	4.63E−02
INGENUITY	IGF-1 signaling	1.80E−03	6.32E−02
PANTHER	DNA repair	1.80E−03	4.61E−01
KEGG	Acute myeloid leukemia	1.90E−03	5.21E−02
KEGG	Type 2 diabetes mellitus	2.20E−03	5.84E−02

B.

Database	Pathway name	GSEA p-value	FDR
KEGG	Glioma	1.00E−04	1.59E−02
GOTERM	Positive regulation of DNA replication	1.12E−04	1.68E−01
KEGG	Acute myeloid leukemia	1.50E−03	5.40E−02
KEGG	Focal adhesion kinase	1.60E−03	1.12E−01
KEGG	Non-homologous end joining	2.80E−03	4.50E−02
KEGG	Citrate cycle, TCA cycle	3.00E−03	6.63E−02
BIOCARTA	IGF1 pathway	3.00E−03	3.08E−01
GOTERM	Centrosome organization	3.20E−03	5.03E−01
GOTERM	Oligodendrocyte development	3.70E−03	6.88E−01
REACTOME	Cell cycle, mitotic	3.80E−03	7.37E−01

Supplementary Table 18. Top ten pathways with lowest p-values in MAGENTA analysis of FI, analyzing all genes (A) and excluding those with known associations with FI (B)

A.

Database	Pathway name	GSEA p-value	FDR
GOTERM	Response to DNA damage stimulus	9.80E−05	1.33E−01
REACTOME	Regulation of IGF activity by IGF binding proteins	3.00E−04	4.80E−02
GOTERM	ATP binding	5.00E−04	2.23E−01
GOTERM	Solute:hydrogen antiporter activity	1.00E−03	1.16E−01
REACTOME	PECAM1 interactions	1.00E−03	1.44E−01
GOTERM	Dephosphorylation	1.70E−03	4.63E−02
GOTERM	Positive regulation of smooth muscle contraction	1.80E−03	6.32E−02
BIOCARTA	ATRBRCA pathway	1.80E−03	4.61E−01
GOTERM	Rab GTPase binding	1.90E−03	5.21E−02
GOTERM	Nucleotide binding	2.20E−03	5.84E−02

B.

Database	Pathway name	GSEA p-value	FDR
GOTERM	Response to DNA damage stimulus	1.00E−04	1.59E−02
GOTERM	ATP binding	1.12E−04	1.68E−02
REACTOME	PECAM1 interactions	1.50E−03	5.40E−02
GOTERM	Protein C-terminus binding	1.60E−03	1.12E−01
GOTERM	Solute:hydrogen antiporter activity	2.80E−03	4.50E−02
GOTERM	Dephosphorylation	3.00E−03	6.63E−01
GOTERM	Positive regulation of smooth muscle contraction	3.00E−03	3.08E−01
GOTERM	Motor activity	3.20E−03	5.03E−01
GOTERM	Rab GTPase binding	3.70E−03	6.88E−01
GOTERM	Cation transport	3.80E−03	7.37E−01

Supplementary Table 19. MAGENTA results for glucometabolic pathways
from curated pathway databases for FG (A) and FI (B)

A.

Database	Pathway name	GSEA p-value	FDR
GOTERM	Cellular metabolic process	3.50E−03	6.55E−02
PANTHER	Metabolism of cyclic nucleotides	1.67E−02	4.33E−01
GOTERM	Positive regulation of insulin secretion	1.77E−02	1.36E−01
REACTOME	Glucose and other sugar SLC transporters	2.83E−02	1.10E−01
KEGG	Alpha linoleic acid metabolism	3.88E−02	5.79E−01
PANTHER	Lipid, fatty acid and steroid metabolism	4.17E−02	4.19E−01
GOTERM	Hydrolase activity, hydrolyzing O-glycosyl compounds	5.24E−02	3.66E−01
GOTERM	Response to glucose stimulus	5.45E−02	2.90E−01
PANTHER	Phospholipid metabolism	5.60E−02	5.65E−01
REACTOME	Metabolism of carbohydrates	5.62E−02	1.00E+00

B.

Database	Pathway name	GSEA p-value	FDR
REACTOME	Regulation of IGF activity by IGF binding proteins	8.00E−04	2.30E−03
GOTERM	Xenobiotic metabolic process	6.60E−03	1.49E−01
GOTERM	Regulation of lipid metabolic process	1.59E−02	1.50E−01
GOTERM	Lipid metabolic process	3.73E−02	5.13E−01
KEGG	Inositol phosphate metabolism	5.22E−02	1.00E+00
GOTERM	Response to glucose stimulus	5.48E−02	2.96E−01
GOTERM	Generation of precursor metabolites and energy	6.08E−02	4.87E−01
GOTERM	Response to glucocorticoid stimulus	6.25E−02	2.35E−01
GOTERM	Glucose homeostasis	7.08E−02	1.73E−01
REACTOME	Peroxisomal lipid metabolism	7.21E−02	8.74E−01

Supplementary Table 20. Identifying coding variants significantly associated with fasting glucose and fasting insulin in known loci

Gene	Index*/ novel	rsID	Chr	Build 37 position	Alleles Effect	Alleles Other	Variation type	EAF	Beta	SE	p	N	r^2	D'
Fasting Glucose														
DNLZ	index	rs3829109	9	139256766	A	G	intronic	0.26	−0.01	0.004	1.4E−03	55633	LD with rs3829109	
GPSM1	nsSNV	rs60980157	9	139235415	T	C	nsSNV	0.24	−0.02	0.004	2.0E−06	55633	0.62	0.88
SLC30A8	index	rs11558471	8	118185733	G	A	UTR3	0.28	−0.04	0.003	5.5E−24	59748	LD with rs11558471	
SLC30A8	nsSNV	rs13266634	8	118184783	T	C	nsSNV	0.27	−0.04	0.004	1.7E−24	59748	0.97	1.00
SLC2A2	index	rs11920090	3	170717521	A	T	intronic	0.17	−0.02	0.004	5.0E−08	59748	LD with rs11920090	
SLC2A2	nsSNV	rs5400	3	170732300	A	G	nsSNV	0.18	−0.03	0.004	3.5E−09	59748	1.00	1.00
													LD with rs17762454*	
	"proxy"	rs675209	6	7102084	T	C	intergenic	0.29	0.00	0.003	8.7E−01	59502	0.46	0.71
RREB1	novel	rs35742417	6	7247344	A	C	nsSNV	0.18	−0.02	0.004	6.9E−07	59748	0.04	0.79
MADD	index	rs7944584	11	47336320	T	A	intronic	0.24	−0.03	0.004	5.1E−12	59748	LD with rs7944584	
MADD	nsSNV	rs35233100	11	47306630	T	C	stopgain	0.05	−0.04	0.007	6.2E−08	59748	0.13	1.00
ACP2	novel	rs2167079	11	47270255	T	C	nsSNV	0.34	0.02	0.003	4.5E−07	59748	0.15	0.93
AGBL2	novel	rs7941404	11	47712213	T	C	nsSNV	0.12	−0.02	0.005	7.0E−06	59748	0.20	0.73
IKBKAP	index	rs16913693	9	111680359	G	T	intronic	0.06	−0.02	0.007	1.9E−02	59748	LD with rs16913693	
FAM206A	novel	rs76817627	9	111696795	T	C	nsSNV	0.02	−0.05	0.010	3.2E−06	59748	1.00	1.00
IKBKAP	novel	rs17853166	9	111679940	C	T	nsSNV	0.02	−0.05	0.010	5.7E−06	59748	1.00	1.00
Fasting Insulin														
	index	rs10195252	2	165513091	C	T	intergenic	0.48	−0.02	0.003	2.3E−09	46332	LD with rs10195252	
COBLL1	index	rs7607980	2	165551201	C	T	nsSNV	0.13	−0.03	0.004	5.9E−11	47388	0.15	0.83

Causal gene for MADD and IKBKAP is undetermined since associations are seen in multiple genes and the SNVs r2<.2. Fasting glucose concentrations were adjusted for sex, age, cohort effects and up to 10 principal components in up to 60,564 (African ancestry n=9664 and European ancestry n=50,900) non-diabetic individuals. Effects are reported per copy of the minor allele. Beta coefficient units are in mmol/L. *No proxy for index (rs17762454) is on the exome chip. p-value threshold: 1.1x10-5=0.05/4513 SNVs analyzed. EAF: effect allele frequency; N: Sample size, UTR: untranslated region; nsSNV:nonsynonymous single nucleotide variant.

Supplementary Table 21. Association of MODY variants with fasting glucose and fasting insulin

Trait	Gene	rsID	Chr	Build 37 Position	Alleles Effect	Other	Variation type	EAF	Beta	SE	*p*	N
FG	*KLF11*	rs34336420	2	10188123	T	C	nonsynonymous	0.007	−0.013	0.021	0.529	59748
	HNF4A	rs139591750	20	43047151	G	A	synonymous	0.002	−0.109	0.036	0.003	59502
FI	*KLF11*	rs34336420	2	10188123	T	C	nonsynonymous	0.008	−0.009	0.019	0.656	47388
	HNF4A	rs139591750	20	43047151	G	A	synonymous	0.003	−0.062	0.033	0.063	47148

Fasting glucose concentrations were adjusted for sex, age, cohort effects and up to 10 principal components in up to 60,564 non-diabetic individuals; beta coefficient units are in mmol/L. Fasting insulin concentrations were log transformed and adjusted for sex, age, BMI, cohort effects and up to 10 principal components in up to 48,118 non-diabetic individuals. Effects are reported per copy of the minor allele. EAF: effect allele frequency; N: sample size.

Supplementary Table 22. Type of variant on the exome chip by allele frequency

Type	All MAF	MAF<1%	MAF>=1%
nonsynonymous	128679	32142	96537
intergenic	8142	8116	26
intronic	5573	5536	37
synonymous	3827	1332	2495
splicing, synonymous	1788	494	1294
splicing, nonsynonymous	2387	491	1896
ncRNA intronic	435	434	1
UTR3	482	431	51
splicing	1079	307	772
stopgain	1862	289	1573
downstream	181	178	3
upstream	177	176	1
ncRNA_exonic	101	92	9
UTR5	69	64	5
exonic;stoploss	125	33	92
exonic;splicing	39	15	24
upstream;downstream	8	8	0
ncRNA_UTR3	7	7	0
exonic;splicing;stopgain	31	4	27
exonic;splicing;stoploss	3	1	2
ncRNA_splicing	1	1	0
ncRNA_UTR5	1	1	0
TOTAL Disruptive	134205	33282	100923
TOTAL	154997	50152	104845

Up to 155,106 SNVs were available for association analyses of fasting glucvose and insulin.
The type of variant is uncategorized for 109 SNVs. MAF: minor allele frequency.

Supplementary Figure 1. Study Design.

Design of CHARGE consortium discovery of novel variants associated with fasting glucose (FG) and fasting insulin (FI), and the type of follow-up performed on novel and known loci
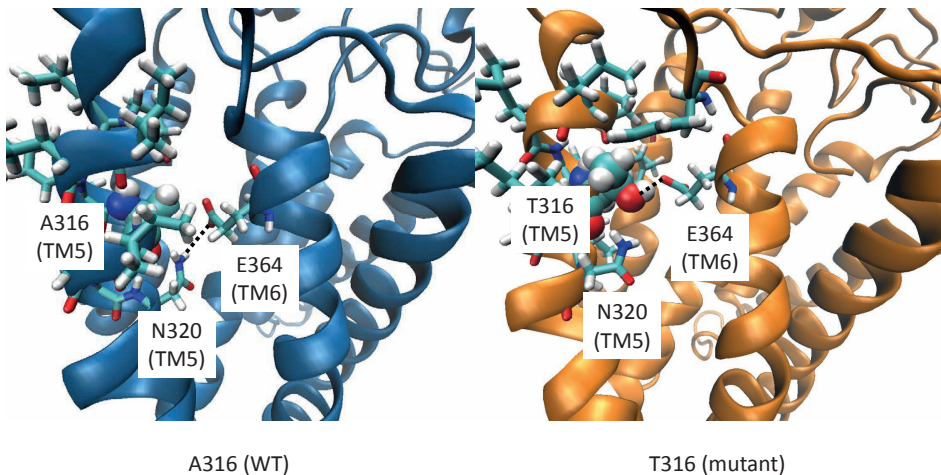
A316 (WT)   T316 (mutant)

Supplementary Figure 2. Detailed Comparison of A316 (WT) with T316 (mutant).

The figure shows that in the wildtype (WT) receptor (A316), residue N320 (transmembrane (TM) 5) is involved in a hydrogen-bonding interaction with E364 (TM6), whereas in the mutant receptor the T316 residue displaces N320 and takes its place to engage in a stable interaction with E364. These changes then affect the positions of TM5 and TM6, as well as the conformation of the intracellular loop 3 (that connects TM5 and TM6 inside the cell).
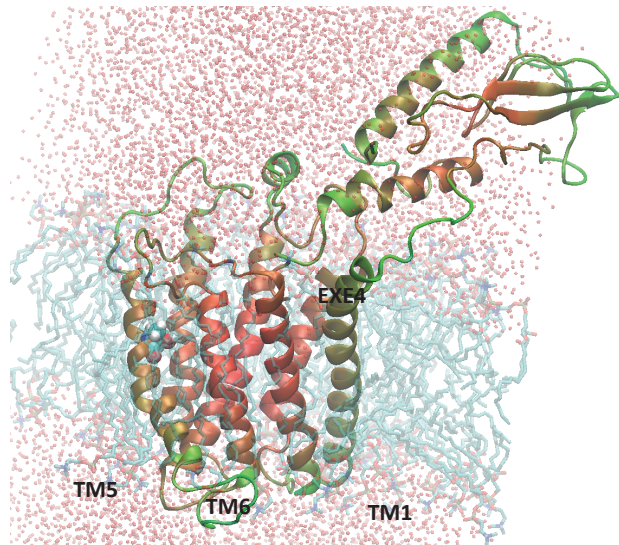
Supplementary Figure 3. Effect of mutant GLP-1 receptor on position in the cell membrane.
This figure displays the receptor mutant embedded in the membrane with the receptor color capturing the fluctuations and deviations (red for less and green/blue for more) in the mutant compared to the wild type (WT) receptor. Transmembrane domain 5 (TM5), which contains A316T, and TM1 show the largest changes in conformation/position.
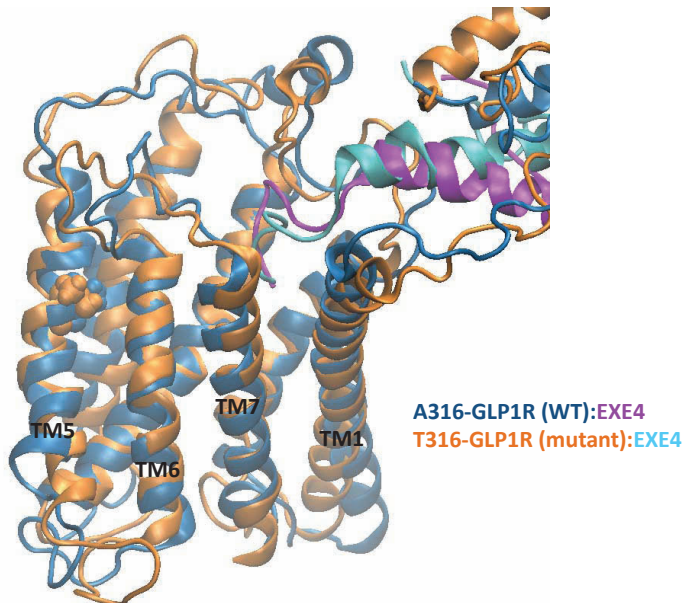
Supplementary Figure 4. Global changes in the transmembrane domains of the mutant GLP1R and exendin-4 (EXE4) system.

This figure compares the global changes in the transmembrane (TM) domains for the A316 (WT) receptor and EXE4 (in blue and purple, respectively) versus the T316 (mutant) receptor and EXE4 (in orange and cyan, respectively), showing that TM5 shifts slightly down towards the cytoplasm and TM6 shifts slightly upward.
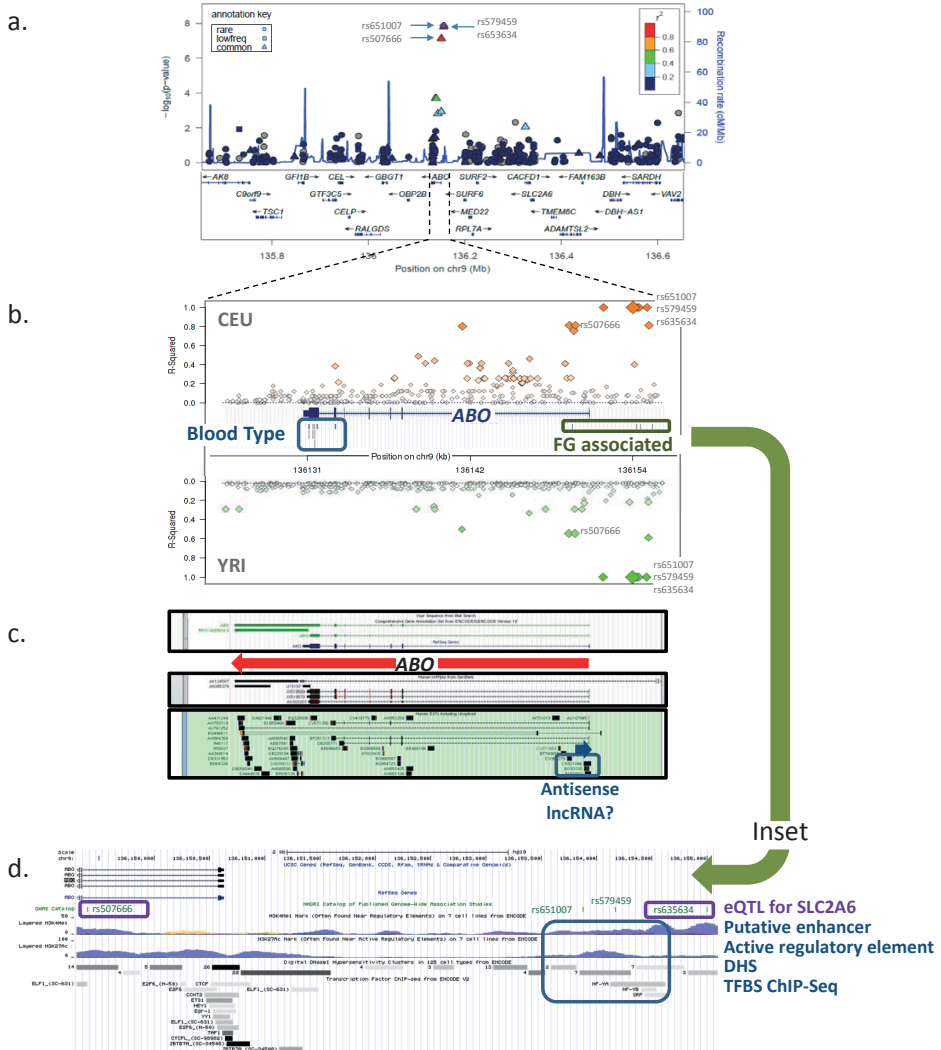
Supplementary Figure 5. Association signals, linkage disequilibrium, transcriptional and epigenetic landscapes of significant SNVs at the *ABO* locus

(a) Regional association results (-log$_{10}p$) for fasting glucose of the *ABO* locus and within 500KB around the lead SNV (rs651007, purple dot); rs579459, rs653634 and rs507666 are also shown and are in strong linkage disequilibrium (LD) with rs651007 (r^2=0.95-1). r^2 indicated by color scale legend. Triangle symbols indicate variants with MAF>5%, square symbols indicate variants with MAF1-5%, and circle symbols indicate variants with MAF <1%.

(b) Inset of *ABO* gene with lead SNVs from FG analysis (labeled "FG associated") depicting low LD (r^2) with the major blood group variants (labeled "Blood Type") in European (CEU, top) and African (YRI, bottom) individuals. Major blood group variants were not genotyped on the exome chip; therefore r^2 was calculated from the 1000 Genomes project (Phase 1, version 3).

(c) An expressed sequence tag (EST)-supported antisense transcript from islets overlaps *ABO* exon 1. RED arrow: Genomic span of the *ABO* gene. The panel highlighted (light green, below) is the UCSC Expressed Sequence Tags (EST) Track. BLUE arrow: Genomic span of the EST-supported antisense transcript. BLUE Ellipse: ESTs supporting antisense transcription.

(d) Inset of the ABO upstream region, promoter, 5' untranslated region (5'UTR) and part of intron 1. The intronic SNV rs507666 is near the transcription start site of the expressed sequence tag (EST) CK821046 from a human islet cDNA library. Two other ESTs, also from human islets, support this antisense non-coding transcript. This EST is antisense to exon 1 of *ABO*, suggesting that rs507666 may function as a promoter SNV of a previously uncharacterized *ABO* antisense non-coding RNA transcript in islet cells. The intergenic SNVs rs651007 and rs579459 reside in a DNAseI hypersensitive site cluster, overlapping an H3K27Ac peak and partially overlapping a transcription factor binding site (TFBS) ChIP-seq peak upstream of the *ABO* promoter. The sequences encompassing these SNVs may, therefore, represent putative active chromatin regulatory elements whose function may be altered by these SNVs. The intergenic SNV rs635634 had less evidence for transcriptional or regulatory activity. Purple Boxes: Two variants (rs507666, rs635634) were annotated as eQTLs for *SLC2A6* (gene to the right of *ABO* in panel A) from GTEx analysis. Blue Box: The ENCODE H3K4Me1 (enhancer), H3K27Ac (active regulatory element), DNAse I hypersensitive sites, and TFBS ChIP-Seq tracks, all with signals overlapping the SNVs, are shown.
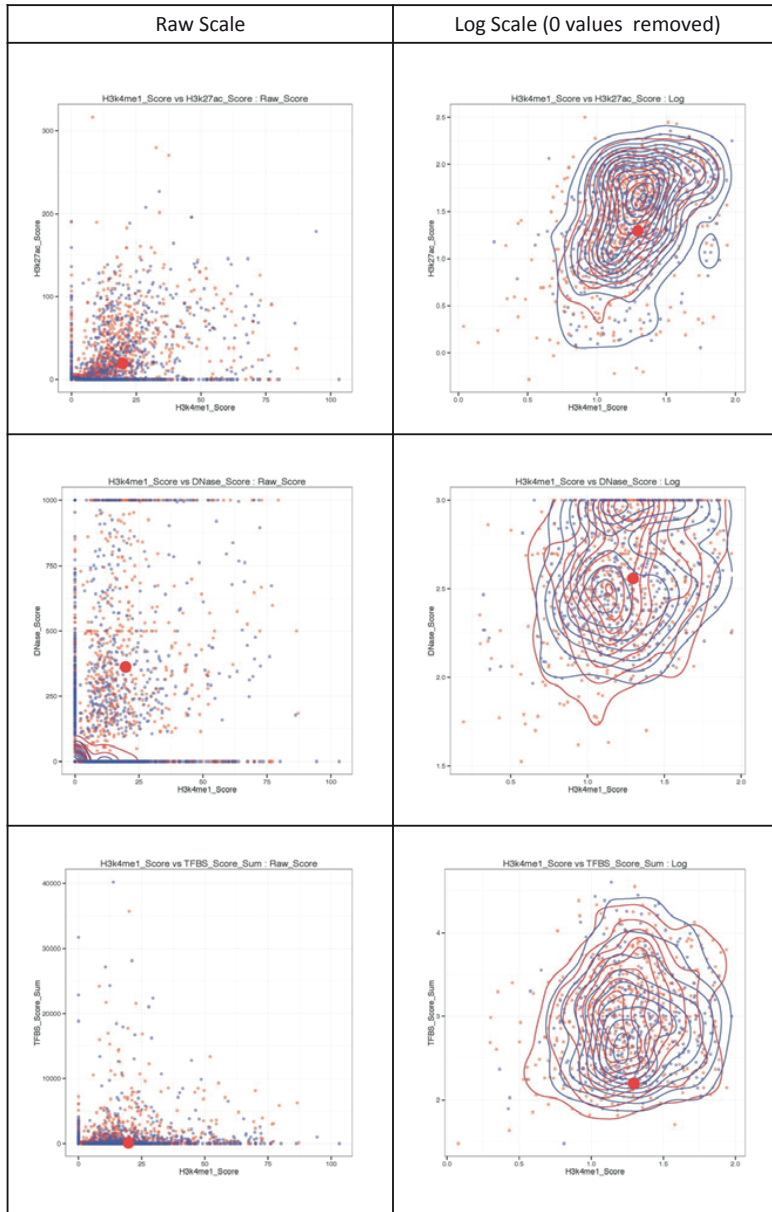
Supplementary Figure 6. ENCODE enrichment analyses.
Overlap of *ABO* locus SNP LD block with ENCODE functional marks. Each plot shows the distribution of all SNPs or SNP LD blocks (*ABO* SNP LD block of interest is the larger red point; smaller red points are GWAS SNPs; blue points are non-GWAS SNPs). Each point represents a single SNP or SNP LD block and its location represents the SNPs overlap (score of each mark overlapped or sum for marks with multiple types; i.e. TFBS) with the ENCODE functional mark denoted. Scores are averaged over LD blocks. Each plot is shown in raw and $\log_{10}$ scale. Contour lines are created with the R function kde2d in order to represent the density of all points in each plot.

Raw Scale	Log Scale (0 values removed)
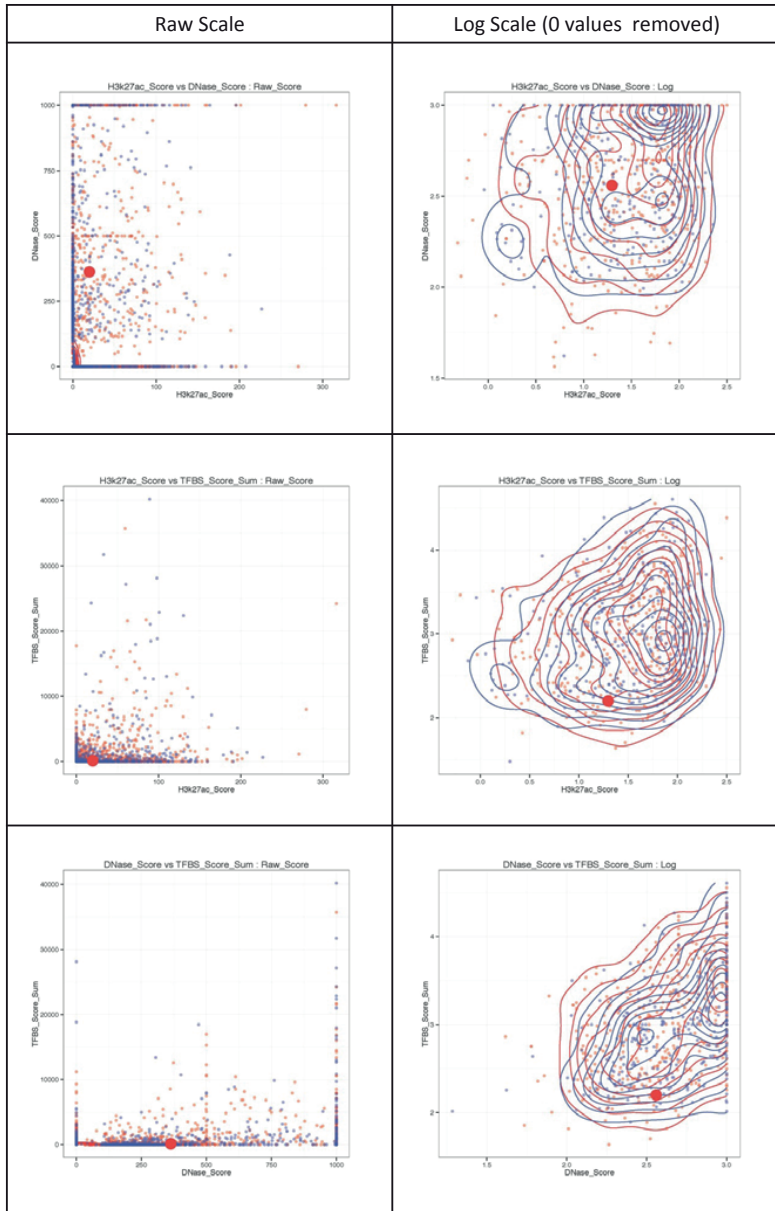
Supplementary Figure 6. Continued

Supplementary Figure 7. The relationship of the common intronic variant rs560887 to epigenetic marks, transcriptional regulation, and splicing at the *G6PC2* locus

(a) Figure of *G6PC2* gene structure showing the location of rs560887 and nearby ENCODE epigenetic signatures.

(b) Zoomed-in plot showing the EST DB031634 and the splice site of *G6PC2* nearest to rs560887.

The intronic SNV rs560887 was assessed as significant in 11 independent NHGRI-catalogued GWAS studies (green, top, A) of serum metabolites, pregnancy-associated glycemia, fasting glucose levels, atherosclerosis, and body mass index (http://genome.ucsc.edu/cgi-bin/hgc?hgsid=369635347_bV0VuuwNQelM7M RwqH5tEyMLIvjx&c=chr2&o=169763147&t=169763148&g=gwasCatalog&i=rs560887). The SNV resides at the 5'end of an ENCODE H3K4Me1 (putative enhancer) signature, suggesting that it may have a regulatory role (Layered H3K4Me1, middle, A). There are several ENCODE transcription factor binding sites evident in ChIP-seq data in the last exon of *G6PC2* (bottom, a), further suggesting that the region may have regulatory functions impacting the expression of *G6PC2* or other genes. The intronic SNV rs560887 is also exonic with respect to the EST DB031634, a positive-strand (same as *G6PC2*) transcript that may represent a cryptic minor isoform of *G6PC2* initiating from an internal promoter in the intron where this SNV resides (b). This SNV is 25 bases upstream of the intron's splice acceptor (b), suggesting that it may also function as a regulator of *G6PC2* splicing.

DNLZ-GPSM1 (FG)
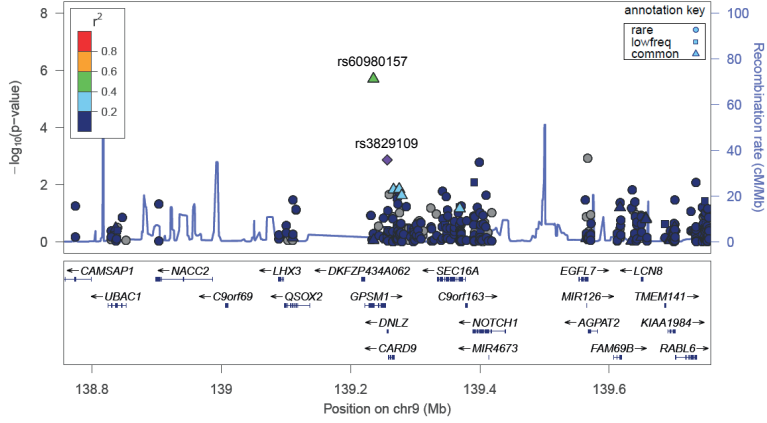
SLC30A8 (FG)
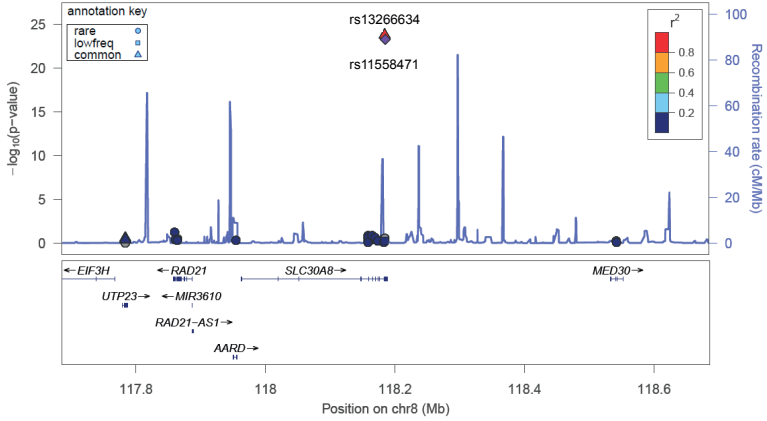
SLC2A2 (FG)
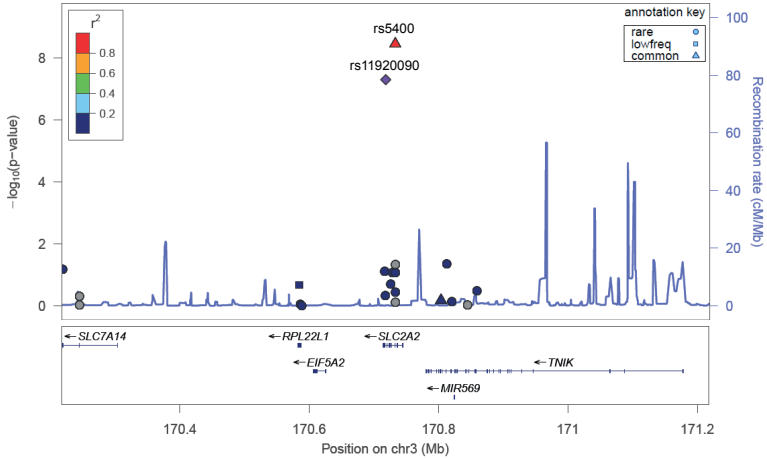
RREB1-SSR1 (FG)
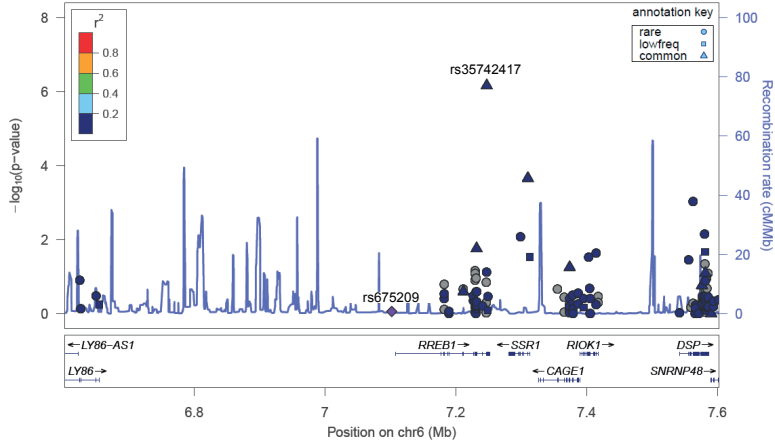
MADD-ACP2 (FG)

IKBKAP-FAM206A (FG)

GRB14-COBLL1 (FI)
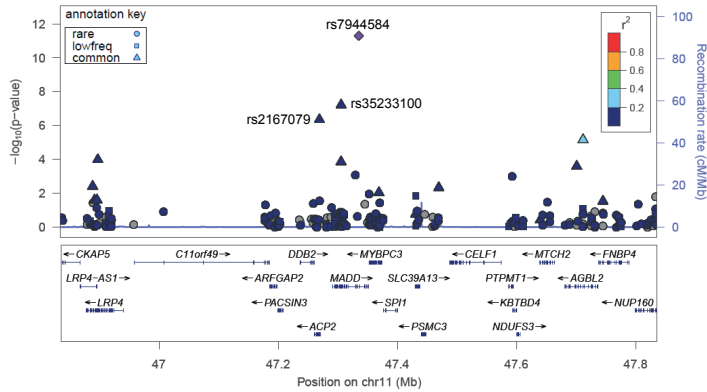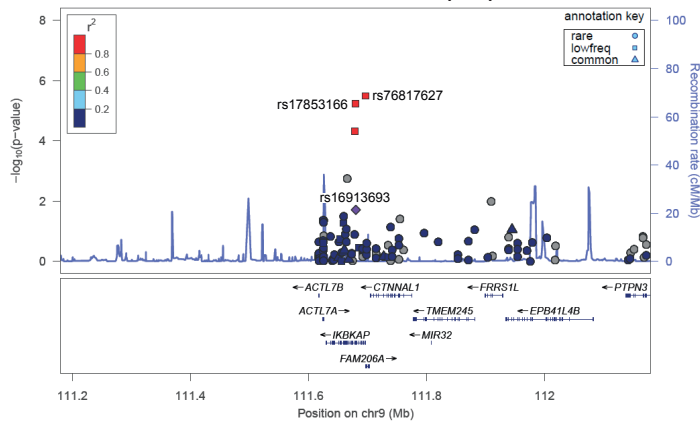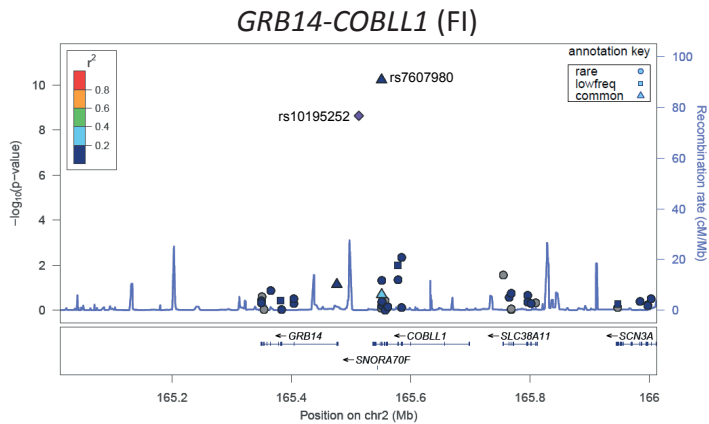
Supplementary Figure 8. Regional association plots for known fasting glucose and fasting insulin loci and including only nonsynonymous, splice or stop-gain/loss variants

Regional association results (-$\log_{10}p$) for fasting glucose (FG) (a-f) and fasting insulin (FI) (g).

(a) DNLZ-GPSM1 (FG)

(b) SLC30A8 (FG)

(c) SLC2A2 (FG)

(d) RREB1 (FG) footnote, rs675209 is the highest quality (but a poor) proxy for the indx GWAS FG SNV (rs17762454; r^2=0.46, D'=0.71) available on the Exome chip

(e) MADD-ACP2 (FG)

(f) IKBKAP-FAM206A (FG)

(g) GRB14-COBLL1 (FI)
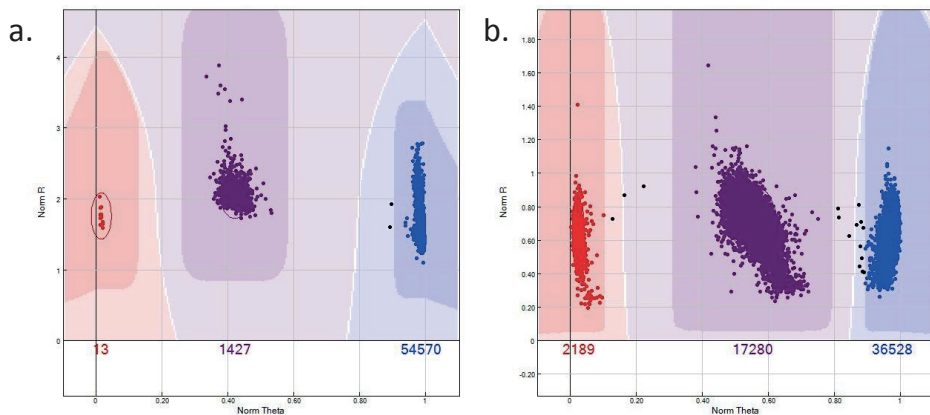
Supplementary Figure 9. Cluster plots of the newly reported variants from CHARGE joint calling.

(a) rs10305492 – *GLP1R* (A316T)

(b) rs651007 – *ABO*
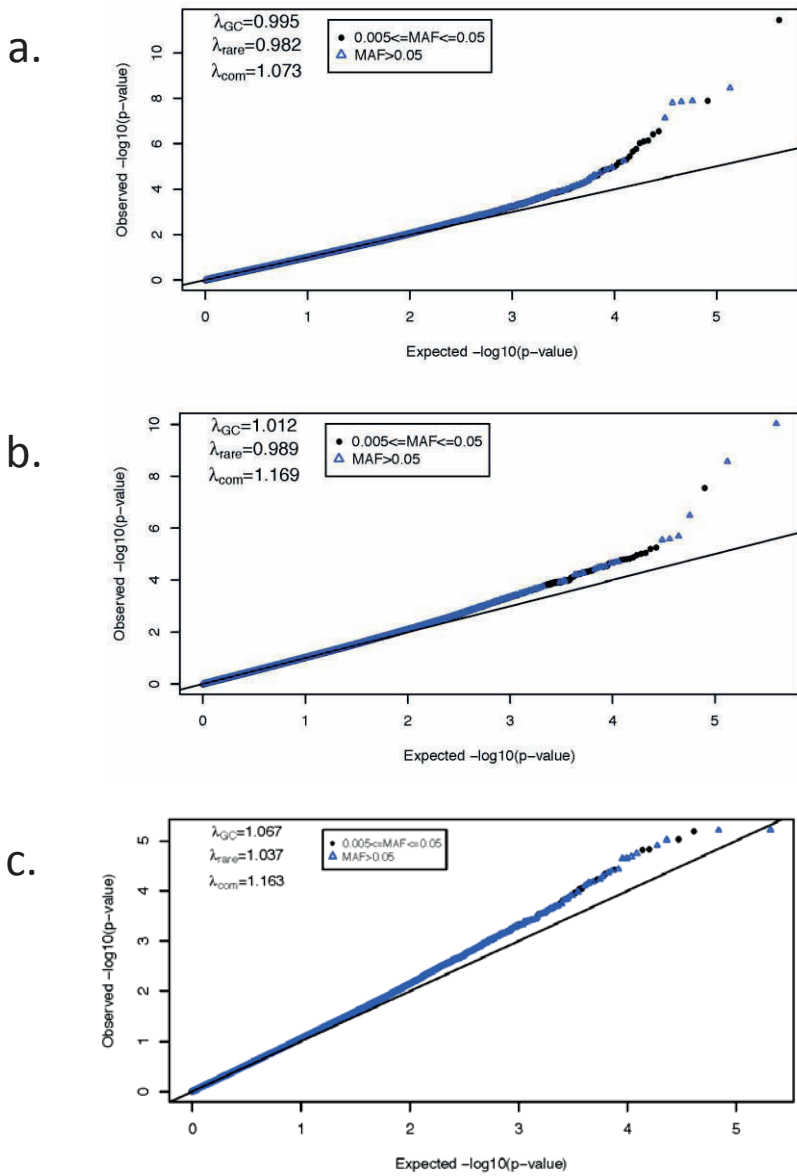
Supplementary Figure 10. Quantile-quantile (QQ) plots from single variant association meta-analysis of (a) fasting glucose, (b) fasting insulin and (c) type 2 diabetes associations without known variants

```
          .TGACTACGTA.                              .CGCGC.
       .CTACTAGCACTGTACACGA.                     .CTACGTACG.
    .CGTACGTAGTCTGATCAAAACGTA.               .GCTATACAGCTA.
   .CGTTGGTACCAGTACGTACGACTGCGA.           .GTACGTACGTACG.
  .GTGTGTCACACAGACACGTGATATAGACCAGATGACACACGTAC
 .GTACGACGGCGCGGCGCGATATATATAAAAGCACGGACACTACGA
 ACGATGACTGACTACGTGCGCGCGTAGCTAGCTACGGATAGGACTA
 TACTGACTACGTACTACTAAAGCTGCTTTTACGTACGTACGGGCATC
 CATGTACTATCGCGGCGCGTAGACTGTACGACCAGACTGACTGACT
 GACTGACTGACGCGCGTACGTAGCTACTGTAGTACGTACGTACGTA
 ITTCAGACTGACTGACTGAGATTACAGCTACGTAGTACTGACACGT
  TACGACGGCGCGGCGCGATATATATAAAAGCACGGACACTACGA
  GATGACTGACTACGTGCGCGCGTAGCTAGCTACGGATAGGACTA
   TGACTACGTACTACTAAAGCTGCTTTTACGTACGTACGGGCATC
    CTGCGCATGTACTATCGCGGCGCGTAGACTGTACGACCAGAC
    ACTGACGCGCGTACGTAGCTACTGTAGTACGTACGTACGTA
    ACTGACTGACTGAGATTACAGCTACGTAGTACTGACACGT
    GCATACGAGTCGAAGCGCGCAATATATATTATATCGGCC
     GTGACTACATCATGCTGACTACGGTCGCGCTCAACG
      TGACTACGTACGTATACGACGTACTGACGGCGCGC
      GCACTGTACACGACTGACTTACTAGCTACGTACG
       TGATCAAAACGTACGCGCGGCTATACAGCTA
        CGTACGACTGCGATACGTACGTACGTACG
         CGTGATATAGACCAGATGACACACGT
          TATATATAAAAGCACGGACACT
           GTAGCTAGCTACGGATA
             CTTTTACGTACGT
               ACGACCAG
                 TAGT
```

Supplemental material Chapter 3.1

The role of common lipid-altering gene variants
in the risk of dyslipidemia through old age

Rotterdam Study
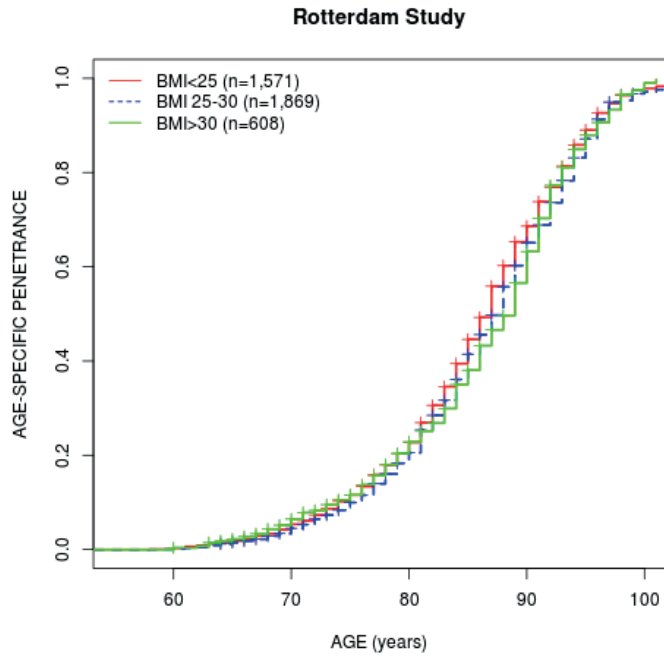
Supplemental Figure 1. Age-specific penetrance of mortality stratified by BMI
BMI: body mass index

TGACTACGTA
CTACTAGCACTGTACACGA   CTACGTACG
CGTACGTAGTCTGATCAAAACGTA   GCTATACAGCTA
CGTTGGTACCAGTACGTACGACTGCGA   GTACGTACGTACGG
GTGTGTCACACAGACACGTGATATAGACCAGATGACACACGTAC
GTACGACGGCGCGGCGCGATATATATAAAAGCACGGACACTACGA
ACGATGACTGACTACGTGCGCGCGTAGCTAGCTACGGATAGGACTA
TACTGACTACGTACTACTAAAGCTGCTTTTACGTACGTACGGGCATC
CATGTACTATCGCGGCGCGTAGACTGTACGACCAGACTGACTGACT
GACTGACTGACGCGCGTACGTAGCTACTGTAGTACGTACGTACGTA
TTCAGACTGACTGACTGAGATTACAGCTACGTAGTACTGACACGT
TACGACGGCGCGGCGCGATATATATAAAAGCACGGACACTACGA
GATGACTGACTACGTGCGCGCGTAGCTAGCTACGGATAGGACTA
TGACTACGTACTACTAAAGCTGCTTTTACGTACGTACGGGCATC
CTGCGCATGTACTATCGCGGCGCGTAGACTGTACGACCAGAC
ACTGACGCGCGTACGTAGCTACTGTAGTACGTACGTACGTA
ACTGACTGACTGAGATTACAGCTACGTAGTACTGACACGT
GCATACGAGTCGAAGCGCGCAATATATATTATATCGGC
GTGACTACATCATGCTGACTACGGTCGCGCTCAACG
TGACTACGTACGTATACGACGTACTGACGGCGCG
GCACTGTACACGACTGACTTACTAGCTACGTACG
TGATCAAAACGTACGCGCGGCTATACAGCTA
CGTACGACTGCGATACGTACGTACGTACG
CGTGATATAGACCAGATGACACACGT
TATATATAAAAGCACGGACACT
GTAGCTAGCTACGGATA
CTTTTACGTACGT
ACGACCAG
TAGT

Supplemental material Chapter 4.1

Risk scores comprised of common lipid-altering genetic variants are associated with lipid levels and suggest an altered role of common genetic variation in type 2 diabetes

Supplemental Table S1. Association of SNP*diabetes interaction with dyslipidemia

lipid	name	locus	OR [95% CI]$_{main}$	P_{main}	OR [95% CI]$_{interact}$	$P_{interact}$	OR$_{interact}$ RS/ERF	T2D related
TC	rs1260326	GCKR	0.91 [0.86 - 0.97]	0.004	0.76 [0.62 - 0.91]	0.004	0.75/0.78	T2D/FG/FI
	rs7239867	LIPG	0.93 [0.86 - 1.01]	0.082	1.40 [1.09 - 1.79]	0.008	1.38/1.53	T2D
	rs2287623	ABCB11	1.05 [0.89 - 1.25]	0.564	1.27 [1.06 - 1.54]	0.012	1.30/1.07	HbA1C
	rs10490626	INSIG2	1.20 [1.06 - 1.36]	0.004	0.72 [0.50 - 1.06]	0.095	0.72/0.73	No
	rs2290159	RAF1	1.05 [0.98 - 1.13]	0.191	1.19 [0.96 - 1.48]	0.119	1.23/0.99	No
	rs11563251	UGT1A1	1.06 [0.95 - 1.17]	0.287	1.26 [0.93 - 1.71]	0.144	1.30/0.96	No
	rs2072183	NPC1L1	0.89 [0.82 - 0.97]	0.007	1.20 [0.93 - 1.56]	0.162	1.21/1.15	No
	rs964184	APOA1-C3-A4-A5	1.38 [1.27 - 1.51]	0.000	1.19 [0.92 - 1.54]	0.178	1.24/0.91	No
	rs7206971	OSBPL7	1.03 [0.97 - 1.09]	0.345	0.88 [0.73 - 1.06]	0.179	0.86/1.07	No
	rs1564348	LPA	0.97 [0.89 - 1.05]	0.481	0.85 [0.66 - 1.08]	0.186	0.84/0.86	No
	rs12027135	LDLRAP1	1.03 [0.97 - 1.10]	0.304	0.89 [0.74 - 1.07]	0.199	0.87/0.96	No
	rs1367117	APOB	0.83 [0.78 - 0.89]	0.000	0.86 [0.69 - 1.08]	0.201	0.83/1.16	No
	rs9488822	FRK	1.00 [0.93 - 1.06]	0.909	1.14 [0.93 - 1.39]	0.209	1.16/1.00	No
	rs4297946	TOP1	0.93 [0.74 - 1.16]	0.534	1.23 [0.87 - 1.73]	0.243	1.10/1.63	FG
	rs3757354	MYLIP	1.14 [0.94 - 1.38]	0.192	0.88 [0.70 - 1.11]	0.275	0.90/0.74	No
	rs4253772	PPARA	0.96 [0.80 - 1.17]	0.711	1.18 [0.88 - 1.58]	0.275	1.17/1.22	No
	rs2255141	GPAM	1.00 [0.85 - 1.18]	0.995	1.30 [0.79 - 2.15]	0.297	1.11/1.94	No
	rs970548	MARCH8-ALOX5	0.98 [0.91 - 1.05]	0.544	1.43 [0.69 - 2.97]	0.333	1.03/2.19	No
	rs1800961	HNF4A	1.06 [0.90 - 1.26]	0.481	0.79 [0.48 - 1.29]	0.343	0.81/0.51	T2D
	rs3177928	HLA	0.96 [0.88 - 1.04]	0.281	1.13 [0.87 - 1.47]	0.344	1.14/1.10	No
	rs4722551	MIR148A	0.83 [0.67 - 1.04]	0.101	1.13 [0.88 - 1.44]	0.346	1.16/0.93	No
	rs12916	HMGCR	0.84 [0.76 - 0.92]	0.000	1.34 [0.72 - 2.47]	0.353	1.03/1.93	No
	rs11220463	ST3GAL4	0.98 [0.89 - 1.08]	0.705	1.15 [0.85 - 1.56]	0.357	1.11/1.38	No
	rs10832963	SPTY2D1	1.07 [1.00 - 1.15]	0.068	1.11 [0.89 - 1.38]	0.368	1.13/0.89	No
	rs11694172	FAM117B	1.05 [0.98 - 1.12]	0.205	0.91 [0.73 - 1.12]	0.370	0.90/0.94	No
	rs651007	ABO	0.89 [0.83 - 0.96]	0.004	1.24 [0.76 - 2.03]	0.385	1.05/1.80	No
	rs2902940	MAFB	0.96 [0.88 - 1.05]	0.380	1.48 [0.61 - 3.62]	0.389	0.98/2.44	No
	rs7941030	UBASH3B	1.06 [0.89 - 1.25]	0.518	1.10 [0.88 - 1.39]	0.405	1.17/0.87	No
	rs629301	SORT1	1.27 [1.08 - 1.50]	0.005	0.58 [0.15 - 2.18]	0.419	1.10/0.28	No
	rs1961456	NAT2	0.94 [0.88 - 1.01]	0.079	1.17 [0.79 - 1.73]	0.425	1.02/1.56	No
	rs1169288	HNF1A	1.06 [1.00 - 1.14]	0.062	0.92 [0.76 - 1.12]	0.430	0.95/0.70	T2D
	rs2000999	HPR	0.88 [0.79 - 0.99]	0.028	0.92 [0.73 - 1.14]	0.435	0.90/1.02	No
	rs2479409	PCSK9	1.15 [1.06 - 1.24]	0.000	1.10 [0.87 - 1.38]	0.437	1.10/1.07	No
	rs1077514	ASAP3	0.96 [0.87 - 1.07]	0.479	1.35 [0.63 - 2.91]	0.438	0.97/2.14	No
	rs10904908	VIM-CUBN	0.94 [0.88 – 1.00]	0.049	1.07 [0.89 - 1.29]	0.462	1.09/0.90	No
	rs2285942	DNAH11	0.94 [0.86 - 1.03]	0.166	1.11 [0.84 - 1.45]	0.464	1.09/1.19	No
	rs11603023	PHLDB1	0.93 [0.82 - 1.06]	0.294	1.07 [0.89 - 1.28]	0.465	1.11/0.84	No
	rs2807834	MOSC1	0.99 [0.86 - 1.14]	0.896	0.93 [0.76 - 1.14]	0.502	0.92/1.05	No
	rs3764261	CETP	0.88 [0.77 – 1.00]	0.048	1.07 [0.88 - 1.30]	0.512	1.10/0.89	No

Supplemental Table S1. (continued)

lipid	name	locus	OR [95% CI]$_{main}$	P_{main}	OR [95% CI]$_{interact}$	$P_{interact}$	OR$_{interact}$ RS/ERF	T2D related
	rs11065987	BRAP	1.03 [0.97 - 1.09]	0.384	1.06 [0.88 - 1.29]	0.522	1.06/1.10	No
	rs6831256	LRPAP1	0.95 [0.89 – 1.00]	0.067	0.94 [0.78 - 1.14]	0.529	0.95/0.90	No
	rs514230	IRF2BP2	1.12 [1.06 - 1.19]	0.000	0.86 [0.52 - 1.40]	0.536	1.05/0.63	No
	rs138777	TOM1	1.08 [0.91 - 1.27]	0.370	0.90 [0.65 - 1.25]	0.541	0.99/0.67	No
	rs9376090	HBS1L	1.01 [0.94 - 1.08]	0.741	1.07 [0.87 - 1.31]	0.546	1.10/0.86	No
	rs2030746	LOC84931	1.00 [0.94 - 1.06]	0.946	0.84 [0.48 - 1.49]	0.554	1.06/0.58	No
	rs2814982	C6orf106	0.97 [0.87 - 1.07]	0.527	1.10 [0.80 - 1.51]	0.564	1.17/0.74	No
	rs4883201	PHC1-A2ML1	0.96 [0.87 - 1.06]	0.404	0.79 [0.35 - 1.77]	0.566	1.12/0.48	No
	rs13315871	PXK	1.05 [0.94 - 1.17]	0.422	1.10 [0.79 - 1.53]	0.571	1.14/0.96	No
	rs314253	DLG4	1.06 [0.99 - 1.13]	0.077	1.13 [0.73 - 1.73]	0.592	0.96/1.53	No
	rs4420638	APOE-C1-C2	1.49 [1.26 - 1.76]	0.000	0.74 [0.23 - 2.42]	0.618	1.27/0.38	No
	rs2126259	PPP1R3B	1.11 [0.96 - 1.28]	0.160	0.82 [0.37 - 1.83]	0.632	1.08/0.45	FG/FI/2hr glu
	rs2277862	ERGIC3	1.08 [0.96 - 1.22]	0.205	0.85 [0.42 - 1.72]	0.658	1.14/0.55	No
	rs4530754	CSNK1G3	0.99 [0.94 - 1.05]	0.784	1.09 [0.75 - 1.57]	0.663	1.24/0.82	No
	rs7640978	CMTM6	1.07 [0.95 - 1.20]	0.294	1.09 [0.74 - 1.61]	0.673	1.19/0.68	No
	rs6882076	TIMD4	0.88 [0.83 - 0.94]	0.000	1.05 [0.83 - 1.32]	0.684	1.00/1.38	No
	rs492602	FLJ36070	1.04 [0.98 - 1.10]	0.248	1.04 [0.86 - 1.26]	0.692	1.03/1.10	No
	rs581080	TTC39B	1.12 [0.77 - 1.63]	0.562	1.08 [0.72 - 1.62]	0.720	1.21/0.75	No
	rs3780181	VLDLR	1.03 [0.91 - 1.17]	0.655	1.07 [0.73 - 1.57]	0.736	1.06/1.19	No
	rs1997243	GPR146	0.95 [0.88 - 1.03]	0.234	0.95 [0.68 - 1.31]	0.744	0.88/1.35	No
	rs1030431	CYP7A1	1.10 [1.03 - 1.17]	0.002	0.97 [0.80 - 1.18]	0.793	0.96/1.12	No
	rs10401969	CILP2	1.29 [1.01 - 1.66]	0.044	1.10 [0.54 - 2.23]	0.801	1.35/0.59	T2D
	rs10102164	SOX17	0.91 [0.84 - 0.98]	0.019	1.04 [0.78 - 1.37]	0.809	0.97/1.40	No
	rs11136341	PLEC1	0.89 [0.77 - 1.02]	0.098	1.03 [0.83 - 1.26]	0.816	1.05/0.87	No
	rs174550	FADS1-2-3	0.99 [0.93 - 1.05]	0.695	1.02 [0.84 - 1.24]	0.823	1.00/1.20	No
	rs2954022	TRIB1	1.14 [1.07 - 1.21]	0.000	1.02 [0.85 - 1.22]	0.848	1.03/0.96	No
	rs1532085	LIPC	1.04 [0.94 - 1.15]	0.417	0.96 [0.63 - 1.48]	0.856	0.82/1.32	No
	rs3850634	ANGPTL3	1.09 [1.02 - 1.16]	0.007	1.02 [0.75 - 1.39]	0.877	0.94/1.37	No
	rs1800562	HFE	1.08 [0.95 - 1.22]	0.239	0.98 [0.68 - 1.40]	0.901	0.96/1.08	HbA1C
	rs2758886	KCNK17	1.00 [0.86 - 1.16]	0.979	1.01 [0.82 - 1.25]	0.911	0.99/1.26	No
	rs1883025	ABCA1	1.03 [0.96 - 1.11]	0.389	1.01 [0.81 - 1.27]	0.912	1.04/0.84	No
	rs2737229	TRPS1	1.10 [0.96 - 1.27]	0.166	1.01 [0.82 - 1.23]	0.963	0.98/1.22	FPI*
	rs4299376	ABCG5/8	0.90 [0.84 - 0.96]	0.001	1.01 [0.83 - 1.22]	0.964	1.03/0.83	No
	rs7515577	EVI5	1.04 [0.97 - 1.12]	0.299	0.99 [0.65 - 1.52]	0.968	1.13/0.70	No
	rs6759321	RAB3GAP1	0.98 [0.91 - 1.04]	0.475	1.00 [0.82 - 1.22]	0.984	1.02/0.89	No
	rs6511720	LDLR	1.35 [1.23 - 1.48]	0.000	1.00 [0.50 - 2.02]	0.994	1.35/0.65	No
LDL	rs2332328	NYNRIN	1.00 [0.94 - 1.06]	0.979	1.25 [1.04 - 1.5]	0.016	1.24/1.31	No
	rs11153594	FRK	0.98 [0.92 - 1.04]	0.483	1.23 [1.01 - 1.49]	0.038	1.26/0.96	No

Supplemental Table S1. (continued)

lipid	name	locus	OR [95% CI]$_{main}$	P_{main}	OR [95% CI]$_{interact}$	$P_{interact}$	OR$_{interact}$	RS/ERF	T2D related
	rs10490626	INSIG2	1.20 [1.06 - 1.36]	0.004	0.72 [0.50 - 1.06]	0.095	0.72/0.73		No
	rs1250229	FN1	0.96 [0.89 - 1.03]	0.221	1.19 [0.97 - 1.46]	0.098	1.17/1.32		FPI*
	rs2328223	SNX5	0.96 [0.88 - 1.03]	0.253	1.22 [0.96 - 1.54]	0.109	1.19/1.43		No
	rs11563251	UGT1A1	1.06 [0.95 - 1.17]	0.287	1.26 [0.93 - 1.71]	0.144	1.30/0.96		No
	rs964184	APOA1-C3-A4-A5	1.38 [1.27 - 1.51]	0.000	1.19 [0.92 - 1.54]	0.178	1.24/0.91		No
	rs1564348	LPA	0.97 [0.89 - 1.05]	0.481	0.85 [0.66 - 1.08]	0.186	0.84/0.86		No
	rs12027135	LDLRAP1	1.03 [0.97 - 1.10]	0.304	0.89 [0.74 - 1.07]	0.199	0.87/0.96		No
	rs1367117	APOB	0.83 [0.78 - 0.89]	0.000	0.86 [0.69 - 1.08]	0.201	0.83/1.16		No
	rs2710642	EHBP1	0.93 [0.81 - 1.07]	0.296	1.26 [0.86 - 1.84]	0.232	1.11/1.71		No
	rs3757354	MYLIP	1.14 [0.94 - 1.38]	0.192	0.88 [0.70 - 1.11]	0.275	0.90/0.74		No
	rs4253772	PPARA	0.96 [0.80 - 1.17]	0.711	1.18 [0.88 - 1.58]	0.275	1.17/1.22		No
	rs1129555	GPAM	1.00 [0.84 - 1.19]	0.980	1.28 [0.81 - 2.05]	0.293	1.12/1.92		No
	rs909802	TOP1	0.94 [0.76 - 1.18]	0.610	1.16 [0.85 - 1.59]	0.336	1.06/1.53		FG
	rs3177928	HLA	0.96 [0.88 - 1.04]	0.281	1.13 [0.87 - 1.47]	0.344	1.14/1.10		No
	rs4722551	MIR148A	0.83 [0.67 - 1.04]	0.101	1.13 [0.88 - 1.44]	0.346	1.16/0.93		No
	rs12916	HMGCR	0.84 [0.76 - 0.92]	0.000	1.34 [0.72 - 2.47]	0.353	1.03/1.93		No
	rs11220462	ST3GAL4	0.99 [0.91 - 1.08]	0.832	1.12 [0.87 - 1.43]	0.384	1.08/1.32		No
	rs629301	SORT1	1.27 [1.08 - 1.50]	0.005	0.58 [0.15 - 2.18]	0.419	1.10/0.28		No
	rs1169288	HNF1A	1.06 [1.00 - 1.14]	0.062	0.92 [0.76 - 1.12]	0.430	0.95/0.70		T2D
	rs2000999	HPR	0.88 [0.79 - 0.99]	0.028	0.92 [0.73 - 1.14]	0.435	0.90/1.02		No
	rs2479409	PCSK9	1.15 [1.06 - 1.24]	0.000	1.10 [0.87 - 1.38]	0.437	1.10/1.07		No
	rs17404153	ACAD11	1.12 [0.95 - 1.33]	0.171	0.70 [0.28 - 1.77]	0.453	1.05/0.40		No
	rs649129	ABO	0.90 [0.84 - 0.97]	0.004	1.21 [0.74 - 1.97]	0.456	1.01/1.73		No
	rs2902941	MAFB	0.98 [0.90 - 1.07]	0.620	1.25 [0.68 - 2.29]	0.464	0.97/1.81		No
	rs2807834	MOSC1	0.99 [0.86 - 1.14]	0.896	0.93 [0.76 - 1.14]	0.502	0.92/1.05		No
	rs11065987	BRAP	1.03 [0.97 - 1.09]	0.384	1.06 [0.88 - 1.29]	0.522	1.06/1.10		No
	rs6831256	LRPAP1	0.95 [0.89 – 1.00]	0.067	0.94 [0.78 - 1.14]	0.529	0.95/0.90		No
	rs247616	CETP	0.87 [0.76 – 1.00]	0.056	1.07 [0.87 - 1.30]	0.533	1.09/0.89		No
	rs12748152	PIGV-NROB2	0.84 [0.64 - 1.09]	0.197	0.90 [0.65 - 1.25]	0.533	0.88/1.01		No
	rs514230	IRF2BP2	1.12 [1.06 - 1.19]	0.000	0.86 [0.52 - 1.40]	0.536	1.05/0.63		No
	rs217386	NPC1L1	0.89 [0.84 - 0.95]	0.000	0.94 [0.78 - 1.14]	0.545	0.96/0.83		No
	rs2030746	LOC84931	1.00 [0.94 - 1.06]	0.946	0.84 [0.48 - 1.49]	0.554	1.06/0.58		No
	rs314253	DLG4	1.06 [0.99 - 1.13]	0.077	1.13 [0.73 - 1.73]	0.592	0.96/1.53		No
	rs267733	ANXA9-CERS2	0.96 [0.89 - 1.04]	0.360	1.15 [0.67 - 1.99]	0.616	1.41/0.78		No
	rs4420638	APOE-C1-C2	1.49 [1.26 - 1.76]	0.000	0.74 [0.23 - 2.42]	0.618	1.27/0.38		No
	rs2126259	PPP1R3B	1.11 [0.96 - 1.28]	0.160	0.82 [0.37 - 1.83]	0.632	1.08/0.45		FG/FI/2hr glu
	rs4530754	CSNK1G3	0.99 [0.94 - 1.05]	0.784	1.09 [0.75 - 1.57]	0.663	1.24/0.82		No
	rs7640978	CMTM6	1.07 [0.95 - 1.20]	0.294	1.09 [0.74 - 1.61]	0.673	1.19/0.68		No

Supplemental Table S1. (continued)

lipid	name	locus	OR [95% CI]$_{main}$	P_{main}	OR [95% CI]$_{interact}$	$P_{interact}$	OR$_{interact}$ RS/ERF	T2D related
	rs6882076	TIMD4	0.88 [0.83 - 0.94]	0.000	1.05 [0.83 - 1.32]	0.684	1.00/1.38	No
	rs364585	SPTLC3	1.03 [0.97 - 1.10]	0.333	0.90 [0.55 - 1.49]	0.689	1.10/0.65	No
	rs1801689	APOH-PRXCA	0.75 [0.46 - 1.22]	0.243	0.72 [0.12 - 4.48]	0.723	1.15/0.11	No
	rs3780181	VLDLR	1.03 [0.91 - 1.17]	0.655	1.07 [0.73 - 1.57]	0.736	1.06/1.19	No
	rs7225700	OSBPL7	0.97 [0.88 - 1.06]	0.476	1.03 [0.85 - 1.24]	0.765	1.00/1.37	No
	rs1030431	CYP7A1	1.10 [1.03 - 1.17]	0.002	0.97 [0.80 - 1.18]	0.793	0.96/1.12	No
	rs10401969	CILP2	1.29 [1.01 - 1.66]	0.044	1.10 [0.54 - 2.23]	0.801	1.35/0.59	T2D
	rs10102164	SOX17	0.91 [0.84 - 0.98]	0.019	1.04 [0.78 - 1.37]	0.809	0.97/1.40	No
	rs11136341	PLEC1	0.89 [0.77 - 1.02]	0.098	1.03 [0.83 - 1.26]	0.816	1.05/0.87	No
	rs2954022	TRIB1	1.14 [1.07 - 1.21]	0.000	1.02 [0.85 - 1.22]	0.848	1.03/0.96	No
	rs4942486	BRCA2	1.06 [1.00 - 1.13]	0.051	0.98 [0.82 - 1.18]	0.866	0.97/1.07	No
	rs174583	FADS1–2–3	0.99 [0.93 - 1.05]	0.736	1.02 [0.84 - 1.23]	0.876	1.00/1.18	No
	rs3850634	ANGPTL3	1.09 [1.02 - 1.16]	0.007	1.02 [0.75 - 1.39]	0.877	0.94/1.37	No
	rs5763662	MTMR3	1.07 [0.85 - 1.34]	0.590	1.06 [0.47 - 2.38]	0.894	1.05/0.83	No
	rs1800562	HFE	1.08 [0.95 - 1.22]	0.239	0.98 [0.68 - 1.40]	0.901	0.96/1.08	HbA1C
	rs12670798	DNAH11	0.94 [0.88 - 1.01]	0.101	1.01 [0.81 - 1.25]	0.939	1.05/0.76	No
	rs4299376	ABCG5/8	0.90 [0.84 - 0.96]	0.001	1.01 [0.83 - 1.22]	0.964	1.03/0.83	No
	rs6511720	LDLR	1.35 [1.23 - 1.48]	0.000	1.00 [0.5 - 2.02]	0.994	1.35/0.65	No
HDL	rs7241918	LIPG	0.93 [0.86 - 1.01]	0.086	1.38 [1.08 - 1.77]	0.010	1.37/1.53	T2D
	rs1042034	APOB	1.19 [1.11 - 1.29]	0.000	0.79 [0.63 - 0.99]	0.036	0.78/0.84	No
	rs7255436	ANGPTL4	0.98 [0.92 - 1.04]	0.421	1.23 [0.97 - 1.57]	0.083	1.16/1.58	No
	rs643531	TTC39B	1.13 [0.76 - 1.67]	0.546	1.23 [0.95 - 1.59]	0.115	1.28/0.85	No
	rs737337	LOC55908	0.99 [0.88 - 1.12]	0.923	0.65 [0.36 - 1.15]	0.136	0.80/0.43	No
	rs11246602	OR4C46	1.11 [1.02 - 1.21]	0.017	0.84 [0.64 - 1.09]	0.176	0.82/1.03	No
	rs964184	APOA1-C3-A4-A5	1.38 [1.27 - 1.51]	0.000	1.19 [0.92 - 1.54]	0.178	1.24/0.91	No
	rs4082919	PGS1	0.95 [0.89 - 1.00]	0.071	1.13 [0.94 - 1.36]	0.194	1.14/1.07	No
	rs2925979	CMIP	1.12 [1.05 - 1.19]	0.000	0.89 [0.73 - 1.08]	0.220	0.92/0.69	No
	rs12328675	COBLL1	0.93 [0.83 - 1.04]	0.200	1.18 [0.90 - 1.54]	0.237	1.16/1.26	FI
	rs13326165	STAB1	1.09 [1.01 - 1.18]	0.023	0.87 [0.69 - 1.10]	0.248	0.88/0.82	No
	rs10019888	C4orf52	0.96 [0.88 - 1.04]	0.307	0.86 [0.67 - 1.11]	0.251	0.83/1.16	No
	rs17145738	MLXIPL	1.04 [0.87 - 1.25]	0.641	1.18 [0.89 - 1.55]	0.251	1.20/0.99	No
	rs4846914	GALNT2	1.13 [0.87 - 1.47]	0.371	0.90 [0.75 - 1.09]	0.270	0.91/0.76	No
	rs499974	MOGAT2-DGAT2	0.99 [0.86 - 1.13]	0.870	1.14 [0.90 - 1.45]	0.285	1.14/1.11	No
	rs3136441	LRP4	0.95 [0.78 - 1.17]	0.640	0.87 [0.67 - 1.13]	0.287	0.88/0.73	No
	rs386000	LILRA3	0.96 [0.88 - 1.04]	0.335	0.87 [0.67 - 1.12]	0.290	0.89/0.68	No
	rs1084651	LPA	1.06 [0.97 - 1.15]	0.212	1.14 [0.89 - 1.47]	0.310	1.12/1.27	No
	rs7134375	PDE3A	1.04 [0.98 - 1.10]	0.219	0.71 [0.36 - 1.39]	0.319	0.96/0.48	No
	rs2602836	ADH5	0.98 [0.89 - 1.06]	0.567	1.09 [0.91 - 1.31]	0.326	1.13/0.90	No
	rs970548	MARCH8-ALOX5	0.98 [0.91 - 1.05]	0.544	1.43 [0.69 - 2.97]	0.333	1.03/2.19	No

Supplemental Table S1. (continued)

lipid name	locus	OR [95% CI]$_{main}$	P_{main}	OR [95% CI]$_{interact}$	$P_{interact}$	OR$_{interact}$	RS/ERF	T2D related
rs1800961	HNF4A	1.06 [0.90 - 1.26]	0.481	0.79 [0.48 - 1.29]	0.343	0.81/0.51		T2D
rs1515100	IRS1	1.00 [0.87 - 1.14]	0.953	1.40 [0.70 - 2.81]	0.343	1.04/2.13		T2D/FI
rs2293889	TRPS1	1.02 [0.96 - 1.09]	0.497	1.09 [0.91 - 1.32]	0.343	1.06/1.29		FPI*
rs12801636	KAT5	1.02 [0.95 - 1.09]	0.630	0.89 [0.71 - 1.13]	0.345	0.90/0.78		No
rs702485	DAGLB	1.01 [0.90 - 1.12]	0.894	1.09 [0.91 - 1.31]	0.361	1.07/1.24		No
rs2606736	ATG7	1.08 [1.02 - 1.15]	0.011	0.92 [0.76 - 1.12]	0.414	0.90/1.10		No
rs2814944	C6orf106	0.97 [0.89 - 1.05]	0.407	1.15 [0.81 - 1.63]	0.422	1.25/0.80		No
rs17404153	ACAD11	1.12 [0.95 - 1.33]	0.171	0.70 [0.28 - 1.77]	0.453	1.05/0.40		No
rs4660293	PABPC4	1.13 [0.91 - 1.42]	0.269	0.86 [0.57 - 1.31]	0.490	1.00/0.63		No
rs4650994	ANGPTL1	1.01 [0.95 - 1.07]	0.760	1.07 [0.89 - 1.27]	0.491	1.05/1.18		No
rs10808546	TRIB1	1.13 [1.06 - 1.19]	0.000	1.07 [0.89 - 1.28]	0.499	1.08/0.97		No
rs2923084	AMPD3	1.13 [0.98 - 1.30]	0.078	0.93 [0.74 - 1.16]	0.507	0.94/0.79		No
rs731839	PEPD	0.99 [0.93 - 1.06]	0.828	1.15 [0.76 - 1.76]	0.507	0.99/1.56		T2D/FI
rs3764261	CETP	0.88 [0.77 – 1.00]	0.048	1.07 [0.88 - 1.30]	0.512	1.10/0.89		No
rs2652834	LACTB	1.00 [0.92 - 1.08]	0.926	0.93 [0.73 - 1.17]	0.530	0.95/0.78		No
rs12748152	PIGV-NROB2	0.84 [0.64 - 1.09]	0.197	0.90 [0.65 - 1.25]	0.533	0.88/1.01		No
rs4765127	ZNF664	1.04 [0.97 - 1.10]	0.294	1.07 [0.87 - 1.30]	0.541	1.06/1.13		No
rs1936800	RSPO3	0.94 [0.83 - 1.06]	0.288	1.10 [0.80 - 1.53]	0.553	1.00/1.44		FI
rs838880	SCARB1	0.97 [0.91 - 1.03]	0.292	1.06 [0.87 - 1.30]	0.554	1.05/1.13		No
rs4917014	IKZF1	0.97 [0.89 - 1.07]	0.560	1.06 [0.87 - 1.29]	0.565	1.09/0.89		No
rs3741414	LRP1	1.01 [0.88 - 1.16]	0.904	1.18 [0.66 - 2.13]	0.579	0.93/1.73		No
rs7115089	UBASH3B	1.05 [0.89 - 1.24]	0.536	1.08 [0.81 - 1.45]	0.588	1.18/0.84		No
rs1121980	FTO	1.11 [0.99 - 1.24]	0.085	0.95 [0.79 - 1.14]	0.593	0.93/1.11		T2D/FI
rs4731702	KLF14	1.08 [0.98 - 1.19]	0.106	1.05 [0.87 - 1.26]	0.610	1.03/1.22		T2D
rs998584	VEGFA	0.92 [0.86 - 0.98]	0.013	1.05 [0.86 - 1.29]	0.611	1.05/1.08		No
rs4420638	APOE-C1-C2	1.49 [1.26 - 1.76]	0.000	0.74 [0.23 - 2.42]	0.618	1.27/0.38		No
rs181362	UBE2L3	1.05 [0.98 - 1.14]	0.165	1.06 [0.84 - 1.34]	0.639	1.05/1.10		No
rs4759375	SBNO1	0.94 [0.81 - 1.09]	0.409	1.09 [0.75 - 1.59]	0.647	1.02/1.65		No
rs9987289	PPP1R3B	1.07 [0.88 - 1.29]	0.499	0.76 [0.24 - 2.46]	0.651	1.19/0.34		FG/FI/2hr glu
rs4148008	ABCA8	1.03 [0.97 - 1.10]	0.350	0.96 [0.79 - 1.16]	0.666	0.96/0.92		No
rs6065906	PLTP	0.93 [0.87 - 1.01]	0.069	0.96 [0.76 - 1.21]	0.728	0.98/0.84		No
rs12145743	HDGF-PMVK	1.04 [0.96 - 1.12]	0.347	1.03 [0.85 - 1.25]	0.755	1.02/1.11		No
rs4142995	SNX13	0.93 [0.80 - 1.07]	0.305	1.03 [0.85 - 1.24]	0.769	0.99/1.34		No
rs881844	STARD3	0.97 [0.91 - 1.03]	0.346	0.97 [0.80 - 1.18]	0.780	0.95/1.15		No
rs1689800	ZNF648	0.92 [0.86 - 0.98]	0.009	0.98 [0.80 - 1.18]	0.800	0.97/0.99		No
rs2013208	RBM5	1.04 [0.98 - 1.10]	0.198	0.98 [0.81 - 1.18]	0.801	0.97/1.02		No
rs6450176	ARL15	0.92 [0.86 - 0.99]	0.025	1.02 [0.83 - 1.26]	0.836	1.02/1.01		FI
rs1532085	LIPC	1.04 [0.94 - 1.15]	0.417	0.96 [0.63 - 1.48]	0.856	0.82/1.32		No

Supplemental Table S1. (continued)

lipid	name	locus	OR [95% CI]$_{main}$	P_{main}	OR [95% CI]$_{interact}$	$P_{interact}$	OR$_{interact}$ RS/ERF	T2D related
	rs16942887	LCAT	1.12 [1.03 - 1.23]	0.014	0.98 [0.74 - 1.28]	0.872	0.96/1.13	No
	rs174601	FADS1-2-3	0.98 [0.92 - 1.05]	0.633	1.02 [0.82 - 1.26]	0.885	1.00/1.20	No
	rs13107325	SLC39A8	1.06 [0.88 - 1.27]	0.564	1.04 [0.54 - 2.02]	0.896	1.36/0.68	No
	rs605066	CITED2	1.06 [1.00 - 1.13]	0.051	0.99 [0.82 - 1.19]	0.901	1.00/0.94	No
	rs1883025	ABCA1	1.03 [0.96 - 1.11]	0.389	1.01 [0.81 - 1.27]	0.912	1.04/0.84	No
	rs2290547	SETD2	1.00 [0.92 - 1.08]	0.978	0.99 [0.78 - 1.26]	0.925	1.02/0.75	No
	rs17695224	HAS1	0.94 [0.87 – 1.00]	0.052	0.99 [0.81 - 1.21]	0.927	1.02/0.78	No
	rs12967135	MC4R	1.03 [0.94 - 1.12]	0.579	1.01 [0.82 - 1.24]	0.927	1.03/0.88	T2D
	rs4983559	ZBTB42-AKT1	0.93 [0.88 - 0.99]	0.022	1.02 [0.66 - 1.58]	0.929	1.20/0.75	No
	rs6805251	GSK3B	0.99 [0.93 - 1.05]	0.717	0.99 [0.82 - 1.20]	0.936	0.99/1.03	No
	rs17173637	TMEM176A	0.90 [0.81 - 0.99]	0.039	1.01 [0.74 - 1.38]	0.956	1.00/1.09	No
	rs7134594	MVK	1.00 [0.94 - 1.06]	0.896	1.00 [0.83 - 1.21]	0.963	0.98/1.20	No
	rs3822072	FAM13A	1.03 [0.97 - 1.10]	0.322	1.00 [0.84 - 1.21]	0.966	1.04/0.80	No
	rs12678919	LPL	1.16 [1.05 - 1.28]	0.003	1.00 [0.66 - 1.52]	0.999	1.14/0.71	No
TG	rs1260326	GCKR	0.91 [0.86 - 0.97]	0.004	0.76 [0.62 - 0.91]	0.004	0.75/0.78	T2D/FG/FI
	rs11649653	CTF1	0.96 [0.90 - 1.03]	0.279	1.26 [1.03 - 1.55]	0.023	1.26/1.30	No
	rs1042034	APOB	1.19 [1.11 - 1.29]	0.000	0.79 [0.63 - 0.99]	0.036	0.78/0.84	No
	rs9686661	MAP3K1	0.99 [0.75 - 1.29]	0.915	0.83 [0.66 - 1.05]	0.118	0.85/0.71	FI
	rs964184	APOA1-C3-A4-A5	1.38 [1.27 - 1.51]	0.000	1.19 [0.92 - 1.54]	0.178	1.24/0.91	No
	rs261342	LIPC	1.05 [0.98 - 1.14]	0.159	0.79 [0.54 - 1.15]	0.212	0.70/1.09	No
	rs10195252	COBLL1	0.90 [0.75 - 1.08]	0.268	1.12 [0.93 - 1.35]	0.213	1.10/1.26	FI
	rs1321257	GALNT2	1.13 [0.88 - 1.44]	0.348	0.89 [0.74 - 1.08]	0.232	0.91/0.76	No
	rs7248104	INSR	1.00 [0.94 - 1.06]	0.874	1.12 [0.93 - 1.34]	0.242	1.14/0.99	No
	rs1832007	AKR1C4	0.93 [0.85 - 1.01]	0.072	1.15 [0.90 - 1.48]	0.267	1.11/1.54	No
	rs5756931	PLA2G6	1.01 [0.95 - 1.08]	0.735	1.11 [0.92 - 1.34]	0.277	1.10/1.21	No
	rs38855	MET	1.04 [0.98 - 1.10]	0.221	0.91 [0.76 - 1.10]	0.328	0.89/1.11	No
	rs2943645	IRS1	0.99 [0.86 - 1.14]	0.920	1.46 [0.67 - 3.18]	0.342	1.03/2.30	T2D/FI
	rs4722551	MIR148A	0.83 [0.67 - 1.04]	0.101	1.13 [0.88 - 1.44]	0.346	1.16/0.93	No
	rs442177	KLHL8	0.94 [0.89 – 1.00]	0.050	1.09 [0.90 - 1.31]	0.366	1.08/1.17	No
	rs10761731	JMJD1C	0.95 [0.81 - 1.12]	0.541	1.15 [0.85 - 1.54]	0.367	1.05/1.49	No
	rs11776767	PINX1	1.01 [0.92 - 1.11]	0.818	0.88 [0.66 - 1.17]	0.380	0.95/0.67	No
	rs1495743	NAT2	0.94 [0.88 - 1.01]	0.117	1.10 [0.88 - 1.37]	0.383	1.06/1.45	No
	rs13238203	TYW1B	1.17 [0.87 - 1.57]	0.303	1.27 [0.72 - 2.26]	0.412	1.32/1.13	No
	rs439401	APOE-C1-C2	1.12 [0.97 - 1.29]	0.126	0.82 [0.50 - 1.34]	0.432	1.00/0.59	No
	rs7811265	MLXIPL	1.03 [0.87 - 1.21]	0.763	1.08 [0.86 - 1.36]	0.483	1.13/0.80	No
	rs731839	PEPD	0.99 [0.93 - 1.06]	0.828	1.15 [0.76 - 1.76]	0.507	0.99/1.56	T2D/FI
	rs2068888	CYP26A1	0.98 [0.93 - 1.04]	0.527	0.94 [0.78 - 1.13]	0.509	0.94/0.94	No
	rs2929282	FRMD5	1.13 [0.99 - 1.30]	0.074	1.38 [0.53 - 3.62]	0.513	0.93/2.56	No
	rs8077889	MPP3	0.93 [0.86 – 1.00]	0.063	1.18 [0.72 - 1.93]	0.521	0.98/1.67	No

Supplemental Table S1. (continued)

lipid	name	locus	OR [95% CI]$_{main}$	P_{main}	OR [95% CI]$_{interact}$	$P_{interact}$	OR$_{interact}$ RS/ERF	T2D related
	rs4810479	PLTP	0.95 [0.89 - 1.01]	0.106	0.93 [0.76 - 1.15]	0.528	0.94/0.87	No
	rs6831256	LRPAP1	0.95 [0.89 – 1.00]	0.067	0.94 [0.78 - 1.14]	0.529	0.95/0.90	No
	rs12748152	PIGV-NROB2	0.84 [0.64 - 1.09]	0.197	0.90 [0.65 - 1.25]	0.533	0.88/1.01	No
	rs1936800	RSPO3	0.94 [0.83 - 1.06]	0.288	1.10 [0.80 - 1.53]	0.553	1.00/1.44	FI
	rs1121980	FTO	1.11 [0.99 - 1.24]	0.085	0.95 [0.79 - 1.14]	0.593	0.93/1.11	T2D/FI
	rs7205804	CETP	0.83 [0.79 - 0.89]	0.000	0.93 [0.72 - 1.21]	0.599	1.00/0.72	No
	rs11613352	LRP1	1.01 [0.89 - 1.15]	0.876	1.16 [0.67 - 2.01]	0.604	0.93/1.66	No
	rs998584	VEGFA	0.92 [0.86 - 0.98]	0.013	1.05 [0.86 - 1.29]	0.611	1.05/1.08	No
	rs12310367	ZNF664	1.04 [0.97 - 1.11]	0.279	1.05 [0.86 - 1.30]	0.615	1.04/1.14	No
	rs1553318	TIMD4	0.87 [0.82 - 0.93]	0.000	1.13 [0.68 - 1.88]	0.641	0.92/1.59	No
	rs2131925	ANGPTL3	1.09 [1.02 - 1.16]	0.007	1.08 [0.74 - 1.57]	0.699	0.95/1.46	No
	rs3198697	PDXDC1	0.99 [0.93 - 1.05]	0.652	0.95 [0.71 - 1.28]	0.755	1.03/0.70	No
	rs2954029	TRIB1	1.13 [1.03 - 1.24]	0.009	1.02 [0.86 - 1.23]	0.794	1.03/0.99	No
	rs10401969	CILP2	1.29 [1.01 - 1.66]	0.044	1.10 [0.54 - 2.23]	0.801	1.35/0.59	T2D
	rs645040	MSL2L1	1.02 [0.95 - 1.09]	0.655	0.97 [0.78 - 1.21]	0.814	0.94/1.26	No
	rs174546	FADS1-2-3	0.99 [0.93 - 1.05]	0.702	1.02 [0.84 - 1.24]	0.824	1.00/1.21	No
	rs2247056	HLA	0.98 [0.80 - 1.19]	0.803	1.02 [0.83 - 1.25]	0.878	1.01/1.09	No
	rs2412710	CAPN3	1.31 [1.07 - 1.61]	0.008	1.04 [0.53 - 2.04]	0.920	1.07/0.92	No
	rs12678919	LPL	1.16 [1.05 - 1.28]	0.003	1.00 [0.66 - 1.52]	0.999	1.14/0.71	No

main: SNP main effect, interact: SNP*diabetes interaction effect, ERF: Erasmus Rucphen Family Study, FPI: fasting pro-insulin

*suggestive

Chapter 8

Epilogue

CTGACTACGTA
CTACTAGCACTGTACACGA   CTACGTACG
CGTACGTAGTCTGATCAAAACGTA   GCTATACAGCTA
CGTTGGTACCAGTACGTACGACTGCGA GTACGTACGTACGG
GTGTGTCACACAGACACGTGATATAGACCAGATGACACACGTAC
GTACGACGGCGCGGCGCGATATATATAAAAGCACGGACACTACGA
ACGATGACTGACTACGTGCGCGCGTAGCTAGCTACGGATAGGACTA
TACTGACTACGTACTACTAAAGCTGCTTTTACGTACGTACGGGCATC
CATGTACTATCGCGGCGCGTAGACTGTACGACCAGACTGACTGACT
GACTGACTGACGCGCGTACGTAGCTACTGTAGTACGTACGTACGTA
TTCAGACTGACTGACTGAGATTACAGCTACGTAGTACTGACACGT
TACGACGGCGCGGCGCGATATATATAAAAGCACGGACACTACGA
GATGACTGACTACGTGCGCGCGTAGCTAGCTACGGATAGGACTA
TGACTACGTACTACTAAAGCTGCTTTTACGTACGTACGGGCATC
CTGCGCATGTACTATCGCGGCGCGTAGACTGTACGACCAGAC
ACTGACGCGCGTACGTAGCTACTGTAGTACGTACGTACGTA
ACTGACTGACTGAGATTACAGCTACGTAGTACTGACACGT
GCATACGAGTCGAAGCGCGCAATATATATTATATCGGCC
CGTGACTACATCATGCTGACTACGGTCGCGCTCAACG
TGACTACGTACGTATACGACGTACTGACGGCGCGC
GCACTGTACACGACTGACTTACTAGCTACGTACG
TGATCAAAACGTACGCGCGGCTATACAGCTA
CGTACGACTGCGATACGTACGTACGTACG
CGTGATATAGACCAGATGACACACGT
TATATATAAAAGCACGGACACT
GTAGCTAGCTACGGATA
CTTTTACGTACGT
ACGACCAG
TAGT

Dankwoord/Acknowledgments

Ik kan het bijna niet geloven, maar het zit er dan toch echt bijna op. Het waren intensieve jaren met bloed, zweet en tranen, maar ook met de meest fantastische momenten en ervaringen, en bovenal heb ik de afgelopen jaren enorm veel geleerd. Ik ben velen dank verschuldigd voor de lessen, de hulp, de mooie kansen, de steun en de mooie en fijne momenten de afgelopen jaren.

Om te beginnen wil ik mijn promotor Prof. Cornelia van Duijn bedanken. Cock, bedankt voor alle geweldige kansen die je me hebt gegeven, zowel op het gebied van de data die we in Rotterdam tot onze beschikking hebben als de samenwerkingsverbanden met de consortia, met als hoogtepunt de drie maanden die ik in Boston heb mogen doorbrengen. Dit was een fantastische ervaring die ik niet had willen missen!

Dear Aaron, very learned copromotor, thank you very much for all your help and input in my thesis and for everything you learned me over the past years. And for enabling a great start of my PhD training with an oral presentation at ESHG, young investigator award nomination and press release. These have been great experiences and I know this would not have happened without your help!

Ook wil ik Prof. Eric Sijbrands, Prof. Oscar Franco en Prof. Nick Wareham, de leden van de kleine commissie, heel hartelijk danken. Eric en Oscar, heel hartelijk dank voor al jullie waardevolle input in een groot deel van de papers in dit proefschrift en ik waardeer het zeer dat jullie bereid waren plaats te nemen in de kleine commissie en mijn proefschrift zo snel te beoordelen. Dear Nick, thank you very much that you are willing to be in the reading committee and for coming to Rotterdam for the defence, I really appreciate that. I'm looking forward to the coming years of my fellowship, now with the PhD title!

Mijn dank gaat ook uit naar Prof. Ben Oostra en Prof. Ko Willems van Dijk, leden van de grote commissie. Ben, heel hartelijk dank dat je ondanks je pensionering toch plaats-neemt in de commissie, ik waardeer dat enorm. Prof. Ko Willems van Dijk, niet zo lang geleden hadden we elkaar gezien in Cambridge toen u daar was om een seminar te geven. U reageerde direct positief toen ik vertelde dat de vraag eraan kon gaan komen of u deel wilde uitmaken van de grote commissie en u gaf me meteen wat nuttige tips mee voor de afronding van het proefschrift, waar ik u ook nog hartelijk voor wil danken.

Prof. Janssens, Cecile, ook jou wil ik heel hartelijk danken. Je hulp en begeleiding bij het schrijven van mijn eerste paper voor dit proefschrift waren heel fijn en inspirerend. De goede schrijfadviezen zitten in mijn hoofd gebakken!

Prof. Albert Hofman, ik heb er veel bewondering voor hoe je de Rotterdam Studie hebt opgezet en vele onderzoekers de kans geeft te leren en bij te dragen aan belangrijke kennis over ziekte en gezondheid. Ik ben dankbaar dat ik een van deze onderzoekers ben! Hartelijk dank ook voor de onvergetelijke colleges in het NIHES zomerprogramma.

Verder ben ik alle deelnemers aan het Erasmus Rucphen Familieonderzoek en de Rotterdam Studie veel dank verschuldigd en iedereen die betrokken is bij het verzamelen van de data, de verwerking en de verdere ondersteuning van het onderzoek. In het bijzonder wil ik noemen Elza, Anneke, Frank, Andrea, Jeannette, Andy, Bernadette en Maaike. Ik vond het enorm fijn samenwerken met jullie. Maaike, de samenwerking tussen ons is kort maar krachtig geweest, ik voelde me erg door je gesteund in de afrondingsprocedure van het proefschrift, hartelijk dank daarvoor.

I would also like to thank the many consortium members from CHARGE, MAGIC, DIAGRAM, GIANT and ENGAGE for the great, inspiring collaborations. I would like to thank in particular Prof. Bruce Psaty and the CHARGE RSC: thank you so much for giving me the fellowship grant to go to Boston for three months to work on the targeted sequencing data. This has been a fantastic experience! Also a very special thanks to Prof. James Meigs and Prof. Josée Dupuis, my supervisors in Boston, for giving me an amazing time. James and Josée, thank you so so much for all your help and support, not only during my months in Boston, but during the past years. James, since I visited your group in 2012 you have been a mentor for me and I'm really greatful for that. And a special thanks to Jennifer Wessel, Audrey Chu, Robert Scott and Mark Goodarzi for their vast contributions to the CHARGE Exome Chip paper included in this thesis. Jennifer, Audrey, Robert, Mark, thank you very much for the great team work and for giving me the opportunity to include the paper in my thesis.

Collega's van de genetische epidemiologie in Rotterdam, veel dank voor alle hulp, steun, prettige samenwerking en gezelligheid de afgelopen jaren. Om te beginnen met jou lieve Elisa. Dankjewel dat je er altijd voor me was, voor de vele fijne ritjes in de trein waarin we op kritische wijze de dag doornamen of je nog even een scriptje voor mij uit je mouw schudde, de vele lekkere mandarijntjes. Ik ben heel blij dat je nu mijn paranymf bent! Carla, jij was net begonnen toen ik startte op de afdeling en in de jaren die volgden hebben we hoogte- en dieptepunten en een grote liefde voor sushi en thee gedeeld. Veel succes en plezier in de kliniek en ik hoop dat je met volle teugen van je gezin aan het genieten bent! Linda, we zijn bijna mijn hele PhD periode kamergenootjes geweest en ik wil je hartelijk danken voor al je hulp en gezelligheid. Adriana and Claudia, we started the NIHES master at the same time and I'm so happy that you joined the group after you finalized the master. Adriana, it was so much fun exchanging Spanish and

Dutch words. The beautiful words 'primavera siempre' and 'camino mente' are still on my fridge! Claudia, good luck with finalizing your PhD, I hope to see you soon at your defence! Sven, jij was er altijd met hulp, raad en steun, tot mijn laatste minuut op de afdeling voor ik naar Cambridge vertrok. En met een goede dosis lol, hoewel het natuurlijk best een traumatische ervaring is als er opeens een sinterklaas voor je neus van de kast af wordt geschoten.. Dina, I'm very happy that you joined the group and thank you for the nice months as roommates. Najaf and Ayse, thank you so much for all your help and for many nice moments in restaurants, at conferences (Ayse, I really enjoyed Paris), the yearly barbecue (Najaf, the meat is amazing). Lennart en Maarten, dankjulliewel voor al jullie hulp, steun en jullie luisterend oor zowel op computergebied als daarbuiten. Ik hoop dat jullie het heel erg naar jullie zin hebben in de nieuwe banen!

Van de afdeling epidemiologie in bredere zin wil ik Abbas, Jory, Raluca en 'de oogmeisjes' in het bijzonder bedanken. Abbas, heel hartelijk dank voor je waardevolle input in verschillende papers in dit proefschrift. Jory, we deden tegelijk onze PhD en ik wil je heel erg bedanken voor de fijne gesprekken die we nu via de App voortzetten. Raluca, thank you so much for your great help with my first paper and the good memories from the course in Egmond aan Zee! Virginie, Gabriëlle en Henriëtte, bedankt voor alle gezelligheid!

Dear colleagues from the MRC Epidemiology unit, a special thank you to you as well! Nick, thank you so much for giving me the opportunity to continue the work in the same research field in such a great team. Robert, I'm really greatful for your support and the time you gave me to finalize this thesis. Ali, thank you very much for your suggestions for the cover. Baska, Sherly, Felix, Luca, Cathy, Laura, Clara, Sundas, Soren, Nicola, Jian'an, John, Jing'Hua, Ken, Simon, Rebecca, Adina and many others: thanks a lot for your support and kind words which were so helpful in finalizing the thesis.

Ik wil ook graag de collega's en ruiters van stichting Ariane de Ranitz bedanken. Ik heb enorm genoten van de maandagavonden met jullie. Deze altijd vrolijke en gezellige avonden vormden ultieme ontspanning naast het werken aan het proefschrift! Ik hoop jullie snel weer te zien!

En dan mijn lieve vrienden Nelleke, Anna, Marthe, Mariette, Paula, Lieselotte, Laura, Sabrine, Esther, Mark, Ruben, Roderick, Marieke, Harmen, Maarten, Peter, Trudelien. Ik vind dat ik jullie veel te weinig heb gezien de afgelopen tijd door alle drukte met het starten van de nieuwe baan in Cambridge en het afronden van het proefschrift. Eindelijk kunnen we nu de tot nu toe uitgestelde weekendjes Cambridge gaan plannen! Ik kijk daar enorm naar uit!

Anne en Sophie, lieve zussen, of het nu is als paranymph, om voor Timmie en Tommie te zorgen, met een peptalk als het even tegenzit of mooie herinneringen makend met Pasen in Kings College Chapel of te paard in de heuvels van Wales, ik weet dat jullie er altijd voor me zijn en dat is me ongelofelijk veel waard. Dankjulliewel!

Lieve Gonneke, lieve Kees, wat hebben jullie veel meegemaakt de afgelopen jaren. En toch ontbrak het niet aan vele fijne momenten, waarvoor ik heel dankbaar ben. Ik hoop dat we er nog vele zullen toevoegen. Gonneke, ik vind het heel jammer dat ik ons weekendje Cambridge steeds moest uitstellen, maar dat gaan we wat mij betreft snel inhalen!

Lieve pappa en mamma, lieve Koen, dit boekje draag ik op aan jullie. Pappa en mamma, ik ben zo dankbaar voor alle kansen die jullie me hebben gegeven, dat jullie altijd achter me staan en er altijd voor me zijn. Ik verheug me er heel erg op jullie Cambridge te laten zien!

Lieve Koen, wat een jaren. Jij hebt als geen ander de pieken en dalen meegemaakt. Ik ben zo blij dat we uit dalen zijn gekomen en dankbaar voor hoe je me op pieken hebt geholpen en voor alle onvergetelijk mooie herinneringen die we hebben gemaakt . Ik weet dat ik altijd op je terug kan vallen en dat maakt dat ik zorgeloos kansen durf te grijpen.

Terwijl ik dit alles opschrijf besef ik weer extra wat een geluksvogel ik ben met al deze mensen om me heen! Dankjulliewel ♥

```
                    CTGACTACGTA
            CTACTAGCACTGTACACGA            CTACGTACG
         CGTACGTAGTCTGATCAAAACGTA        GCTATACAGCTA
       CGTTGGTACCAGTACGTACGACTGCGA     GTACGTACGTACG
      GTGTGTCACACAGACACGTGATATAGACCAGATGACACACGTAC
      GTACGACGGCGCGGCGCGATATATATAAAAGCACGGACACTACGA
     ACGATGACTGACTACGTGCGCGCGTAGCTAGCTACGGATAGGACTA
     TACTGACTACGTACTACTAAAGCTGCTTTTACGTACGTACGGGCATC
     CATGTACTATCGCGGCGCGTAGACTGTACGACCAGACTGACTGACT
     GACTGACTGACGCGCGTACGTAGCTACTGTAGTACGTACGTACGTA
     TTCAGACTGACTGACTGAGATTACAGCTACGTAGTACTGACACGT
      TACGACGGCGCGGCGCGATATATATAAAAGCACGGACACTACGA
       GATGACTGACTACGTGCGCGCGTAGCTAGCTACGGATAGGACTA
       TGACTACGTACTACTAAAGCTGCTTTTACGTACGTACGGGCATC
        CTGCGCATGTACTATCGCGGCGCGTAGACTGTACGACCAGAC
         ACTGACGCGCGTACGTAGCTACTGTAGTACGTACGTACGTA
         ACTGACTGACTGAGATTACAGCTACGTAGTACTGACACGT
          GCATACGAGTCGAAGCGCGCAATATATATTATATCGGCC
           GTGACTACATCATGCTGACTACGGTCGCGCTCAACG
            TGACTACGTACGTATACGACGTACTGACGGCGCGC
             GCACTGTACACGACTGACTTACTAGCTACGTACG
              TGATCAAAACGTACGCGCGGCTATACAGCTA
               CGTACGACTGCGATACGTACGTACGTACG
                CGTGATATAGACCAGATGACACACGT
                 TATATATAAAAGCACGGACACT
                  GTAGCTAGCTACGGATA
                   CTTTTACGTACGT
                    ACGACCAG
                     TAGT
```

About the author

GSTCG
GACTGACTGC
CAACGACTGATC
GACTGTACGCGCTA
GATATAGACAGACGA
ACTGACTGACTGACTG
ACTCAGACTCGATGAC
GACATATAGCTACGTAC
TGACGTACGTACGTACG
GTACTACTACGGTACT
AAACGGGTGTGTGTC
ACTGACTGACTGACT
ACTCAGACTCGATGA
GACATATAGCTACG
TGACTGACTGAC
GTACTACTACG
AAACGGGTG
GCATGATGA
TACGTAC
GSTCC
GACT
CA

Sara Marie Willems was born in Leiden, the Netherlands, on July 24, 1984. In 2003 she completed her gymnasium (pre-university) education at the Stedelijk Gymnasium Leiden, Leiden, the Netherlands and started medical school at Utrecht University, Utrecht, the Netherlands. In 2010 she obtained her MD title and started the work presented in this PhD thesis under supervision of Prof. Cornelia van Duijn at the Genetic Epidemiology Unit, Department of Epidemiology, Erasmus Medical Center Rotterdam, the Netherlands. In 2012 she received a grant from the Cohorts for Heart and Aging Research in Genomic Epidemiology (CHARGE) Consortium to go to Boston for three months to collaborate with Framingham Heart Study researchers under supervision of Prof. James Meigs at the Division of General Internal Medicine, Massachusetts General Hospital, Boston, MA, USA and under supervision of Prof. Josée Dupuis at the Department of Biostatistics, Boston University School of Public Health, Boston, MA, USA. In August 2012 she obtained a Master of Science degree in Health Sciences, specialisation Genetic Epidemiology from the Netherlands Institute for Health Sciences (NIHES), Rotterdam, the Netherlands. She is currently working as a postdoctoral researcher at the MRC Epidemiology Unit, University of Cambridge School of Clinical Medicine, Cambridge, UK.

ATGACTACGTA
CTACTAGCACTGTACACGA   GCTACGTAC
CGTACGTAGTCTGATCAAAACGTA   GCTATACAGCTA
CGTTGGTACCAGTACGTACGACTGCGA   GTACGTACGTACG
TGTGTGTCACACAGACACGTGATATAGACCAGATGACACACGTAC
GTACGACGGCGCGGCGCGATATATATAAAAGCACGGACACTACGA
ACGATGACTGACTACGTGCGCGCGTAGCTAGCTACGGATAGGACTA
TACTGACTACGTACTACTAAAGCTGCTTTTACGTACGTACGGGCATC
CATGTACTATCGCGGCGCGTAGACTGTACGACCAGACTGACTGACT
GACTGACTGACGCGCGTACGTAGCTACTGTAGTACGTACGTACGTA
TTCAGACTGACTGACTGAGATTACAGCTACGTAGTACTGACACGT
TACGACGGCGCGGCGCGATATATATAAAAGCACGGACACTACGA
GATGACTGACTACGTGCGCGCGTAGCTAGCTACGGATAGGACTA
TGACTACGTACTACTAAAGCTGCTTTTACGTACGTACGGGCATC
CTGCGCATGTACTATCGCGGCGCGTAGACTGTACGACCAGAC
ACTGACGCGCGTACGTAGCTACTGTAGTACGTACGTACGTA
ACTGACTGACTGAGATTACAGCTACGTAGTACTGACACGT
GCATACGAGTCGAAGCGCGCAATATATATTATATCGGC
GTGACTACATCATGCTGACTACGGTCGCGCTCAACG
TGACTACGTACGTATACGACGTACTGACGGCGCG
GCACTGTACACGACTGACTTACTAGCTACGTACG
TGATCAAAACGTACGCGCGGCTATACAGCTA
CGTACGACTGCGATACGTACGTACGTACG
CGTGATATAGACCAGATGACACACGT
TATATATAAAAGCACGGACACT
GTAGCTAGCTACGGATA
CTTTTACGTACGT
ACGACCAG
TAGT

List of publications

1. **Willems SM**, Koekkoek PS, Kwee TC, van den Bosch MA
Diffusion-weighted MRI of the liver for early tumor response assessment: Promising technique but evidence is still lacking
Acta Oncol. (2010) 49:252-5

2. **Willems SM**, Mihaescu R, Sijbrands EJ, van Duijn CM, Janssens AC
A methodological perspective on genetic risk prediction studies in type 2 diabetes: recommendations for future research
Curr Diab Rep. (2011) 11:511-518

3. Dastani Z, Hivert MF, Timpson N, Perry JR, Yuan X, Scott RA, Henneman P, Heid IM, Kizer JR, Lyytikäinen LP, Fuchsberger C, Tanaka T, Morris AP, Small K, Isaacs A, Beekman M, Coassin S, Lohman K, Qi L, Kanoni S, Pankow JS, Uh HW, Wu Y, Bidulescu A, Rasmussen-Torvik LJ, Greenwood CM, Ladouceur M, Grimsby J, Manning AK, Liu CT, Kooner J, Mooser VE, Vollenweider P, Kapur KA, Chambers J, Wareham NJ, Langenberg C, Frants R, Willems-Vandijk K, Oostra BA, **Willems SM**, Lamina C, Winkler TW, Psaty BM, Tracy RP, Brody J, Chen I, Viikari J, Kähönen M, Pramstaller PP, Evans DM, St Pourcain B, Sattar N, Wood AR, Bandinelli S, Carlson OD, Egan JM, Böhringer S, van Heemst D, Kedenko L, Kristiansson K, Nuotio ML, Loo BM, Harris T, Garcia M, Kanaya A, Haun M, Klopp N, Wichmann HE, Deloukas P, Katsareli E, Couper DJ, Duncan BB, Kloppenburg M, Adair LS, Borja JB; DIAGRAM+ Consortium; MAGIC Consortium; GLGC Investigators; MuTHER Consortium, Wilson JG, Musani S, Guo X, Johnson T, Semple R, Teslovich TM, Allison MA, Redline S, Buxbaum SG, Mohlke KL, Meulenbelt I, Ballantyne CM, Dedoussis GV, Hu FB, Liu Y, Paulweber B, Spector TD, Slagboom PE, Ferrucci L, Jula A, Perola M, Raitakari O, Florez JC, Salomaa V, Eriksson JG, Frayling TM, Hicks AA, Lehtimäki T, Smith GD, Siscovick DS, Kronenberg F, van Duijn C, Loos RJ, Waterworth DM, Meigs JB, Dupuis J, Richards JB, Voight BF, Scott LJ, Steinthorsdottir V, Dina C, Welch RP, Zeggini E, Huth C, Aulchenko YS, Thorleifsson G, McCulloch LJ, Ferreira T, Grallert H, Amin N, Wu G, Willer CJ, Raychaudhuri S, McCarroll SA, Hofmann OM, Segrè AV, van Hoek M, Navarro P, Ardlie K, Balkau B, Benediktsson R, Bennett AJ, Blagieva R, Boerwinkle E, Bonnycastle LL, Boström KB, Bravenboer B, Bumpstead S, Burtt NP, Charpentier G, Chines PS, Cornelis M, Crawford G, Doney AS, Elliott KS, Elliott AL, Erdos MR, Fox CS, Franklin CS, Ganser M, Gieger C, Grarup N, Green T, Griffin S, Groves CJ, Guiducci C, Hadjadj S, Hassanali N, Herder C, Isomaa B, Jackson AU, Johnson PR, Jørgensen T, Kao WH, Kong A, Kraft P, Kuusisto J, Lauritzen T, Li M, Lieverse A, Lindgren CM, Lyssenko V, Marre M, Meitinger T, Midthjell K, Morken MA, Narisu N, Nilsson P, Owen KR, Payne F, Petersen AK, Platou C, Proença C, Prokopenko I, Rathmann W, Rayner NW, Robertson NR, Rocheleau G, Roden M, Sampson MJ, Saxena R, Shields BM, Shrader P, Sigurdsson G, Sparsø T, Strassburger K, Stringham HM, Sun Q, Swift AJ, Thorand B, Tichet J, Tuomi T, van Dam RM, van Haeften TW, van Herpt T, van Vliet-Ostaptchouk JV, Walters

GB, Weedon MN, Wijmenga C, Witteman J, Bergman RN, Cauchi S, Collins FS, Gloyn AL, Gyllensten U, Hansen T, Hide WA, Hitman GA, Hofman A, Hunter DJ, Hveem K, Laakso M, Morris AD, Palmer CN, Rudan I, Sijbrands E, Stein LD, Tuomilehto J, Uitterlinden A, Walker M, Watanabe RM, Abecasis GR, Boehm BO, Campbell H, Daly MJ, Hattersley AT, Pedersen O, Barroso I, Groop L, Sladek R, Thorsteinsdottir U, Wilson JF, Illig T, Froguel P, van Duijn CM, Stefansson K, Altshuler D, Boehnke M, McCarthy MI, Soranzo N, Wheeler E, Glazer NL, Bouatia-Naji N, Mägi R, Randall J, Elliott P, Rybin D, Dehghan A, Hottenga JJ, Song K, Goel A, Lajunen T, Doney A, Cavalcanti-Proença C, Kumari M, Timpson NJ, Zabena C, Ingelsson E, An P, O'Connell J, Luan J, Elliott A, McCarroll SA, Roccasecca RM, Pattou F, Sethupathy P, Ariyurek Y, Barter P, Beilby JP, Ben-Shlomo Y, Bergmann S, Bochud M, Bonnefond A, Borch-Johnsen K, Böttcher Y, Brunner E, Bumpstead SJ, Chen YD, Chines P, Clarke R, Coin LJ, Cooper MN, Crisponi L, Day IN, de Geus EJ, Delplanque J, Fedson AC, Fischer-Rosinsky A, Forouhi NG, Franzosi MG, Galan P, Goodarzi MO, Graessler J, Grundy S, Gwilliam R, Hallmans G, Hammond N, Han X, Hartikainen AL, Hayward C, Heath SC, Hercberg S, Hillman DR, Hingorani AD, Hui J, Hung J, Kaakinen M, Kaprio J, Kesaniemi YA, Kivimaki M, Knight B, Koskinen S, Kovacs P, Kyvik KO, Lathrop GM, Lawlor DA, Le Bacquer O, Lecoeur C, Li Y, Mahley R, Mangino M, Martínez-Larrad MT, McAteer JB, McPherson R, Meisinger C, Melzer D, Meyre D, Mitchell BD, Mukherjee S, Naitza S, Neville MJ, Orrù M, Pakyz R, Paolisso G, Pattaro C, Pearson D, Peden JF, Pedersen NL, Pfeiffer AF, Pichler I, Polasek O, Posthuma D, Potter SC, Pouta A, Province MA, Rayner NW, Rice K, Ripatti S, Rivadeneira F, Rolandsson O, Sandbaek A, Sandhu M, Sanna S, Sayer AA, Scheet P, Seedorf U, Sharp SJ, Shields B, Sigurðsson G, Sijbrands EJ, Silveira A, Simpson L, Singleton A, Smith NL, Sovio U, Swift A, Syddall H, Syvänen AC, Tönjes A, Uitterlinden AG, van Dijk KW, Varma D, Visvikis-Siest S, Vitart V, Vogelzangs N, Waeber G, Wagner PJ, Walley A, Ward KL, Watkins H, Wild SH, Willemsen G, Witteman JC, Yarnell JW, Zelenika D, Zethelius B, Zhai G, Zhao JH, Zillikens MC; DIAGRAM Consortium; GIANT Consortium; Global B Pgen Consortium, Borecki IB, Meneton P, Magnusson PK, Nathan DM, Williams GH, Silander K, Bornstein SR, Schwarz P, Spranger J, Karpe F, Shuldiner AR, Cooper C, Serrano-Ríos M, Lind L, Palmer LJ, Hu FB 1st, Franks PW, Ebrahim S, Marmot M, Kao WH, Pramstaller PP, Wright AF, Stumvoll M, Hamsten A; Procardis Consortium, Buchanan TA, Valle TT, Rotter JI, Penninx BW, Boomsma DI, Cao A, Scuteri A, Schlessinger D, Uda M, Ruokonen A, Jarvelin MR, Peltonen L, Mooser V, Sladek R; MAGIC investigators; GLGC Consortium, Musunuru K, Smith AV, Edmondson AC, Stylianou IM, Koseki M, Pirruccello JP, Chasman DI, Johansen CT, Fouchier SW, Peloso GM, Barbalic M, Ricketts SL, Bis JC, Feitosa MF, Orho-Melander M, Melander O, Li X, Li M, Cho YS, Go MJ, Kim YJ, Lee JY, Park T, Kim K, Sim X, Ong RT, Croteau-Chonka DC, Lange LA, Smith JD, Ziegler A, Zhang W, Zee RY, Whitfield JB, Thompson JR, Surakka I, Spector TD, Smit JH, Sinisalo J, Scott J, Saharinen J, Sabatti C, Rose LM, Roberts R, Rieder M, Parker AN, Pare G, O'Donnell CJ, Nieminen MS, Nickerson DA, Montgomery GW, McArdle W, Masson D, Martin NG, Marroni F, Lucas G, Luben R, Lokki ML, Lettre G, Launer LJ, Lakatta

EG, Laaksonen R, Kyvik KO, König IR, Khaw KT, Kaplan LM, Johansson Å, Janssens AC, Igl W, Hovingh GK, Hengstenberg C, Havulinna AS, Hastie ND, Harris TB, Haritunians T, Hall AS, Groop LC, Gonzalez E, Freimer NB, Erdmann J, Ejebe KG, Döring A, Dominiczak AF, Demissie S, Deloukas P, de Faire U, Crawford G, Chen YD, Caulfield MJ, Boekholdt SM, Assimes TL, Quertermous T, Seielstad M, Wong TY, Tai ES, Feranil AB, Kuzawa CW, Taylor HA Jr, Gabriel SB, Holm H, Gudnason V, Krauss RM, Ordovas JM, Munroe PB, Kooner JS, Tall AR, Hegele RA, Kastelein JJ, Schadt EE, Strachan DP, Reilly MP, Samani NJ, Schunkert H, Cupples LA, Sandhu MS, Ridker PM, Rader DJ, Kathiresan S
Novel loci for adiponectin levels and their influence on type 2 diabetes and metabolic traits: a multi-ethnic meta-analysis of 45,891 individuals
PLoS Genet. (2012) 8(3):e1002607

4. Manning AK, Hivert MF, Scott RA, Grimsby JL, Bouatia-Naji N, Chen H, Rybin D, Liu CT, Bielak LF, Prokopenko I, Amin N, Barnes D, Cadby G, Hottenga JJ, Ingelsson E, Jackson AU, Johnson T, Kanoni S, Ladenvall C, Lagou V, Lahti J, Lecoeur C, Liu Y, Martinez-Larrad MT, Montasser ME, Navarro P, Perry JR, Rasmussen-Torvik LJ, Salo P, Sattar N, Shungin D, Strawbridge RJ, Tanaka T, van Duijn CM, An P, de Andrade M, Andrews JS, Aspelund T, Atalay M, Aulchenko Y, Balkau B, Bandinelli S, Beckmann JS, Beilby JP, Bellis C, Bergman RN, Blangero J, Boban M, Boehnke M, Boerwinkle E, Bonnycastle LL, Boomsma DI, Borecki IB, Böttcher Y, Bouchard C, Brunner E, Budimir D, Campbell H, Carlson O, Chines PS, Clarke R, Collins FS, Corbatón-Anchuelo A, Couper D, de Faire U, Dedoussis GV, Deloukas P, Dimitriou M, Egan JM, Eiriksdottir G, Erdos MR, Eriksson JG, Eury E, Ferrucci L, Ford I, Forouhi NG, Fox CS, Franzosi MG, Franks PW, Frayling TM, Froguel P, Galan P, de Geus E, Gigante B, Glazer NL, Goel A, Groop L, Gudnason V, Hallmans G, Hamsten A, Hansson O, Harris TB, Hayward C, Heath S, Hercberg S, Hicks AA, Hingorani A, Hofman A, Hui J, Hung J, Jarvelin MR, Jhun MA, Johnson PC, Jukema JW, Jula A, Kao WH, Kaprio J, Kardia SL, Keinanen-Kiukaanniemi S, Kivimaki M, Kolcic I, Kovacs P, Kumari M, Kuusisto J, Kyvik KO, Laakso M, Lakka T, Lannfelt L, Lathrop GM, Launer LJ, Leander K, Li G, Lind L, Lindstrom J, Lobbens S, Loos RJ, Luan J, Lyssenko V, Mägi R, Magnusson PK, Marmot M, Meneton P, Mohlke KL, Mooser V, Morken MA, Miljkovic I, Narisu N, O'Connell J, Ong KK, Oostra BA, Palmer LJ, Palotie A, Pankow JS, Peden JF, Pedersen NL, Pehlic M, Peltonen L, Penninx B, Pericic M, Perola M, Perusse L, Peyser PA, Polasek O, Pramstaller PP, Province MA, Räikkönen K, Rauramaa R, Rehnberg E, Rice K, Rotter JI, Rudan I, Ruokonen A, Saaristo T, Sabater-Lleal M, Salomaa V, Savage DB, Saxena R, Schwarz P, Seedorf U, Sennblad B, Serrano-Rios M, Shuldiner AR, Sijbrands EJ, Siscovick DS, Smit JH, Small KS, Smith NL, Smith AV, Stančáková A, Stirrups K, Stumvoll M, Sun YV, Swift AJ, Tönjes A, Tuomilehto J, Trompet S, Uitterlinden AG, Uusitupa M, Vikström M, Vitart V, Vohl MC, Voight BF, Vollenweider P, Waeber G, Waterworth DM, Watkins H, Wheeler E, Widen E, Wild SH, **Willems SM**, Willemsen G, Wilson JF, Witteman JC, Wright AF, Yaghootkar H, Zelenika D, Zemunik

T, Zgaga L; DIAbetes Genetics Replication And Meta-analysis (DIAGRAM) Consortium; Multiple Tissue Human Expression Resource (MUTHER) Consortium, Wareham NJ, McCarthy MI, Barroso I, Watanabe RM, Florez JC, Dupuis J, Meigs JB, Langenberg C

A genome-wide approach accounting for body mass index identifies genetic variants influencing fasting glycemic traits and insulin resistance

Nat Genet. 2012 May 13;44(6):659-69

5. Scott RA, Lagou V, Welch RP, Wheeler E, Montasser ME, Luan J, Mägi R, Strawbridge RJ, Rehnberg E, Gustafsson S, Kanoni S, Rasmussen-Torvik LJ, Yengo L, Lecoeur C, Shungin D, Sanna S, Sidore C, Johnson PC, Jukema JW, Johnson T, Mahajan A, Verweij N, Thorleifsson G, Hottenga JJ, Shah S, Smith AV, Sennblad B, Gieger C, Salo P, Perola M, Timpson NJ, Evans DM, Pourcain BS, Wu Y, Andrews JS, Hui J, Bielak LF, Zhao W, Horikoshi M, Navarro P, Isaacs A, O'Connell JR, Stirrups K, Vitart V, Hayward C, Esko T, Mihailov E, Fraser RM, Fall T, Voight BF, Raychaudhuri S, Chen H, Lindgren CM, Morris AP, Rayner NW, Robertson N, Rybin D, Liu CT, Beckmann JS, **Willems SM**, Chines PS, Jackson AU, Kang HM, Stringham HM, Song K, Tanaka T, Peden JF, Goel A, Hicks AA, An P, Müller-Nurasyid M, Franco-Cereceda A, Folkersen L, Marullo L, Jansen H, Oldehinkel AJ, Bruinenberg M, Pankow JS, North KE, Forouhi NG, Loos RJ, Edkins S, Varga TV, Hallmans G, Oksa H, Antonella M, Nagaraja R, Trompet S, Ford I, Bakker SJ, Kong A, Kumari M, Gigante B, Herder C, Munroe PB, Caulfield M, Antti J, Mangino M, Small K, Miljkovic I, Liu Y, Atalay M, Kiess W, James AL, Rivadeneira F, Uitterlinden AG, Palmer CN, Doney AS, Willemsen G, Smit JH, Campbell S, Polasek O, Bonnycastle LL, Hercberg S, Dimitriou M, Bolton JL, Fowkes GR, Kovacs P, Lindström J, Zemunik T, Bandinelli S, Wild SH, Basart HV, Rathmann W, Grallert H; DIAbetes Genetics Replication and Meta-analysis (DIAGRAM) Consortium, Maerz W, Kleber ME, Boehm BO, Peters A, Pramstaller PP, Province MA, Borecki IB, Hastie ND, Rudan I, Campbell H, Watkins H, Farrall M, Stumvoll M, Ferrucci L, Waterworth DM, Bergman RN, Collins FS, Tuomilehto J, Watanabe RM, de Geus EJ, Penninx BW, Hofman A, Oostra BA, Psaty BM, Vollenweider P, Wilson JF, Wright AF, Hovingh GK, Metspalu A, Uusitupa M, Magnusson PK, Kyvik KO, Kaprio J, Price JF, Dedoussis GV, Deloukas P, Meneton P, Lind L, Boehnke M, Shuldiner AR, van Duijn CM, Morris AD, Toenjes A, Peyser PA, Beilby JP, Körner A, Kuusisto J, Laakso M, Bornstein SR, Schwarz PE, Lakka TA, Rauramaa R, Adair LS, Smith GD, Spector TD, Illig T, de Faire U, Hamsten A, Gudnason V, Kivimaki M, Hingorani A, Keinanen-Kiukaanniemi SM, Saaristo TE, Boomsma DI, Stefansson K, van der Harst P, Dupuis J, Pedersen NL, Sattar N, Harris TB, Cucca F, Ripatti S, Salomaa V, Mohlke KL, Balkau B, Froguel P, Pouta A, Jarvelin MR, Wareham NJ, Bouatia-Naji N, McCarthy MI, Franks PW, Meigs JB, Teslovich TM, Florez JC, Langenberg C, Ingelsson E, Prokopenko I, Barroso I

Large-scale association analyses identify new loci influencing glycemic traits and provide insight into the underlying biological pathways

Nat Genet. 2012 Sep;44(9):991-1005

6. van der Loos MJ, Rietveld CA, Eklund N, Koellinger PD, Rivadeneira F, Abecasis GR, Ankra-Badu GA, Baumeister SE, Benjamin DJ, Biffar R, Blankenberg S, Boomsma DI, Cesarini D, Cucca F, de Geus EJ, Dedoussis G, Deloukas P, Dimitriou M, Eiriksdottir G, Eriksson J, Gieger C, Gudnason V, Höhne B, Holle R, Hottenga JJ, Isaacs A, Järvelin MR, Johannesson M, Kaakinen M, Kähönen M, Kanoni S, Laaksonen MA, Lahti J, Launer LJ, Lehtimäki T, Loitfelder M, Magnusson PK, Naitza S, Oostra BA, Perola M, Petrovic K, Quaye L, Raitakari O, Ripatti S, Scheet P, Schlessinger D, Schmidt CO, Schmidt H, Schmidt R, Senft A, Smith AV, Spector TD, Surakka I, Svento R, Terracciano A, Tikkanen E, van Duijn CM, Viikari J, Völzke H, Wichmann HE, Wild PS, **Willems SM**, Willemsen G, van Rooij FJ, Groenen PJ, Uitterlinden AG, Hofman A, Thurik AR
The molecular genetic architecture of self-employment
PLoS One. 2013 Apr 4;8(4):e60542

7. Isaacs A*, **Willems SM***, Bos D, Dehghan A, Hofman A, Ikram MA, Uitterlinden AG, Oostra BA, Franco OH, Witteman JC, van Duijn CM
Risk scores of common genetic variants for lipid levels influence atherosclerosis and incident coronary heart disease
Arterioscler Thromb Vasc Biol. 2013 Sep;33(9):2233-9

8. Fall T, Hägg S, Mägi R, Ploner A, Fischer K, Horikoshi M, Sarin AP, Thorleifsson G, Ladenvall C, Kals M, Kuningas M, Draisma HH, Ried JS, van Zuydam NR, Huikari V, Mangino M, Sonestedt E, Benyamin B, Nelson CP, Rivera NV, Kristiansson K, Shen HY, Havulinna AS, Dehghan A, Donnelly LA, Kaakinen M, Nuotio ML, Robertson N, de Bruijn RF, Ikram MA, Amin N, Balmforth AJ, Braund PS, Doney AS, Döring A, Elliott P, Esko T, Franco OH, Gretarsdottir S, Hartikainen AL, Heikkilä K, Herzig KH, Holm H, Hottenga JJ, Hyppönen E, Illig T, Isaacs A, Isomaa B, Karssen LC, Kettunen J, Koenig W, Kuulasmaa K, Laatikainen T, Laitinen J, Lindgren C, Lyssenko V, Läärä E, Rayner NW, Männistö S, Pouta A, Rathmann W, Rivadeneira F, Ruokonen A, Savolainen MJ, Sijbrands EJ, Small KS, Smit JH, Steinthorsdottir V, Syvänen AC, Taanila A, Tobin MD, Uitterlinden AG, **Willems SM**, Willemsen G, Witteman J, Perola M, Evans A, Ferrières J, Virtamo J, Kee F, Tregouet DA, Arveiler D, Amouyel P, Ferrario MM, Brambilla P, Hall AS, Heath AC, Madden PA, Martin NG, Montgomery GW, Whitfield JB, Jula A, Knekt P, Oostra B, van Duijn CM, Penninx BW, Davey Smith G, Kaprio J, Samani NJ, Gieger C, Peters A, Wichmann HE, Boomsma DI, de Geus EJ, Tuomi T, Power C, Hammond CJ, Spector TD, Lind L, Orho-Melander M, Palmer CN, Morris AD, Groop L, Järvelin MR, Salomaa V, Vartiainen E, Hofman A, Ripatti S, Metspalu A, Thorsteinsdottir U, Stefansson K, Pedersen NL, McCarthy MI, Ingelsson E, Prokopenko I; European Network for Genetic and Genomic Epidemiology (ENGAGE) consortium
The role of adiposity in cardiometabolic traits: a Mendelian randomization analysis
PLoS Med. 2013;10(6):e1001474

9. Pirastu N, Kooyman M, Traglia M, Robino A, **Willems SM**, Pistis G, d'Adamo P, Amin N, d'Eustacchio A, Navarini L, Sala C, Karssen LC, van Duijn C, Toniolo D, Gasparini P
Association analysis of bitter receptor genes in five isolated populations identifies a significant correlation between TAS2R43 variants and coffee liking
PLoS One. 2014 Mar 19;9(3):e92065

ATGACTACGTA
CTACTAGCACTGTACACGA   GCTACGTACG
CGTACGTAGTCTGATCAAAACGTAC   GCTATACAGCTA
CGTTGGTACCAGTACGTACGACTGCGA   GTACGTACGTACG
TGTGTGTCACACAGACACGTGATATAGACCAGATGACACACGTAC
GTACGACGGCGCGGCGCGATATATATAAAAGCACGGACACTACGA
ACGATGACTGACTACGTGCGCGCGTAGCTAGCTACGGATAGGACTA
TACTGACTACGTACTACTAAAGCTGCTTTTACGTACGTACGGGCATC
CATGTACTATCGCGGCGCGTAGACTGTACGACCAGACTGACTGACT
GACTGACTGACGCGCGTACGTAGCTACTGTAGTACGTACGTACGTA
TTCAGACTGACTGACTGAGATTACAGCTACGTAGTACTGACACGT
TACGACGGCGCGGCGCGATATATATAAAAGCACGGACACTACGA
GATGACTGACTACGTGCGCGCGTAGCTAGCTACGGATAGGACTA
TGACTACGTACTACTAAAGCTGCTTTTACGTACGTACGGGCATC
CTGCGCATGTACTATCGCGGCGCGTAGACTGTACGACCAGAC
ACTGACGCGCGTACGTAGCTACTGTAGTACGTACGTACGTA
ACTGACTGACTGAGATTACAGCTACGTAGTACTGACACGT
GCATACGAGTCGAAGCGCGCAATATATATTATATCGGCC
CGTGACTACATCATGCTGACTACGGTCGCGCTCAACG
TGACTACGTACGTATACGACGTACTGACGGCGCGC
GCACTGTACACGACTGACTTACTAGCTACGTACG
TGATCAAAACGTACGCGCGGCTATACAGCTA
CGTACGACTGCGATACGTACGTACGTACG
CGTGATATAGACCAGATGACACACGT
TATATATAAAAGCACGGACACT
GTAGCTAGCTACGGATA
CTTTTACGTACGT
ACGACCAG
TAGT

8.4

PhD portfolio summary

Erasmus MC
Universitair Medisch Centrum Rotterdam

SUMMARY OF PhD TRAINING AND TEACHING ACTIVITIES

Name PhD student: Sara Willems	PhD period: 2010-2014
Erasmus MC Department: Epidemiology	Promotor(s): Prof. dr. Cornelia M. van Duijn
Research School: NIHES	Supervisor: Dr. Aaron Isaacs

1. PhD training

	Year	Workload (ECTS)*
In-depth courses (e.g. Research school, Medical Training)		
NIHES Master of Science in Health Sciences	2010-2012	
- Study Design		4.3
- Classical Methods for Data-analysis		5.7
- Modern Statistical methods		4.3
- Genetic-Epidemiologic Research Methods		5.7
- SNPs and Human Diseases		1.4
- Courses for the Quantitative Researcher		1.4
- Introduction to Clinical and Public health Genomics		1.4
- Advances in Genome-Wide Association Studies		1.4
- Family-based Genetic Analysis		1.4
- Principles of Research in Medicine		0.7
- Clinical Decision Analysis		0.7
- Topics in Meta-analysis		0.7
- Health Economics		0.7
- Genome Wide Association Analysis		1.4
- Conceptual Foundation of Epidemiologic Study Design		0.7
- Principles of Genetic Epidemiology		0.7
- Primary and Secondary Prevention Research		0.7
- Genomics in Molecular Medicine		1.4
- Markers and Prognostic Research		0.7
- Advances in Genomics Research		0.4
- ENGAGE Summer Institute "Genetics, Ethics and Clinical Translation", Egmond aan Zee, the Netherlands	2010	1

Presentations

International conferences and meetings

- European Human Genetics Conference, Amsterdam, the Netherlands: "*Accumulation of common lipid variants influences atherosclerosis, and incident cardiovascular disease*" (oral)	2011	1
- CHARGE investigator meeting, Reykjavik, Iceland: "*Association of the IGF1 gene with fasting insulin levels*" (poster)	2012	0.5
- ENGAGE investigator meeting, Rotterdam, the Netherlands: "*Association of type 2 diabetes with plasma sphingomyelin, phosphatidylcholine, ceramide and phosphatidylethanolamine concentrations*" (poster)	2012	0.5
- CHARGE investigator meeting, Houston, USA: "*Association of the IGF1 gene with fasting insulin levels*" (poster)	2012	0.5
- European Human Genetics Conference, Paris, France: "*Risk scores derived from known lipid variants improve prediction of hypercholesterolemia*" (poster)	2013	0.5
- CHARGE investigator meeting, Rotterdam, the Netherlands: "*Association of the IGF1 gene with fasting insulin levels*" (oral)	2013	1
- Framingham Heart Study Scientific Retreat, Waltham, MA, USA: "*Association of the IGF1 gene with fasting insulin levels*" (poster)	2012	0.5

Oral presentations at lab meetings

- Genetic Epidemiology Unit, Erasmus MC, Rotterdam, the Netherlands	2010-2013	3
- Division of General Internal Medicine, Massachusetts General Hospital, Boston, MA, USA	2012	0.5
- Department of Biostatistics, Boston University School of Public Health, Boston, MA, USA	2012	0.5
- Department of Epidemiology and Population Health, Albert Einstein College of Medicine, New York, USA	2012	0.5

Seminars, symposia and workshops

- Regular seminars at the Department of Epidemiology, Erasmus MC, Rotterdam, the Netherlands	2010-2013	1.5
- Annual Centre for Medical Systems Biology Symposium	2011-2013	0.7
- Regular seminars from the Program in Medical and Population Genetics, Broad Institute of MIT and Harvard, Cambridge, MA, USA	March-June 2012	0.3

Other

- Reviewer of papers for international journals	2011-2014	1
- Radio interview following press release of the project "*Accumulation of common lipid variants influences atherosclerosis, and incident cardiovascular disease*"	2011	0.3
- Research fellow at the Division of General Internal Medicine, Massachusetts General Hospital and the Department of Biostatistics, Boston University School of Public Health, Boston, MA, USA, supervised by Prof. James Meigs and Prof. Josée Dupuis and supported by a grant from the CHARGE consortium	March-June 2012	3 months

2. Teaching activities		
- Teaching assistant for the NIHES Summer Programme course "Principles of Genetic Epidemiology"	2012-2013	1.5

3. Other		
Creutzfeldt Jakob disease registry		
- Maintenance of the Dutch CJD registration, patient visits and inclusion in biobank, information service	2010-2013	5

*1 ECTS (European Credit Transfer System) equals a workload of 28 hours